Using Statistical Prediction for Estimating Student Ensemble Ratings: Preliminary Report

Prepared for the

Berklee College of Music

Boston, MA

September 30, 2019

Executive Summary

Overview

This report summarizes the findings and analysis activities of a preliminary study intended to examine the feasibility of predicting student ensemble placement ratings using a small set of variables obtained prior to fall ensemble auditions.

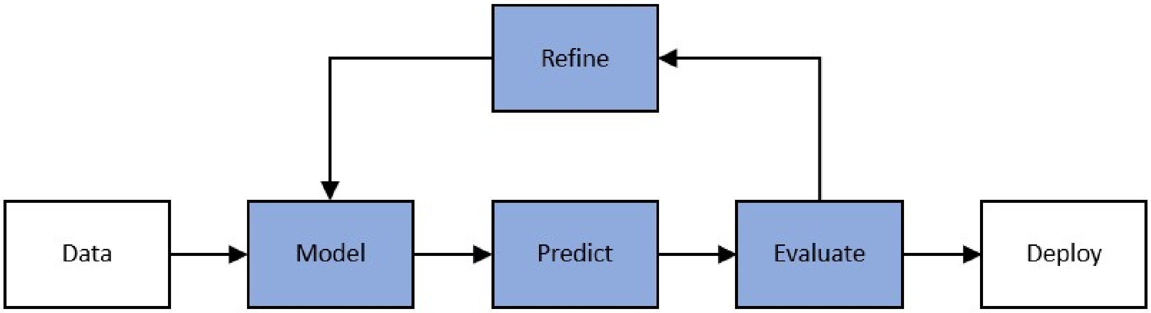
The main findings of this study suggest that, while the prediction models examined exhibited some shortcomings, it is likely that additional data and analyses will produce prediction models that could be useful for informing ensemble placements for a large proportion of students.

However, a non-trivial number of substantive and policy questions should be resolved before operationalizing a statistical prediction system. Additionally, supplemental analyses revealed some potential statistical issues that should be addressed as part of the development of an operational system.

Methodological Approach

The main analysis activities for this study can be viewed as part of an iterative process, as illustrated in the shaded boxes below. A model is selected and fit to the data, predictions are produced from the model, and the prediction results are evaluated. After evaluation, the model is refined, by changing parameters, input variables, model type, or a combination of the three, and the process is repeated until some set of criteria are satisfied.

Figure 1. Iterative process for statistical prediction modeling.



More than 30 models were attempted as part of this study, of which 14 are described and presented in this report. The models were selected to illustrate a range of methods (simple to moderately complex) and prediction outcomes (poor to moderate accuracy).

Key Results

Models were evaluated based on an array of accuracy statistics and an examination of overfitting. While many subjective decisions are typically required to choose a "best" model, one model that showed promise was an ordered probit model with polynomial predictors. Selected results from this model are shown in Table 1.

Table 1. "Within-one" recall rates and weighted *kappa* for ordered probit prediction model with polynomial predictors.

Rating	Full sample	Cross-validated
1	94.8%	94.6%
2	99.1%	99.1%
3	97.2%	97.2%
4	77.2%	77.0%
5	48.1%	47.6%
6	27.3%	25.8%
kappa	0.677	0.673

"Rating" rows display the percentages of students with the indicated rating whose predicted ratings were within one point of their actual rating. *K* (weighted *kappa*) is a measure of how well the model predicts, relative to random assignment. "Cross-validated" values are average values from out-of-sample predictions (5-fold).

Table 1 shows that, among students with actual ensemble ratings ranging from 1 to 3, the model predicted ratings that were within one point of actual ratings for more than 90% of students. The *kappa* statistic, which measures the model's performance relative to random assignment and generally ranges from 0 to 1, is also fairly high for this model, at 0.677. Furthermore, the fact that the "cross-validated" values are not appreciably lower than their corresponding full-sample values suggests that the model may be useful for predictions using future data.

On the other hand, the model is noticeably worse at predicting highly-rated students than students with ratings from 1 to 3, with recall rates declining to 27.3% for students in the highest category. Nearly all of the attempted models had this issue; the only exceptions were those that also suffered from substantial overfitting for these ratings.

While these latter issues are not ideal, they can potentially be addressed by incorporating additional data and attempting to fit additional models. Only eight variables were used to produce the above predictions, and only 30 of potentially hundreds of available models were examined. Given these limitations, it seems plausible that a statistical prediction model can be developed that would be helpful for informing ensemble placements.

Limitations and Caveats

It should be emphasized that "perfect" prediction models are often impossible to produce, and predicted categories should therefore not be used as the sole basis for high-stakes decisions. Such decisions should, at a minimum, take into account measures of prediction uncertainty, and they should also incorporate professional judgment and relevant information that may be difficult to quantify (e.g., notes taken by auditioners).

Additionally, supplemental statistical analyses revealed some potentially problematic findings:

- Cohort effects
- Construct validity issues
- Issues related to prediction error estimation

These issues are detailed in the Key Issues section of this report.

