



When Youth Enter The Chat: An Epistemic Shift in the Validation of LLM-Based Measures of Student Talk

Liliana
Santos-Deonizio
Stanford University

James Malamut
Stanford University

Ramón Martínez
Stanford University

Dorottya Demszky
Stanford University

Large Language Models (LLMs) are being used increasingly to measure aspects of student discourse (e.g. talk moves, collaboration, equity of voice) at scale. Typically, LLM-based measures of student talk use transcriptions of classroom conversations that only include verbal contributions, which de-contextualize student language. Common practices for validating these measures include comparing outputs against expert annotations by adults, using held out evaluation sets and F1 scores. We argue that these approaches are insufficient to ensure that such measures are meaningful and equitable for teaching and learning, particularly for racially and linguistically marginalized youth. In order to center the youth whose talk is being analyzed, re-contextualizing these classroom conversations and engaging youth in the research process is necessary. Sharing epistemic authority with youth, ultimately, centers their point of view and adds crucial nuance to the analysis of their talk that adult experts, researchers, and LLMs cannot provide. In a case study of multilingual youth in one 8th-grade math classroom, we address the epistemic exclusion of youth by employing multiple ethnographically-oriented methods to re-contextualize student conversations and center youth as epistemic authorities in conversation with researchers and LLMs. We conducted participant observations, interviews, focus groups, and member checks with four focal students. Findings reveal that there were misalignments between students' interpretations of their own math talk experiences and the LLM-based measures of their talk. Students contested both the LLM classifications and the coding scheme used to measure their talk, highlighting the need for youth to be involved in the epistemic process of producing knowledge about their experiences.

VERSION: September 2026

Suggested citation: Santos-Deonizio, Liliana, James Malamut, Ramón Martínez, and Dorottya Demszky. (2026). When Youth Enter The Chat: An Epistemic Shift in the Validation of LLM-Based Measures of Student Talk. (EdWorkingPaper: 26-1564). Retrieved from Annenberg Institute at Brown University: <https://doi.org/10.26300/fgpa-9447>

When Youth Enter The Chat: An Epistemic Shift in the Validation of LLM-Based Measures of Student Talk

LILIANA SANTOS-DEONIZIO, Stanford University, United States

JAMES MALAMUT, Stanford University, United States

RAMÓN ANTONIO MARTÍNEZ, Stanford University, United States

DOROTTYA DEMSZKY, Stanford University, United States

Large Language Models (LLMs) are being used increasingly to measure aspects of student discourse (e.g. talk moves, collaboration, equity of voice) at scale. Typically, LLM-based measures of student talk use transcriptions of classroom conversations that only include verbal contributions, which de-contextualize student language. Common practices for validating these measures include comparing outputs against expert annotations by adults, using held out evaluation sets and F1 scores. We argue that these approaches are insufficient to ensure that such measures are meaningful and equitable for teaching and learning, particularly for racially and linguistically marginalized youth. In order to center the youth whose talk is being analyzed, re-contextualizing these classroom conversations and engaging youth in the research process is necessary. Sharing epistemic authority with youth, ultimately, centers their point of view and adds crucial nuance to the analysis of their talk that adult experts, researchers, and LLMs cannot provide. In a case study of multilingual youth in one 8th-grade math classroom, we address the epistemic exclusion of youth by employing multiple ethnographically-oriented methods to re-contextualize student conversations and center youth as epistemic authorities in conversation with researchers and LLMs. We conducted participant observations, interviews, focus groups, and member checks with four focal students. Findings reveal that there were misalignments between students' interpretations of their own math talk experiences and the LLM-based measures of their talk. Students contested both the LLM classifications and the coding scheme used to measure their talk, highlighting the need for youth to be involved in the epistemic process of producing knowledge about their experiences.

CCS Concepts: • **Human-centered computing**; • **Applied computing** → **Computer-assisted instruction**; **Collaborative learning**;

Additional Key Words and Phrases: education, student discourse, large language models, LLMs, member-checking, interviews

ACM Reference Format:

Liliana Santos-Deonizio, James Malamut, Ramón Antonio Martínez, and Dorottya Demszky. 2027. When Youth Enter The Chat: An Epistemic Shift in the Validation of LLM-Based Measures of Student Talk. In *Proceedings of TBD (CSCW '27)*. ACM, New York, NY, USA, 18 pages. <https://doi.org/XXXXXXX.XXXXXXX>

1 Introduction

"It's why we don't do math. Math pisses me off." Ximena, an 8th-grade Mexican American girl, said to her friend and classmate, Destiny, an 8th-grade Dominican and Haitian American girl. They were working on a problem and

Authors' Contact Information: Liliana Santos-Deonizio, Stanford University, Stanford, CA, United States, lilianas@stanford.edu; James Malamut, Stanford University, Stanford, CA, United States, jmalamut@stanford.edu; Ramón Antonio Martínez, Stanford University, Stanford, CA, United States, ramon.martinez@stanford.edu; Dorottya Demszky, Stanford University, Stanford, CA, United States, ddemszky@stanford.edu.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2027 Copyright held by the owner/author(s). Publication rights licensed to ACM.

Manuscript submitted to ACM

Table 1. Excerpt 1 from Classroom Conversation between Ximena and Destiny

Speaker	Line	Dialogue
Ximena	1	"But I subtract th-, that, that to that."
Destiny	2	"No, you minus on these sides, right?"
Ximena	3	"Yeah."
Destiny	4	"That's how we learned yesterday, and then do the same thing on these numbers once you figure out this answer."
Ximena	5	"Oh, oh it's ninety one. That's why."
	6	"This is ninety one."
	7	"Ninety one divided by nine."
	8	"It's why we don't do math."
	9	"Math pisses me off."

