What does a good teacher sound like?

Using machine learning and voice to predict teacher effectiveness

William Arbour
University of Toronto

Marlène Koffi
University of Toronto

Philip Oreopoulos
University of Toronto

General outline for today's talk

- Quick review of research on teacher evaluation
- Motivation for using machine learning and audio
- Data and methodology
- Descriptive statistics
- Results: how audio variables predict teacher VA
- Comparison with human raters
- Directions for future research

Motivation

- Variance decomposition of student test scores suggests that teachers are the most important component of a school's effect on student learning (e.g. Hattie, 1999)
- Teacher value-added (expected test score change from teacher assignment) often used as overall measure of teacher effectiveness and exhibits considerable heterogeneity within and across schools ($\sigma \approx 0.2$)
- Identifying variables that predict teacher effectiveness is essential for hiring, promotion, accountability, professional development, and for suggesting mechanisms behind education production function

Past efforts to evaluate teacher effectiveness

- 1910-60s: teacher evaluations primarily based on personal attributes measured outside the classroom (e.g. education, experience, demeanor) [Goe et al. 2008]
- 1960-80s: greater focus on classroom activity and lessons, with series of small experimental and correlational studies examining isolated effects on test score gains (e.g. quality of feedback, time management, homework) [Hattie, 1999; Brophy et al, 1984]
- 1980-90s: evaluations based on detailed classroom observation and aggregating across multiple categories (e.g. environment of respect and rapport, frequency and quality of engagement, feedback, etc...) [Downer et al., 2008]
- 2000-20s: teacher value-added used as validity check for whether evaluations or characteristics are good predictors of teacher effectiveness [e.g. Harris and Sass, 2009]

Example from the Measures of Effective Teaching (MET) Project

- 3,000 teacher volunteers across 6 U.S. districts in 2009-2011, video recorded 4-8 times with 360 deg. camera and evaluated by principals, students, and certified assessors
- Common observation instruments used: e.g. Classroom Assessment Scoring System (CLASS), assesses quality of instructional emotional support, classroom organization, and student engagement
- Math and reading assessments administered to grades 4-8 students in participating classes
- Random assignment of teachers within school and grade in year 2 used to validate value-added estimates in year 1

> *Each of the observation instruments is related to student achievement gains, but the correlations were modest.*
>
> *MET report*

e.g. CLASS scores correlated 0.18 with prior-year value added; difference in achievement gains between top and bottom 25 percent in CLASS score: 8% SD

> *We did not find evidence that any particular set of competencies within the five instruments were more important than the others or that some should be dropped.*
>
> *RAND study*

Machine learning (ML) to predict teacher effectiveness

Conditions are ripe for applying ML to predict teacher effectiveness:

- Computers may be able to better predict patterns than humans
- New techniques have made this easier to do
  - Supervised techniques can be used for *reducing dimensionality* and more objectively identifying which observed characteristics best predict VA, which can be clustered into similar categories, and which matter least
  - Unsupervised techniques can be used for processing data too large for humans to assess (e.g. video, audio, text) to find patterns that assist with accurate prediction
- Prediction methods do not involve costly, timely, and subjective assessors (the CLASS rating system requires 17-25 hours of training)

Text-as-data examples

- Liu and Cohen (2021)
  - transcribes 976 videos of 258 Grade 4-5 ELA teachers from MET data
  - teacher-centric var. (greater use of "I", higher proportion of time teacher talks) slightly negatively correlated with VA (1 SD decrease = 0.04 VA increase)
  - classroom management and interactive instruction variables not sign. related with VA
- Kelly et al. 2018
  - automates transcription of 132 ELA classroom lectures in 2013-16
  - trains computer to detect *authentic* (engaging?) questions, which many evaluation protocols value as good teacher traits
  - 0.68 correlation between human and computer coders

Using audio

- Wide variance in phys. features of human voice; e.g., frequency, rhythm, continuity, volume, brightness, speed, pitch, ...
- Humans convert these physical features to perceptual features; e.g., emotion like anger, confidence, happiness, tenderness, masculinity, dominating, ...
- Small number of studies have tried to make the connection:
  - music (Friberg et al. 2014)
  - passion and masculinity from politicians (Dietrich and Juelich, 2018, Anderson and Klofstad 2012)
  - trust (Klofstad and Peters (2012)

Studies on audio and education

- Some story teller's voices easier to pay attention to (Lyberg-Åhlander et al., 2015)
- Slower speech increases language comprehension (Haake et al., 2014)
- Native-English-speaking TAs increase accessibility of lecture and rapport (Pickering, 2001)

Why focusing on voice?

- Potentially a key factor of the teaching-learning process (Rodrigues et al., 2016)
  - 50% to 90% of time spent listening to the teacher
  - Some voices found to be more effective (Schmidt et al., 1998)
  - Affects comprehension (e.g., speed in Lyberg-Åhlander et al., 2015; frequency of pauses in Schmidt et al., 1998)
- Could change perceptions and behavior (Amir and Levine-Yundof, 2013)
- Corresponds to different ways of teaching
  - proportion of time spent speaking
  - interruptions

This study: our contributions

- Explore predictive power of physical features in teacher voices on value-added
- Use ML to determine strongest relationships (supervised and unsupervised*)
- Compare predictive power with that from classroom observation assessments
- Focus on top and bottom value-added teachers

* not shown today

Process and procedure

- Obtain data with classroom recordings and value added estimates
- Convert videos to audio files and isolate teacher's voice
- Convert teacher voice to physical feature variables
- Train computer to predict VA based on these variables
- Examine predictive power of audio variables
- Train computer to predict VA based on human assessor observations
- Examine predictive power of these variables and compare with audio results

MET data

Largest study of classroom teaching ever conducted, sponsored by Gates Foundation

- Began in 2009-10, lasting 2-4 years
- 6 large U.S. school districts
- 315 schools
- 3,000 volunteer teachers, Gr 4-9, Math and ELA
- 6,500 sections
- 45,000 students

MET data

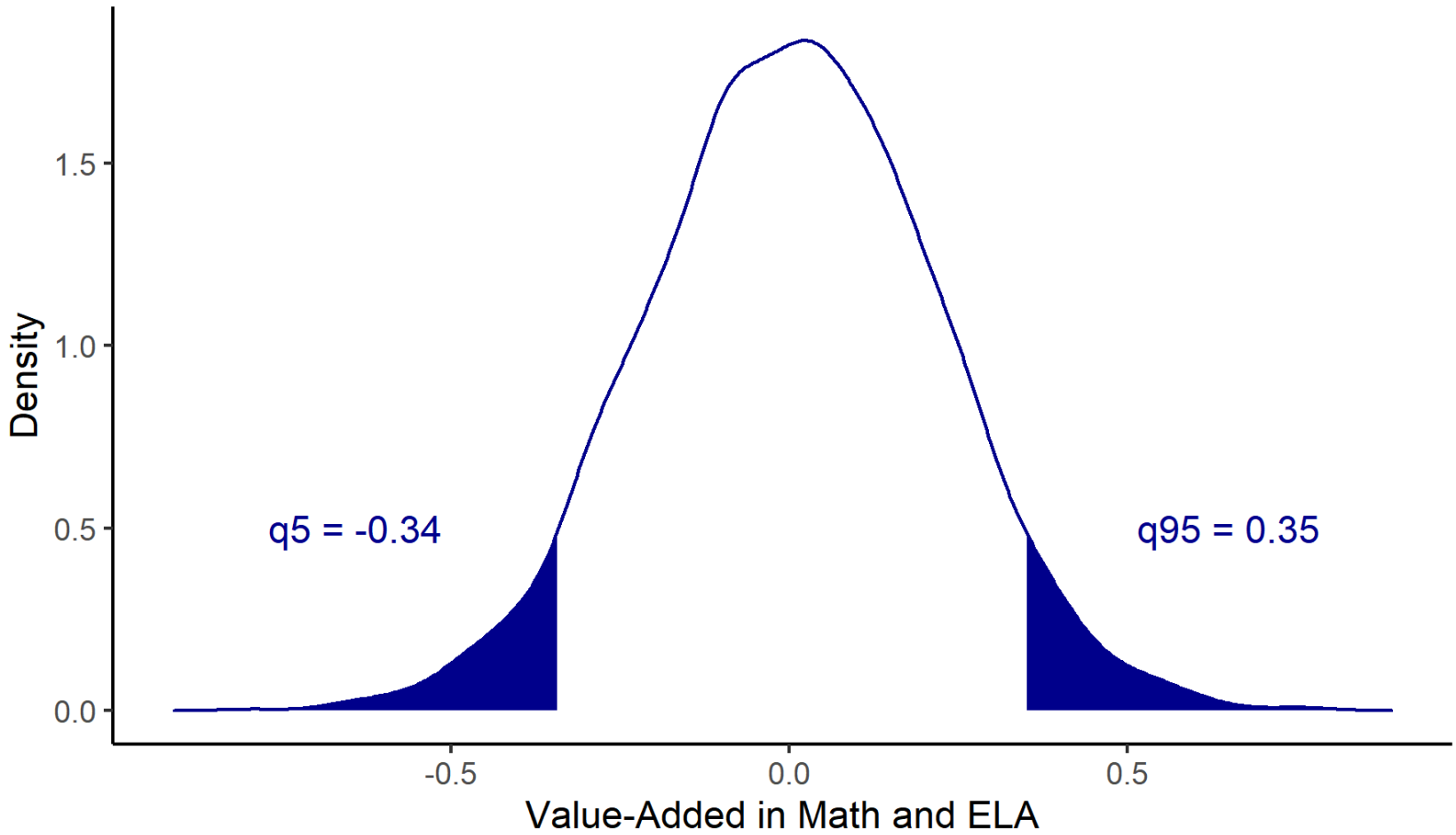
Classroom recordings

- 11,541 videos of 1,470 consenting MET teachers
- Some videos dropped or omitted due to lack of quality or recording issues
- Videos used to calculate multiple classroom observation instruments
- 1,000 assessors, extensive training, multiple validation checks, shift work

