

CS 293/EDUC 473

Measure Validation

Today's class

- Pitch rubric
- Pitch example video
- Case study on measurement & validation
- Discussion by Kathy and Riz

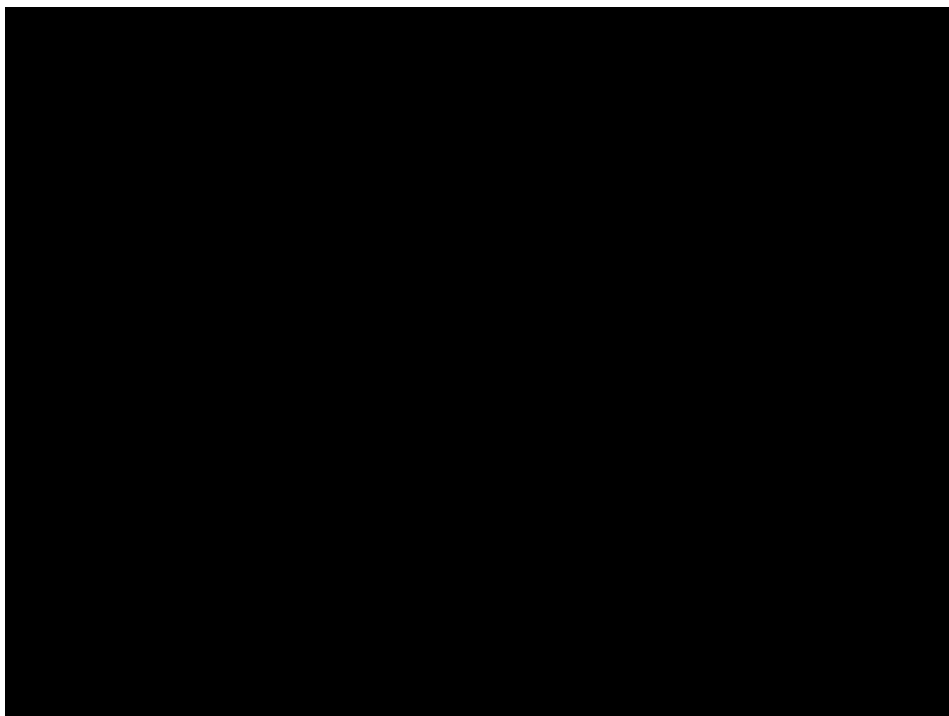


Practice pitch structure

- You will give a 4 minute pitch
- Everyone will receive feedback from 2 students + the instructors
 - Students giving feedback on the same pitch can discuss their feedback but have to submit individually
- Quality of feedback is part of your practice pitch grade