Ximena had asked Destiny for help. She expressed her frustration with the process of solving the problem. All the while, two small microphones captured their words and an iPad filmed while they worked. Later, this conversation would be transcribed (automatic transcription reviewed and edited by human transcribers) and an LLM would be used to annotate it for mathematic talk moves. These two sentences from Ximena would be coded as "Off-task," meaning they did not relate to the math task at hand or the process of solving the problem. This definition of Off-task had come from educational research about math and collaboration, and it distinguished other academic talk moves from Off-task talk. A few weeks later, the first author sat with Ximena to watch a clip of the exchange. When asked whether she agreed that those lines were off-task, Ximena pushed back: "Technically I wasn't really off task, I was talking about math but not in a good way... So, Chat PT [sic] needs to get his life together."

In this paper we explore the following question: **what can be learned by involving students in the epistemological process of validating LLM-based measures of student talk?** Validation itself is a practice of knowledge production, and it raises the question of who holds *epistemic authority* over how student language is interpreted. Students are experts with respect to their own experiences and bringing them into the validation process can deepen the understanding of their talk in ways that teachers, researchers, and LLMs cannot achieve alone. We argue for a two-part *epistemic shift*. First, language should not be treated solely as a text-based representation, but as a situated, social, cultural, and embodied experience — which requires *re-contextualizing* the transcripts that LLM-based measures rely on. Second, youth should be brought into the validation process as epistemic authorities on their own talk, not just as the subject of expert annotation. Figure 1 maps this shift from excluding to including students in the validation process.

We illustrate this shift through a case study of multilingual youth¹ in one 8th-grade math classroom, drawing on participant observations, interviews, focus groups, and member checks with four focal students. Re-contextualizing the transcripts surfaces dimensions of participation — relationships, classroom routines, physical space, gesture, prosody, language ideologies — that text-only representations obscure and that shape potential interpretations of student talk. We also show how sharing epistemic authority with the students reveals misalignments between how the LLM classified their talk and how they themselves understood it. These misalignments go beyond errors that can be fixed with prompt-engineering and instead require modifications of the coding scheme that delineates what counts as math talk.

¹We use the term multilingual youth expansively to refer to youth that speak more than one language or language variety, i.e. African American English or Spanish-English varieties [24].

Together, the two shifts point toward a model of validation in which youth are not only the source of the data, but also participants in deciding what the data means.

The remainder of the paper illustrates these two shifts. We start in Section 2 by reviewing related work on computational measures of classroom discourse, ethnographically oriented methods and participatory approaches to AI evaluation. Next, in Section 3, we provide a brief overview of the LLM-based measures for context. In Section 4, we describe the ethnographic methods we used for re-contextualization and member checking. Section 5 presents what each surfaces, including the context the transcripts miss and classifications the students themselves contest.

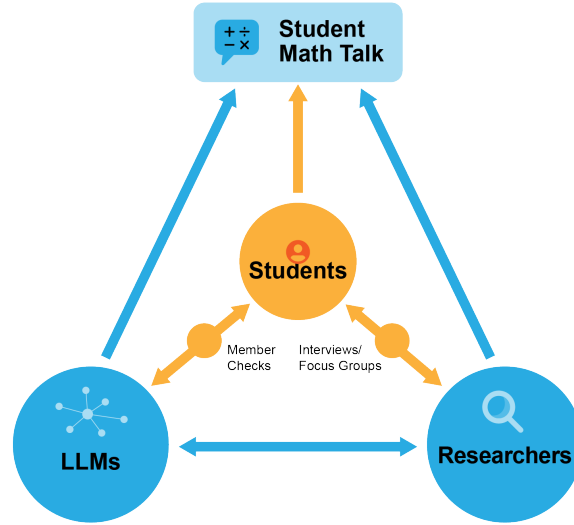


Fig. 1. When Youth Enter the Chat: Before (blue) and After (yellow) Involving Youth in the Knowledge Production Process

2 Related Work

2.1 Computational Measures of Classroom Discourse

Recent work in Natural Language Processing (NLP) and learning analytics has used computational methods to analyze student and teacher discourse, including talk moves, collaboration, and patterns of participation in classroom and collaborative problem-solving settings [1, 3, 6, 28, 34, 36–38]. Prior studies have used methods such as Linguistic Inquiry and Word Count (LIWC) [34, 36], BERT [36, 37], and other classifiers to identify discourse features and relate them to learning outcomes or collaborative behaviors. These studies have advanced the ability to analyze classroom talk at scale and have typically validated models using expert-annotated transcripts, held-out test sets, and performance metrics such as F1 scores, AUROC, or agreement with human experts. While these approaches are useful for assessing how well models reproduce expert labels, they generally treat transcripts as sufficient representations of classroom interaction and rarely incorporate validation from the perspectives of the students whose discourse is being measured. As a result, current validation practices say little about whether model outputs align with how students themselves understand their participation or with the broader contexts in which that talk occurs.

2.2 Ethnographically-Oriented Methods and Context in Multilingual Math Classrooms.

What ethnographic research, in particular, has long shown is that classroom discourse is shaped by social relationships, institutional norms, identity, and embodied interaction. Studies focused on multilingual youth translanguaging practices [14, 35, 46], and language ideologies [11, 12, 23] demonstrate that meaning cannot be understood from words alone, but must be interpreted in relation to context, audience, language ideologies and the racialized experiences of interlocutors. The context of multilingual mathematics classrooms makes these limitations especially important. In the United States, the language practices of multilingual youth have historically been framed through a deficit lens [17], positioning their linguistic repertoires as requiring remediation rather than serving as resources for learning [16, 39]. Research in mathematics education has shown that talk and collaboration are central to learning, [8], while scholarship on identity and language in math classrooms has demonstrated that participation is shaped by broader histories of race, gender, language, and belonging [10, 15, 29, 32, 33]. In classrooms where standardized English is often treated as the norm [5, 19], multilingual students may draw on broad linguistic repertoires across settings [14], and their participation may not be fully legible through transcript-based measures alone. This makes multilingual math classrooms a particularly revealing site for examining what LLM-based discourse measures capture, what they miss, and what is at stake if youth are not involved in the validation process.

