Evaluating Student Writing
A preliminary investigation

Courtney Napoles
Johns Hopkins University

Chris Callison-Burch
University of Pennsylvania

Outline

1. FWC corpus

    - A new corpus of student writing

2. Automatic scoring

    - A topic-independent model for this type of writing

    - Present a model that can handle grading differences between teachers

3. Conclusion and future directions

Motivation

1. Provide feedback to help teachers evaluate students
    - Can automatic writing evaluation be used on classroom writing assignments?
2. Provide feedback to help teachers grade better
    - Can we overcome different grading tendencies between teachers?

Freshman Writing Corpus

Freshman Writing Corpus

1. Take-home essays

Freshman Writing Corpus

1. Take-home essays
2. Long(er)-form

	Kaggle	FWC
# essays	22k	25k
avg. # tokens	250	900
avg. # grafs	1.5	5.5

Freshman Writing Corpus

1. Take-home essays

2. Long(er)-form

3. Open-ended topic

Freshman Writing Corpus

1. Take-home essays

2. Long(er)-form

3. Open-ended topic

4. Aligned drafts

Freshman Writing Corpus

1. Take-home essays
2. Long(er)-form
3. Open-ended topic
4. Aligned drafts
5. Detailed rubric scores

Freshman Writing Corpus

1. Take-home essays
2. Long(er)-form
3. Open-ended topic
4. Aligned drafts
5. Detailed rubric scores
6. Teacher comments

Freshman Writing Corpus

Freshman Writing Corpus

Freshman Writing Corpus (FWC)

- English Composition I
- 4 writing assignments (“projects”)
- Students submitted Intermediate and Final drafts for each assignment
- Each draft graded

Syllabus

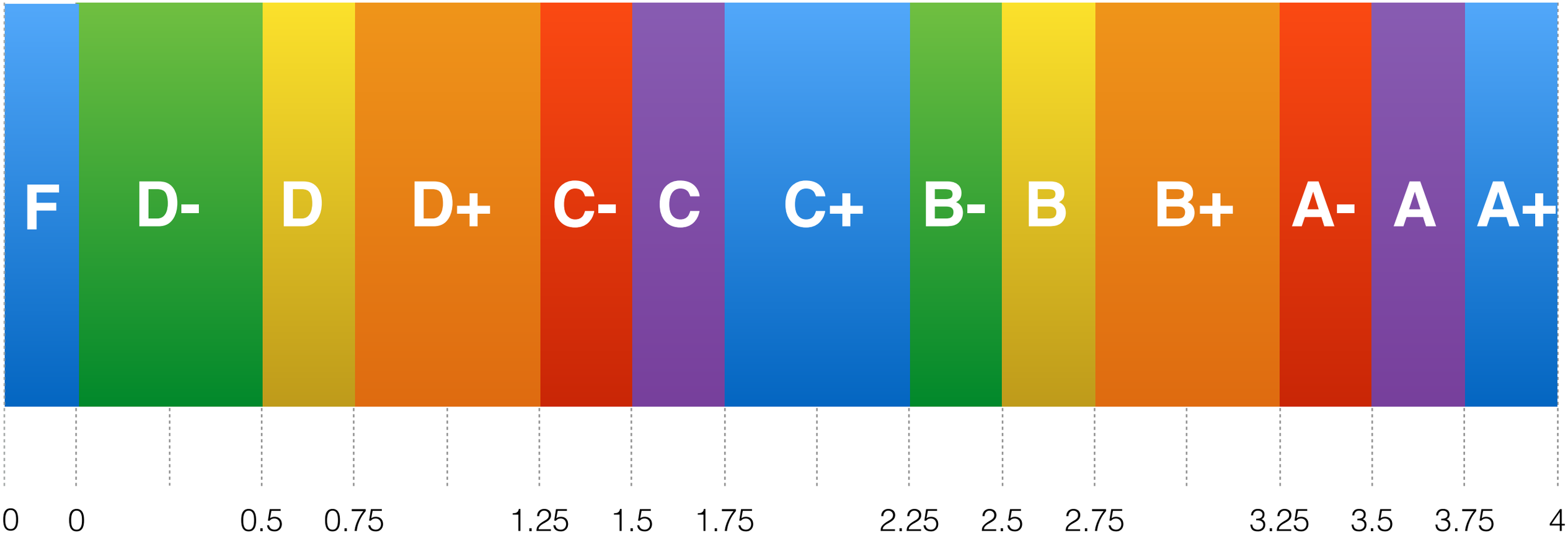
Project	Target # words	Brief description
1	600-770	A personal narrative that describes an experience and uses that experience to tell readers something important about the writer.
2	600	A bibliographic essay that asks you to understand the conversation surrounding your chosen topic by examining four relevant sources. …
3	600-800	A reflection that asks you to think carefully about how audience and purpose, as well as medium and genre, affect your choices as composers and reflect carefully on a new dimension of your topic.
4	1000-1200	A polished essay that asserts an arguable thesis that is supported by research and sound reasoning.

Rubric

Category	Weight	Level	Points	Brief Description
Focus	25%	Basics	0-4	Meeting assignment requirements
		Critical thinking	0-4	Meeting assignment requirements Strength of thesis and analysis
Evidence	25%	Basics	0-4	Quality of sources and how they are presented
Organization	25%	Basics	0-4	Introduction, supporting sentences, transitions, and conclusion
		Critical thinking	0-4	Progression and cohesion of argument
Style	25%	Basics	0-4	Grammar, punctuation, and consistent point of view
		Critical thinking	0-4	Syntax, word choice, and vocabulary
Format	5%	Basics	0-4	Paper formatting and conformance with style guide

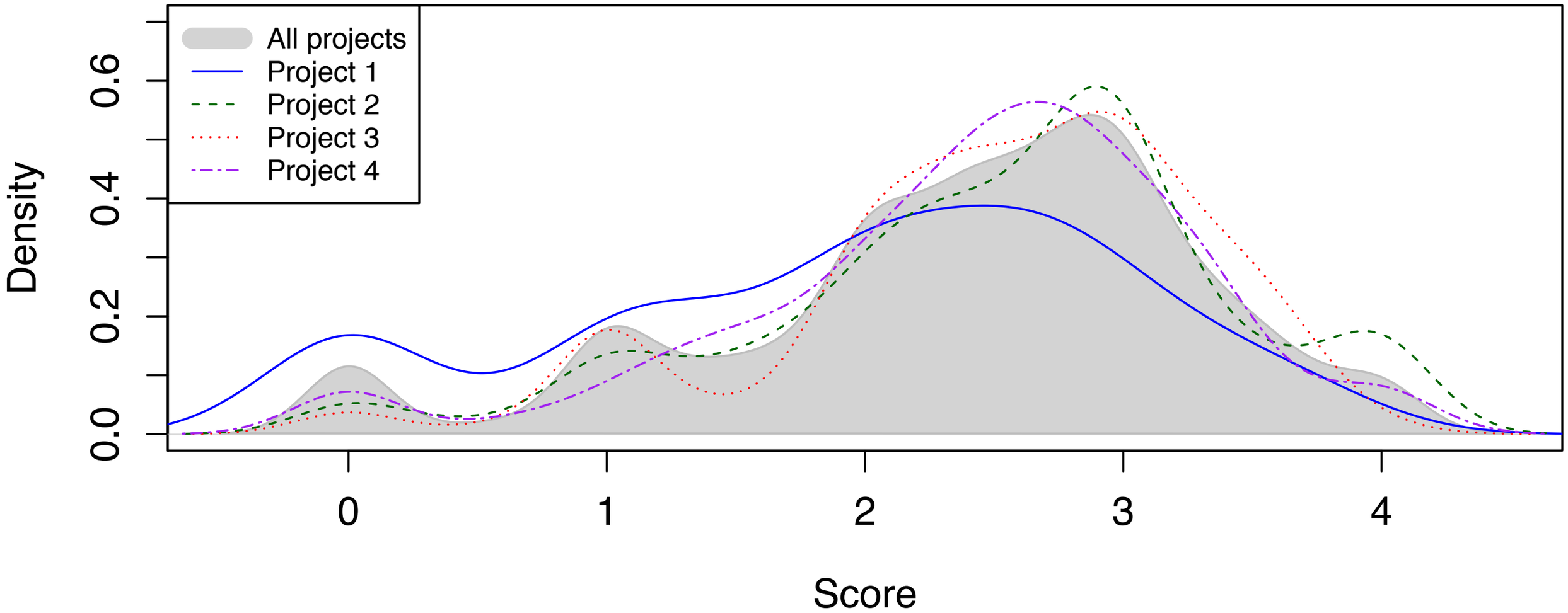
Rubric

Weighted average of rubric scores corresponds to letter grade



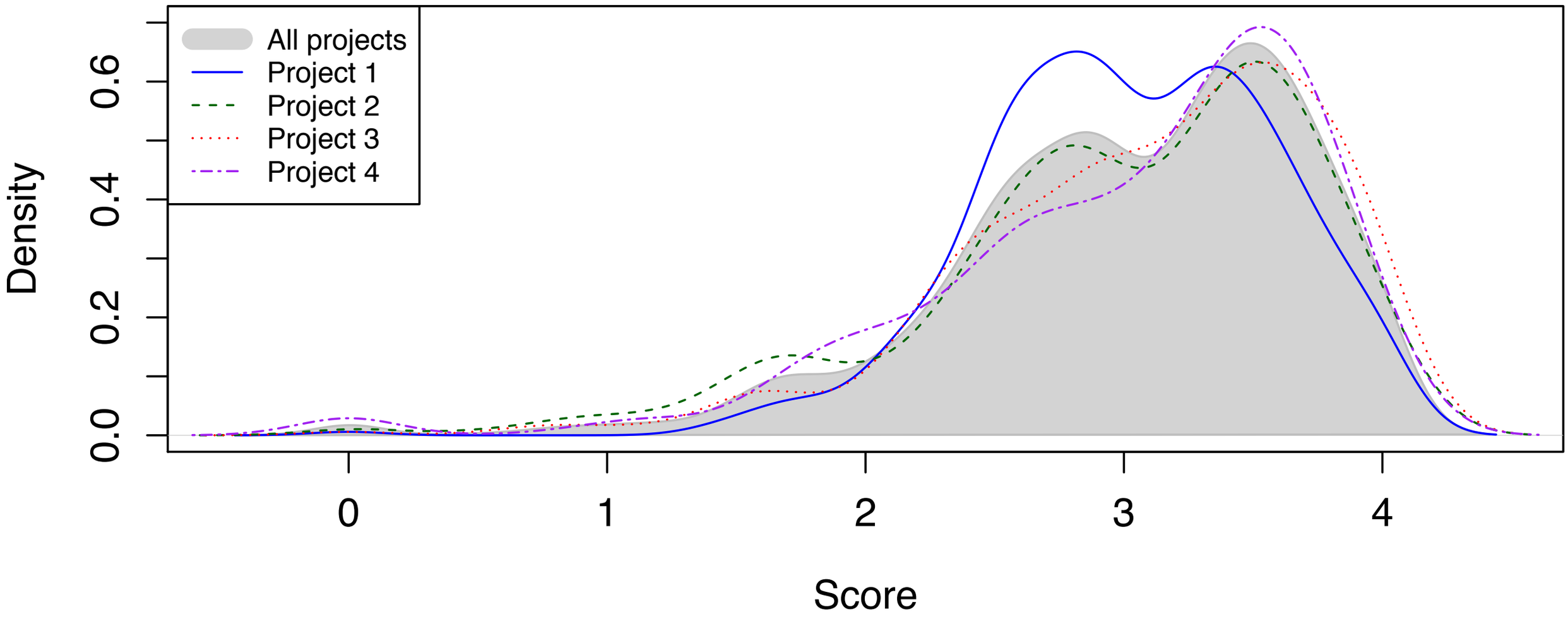
Scores

Intermediate drafts (M=2.4, SD=0.9)



Scores

Final drafts (M=3.0, SD=0.7)



Teacher Feedback

General comments

Your introduction offers some general introduction to the topic. You introduce one of the sources completely. Your introduction requires a stronger thesis statement that draws the connections between the two.

Although I appreciate that you changed one of the sources, there still remains not much substance to summarize. Both sources are very brief, and the arguments are not complex.

...

Teacher Feedback

Inline

Each start of the new school year, headlines bear the names of a handful of young,

seemingly healthy athletes who die suddenly on the basketball court, the football field or the

no comma

track. Most of the time the reason why, is unknown. This happens so often and yet no one ever

sees it coming. Athletes train and work out for years with no problems, until one day they

collapse and die. One minute Reggie Garrett was making a touchdown, and the next minute he

collapsed and died, according to a NBC news report. About ten to twenty-five sports related citation?

sudden cardiac deaths in young athletes occur annually in the United States. Robin J Northcote

quotes for articles

who wrote the article, Sudden Cardiac Death in Sport, believes that these athletes had to have

alone

had a previous medical problem and that exercise along would not cause them to die. Dr. what's the name of the article?

Milton Greenwich wrote an article on young athletes as well and also says that exercise alone

what is your thesis? The connection between the two?

would not cause death.