Rubric

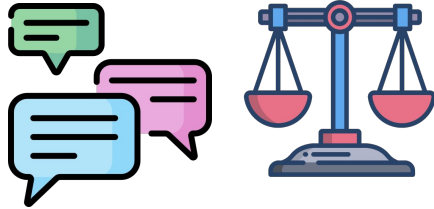




Unsupervised Measurement Case study: Teachers' uptake of student ideas

1

Measure an educationally important discourse phenomenon



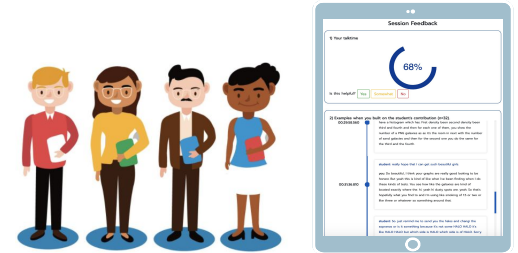
2

Validate the measure using existing data



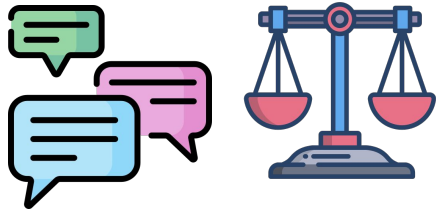
3

Deploy the measure to give teachers feedback



1

Measure an educationally important discourse phenomenon



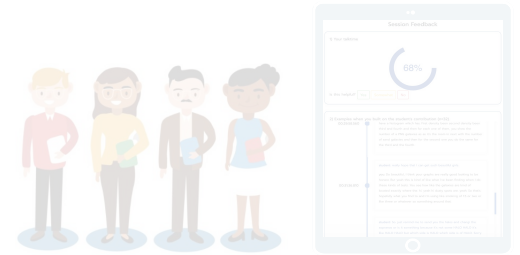
2

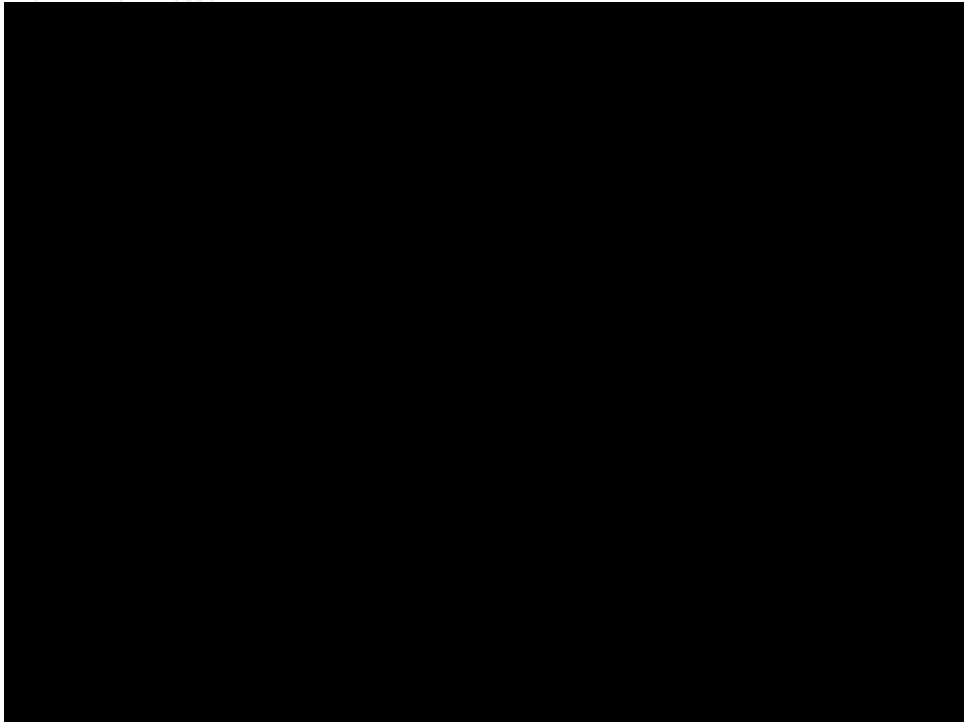
Validate the measure using existing data



3

Deploy the measure to give teachers feedback





uptake [Collins, 1982; Nystrand et al., 1997]

~ revoicing [O'Connor & Michaels, 1993; Herbel-Eisenmann et al., 2009]

Uptake is to build on the interlocutor's contribution.

S

I added 30 to 70...

acknowledgment

Okay.

t₁

collaborative completion

And you got what?

t₂

repetition

Okay, you added 30 to 70.

t₃

reformulation

Good, you did the first step.

t₄

elaboration

Where did the 70 come from?

t₅

Uptake serves several functions

STRUCTURAL

it creates **coherence**

[Halliday & Hasan, 1976; Grosz et al., 1977; Hobbs, 1979]

PRAGMATIC

it enables **grounding**

[Clark & Schaefer, 1989]

SOCIAL

it promotes **collaboration** and makes the interlocutor **feel heard**

[Bakhtin, 1981; Nystrand et al., 1997]

demonstrating **understanding** of the interlocutor's contribution by accepting it as part of the common ground

When teachers take up student ideas, ...

- they **amplify student voices** and promote **dialogic instruction**

[Wells, 1999; Nystrand et al., 1997]

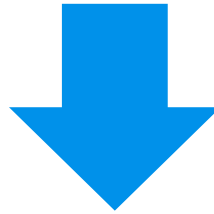
- **students learn and do better**

[Brophy, 1984; O'Connor & Michaels, 1993; Nystrand et al., 2003]





Existing methods for measuring and improving teachers' uptake at scale are **prohibitively resource-intensive**



Fully-automated measure:

- domain-transferable
- resource-efficient
- protects privacy

How can we measure uptake?

S

I added 30 to 70...

acknowledgment

Okay.

t_1

collaborative completion

And you got what?

t_2

repetition

Okay, you added 30 to 70.

t_3

reformulation

Good, you did the first step.

t_4

elaboration

Where did the 70 come from?

t_5

How can we measure uptake?

S

I added 30 to 70...

acknowledgment

Okay.

t_1

collaborative completion

And you got what?

t_2

utterance similarity?
word overlap?

repetition

Okay, you added 30 to 70.

t_3

reformulation

Good, you did the first step.

t_4

elaboration

Where did the 70 come from?

t_5

How can we measure uptake?

S

I added 30 to 70...

How can we capture these strategies?

acknowledgment

Okay.

t_1

collaborative completion

And you got what?

t_2

repetition

Okay, you added 30 to 70.

t_3

reformulation

Good, you did the first step.

t_4

elaboration

Where did the 70 come from?

t_5

Uptake as dependence

How easily can we tell that T is a response to S and not some random response T'?

S

I added 30 to 70...

Where did the 70 come from?

t_1

?

Okay.

t_2

???