2.3 Participatory Approaches to AI Evaluation

We engage youth through participatory approaches similar to work done in AI and HCI evaluation. Related work in HCI and responsible AI has argued that evaluating AI systems requires attention to the perspectives of the communities most affected by them [18, 40, 41, 43]. Participatory approaches, including audits [40], co-design and co-creation [30], teacher and youth centered evaluation [18, 43], have been used to surface harms, biases, and mismatches that are not visible through standardized metrics alone. Additionally, youth have been engaged through Youth Participatory Action Research to reflect on technology and its impact on students [13], and college students have been engaged in participatory design processes [31]. Our work builds on the evaluation strand of this literature: we ask what happens when youth are brought into the validation of LLM-based measures that researchers have already built — a stage in the AI lifecycle where their perspectives remain especially rare.

2.4 Epistemic Authority and Exclusion

Finally, we draw on theoretical frameworks of epistemic agency, authority and exclusion to analyze what is missed when students are not involved in the validation process. *Epistemic agency* is the ability for youth to participate in the disciplinary practices and intellectual activities of knowledge production [27]. *Epistemic authority* relates to who has the power to engage in the intellectual process of knowledge building, which traditionally has been held by teachers in classrooms, and scholars and their tools in research. *Epistemic exclusion* relates to who is excluded from the knowledge-building process [4]. Scholars in math and science education have used these constructs to examine power and participation in classroom collaboration [20, 21], showing that epistemic agency is dynamic and interactionally constituted, and that authority does not redistribute automatically in collaborative or student-centered settings. Additionally, scholars have applied this framework to show how technology can also mediate epistemic injustice [2].

We extend this body of work in two ways. First, while these frameworks have largely been applied to classroom interactions, we apply them to the research process itself. We ask how youth are epistemically excluded not only when

learning, but also when their learning is being measured, interpreted, and represented. Second, building on work that positions youth as philosophers of technology [44], we engage youth as experts of their own experiences who can interrogate technological artifacts, in this case the LLM classifications of their talk.

3 LLM-Based Measures of Student Talk

The context of our study is a two-year project, one goal of which has been to develop scalable measures of multilingual students' math language practices. Data from the first year of the project (2024–2025 school year) was used to develop and validate preliminary LLM-based measures via prompt-engineering. We treat these LLM measures as central artifacts in the ethnographic work presented in this paper, both for re-contextualization and member-checking. The ethnographic work was conducted during the 2025–2026 school year. Here we provide a short overview of these LLM-based measures and their initial validation.

In the first project year, the research team partnered with two large urban school districts in the West Coast, both serving a high number of multilingual students, to collect transcripts from middle school math classrooms. During recordings, students worked on math tasks with peers, with at least one student per group reporting Spanish as a home language. Recordings were transcribed by automated speech recognition software with human cleanup and review prior to annotation.

Codebook. We used a multi-theoretic coding scheme that covered topic and talk move codes from collaborative problem solving (CPS) and mathematics learning frameworks [7, 9, 37, 38, 45]. LLMs were used to code for collaborative and academic talk moves, including the topic codes—Offtask, Recording, Competent, Understanding, Tool—and the talk move codes—Question, Apology, Claim, Reasoning, Next Step, Monitor, Disagree, Agree, Compare, Add-On, and Revoice. The codebook definitions were developed with human experts who also helped create the validation set. In this paper, we focus on the seven codes with the highest frequency in our analytic sample: Off-task, Understanding, Recording, Question, Claim, Next-step, Disagree. This analytic codebook is shown in Table 2.

Validation set. To develop a validation set for the LLM models, five transcripts totaling 41,196 utterances were annotated by human experts—current and former teachers and educators, including members of the research team. Specifically, the annotators were one Latina woman (former teacher and first author of this paper), one Asian woman (current curriculum specialist), three white women (a current teacher, an instructional coach, and a teacher educator), and one white man (former teacher and second author of this paper). Utterances were split at the sentence level that served as the unit of analysis for the LLM-based talk measures. Each utterance was additionally tagged for whether the utterance was in English, Spanish or both. We evaluated several LLMs against this validation set (Claude 3.7, GPT 5.0, GPT 5.1) and selected GPT-5.1, which performed best based on F1 scores.

LLM prediction. We used GPT-5.1 to annotate the transcripts chosen for the student member checks. The model was prompted with the code definitions and three positive examples and three non-examples per code (chosen by human experts), and instructed to assign each utterance a binary (0/1) label for each code. The analytic sample consisted of two transcripts from the focal classroom (2025–2026).