By the numbers

- Full corpus: 2 years of Composition I and II
  - Fall, Spring, Summer
  - > 25k essays
- This study: 1 semester of Composition I
  - 3,362 essays
  - 639 students, 55 sections, 21 instructors

By the numbers

Draft	Count	Tokens	Sentences	Paragraphs
Inter.	1200	840.3	35.6	5.2
Final	1762	938.5	39.6	5.7

Automatic Scoring

Introduction

Previous work: Test writing

- Short answers or short essays
- In response to a prompt or passage
- Timed
- Limit on outside sources

Introduction

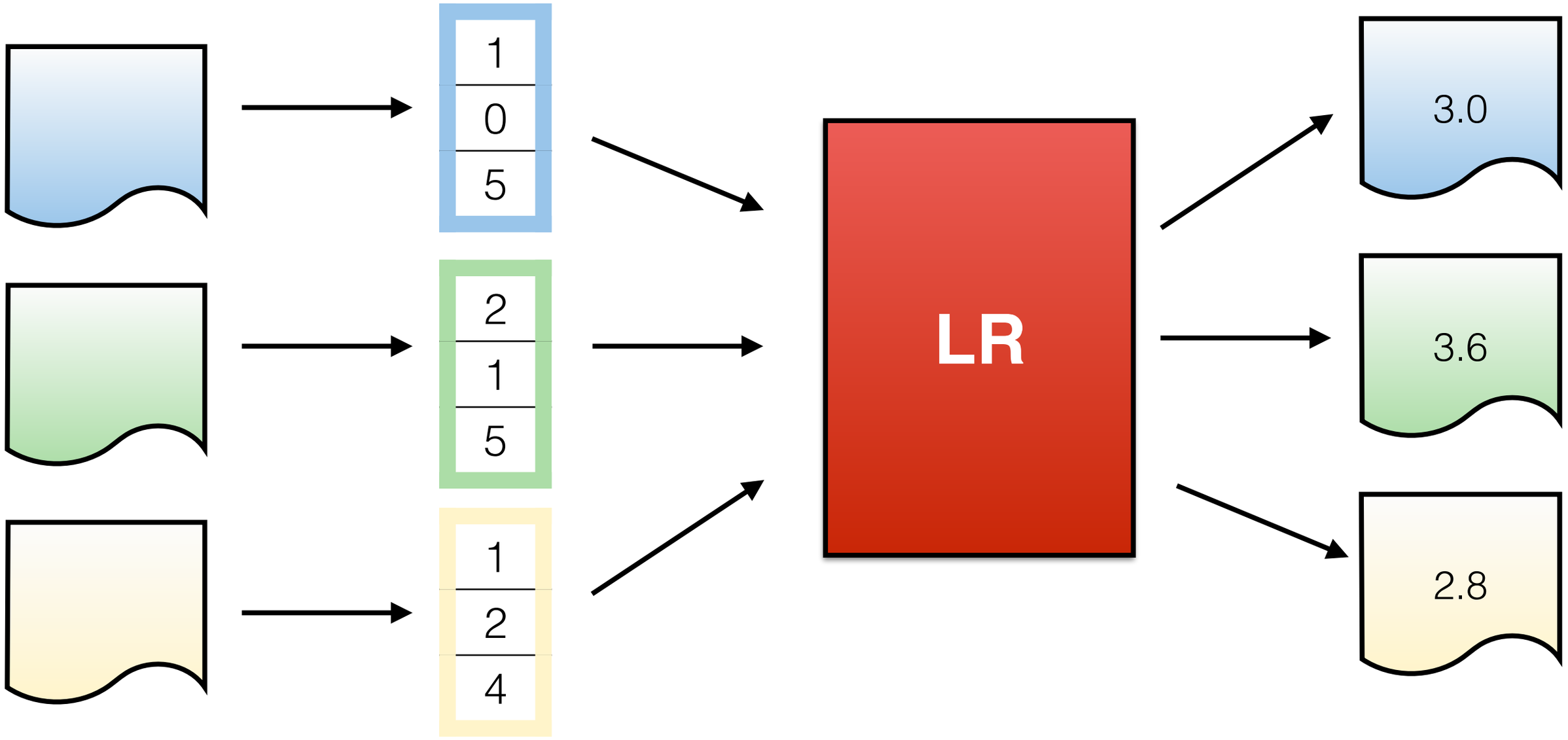
This work: Classroom writing

- Take-home assignments
- Open-ended topics
- Longer
- More polished (?)
- Different scoring criteria (?)

Experimental Setup

- Linear regression to predict score 0-4
  - Round to nearest 0.05
- Different models for Intermediate and Final drafts
- Data
  - Training: 1,200 Intermediate, 1762 Final essays
  - Testing: 100 Intermediate, 100 Final essays

Experimental Setup

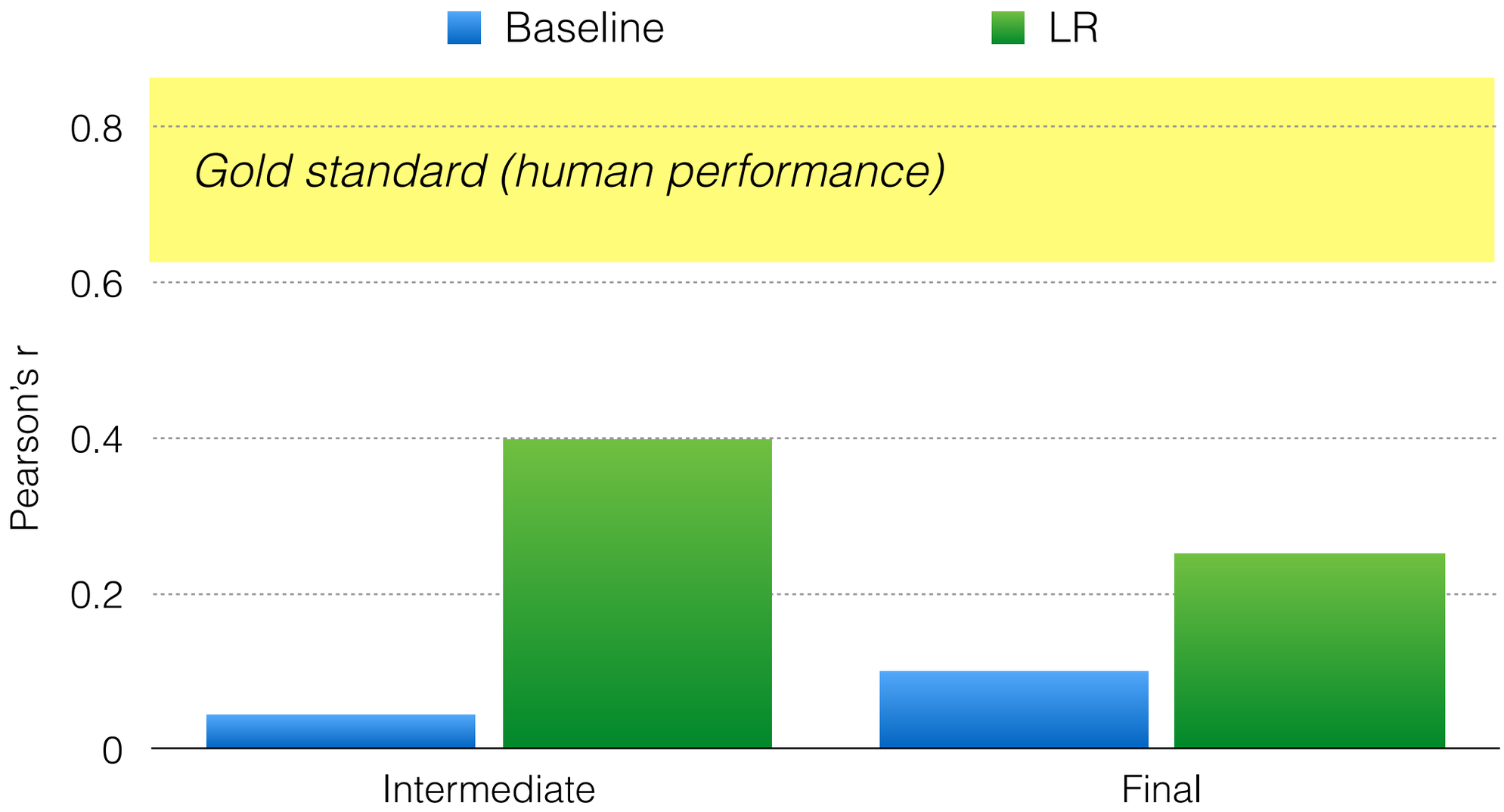


Features

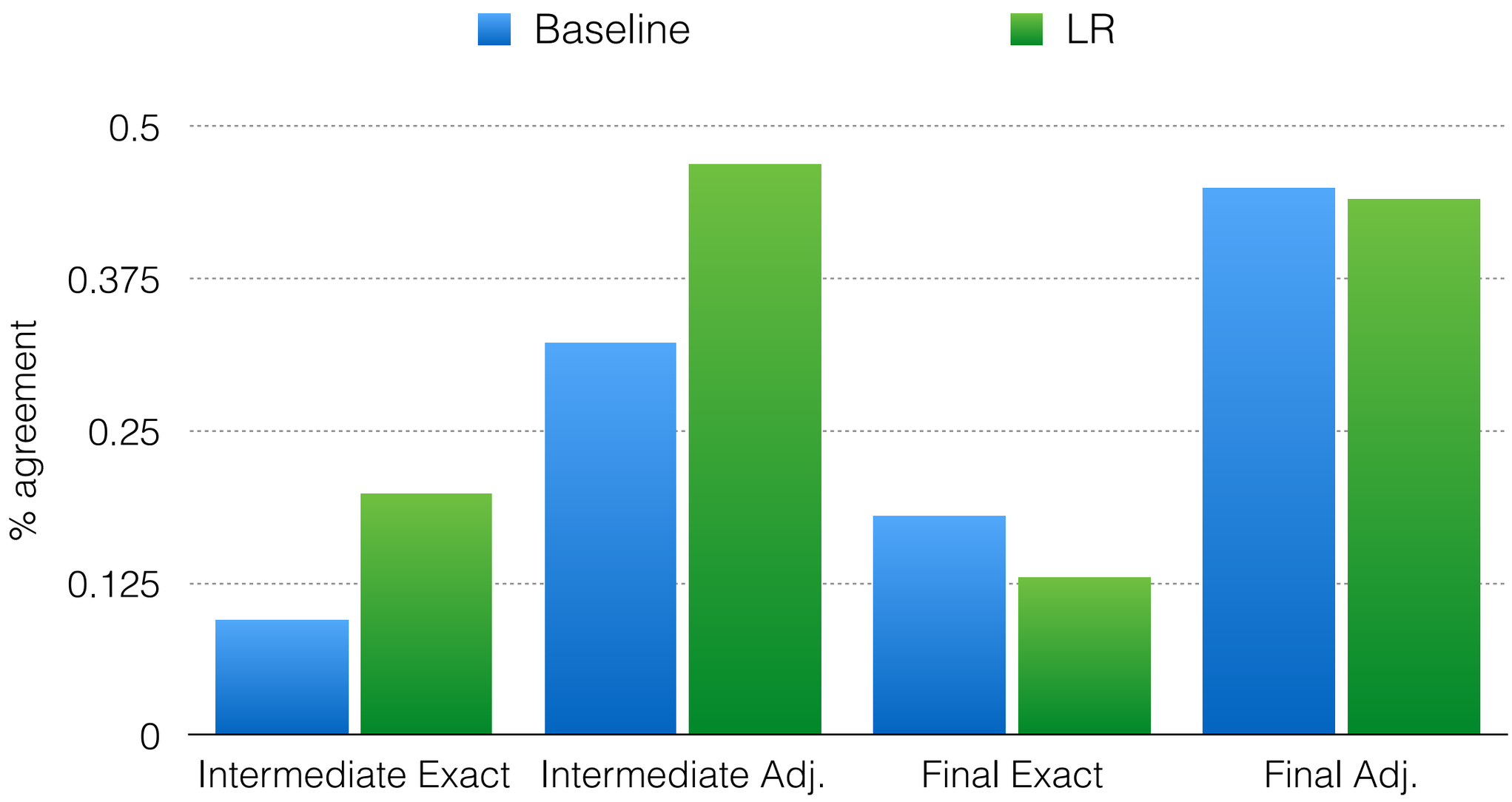
- Categories:
  - surface
  - structural
  - lexical
  - syntactic
  - grammatical
- 57 features + n-gram features

Results

Results


Baseline
LR
0.8
Gold standard (human performance)
0.6
Pearson's r
0.4
0.2
0
Intermediate
Final

Results


Baseline
LR
0.5
0.375
0.25
0.125
0
% agreement
Intermediate Exact
Intermediate Adj.
Final Exact
Final Adj.