Let's draw a circle.

t_3

→ **Formal goal:** estimate how far is $T|S$ from $T'|S$

→ **Formal goal:** estimate how far is $T|S$ from $T'|S$

Pointwise Jensen Shannon Divergence (PJSD)

$$pJSD(t, s) := -\frac{1}{2} \left(\log P(Z = 1 | M = t, s) + \mathbb{E} \log(1 - P(Z = 1 | M = T', s)) \right) + \log(2)$$

where (\mathbf{S}, \mathbf{T}) is a teacher-student utterance pair, \mathbf{T}' is a randomly sampled teacher utterance and $M := Z\mathbf{T} + (1 - Z)\mathbf{T}'$ is a mixture of the two with a binary indicator variable $\mathbf{Z} \sim \mathbf{Bern}(\mathbf{p}=0.5)$.

→ **Formal goal:** estimate how far is $T|S$ from $T'|S$

Pointwise Jensen Shannon Divergence (PJSD)

$$pJSD(t, s) := -\frac{1}{2} \left(\log \right)$$

$$\mathbb{E} \log(1 - P(Z = 1|M))$$

where (S, T) is a teacher-student utterance pair, T is a teacher utterance and $M := ZT + (1 - Z)S$ is a binary indicator variable $Z \sim \text{Bern}(p)$

ACL 2021

Measuring Conversational Uptake: A Case Study on Student-Teacher Interactions

Dorottya Demszky¹ Jing Liu² Zid Mancenido³ Julie Cohen⁴
Heather Hill³ Dan Jurafsky¹ Tatsunori Hashimoto¹

¹Stanford University ²University of Maryland ³Harvard University ⁴University of Virginia
{ddemszky, thashim}@stanford.edu

Abstract

In conversation, *uptake* happens when a speaker builds on the contribution of their interlocutor by, for example, acknowledging, repeating or reformulating what they have said. In education, teachers' uptake of student contributions has been linked to higher student achievement. Yet measuring and improving teachers' uptake at scale is challenging, as existing methods require expensive annotation by experts. We propose a framework for computationally measuring uptake, by (1) releasing a dataset of student-teacher exchanges extracted from US math classroom transcripts annotated for uptake by experts; (2) formalizing uptake as pointwise Jensen-Shannon Divergence (PJSD), estimated via next utterance classification; (3) conducting a linguistically-



Figure 1: Example student utterance s and possible teacher replies t , illustrating different uptake strategies.

which is especially important in contexts like edu

Pointwise Jensen Shannon Divergence (PJSD)

~ **Next utterance classification task**

s	t	label
I added 30 to 70...	Where did the 70 come from?	1
I added 30 to 70...	Let's draw a circle.	0
I added 30 to 70...	Okay.	0

Model's predicted score for $(s, t) =$
Estimate for t 's uptake of s

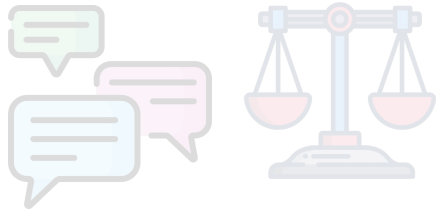
Model training via next utterance classification

- BERT [Devlin et al., 2019]
- Combination of 3 training datasets:
 - Switchboard
 - Elementary math dataset (NCTE)
 - Tutoring dataset