Next Steps

If stakeholders decide to further explore the use of statistical modeling to inform ensemble placement decisions, the following steps are recommended:

1. Discuss a set of procedures for ensemble placement that specifies how statistical predictions might be incorporated. Such procedures might include, for example, individual review of ratings for students with high prediction uncertainty; opportunities for students to request an audition both before and after receiving estimated placement ratings; and a process for diagnosing disagreements between initial predictions and final ratings. This will help to determine a set of properties and features that a minimally-acceptable prediction model should include.
2. Gather, clean, and merge additional data that is likely to improve the quality of predictions, such as auditioner IDs and student background characteristics.
3. If possible, hire or train in-house analysts who are capable of developing, analyzing, and refining the prediction model as more data are observed. Relationships between predictors and outcomes are likely to change over time, especially if participants change their behavior in response to policy changes. Operational prediction models should be adjusted or replaced regularly in order to adjust to such changes. Berklee staff are more likely than external analysts to develop a deep understanding of the college's business processes and policy goals, which will be crucial for assessing the adequacy of an operational prediction system.

The remainder of this report is organized as follows:

- The Data and Prediction Methods sections provide high-level overviews of the analysis dataset, modeling methods, and prediction evaluation criteria.
- The Prediction Results section presents results from a selected set of relatively well-performing prediction models.
- The Key Issues section details statistical and substantive issues that should be addressed.
- The Appendix contains technical details and complete results for the models selected for this study.

The author gratefully acknowledges the assistance of Robert Lagueux and Michael Maieli in producing this study and hopes that readers find the results useful.

Sincerely,

Darrick S. Yee, Ed.D.

Data

The dataset for this study comprised 2,965 student records from cohorts entering in the fall of 2016, 2017, and 2018, of which 2,911 (98.2%) had complete data on all of the required variables. The overall ensemble rating was the outcome of interest, while the predictors consisted of six audition scores, one "scholarship rating" score, and variable indicating the student's principal instrument. The analysis variables are described in more detail below.

Analysis Variables

Overall Ensemble Rating: The outcome of interest for all models. Ensemble ratings are integer values with a theoretical range of 1 to 8. In the dataset, no students received an overall rating of 8, and only six students obtained a rating of 7; none of these were in the 2018 cohort, and all were Piano players. These students were recoded as having a rating of 6, due to their scarcity.

Instrument: A categorical variable indicating the student's principal instrument. Possible values were Bass, Brass, Guitar, Percussion, Piano, Strings, Voice, and Woodwind.

Audition Improv, Prepared Piece, and Overall Score: These audition ratings ranged from 1 to 8 and included decimal values. These were treated as continuous variables.

Audition Reading, Melodic, and Rhythmic Rating: Values for these ratings were None, Beginner, Intermediate, Strong, and Advanced. A single student received one rating labeled "Low-Intermediate," which was recoded as "Beginner." These were treated as either continuous (ranging from 0 to 4) or categorical, depending on the model.

Scholarship Rating: This variable ranged from 2 to 8 and included decimal values. Scholarship rating was treated as continuous.

According to the dataset's accompanying data dictionary, the Overall Ensemble Rating is a composite of four sub-ratings: Reading, Instrumental Skills, Improvisation, and Rhythmic Interpretation. These ratings were not used in the prediction models, as they would not be obtainable prior to fall ensemble auditions. However, they were examined as part of a set of supplemental analyses, as described in the Key Issues section.

Prediction Methods

Broadly speaking, the prediction methods employed below can be described as varying along two dimensions: the *model type*, which specifies the relationship between predictors and ensemble rating; and the *variable set*, the set of predictors used.

Model Types

Ordinary Least Squares regression (OLS)

Classical linear regression, in which the ensemble rating is modeled as a *continuous* variable. Predicted ratings are obtained by rounding (decimal) predictions to the nearest integer and, if necessary, clamping them to range from 1 to 6.

Ordered Probit regression (OP)

Ratings are treated as *ordered categories* (i.e., ranging from lowest to highest), and for each student, the model estimates a probability of assignment to each category. Each student's predicted rating is taken to be the rating with the highest probability. Unlike the OLS model, the OP model does not implicitly assume that ratings are equally-spaced, and decimal or out-of-range values are not possible.

Multinomial Logistic regression (MLOG)

Ratings are treated as *unordered categories*. As with OP regression, the MLOG model produces an estimated probability of assignment to each category for each student. A disadvantage of this method is that it can easily fail to produce a solution (i.e., achieve convergence) when insufficient variation exists in the data.

Variable Sets

Base (14 predictors)

The base set of variables consists of indicator variables for the student's principal instrument, along with audition and scholarship ratings treated as linear, continuous variables. This represents the simplest variable set employed in the study.

Indicator Expansion (23 predictors)

Identical to the Base set, except that audition Reading, Melodic, and Rhythmic ratings are treated as categorical, rather than continuous. This "frees" the model to estimate arbitrary relationships between these three predictors and ensemble ratings. This added flexibility can improve prediction accuracy, but it may also lead to overfitting and convergence difficulties.