Student achievement outcomes

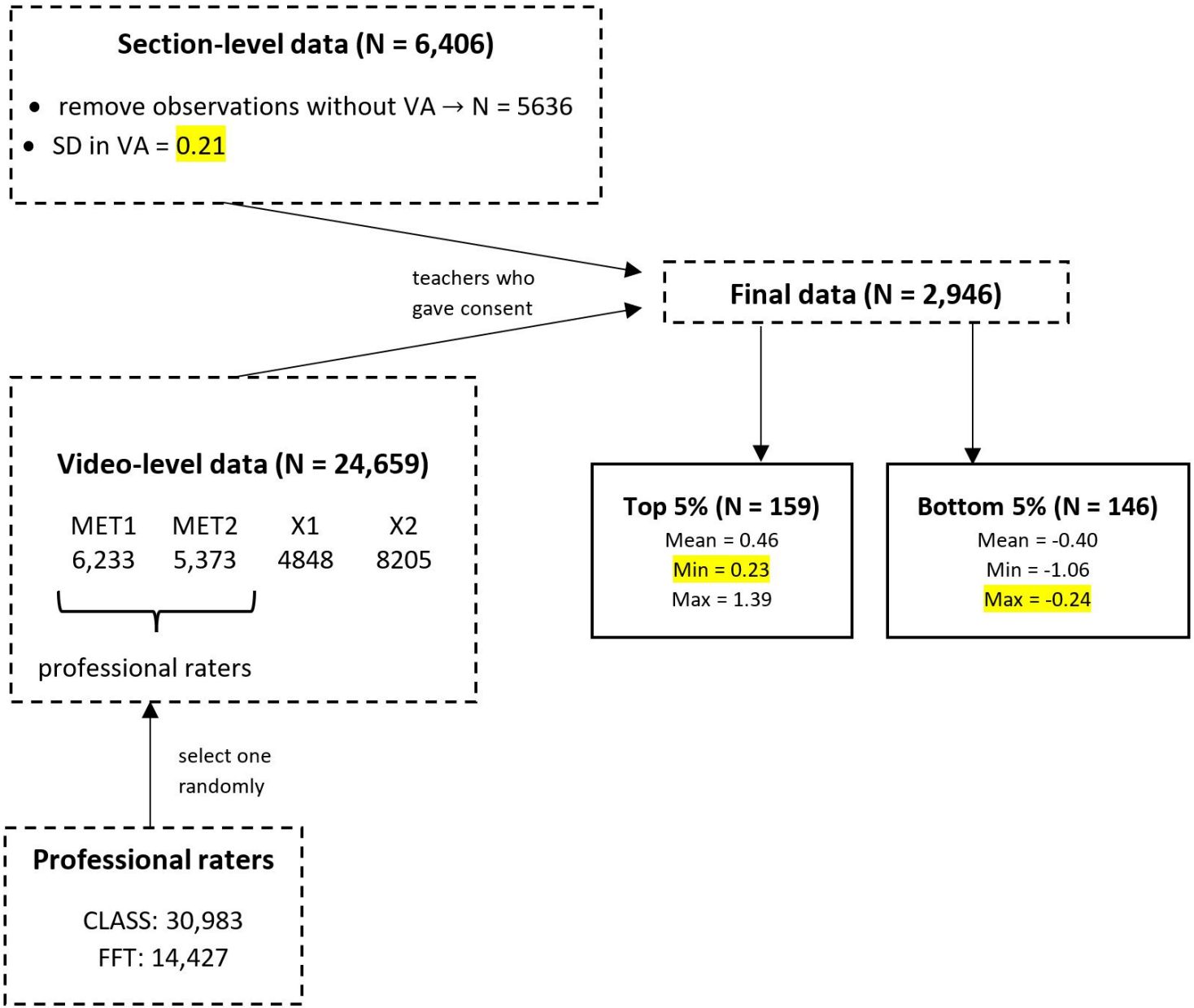
- Mostly from admin. state test scores in Math and ELA
- Used to compute value-added, within district, grade, and subject
- Also recorded student and principal assessments

