Political Analysis (2004) 12:400–424
doi:10.1093/pan/mph026

What Do We Learn from Graduate Admissions Committees? A Multiple Rater, Latent Variable Model, with Incomplete Discrete and Continuous Indicators

Simon Jackman
Department of Political Science, Stanford University, Stanford, CA 94305-6044
e-mail: jackman@leland.stanford.edu

What do we really know about applicants to graduate school? How much information is in an applicant's file? What do we learn by having graduate admissions committees read and score applicant files? In this article, I develop a statistical model for measuring applicant quality, combining the information in the committee members' ordinal ratings with the information in applicants' GRE scores. The model produces estimates of applicant quality purged of the influence of committee members' preferences over ostensibly extraneous applicant characteristics, such as gender and intended field of study. An explicitly Bayesian approach is adopted for estimation and inference, making it straightforward to obtain confidence intervals not only on latent applicant quality but over rank orderings of applicants and the probability of belonging in a set of likely admittees. Using data from applications to a highly ranked political science graduate program, I show that there is considerable uncertainty in estimates of applicant quality, making it impossible to make authoritative distinctions as to quality among large portions of the applicant pool. The multiple rater model I develop here is extremely flexible and has applications in fields as diverse as judicial politics, legislative politics, international relations, and public opinion.

1 Introduction

Almost all readers of this article have been through the process of applying to graduate school. For some readers, this happy event was not so long ago. For others, myself included, applying to graduate school only *seems* as though it was not so long ago, and we're actually much closer to a different process: serving on graduate admissions committees and therefore deciding who gets into graduate school. As academic committee assignments go, graduate admissions is an odd one. Different colleagues bring different criteria to bear on the task. How does one weigh test scores, undergraduate records, writing samples, letters of recommendation, and the intended field of study nominated by the applicant? How does the committee divide the workload among its members? How does the committee aggregate the opinions of its members? And, since we are political scientists, to what extent are these procedures open to manipulation by committee members attempting to have their preferred applicants admitted?

Author's note: Tim Büthe, Jim Fearon, Doug Rivers, Mike Tomz, Jonathan Wand, and participants at the American Politics seminar at Harvard University provided useful comments and suggestions; the standard disclaimer applies.

Political Analysis, Vol. 12 No. 4, © Society for Political Methodology 2004; all rights reserved.

I have served on graduate admissions committees multiple times and at multiple universities. Each time the process has taken place in stages. There is always a first cut of sorts, reducing the committee's burden from the initial set of hundreds of files to a more manageable set of files that will receive longer and more serious consideration. The (usually tacit) justification for the multistage procedure is that the task of selecting admittees is easier if there are fewer alternatives on the table: choosing 25 admittees from 50 "finalists" is easier than choosing 25 from 300 (or more). However, having first selected 50 finalists from the entire applicant pool, the task of then selecting 25 admittees is often hard going. My experience is that committees often struggle to discriminate among finalists, there simply being insufficient information in the files with which to make reliable judgments as to whom to admit. All 50 finalists look good. All look as though they could do well in one's Ph.D. program. To paraphrase a colleague, "We might as well toss these files down the stairs and admit the first 25 we pick up".

In this article I assess the extent to which we can reliably distinguish among applicants for graduate school. I am fortunate to have real graduate admissions data. I use a latent variable model and Bayesian statistical procedures to attempt to rigorously assess how much we learn about any given applicant, relative to other applicants. Although this article focuses on the task of graduate admissions, the statistical methodology I employ and elaborate here is quite general and is applicable to a wide range of measurement problems encountered in the social sciences.

2 Multiple Rater Data in Political Science

Analysis of the graduate admissions data poses methodological problems of more general interest. As we shall see below, graduate admissions data are usefully considered as a form of multiple rater data: each subject $i = 1, \ldots, n$ is assigned scores by raters $j = 1, \ldots, m$, and usually $m \ll n$. Examples in political science include:

- country experts assigning scores to the platforms of various political parties (e.g., Castles and Mair 1984; Laver and Hunt 1992; Huber and Inglehart 1995)
- judges hearing cases (e.g., Martin and Quinn 2002)
- legislators voting on roll calls (Poole and Rosenthal 1997; Clinton et al. 2004)
- survey respondents rating political entities (in this case, $m \gg n$) (Aldrich and McKelvey 1977; Erikson 1990; Franklin 1991)
- newspaper editorials rating political entities (e.g., Noel 2004)
- experts rating the importance of legislation (Clinton and Lapinski 2004).

The data generated by multiple rater designs may also present other challenges for statistical analysis. First, there is no guarantee that the ratings will be continuous variables of a sort amenable to correlational analysis (e.g., factor analysis); ratings data may be binary, ordinal, nominal, or continuous indicators of a latent trait. Second, the data generated by multiple rater designs are not guaranteed to be balanced: in other words, not all raters rate all entities. This feature of the data is especially pernicious in the case of graduate admissions data, where, by design, any one committee member might see only a third of the files in the applicant pool. But many of the examples discussed above have this feature: e.g., not all judges hear all cases,[1] not all legislators vote on all roll calls, not

[1] For instance, in a study of asylum decisions heard by three-judge appeals panels drawn from the U.S. Ninth Circuit (Law 2004), overlap among raters (judges) is the exception, not the norm.

all editorial writers offer opinions about all political entities, not all experts rate every political party in every country. In many situations, it is not at all uncommon for there to be no overlap whatsoever between pairs of raters: e.g., none of the experts on country *x* rated parties from country *y*. Many of the methodological problems I confront in assessing graduate admissions data crop up in an array of measurement problems, spanning a broad range of contemporary political science. In short, although my substantive focus here is the graduate admissions problem, the methodological tools clearly have applications in real political science settings.

3 Graduate Admissions Data and Context

Graduate admissions data are sensitive, and so all I will say is that these data are for admissions to Stanford's political science Ph.D. program. Data that would greatly increase the chances that a particular applicant could be identified are not available for the analysis I report here: these include undergraduate GPA, undergraduate institution, date of birth, and race and ethnicity.

The program received 279 applications. For each applicant we have GRE scores, intended field of study (American politics, comparative politics, international relations, political theory), and indicators of gender and U.S. residence. Of course, data available to the committee (but unavailable for my analysis) include everything else one sees in an applicant's file: the undergraduate institution and transcript (including classes taken and grades received), records of any graduate-level studies or work experience, letters of recommendation, writing samples, and statements of purpose.

The graduate admissions committee consisted of five faculty members (F1–F5) and three graduate students (S1–S3). The workload of reading the files was divided among the committee members. Not every committee member read every file, at least not in the initial pass through the applicant pool that I analyze. The workload was staggered across the committee such that almost every file was read by at least two faculty and one graduate student. Committee members were asked to give a one (low) through five (high) rating to each file; only integer scores were recorded. Sensitive to the fact that different committee members might use the five-point scale differently, the committee aggregated the 1–5 scores via standardization, normalizing each rater's scores to have mean zero and unit variance. The applicants were then ranked on the sum of the normalized scores from their three respective readers. The top 56 applicants on this summary measure were designated as "semifinalists," of whom approximately 30 were ultimately offered admission to the Ph.D. program.

4 Preliminary Data Analysis

Table 1 shows the correlation matrix and descriptive statistics for the 1–5 scores assigned by the eight raters (F1–F5 and S1–S3) and the three GRE scores. The correlations are ordinary Pearson correlations and are computed in pairwise fashion. Staggering the workload across the committee means that the judges were rating disjoint subsets of the applicant pool (e.g., F1 and F3 saw different sets of applicants), and correlations can not be computed in these instances. Factor analytic procedures are therefore difficult to implement in this context: we have no regular correlation matrix to factor analyze, and the latent variable model and methods I employ are more appropriate for this context.

Ordinarily if two judges are rating different objects, there is no way to compare their ratings: if I judge apples and you judge oranges, the fact that we might be allocating 1–5 scores on a "tastiness" scale in different ways (problem one) is confounded with the fact

Table 1 Pairwise correlation matrix and descriptive statistics, committee ratings (1–5) and GRE scores

	Correlation matrix										
	F1	*F2*	*F3*	*F4*	*F5*	*S1*	*S2*	*S3*	*Vrb*	*Qnt*	*Anl*
F1	1.00	.69			.23	.81	.54		.46	.65	.53
F2	.69	1.00	.55			.68		.72	.58	.46	.50
F3		.55	1.00	.75			.65	.65	.54	.52	.50
F4			.75	1.00	.46		.62	.68	.41	.44	.50
F5	.23			.46	1.00		.43	.53	.32	.11	.22
S1	.81	.68				1.00			.50	.74	.72
S2	.54		.65	.62	.43		1.00		.69	.30	.45
S3		.72	.65	.68	.53			1.00	.38	.42	.36
Vrb	.46	.58	.54	.41	.32	.50	.69	.38	1.00	.27	.46
Qnt	.65	.46	.52	.44	.11	.74	.30	.42	.27	1.00	.44
Anl	.53	.50	.50	.50	.22	.72	.45	.36	.46	.44	1.00

	Descriptive statistics										
	F1	*F2*	*F3*	*F4*	*F5*	*S1*	*S2*	*S3*	*Vrb*	*Qnt*	*Anl*
Mean	2.99	3.25	3.32	3.00	2.44	2.69	2.95	3.37	604	680	658
StdDev	1.50	1.00	1.24	1.37	1.24	1.37	1.08	1.24	111	91	109
5%	1	2	1	1	1	1	1	1	400	510	480
95%	5	5	5	5	4.45	5	5	5	750	790	800
n	111	110	111	112	112	52	107	106	264	264	264

Note. Inter-rater correlations are Spearman correlations; correlations between the ordinal ratings and GRE scores are polyserial correlations (Cox 1974; Olsson et al. 1982); correlations between the GRE scores are Pearon correlations. Vrb = Verbal, Qnt = Quantitative, Anl = Analytic.

that we are rating different things (problem two). We get around problem two in the current context via the overlap in the data: F1 and F3 do not rate any of the same applicants, but F1 and F2 do, and F2 rates some of the files that F3 did, and so on. An analogy comes from the literature on measuring preferences in Congress over time: Paul Wellstone served in the U.S. Senate from 1991 until his death in 2002 and did not serve with Barry Goldwater (1953–1965 and 1969–1987), but the terms of Ted Kennedy (1962–present), Strom Thurmond (1956–2003), and Robert Byrd (1959–present) overlap both Wellstone's and Goldwater's terms. Subject to some identifying restrictions on how often legislators' preferences change, the overlapping generations structure of the data lets us compare legislators with temporally disjoint terms. Note also that GRE scores are available for all but 15 applicants, providing additional leverage on our ability to jointly assess all applicants, even though not every committee member reads every file (indeed, one might argue that that is the very point of GRE scores).