Challenges

639 students, 55 sections, 21 instructors

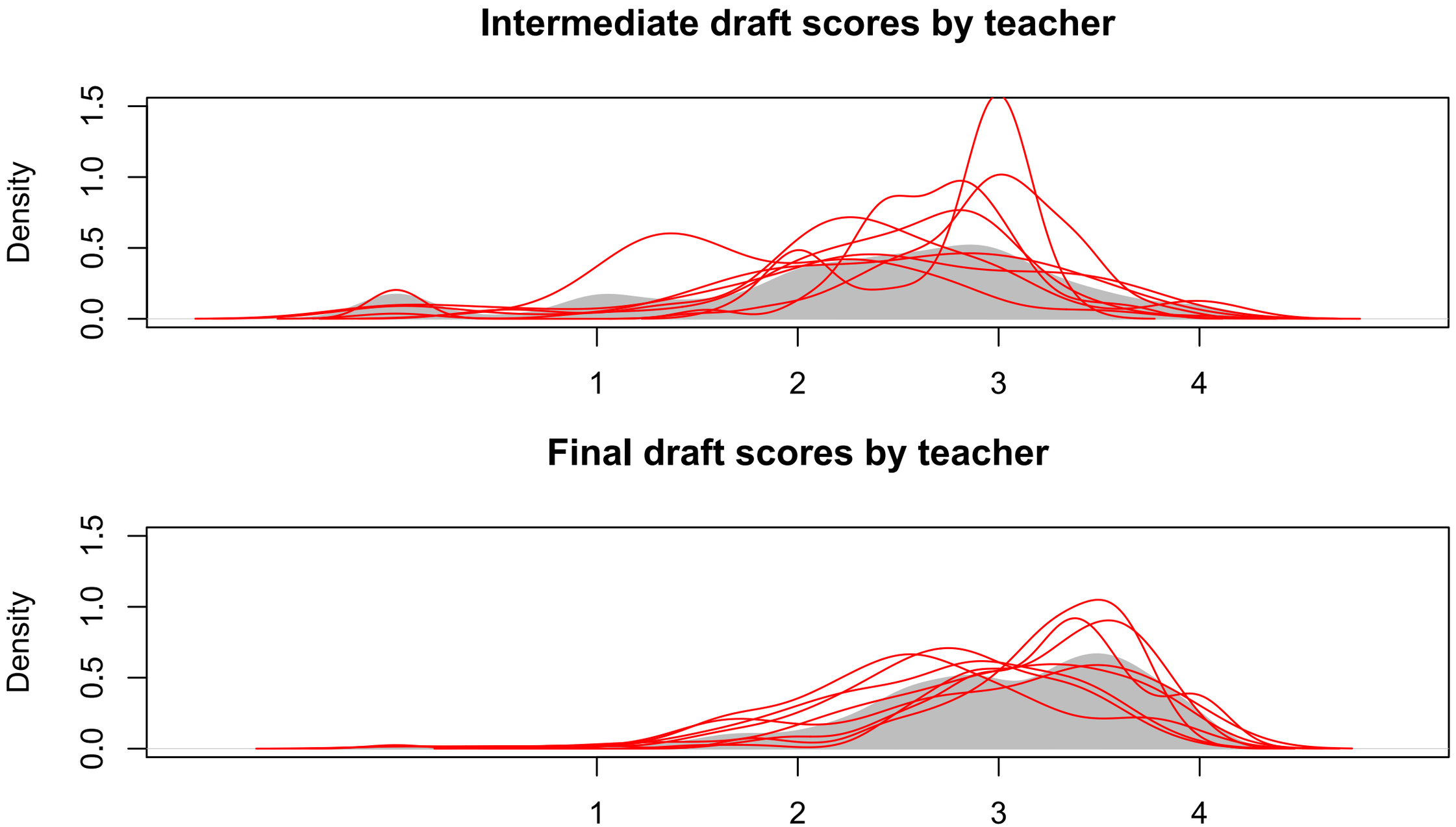
Instructors

- Standardized tests graded by multiple instructors
- Validated scores = trustworthy scores
- FWC scores are NOT validated
- Even when graders are trained and score on the same rubric, they *may* be inconsistent

Instructors

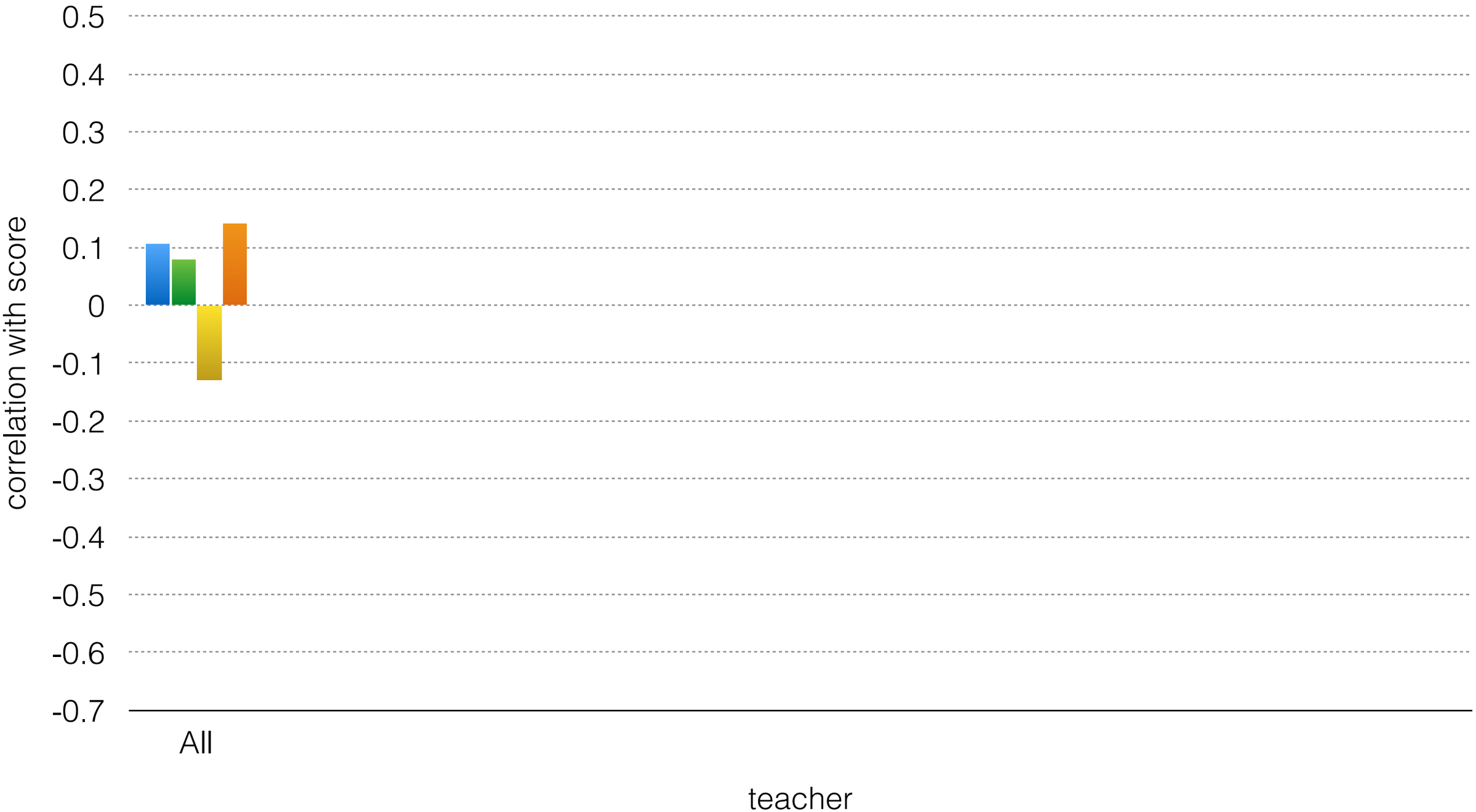
Do teachers grade differently?

Instructors

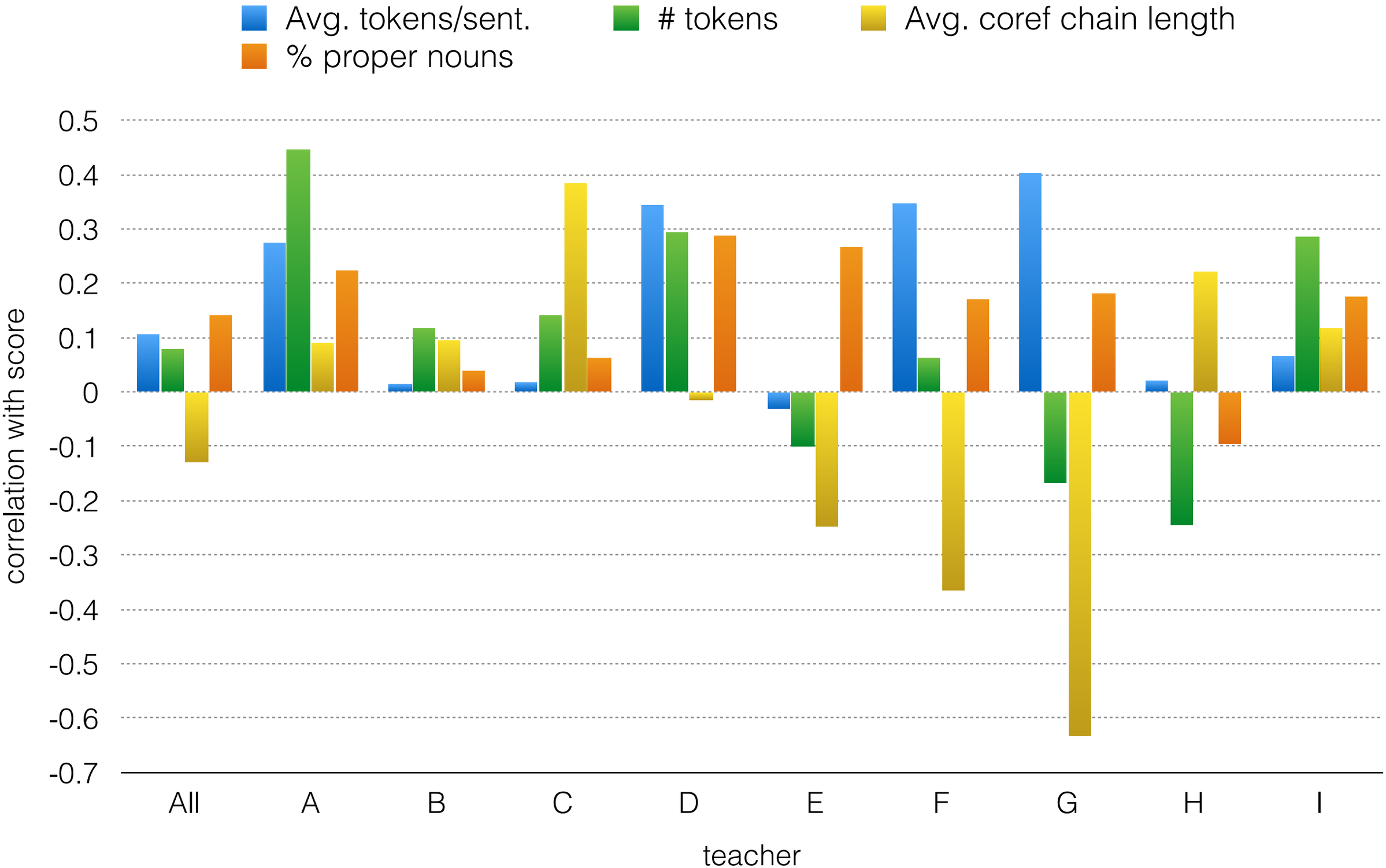


Instructors

Avg. tokens/sent.
tokens
Avg. coref chain length
% proper nouns



Instructors



Instructors

Do teachers grade differently?

…maybe

Single-task vs. Multi-task Learning

Single-task Learning

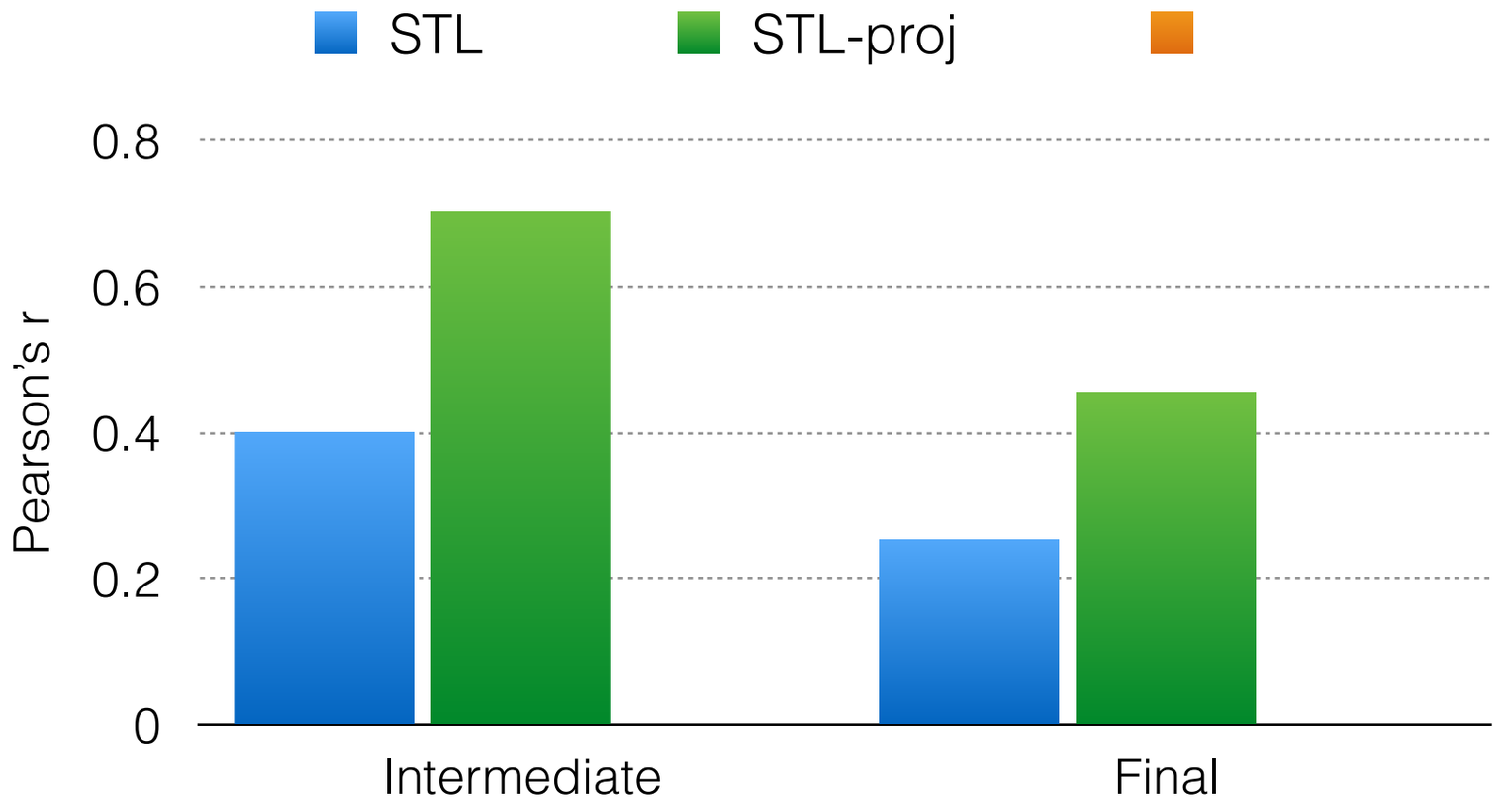
- Original model was single-task
  - learns one task at a time
  - scoring all essays