1

Measure an educationally important discourse phenomenon



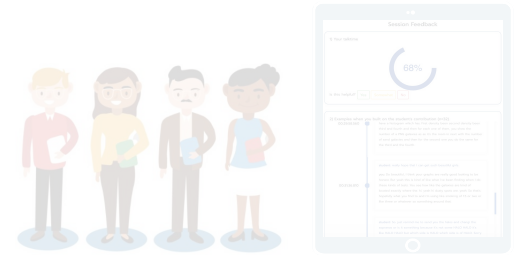
2

Validate the measure using existing data

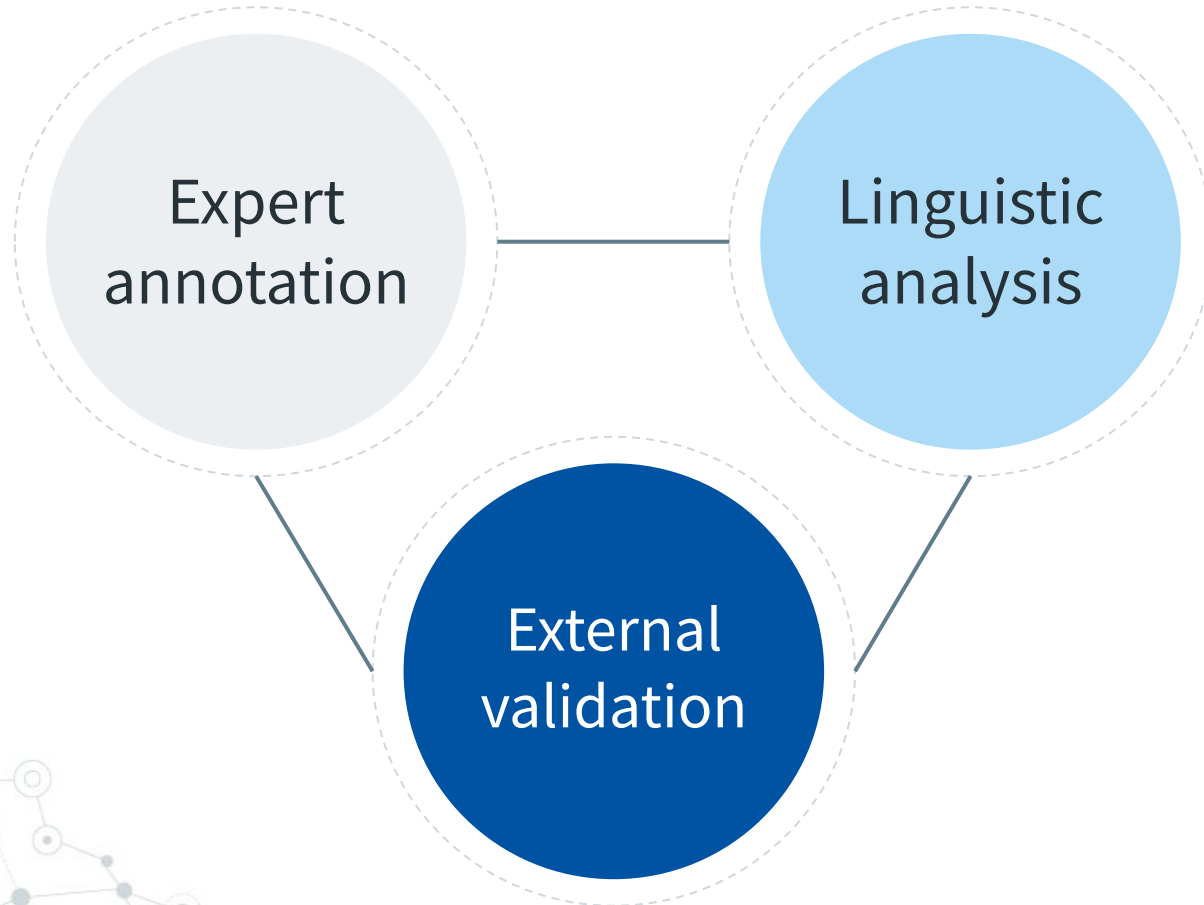


3


Deploy the measure to give teachers feedback



Validation methods



Validation methods



Expert
annotation

Validation #1: Comparison to expert labels

- annotated **2246 student-teacher (S, T) utterance pairs**
 - from the NCTE elementary math classroom dataset
- 3 expert raters / example
- given an (S, T) pair, rate T for “low”, “mid” or “high” uptake

Interrater agreement

Leave-out Spearman ρ is .474 on the full dataset (.539 on a subset of the data that all 13 raters rated during the pilot (n=70)). Fleiss $\kappa = .286$.

→ **comparable to those obtained in widely-used classroom observation protocols** such as Classroom Assessment Scoring System (CLASS) and Mathematical Quality of Instruction (MQI) that include parallel measures to our uptake construct (see Kelly et al., 2020 for a summary).

Validation #1: Comparison to expert labels

Example	Label
S: 'Cause you took away 10 and 70 minus 10 is 60. T: Why did we take away 10?	high
S: There's not enough seeds. T: There's not enough seeds. How do you know right away that 128 or 132 or whatever it was you got doesn't make sense?	high
S: Teacher L, can you change your dimensions like 3-D and stuff for your bars? T: You can do 2-D or 3-D, yes. I already said that.	mid
S: The higher the number, the smaller it is. T: You got it. That's a good thought.	mid
S: An obtuse angle is more than 90 degrees. T: Why don't we put our pencils down and just do some brainstorming, and then we'll go back through it?	low
S: Because the base of it is a hexagon. T: Student K?	low

Validation #1: Comparison to expert labels

Our uptake measure

Correlation
with raters

0.540***

*** $p < 0.001$

This score is considered high for a construct as subjective and heterogeneous as uptake! [Kelly et al., 2020]
(leave-out interrater correlation = 0.539)

Validation #1: Comparison to expert labels

Our uptake measure

Correlation
with raters

0.540***

*** $p < 0.001$

✓ Does better than several NLP baselines!