The summaries reported in Table 1 reveal some interesting substantive features of the data. It is clear that F5 is something of a maverick: F5's scores display low to weak Spearman correlations with the other raters (.23 to .53) and with the GRE scores (from .11 for the quantitative score to .32 for the verbal score). The other inter-rater Spearman correlations are in the range of .54 to .81, and the other raters have polyserial correlations with the GRE scores ranging from .30 to .74. Nonetheless, this general pattern of correlation provides support for two propositions that will be helpful in the statistical modeling I report below.

First, applicant quality can be considered a unidimensional trait. Typically one would make this judgment via an eigen decomposition of the inter-rater correlation matrix (e.g., a large first eigenvalue and all other eigenvalues being small). But with these non-overlapping ratings data, the inter-rater correlation matrix is incomplete and cannot be decomposed (as shown in Table 1). Moreover, my goal here is not primarily to model the observed ratings so as to maximize goodness of fit: clearly, adding a second or third dimension to the latent trait will improve goodness of fit. However, using the same data to assign not just one score to an applicant, but two or more, will result in a substantial increase in imprecision. Each applicant supplies a limited amount of information: the three ratings, three GRE scores, and some other information such as gender and intended field of study (presumably not related to applicant quality). Given this limited information, trying to learn two or more things about each applicant (i.e., each applicant's score on multiple dimensions) as opposed to one will result in a considerable loss of precision in estimates of applicant quality. Perhaps most important, the committee's procedures suggest that the members had in mind a unidimensional concept of applicant quality: they themselves reduced all the available information for each applicant to a single number and rank ordered the applicants on this summary measure. My own experience on graduate admissions committees suggests that while there are different indicators of applicant quality, the underlying construct is itself unidimensional: seldom did the committees I have served on struggle with an applicant who, say, showed excellent quantitative ability but poor writing skills (such an applicant would be considered an applicant of low to moderate quality). Moreover, I am primarily interested in questions of reliability, i.e., how much uncertainty attaches to the measure of applicant quality we generate using a more rigorous and formal procedure than that used by the committee, and how confident we might be in a rank ordering of the applicants. I am less concerned in defining applicant quality or the validity of a unidimensional measure. As we shall see, there is considerable uncertainty in estimates of applicant quality that result from a unidimensional operationalization; we would have even less precision if we operationalized applicant quality as a multidimensional latent trait.

Second, the GRE scores are informative about applicant quality. The polyserial correlations in Table 1 between GRE scores and ratings range from .11 (between F5's ratings and quantitative GRE score) to .74 (between S1's ratings and quantitative GRE score). The median of the 24 polyserial correlations between rating and GRE scores is .48. These results suggest that as a general proposition, the raters seem to be basing their assessments of candidate quality on something that is also being tapped by the GRE scores. To be sure, the degree to which GRE scores and the ratings are determined by applicant quality may well vary over the components of the GRE and across raters. But it seems that modeling both GRE scores and the ratings as functions of the same latent trait (applicant quality) is a reasonable strategy.

Additional preliminary analysis appears in Figure 1, showing the distribution of ratings by rater. Quite different patterns of ratings are evident. F1 gave just two "3" ratings, producing an approximately bimodal distribution of ratings, while F2 gave just two "1" ratings. F5's ratings have a pronounced skew to the right: F5's modal rating is a "1" and F5 also issued just six "5" ratings. This variation in the marginal distribution of ratings strongly suggests that the raters used the five-point rating scale differently. Indeed, there is some evidence to suggest some strategic behavior by the raters. It can be reasonably assumed that the raters knew the aggregation mechanism that would be applied to their ratings, the first step being to convert their ratings to z scores; raters looking to increase the impact of their high ratings given this rule would seek to give lots of low scores (driving

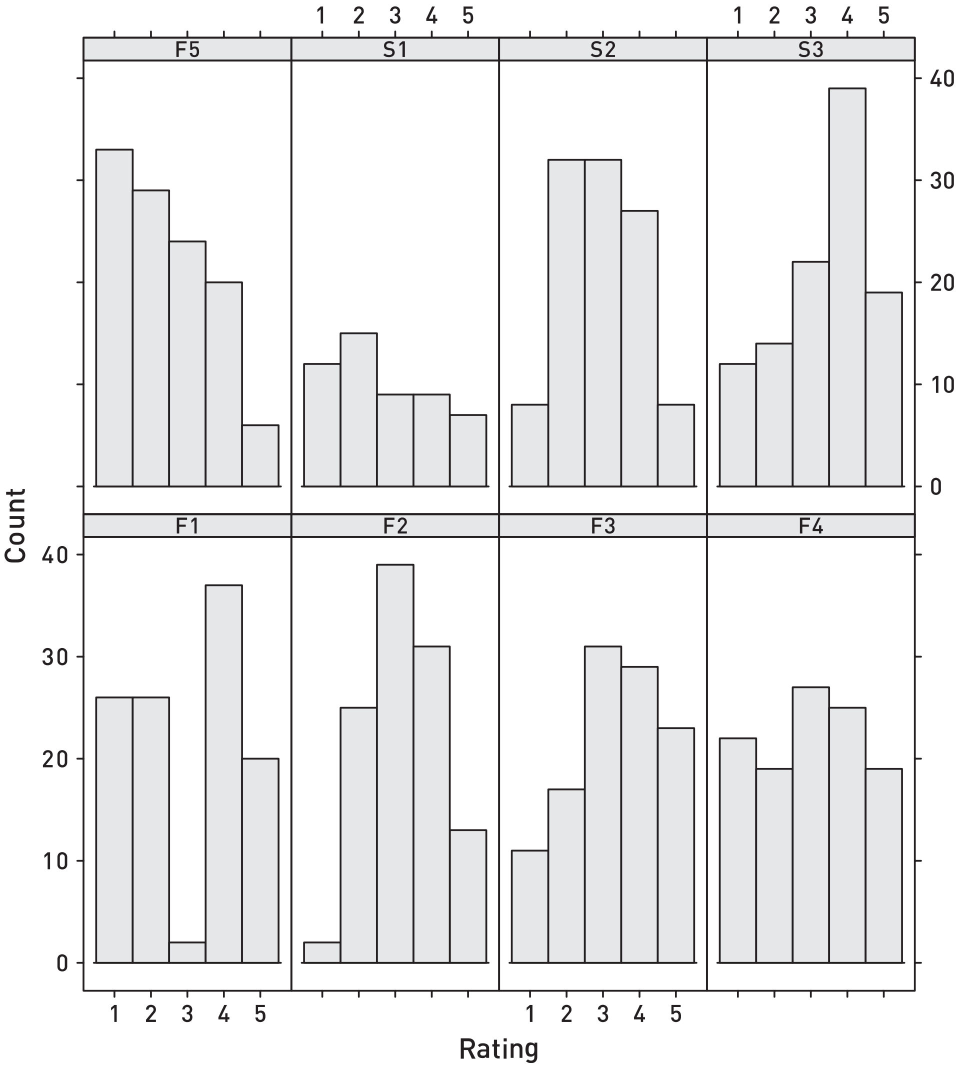

Fig. 1 Histograms of ratings, by rater.

down their mean) and save the "5" rating for their most preferred applicants. F5's "5" rating generated a z score of 2.07, the highest such z score in the data. F1's bimodal pattern of ratings is consistent with trying to maximize influence on outcomes: although all raters' z scores sum to zero (by the definition of z scores), relatively fewer of F1's z scores will be "small" in magnitude, since few of F1's ratings are close to F1's mean rating; accordingly, the mean of the absolute values of F1's z scores is .93, while for other raters this statistic is around .80 to .87. A detailed discussion of the strategic possibilities created by the committee's "rate-then-normalize" procedure is beyond the scope of this paper. Nonetheless, it is clear that different raters used the scales in different ways, and the statistical modeling will capture these differences.

Finally, what of applicant-specific covariates: intended field of study, U.S. residence, and gender? In and of themselves, these variables are neither predictors nor indicators of applicant quality. But they may well be predictors of the ratings assigned by specific raters.

Table 2 Preliminary ordered logit analysis, 1–5 ratings (first threshold set to zero)

	F1	*F2*	*F3*	*F4*	*F5*	*S1*	*S2*	*S3*
Intercept	−15.73*	−7.85*	−11.53*	−9.72*	−3.86*	−28.01*	−6.44*	−6.63*
	(2.47)	(1.97)	(2.22)	(1.98)	(1.90)	(5.44)	(1.93)	(2.06)
Female	0.68	−0.44	−0.064	1.03*	0.35	0.55	−0.17	0.39
	(0.42)	(0.41)	(0.40)	(0.39)	(0.37)	(0.70)	(0.41)	(0.38)
Foreign	1.52*	0.47	0.70	0.44	−0.069	2.87*	−0.72*	−0.10
	(0.45)	(0.43)	(0.44)	(0.45)	(0.40)	(0.90)	(0.44)	(0.46)
GRE total/100	0.89*	0.70*	0.78*	0.59*	0.24*	1.59*	0.54*	0.48*
	(0.13)	(0.11)	(0.12)	(0.10)	(0.092)	(0.29)	(0.10)	(0.10)
American politics	2.21*	0.93	−0.93	−0.08	0.31	0.83	0.80	−0.19
	(0.73)	(0.70)	(0.71)	(0.63)	(0.69)	(1.01)	(0.67)	(0.70)
Political theory	−0.31	−0.26	−0.60	−1.02	0.13	0.34	−0.51	−0.39
	(0.72)	(0.70)	(0.66)	(0.66)	(0.68)	(1.03)	(0.67)	(0.75)
International relations	−0.61	−0.45	−0.086	−0.75*	−0.39	0.89	−0.53	−0.55
	(0.44)	(0.44)	(0.41)	(0.41)	(0.39)	(0.77)	(0.42)	(0.44)
λ_2	1.97	3.52	1.84	1.15	1.15	3.75	2.79	1.21
	(0.37)	(0.84)	(0.45)	(0.26)	(0.20)	(0.91)	(0.51)	(0.31)
λ_3	2.12	5.91	3.78	2.59	2.20	5.38	4.56	2.32
	(0.37)	(0.90)	(0.54)	(0.35)	(0.27)	(1.03)	(0.58)	(0.37)
λ_4	4.89	8.38	5.49	4.02	3.86	7.49	6.77	4.49
	(0.57)	(1.02)	(0.61)	(0.43)	(0.46)	(1.26)	(0.71)	(0.48)

Note. Table entries are maximum likelihood estimates, with standard errors in parentheses. Coefficients marked with an asterisk have magnitudes more than 1.65 times their standard errors.