Single-task Learning

- Other single-task variations:
  - Model each project separately

Single-task Learning

- Other single-task variations:
  - Model each project separately

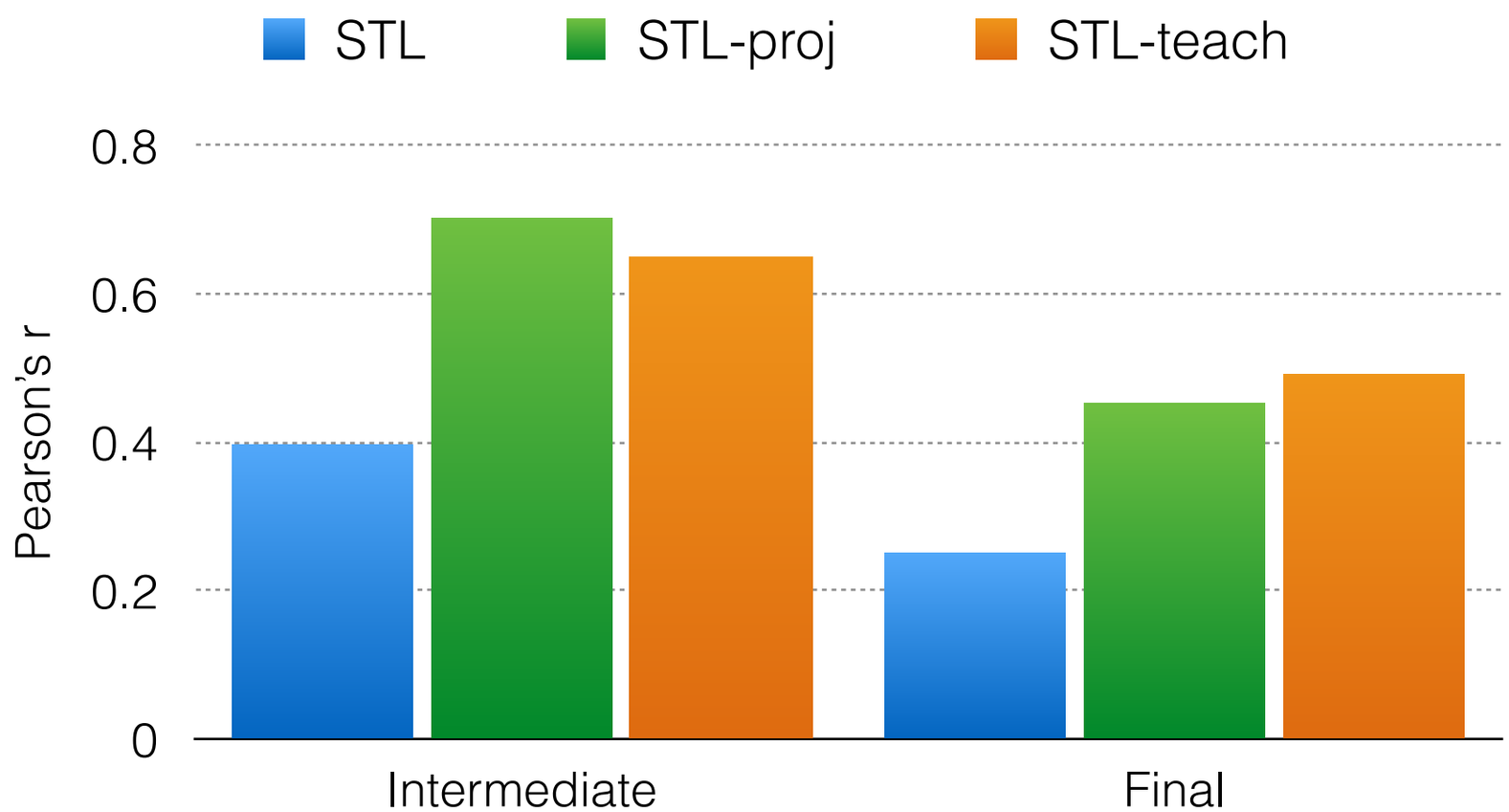


Single-task Learning

- Other single-task variations:
  - Model each project separately
  - Model each teacher separately

Single-task Learning

- Other single-task variations:
  - Model each project separately
  - Model each teacher separately



Multi-task Learning

- Multi-task learning
  - learns many problems at the same time
    - how each teacher scores
- jointly models the scores given by each teacher
- takes advantage of shared knowledge

Multi-task Learning

How?

Multi-task Learning

How?

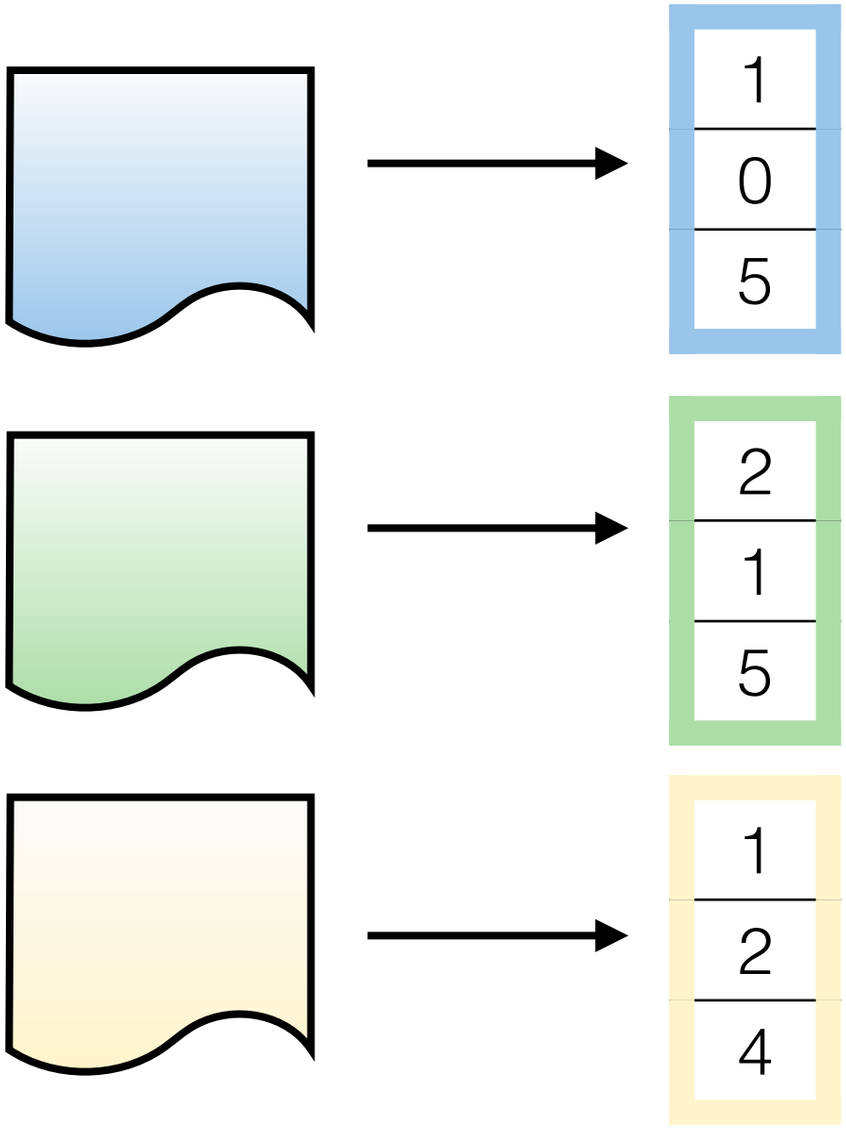
- Enlarge the feature space
  - Extracted *m* features for each essay
  - Add teacher-specific features
    - Each feature has a global copy and a teacher-specific copy
  - Now, *m* * (1 + # teachers) features

Multi-task Learning

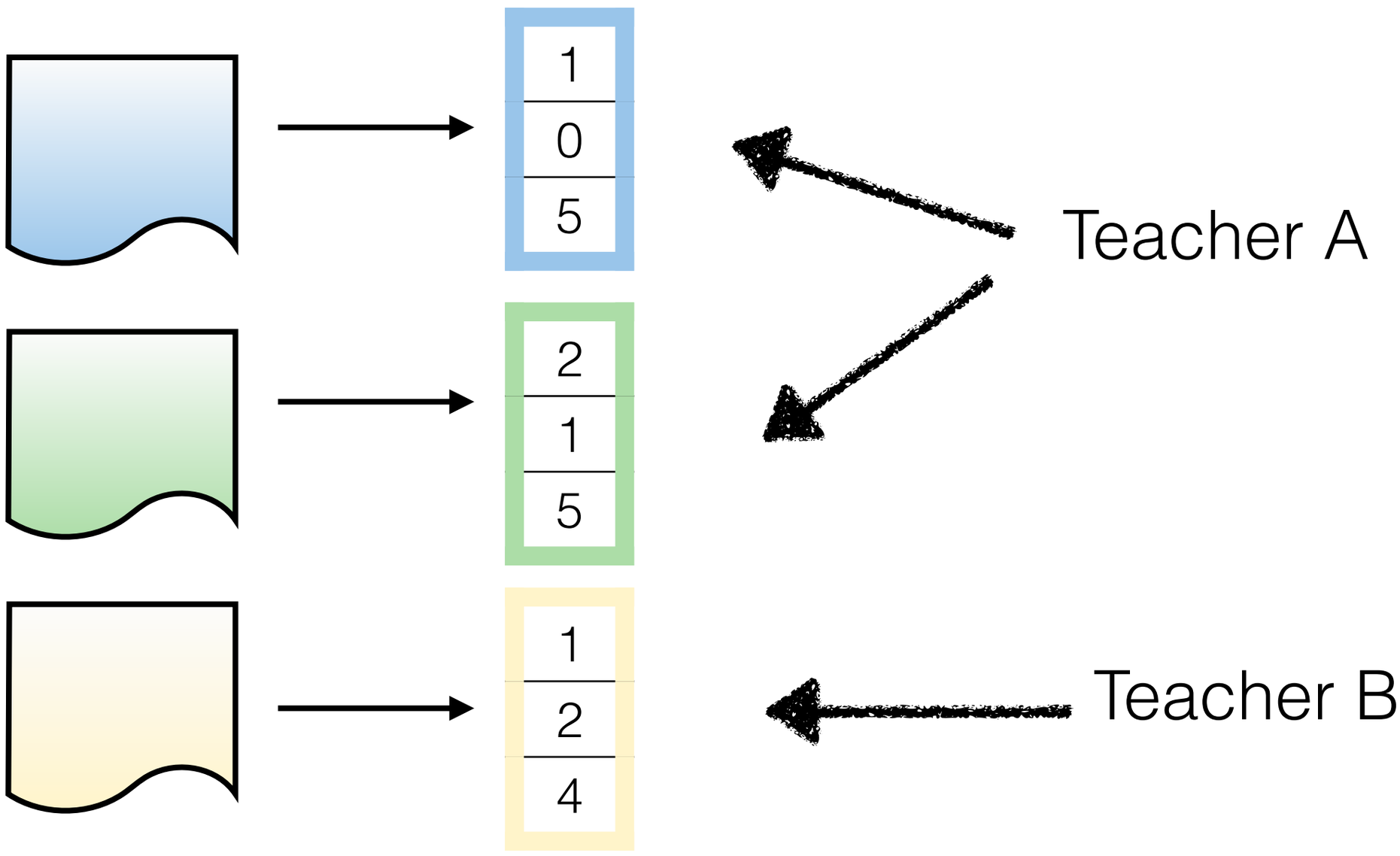
How?