Validation #1: Comparison to expert labels

	Model	Correlation with raters
word overlap	%-IN-S	0.449
word overlap	Jaccard	0.450
word overlap	BLEU	0.510
word overlap	%-IN-T	0.523
Our uptake measure		0.540***

Validation #1: Comparison to expert labels



	Model	Correlation with raters
utterance similarity	Sentence-Bert	0.390
utterance similarity	Glove	0.424
word overlap	%-IN-S	0.449
utterance similarity	Universal Sentence Encoder	0.448
word overlap	Jaccard	0.450
word overlap	BLEU	0.510
word overlap	%-IN-T	0.523
Our uptake measure		0.540***

Validation #1: Comparison to expert labels

Our uptake measure

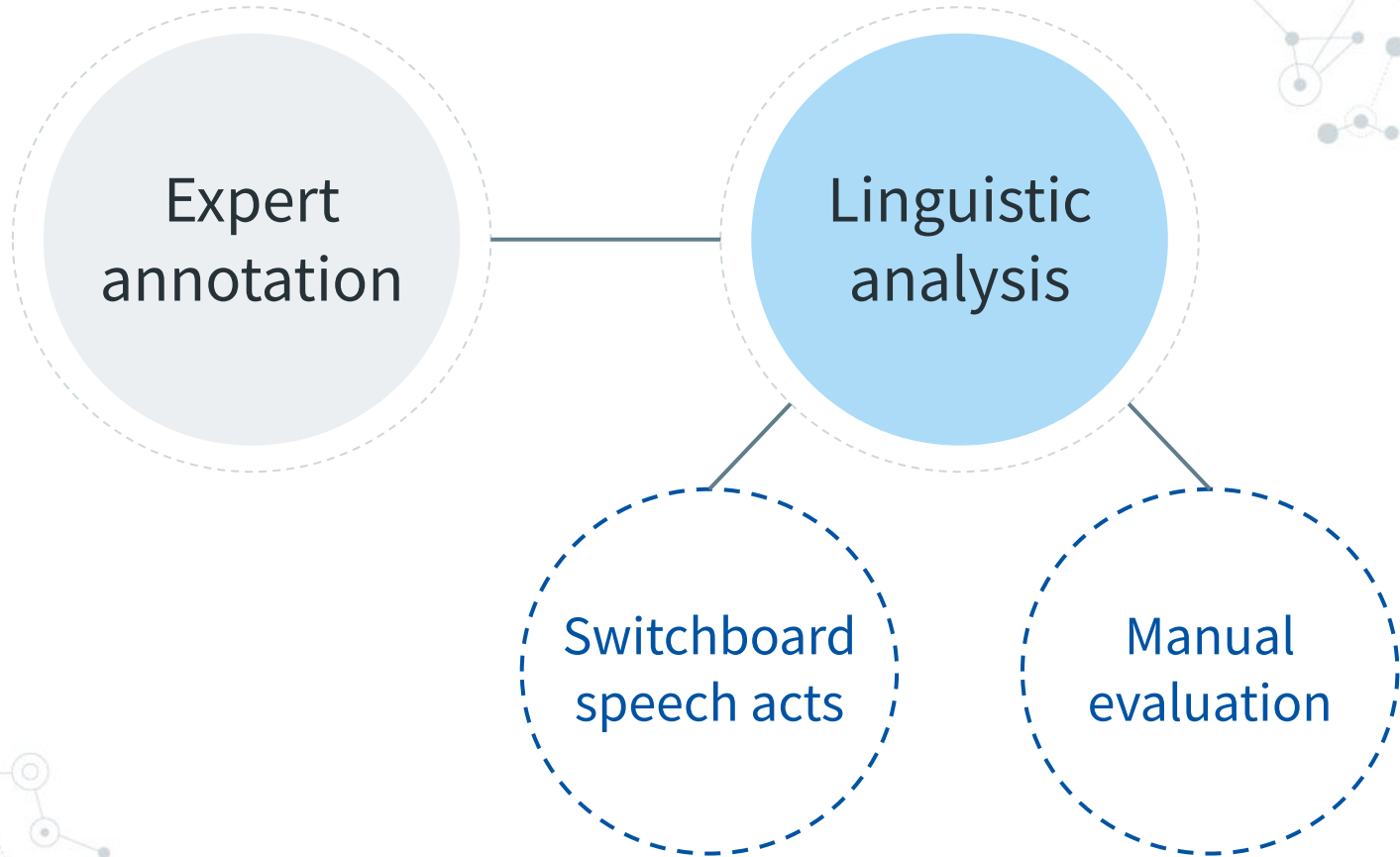
**Correlation
with raters**

0.540***

*** $p < 0.001$

What kind of linguistic phenomena does the measure capture?

Validation methods



Validation #2: Qualitative comparison via dialog acts

 2nd best model: percentage of tokens from S that are in

T

	Model	Correlation with raters
word overlap	%-IN-T	0.523
Our uptake measure		0.540***

Validation #2: Qualitative comparison via speech acts

Switchboard corpus

Do you have a pet, Randy?

yes-no question

Yeah, we currently have a poodle.

answer

...

It just turned two I believe.

statement

Oh, it's still just a pup.

reformulation

...