MET data

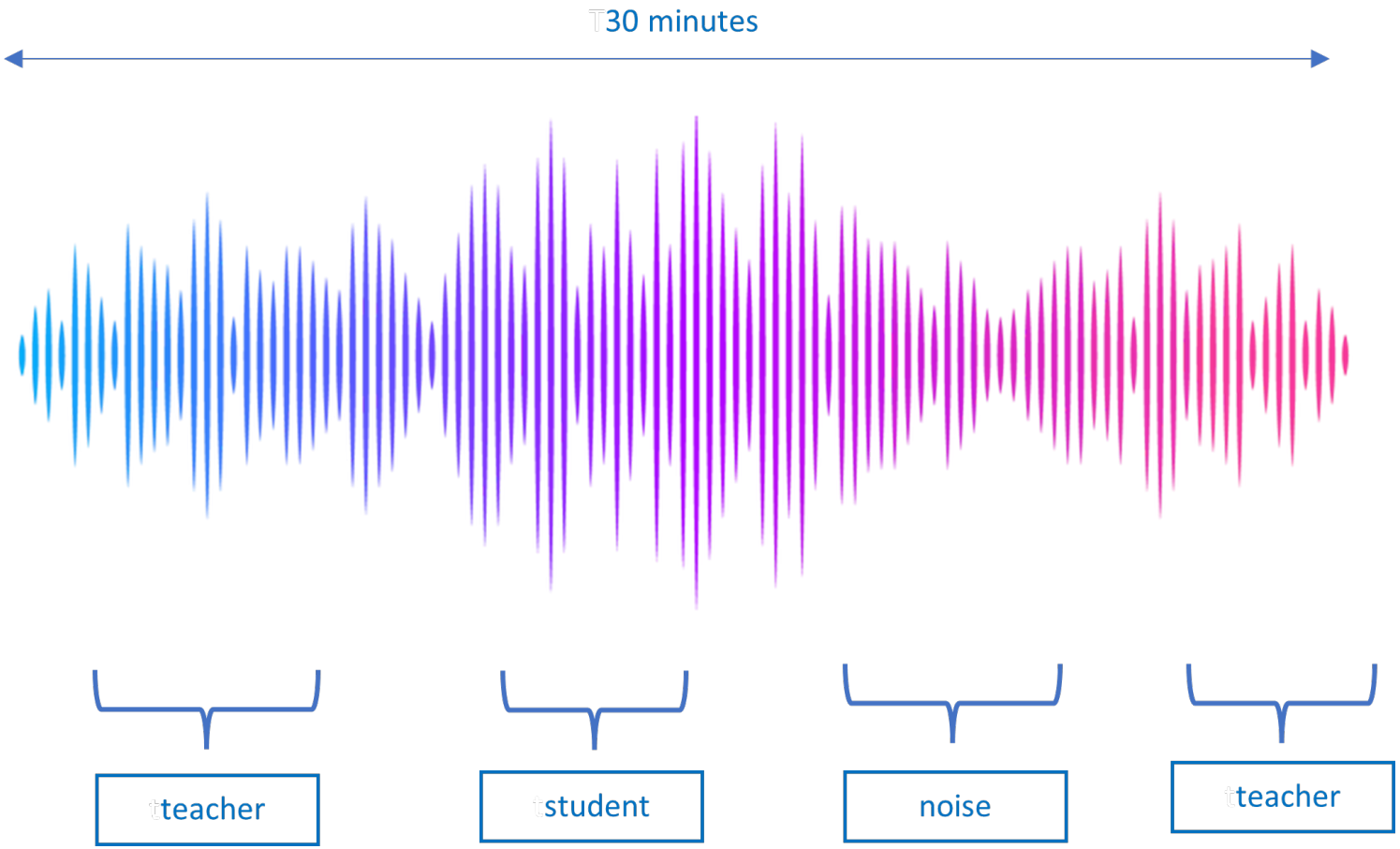


Mathematics | English | Details on VA calculations

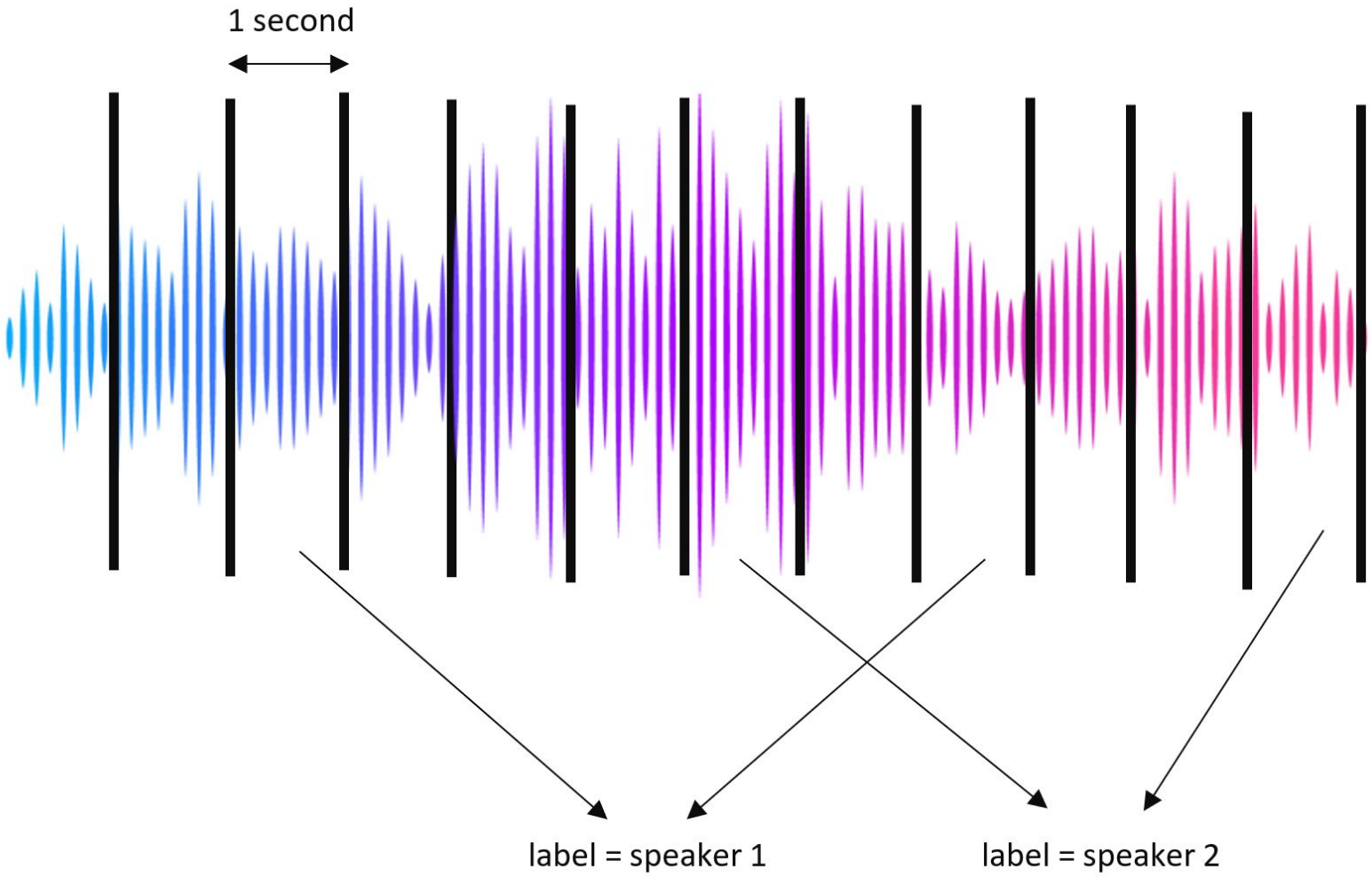
MET data: sample



Methodology: isolating the teachers' voices



Methodology: isolating the teachers' voices



Hypothesis: teacher $\equiv$ most frequent speaker

More details on the diarization process

Methodology: from audio files to variables

- Once the teacher is identified $\rightarrow$ select a clip of 15 seconds
- We extract features from the voice ($\approx$ 30 variables + higher moments) Full list
  - Pitch: how *high* the voice is
  - Spectral centroid: how *bright* the voice is
  - Zero-crossing rate: how *smooth* the voice is
- From the overall 30-minute segment, we measure the proportion of time spoken by the teacher (v.s. students)

How audio relates to CLASS variables

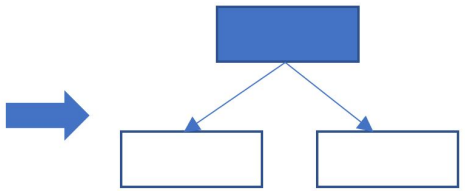
Table: Correlations between voice features and CLASS

	Pitch	Spectral Centroid	Zero-Crossing Rate	Proportion Time Spoken
Positive Climate	-0.01	0.02	-0.00	-0.06
Negative Climate	0.17**	0.08**	0.14**	-0.02
Teacher Sensitivity	-0.00	0.02	-0.01	-0.07
Regard for Student Perspectives	0.09**	0.10**	0.10**	-0.11**
Behavior Management	-0.15**	-0.09**	-0.13**	0.052
Productivity	-0.03	-0.04	-0.03	0.02
Instructional Learning Format	0.05	0.02	0.03	-0.05
Content Understanding	0.04	0.06	0.07	-0.11**
Analysis and Problem Solving	0.03	0.05	0.06	-0.09**
Quality of Feedback	0.02	0.03	0.03	-0.12**
Student Engagement	0.01	0.05	0.03	-0.0009

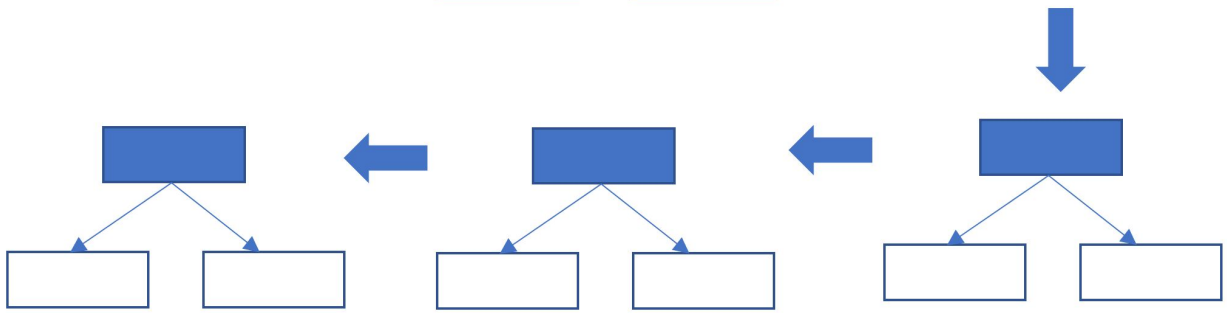
Methodology: gradient boosting

Random forest: bagging (independent trees)
Gradient trees: boosting (dependent trees)

teacher	effective	voice feature
A	1	-18
B	0	12
C	0	3

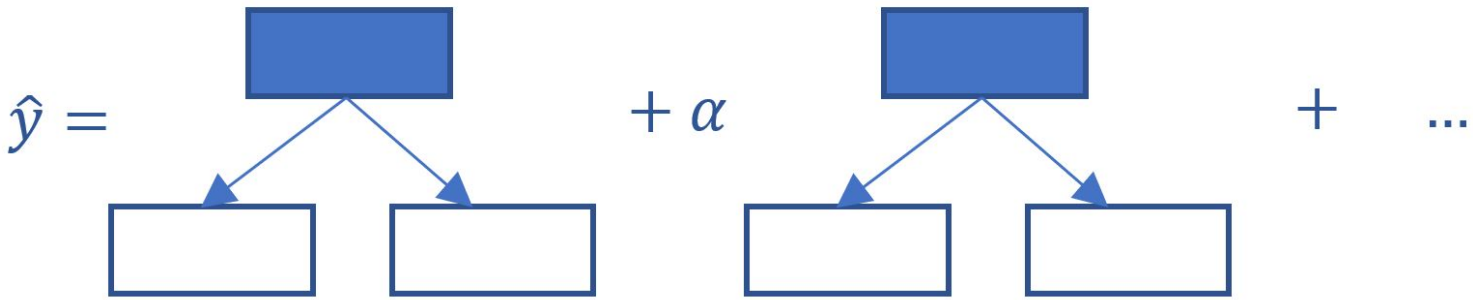


teacher	effective	voice feature	residual
A	1	-18	$1 - \hat{y}$
B	0	12	$0 - \hat{y}$
C	0	3	$0 - \hat{y}$



Methodology: gradient boosting

Trees are weighted by a learning rate (α)



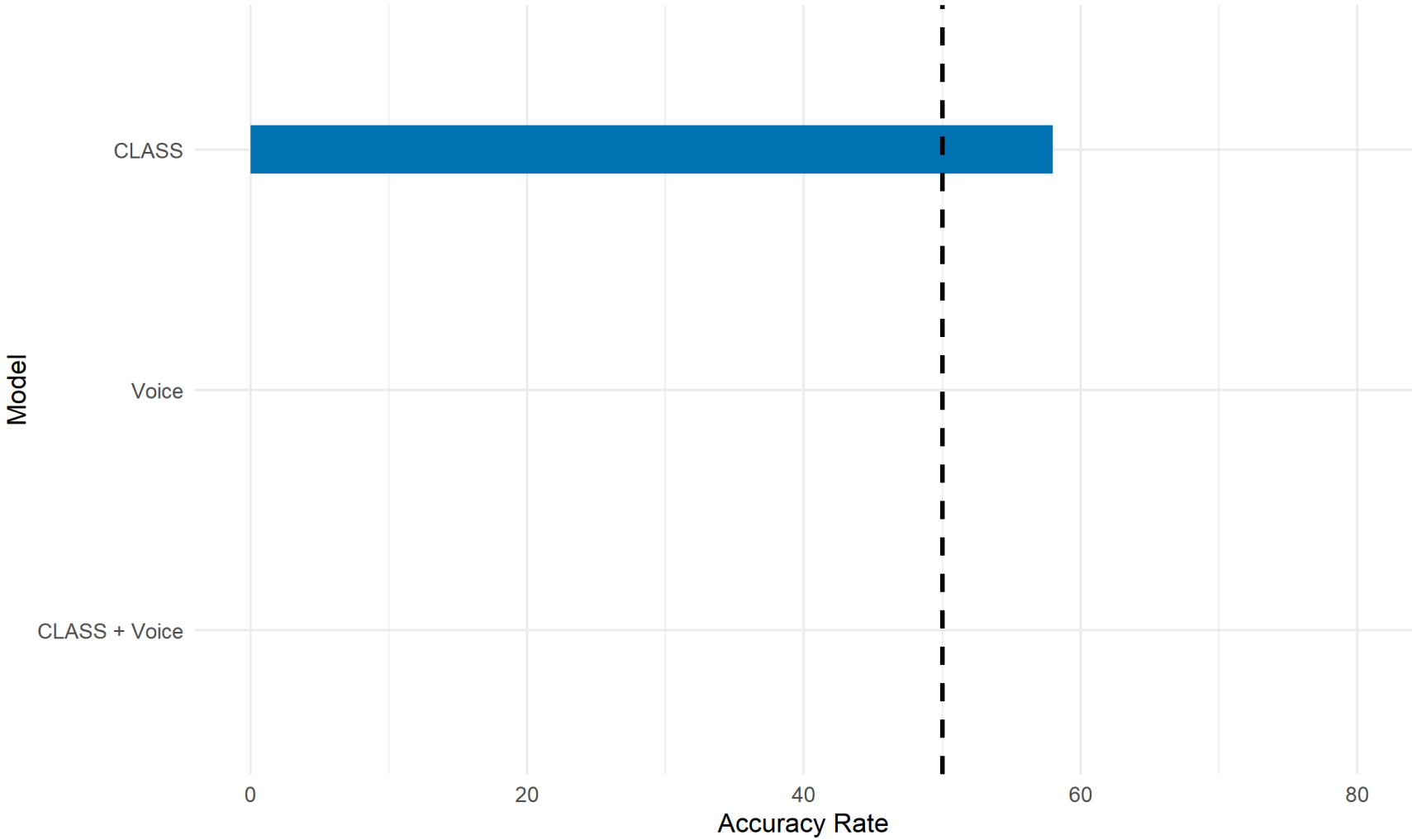
Outcome: individual probability of being an effective teacher solely based on voice variables

Do audio variables predict value-added?