- each feature has a global feature and a teacher-specific feature for each teacher
- replicate feature values for the teacher-specific features if that teacher graded the essay (0 otherwise)
- STL: *m* features
- MTL: *m* * (1 + # teachers) features
- dimensionality reduction with PCA
- linear regression

Multi-task Learning



Multi-task Learning



Multi-task Learning

global {

1
0
5

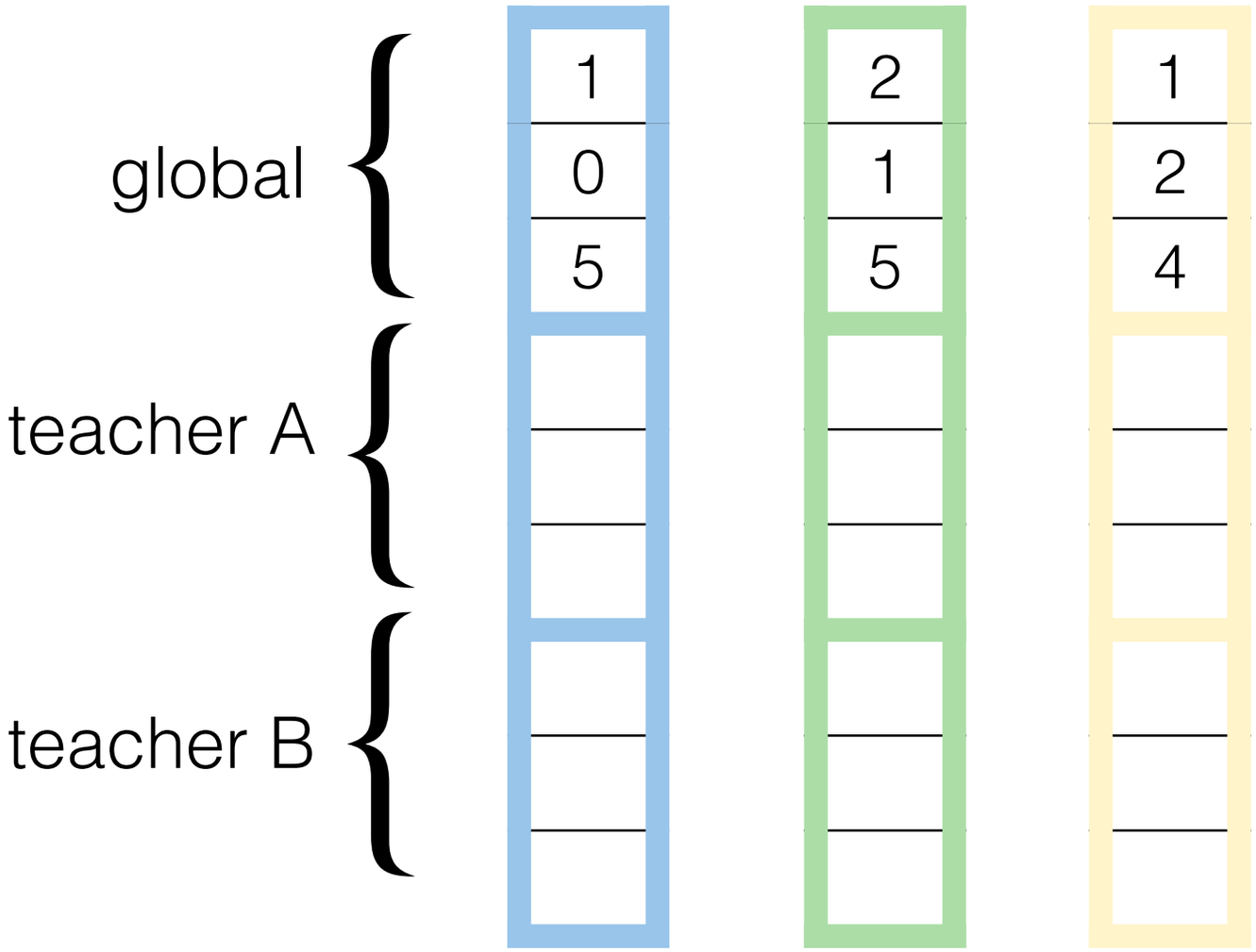
2
1
5

1
2
4

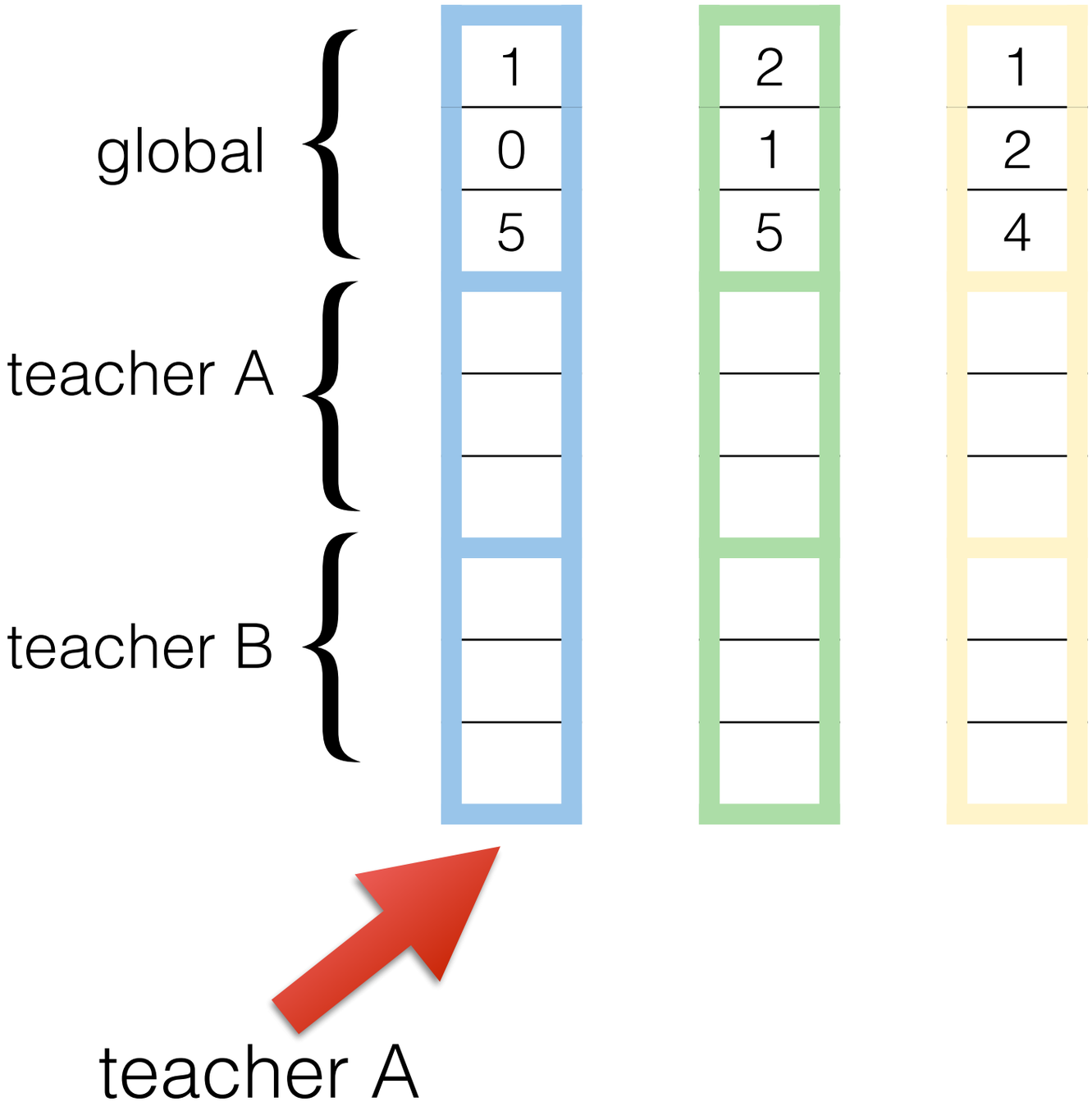
Multi-task Learning

global	1	2	1
	0	1	2
	5	5	4
teacher A			

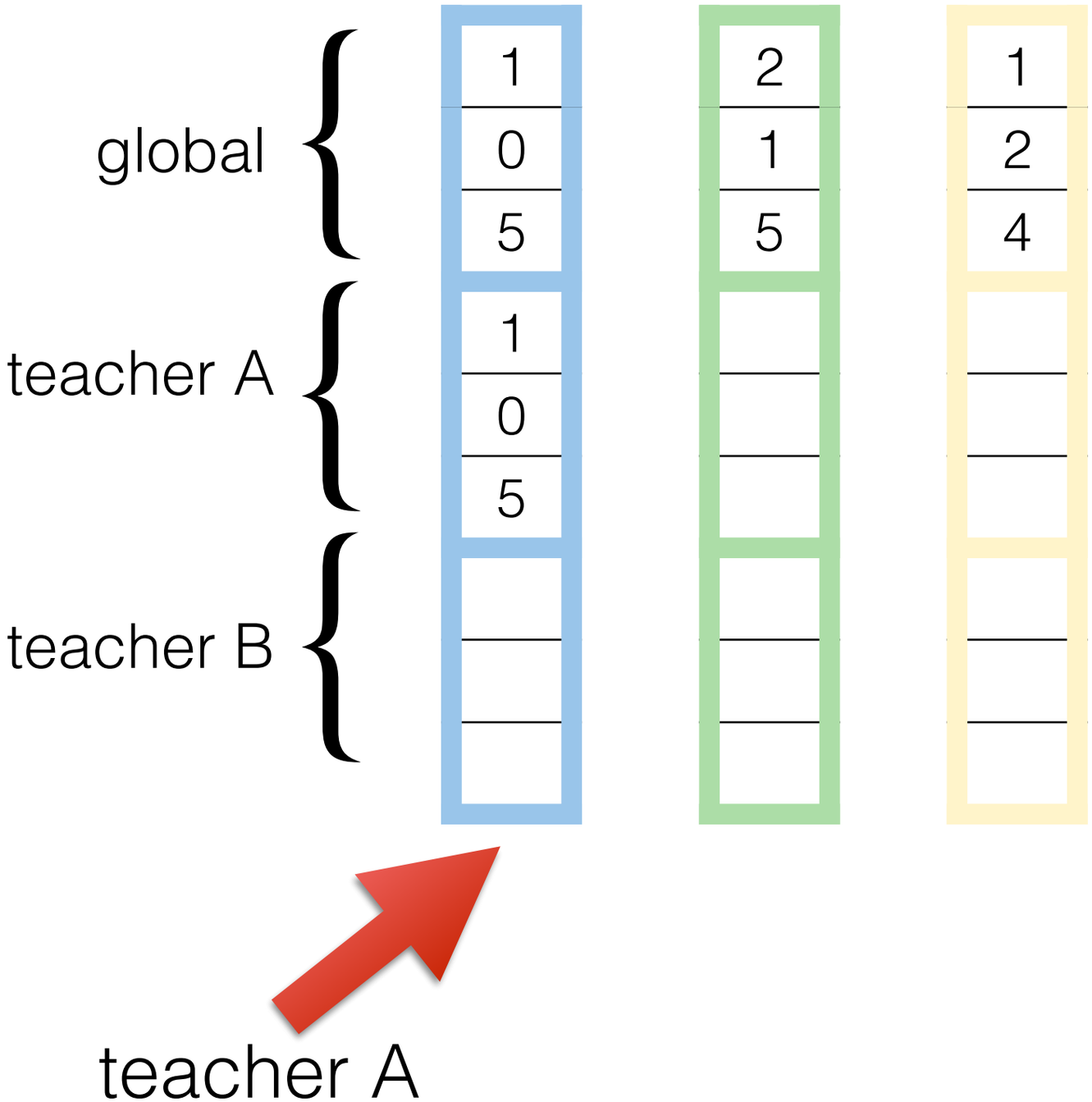
Multi-task Learning



Multi-task Learning



Multi-task Learning



Multi-task Learning

global	1	2	1
	0	1	2
	5	5	4
teacher A	1		
	0		
	5		
teacher B	0		
	0		
	0		

teacher A

Multi-task Learning

global	1	2	1
	0	1	2
	5	5	4
teacher A	1		
	0		
	5		
teacher B	0		
	0		
	0		

teacher A

Multi-task Learning

global	1	2	1
	0	1	2
	5	5	4
teacher A	1	2	
	0	1	
	5	5	
teacher B	0		
	0		
	0		

teacher A

Multi-task Learning

global	1	2	1
	0	1	2
	5	5	4
teacher A	1	2	
	0	1	
	5	5	
teacher B	0	0	
	0	0	
	0	0	

teacher A

Multi-task Learning

global	1	2	1
	0	1	2
	5	5	4
teacher A	1	2	
	0	1	
	5	5	
teacher B	0	0	
	0	0	
	0	0	

teacher B

Multi-task Learning

global	1	2	1
	0	1	2
	5	5	4
teacher A	1	2	0
	0	1	0
	5	5	0
teacher B	0	0	
	0	0	
	0	0	