What do you call the dog?

wh-question

Uh, it's Mitzi.

answer

Mitzi.

repetition

Validation #2: Qualitative comparison via speech acts

Switchboard corpus

Do you have a pet, Randy?

yes-no question

Yeah, we currently have a poodle.

answer

...

It just turned two I believe.

statement

Oh, it's still just a pup.

reformulation

...

What do you call the dog?

wh-question

similar predictions!

Uh, it's Mitzi.

answer

Mitzi.

repetition



%-IN-T OUR MEASURE

Validation #2: Qualitative comparison via speech acts

Switchboard corpus

Do you have a pet, Randy?

Yeah, we currently have a poodle.

...

It just turned two I believe.

Oh, it's still just a pup.

...

What do you call the dog?

Uh, it's Mitzi.

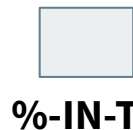
Mitzi.

yes-no question

answer

statement

reformulation



OUR MEASURE

wh-question

answer

repetition

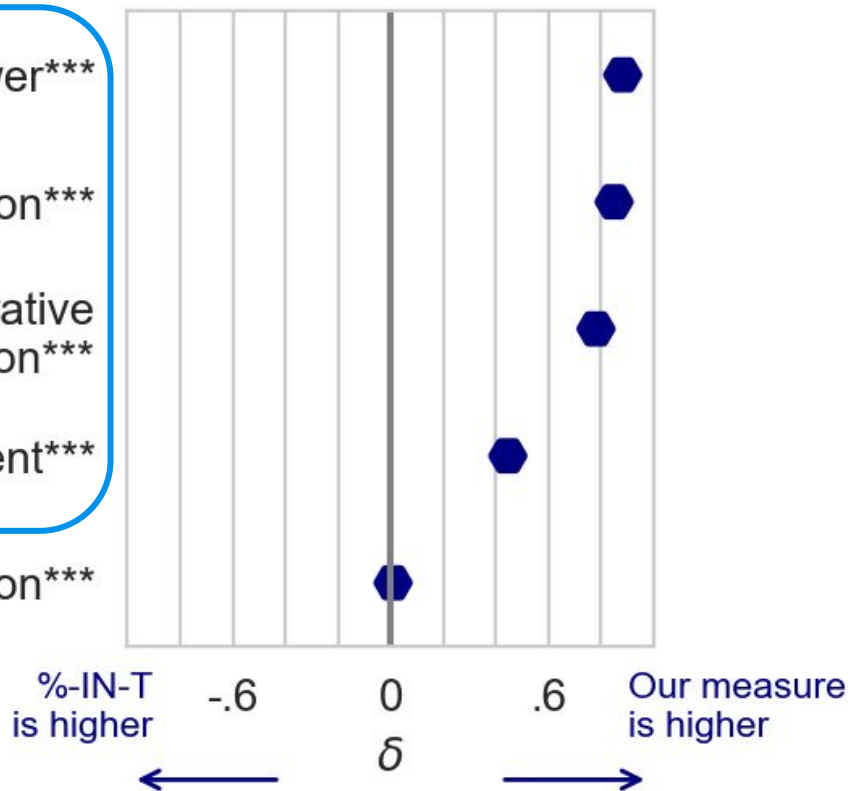
Validation #2: Qualitative comparison via speech acts

Switchboard corpus

Our measure captures a richer range of uptake strategies than %-IN-T

answer***
reformulation***
collaborative completion***
acknowledgment***

repetition***



Validation #2: Qualitative comparison via dialog acts

Compare predictions for **Switchboard-DAMSL dialog act tags** [Jurafsky et al., 1997]

a

That was almost an example of an invasion that turns out to be not invasive.

reformulation

Right, it turned out to be, uh, uh, an invitation.

b

No world overlap!

$\% \text{-IN-T} = 0 \lll \text{PJSD} = 0.99$

Validation #2: Qualitative comparison via speech acts

Switchboard corpus

Do you have a pet, Randy?

Yeah, we currently have a poodle.

...

It just turned two I believe.

Oh, it's still just a pup.

...

What do you call the dog?

Uh, it's Mitzi.

Mitzi.

yes-no question

answer

statement

reformulation

wh-question

answer

repetition

%-IN-T

OUR MEASURE

%-IN-T

OUR MEASURE

%-IN-T

OUR MEASURE

PJSD captures **elaboration prompts** better than %-IN-T

In the NCTE data, manually label high uptake examples where PJSD significantly outperforms %-IN-T (N=67). **N is very small!**

Category	Example	Odds ratio
elaboration prompt	S: so it means that the whole equation is only the same. T: what does it mean? i still don't understand what is it?	4.25*
reformulation	S: multiplication is like, say, for instance, nine times twenty. you just take - nine just nine times and add it up. T: okay, so repeated addition.	2.6
answer	S: do we look at the d or the m first? T: the m. what's this called, that i'm writing?	2.67
collaborative completion	S: we had to add twenty-four plus twenty-four. T: because there are how many triangles?	0

Validation #2: Qualitative comparison via **speech acts**

S

I added 30 to 70...

acknowledgment

Okay.

t_1

collaborative completion

And you got what?

t_2

word overlap can only capture repetition

repetition

Okay, you added 30 to 70.

t_3

reformulation

Good, you did the first step.

t_4

elaboration

Where did the 70 come from?

t_5

Validation #2: Qualitative comparison via **speech acts**

S

I added 30 to 70...