Instrument Interaction (63 predictors)

The Base set with interaction terms between instrument indicators and each of the other ratings. This enables the model to estimate a different linear relationship between ratings and the outcome for each instrument; this is substantively similar to estimating a distinct set of relationships between ratings and the outcome for each instrument type.

Polynomial (28 predictors)

The Base set with quadratic and cubic polynomial terms added for each audition and scholarship rating. This enables the estimation of non-linear relationships between ratings and the outcome. This provides more flexibility than the Base set but less than the Indicator Expansion set for Reading, Melodic, and Rhythmic ratings.

An additional modeling approach, called a *support-vector machine*, was also employed as an experiment. The SVM assigns students to rating categories using algorithms that automate the selection of variable transformations and search for variable combinations that best "separate" students into their observed categories.

One drawback of the SVM approach is that results are highly sensitive to user-defined parameters that are difficult to express in substantive terms. Two SVM models were fit for illustrative purposes; a full exploration of SVMs and other automated approaches may yield useful results but is beyond the scope of this study.

Evaluating Prediction Quality

Overall accuracy – the proportion of students whose predicted scores match their actual scores – may be an intuitive statistic for evaluating the quality of predictions, but this statistic is often misleading. In Fall 2018, for example, about 40% of students were assigned an ensemble score of 2; a model that simply predicted a score of 2 for *all* students would therefore have an overall accuracy of 40%, despite using no student information at all to produce its predictions.

Instead, a more detailed set of statistics can be used to assess the accuracy and quality of predictions, as described below.

Accuracy

The primary measures used to assess prediction accuracy in this study were *recall*, *precision*, and Cohen's *kappa*. The first two of these are category-specific, in that they summarize accuracy for each ensemble rating level (1 to 6), while the *kappa* is an "adjusted" measure of overall accuracy.

It is often not possible to find a model that outperforms all others in every one of these accuracy statistics. Because of this, there is usually not a single "best" model, as the relative importance of different metrics is subject to judgment; a model that predicts moderately well for all students, for instance, may be preferable to a model that predicts much better for low-rated students but much worse for highly-rated students.

The accuracy statistics are defined below and described in more detail in the appendix. First, however, an introduction to the *confusion matrix* may be helpful.

Confusion Matrix

The *confusion matrix* is a cross-tabulation of students' predicted ratings vs. their actual ratings. The confusion matrix for the ordered probit model with polynomial predictors is shown in Table 2.

Table 2. Confusion matrix for ordered probit model with polynomial predictors.

		Predicted						
		1	**2**	**3**	**4**	**5**	**6**	*Total*
Actual	**1**	138	338	26	0	0	0	*502*
	2	85	812	209	10	0	0	*1,116*
	3	18	417	300	39	4	0	*778*
	4	0	76	172	78	11	1	*338*
	5	0	14	55	42	17	5	*133*
	6	0	0	13	19	8	4	*44*
	Total	*241*	*1,657*	*775*	*188*	*40*	*10*	*2,911*

The diagonal elements of the confusion matrix in Table 2 (highlighted in blue) contain the counts of students for whom predicted ratings matched actual ratings. Students to the left and right of the diagonal are students for whom predicted ratings were lower and higher, respectively, than their actual ratings.

The confusion matrix for each prediction model is used as the basis for computing statistics for assessing model accuracy.

Recall

The *recall* (also called *sensitivity* in some contexts) for an ensemble rating of *k* is the proportion of students with an actual rating of *k* who whose predicted score was also *k*. In Table 2, this is the number of accurate predictions in a row divided by the total students in the row. For example, 300 students were accurately predicted as having a rating of 3, and the total number of students with an actual rating of 3 was 778 (the row total), resulting in a recall rate of 300/778 = 38.6%. Each model has six recall rates – one for each ensemble rating.

Precision

The *precision* (also called *positive predicted value*) for a rating of *k* is the proportion of students with a predicted rating of *k* whose actual score was also *k*. In Table 2, this would be the number of accurate predictions in a column divided by the column total. As with recall, each model has six precision rates. The precision is technically undefined when a model fails to assign *any* students to a particular rating level. For this study, these outcomes are treated as having a precision of zero.

Cohen's kappa

Cohen's *kappa* statistic, often used as a measure of inter-rater agreement, can be employed to describe the degree to which predicted and actual outcomes match beyond the level that would be expected due to random chance. Its value typically ranges from 0 to 1 (values below 0 are possible, but rare), with values near 0 indicating predictions are not appreciably better than random assignment and 1 indicating perfect prediction.

Weighted kappa

The *weighted kappa* statistic can be used with ordered outcome categories (such as ensemble ratings) to measure how close predicted ratings are to actual ratings for cases where prediction is incorrect. For this study, the weighted *kappa* is used to evaluate how successful each model is at predicting a value within one point of a student's actual ensemble rating.

Overfitting

Prediction models may suffer from *overfitting*, especially when they rely heavily on automation. Overfitting occurs when a model predicts well for the dataset on which it was fit but performs poorly on similar data that was not originally included. A model might, for example, fit past stock prices almost perfectly, while being useless for predicting future stock prices.