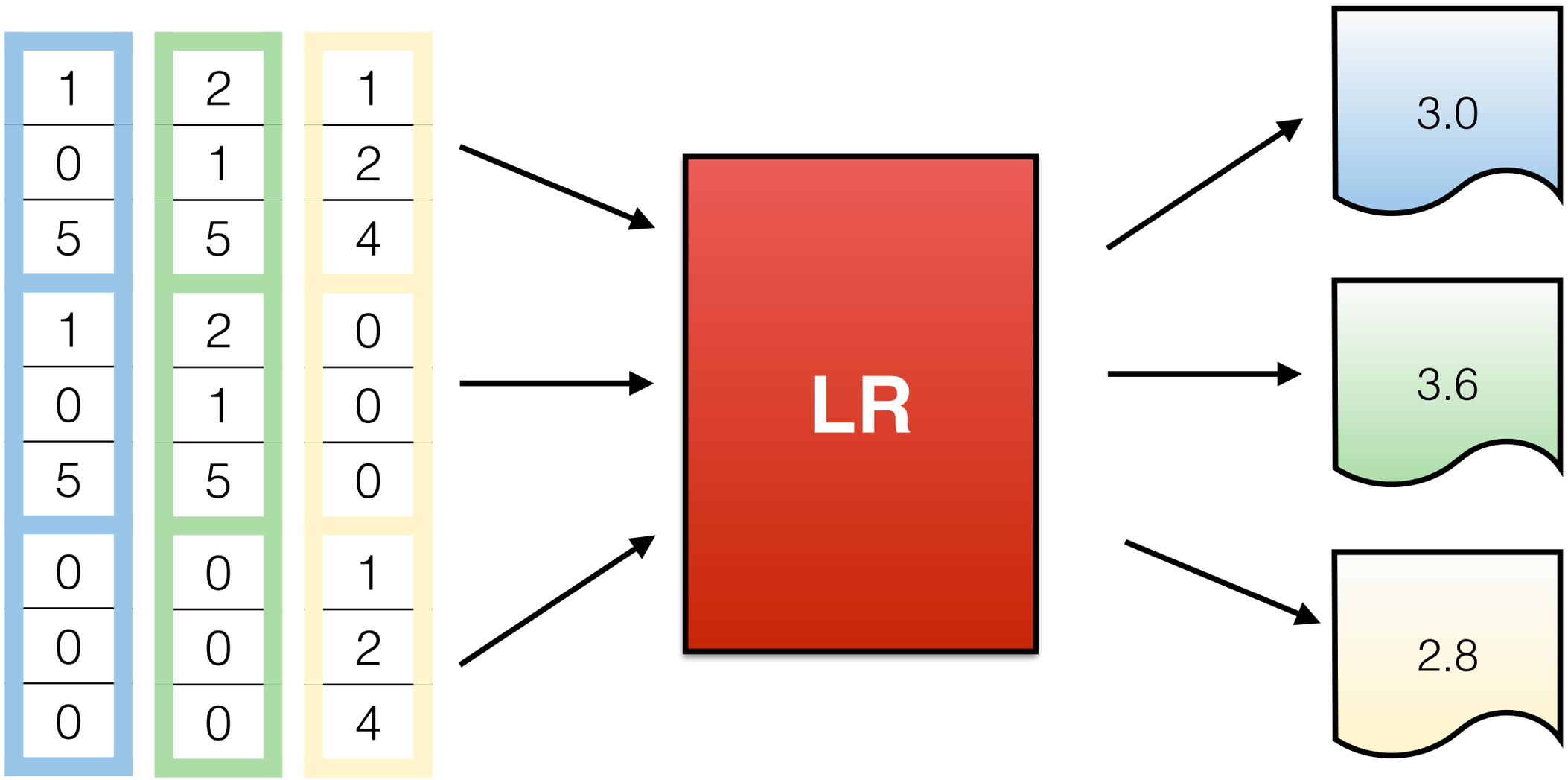
teacher B

Multi-task Learning

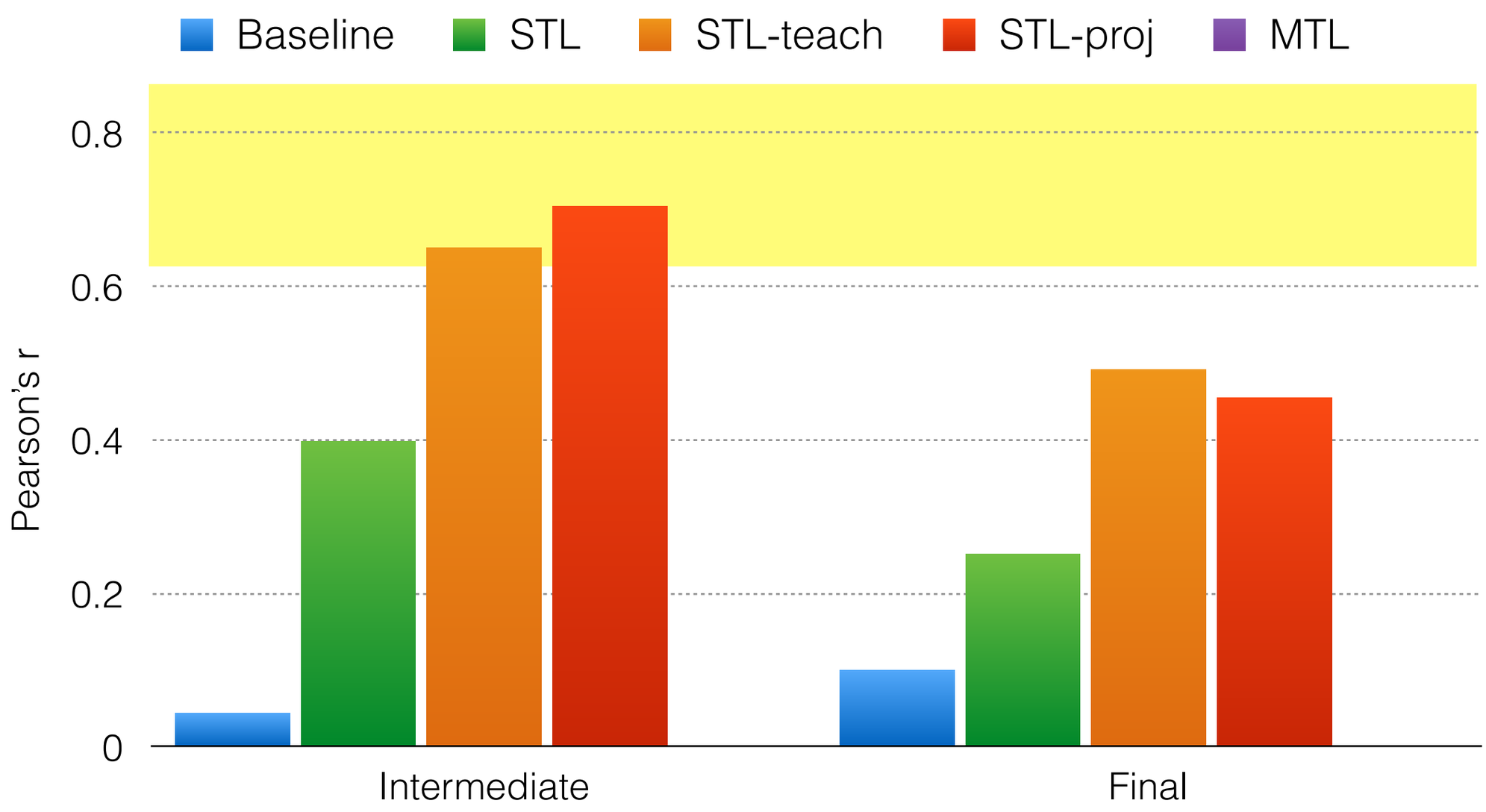
global	1	2	1
	0	1	2
	5	5	4
teacher A	1	2	0
	0	1	0
	5	5	0
teacher B	0	0	1
	0	0	2
	0	0	4