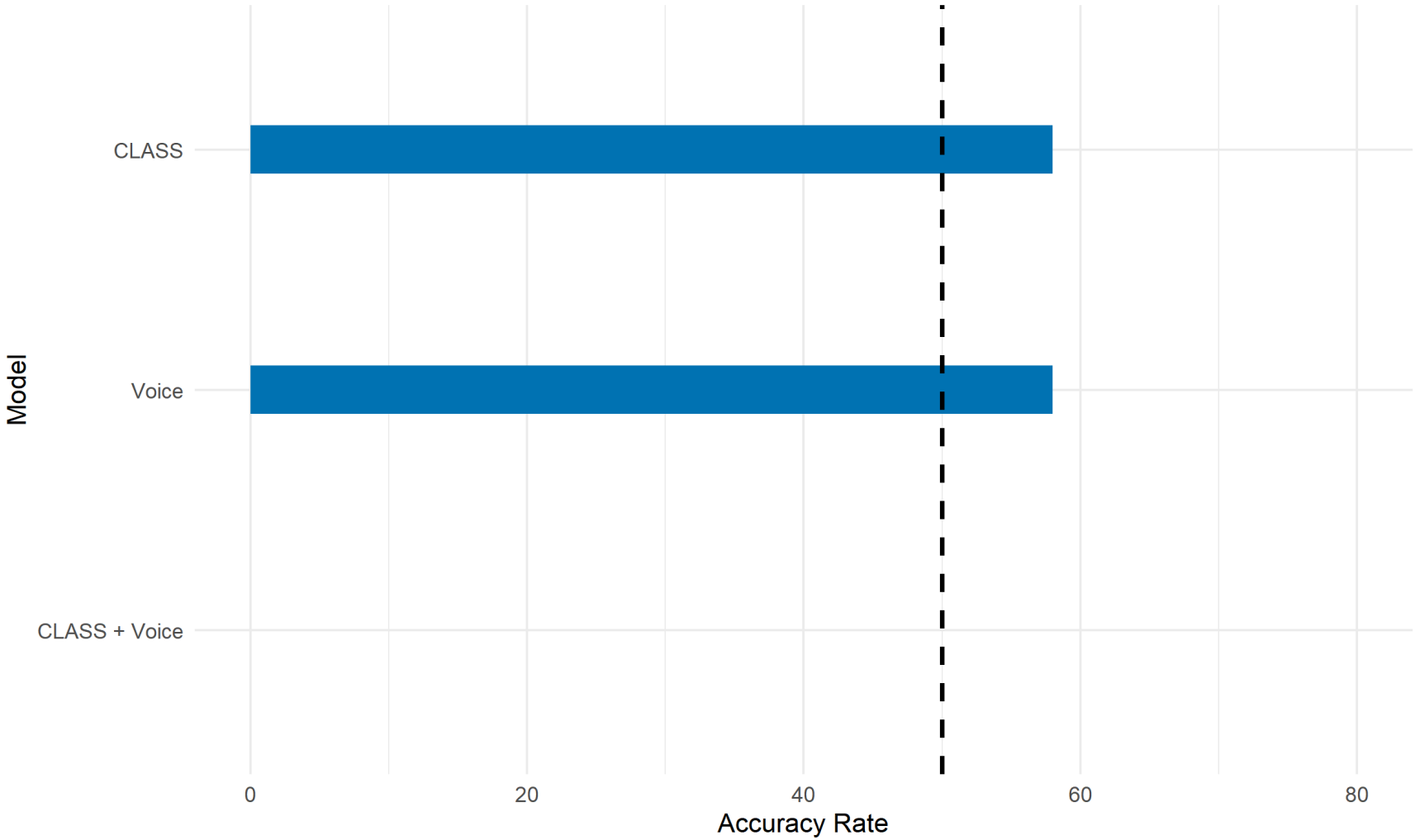
Table: Correlations between voice features, CLASS and value-added

	Pitch	Spectral Centroid	Zero-Crossing Rate	Proportion Time Spoken	Predicted Prob.
Positive Climate	-0.01	0.02	-0.00	-0.06	0.06
Negative Climate	0.17**	0.08**	0.14**	-0.02	-0.03
Teacher Sensivity	-0.00	0.02	-0.01	-0.07	-0.03
Regard for Student Perspectives	0.09**	0.10**	0.10**	-0.11**	-0.07
Behavior Management	-0.15**	-0.09**	-0.13**	0.052	0.11**
Productivity	-0.03	-0.04	-0.03	0.02	0.09**
Instructional Learning Format	0.05	0.02	0.03	-0.05	0.02
Content Understanding	0.04	0.06	0.07	-0.11**	-0.11**
Analysis and Problem Solving	0.03	0.05	0.06	-0.09**	-0.07
Quality of Feedback	0.02	0.03	0.03	-0.12**	-0.03
Student Engagement	0.01	0.05	0.03	-0.0009	0.08*
High/low value-added (dummy)	-0.05	0.018	-0.01	-0.02	0.50***

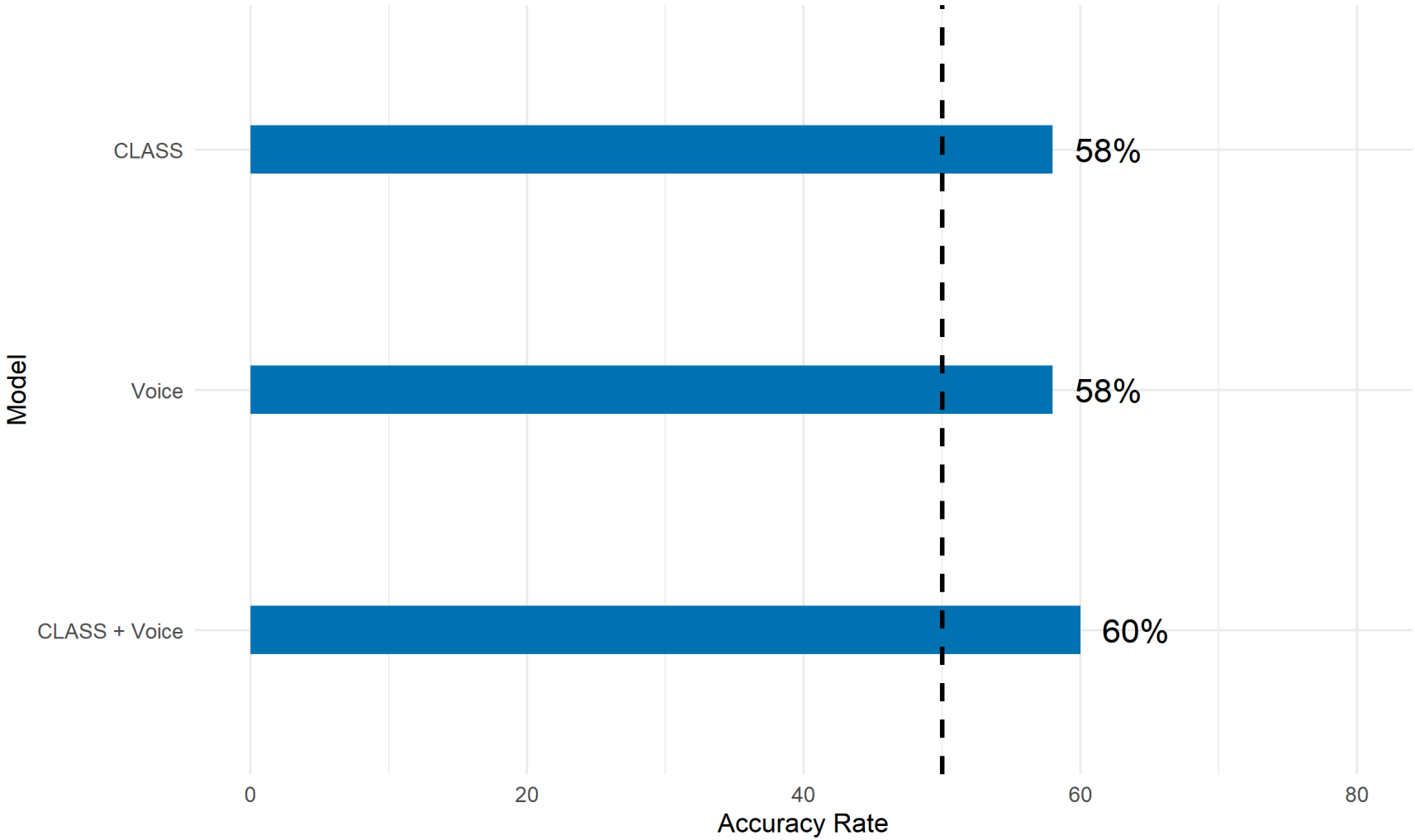
Results: Accuracy (out-of-sample)