A common technique for investigating overfitting is *cross-validation*, in which a model is trained on some subset of the available data, then tested on the remaining "out-of-sample" data. Accuracy statistics from both sets of data can then be compared. Large discrepancies between the original and out-of-sample predictions may indicate the model is overfit.

Cross-validation was conducted for each of the primary models in this study (excluding SVM) using five-fold cross-validation, in which the sample is randomly partitioned into five groups, and accuracy statistics are averaged from five replications of the model, with each replication using four of the five groups as the "training" set and the remaining group as the "test" set. Differences between the original accuracy statistics and the averaged cross-validated statistics may then be used to assess overfitting.

Prediction Results

A total of 14 models are presented in this report: one for each combination of the four variable sets and three primary model types, plus an additional two support-vector machine models. For each of the primary model/variable combinations, an additional five models were fit for cross-validation purposes. Full accuracy results for exact-match predictions are presented in appendix tables A1 and A2.

Exact Matching

Overall, the Ordered Probit (OP) and Multinomial Logistic (MLOG) model types and the Instrument Interaction (Int) and Polynomial (Poly) variable sets tended to produce the most accurate predictions. Selected statistics for two of these models are presented in Table 3.

Table 3. Recall rates and *kappa* statistics for Ordered Probit/Polynomial model and Multinomial Logit/Instrument Interaction model predictions.

	Full sample		Cross-validated		Difference	
R.	OP/Poly	MLOG/Int	OP/Poly	MLOG/Int	OP/Poly	MLOG/Int
1	27.5%	37.6%	26.5%	35.0%	-1.0%	-2.6%
2	72.8%	71.1%	71.4%	67.7%	-1.4%	-3.4%
3	38.6%	37.0%	37.8%	32.3%	-0.8%	-4.7%
4	23.1%	28.7%	21.6%	19.4%	-1.5%	-9.3%
5	12.8%	29.3%	12.8%	17.7%	0.0%	-11.6%
6	9.1%	40.9%	8.1%	10.6%	-1.0%	-30.4%
K	0.220	0.274	0.203	0.203	-0.017	-0.071

Full sample statistics are based on all observations. *Cross-validated* are averages from five-fold cross-validation. *Difference* columns display cross-validated minus full-sample values. Negative differences indicate cross-validated predictions are less accurate than full-sample. R. = Ensemble rating. *K* = Cohen's *kappa*.

Most of the models suffered from low recall rates for students with high ensemble ratings, meaning that students with actual ratings of 5 or 6 were less likely to be accurately predicted than students with lower ratings. For example, in the ordered probit model with polynomial predictors, only about 9% of students who were actually assigned a rating of 6 were accurately predicted.

The multinomial logistic models tended to be the least affected by this issue. However, they also were the most likely to suffer from overfitting. For example, while the multinomial logistic/instrument interaction model accurately predicted ratings for about 41% of students with a rating of 6 in the full sample, this rate declined to less than 11%, on average, in the cross-validation samples – a rate similar to the ordered probit/polynomial model. Large accuracy declines like this are often indicative of overfitting.

Finally, the overall *kappa* statistic was somewhat low for all models (around 0.2 in most cases), indicating that, with respect to exact matches, the prediction models were only somewhat better than random assignment.

Within-1 Matching

Results are more promising when considering the proportion of predictions that are within one point of actual ratings. Table 4 presents recall and *kappa* statistics for the two models in Table 3, but with results modified to use the proportion of students whose predicted scores were within one point of actual scores.

Table 4. Recall rates and *kappa* statistics for Ordered Probit/Polynomial model and Multinomial Logit/Instrument Interaction model predictions, with predicted ratings within one point given full weight.

	Full sample		Cross-validated		Difference	
R.	OP/Poly	MLOG/Int	OP/Poly	MLOG/Int	OP/Poly	MLOG/Int
1	94.8%	92.8%	94.6%	92.9%	-0.2%	0.0%
2	99.1%	97.6%	99.1%	96.8%	0.0%	-0.8%
3	97.2%	93.7%	97.2%	91.1%	0.1%	-2.6%
4	77.2%	74.9%	77.0%	71.6%	-0.2%	-3.3%
5	48.1%	57.9%	47.6%	52.5%	-0.5%	-5.4%
6	27.3%	59.1%	25.8%	48.7%	-1.5%	-10.4%
K	0.677	0.666	0.673	0.605	-0.004	-0.060

See notes for Table 3.
Predicted ratings are considered "correct" if they are within one point of actual ratings.

For both models, recall rates exceeded 90% for ratings of 3 or below, and overall *kappa* values were substantially higher, indicating that the models performed far better than would be expected if predicted scores were randomly assigned. While recall rates were still lower for higher-rated students and the MLOG model still appeared to be overfit, the results suggest that ratings can be closely approximated for most students.

Key Issues

Analyses conducted for this study revealed a number of statistical and substantive issues that should be addressed before using statistical prediction models to inform high-stakes decision making.

Range restriction

As noted in the Data section, overall ensemble ratings have a theoretical maximum of 8, but no students among the three analysis cohorts obtained this rating. Classification prediction models, including ordered probit and multinomial logistic models, cannot predict categories that are not observed in the data, and often have difficulty predicting categories that are highly unusual. (This latter fact motivated the re-coding of the six students with a rating of 7 to a rating of 6.)