Raters may well be favorably predisposed toward applicants on the basis of their intended field of study or toward applicants with undergraduate degrees from U.S. universities (e.g., the undergraduate transcript is easier to decipher, the rater may be familiar with the applicant's undergraduate institution or its faculty, and may even know some of the applicant's referees). Finally, raters may differ in the extent to which affirmative action considerations influence their ratings; my own experience is that graduate admissions committees make a tremendous effort to ensure that appropriately qualified women and minority applicants are admitted to their respective department's Ph.D. programs. Nonetheless, raters may differ in the extent to which these types of considerations enter in the ratings they give at this first pass through the applicant pool. The ethnic and racial identity of the applicants is not available for analysis, but gender is available.

Table 2 presents a final piece of preliminary data analysis, an ordered logit analysis of each rater's scores, using the other available indicators and covariates as predictors. First, total GRE score is a statistically significant predictor for all raters (at conventional levels of significance), although as the correlation analysis suggests, the impact of test scores on ratings varies substantially across raters (from a low of .24 for F5 to a high of 1.59 for S1).

If we accept total GRE score as a temporary proxy for candidate ability, then the coefficients on the dummy variables for applicant attributes not related to quality can be interpreted as bias parameters. For instance, rater F1 gives applicants in American politics a substantial boost net of GRE scores. We can gauge the size of this boost by comparing the 2.21 coefficient on the American politics dummy variable with the distance between the threshold parameters; the 2.21 boost associated with being an American politics applicant is roughly the same as moving from a borderline 3–4 rating to a borderline 4–5 rating (put differently, an applicant who ceteris paribus F1 would place in the middle of the "4" category would be given a "5" if that applicant indicated a desire to study American politics). F4 has a moderate but statistically significant tendency to give applicants in international relations lower ratings. These two biases are the only statistically significant field biases revealed by this preliminary analysis, but they suggest that the measurement model I deploy below ought to be sensitive to the possibility of these types of bias.

Other statistically significant bias parameters are apparent in Table 2. For instance, S1 gives an even larger boost to foreign applicants with an ordered logit coefficient of 2.87 (the largest single source of bias revealed by this preliminary analysis). F1 also appears to be favorably disposed toward foreign applicants (1.52); contrast S2, who appears biased against foreign applicants (−.54). F4 gives a hefty boost to female applicants (1.03 on the logit scale), the only statistically significant gender bias we see at this stage of the analysis. Accordingly, the measurement model to be deployed below will include parameters to model the impact of these applicant characteristics on ratings.

5 Measurement Model

The preliminary ordered logit analysis reported in the previous section provides no estimates of applicant quality or any rigorous assessment of our uncertainty over applicant quality. I now outline a measurement model that will generate estimates of applicant quality accompanied by uncertainty assessments.

The indicators of applicant quality are of two types: the ordinal ratings provided by the committee members and GRE scores. Applicant characteristics such as gender, U.S. residence, and intended field of study are not considered indicators of applicant quality but are predictors of the committee ratings nonetheless. The committee members see the GRE scores in each applicant's file, so just as in the preliminary analysis, GRE scores are both indicators of applicant quality in their own right and determinants of the committee members' ordinal ratings.

For the ordinal ratings, an ordinal logit model is convenient and can be motivated via generalization of an item-response model for binary responses, known in the educational testing literature as a *graded response model* and extended here for the case of *multiple raters*; see Johnson and Albert (1999) for a useful introduction from a Bayesian perspective. Raters differ in the extent to which they rely on latent quality in coming up with ratings (i.e., we probably have differences in rater *discrimination* or *reliability*) and also in the way they interpret and use the five-point rating scheme (giving rise to different sets of rater *difficulty* parameters, the thresholds between the ordinal response categories). The preliminary data analysis suggests that both of these inter-rater differences are present in the data. To make these ideas more rigorous, consider the following model:

$$\Pr(y_{ir} \leq d) = \pi_{ird} = F(\tau_{rd} - \mu_{ir}), \tag{1}$$

where

- $y_{ir} \in \{1, \ldots, 5\}$ is the ordinal score given by rater r to applicant i;
- r indexes $\mathcal{R}_i$, the subset of committee members reading the file of the ith applicant: i.e., $r \in \mathcal{R}_i \subset \mathcal{R}$, where $\mathcal{R} =\{$ "F1", ..., "S3"$\}$;
- $\tau_r = (\tau_{r1}, \ldots, \tau_{r5})'$ are thresholds specific to each judge (note that Eq. [1] implies that $\tau_{r5} = \infty$ for all raters); and

- F is the cumulative distribution function of the logistic distribution, i.e., $F(\cdot) = \exp(\cdot)/(1 + \exp(\cdot))$.

The latent, applicant-specific trait x_i is embedded in the following linear model for μ_{ir}:

$$\mu_{ir} = x_i\beta_r + \mathbf{g}_i\gamma_r + \mathbf{z}_i\boldsymbol{\delta}_r, \qquad (2)$$

where

- β_r is the discrimination parameter of committee member r (higher values indicate that the ratings assigned by the committee member are more sensitive to changes in the latent trait);
- $\mathbf{g}_i$ is a vector containing the three GRE scores for applicant i;
- γ_r is a vector of three unknown coefficients, tapping the sensitivity of judge r to each component of the GRE;
- $\mathbf{z}_i$ is a 1-by-q vector of applicant characteristics, ostensibly unrelated to applicant quality, but nonetheless plausible sources of variation in ratings (e.g., gender, intended field of study, residency); and
- $\boldsymbol{\delta}_r$ is a q-by-1 vector of parameters that tap the sensitivity of rater r to the q characteristics in $\mathbf{z}$.

This model would be a conventional ordered logit model save for the fact that the applicant-specific latent trait x_i is unobserved.

In addition, I impose some additional restrictions on the threshold parameters: aside from the ordering constraint implied by the ordinal responses, I estimate a "baseline" set of thresholds $\boldsymbol{\lambda}$ for a single committee member and then estimate the other committee members' thresholds as a uniform shift from the baseline thresholds; i.e., $\tau_r = \boldsymbol{\lambda} + \eta_r$, $r = 2, \ldots, 8$ (the choice as to whose thresholds are the baseline thresholds is arbitrary; I use committee member F1). This restriction means there are just $4 + 7 = 11$ parameters required to estimate thresholds, versus the $4 \times 8 = 32$ parameters needed to fit a full set of four thresholds per committee member. In analysis not reported here I found that that there is simply not enough data (nor prior information) to estimate four thresholds for each committee member as well as all the other parameters in the model. The restriction means that the "widths" of the ordinal categories are equal across committee members, but the location of the thresholds varies across committee members, depending on the difficulty contrasts η_r; i.e., a committee member who is a "tougher grader" relative to the baseline committee member F1 will have $\eta_r > 0$ and thresholds shifted to the right.

The GRE scores closely approximate continuous variables and are modeled via a much simpler setup, i.e., linear regression functions of latent quality with normal iid errors. Let $j \in \{\text{Verbal, Quant, Analytic}\}$ index the GRE scores. Then the model is

$$g_{ij} \sim N(\nu_{j1} + \nu_{j2}x_i, \sigma_j^2), \qquad (3)$$

where $\mathbf{v}_j = (\nu_{j1}, \nu_{j2})'$ and σ_j^2 are additional unknown parameters to be estimated. Note that the applicant characteristics in $\mathbf{z}_i$ are excluded from the model for the GREs: preliminary analysis unreported here suggests that these characteristics are not significant sources of variation in GRE scores. Note also the assumption that conditional on latent quality, GRE scores are independent. Implicit here is the notion that GRE scores can be modeled with a unidimensional latent trait (the same latent trait that I assume to underlie the committee ratings). Of course, the GRE is administered with three different components (verbal, quantitative, and analytical) because extensive psychometric research indicates that these are separate domains of ability CITE. To some extent, this is more or less confirmed in the data from the applicants: Table 1 reports a modest correlation (.27) between verbal and quantitative GRE scores among the applicants, and stronger but still moderate correlations among the analytic and verbal components (.46) and the analytic and quantitative components (.44), suggesting that there is more than one ability dimension underlying the three components. However, the GRE components are far from being orthogonal, suggesting that there is some redundancy among the scores and that they do tap at least one common trait. Whatever view one takes concerning the dimensionality of the GRE data, it should be remembered that my goal here is not to model the GREs per se, but rather to exploit them as an additional source of information about the latent quality of applicants. By definition and by construction, latent quality is unidimensional; should this not be an especially good model for the GREs, this will be reflected in the relevant parameters (e.g., small ν_{j2}, perhaps indistinguishable from zero, and large σ_j^2, indicating a poor fit to the GRE data), and in turn less precision concerning the latent traits of the applicants.