LLM performance. Table 3 reports GPT-5.1's validation performance (F1, precision, recall) on the seven most frequent codes in the analytic sample. We also conducted an error analysis, which surfaced misclassification patterns that we attempted to address by revising code definitions in the prompts and resolving inconsistencies in the human-annotated validation set. After updating definitions, we re-ran the GPT 5.1 on the validation set. The F1 scores increased for each

Table 2. Code Definitions and Examples

Code Type	Code	Definition	Example
Topic	Off-Task	Talk not pertaining to the assigned task and how to engage with it.	“I’m just gonna go back there and play Fortnite. I somehow know how to play Fortnite.”
	Understanding	Discussing group or individual’s understanding of a mathematical idea, question, or problem. Discussing whether or not or to the extent to which the group or individual feels like they understand or do not understand.	“Okay, ya lo entiendo.” (Okay, I understand it now)
	Recording	Talk pertaining to the recording equipment or the university researchers using it to observe them.	“Estoy diciendo cosas y tengo el micrófono.” (I am saying things and I have the microphone.)
Talk Move	Question	Asks a question	“Why?”
	Claim	A mathematical statement of fact or conjecture that can be proven false or not.	“Veinte por cuarenta es ochenta.” (Twenty times forty is eighty.)
	Next-Step	Suggests the next step for the group or an individual to take.	“Try and see if eight can go into twenty-four on this.”
	Disagree	Expresses disagreement with another students’ mathematical reasoning expressed in a prior turn.	“No, you can’t put the five because you never know how many friends there is.”

feature, as shown in Table 3, which reports the F1 scores before and after updating definitions and human annotations. Two codes saw the largest change in F1 scores due to updating definitions, Claim (+0.104), and updating inconsistencies in human annotation, Question (+0.23). The remaining codes used for this analytic sample saw a smaller increase in F1 performance (+0.021 - +0.061).

Table 3. GPT-5.1 Scores by Feature Before and After Definition Update

Feature	Old F1	New F1	New Precision	New Recall
Off-Task	0.764	0.817	0.734	0.921
Understanding	0.639	0.700	0.605	0.830
Recording	0.661	0.682	0.564	0.863
Question	0.717	0.947	0.955	0.938
Claim	0.653	0.757	0.894	0.656
Next-Step	0.586	0.631	0.515	0.815
Disagree	0.429	0.473	0.481	0.464

The process of creating the validation set surfaced limitations that motivate the rest of the paper. The annotation team found that disambiguating certain talk moves required context beyond text—intonation, students’ intent, what they were pointing to or referring to. This was difficult even for human experts, and more so for LLMs, which lacked the annotators’ background as educators and their experience listening to youth talk. In early runs of the LLM annotation, the models struggled to reach a 0.7 F1 score on most features. This gap was exacerbated by certain features being infrequent and more conceptual, requiring inference about intent rather than recognition of surface features. After

adjusting definitions, the F1 scores improved but for some features—most notably Next-Step and Disagree (Table 3)—they remained low. This validation process motivated the need to find other ways to validate LLM-based measures of student talk and to re-contextualize the transcripts on which they depend.

4 Ethnographically-Oriented Methods

The epistemic shift in the validation of the LLM-based measures of student talk consists of two shifts, which include re-contextualizing student talk and bringing youth into the validation process. We employed a set of ethnographically-oriented methods as part of these shifts: participant observations, student interviews, focus groups, and member checks.

Focal classroom and participants. For the 2025–2026 case study, we selected one 8th-grade math classroom from a large urban school district on the West Coast. The classroom was selected from participating classrooms, as the school primarily serves multilingual youth from African-American and Latinx families. The classroom teacher had participated in the broader study during the prior year, but her students were new to it. From the students who consented to be recorded, we selected four focal students because they worked in the same pairs frequently, which allowed us to discuss their collaborative dynamics: Ximena and Destiny, two girls (Ximena is Latina and a designated English Learner; Destiny is Latina and Haitian), and Diego and Drake, two boys (Diego is Latino; Drake is African-American). Across the school year, I, the first author, conducted seven participant observations, four interviews, two focus groups, and four member checks with these students. We also selected two transcripts from the focal classroom as the analytic sample for the case study; the same excerpts were used in the interviews, focus groups, and member checks. The transcripts were the two most recent transcripts from lessons the students had done, so that the conversations would be more recent in the students’ memories for the participant retrospection, interviews, and focus groups.

Positionality. I, the first author, am an immigrant Latina woman who is multilingual and grew up in large urban city on the West Coast. I also previously taught as a K-5 classroom teacher. This background allowed me to speak in both Spanish and English with the youth, and shaped how I approached relationship-building with the teacher and students.

Participant observations. I sat with various student groups during observations to get to know the youth and notice how their group work unfolded. After each visit, I wrote field notes capturing the classroom environment, the math activities, and my own reflections on the interactions I observed. These initial participant observations supported relationship-building between the teacher, the youth, and me, which helped create a more comfortable environment in the interviews and focus groups.

Student interviews. In one-on-one interviews, students participated in two activities: a language mapping activity and participant retrospection. Building on Martínez and Mejía, [25], the language mapping activity asked youth to reflect on the spaces and places they traverse daily and weekly, the people they interact with, and how they use language across those audiences and interlocutors. The purpose of this activity was to build a more expansive picture of students’ linguistic repertoires both within and beyond math class.

Participant retrospection and member checks. During retrospection, youth revisited an excerpt of one of their math class conversations and reflected on what they were doing, thinking, and feeling at the time. They were then shown the LLM’s annotations of that same excerpt and asked whether they agreed or disagreed with how the model had classified



Fig. 2. Example of Language Mapping Activity

their talk, what they would change, and what they felt the model had missed. The full protocol is included in Table 6 in the Appendix.

Focus groups. Focus groups paired each focal student with a peer with whom they had worked before. They revisited the same transcript excerpts used in the interviews and reflected on their group dynamics, collaboration, and what it was like to talk with one another in math class. Finally, students were asked to reflect on their group dynamics and collaboration with their classmates.

Analysis. From the participant observations, I thematically grouped field notes into contextual factors that LLM-based measures miss (e.g., environmental, social, multimodal, and relational dimensions). These contextual factors, which informed how students participated, included information that was not supplied to the LLMs about the multimodal context of the classrooms and would not be represented in the transcription alone. For the member checks, I tabulated each instance where a student disagreed with an LLM classification (Table 4) and sorted the disagreements into two categories: disagreement with the LLM’s classification of a given utterance, or disagreement with the coding scheme. These groupings surface misalignments between how students saw their talk and how the LLMs and coding scheme classified their talk.