Stakeholders should determine the extent to which such range restrictions are acceptable, or if the rating scale may be altered to better align with observed ranges. If no amount of restriction is acceptable, the set of candidate models for prediction may need to be narrowed greatly.

Cohort effects

Ensemble ratings were generally higher for the Fall 2016 and Fall 2017 cohorts, even after restricting scores for these cohorts to a maximum of 6, as described in the Data section. These "cohort effects" persisted even after controlling for the predictor variables. In other words, it appears that, on average, students in cohorts prior to 2018 received higher ensemble ratings than 2018 students with the same principal instrument, audition ratings, and scholarship ratings.

Table 5. Estimated cohort effects on ensemble ratings, relative to Fall 2018 cohort.

	T-tests	Model
Fall 2016	0.250*	0.146*
Fall 2017	0.159*	0.216*

* - Statistically significant at the 0.05 level.
"Model" column displays effects after controlling for main effects of predictor variables.

The presence of cohort effects may be problematic for prediction. Although such effects can be estimated for existing data, they cannot be estimated for future cohorts.

The ideal solution would be to eliminate cohort effects to the extent possible, perhaps via auditioner training or development/improvement of a rating rubric. Statistically, it is possible to predict future values using cohort effects set to zero, set to an average effect, or using other options. Each of these options is likely to introduce some degree of bias, but this may not strongly impact prediction quality if the bias is small.

Measurement issues

The analysis dataset includes four ensemble "sub-scores" for each student, from which overall scores were computed. Although not used in the prediction models, these ratings can be used to examine issues related to measurement, including reliability and validity.

Based on a preliminary principal component analysis (results available on request), it appears that ensemble ratings may be measuring a different construct from the ratings used in the prediction models. This suggests that the construct validity of the prediction models may be limited.

A wide range of factors may explain this finding. For example, the distribution of student abilities may change between the initial audition and ensemble audition, with some students improving more than others. As another example, auditioners may assign ratings using different criteria during the initial audition compared to the ensemble audition, especially if they are looking for different types of skills.

Where reliability is concerned, while the four ensemble sub-scores appear to have high reliability (r > 0.9), the results suggest that ensemble Reading scores may be measuring a different construct from the other three. Based on the provided data, excluding the Reading score from overall ensemble ratings would actually slightly increase the reliability of the average of the remaining three. This may be an indication that the Reading score does not measure the same construct as the others.

The extent to which such validity issues might bias predictions is not immediately clear. In the ideal case, sufficient additional predictors would be added to the models, or auditioner/auditionee behaviors would change, so that the bias would be minimal. In the worst case, ignoring the possible "mismatch" between predictor and outcome constructs may systematically disadvantage some students in favor of others for reasons unrelated to their actual skill.

Prediction error

As with most classification-based prediction models, the results presented above assign students a predicted score based on their “most probable” category. These results do not provide information regarding the estimated probability of an individual student belonging to a different category; however, if predictions are to be used in high-stakes decisions, it is strongly recommended to take this information into account.

Measures of prediction uncertainty can be computed for most prediction models, including all of the models presented above. In some cases, however, additional methods may be required to accurately produce these measures. For example, initial analyses indicate that prediction uncertainty may be higher for some instruments than others, which is not accounted for in the model types above. Extended techniques, such as multi-level random-effects or heteroskedastic ordered probit modeling, may be required to accurately estimate prediction uncertainty.

Changes in participant behavior

Major policy changes have the potential to impact the behavior of participants, including auditioners and auditionees. The latter may prepare differently for auditions if they know that the audition’s goals have changed, and the former may similarly adjust how they assign audition ratings if auditions are to be used for new purposes. Such changes may alter the relationships between predictor variables and “correct” ensemble ratings in ways that cannot be anticipated using purely quantitative methods.

As a result, models based on data acquired before the policy change may predict poorly for students after the policy change. Analysts should be prepared to investigate these effects closely (at a minimum, by comparing pre-change and post-change results) and to regularly review modeling and variable selection algorithms. Additionally, this underlines the importance of refraining from using predicted scores as the sole or primary basis for high-stakes decisions until they can be shown to be robust.

Additional data

Additional data will almost certainly increase the overall quality of predictions. Data in the following categories may be especially useful:

> *Rater identities*. Some auditioners are likely to be more lenient or harsh than others in assigning ratings. Including auditioner identity information in the dataset would enable models to adjust for these effects.
>
> *Student background information*. Student background data may be relevant to ensemble ratings but not captured fully by audition and scholarship ratings. Examples might include the length of time the student has studied the instrument, measures of the student’s proficiency in music theory (e.g., grades or AP Music Theory scores), and the number and types of previous ensembles in which the student has participated.

Appendix

Technical Notes

Models

This section provides formal descriptions of the model types used in this study. For each of the models below, Y_i (an integer ranging from 1 to 6) denotes the overall ensemble rating for student i, X_i is a vector of predictor variables, ε_i is a student-specific error term, and β is a vector of model parameters.