It is also worth pointing out that the model has latent quality influencing the committee members' ratings directly, and indirectly via the GRE scores. This poses no difficulty for identification, estimation, or inference, since the resulting system of equations is recursive: the likelihood for the endogenous variables (ratings and GRE scores) factors into two pieces, an ordered logistic likelihood for the ratings conditional on GRE scores and a normal likelihood for the GRE scores. In interpreting the parameter estimates I will focus on the total effect of a unit change in latent quality on ratings for committee member r, the sum of indirect and direct effects:

$$\mathrm{TE}(x)_r = \sum_{j=1}^{3} \nu_{j2}\gamma_{rj} + \beta_r. \tag{4}$$

5.1 *Identification and Parameter Restrictions*

Before proceeding to estimation, the scale and location of the unidimensional latent trait needs to pinned down. The identification problem created by scale invariance is obvious and arises frequently in models with latent variables. Inspection of Eqs. (2) and (3) makes it clear that the data cannot distinguish among arbitrary rescalings of the unobserved x_i and offsetting rescalings of the unknown β_r (Eq. 2) and ν_{t2} and σ_t^2 (Eq. 3). The problem of location invariance is that the data cannot distinguish among arbitrary shifts in the x_i, say $x_i + c$, and offsetting shifts in the threshold parameters $\boldsymbol{\tau}$ (see Eq. 1) and the intercepts ν_{j1} in the GRE model (Eq. 3).

Of course, in a Bayesian analysis, one can always choose to ignore the lack of identification; this poses no formal problem for Bayesian inference.[2] In the specific case of latent variable modeling, the lack of identification stems from a lack of an established metric for assessing applicant quality (analogously, consider trying to measure and model temperature without having first agreed on whether we are working in Fahrenheit or Celsius); this is more a technical annoyance than a deep issue with substantive implications.

[2] Since the posterior is proportional to the prior times a likelihood, if a likelihood is uniform over the parameter space (i.e., the data are uninformative about the model parameters), then the posterior and the prior coincide (nothing has been learned about the model parameters).

Two simple routes to identification are available for the case of a unidimensional latent trait:[3] (1) normalize the x_i to have a fixed scale and location (e.g., standardize so the x_i have, say, mean zero and variance one); (2) a priori set two of the unknown x_i. Note that the former approach guarantees local identification, while the latter supplies global identification (with only local identification we could always "flip" the unidimensional scale onto its mirror image—a rotation of 180°—and obtain an identical fit to the data; or, in Bayesian language, with only local identification of the likelihood and uninformative priors there are two mirror-image posterior modes. For all practical purposes local identification is usually sufficient to proceed with data analysis (and ex post we deal with the trivial distinction between an applicant quality scale that runs from low to high or vice versa). I impose the former restriction (the latent traits have mean zero and standard deviation one across applicants).

In summary, the statistical problem here is to learn about the following 371 parameters:

- 279 latent traits, $\mathbf{x} = (x_1, \ldots, x_{279})'$
- 8 rater-specific discrimination parameters, $\boldsymbol{\beta}$
- $8 \times 3 = 24$ γ parameters, stacked in the matrix $\boldsymbol{\Gamma}$, the impact of each of the three GRE scores on each of the eight committee members' ratings
- $8 \times 5 = 40$ δ parameters, stacked in the matrix $\boldsymbol{\Delta}$, the impact of each of the applicant characteristics (unrelated to quality) on each of the eight committee members' ratings
- 4 threshold parameters $\boldsymbol{\tau}$, plus seven contrasts $\boldsymbol{\eta}$, for committee members 2, …, 8.
- $3 \times 2 = 6$ ν parameters, stacked in the matrix $\mathbf{N}$, the slope and intercept parameters in the linear model relating the three GRE scores to the latent trait
- 3 error variance parameters σ_i^2, the error variances in the GRE equations.

5.2 *Priors*

For the latent traits, $\mathbf{x}$, the choice of prior is made redundant by the fact that I impose the identifying restriction that the x_i have mean zero and standard deviation one across applicants. The discussion below on postprocessing (Section 5.4) clarifies how this constraint was imposed. Vague, normal priors are also used for the discrimination parameters $\boldsymbol{\beta}$; the normalization that the latent traits have standard deviation one suggests that the likely values of β (slopes in an ordered logit analysis) will not be massive; I express my prior uncertainty over the β_r parameters with independent $N(0, 10^2)$ priors.

Likewise, the scaling of the latent trait suggests priors for the $\mathbf{N}$ parameters, linking the latent traits x_i to the GRE scores. Each GRE score is divided by 100 in this analysis, so if a unit change in x_i (a big change) brought about a big change in terms of GRE points (e.g., 200 points), then we might expect ν_{j2} to be around 2.0; similarly, if the average applicant has $x_i = 0$ (halfway between the anchors) and the average GRE score is, say, 650, then we might expect the intercept terms ν_{j1} to be around 6.5. Of course, I want the data to dominate inferences for these parameters, so I adopt independent $N(0, 25^2)$ priors that are approximately uniform over the range of plausible values for the various ν parameters. Likewise, the error variances in the GRE equations σ_j^2 cannot be larger than the variances in the normalized GRE scores themselves (the variances range from .83 for the normalized quantitative scores to 1.23 for normalized verbal scores); I employ diffuse, independent inverse-Gamma priors on the σ_j^2 parameters, which are approximately uniform over the

[3] See Rivers (2003) on identification conditions for the more general case.

plausible range of values for these parameters. An additional feature of the inverse-Gamma prior is that conditional on the latent x_i, the normal priors for the **N** parameters and the normal model in Eq. (3) mean that this part of the analysis is conjugate (the posteriors are in the same parametric family as the priors), and the computation is reasonably straightforward (although with modern computational tools, conjugacy in and of itself is less of a compelling reason for choosing particular parametric forms for priors).

I also employ independent, vague normal priors for the $\mathbf{\Delta}$ and $\mathbf{\Gamma}$ parameters, which tap the effects of the applicant characteristics unrelated to quality (**Z**) and the GRE scores on the committee members' ratings. The **Z** are all dummy variables (indicators of various attributes) and the GRE scores are divided by 100 (and range from 4.0 to 8.0), so I do not expect the (ordered logit) coefficients on these quantities to be massive. Uncertainty over the value of these parameters is represented with normal priors with mean zero and variance 10^2.

Finally, for the threshold parameters $\boldsymbol{\tau}$ I have no prior information other than the ordering constraint implied by the model. Accordingly, I use the following vague priors: $\tau_{r1} \sim N(0, 10^2)\mathcal{I}(\tau_{r1} < \tau_{r2})$, $\tau_{r2} \sim N(0, 10^2)\mathcal{I}(\tau_{r1} < \tau_{r2} < \tau_{r3})$, $\tau_{r3} \sim N(0, 10^2)\mathcal{I}(\tau_{r2} < \tau_{r3} < \tau_{r4})$, and $\tau_{r4} \sim N(0, 10^2)\mathcal{I}(\tau_{r4} > \tau_{r3})$, where $\mathcal{I}(z)$ is an indicator function set to one if the event z is true and zero otherwise (truncating the prior densities to those regions where the ordering constraints are true).

5.3 *Estimation via Markov chain Monte Carlo*

I use Markov chain Monte Carlo (MCMC) methods to generate a random tour of the posterior density of the parameters listed above. MCMC is an attractive computational strategy for this problem for several reasons. Given the large number of parameters to estimate, brute force optimization of the likelihood for this problem is not a particularly attractive strategy. In addition, beyond the sheer size of the optimization problem, there is also the issue of missing data. Fifteen applicants (5.3% of the applicant pool) have no GRE data. Nonetheless, these applicants' files were read and rated by committee members, and so are not entirely uninformative about rater discrimination and the weight that raters attach to the observable applicant characteristics aside from GREs. The missing data and the large number of parameters pose no difficulty in the MCMC context; if we are willing to treat the missing GRE data as missing at random, then over iterations of the MCMC algorithm (described below) we obtain multiple imputations for the missing data, with the additional uncertainty that this generates automatically propagated into inferences about other model parameters (e.g., the latent trait for these applicants). As I show below, it is not that we know nothing about applicants with incomplete files; rather, we have less certainty about the quality of applicants with incomplete files.

Finally, MCMC methods let us explore the joint posterior density of the model parameters; *any* feature of that posterior density can be obtained up to *any* degree of precision. This is particularly helpful in the context of graduate admissions. For instance, as I show below, MCMC methods make it simple to estimate and make inferences about the order statistics of the applicant on the latent quality dimension. An explicit goal of most graduate admissions committees is to come up with a rank ordering of the candidates, e.g., a "top-20" list of "must admit" applicants and/or a "top-50" list of those "worth a second read." The Bayesian approach I adopt here lets us compute confidence intervals on the ranks of each applicant on the latent quality dimension; moreover, we can then go on to consider the posterior probability of a given candidate belonging in the top 20 or top 50 and so on. These types of inferences over transformations and comparisons of

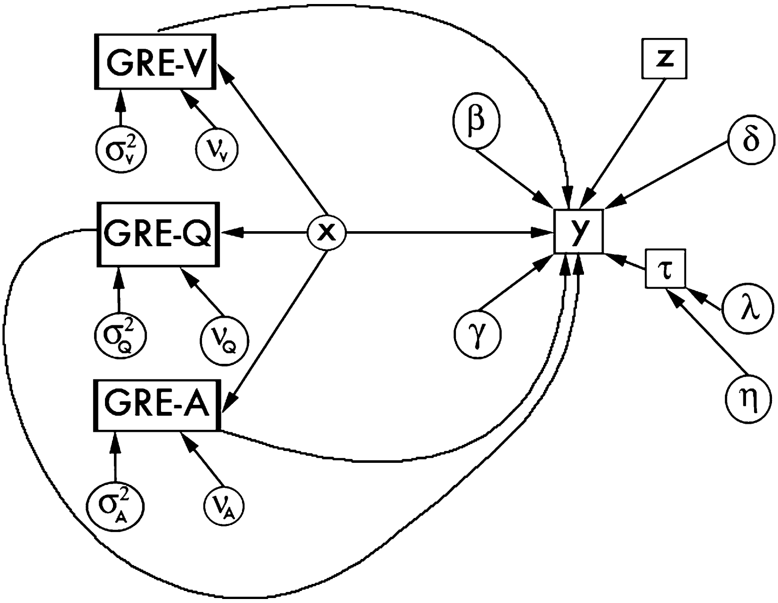

Fig. 2 Directed graph for measurement model. Circular nodes denote unknown parameters; rectangular nodes denote fixed or deterministic quantities. y represents the ordinal ratings given by the committee members; for clarity, just one y is included in the graph.

parameters are extremely difficult to implement in a classical framework but are generated more or less automatically in the fully Bayesian, MCMC approach.