**our measure can
better capture
all of these
phenomena**

acknowledgment

Okay.

t_1

collaborative completion

And you got what?

t_2

repetition

Okay, you added 30 to 70.

t_3

reformulation

Good, you did the first step.

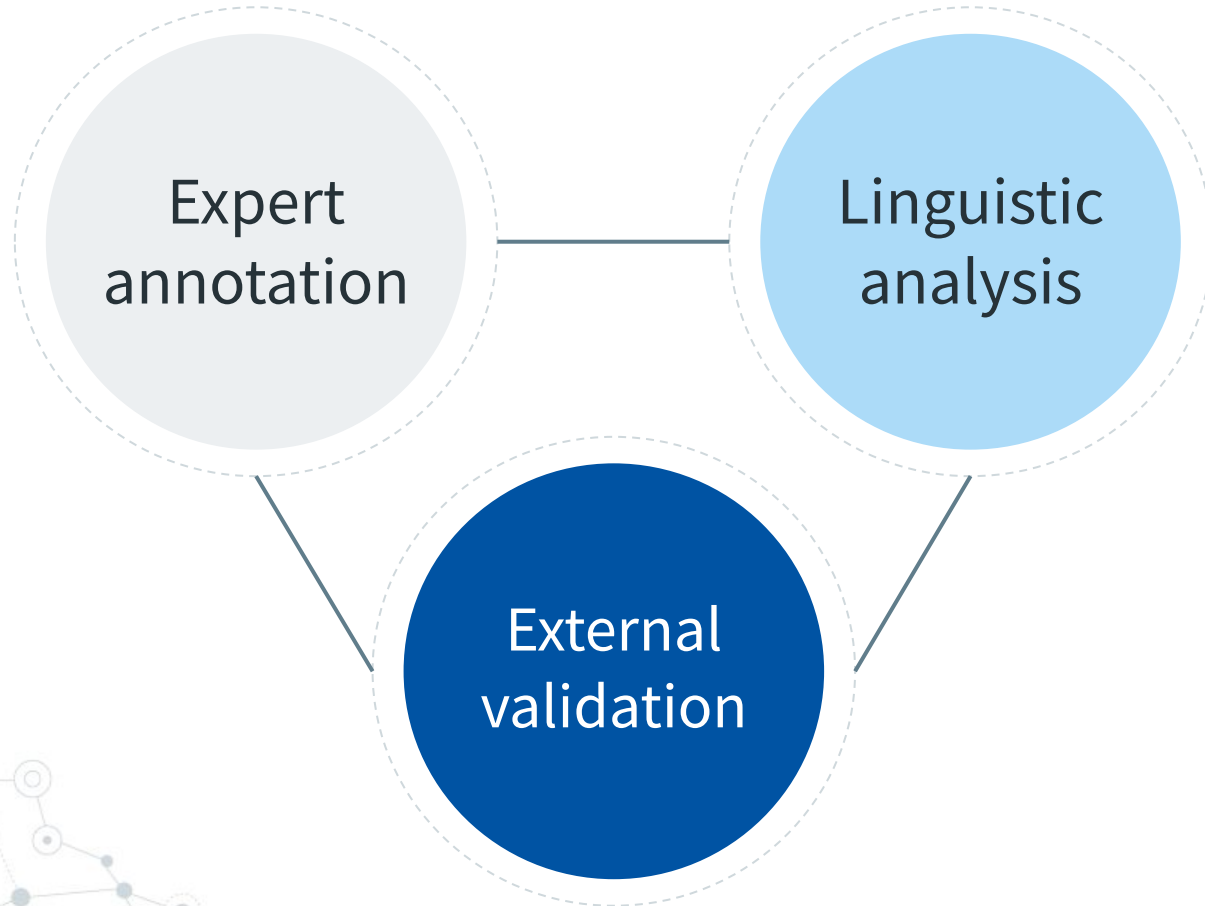
t_4

elaboration

Where did the 70 come from?