OLS regression

The OLS model can be written as follows:

$$Y_i = \text{round}(Z_i) \tag{1}$$

$$Z_i = \beta' X_i + \varepsilon_i, \tag{2}$$

where Z_i is a continuous version of the ensemble rating, round$(\cdot)$ is the rounding function, and ε_i is drawn from a normal distribution with mean 0, independently and identically distributed (i.i.d.) across students. In the OLS model, β is estimated by minimizing the sum of squared residuals between the observed values of Y_i and estimated values of Z_i.

Ordered probit regression

The ordered probit model can be written:

$$\Pr(Y_i \leq k) = \Phi(V_k - Z_i)$$

$$Z_i = \beta' X_i,$$

where $k \in \{1, \ldots, 6\}$ is an ensemble rating, V_k is a rating-specific cutoff (with $V_6 = \infty$), and $\Phi(\cdot)$ is the standard normal cumulative distribution function (CDF). Equivalently, it can be conceptualized as consisting of a set of cutoff scores, $V_1, \ldots, V_5$, that partition the real line, with students assigned to category k if $Z_i + \varepsilon_i$ falls between V_k and V_{k-1}, where ε_i is a standard normal random variable. β and the five values of V_k are estimated using maximum likelihood.

Multinomial logistic regression

The multinomial logistic model can be described as modeling, for each student, a separate latent variable for each possible rating:

$$Z_{i,1} = \beta_1' X_i + \varepsilon_{i,1}$$

$$Z_{i,2} = \beta_2' X_i + \varepsilon_{i,2}$$

$$\ldots$$

$$Z_{i,6} = \beta_6' X_i + \varepsilon_{i,6},$$

where the $\varepsilon_{i,k}$ are drawn from a type I extreme-value distribution and i.i.d. across both students and ratings. The difference between $Z_{i,1}$ and $Z_{i,2}$ is then

$$Z_{i,1} - Z_{i,2} = \beta_1' X_i + \varepsilon_{i,1} - \left(\beta_2' X_i + \varepsilon_{i,2}\right)$$

$$= (\beta_1 - \beta_2)' X_i + \varepsilon_{i,1} - \varepsilon_{i,2}$$

$$= \beta'_{1v2} X_i + \epsilon_{12} = Z_{i,1v2},$$

where ϵ_{12} takes a standard logistic distribution. The quantity $Z_{i,1v2} - \epsilon_{12}$ is then interpretable as a latent construct reflecting a student's level of "belonging" to rating category 1 relative to 2; similar equations are applied for categories 3 through 6, and the student's observed category is assumed to be the one for which $Z_{i,kv2}$ is highest.

The student's probability of assignment to category $k \neq 2$ is then $\Lambda(Z_{kv}\ )$, where $\Lambda(\cdot)$ is the standard logistic CDF, and the probability of assignment to $k = 2$ is one minus the sum of probabilities for the other categories. A rating of 2 was chosen as the reference category because it was the modal category for the sample. β values were estimated using maximum likelihood.

Cohen's *kappa*

Consider two raters tasked with assigning each student a score of 0 or 1 on a test question. Rater A gives 70% of students a score of 0 and the remaining 30% a score of 1, while Rater B simply flips a coin to determine each student's score, resulting in 50% receiving 0 and 50% receiving 1. Note that the raters' scores are completely independent – they are not "related" in any meaningful way, since Rater B's scores are entirely random.

On average, we would expect 35% of students to receive a score of 0 from *both* raters, since on average half of Rater A's 0-score students would also be given a 0 by Rater B's coin flip. Similarly, 15% of students would be expected to receive a score of 1 from both raters, resulting in an overall "agreement" percentage of 35 + 15 = 50%.

The *kappa* statistic uses this "chance agreement" percentage as its starting point – called the *expected* agreement rate – and measures how much closer the *observed* agreement rate is to 100% (perfect agreement) than the expected rate. Formally, its formula is

$$\kappa = \frac{p_{OBS} - p_{EXP}}{1 - p_{EXP}},$$

where p_{OBS} is the observed overall agreement proportion, and p_{EXP} is the expected proportion. Note that *kappa* is 0 when the observed agreement is the same as the expected ($p_{OBS} = p_{EXP}$) and 1 when the observed agreement is 100% ($p_{OBS} = 1$).

Weighted *kappa*

The classical Cohen's *kappa* statistic effectively awards "full credit" for exact matches between predicted and actual values and "zero credit" for students whose predicted and actual values differ. This is usually desirable when categories are unordered and "distances" between categories are not meaningful. For ordered categories, however, it may be helpful to account in some way for predictions that are close to actual values, but not exact matches.

To achieve this, weights can be assigned to non-exact matches – for example, assigning a weight of 0.5 to non-exact predictions that are within one point of the actual value effectively treats these as a "half of a match" – to produce a *weighted kappa* statistic. Values of 0 and 1 have the same interpretation as for the classical *kappa*; the main difference is that the "baseline" agreement rate is different.