5 Results

5.1 Re-Contextualization: Beyond Text-Only Representations of Language

When classroom interactions are reduced to transcripts of verbal contributions, key dimensions of participation are lost. The physical and socio-cultural environment, multimodal communication (silence, gesture, tone), teacher and student characteristics and relationships, the curriculum and lesson structure, and the broader language ecology of the classroom are missing from transcripts. Due to this, LLM-based measures that rely on such transcripts alone operate on an incomplete representation of classroom activity.

By using ethnographically-oriented methods, we recover much of what transcripts lose by treating student talk as a situated, embodied, social, and cultural experience. In the sections that follow, we describe five dimensions of context surfaced through participant observations, interviews, and focus groups, each illustrated with examples from the focal classroom. These contextual factors allow us to (re)interpret the transcripts and create more holistic representations of student math talk and participation, and ultimately, center the student experience.

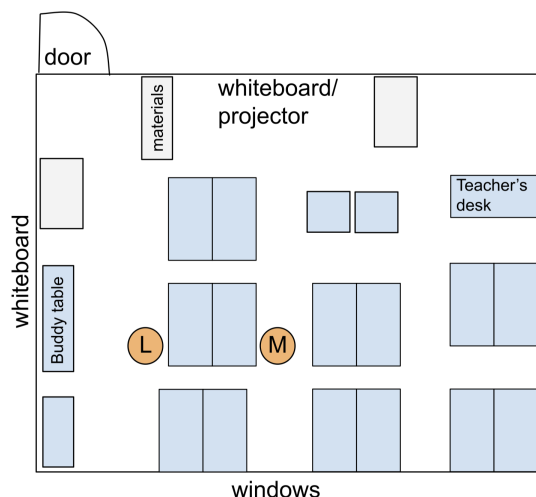


Fig. 3. Layout of Ms. Richer's classroom mapping where Luna and Malik sat during a class session.

5.1.1 School and Classroom Environment. The focal school is situated in a large urban city on the West Coast in a residential neighborhood with mostly single story homes. Down the street from the school, there were businesses with signs written in Spanish and English. In the afternoons, multiple paleteros (ice cream sellers) and snack carts lined up around the front left corner of the school to sell paletas (popsicles), fruit or chicharrones de harina (wheat pinwheels) with lime and hot sauce, chips and sweets. Ms. Richer's 8th-grade math classroom was on the first floor of a two-story building. At the time that data was collected for this study, a decoration of La Llorona (The Weeping Woman) was visible on the classroom door. Inside, the classroom was often dimly lit with shades pulled down and light streaming in from a narrow gap below the shades. Small blue string lights highlighted the contour of the whiteboard at the front of the classroom. Students were assigned seats in groups of about four per table with the only exception being the two single seat desks by the front of the classroom. The classroom was tight with only about one to two feet between any tables—Ms. Richer measured the spacing to be just right for her to squeeze through between tables. All along the walls there were posters and signs including math concepts, classroom expectations, voice level descriptors, and hand signals for the bathroom, questions, and other requests students might have. In the middle of the whiteboard at the front of the class was an agenda for the day. To its side there were class point and seating chart posters.

The physical space in the room impacted how students worked together. In one observation, I sat with two students, Malik, an African-American boy, and Luna, a Latina girl, who were sitting across from each other. The group tables were actually two smaller rectangular tables set up next to each other to form a square (see Figure 3). The distance between the two tables created a wide enough gap that if two students sat across from each other it was harder to hear

with other conversations going on, and it made talking across the two tables less inviting. Instead of talking to one another during the pair-share time, Malik and Luna continued working in silence independently.

5.1.2 Classroom Curriculum and Lesson Structure. Classroom routines further structure when and how students are expected to talk. The focal classroom had structured time for when to talk and when not to talk. Ms. Richer followed a similar agenda during observations starting with a warm-up exercise, then an activity covering the current material the class was learning with designated “mild,” “medium,” and “spicy” challenge levels, and closing with an exit ticket. A typical routine included Ms. Richer starting the class with instructions and 3-5 minutes of independent work at “voice level zero,” meaning there should be no talking. Then, Ms. Richer asked students to turn to their group or partner and share for 1-2 minutes at a normal talking voice level. Finally, Ms. Richer asked students to share-out in a whole class discussion for an additional 2-3 minutes. Ms. Richer typically asked students to share with a partner or share out with the class in between each activity, with lo-fi music playing in the background and a timer counting down for most activity sections. During my observations, the majority of students were quiet during the class, both in group work time and during the whole class share-out. Even during partner talk time or class share-outs, a majority of youth had their heads down looking at their worksheet continuing to work independently. Some of this was structural: parts of the curriculum directed students to complete activities individually on computers, which did not invite collaboration. For this project, we focused on student-to-student conversation, so the whole class share-outs were not part of the final transcript.

Knowing the lesson structure matters for interpretation. Students were constrained by how much time and how many opportunities they had to talk with peers, and those constraints shaped the distribution of talk in ways that are invisible from the transcripts alone. This makes it difficult to interpret measures of participation without the context of the lesson plan.

5.1.3 Multimodal Communication. Two kinds of multimodal information are lost in the transcripts: non-verbal demonstrations of mathematical thinking and communication, and the communicative work done by tone, prosody, and silence.

During class, students engaged in many non-verbal actions that demonstrate their learning and mathematical thinking, which was not captured by transcribed speech. As mentioned above, during my observations, the majority of the class worked independently and quietly on math activities both during the independent work time and the pair-share time. Despite not engaging in verbal communication, most students continued to write equations and solve problems on their handouts.