The workhorse MCMC algorithm—the Gibbs sampler—involves breaking the joint posterior density into its component conditional densities and sampling from each. At iteration t, the MCMC algorithm samples from the conditional density of each parameter, where the conditioning follows from the structure of the model. Spiegelhalter et al. (1996) show that for models that can be represented as *directed acyclic graphs* (DAGs), the conditional distribution for stochastic node v in a DAG $\mathcal{G}$ is

$$p(v \mid \mathcal{G}_{-v}) = p(v \mid \text{parents}[v]) \prod_{w \in \text{children}[v]} p(w \mid \text{parents}[w]). \tag{5}$$

See also Spiegelhalter and Lauritzen (1990). The measurement model I use in this context can be represented as a DAG; a simplified version of the DAG for my model appears in Fig. 2. The DAG is an excellent way to convey the important features of the model; e.g., (1) the latent trait x has no parent nodes; (2) the GRE scores and the committee members' ratings y are children nodes of x; (3) ν and σ^2 depend only on the GRE scores and the latent trait x; (4) parameters in the model for the committee members' ratings, such as β, δ, γ and the threshold parameters τ and η depend on x and the GRE scores, but not on the other parent nodes of the GRE scores (ν and σ^2).

The free *WinBUGS* program (Spiegelhalter et al. 2003) implements MCMC for models that can be represented as DAGs, exploiting the result in Eq. (5) via a variety of sampling algorithms. In particular, *WinBUGS* is especially well suited for Bayesian inference for the measurement model I use here: the mix of continuous data (GREs) and ordinal data (committee members' ratings) means that some of the conditional distributions that underlie the MCMC algorithm are not standard (e.g., the conditional distributions of the latent traits x_i) but can be sampled from using the more advanced sampling algorithms available in *WinBUGS*, such as slice sampling (Neal 1997).

I initialized the MCMC algorithm with zeros for all parameters except the threshold parameters. The MCMC algorithm was run for 250,000 iterations, discarding the first 25,000 iterations as a lengthy burn-in phase, ensuring that the MCMC algorithm has transitioned away from the arbitrary initial values. Inspection of trace plots and within-chain autocorrelation functions shows the MCMC algorithm to be slow mixing with respect to certain parameters (in particular, the parameters tapping the effect of the GRE scores on the committee members' ratings). Accordingly, every 250th iteration was then retained for inference, providing 900 approximately independent samples from the joint posterior density of the model parameters. The *WinBUGS* program was run with *WinBUGS* version 1.4 on a machine with a 1.8 GHz Pentium processor that took approximately 10.5 hours to perform 250,000 iterations. All code is available on request.

5.4 *Postprocessing the MCMC output*

Recall that I impose the identifying restriction that the latent traits have zero mean and standard deviation one across applicants. This constraint is reasonably straightforward to implement in *WinBUGS* but causes the program to run extremely slowly. Accordingly I use *WinBUGS* to analyze the unidentified model, placing independent $N(0, 1)$ priors on the x_i. After the MCMC iterations terminate, I then "postprocess" the MCMC output, normalizing the sampled x_i to have mean zero and standard deviation. Other parameters in the model must be transformed accordingly. For instance, at iteration t the MCMC algorithm produces $\mathbf{x}^{(t)} = (x_1^{(t)}, \ldots, x_n^{(t)})'$, which is then renormalized to $\tilde{x}^{(t)} = b^{(t)}(\mathbf{x}^{(t)} - a^{(t)})$, where $a^{(t)} = \bar{\mathbf{x}}^{(t)}$ and $b^{(t)} = 1/\mathrm{sd}(\mathbf{x}^{(t)})$. Any linear transformation of the x_i requires offsetting linear transformations of parameters that govern how x_i is fit to the given data, if we are to generate the same fit to the data. Inspection of Eq. 1 indicates that we require the transformations $\tilde{\tau}_r^{(t)} = b^{(t)}(\tau_r^{(t)} - \beta_r^{(t)} a^{(t)})$ and $\tilde{\beta}_r^{(t)} = \beta_r^{(t)}/b^{(t)}$. Inspection of Eq. 3 indicates that we also need the transformations $\tilde{\nu}_1^{(t)} = \nu_1^{(t)} + \nu_2^{(t)} a^{(t)}$ and $\tilde{\nu}_2^{(t)} = \nu_2^{(t)}/b^{(t)}$. These transformations are a mapping from an unidentified parameter space into a (lower-dimensional) space of identified parameters (with the location and scale restrictions on the x_i reducing the dimensionality of the parameter space by two). In describing the results of the analysis below, references to x_i should be understood as references to $\tilde{x}_i$ etc.

It bears repeating that neither Bayesian inference nor MCMC algorithms need likelihoods to be identified; analyses of unidentified models may be not particularly interesting but technically feasible and, depending on the computational burden of imposing a given set of identification constraints, analysis of the unidentified model may be more computationally tractable than analysis of the identified model. The strategy I adopt here—running MCMC over a set of unidentified model parameters and then postprocessing the output so as to map into the identified set of parameters—is reasonably novel but seems quite helpful for Bayesian analysis of latent variable models (e.g., Hoff et al. 2002; Edwards and Allenby 2003).

6 Results

Figure 3 summarizes the posterior densities of the latent traits, x_i; each plotted point is a posterior mean, and the lines cover 95% confidence interval for each x_i (the posterior means are simply the means of the corresponding parameter from the 900 MCMC

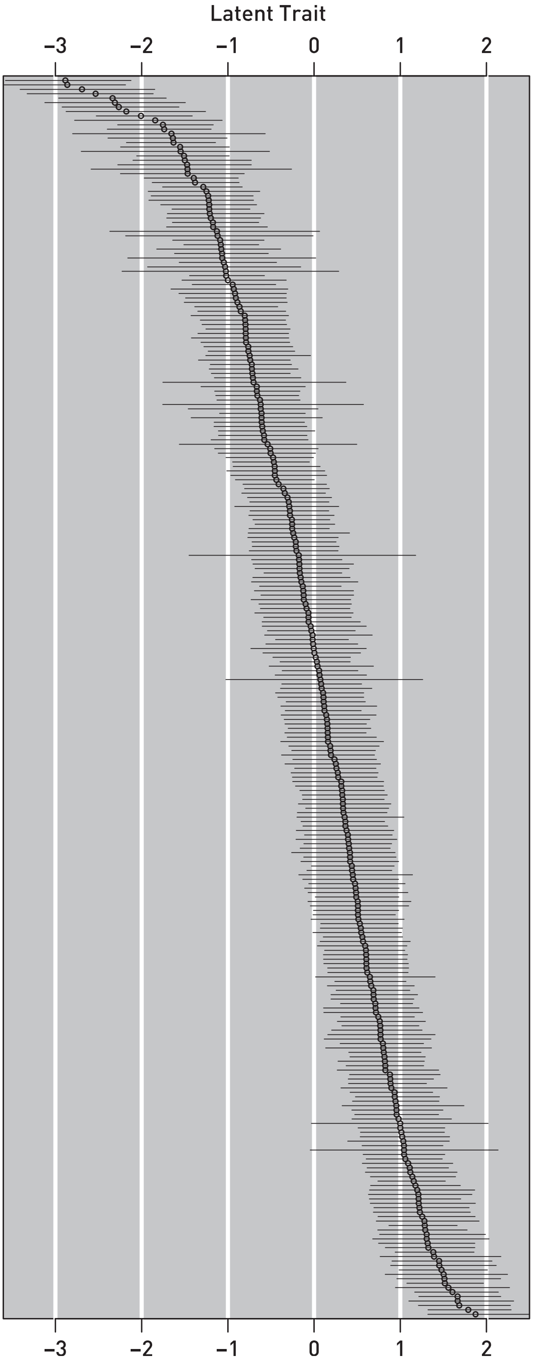

Fig. 3 Posterior densities, latent quality. Circles are posterior means; the lines show the width of 95% confidence interval (2.5th to 97.5th percentiles of the respective posterior density).

iterations retained for inference, and the 95% confidence intervals are computed as the 2.5 and 97.5 percentiles of the 1,000 sampled values). The applicants have been sorted by posterior means for x_i. The posterior means for x_i range from -2.88 to 1.87, and their distribution is slightly left-skewed (a longer lower tail than upper tail).

For applicants without GRE scores we are more uncertain about their latent quality a posteriori. The applicants with appreciably wider confidence intervals in Fig. 3 are almost universally those for whom we lack GRE scores. The posterior standard deviation of x_i for applicants with GRE scores averages .28, while the average posterior standard deviation for applicants without GRE scores is .56. Recall that I use a $N(0, 1)$ prior for each x_i, so the information gain from the data and modeling is quite large: the ratio of posterior precision to prior precision in x_i is (on average) 12.7 for applicants with GRE scores and 3.19 for applicants without GRE scores.

It bears noting that the confidence intervals in Fig. 3 are simple pointwise confidence intervals; they cannot be used for directly comparing the latent quality of two applicants (such a comparison requires knowledge of the joint distribution and, in particular, the covariance of the two random variables). One of the advantages of using MCMC methods to sample from the joint posterior density of all model parameters is that it is straightforward to induce a posterior density on any function of the parameters; i.e., if the MCMC algorithm produces a sequence of sampled values $\langle \boldsymbol{\theta}^{(t)} \rangle$ from the posterior of $\boldsymbol{\theta}$, then the sequence $\langle h(\boldsymbol{\theta}^{(t)}) \rangle$ is a sample from the posterior for the estimand $h(\boldsymbol{\theta})$. The $h(\cdot)$ function can be whatever the analyst is interested in, such as pairwise comparisons of specific applicants' latent traits or the entire set of order statistics.

Figure 4 summarizes the posterior density on a rank ordering of the applicants' x_i. The plotted points show the posterior mean of the order statistic for each applicant's x_i, and the lines cover 95% confidence intervals. The confidence intervals shrink and become asymmetric in the extremes of the data, from the natural "floor" and "ceiling" effects associated with a rank ordering (e.g., for excellent applicants, the uncertainty as to their rank tends to be on the downside). In general there is considerable uncertainty as to the rank of any given applicant: the confidence intervals on the ranks are quite wide, averaging 88 places, or $88/279 = 31\%$ of the applicant pool. Although uncertainty over ranks is smaller in the extremes of the data, as many as 12 applicants can claim to being the highest quality applicant (i.e., the 95% confidence interval on their rank ordering includes the highest rank). In addition, the uncertainty created from an incomplete application file is quite large; among applicants who supply GREs, the 95% bound on their rank ordering averages 85 places, but among applicants for whom we lack GRE scores, the average 95% bound is 142 places, or roughly half the applicant pool.