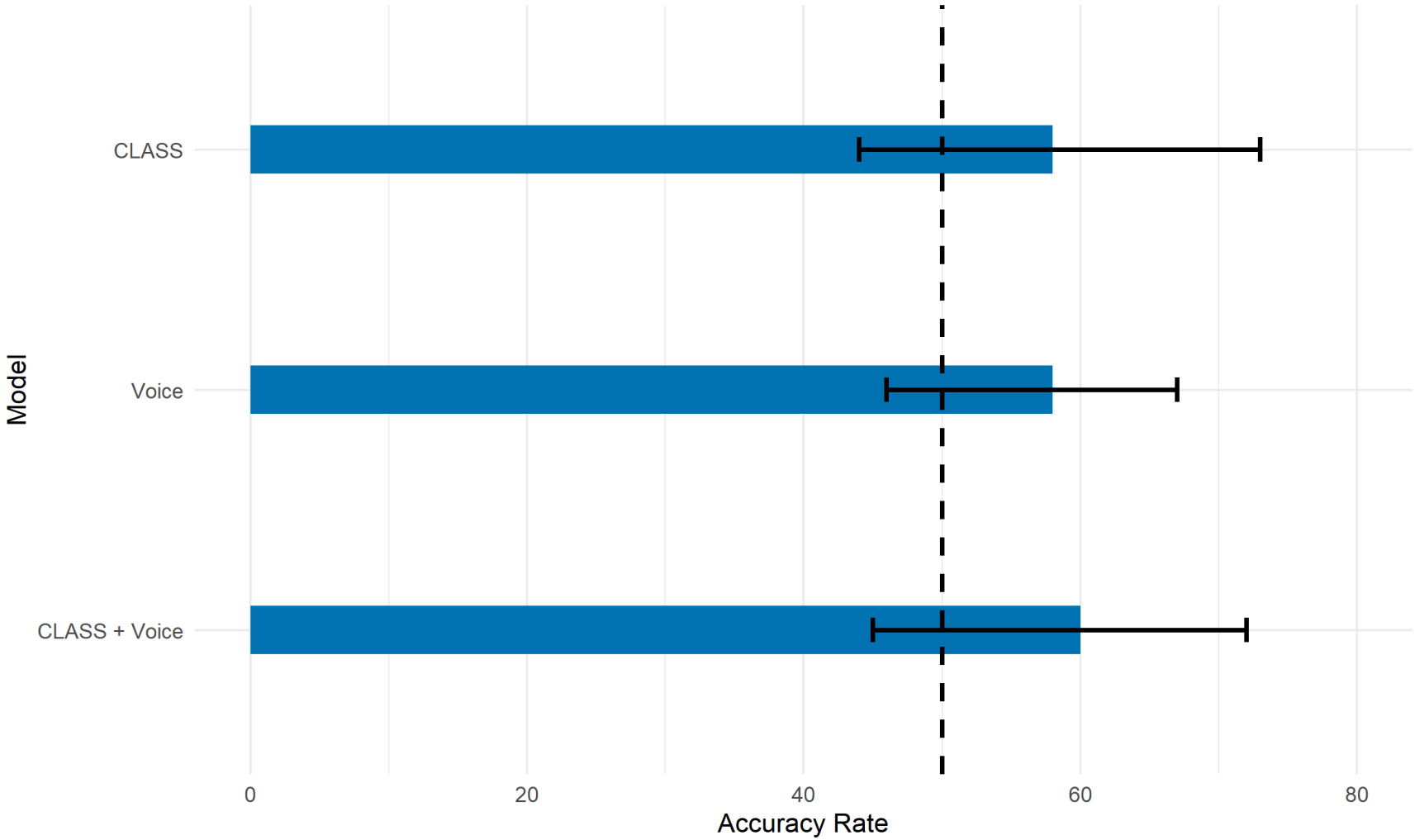
Results: Accuracy (out-of-sample)



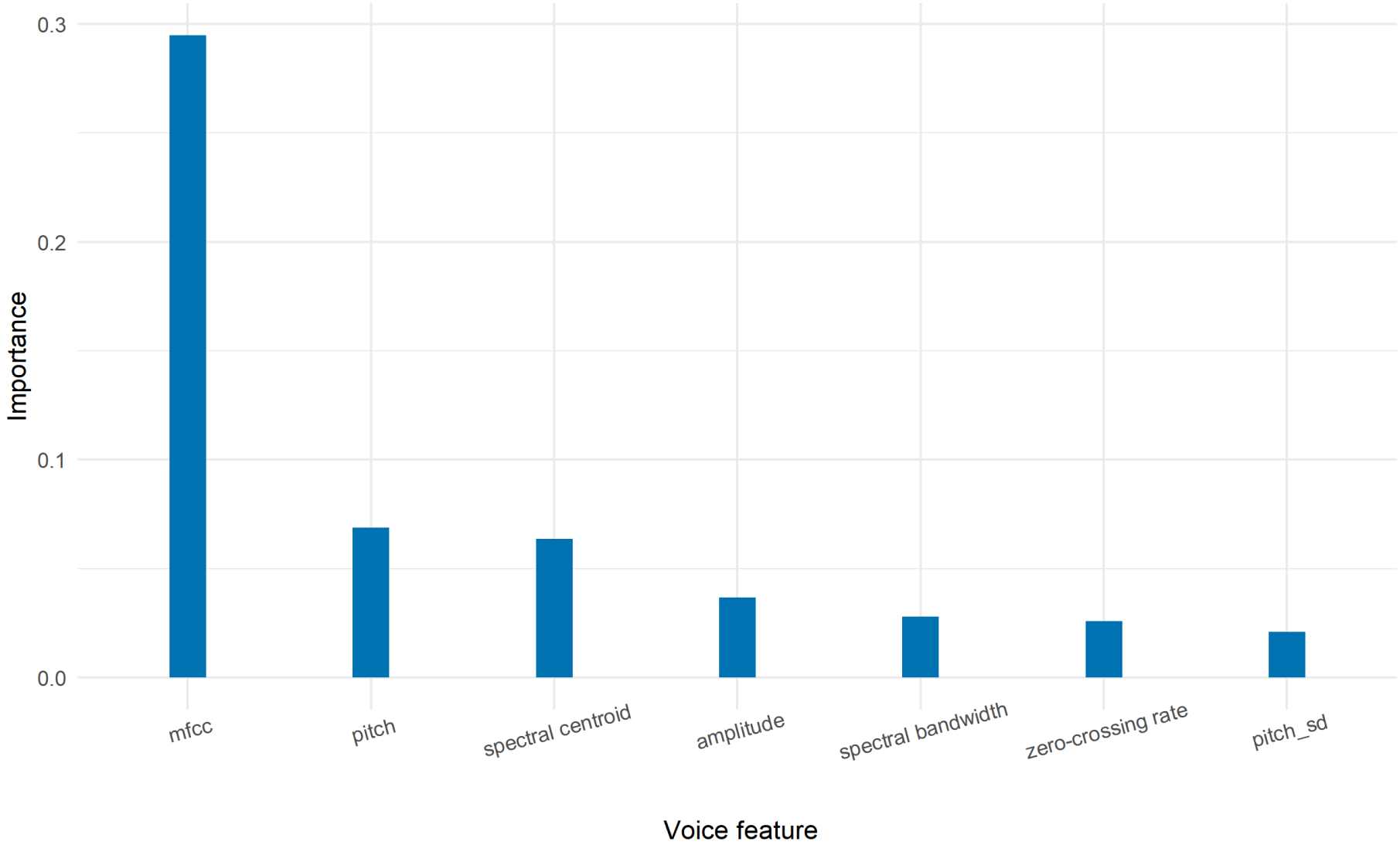
Results: Accuracy (out-of-sample)



Results: Accuracy (out-of-sample)



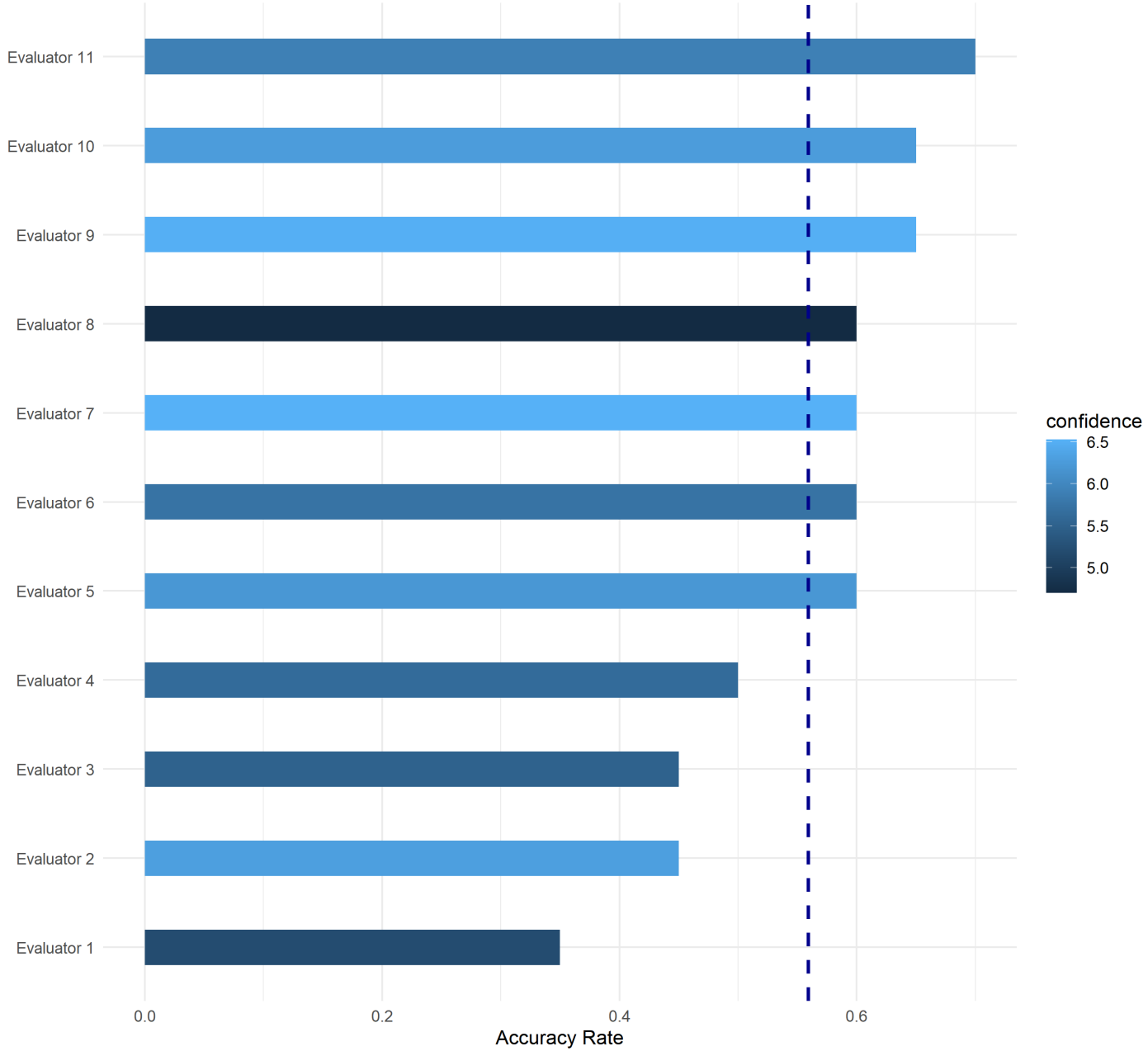
Which voice variables are important?