Multimodal data about student conversations can also reflect individual traits and dispositions that shape participation in ways that coding schemes and LLM-based measures do not capture. For example, Carlos, a Latino boy, rarely spoke in whole-class or group conversations but he worked diligently and quietly on his own. Drake, an African-American boy, was charismatic and outgoing, frequently contributing both in groups and during whole class share-outs. Interpreted only through what an LLM has been told counts as academic talk via the coding scheme, Carlos and Drake look like opposite ends of a participation scale. This difference between them could easily be mischaracterized as reflecting differences in engagement, whereas it may, in fact, be reflecting differences in their preferred *mode* of engagement. Coding schemes built around spoken academic talk risk mischaracterizing those differences as gaps in ability or attention.

Multimodal data about the classroom conversations also helps when interpreting what students meant in the context of their conversations. Tone, silence, prosody, and body movements all help make meaning and these communicative

dimensions are lost in transcription. Within transcribed speech, attending to tone or prosody can disambiguate a student's intent. For example, when "Yeah" is said in a transcript, it can be a signal of agreement or a filler word, and the disambiguating signal is in the voice, not the transcript. Gesture and posture similarly shape what an utterance means in context yet remain unavailable to text-based LLM measures.

5.1.4 Language Ideologies and Norms. Institutional language ideologies and norms also shape student discourse. Although many students in the classroom spoke Spanish as a home language, instruction was conducted in English, and the use of other languages was not explicitly encouraged. As a result, transcripts may not capture the broader linguistic repertoires students draw on in other contexts or outside of recording times. Ms. Richer is a white woman in her mid 30s who had been teaching math for 10 years at the time of this study. She speaks English and shared that she understands "some" Spanish. Ms. Richer explained that while the school did not explicitly prohibit using other named languages like Spanish, they did not explicitly encourage it in the classroom either. Most class discussion used English, with some colloquial language from Ms. Richer. The class was majority Latine and bilingual in Spanish and English; most of those students were Reclassified, with a handful designated English Learners. The students designated English-Only were primarily African-American. This was reflective of the school's broader demographics, where the majority of students were Latine or African-American.

Students' use of home languages was discussed in student interviews and the language mapping activity. For example, Ximena, who is a designated English Learner, shared that even though she speaks Spanish, she would only use it in certain contexts at school, such as during Physical Education class, where she helped translate for Spanish-speaking peers. However, in math class, Ximena shared that she did not use Spanish. During my observations, I only heard Ximena use Spanish a handful of times. Ximena's classmate and friend, Destiny, also shared that she knew some Spanish and Haitian Creole, but did not use these at school. Both students also used features of African-American English though neither named this language variety in their language mapping activity. When I asked the students about whether they thought they used African-American English, Ximena looked over at Destiny and said, "No, I thought I was just talkin' normal English." Destiny smiled and added, "Right, like I never noticed anything different." Looking only at LLM-based measures of student talk based on transcripts would fail to capture the expansiveness of students' linguistic repertoires, the context for why they may change the way they speak with different audiences and in different places, and the way their talk is informed and constrained by language ideologies.

5.1.5 Relationships. In focus groups, students emphasized that interpersonal relationships affected how comfortable they felt talking to their peers in class. In response to a question about what it was like to be paired with Ximena, Destiny shared: "I feel like sitting next to her, I just feel comfortable." Elaborating on whether it was important to feel comfortable with her peer, Destiny said, "Yes because I'm a really anxious person, like I have lots of anxiety and like social anxiety, and so like being around people that I'm comfortable with instead of sitting next to somebody I don't know is a lot easier for me." These relationships are often not explicit in recordings and are not captured in LLM-based coding schemes. Students are aware that their experience collaborating with peers may be different based on preexisting relationships.

Students described their relationship with their teacher as similarly important. In student interviews and focus groups, students expressed feeling more comfortable with Ms. Richer than with other teachers at their school. They reported that this impacted how they talked in the classroom, including their willingness to engage in off-task talk. Ximena shared in an interview that Ms. Richer was "more calmer and that's why everybody's her favorite teacher". Ms. Richer's relationship with her students and her disciplinary choices made Ximena feel more comfortable talking in her

class and made Ms. Richer her favorite teacher. For Ximena and Destiny, that comfort shaped what they felt able to say. Such dynamics are not captured in LLM-based coding schemes, but they help explain variation in participation that transcript-based measures would otherwise obscure.

5.2 "ChatPT [sic] needs to get his life together": Students Engaging with LLM Based Talk Measures

Re-contextualizing transcripts goes some distance toward centering student experience, but it does not fully bring students into the validation process. Students are experts on their own talk and bringing them into the validation process can provide nuance, depth, and contrasting viewpoints that adult annotators, researchers, and LLMs cannot provide alone. Student member checks revealed that discrepancies between LLM classifications and student interpretations were not limited to isolated errors. They reflected deeper misalignments between how coding schemes define academic talk and how students understand their own participation. While students generally agreed with the LLM classifications, the moments of contestation—and the reasons they gave for them—highlighted misalignments with the LLMs' classification of a given utterance, and with the coding scheme itself. Table 4 shows the highest-frequency disagreements during member checks. Students were especially vocal about the Claim and Off-task codes and pushed back on these labels by foregrounding the intentions behind their utterances.

Table 4. Disagreement Rates During Student Member Checks for Analytic Codes

Student	Off-Task	Understanding	Recording	Question	Claim	Next-Step	Disagree
Diego	0/2	0/6	–	0/10	5/12	0/1	1/1
Drake	1/5	1/4	–	0/6	0/6	1/1	0/1
Destiny	0/8	0/5	1/2	0/8	1/4	1/3	0/2
Ximena	5/8	0/4	–	1/4	3/17	0/1	–

Note. Cells show the disagreement counts divided by the total count for each code. Dashes indicate that the code did not occur for that student in the analytic excerpt.