The weights for these computations can be laid out in a matrix with the same dimensions as the confusion matrix. The weight matrix can then be multiplied element-wise with the confusion matrix to produce a count of “accurate” predictions. Exact matching is used when the diagonal elements are 1 and all other elements 0; when elements immediately above and below the diagonal are assigned a weight of 1, predicted values are counted as “accurate” when they are within one point of the actual value.

Table A. Weight matrices for weighted *kappa* computations.

		Predicted					
		1	**2**	**3**	**4**	**5**	**6**
Actual	**1**	1	0	0	0	0	0
	2	0	1	0	0	0	0
	3	0	0	1	0	0	0
	4	0	0	0	1	0	0
	5	0	0	0	0	1	0
	6	0	0	0	0	0	1

		Predicted					
		1	**2**	**3**	**4**	**5**	**6**
Actual	**1**	1	1	0	0	0	0
	2	1	1	1	0	0	0
	3	0	1	1	1	0	0
	4	0	0	1	1	1	0
	5	0	0	0	1	1	1
	6	0	0	0	0	1	1

Weight matrices are multiplied element-wise with the confusion matrix to produce a count of “accurate” predictions.
Top: Exact matching.
Bottom: “Within-one” matching.

Note that weighting is highly flexible, in that customized weights can be assigned as appropriate. For example, if stakeholders find over-prediction acceptable but not under-prediction, weights can be assigned to reflect this; weights can be made to decline as predicted values get farther away from actual values; and so on.

Result Tables

Table A1. Prediction accuracy statistics by model type and variable set.

	OLS Regression					Ordered Probit					Multinomial Logistic					S-V Machine	
																Gamma	Gamma
	Base	Indic.	Inter.	Poly.		Base	Indic.	Inter.	Poly.		Base	Indic.	Inter.	Poly.		0.05	0.25
Rat.	Recall				Rat.	Recall				Rat.	Recall				Rat.	Recall	
1	0.227	0.211	0.219	0.155	1	0.299	0.285	0.279	0.275	1	0.357	0.353	0.376	0.349	1	0.307	0.345
2	0.601	0.614	0.651	0.643	2	0.715	0.717	0.743	0.728	2	0.711	0.706	0.711	0.714	2	0.832	0.782
3	0.530	0.521	0.486	0.491	3	0.388	0.384	0.377	0.386	3	0.341	0.346	0.370	0.361	3	0.217	0.362
4	0.320	0.311	0.317	0.308	4	0.216	0.210	0.210	0.231	4	0.254	0.275	0.287	0.275	4	0.098	0.222
5	0.083	0.105	0.143	0.128	5	0.098	0.113	0.158	0.128	5	0.158	0.165	0.293	0.173	5	0.008	0.218
6	0.000	0.000	0.023	0.023	6	0.091	0.091	0.114	0.091	6	0.295	0.273	0.409	0.318	6	0.000	0.341
	Precision					Precision					Precision					Precision	
1	0.582	0.596	0.655	0.553	1	0.591	0.588	0.601	0.573	1	0.556	0.562	0.561	0.545	1	0.512	0.579
2	0.503	0.501	0.505	0.494	2	0.491	0.487	0.498	0.490	2	0.507	0.501	0.514	0.505	2	0.459	0.508
3	0.382	0.381	0.390	0.381	3	0.368	0.372	0.389	0.387	3	0.369	0.377	0.417	0.401	3	0.335	0.427
4	0.382	0.378	0.369	0.369	4	0.415	0.397	0.378	0.415	4	0.381	0.401	0.449	0.404	4	0.423	0.500
5	0.611	0.560	0.442	0.531	5	0.481	0.469	0.356	0.425	5	0.429	0.478	0.453	0.460	5	0.500	0.558
6	0.000	0.000	0.500	1.000	6	0.400	0.364	0.357	0.400	6	0.406	0.387	0.514	0.452	6	0.000	0.517
	Kappa					Kappa					Kappa					Kappa	
	0.219	0.217	0.229	0.206		0.216	0.211	0.227	0.220		0.236	0.238	0.274	0.249		0.162	0.269
	Weighted kappa					Weighted kappa					Weighted kappa					Weighted kappa	
	0.689	0.695	0.732	0.690		0.677	0.676	0.706	0.677		0.654	0.660	0.666	0.648		0.448	0.614

N=2,911 for all models. Darker cell shades correspond to higher within-row values for each model. Base, Indic. (Indicator Expansion), Inter. (Instrument Interaction), and Poly. (Polynomial) refer to variable sets. Weighted kappa assigns a weight of 1 to students for whom predicted ratings were within 1 point of actual ratings (in addition to students whose actual and predicted ratings were equal).

Table A2. Prediction accuracy statistics by variable set and model type.