Human evaluators: setup

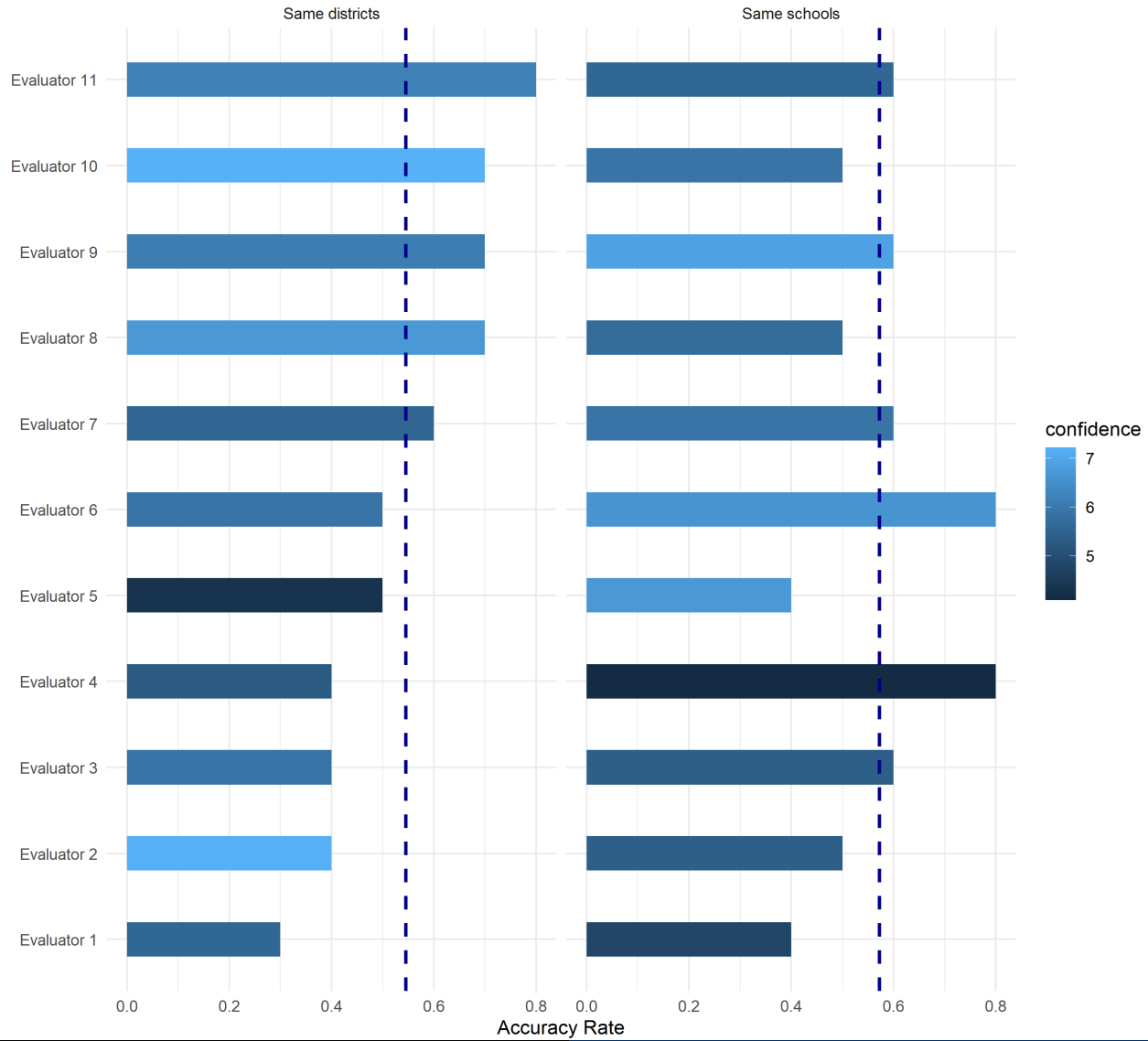
- 11 undergraduates students from various backgrounds
- Went through 20 randomized pairs of teachers
  - 1 pair = one highly effective vs. highly ineffective teacher
  - top 5% vs. bottom 5% of VA distribution
- Listen to audio only, 2×2 minutes per teacher
- Asked to rate technical (e.g. pitch, speed, rhythm) and psychological features (e.g. warm, competent, authoritarian)
- Guessed the most effective teacher Form example

Human evaluators: results (accuracy = 55%)

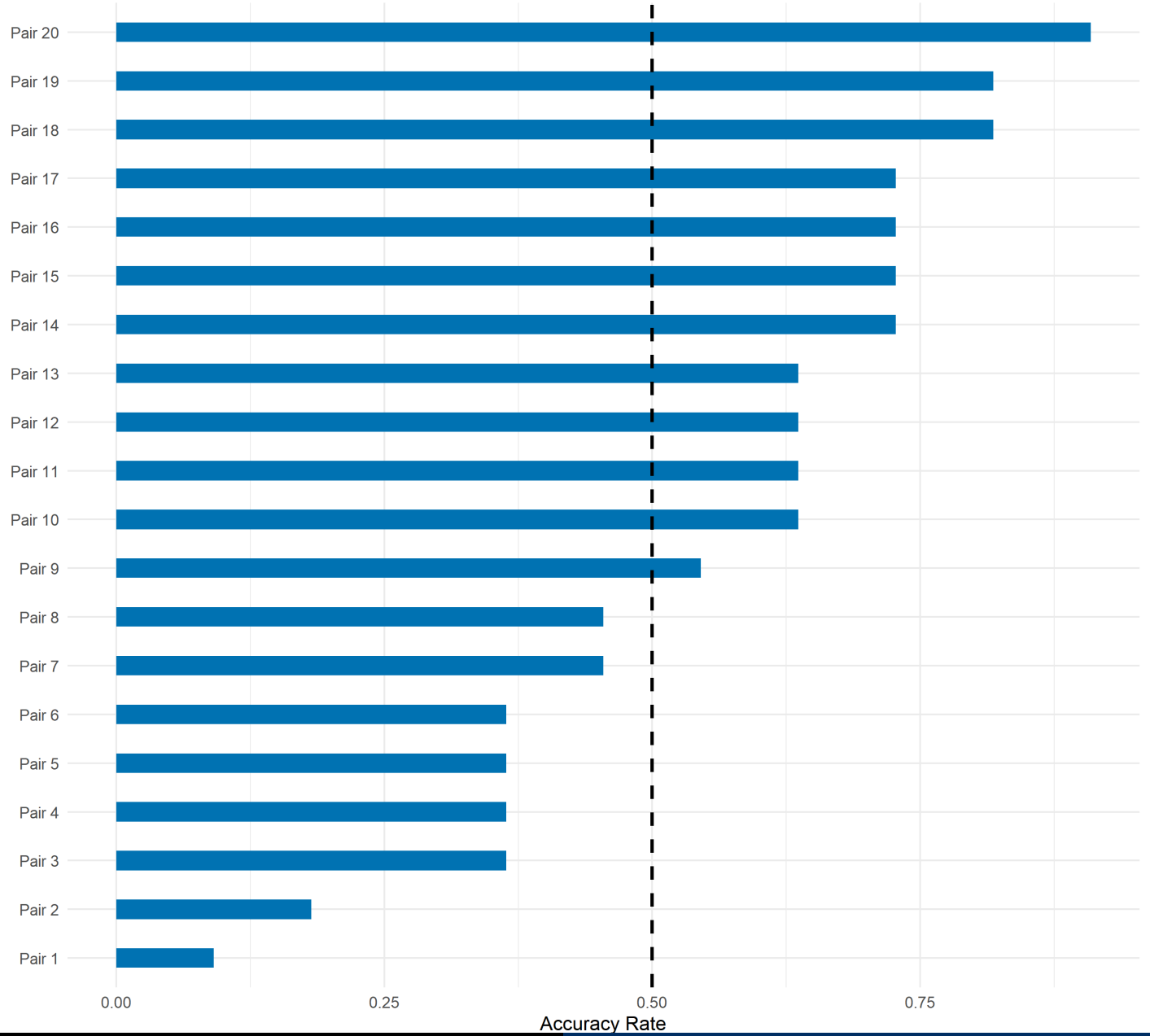


Human evaluators: results

- Accuracy increased slightly (57%) when comparing teachers from the same schools



Human evaluators: accuracy by teacher pair



Human evaluators: who did they choose?