The primary goal of the committee scoring procedure was to make a "first cut" through the applicant pool: identifying approximately 50 applicants to be given closer consideration. It is straightforward to compute a posterior probability that a given applicant is in the top 50: at each iteration, the MCMC algorithm produces $\mathbf{x}^{(t)}$, a sample from the joint posterior distribution of the latent traits of the applicants; these can be sorted to produce a vector of ranks $\mathbf{r}(\mathbf{x}^{(t)}) = (r_1^{(t)}, \ldots, r_n^{(t)})'$. To compute the posterior probability that a given applicant is in the top 50, I simply note the proportion of times each applicant's rank $r_i^{(t)} < 51$ (say, if we rank the x_i in descending order, since quality is increasing in x_i).

These posterior probabilities are plotted in Fig. 5. Just 15 applicants have posterior probabilities of being in the top 50 of 0.95 or higher (the default level of statistical significance in the social sciences); this result might seem paradoxical—only 15 applicants unambiguously ($p > .95$) belong in the set of the top 50 applicants—but it is

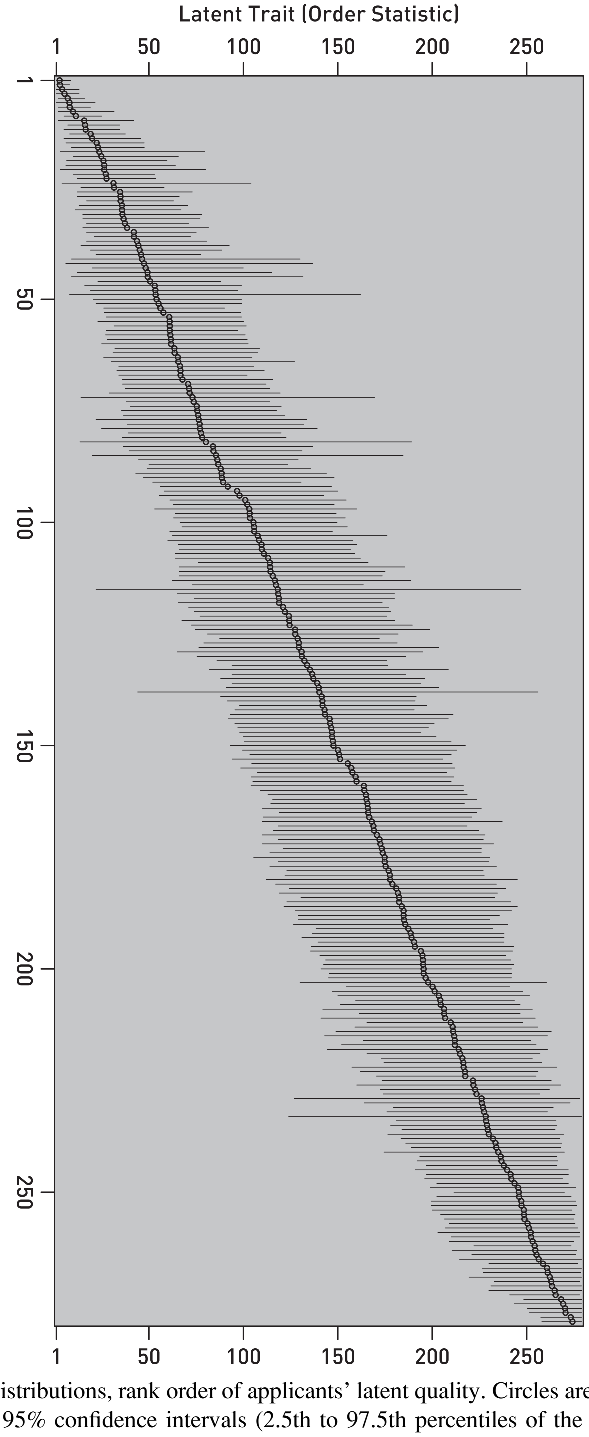

Fig. 4 Posterior distributions, rank order of applicants' latent quality. Circles are posterior means on ranks; lines cover 95% confidence intervals (2.5th to 97.5th percentiles of the respective posterior density).

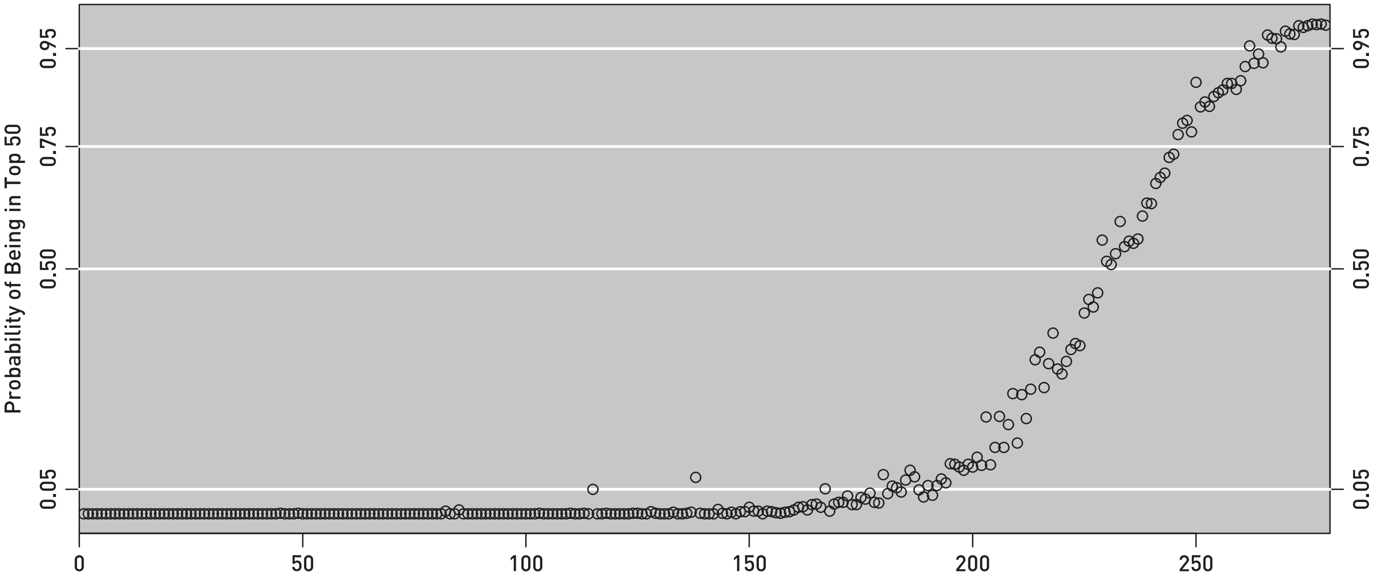

Fig. 5 Posterior probability of being in top 50 applicants.

a consequence of dealing with the considerable uncertainty that accompanies the measures of applicant quality supported by these data.

Of course, graduate admissions committees might also be keen to avoid making errors of the "false negative" sort, rejecting applicants who might actually be worth a second look. If one were to apply the conventional $p = .05$ significance level in this direction, rejecting only those applicants whose posterior probability of being in the top 50 fell below $p = .05$, then the set of top 50 applicants would grow to encompass 98 applicants, a set twice as large as the one the committee might have thought they were trying to create. Again, this is a consequence of measuring applicant quality so noisily.

Recall that the committee ranked the applicants by standardizing each committee members' ordinal scores and then assigning a score to each applicant by averaging the standardized scores given to that applicant by the committee members reading that applicant's file. An initial set of semifinalists was chosen via a simple rank ordering on this measure, and then admittees were chosen from the semifinalists.

Figure 6 presents a comparison of the committee's summary measure and the applicant quality measure produced by the measurement model. There is a strong, positive correlation between the two measures: the posterior means of the applicant quality measure developed here and the committee's summary measure correlate at 0.81; put differently, approximately 66% of the variation in the committee's summary measure is explained by variation in my measure of applicant quality. This leaves the question of what accounts for differences between my measure of applicant quality and the committee's simple summary measure. The answer is in two parts: (1) committee members varying in the way their ratings reflect differences in applicant quality (differential reliability); (2) the influence of characteristics that are ostensibly extraneous to applicant quality on committee members' ratings (bias). The estimates of applicant quality produced by my measure are, by construction, sensitive to variation in discrimination across the committee and are purged of these observable sources of potential bias.

Table 3 presents summaries of the posterior distributions of parameters specific to each committee member. A variety of parameters is reported: the discrimination parameters, parameters tapping possible sources of bias (e.g., gender, foreign residence, intended field of study), the direct effects of the GRE scores, the total effect of the applicants' latent traits on the committee members' ratings (see Eq. 4), and threshold parameters.

Perhaps the most compelling feature of these results is the difference in discrimination across committee members: F1 is the most reliable committee member, with a large β_r parameter (direct effect of applicant quality) and the largest posterior mean for the total effect of latent applicant quality on ratings (posterior mean of 4.15). Committee members F2 and S2 tie for the smallest total effects of applicant quality, with posterior means of 2.04, or roughly half that of F1's total effect. Note also that none of the direct effects of applicant quality nor the total effects have confidence intervals overlapping zero: all committee members' assessments were tapping latent quality in making their assessments of the applicant files, albeit in varying degrees.