	Base				Indicator Expansion				Instrument Interaction				Polynomial		
	OLS	OP	MLOG		OLS	OP	MLOG		OLS	OP	MLOG		OLS	OP	MLOG
Rat.	Recall			Rat.	Recall			Rat.	Recall			Rat.	Recall		
1	0.227	0.299	0.357	1	0.211	0.285	0.353	1	0.219	0.279	0.376	1	0.155	0.275	0.349
2	0.601	0.715	0.711	2	0.614	0.717	0.706	2	0.651	0.743	0.711	2	0.643	0.728	0.714
3	0.530	0.388	0.341	3	0.521	0.384	0.346	3	0.486	0.377	0.370	3	0.491	0.386	0.361
4	0.320	0.216	0.254	4	0.311	0.210	0.275	4	0.317	0.210	0.287	4	0.308	0.231	0.275
5	0.083	0.098	0.158	5	0.105	0.113	0.165	5	0.143	0.158	0.293	5	0.128	0.128	0.173
6	0.000	0.091	0.295	6	0.000	0.091	0.273	6	0.023	0.114	0.409	6	0.023	0.091	0.318
	Precision				Precision				Precision				Precision		
1	0.582	0.591	0.556	1	0.596	0.588	0.562	1	0.655	0.601	0.561	1	0.553	0.573	0.545
2	0.503	0.491	0.507	2	0.501	0.487	0.501	2	0.505	0.498	0.514	2	0.494	0.490	0.505
3	0.382	0.368	0.369	3	0.381	0.372	0.377	3	0.390	0.389	0.417	3	0.381	0.387	0.401
4	0.382	0.415	0.381	4	0.378	0.397	0.401	4	0.369	0.378	0.449	4	0.369	0.415	0.404
5	0.611	0.481	0.429	5	0.560	0.469	0.478	5	0.442	0.356	0.453	5	0.531	0.425	0.460
6	0.000	0.400	0.406	6	0.000	0.364	0.387	6	0.500	0.357	0.514	6	1.000	0.400	0.452
	Kappa				Kappa				Kappa				Kappa		
	0.219	0.216	0.236		0.217	0.211	0.238		0.229	0.227	0.274		0.206	0.220	0.249
	Weighted kappa				Weighted kappa				Weighted kappa				Weighted kappa		
	0.689	0.677	0.654		0.695	0.676	0.660		0.732	0.706	0.666		0.690	0.677	0.648

N=2,911 for all models. OLS = Ordinary least squares. OP = Ordered probit. MLOG = Multinomial logistic. Darker cell shades correspond to higher within-row values for each model. Weighted kappa assigns a weight of 1 to students for whom predicted ratings were within 1 point of actual ratings (in addition to students whose actual and predicted ratings were equal).

Table A3. Mean differences between primary and 5-fold cross-validated fit statistics, by variable set and model type.

	Base				Indicator Expansion				Instrument Interaction				Polynomial		
	OLS	OP	MLOG		OLS	OP	MLOG		OLS	OP	MLOG		OLS	OP	MLOG
Rat.		Recall		Rat.		Recall		Rat.		Recall		Rat.		Recall	
1	-0.017	-0.011	-0.021	1	-0.003	0.001	-0.020	1	-0.011	-0.010	-0.026	1	-0.004	-0.010	-0.007
2	-0.006	-0.003	-0.009	2	0.002	-0.009	-0.008	2	-0.006	-0.020	-0.034	2	-0.002	-0.014	-0.029
3	-0.003	0.007	-0.032	3	-0.015	0.009	-0.032	3	-0.012	-0.002	-0.047	3	0.000	-0.008	-0.036
4	-0.006	0.006	-0.018	4	-0.014	0.003	-0.015	4	-0.017	-0.006	-0.093	4	-0.008	-0.015	-0.006
5	-0.009	0.030	-0.029	5	-0.001	0.019	-0.034	5	-0.023	-0.022	-0.116	5	0.006	0.000	-0.025
6	0.000	-0.010	-0.056	6	0.033	0.008	-0.147	6	0.011	-0.033	-0.304	6	0.011	-0.010	-0.148
		Precision				Precision				Precision				Precision	
1	-0.013	-0.006	-0.027	1	0.004	-0.013	-0.025	1	-0.057	-0.038	-0.056	1	0.003	-0.004	-0.028
2	-0.003	-0.001	-0.014	2	-0.002	0.001	-0.011	2	-0.002	-0.009	-0.021	2	-0.001	-0.009	-0.015
3	-0.007	0.001	-0.025	3	-0.010	0.003	-0.023	3	-0.011	-0.004	-0.047	3	-0.001	-0.014	-0.041
4	-0.007	0.019	-0.015	4	-0.010	0.021	-0.024	4	-0.017	-0.027	-0.129	4	-0.001	-0.007	-0.005
5	-0.051	0.053	-0.097	5	-0.060	-0.033	-0.171	5	-0.066	-0.016	-0.227	5	-0.043	-0.032	-0.122
6	0.000	-0.067	-0.128	6	0.200	0.153	-0.207	6	-0.400	-0.200	-0.412	6	-0.900	-0.183	-0.182
		Kappa				Kappa				Kappa				Kappa	
	-0.010	0.000	-0.031		-0.008	0.002	-0.030		-0.014	-0.016	-0.071		-0.003	-0.017	-0.036

Cell values display means across five test sets. Negative values indicate lower accuracy for test sets than for the full-sample model. Darker shades correspond to lower within-row values.