teacher B

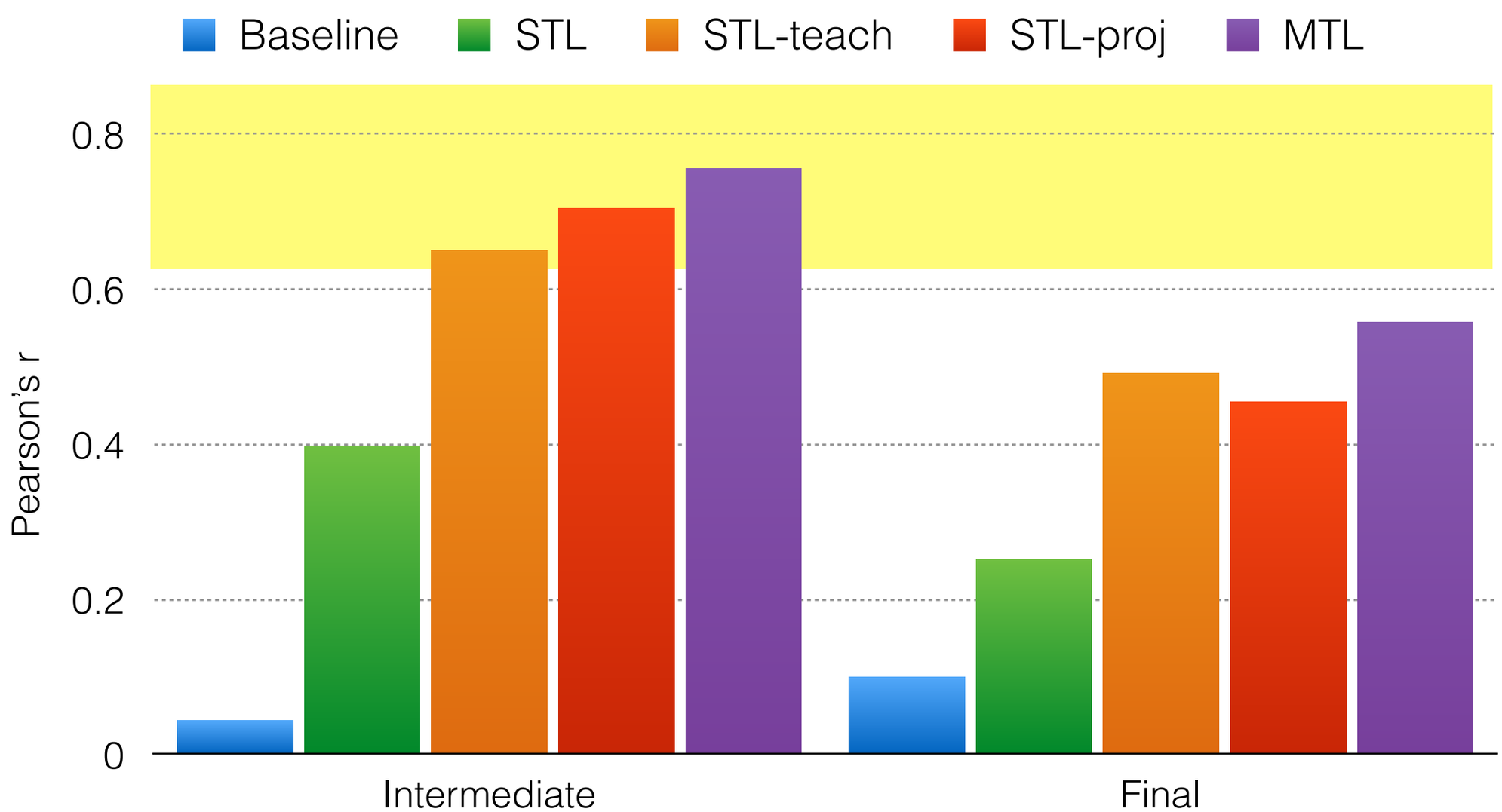
Multi-task Learning



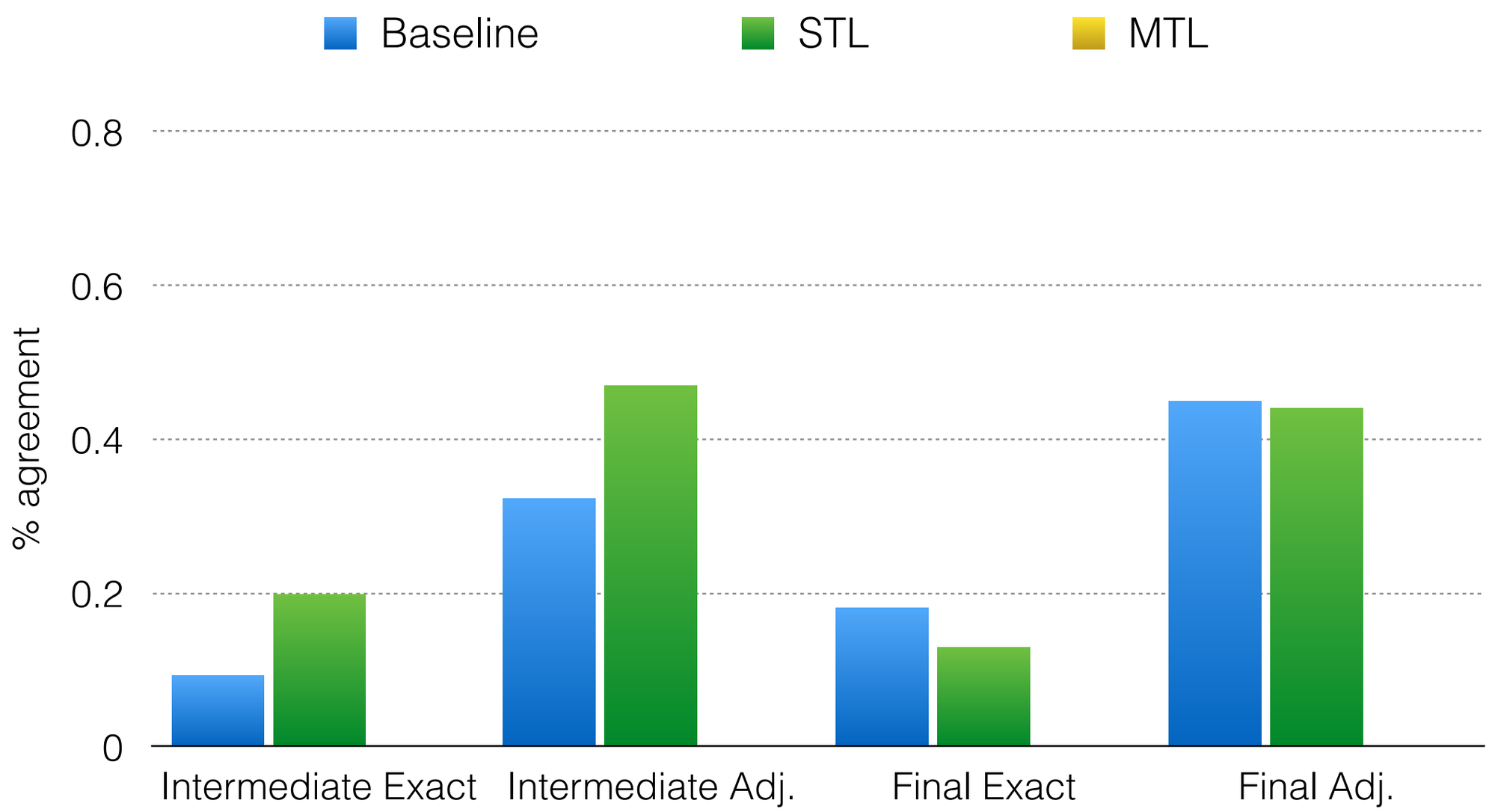
MTL Results


Baseline
STL
STL-teach
STL-proj
MTL
Pearson's r
0.8
0.6
0.4
0.2
0
Intermediate
Final

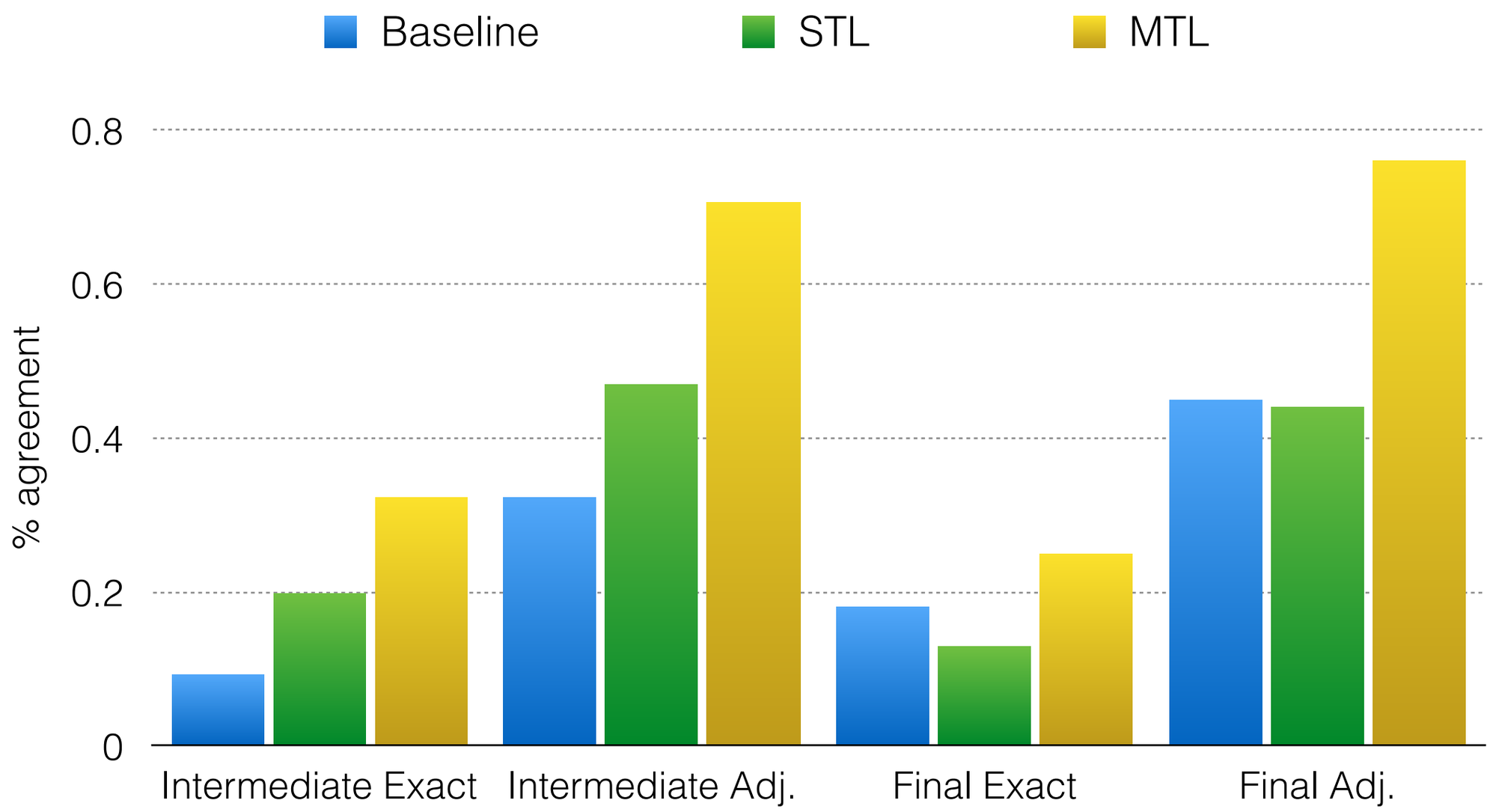
MTL Results


Baseline
STL
STL-teach
STL-proj
MTL
0.8
0.6
0.4
0.2
0
Pearson's r
Intermediate
Final

MTL Results


Baseline
STL
MTL
0.8
0.6
0.4
0.2
0
% agreement
Intermediate Exact
Intermediate Adj.
Final Exact
Final Adj.

MTL Results

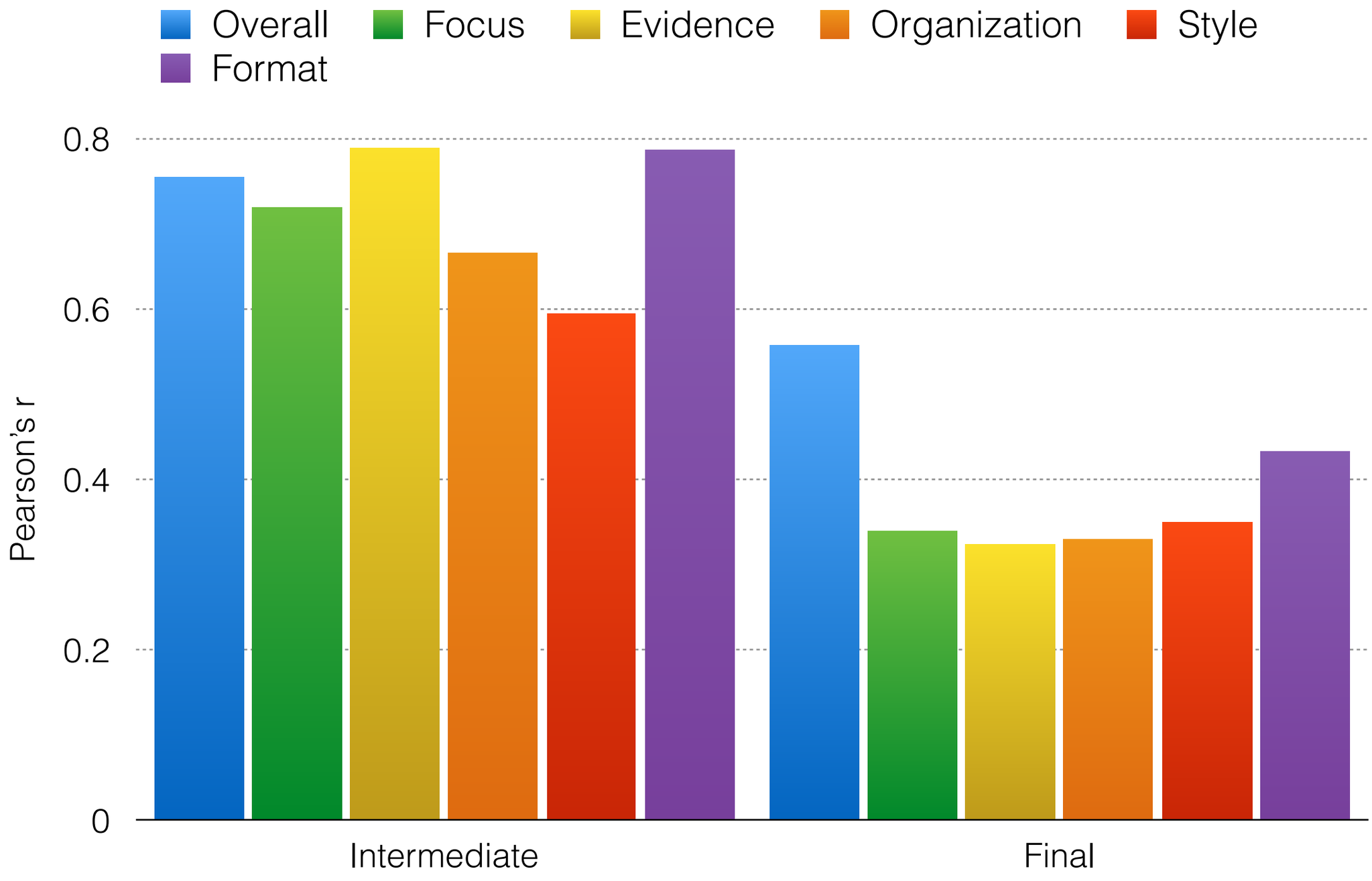

Baseline
STL
MTL
0.8
0.6
0.4
0.2
0
% agreement
Intermediate Exact
Intermediate Adj.
Final Exact
Final Adj.

Other experiments

Can we predict…

- specific rubric scores?
- the improvement/decline between aligned drafts?
- scores given by unseen teachers?

Predict rubric scores



Predict improvement

Can we predict the score change between aligned drafts?

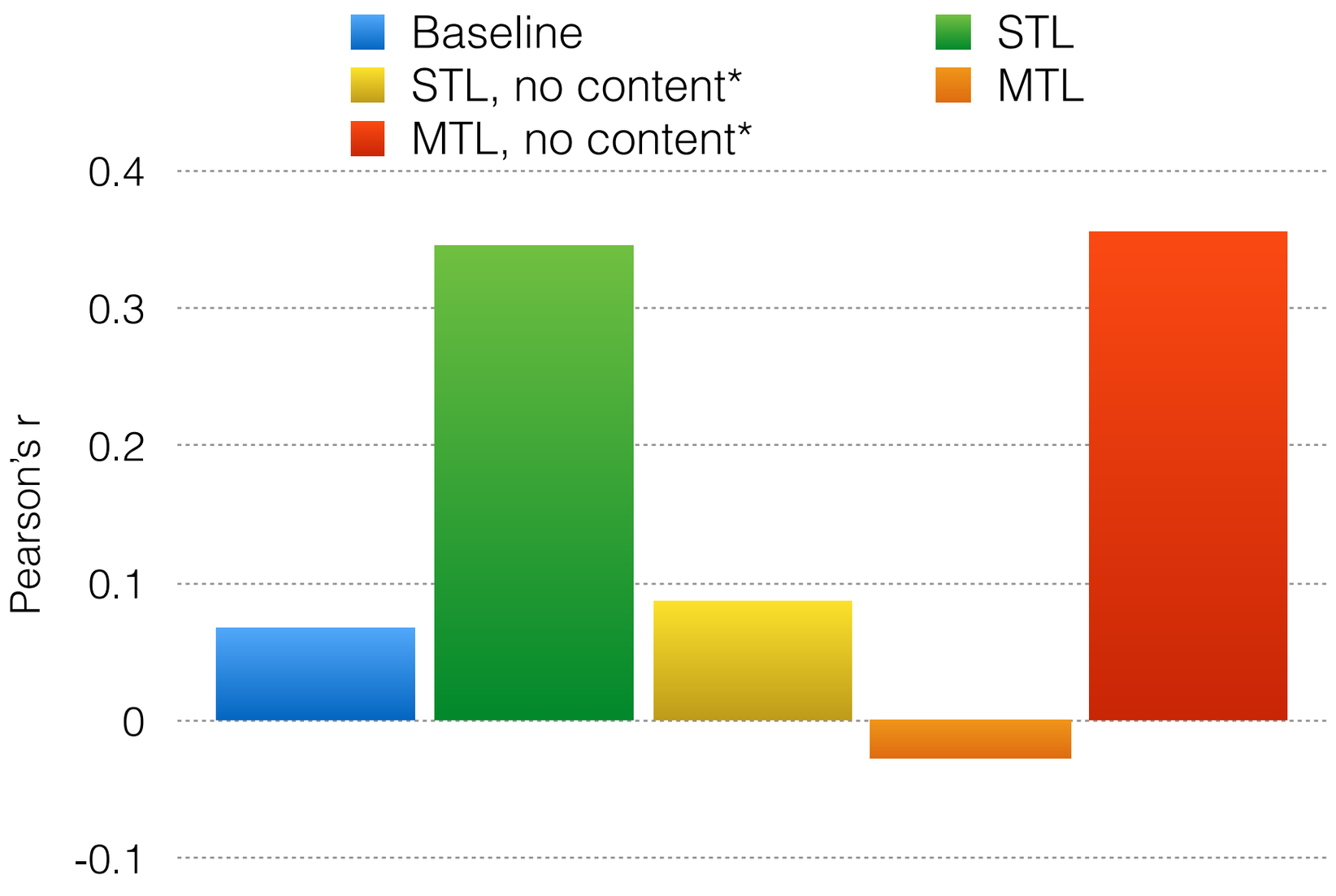
- Train: 794 draft pairs
- Test: 50 pairs