Students argued for more expansive definitions of what counted as math talk. Ximena pushed back on the Off-task label three times because these were moments when she explicitly said "math" and related to her math learning experiences. In the first excerpt (Table 1), where Ximena and Destiny discussed how to solve a problem, the LLM annotated lines 8-9 ("It's why we don't do math. Math pisses me off.") as off-task. From Ximena's perspective, expressing frustration was part of working on the math task. The pattern repeated in a second excerpt (Table 5), where Destiny checks her answer with Ximena and Ximena replies, "I don't know see honey bunch, you're asking the wrong person over here. Let's be for real. you're asking the wrong person." The LLM marked both lines as off-task. Ximena argued they were not, since she was "still talking about the work." A third instance occurred mid-conversation about a math problem, when Ximena said, "[Teacher], I think I'm actually smart. I don't even know why I'm in eighth grade. I'm actually supposed to be in high school." Her position was consistent: "I think it was on task because I was talking about math." In all three of these instances, Ximena engaged actively with math tasks and made comments about her personal math experience. Ximena was quick to say, "ChatPT [sic] needs to get his life together," in response to how the LLM was annotating her talk. Not only was Ximena offering a more expansive idea of what could be considered math talk, she was also pointing out the LLM's role in this process. By engaging Ximena in the conversation about her own talk, the epistemic authority shifted to include her as a part of the knowledge production process.

Students also added more clarity around their intentions and what was missed by the coding scheme used for LLM classification. Ximena and Diego had similar utterances that were coded by the LLM as Claim, but they disagreed with

Table 5. Excerpt 2 from Classroom Conversation between Ximena and Destiny

Speaker	Line	Dialogue
Destiny	1	"Divided by six."
	2	"Negative one. Right? Am I, am I dumb or am I dumb?"
Ximena	3	"I don't know see honey bunch, you're asking the wrong person over here."
	4	"Let's be for real. you're asking the wrong person."

this code. In a recording, Ximena had said, "Ten, twenty, thirty, forty," and reflecting on this she thought this should have been marked as showing reasoning: "I say reasoning...because I was trying to figure out why it's this number, so I was counting." Diego had an instance where he was also counting out loud: "Sixty. Seventy." and similarly disagreed with the Claim label: "Nah I was just like counting...I think it's different because a claim.. you're saying that's the answer or something." For both students, the intention behind their utterance was different from a claim. As Diego explained, a claim involves asserting a final answer, whereas counting reflects an intermediate step. This distinction, while captured by the coding scheme, was not replicated by the LLM-based measures, which classify utterances based on the raw data rather than intended function.

Reflecting on their original intentions in conversations, students described possible new codes for their talk that were missing from the coding scheme. Diego had another instance where he was talking to Drake about a possible solution and said: "I feel like it's twenty-eight for some reason." The LLM classified this as a Claim and Disagree. Diego clarified, "No, it was just like a guess, I was just guessing I wasn't claiming that it was.. wasn't denying it wasn't also 28." In this moment, Diego was able to make a guess about a possible solution and hold multiple possibilities in his head. When I asked how he would categorize this instead, Diego asked, "There's no guess?" There were no codes in the coding scheme for brainstorming or guessing, so this function was missed.

Member checks reveal that coding schemes embed assumptions about what counts as meaningful participation, and that those assumptions do not always align with how students themselves understand their talk. Centering student interpretations can both help correct misalignments and surface categories the scheme failed to anticipate.

6 Discussion

Our findings highlight two epistemic shifts in the design of LLM-based measures of student talk. The first concerns the data: by reducing rich, embodied classroom interactions to text and then to discrete labels, LLM-based systems omit relational, spatial, socio-cultural, and multimodal dimensions of participation by design. These omissions are not errors that can be corrected through improved prompting or larger datasets; they are structural limitations of applying text-based classification to complex social interaction [42, 47]. Because they take conversations out of context and exclude student input, these omissions also de-center the youth whose talk is being measured.

The second concerns validation. Standard validation practices ask whether a model reproduces expert labels, but validation is itself a knowledge-production practice: it determines whose interpretation of the data counts. By involving youth in validation, we shift epistemic authority over interpretation from researchers and models to students themselves (Figure 1). This re-centering of youth allows us to better understand not only how students see their own math talk, but also whether these talk measures are meaningful to them. Asking for the perspective of students allows us to validate if the talk measures are accurate to their original intentions and, beyond this, gives insight as to what youth see as part of the math learning experience. The disagreements the students in this study raised were not simply classification errors,

but alternative interpretations of what their talk meant. In other words, students were not simply interested in using the available codes provided by the research team to improve model performance, they wanted to change the interpretations of existing codes, and add codes they felt were missing. When Ximena contested her talk being labeled as off-task, for instance, she was asserting that her math conversations include social and affective dimensions—an assertion that aligns with prior work showing that off-task talk can serve a range of social and communicative purposes [22, 26]. Students also added nuance to their original intentions, which highlighted misalignments with the coding scheme. These misalignments might cause LLMs to misattribute talk moves and miss significant contributions. Incorporating student perspectives is, therefore, not just a methodological choice, but a necessary step toward creating more equitable and representative forms of AI in education.