This analysis also finds some statistically significant sources of bias, echoing some of the results of the preliminary data analysis. F1's strong preference for American politics applicants remains, and the implied boost is the equivalent of over a standard deviation movement on the latent quality scale (i.e., a 5.10 boost on the logit scale for American politics applicants versus the 4.15 estimate of the total effect of a unit change in applicant quality, recalling that by construction, the applicant quality scale has standard deviation one). F4 displays a smaller but statistically significant bias toward candidates intending to study international relations (posterior mean of -1.48, contrasted with F4's total effect of

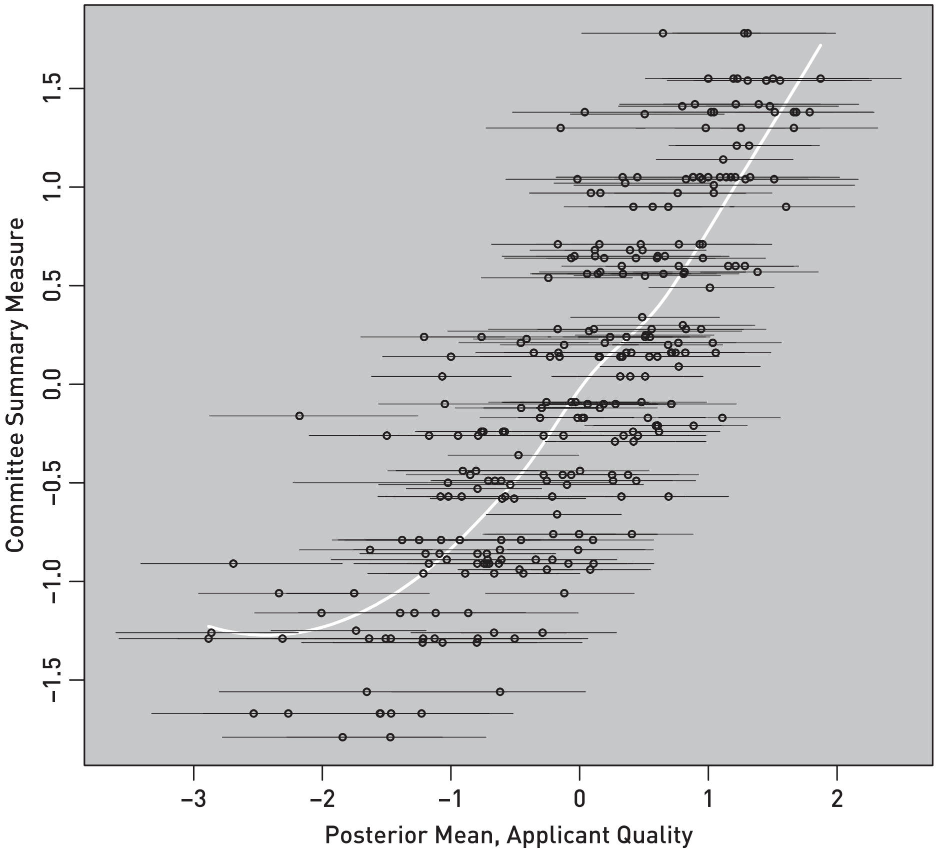

Fig. 6 Comparison of committee summary measure and applicant quality from measurement model. Horizontal bars show 95% confidence intervals for estimates of applicant quality produced by the measurement model. The correlation between the posterior means for applicant quality and the committee's summary measure is .81; the solid white line is a nonparametric smoother fit by local fitting (Loader 1999).

applicant quality of 3.81). F1 and F4 also display gender biases, preferring female applicants by roughly the same amount (posterior means on the effects of 1.89 and 1.81, respectively).

The estimates of the effects of the GRE scores on ratings also warrant comment. Each GRE effect in Table 3 speaks to a rather odd counterfactual: holding applicant quality constant, and the other GRE scores and other characteristics of an applicant constant, what is the impact of a 100-point increase in the particular GRE component? This is an odd counterfactual since elsewhere in the model, GREs are modeled as a function of latent applicant quality (and indeed are an important source of information about applicant quality); estimates of this part of the model appear in Table 4. The analytic component of the GRE is estimated to have the greatest discrimination with respect to the latent trait, which has the largest slope coefficient in the model for the analytic GRE component and the smallest residual standard error. One standard deviation increase in applicant quality produces an estimated 93-point increase on the analytic component of the GRE, versus a 61-point increase on the verbal component and a mild 47-point boost on the quantitative component (although the quantitative component has the highest intercept parameter). The

Table 3 Posterior summaries, ordinal logistic multiple rater model

	F1	*F2*	*F3*	*F4*	*F5*	*S1*	*S2*	*S3*
Discrimination β_r	6.17	2.77	4.73	5.95	5.15	4.35	2.89	5.35
	[4.43, 8.02]	[1.69, 4.05]	[3.30, 6.53]	[4.40, 7.99]	[3.46, 6.99]	[2.56, 6.41]	[1.71, 4.25]	[3.80, 7.08]
Verbal/100 γ_{r1}	0.32	0.74	0.52	−0.24	0.19	0.88	0.72	0.25
	[−0.63, 1.18]	[0.21, 1.28]	[−0.26, 1.23]	[−1.24, 0.70]	[−0.69, 0.96]	[−0.19, 1.94]	[0.13, 1.24]	[−0.57, 0.98]
Quant/100 γ_{r2}	2.14	0.36	0.45	0.08	−0.71	1.64	−0.15	0.55
	[1.16, 3.16]	[−0.33, 0.99]	[−0.54, 1.32]	[−1.13, 1.12]	[−1.69, 0.26]	[0.61, 2.81]	[−0.81, 0.48]	[−0.46, 1.52]
Analytic/100 γ_{r3}	−3.39	−1.41	−2.02	−2.12	−2.71	−1.72	−1.27	−3.04
	[−5.31, −1.49]	[−2.65, −0.32]	[−3.76, −0.44]	[−4.23, −0.27]	[−4.69, −0.99]	[−3.93, 0.11]	[−2.54, −0.17]	[−5.06, −1.27]
Female δ_{r1}	1.89	−0.35	0.16	1.81	0.85	0.14	−0.40	0.94
	[0.53, 3.43]	[−1.19, 0.55]	[−1.08, 1.32]	[0.56, 3.04]	[−0.28, 1.97]	[−1.48, 1.91]	[−1.38, 0.58]	[−0.32, 2.05]
Foreign δ_{r2}	1.13	0.45	0.69	0.64	0.92	1.90	−0.05	0.08
	[−0.70, 3.12]	[−0.69, 1.62]	[−0.68, 2.04]	[−0.94, 2.27]	[−0.47, 2.44]	[−0.37, 4.16]	[−1.19, 1.05]	[−1.32, 1.65]
American politics δ_{r3}	5.10	1.16	−1.77	−1.10	1.50	1.09	1.30	0.09
	[2.72, 7.77]	[−0.27, 2.76]	[−3.78, 0.18]	[−3.22, 1.17]	[−0.46, 3.44]	[−1.48, 3.69]	[−0.31, 2.70]	[−1.99, 2.17]
Political theory δ_{r4}	−0.85	−0.31	−0.77	−1.22	0.72	−1.18	−0.65	0.85
	[−3.23, 1.54]	[−1.78, 1.19]	[−2.88, 1.31]	[−3.28, 0.87]	[−1.52, 3.02]	[−3.82, 1.60]	[−2.15, 0.82]	[−1.37, 3.01]
International relations δ_{r5}	−0.72	−0.26	−0.36	−1.48	−0.53	0.51	−0.64	−1.21
	[−2.29, 0.79]	[−1.23, 0.76]	[−1.64, 0.87]	[−2.94, −0.06]	[−1.81, 0.59]	[−1.37, 2.40]	[−1.61, 0.30]	[−2.51, 0.14]
Total effect of latent trait, Eq. (4)	4.15	2.04	3.32	3.81	2.34	4.01	2.04	2.87
	[3.13, 5.30]	[1.50, 2.59]	[2.64, 4.12]	[3.02, 4.75]	[1.52, 3.24]	[3.02, 5.13]	[1.41, 2.67]	[2.06, 3.73]
Threshold τ_{r1}	−8.55	−9.82	−4.65	1.14	5.38	−18.92	−7.91	2.43
	[−16.74, −1.88]	[−24.17, 3.44]	[−18.32, 8.32]	[−11.46, 14.33]	[−8.12, 20.10]	[−33.49, −4.58]	[−21.08, 4.04]	[−11.52, 15.43]
Threshold τ_{r2}	−5.87	−7.14	−1.97	3.82	8.06	−16.24	−5.24	5.11
	[−14.09, 0.93]	[−21.68, 6.11]	[−15.60, 11.04]	[−8.87, 16.91]	[−5.56, 22.66]	[−30.87, −1.70]	[−18.43, 6.66]	[−8.82, 18.35]
Threshold τ_{r3}	−3.62	−4.89	0.28	6.07	10.31	−13.99	−2.99	7.36
	[−11.89, 3.19]	[−19.35, 8.34]	[−13.36, 13.32]	[−6.49, 19.24]	[−3.34, 24.92]	[−28.63, 0.52]	[−16.10, 8.95]	[−6.52, 20.53]
Threshold τ_{r4}	−0.24	−1.50	3.67	9.45	13.69	−10.60	0.40	10.74
	[−8.39, 6.58]	[−15.86, 11.78]	[−10.10, 16.73]	[−2.95, 22.87]	[0.16, 27.99]	[−25.11, 3.98]	[−12.69, 12.16]	[−3.16, 23.97]

Note. Table entries are posterior means; 95% confidence intervals are shown in square brackets.

Table 4 Posterior summaries, model for GREs

	Verbal	*Quantitative*	*Analytic*
Intercept ν_{t1}	6.02 [5.91, 6.13]	6.78 [6.68, 6.87]	6.55 [6.48, 6.63]
Slope ν_{t2}	0.61 [0.46, 0.75]	0.47 [0.36, 0.59]	0.93 [0.77, 1.04]
σ_t	0.93 [0.84, 1.03]	0.79 [0.71, 0.88]	0.58 [0.43, 0.75]

Note. Table entries are posterior means of indicated parameters; 95% confidence intervals (2.5th and 97.5th percentiles of the respective posterior density) appear in brackets.

upshot is that increases in the GRE analytic score that are somehow unaccompanied by an increase in applicant quality or the other GRE scores seem to lead to lower ratings from the committee members, or at least this is the interpretation of the consistently negative γ_{r3} parameters in Table 3. The more useful way to interpret the GRE effects in Table 3 is to recall that they are part of the total effect of applicant quality on committee members' ratings; as Eq. 4 shows, the total effect of latent applicant quality is spread out over seven parameters, a direct effect and three indirect effects from the three GRE components. The fact that one of the components of this combination of effects is negative is difficult to interpret in and of itself, and I prefer to focus on the total effects of candidate quality.