Predict improvement

Calculate the difference between the paired feature vectors

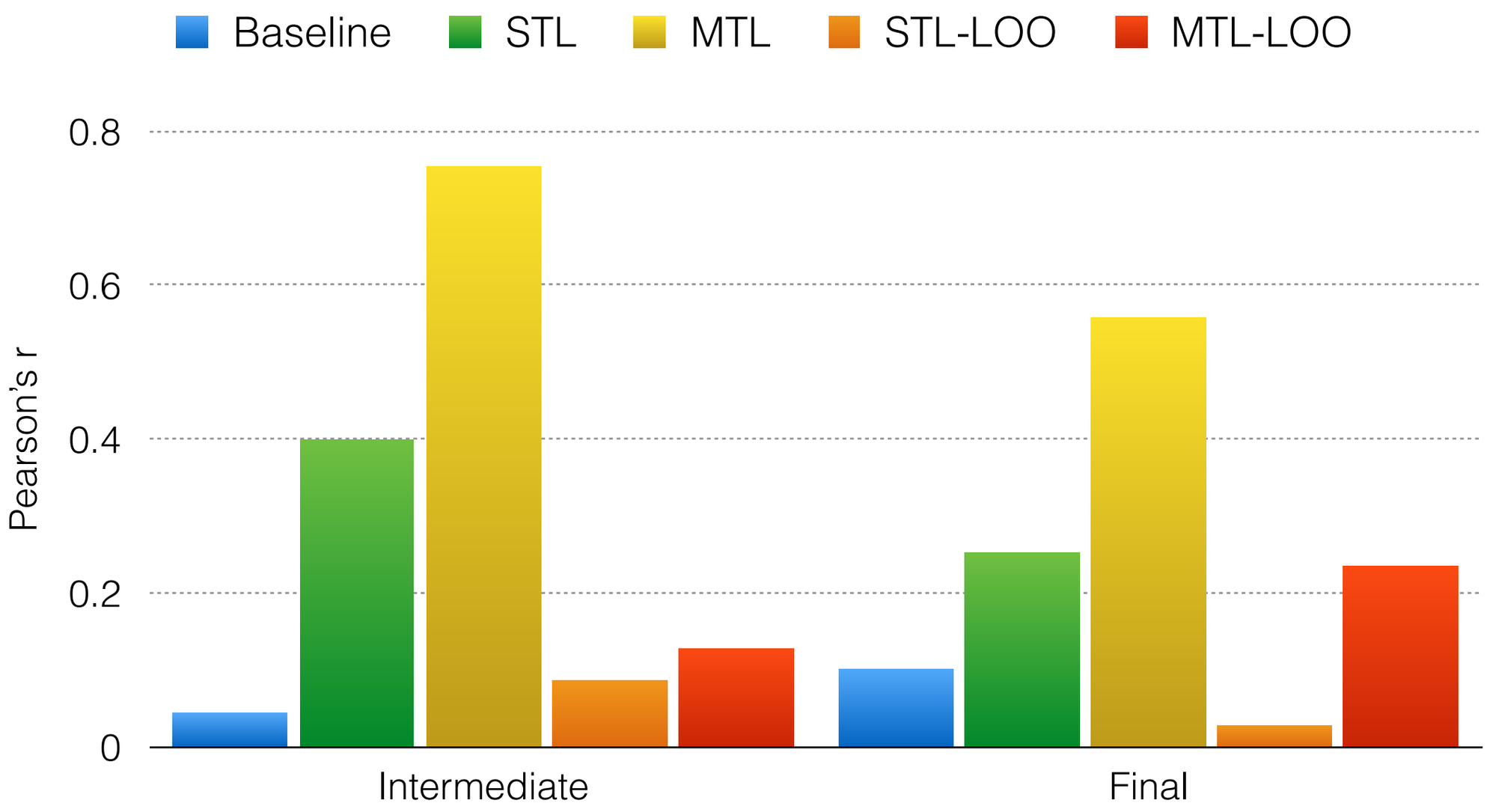
Intermediate		Final		Delta
2	-	1	=	1
1		0		1
4		6		-2

Predict improvement



* no token unigram or trigram features

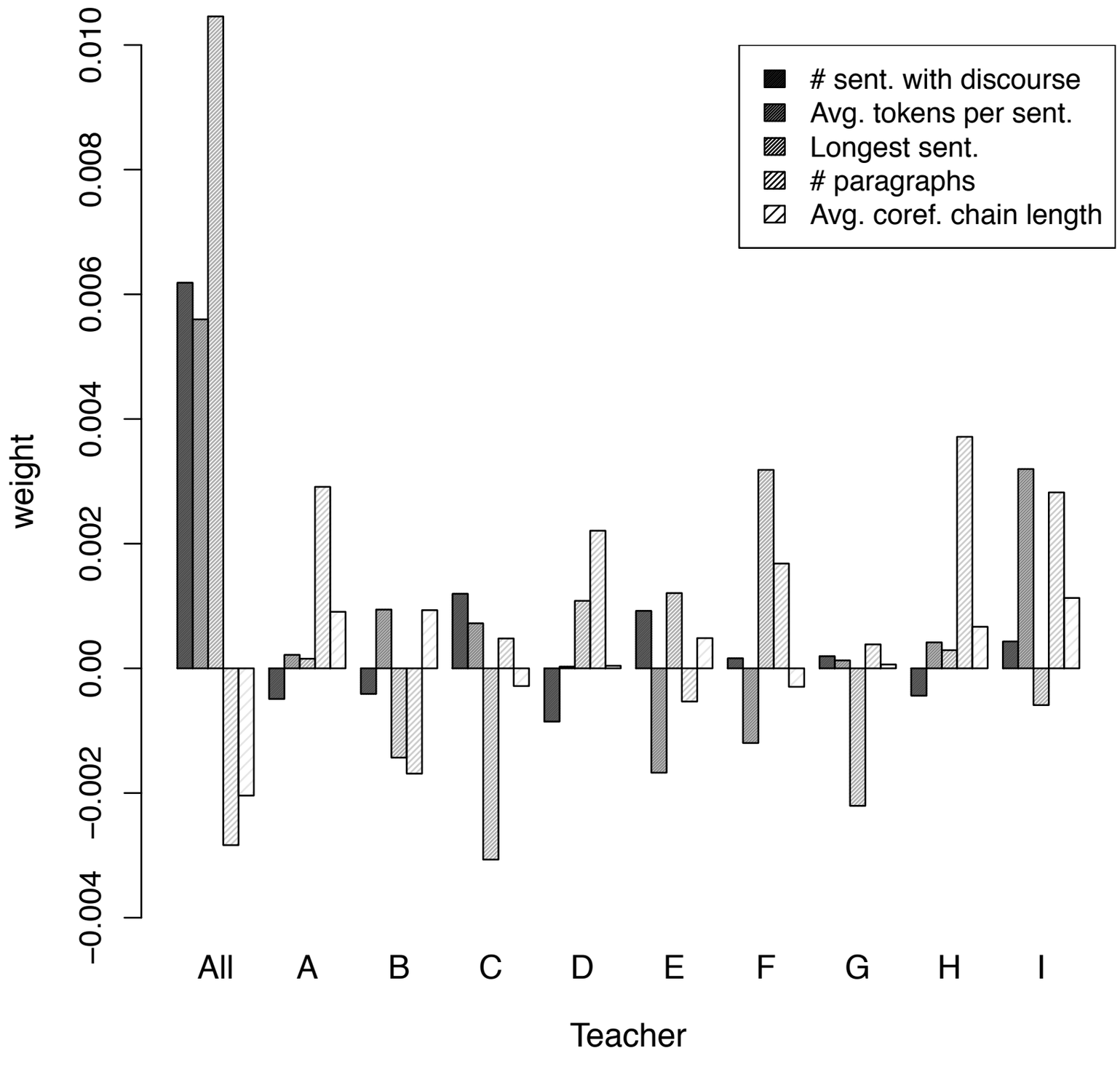
Unseen teachers



Potential applications

- Examine feature weights across individual teacher models
- Potentially share this information to help teachers grade more consistently

Potential application



Summary

- A new corpus of student essays
  - more representative of college writing
- Multi-task learning to account for differences across teachers

Future Work

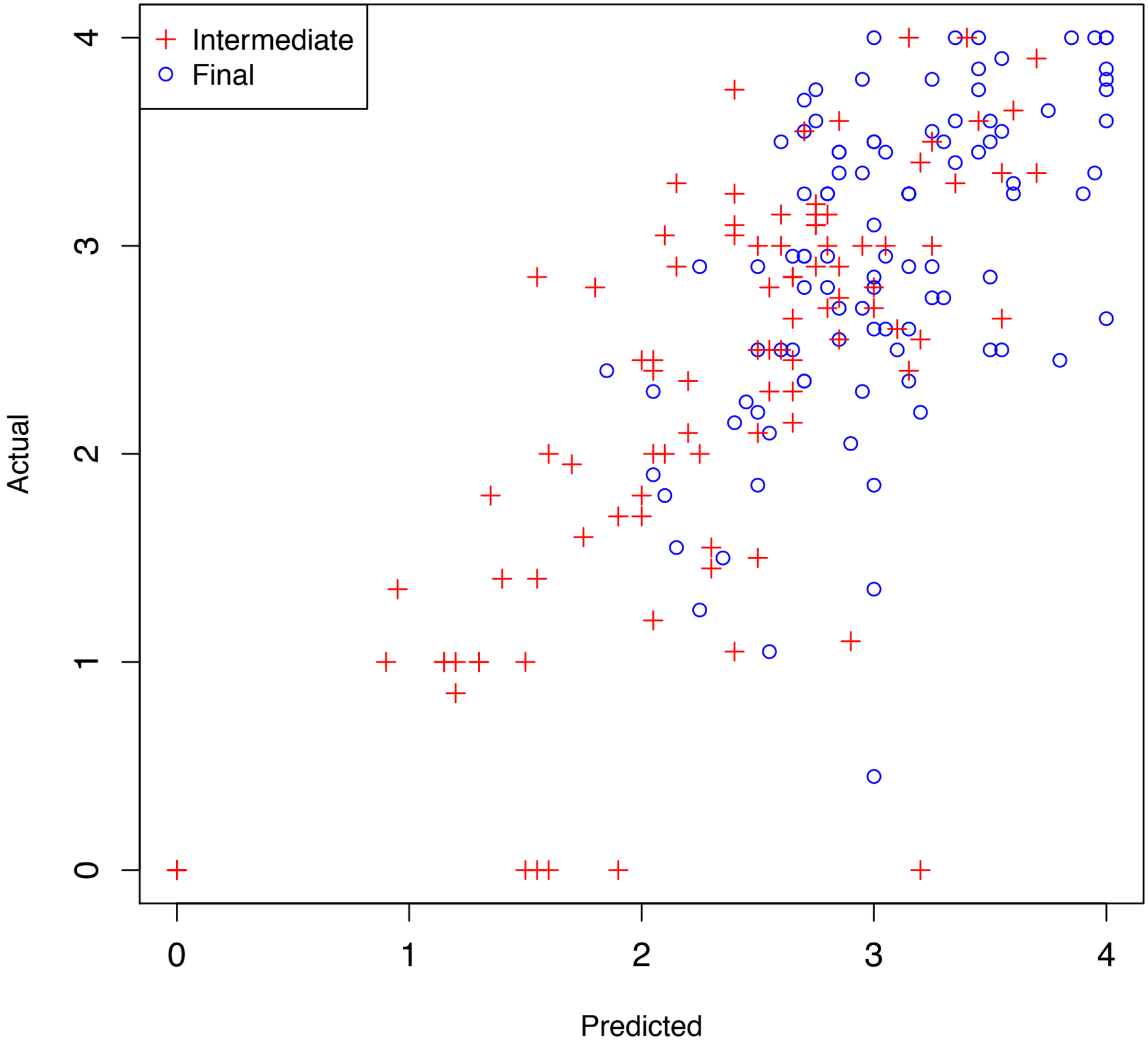
This task

- Tailor features for specific rubric categories
- Better model for unseen teachers
- Validate scores
- Test MTL on different writing corpora

This corpus

- Examine types of revisions made
- Categorize teacher comments
- Align teacher comments to spans of text

Thank you


Intermediate
Final
Actual
Predicted
0
1
2
3
4