t_5

Validation methods



Validation #3: Correlation with external measurements

1

Obtain datasets with transcript-level external measurements

- classroom observation scores
- student satisfaction scores

2

Generate aggregate uptake score for each transcript

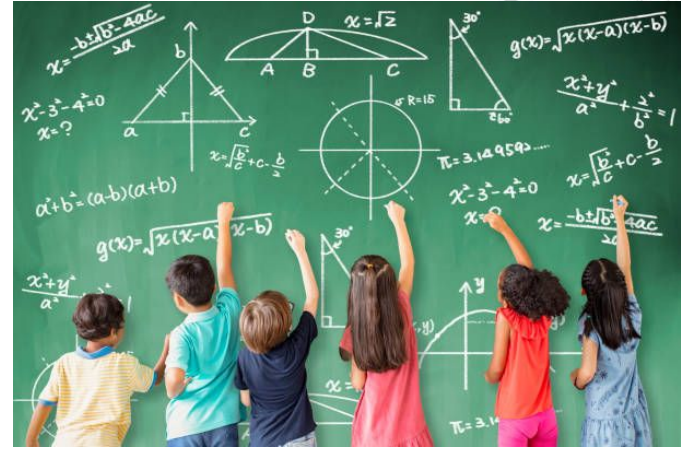
3

Correlate aggregate uptake score with external measurements

Validation #3: Correlation with external measures

NCTE dataset [Kane et al., 2015]

- elementary math classrooms
- spoken (in-person)
- whole class (20-30 students)
- external measures:
 - use of student contributions
 - math instruction quality

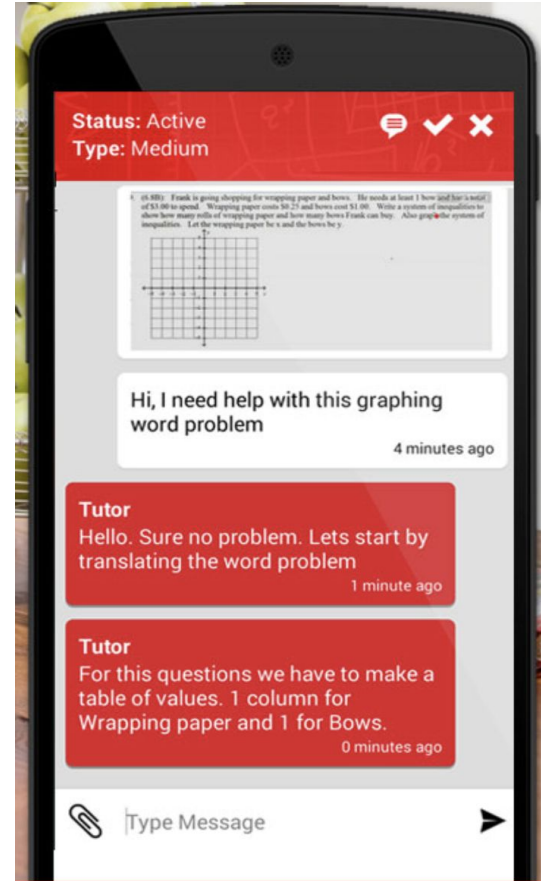


Mathematical Quality of Instruction (MQI)
instrument

Validation #3: Correlation with external measures

Tutoring dataset

- math and science
- written
- 1:1
- external measures:
 - external reviewer rating
 - student satisfaction



Validation #3: Correlation with external measures

SimTeacher [Cohen et al., 2020]

- **not part of training data!**
- elementary literature & arts
- spoken (virtual)
- small groups
- external measures:
 - quality of feedback



Validation #3: Correlation with external measures

NCTE dataset

Tutoring dataset

SimTeacher

External measure	Beta
use of student contributions	.101***
math instruction quality	.091***
student satisfaction	.069***
external reviewer rating	.063***
quality of feedback	.127*

*** $p < 0.001$, * $p < 0.05$

comparable to average effect sizes for an effective educational intervention [Kraft, 2020]
→ uptake is a promising intervention (scalable & easily quantified)!

Extra

Collins (1981): canonical example of uptake

EXAMPLE 1: Incorporations of answer into question (+)

- T *Alright, what are they looking for?*
- C *Signals.*
- T *What signals? (+)*
- C1 *Red.*
- C2 *Red light and green.*
- C3 *Three signals.*
- T *Alright, traffic signals.*
- T *Where do you find those? (+)*
- C *On the street.*
- T *Alright, where on the street? (+)*
- C1 *Corners.*
- C2 *Uh, corners.*
- T *The corner of the street...*
- T *At the corner of what kind of street? (+)*

Collins (1981): canonical example of uptake

EXAMPLE 1: Incorporations of answer into question (+)

- T *Alright, what are they looking for?*
- C *Signals.*
- T *What signals? (+)*
- C1 *Red.*
- C2 *Red light and green.*
- C3 *Three signals.*
- T *Alright, traffic signals.*
- T *Where do you find those? (+)*
- C *On the street.*
- T *Alright, where on the street? (+)*
- C1 *Corners.*
- C2 *Uh, corners.*
- T *The corner of the street...*
- T *At the corner of what kind of street? (+)*

Lots of repetition!