7 Improving Graduate Admissions?

A number of implications for the graduate admissions process stem from my analysis. First, analysis of the available data from this particular graduate admissions process reveals tremendous uncertainty in assessments of latent quality. Our ability to draw believable distinctions between applicants is limited to comparisons no finer than the difference between exceptionally good applicants and average applicants. The average 95% confidence interval on the rank of an applicant is so wide that it covers 31% of the applicant pool. Just 15 of 279 applicants have greater than 95% probability of being among the top 50 applicants. The committee's task—selecting *m* applicants for admission—seems to be fraught with risk: the analysis reveals that there will always be many applicants for whom our best guess is that they lie below the *m* threshold but nonetheless have appreciable probability of actually lying above the *m* threshold. On the balance of probabilities we make the right decision (dropping applicants whose estimated rank is below *m*). But in the neighborhood of any threshold, the accept/reject decision smacks of caprice.

Table 5 provides one final demonstration of the risk inherent in rejecting applicants lying below a particular threshold. Applicant 141 is the 50th best applicant (based on a rank ordering of the posterior means of the latent trait). Applicants ranked 51 and higher will be rejected if we restrict admission to the top 50 applicants. But as Table 5 shows, these accept/reject decisions are based on extremely flimsy evidence; in fact, the probability that the 51st ranked applicant has a higher score than the 50th ranked applicant is slightly better than 50%.[4] The probability that applicants ranked 52 through 55 have higher values on the latent trait than applicant 50 is less than .50, but barely, ranging from .46 to .41, far from the levels of statistical significance we typically apply in our research.

The risk of a wrong decision (admitting someone who is really below a threshold, or not admitting someone who is actually above a particular threshold) is a function of the

[4] This result is not as odd as it might seem, and it follows from the fact that there are different levels of skew in the posterior densities for the two applicants, although $E(x_i) > E(x_j)$, $\Pr(x_i > x_j) < .5$.

Table 5 Comparison of latent trait in the neighborhood of the Top-50 threshold

Rank of $E(x_i)$	*Applicant*	*$E(x_i)$*	*$Pr(x_i > x_{141})$*
45	235	0.98	.53
46	138	0.96	.54
47	222	0.96	.55
48	198	0.95	.50
49	8	0.94	.53
50	141	0.93	-
51	181	0.93	.51
52	39	0.90	.46
53	84	0.89	.46
54	6	0.88	.44
55	25	0.88	.41

posterior variances of the latent traits. The bigger the posterior variances, the more difficult it is to distinguish among applicants. Quite simply, we need more information, either about the applicants (rather obviously) or about the way the committee members rate applicants. Note that uncertainty in the parameters specific to committee members makes their ratings less informative about applicant quality. Thus getting more information about the committee members is one route to getting more information about the applicants. One way to get more information about both applicants and raters is to work the committee harder: if each committee member read more files, then we would have not only more information per applicant but better estimates of rater-specific biases and reliabilities. Committee members might be understandably resistant to this recommendation, so let me offer some other suggestions that could bring about the same end.

One possibility is to pool information over time. If committee members serve on graduate admissions committees in multiple years, we have the possibility of pooling, so as to learn about their biases and reliabilities (subject to the assumption that these features of the raters are constant over time). Another possibility is to make better use of the committee's resources: rather than having, say, all of n files read by m of the committee members, we might consider a system in which an initial cut is made relatively cheaply (e.g., discarding the bottom half of the applicant pool based on GRE scores and other observable characteristics of applicants known to be associated with applicant quality),[5] and the remaining $n/2$ files are read by $2m$ committee members, producing more information about fewer applicants. This strategy has the strength of directing committee resources toward that set of the applicant pool where we will find excellent to marginal applicants, where difficult decisions have to be made, and where the ability to confidently rank order applicants is most highly valued.

Constraints of time and space mean that numerous questions and interesting extensions remain unaddressed here. First, a sensitivity analysis would let us assess the consequences of having more information about the rater-specific parameters; this is reasonably simple to implement in the Bayesian context by increasing the precision in the prior densities I used for the rater-specific parameters. Second, I do not formally explore the implications of my analysis along decision-theoretic lines. In this conclusion I have used terms such as *risk* quite loosely, without formalizing the decision problem actually faced by graduate

[5] See King et al. (1993) for a report on developing such a procedure for admissions to the Ph.D. program in Harvard's Department of Government.

admission committees. Third, I have not confronted a nagging question underlying the entire graduate admissions process: do we have any ability to assess success in graduate school (and afterward) with the information available to us in graduate admissions files? Answers to this question potentially overshadow the present analysis, but with the exception of work referred to by King et al. (1993), I am not aware of systematic research on this question.[6]

8 Conclusion

Graduate admissions data are sparse and discrete. Different committee members read different applicants' files and bring their own sensitivities and preferences to bear on the process. I have developed a measurement model appropriate to the data, combining the information in the committee members' ordinal ratings with the information in the continuous GRE measures. The model was augmented to estimate the influence of extraneous applicant characteristics on the committee members' ratings. The usual way to estimate a model such as this would be some kind of analysis of covariance structure model (e.g., Jöreskog and Sörbom 1993); but those models are not well suited for small data sets with discrete data that are decidedly nonnormal, and flounder on the fact that not every committee member read every file (certain pairings of committee members never appear in the data, and a complete correlation matrix cannot be formed). Other forms of missing data pose no problem for this modeling approach. For instance, the fact that some applicants did not supply GRE scores requires no special handling or preprocessing of the data; in a Bayesian setting, the relative lack of information for these applicants manifests in greater a posteriori uncertainty as to their latent quality.

The measurement model I develop here is quite general and open to further elaboration (e.g., multidimensional latent traits, exploitation of known characteristics of judges/raters). This application provides a vivid demonstration of the flexibility and extensibility of latent variable models tackled in a fully Bayesian setting. Indicators of different types (binary, ordinal, continuous, counts) and subject to quite extensive patterns of missingness pose no special problems for the approach presented here. The model is easy to apply to the extensive list of measurement challenges confronting political science listed in Section 2.

References

Aldrich, John H., and Richard D. McKelvey. 1977. "A Method of Scaling with Applications to the 1968 and 1972 Presidential Elections." *American Political Science Review* 71:111–130.

Castles, Francis G., and Peter Mair. 1984. "Left-Right Political Scales: Some 'Expert' Judgements." *European Journal of Political Research* 12:73–88.

Clinton, Joshua, Simon Jackman, and Douglas Rivers. Forthcoming. "The Statistical Analysis of Roll Call Data." *American Political Science Review* 98(2).

Clinton, Joshua D., and John S. Lapinski. 2004. "Measuring Significant Legislation: Assessing Congressional Output from 1877–1948." Presented at the Annual Meetings of the Midwestern Political Science Association, Chicago, IL.

Cox, N. R. 1974. "Estimation of the Correlation between a Continuous and a Discrete Variable." *Biometrics* 30:171–178.

Edwards, Yancy D., and Greg M. Allenby. 2003. "Multivariate Analysis of Multiple Response Data." *Journal of Marketing Research* 40:321–334.

Erikson, Robert S. 1990. "Roll Calls, Reputations, and Representation in the U.S. Senate." *Legislative Studies Quarterly* 15:623–642.

[6] Happily, King et al. (1993) conclude that graduate admissions committees are able to predict how well applicants do in graduate school.

Franklin, Charles. 1991. “Eschewing Obfuscation? Campaigns and the Perception of U.S. Senate Incumbents.” *American Political Science Review* 85:1193–1214.

Hoff, Peter, Adrian E. Raftery, and Mark S. Handcock. 2002. “Latent Space Approaches to Social Network Analysis.” *Journal of the American Statistical Association* 97:1090–1098.

Huber, John, and Ronald Inglehart. 1995. “Expert Interpretations of Party Space and Party Locations in 42 Societies.” *Party Politics* 1:73–111.

Jöreskog, Karl G., and Dag Sörbom. 1993. *New Features in LISREL 8.* Chicago: Scientific Software International.

Johnson, Valen E., and James H. Albert. 1999. *Ordinal Data Modeling.* New York: Springer-Verlag.

King, Gary, John M. Bruce, and Michael Gilligan. 1993. “The Science of Political Science Graduate Admissions.” *PS: Political Science and Politics* 26:772–778.

Laver, Michael, and W. Ben Hunt. 1992. *Policy and Party Competition.* London: Routledge.

Law, David S. 2004. Strategic Judicial Lawmaking: An Empirical Investigation of Ideology and Publication on the U.S. Court of Appeals for the Ninth Circuit PhD thesis Department of Political Science, Stanford University.

Loader, C. 1999. *Local Regression and Likelihood.* New York: Springer.

Martin, Andrew D., and Kevin M. Quinn. 2002. “Dynamic Ideal Point Estimation via Markov Chain Monte Carlo for the U.S. Supreme Court, 1953–1999.” *Political Analysis* 10:134–153.

Neal, Radford M. 1997. “Markov Chain Monte Carlo Methods Based on ‘Slicing’ the Density Function.” Technical Report No. 9722. Department of Statistics, University of Toronto.

Noel, Hans. 2004. “The Spatial Structure of Ideological Discourse in America: A Hierarchical Model for Estimating Ideal Points with a Paucity of Data.” Presented at the Annual Meetings of the Midwestern Political Science Association, Chicago, IL.

Olsson, Ulf, Fritz Drasgow, and Neil J. Dorans. 1982. “The Polyserial Correlation Coefficient.” *Psychometrika* 47:337–347.

Poole, Keith T., and Howard Rosenthal. 1997. *Congress: A Political-Economic History of Roll Call Voting.* New York: Oxford University Press.

Rivers, Douglas. 2003. “Identification of Multidimensional Item-Response Models.” Typescript. Department of Political Science, Stanford University.

Spiegelhalter, David J., and S. L. Lauritzen. 1990. “Sequential Updating of Conditional Probabilities on Directed Graphical Structures.” *Networks* 20:579–605.

Spiegelhalter, David J., Andrew Thomas, and Nicky G. Best. 1996. “Computation on Bayesian Graphical Models.” In *Bayesian Statistics 5*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith. Oxford: Oxford University Press, pp. 407–425.

Spiegelhalter, David J., Andrew Thomas, Nicky Best, and Dave Lunn. 2003. *WinBUGS User Manual Version 1.4.* Cambridge, UK: MRC Biostatistics Unit.