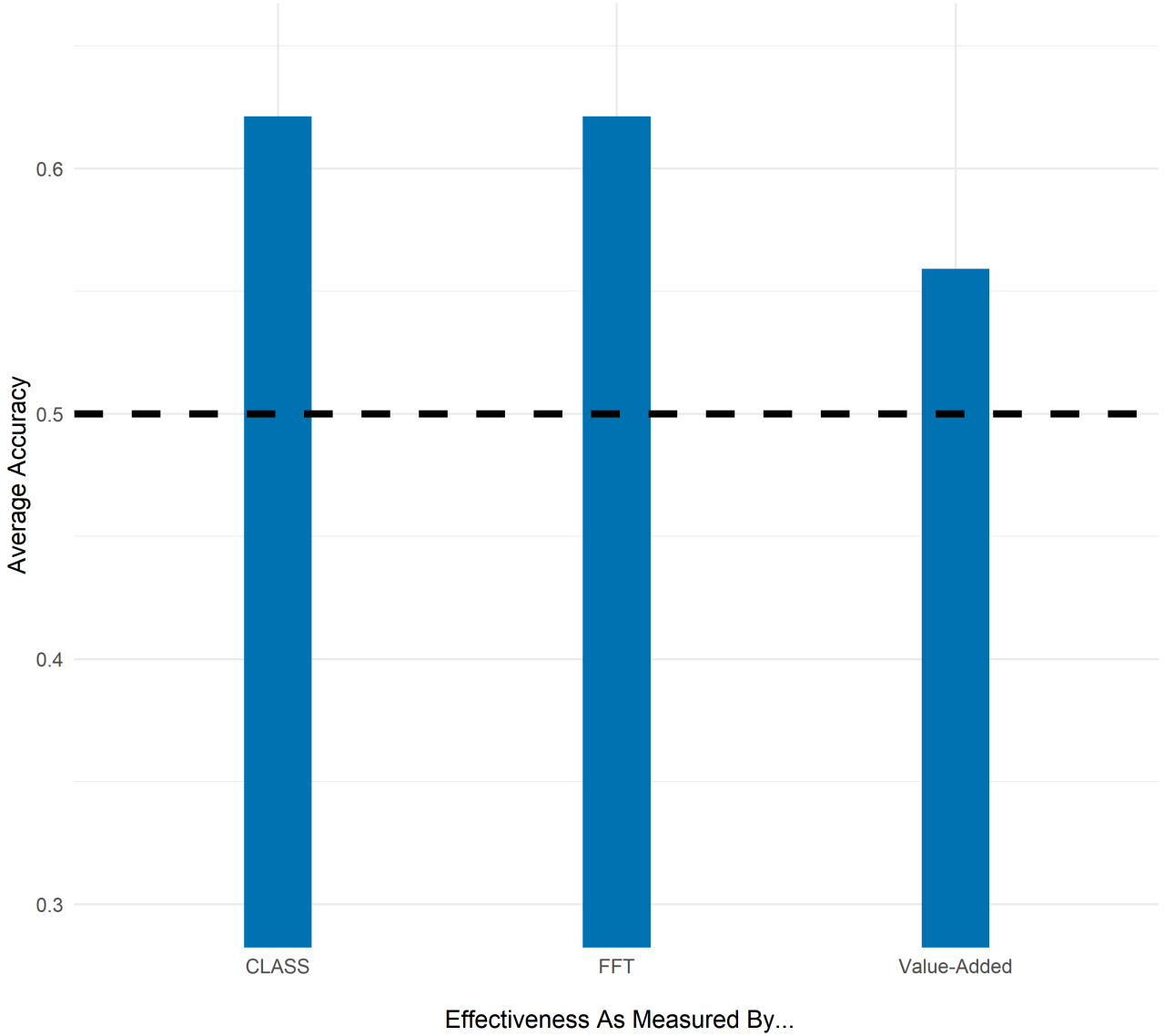
- According to evaluators: psychological features weigh more
  - Psychological features (e.g. warm, competent): 66%
  - Technical features (e.g. pitch, speed): 33%
  - $\rightarrow$ inferred psychological traits from voice Word cloud Regressions
- Conditional logit model: most important variables to explain selection Table
  - $(-)$ authoritarian
  - $(+)$ communicative, competent, stern, warm
  - technical features have low predictive power

Human evaluators: who did they choose?

- Effectiveness is multidimensional
  - Value-added
  - Classroom Assessment Scoring System (CLASS)
  - Framework for Teaching (FFT)
- CLASS: emotional support, classroom organization, instructional support
- FFT: planning and preparation, classroom environment, instruction, professional responsibilities

Human evaluators: other definitions of effectiveness

- Effectiveness as defined by CLASS or FFT instead



Takeaways

- ML is very well suited for exploring what best predicts value added (at least compared to humans)
- Focusing on VA extremes helps avoid classification and exaggerates differences
- CLASS scores modestly predicts extreme VA differences (60% accuracy)
- Audio information from 15 second recording of teacher speaking in class predicts as well as using info from CLASS score (60% accuracy)
- Audio correlated with teacher's class and behavior management, and student engagement
- combining audio and class boosts prediction power a little

Futures directions

- Use and predict the entire value-added distribution (e.g., top 10%, top 15%, ...)
- Exploit all the information in the videos
  - Voice + Text + Images
- Can human evaluators *learn* to predict effectiveness?

- Thank you!

Full list of extracted features

pitch	spectral bandwidth
amplitude	spectral flatness
tempo	rolloff
chroma	zero-crossing rate
root-mean sqr. energy	mel-frequency cepstrum (1-20)
spectral centroid	

- Most variables are extracted as vectors $\rightarrow$ compute the mean
- To capture the distribution $\rightarrow$ include the variance

◂ Back

Estimating Value-Added Measures

For a student i in group g on year t, estimate

$$\underbrace{y_{igt}}_{\text{test score}} = \overbrace{x_{it}}^{\text{characteristics}} \beta + \bar{x}_{gt}\gamma + \theta y_{i,t-1} + \lambda \bar{y}_{g,t-1} + \epsilon_{igt}.$$

x_{it}: ethnicity, age, gender, special education, free lunch

Let the students of teacher j in group g on year t be represented by the set $\mathcal{N}_{jgt}$.

The value-added of teacher j is

$$VA_j = \frac{1}{|\mathcal{N}_{jgt}|} \sum_{i \in \mathcal{N}_{jgt}} \hat{\epsilon}_{igt}.$$

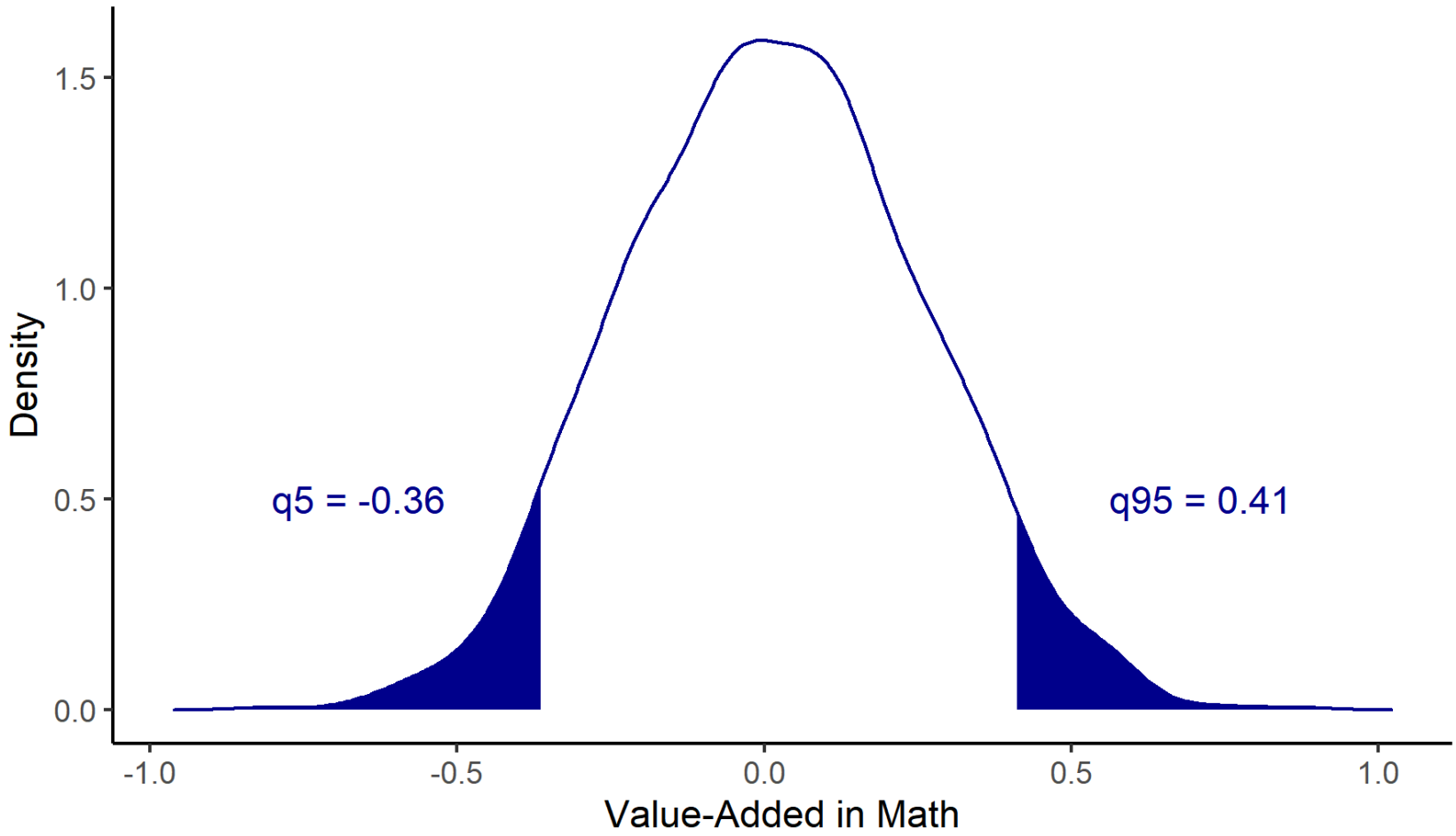
◂ Back

Details on diarization

- Research assistants *diarized* 130 videos
  - *Diarization*: who speaks when, at each second
- Procedure
  1. Supervised and unsupervised noise reducer (to reduce ambient noise)
  2. Trimming (e.g. when students are chatting at the beginning of the class)
  3. Extract the MFCC coefficients at each segment (1 sec)
  4. Perform k-means clustering as in Giannakopoulos et al. (2012)
- Benchmark with several different metrics
  - Most frequent speaker, number of speakers changes, total teacher time, etc.
- Accuracy in identifying the teacher: > 80%

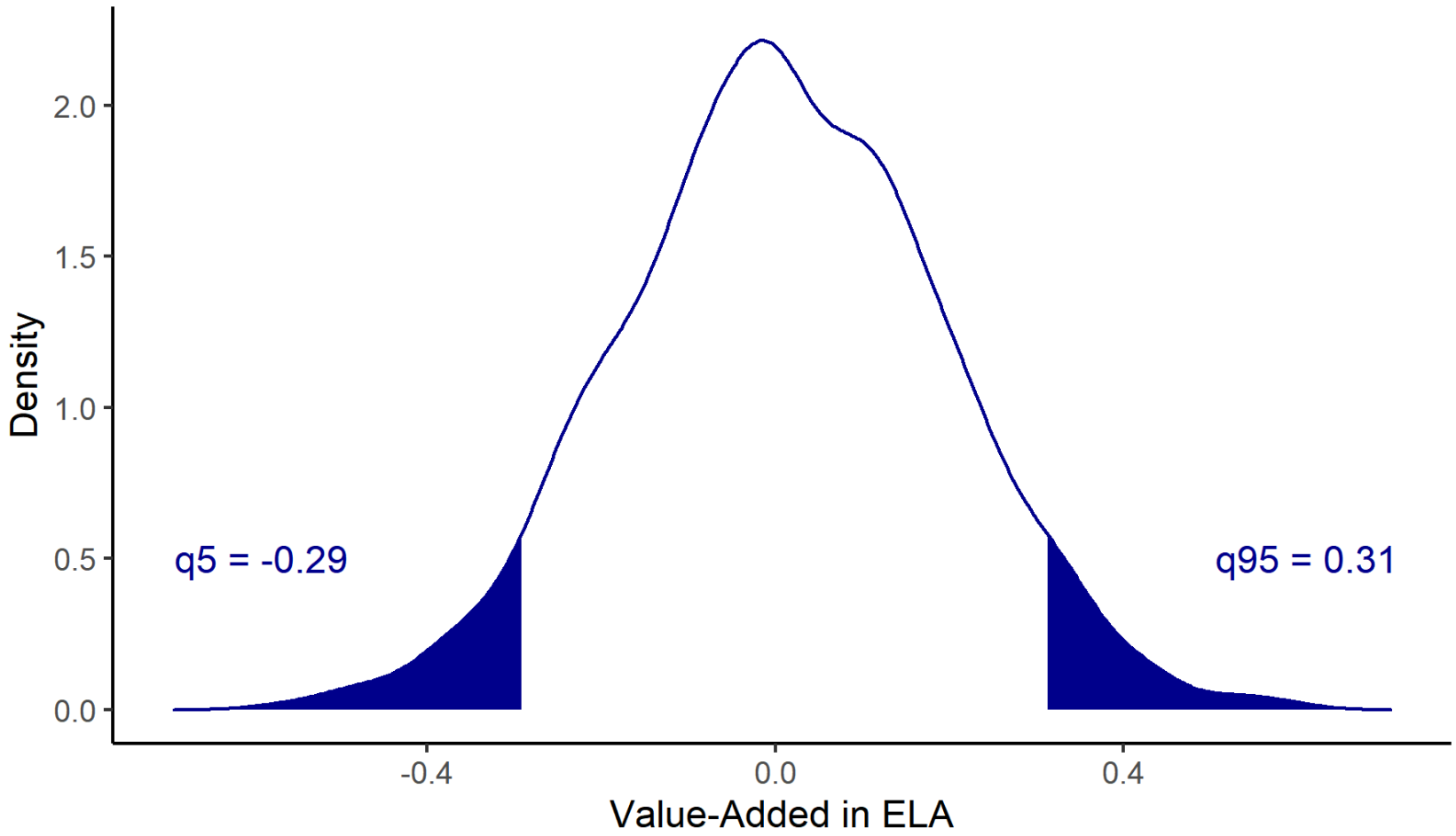
◂ Back

Value-Added Estimates: Mathematics



Back

Value-Added Estimates: ELA



Back

Word Cloud: selected teacher



◂ Back

Example of a form

Effective Teachers – Audio (form #131)

Enter your name:

Motivation

Researchers have explored several ways to evaluate teachers' effectiveness, one of which measures how teachers affect students on standardized tests (e.g. math and English). Given this definition, students with an effective teacher should, therefore, expect to receive a higher test score than if they were with an average teacher.

The Experiment

You will be asked to *listen* to audio samples of teachers in their classroom and answer various questions. More specifically, you will listen to two distinct teachers, one measured to be highly effective and the other measured to be highly ineffective. The order of these two teachers you are asked to listen to is random. Some questions from the present form can be answered while you are listening. *Question 1* (and its subquestions) can be answered while you are listening to the audio from the first teacher, whereas *Question 2* (and its subquestions) can be answered while you listen to the second teacher. The list of videos is available on this website once you log in with your email address: https://www.icpsr.umich.edu/web/pages/about/metldb.html.

This Session

In this session, you will be reviewing the teachers appearing in the following samples: **enk7** (Teacher A) and **dgg8** (Teacher B). For each teacher, please listen for 2 minutes (no more, no less) starting at 10:00 and for 2 minutes (no more, no less) starting at 30:00. That is, for each teacher, you will listen for a total of 4 minutes.

pair of teachers, randomized

◂ Back

1

timer

Questionnaire

Question 1: **Start the first audio.** You can answer the four following questions as the audio plays. You can use this stopwatch as a timer: 00:01:55 reset. After 2 minutes, remember to switch to the second segment.

a) The following properties correspond to ways that humans perceive sound. For each, assess Teacher A's voice.

- Pitch (1=very low, 7=very high)
- Volume (1=very quiet, 7=very loud)
- Brightness (1=very dark, 7=very bright)
- Speed (1=very slow, 7=very fast)
- Rhythm (1=very monotone, 7=very rhythmic)
- Continuity (1=long breaks and pauses, 7=no breaks and pauses)

b) The following characteristics have been documented to affect teachers' effectiveness and to be inferred through voices. On a scale from 1 (not at all) to 7 (very much), how would you rate Teacher A with respect to these characteristics? For each characteristic, please indicate how confident you are about your assessment. Select "0 (have no idea)" if the sample did not provide enough information.

Teacher A sounds...

- Aggressive
- Competent
- Confident
- Dominant
- Likeable
- Trustworthy
- Warm
- Authoritarian
- Stressed
- Stern (or strict/severe)
- Communicative
- Masculine (vs. feminine)
- Enthusiastic (or energetic/dynamic)
- Caring

characteristics and confidence levels

c) On a scale from 1 (very quiet) to 7 (very noisy), how noisy was the classroom? Noise refers to anything not teaching-related (students interrupting, chatting or whispering, for instance).

d) In your own words, please describe what particular features or elements from the audio seem important for your assessment about Teacher A's effectiveness.

Answer:

2

Conditional logit

Table: Choice model

	(1)	(2)	(3)
brightness	0.375**		-0.091
continuity	0.195		0.105
pitch	-0.074		0.162
rhythm	0.248*		-0.037
speed	-0.248		-0.367**
volume	0.207		0.185*
aggressive		-0.012	-0.037
authoritarian		-0.175**	-0.231*
caring		0.010	0.039
communicative		0.316**	0.330**
competent		0.769***	0.736***
confident		-0.054	-0.020
dominant		0.146	0.158
enthusiastic		0.309	0.335
likeable		0.077	0.172
masculine		0.180*	0.255**
stern		0.376**	0.352*
stressed		-0.162	-0.119
trustworthy		-0.163	-0.201
warm		0.482**	0.475**
Observations	422	422	422

Standard errors clustered at the evaluator level

◂ Back

Technical and psychological features

Characteristic	aggressive	authoritarian	caring	communicative	competent	confident	dominant
brightness	-0.548***	-0.278	0.579***	0.284***	0.150**	0.132*	-0.289**
	(0.118)	(0.165)	(0.065)	(0.074)	(0.057)	(0.069)	(0.123)
continuity	-0.019	0.074	-0.019	0.043	0.106**	0.146**	0.078
	(0.110)	(0.144)	(0.051)	(0.042)	(0.045)	(0.051)	(0.122)
pitch	0.079	0.005	-0.083	-0.011	-0.047	-0.080	-0.100
	(0.105)	(0.087)	(0.070)	(0.056)	(0.055)	(0.071)	(0.098)
rhythm	-0.071	-0.103	0.201***	0.283***	0.334***	0.335***	0.058
	(0.089)	(0.090)	(0.051)	(0.080)	(0.053)	(0.049)	(0.084)
speed	0.132	-0.020	-0.009	0.062	0.022	0.021	0.040
	(0.109)	(0.078)	(0.086)	(0.100)	(0.073)	(0.101)	(0.107)
volume	0.512***	0.514***	-0.208**	0.108	0.059	0.269***	0.494***
	(0.073)	(0.068)	(0.068)	(0.082)	(0.059)	(0.070)	(0.082)
Characteristic	enthusiastic	likeable	masculine	stern	stressed	trustworthy	warm
brightness	0.486***	0.680***	-0.070	-0.497***	-0.218**	0.391***	0.720***
	(0.068)	(0.117)	(0.110)	(0.091)	(0.083)	(0.102)	(0.108)
continuity	-0.038	-0.024	-0.073	0.050	-0.161	0.062	0.062
	(0.059)	(0.050)	(0.101)	(0.148)	(0.195)	(0.055)	(0.057)
pitch	-0.068	-0.110	-0.454***	-0.040	0.046	-0.008	0.023
	(0.079)	(0.066)	(0.125)	(0.088)	(0.046)	(0.079)	(0.045)
rhythm	0.261***	0.140	-0.022	0.112	-0.105	0.239*	0.154*
	(0.061)	(0.090)	(0.077)	(0.100)	(0.087)	(0.112)	(0.073)
speed	0.244**	0.046	0.310***	0.062	0.263*	-0.032	-0.006
	(0.106)	(0.091)	(0.094)	(0.107)	(0.139)	(0.116)	(0.099)
volume	0.170**	-0.181**	0.311***	0.500***	0.200	-0.082	-0.356***
	(0.074)	(0.064)	(0.088)	(0.083)	(0.110)	(0.072)	(0.058)
Observations	430	430	430	430	430	430	430

Standard errors clustered at the evaluator level

◂ Back