Taken together, these findings argue for participatory validation as a core component of responsible AI in education. Tools and research that measure student discourse can shape how participation, collaboration, and learning are understood and evaluated. Without incorporating the perspectives of the communities being measured, these outcomes risk reinforcing narrow or misaligned definitions of meaningful participation, particularly for multilingual and marginalized students. Centering youth as epistemic authorities additionally improves the creation and interpretation of measures by grounding these measures in the lived experiences of the youth. We recommend that scholars using LLM-based discourse measures triangulate data from multiple methods and seek input from participating communities to validate and improve measures. This is especially important for combating the biases and harms that LLMs can reproduce and that disproportionately affect marginalized youth.

7 Limitations and Ethical Considerations

Several limitations bound the claims of this paper. First, our framing intentionally treats student learning as more than spoken contributions, and we address this through methodological triangulation. But the same methods that allow us to re-contextualize a small number of conversations also constrain how many we can analyze. Part of the work in the focal classroom included building relationships slowly and intentionally over the course of months, which limits the total number of students and classrooms under consideration. The tradeoff is that we cannot generalize from one classroom and four students. A larger or more varied sample would likely surface participation patterns and disagreements we did not observe.

Second, our participatory engagement was scoped to the validation phase. Youth provided feedback on the LLM’s outputs and on the coding scheme, but they were not involved in designing the coding scheme or the LLM prompts in the first place. A fuller participatory approach, in which youth shape the underlying definitions of math talk before classification begins, would likely surface further mismatches that our member checks missed.

Finally, the use of LLM-based measures raises concerns not only about the bias and harms that LLMs have been shown to cause for users, but also about the environment and communities near data centers. This is especially concerning for the larger models that cannot be locally run. Our team conducted two annotation runs with GPT 5.0 via API calls, one for the full set of features, one for a reduced set of features with updated definitions, and one for the two analytic excerpts. While our validation set and analytic sample is small, the impact of our research must also be taken into account as we look towards future work. As we center equity in our work, we must also account for the impact of our research on communities beyond our focal classroom, including those most affected by the infrastructure that LLMs depend on.

8 Future work

The limitations of this work point to natural extensions: a broader sample, fuller co-design and participatory approaches with youth, and smaller, locally-runable models. Beyond these, our findings open several further directions for inquiry. First, extending this work beyond mathematics classrooms, into after-school programs, extended-day spaces, or other subject-area classrooms, would help us understand how youth language practices and the categories that describe them shift across settings. Second, future research could examine what happens downstream of measure development. When teachers, students, or administrators see LLM-based discourse outputs, how do those outputs shape participation, instruction, and student-teacher relationships? Finally, the participatory validation approach we describe could be applied to other AI tools used in educational settings, such as automated feedback systems, assessment tools, or behavior detection, to test whether the misalignments we observed are specific to discourse measurement or reflect broader gaps between AI-generated representations and the experiences of the students they describe.

9 Conclusion

This paper argues that standard validation practices for LLM-based measures of classroom discourse (i.e., expert annotation, F1 scores, held-out test sets) are insufficient for capturing the full meaning of student participation and ensuring these measures are meaningful and equitable. Through a case study of multilingual youth in one 8th-grade math classroom, we showed that ethnographically-oriented methods recover dimensions of classroom interaction that text-only transcripts cannot represent, and that students themselves contest both model classifications and the coding schemes used to produce them. Treating youth as epistemic authorities on their own talk reframes validation as a knowledge-production practice in which the people being measured help decide what the measurements mean. As LLM-based tools become more widely used in educational contexts, participatory and community-engaged approaches that center the voices of students are essential for ensuring that these systems are both accurate and equitable. Students are experts with respect to their own experiences, and epistemic authority should be shared with them when evaluating and designing tools to interpret their talk.

Acknowledgments

I would like to thank my husband, Reydrick Santos-Deonizio, and family for their support during the writing of this paper. We thank all the participants in the study and the students and teacher in the focal classroom. Thank you to Ms. Richer’s class for welcoming us into your community and letting us observe the brilliant ways you all engage in math. We also thank Helen Higgins who helped set up the study infrastructure, Lucía Langlois who helped with editing, Dr. Eujin Park who provided thoughtful feedback, and Dr. Christina Krist who supported through thought partnering. This work was funded by the Bill and Melinda Gates Foundation.

References

- [1] Bakhtawar Ahtisham, Kirk Vanacore, and Rene F. Kizilcec. 2026. Optimizing LLM annotation of classroom discourse through multi-agent orchestration. [doi:10.48550/arXiv.2603.13353](https://doi.org/10.48550/arXiv.2603.13353) arXiv:2603.13353 [cs].
- [2] Leah Hope Ajmani. 2025. Power, participation, and knowledge production: Technology mediated epistemic (in)justice. In *Companion Publication of the 2025 Conference on Computer-Supported Cooperative Work and Social Computing (CSCW Companion '25)*. Association for Computing Machinery, New York, NY, USA, 51–53. [doi:10.1145/3715070.3747343](https://doi.org/10.1145/3715070.3747343)
- [3] Emma Anderson, Grace C. Lin, Amelia Farid, Mic Fenech, Brandon Hanks, Eric Klopfer, Emily Doherty, Leanne Hirshfield, Mon-Lin Monica Ko, Peter Foltz, Ha Nguyen, Victoria Nguyen, Sara Ludovise, Rossella Santagata, Lydia Cao, Marlene Scardamalia, Dina Soliman, Monica Resendes, Ahmad Khanlari, Stacy Costa, Carol K. K. Chan, and Andy Nguyen. 2025. Exploring GenAI technologies within collaborative learning. (2025). <https://repository.isls.org/handle/1/11900>

- [4] Christina Baze and María González-Howard. 2025. A call to explicitly name and account for power in epistemic agency research. *Science Education* 109, 5 (Sept. 2025), 1499–1505. doi:10.1002/sce.21966
- [5] Anne H. Charity Hudley and Christine Mallinson. [n. d.]. Understanding English language variation in U.S. schools. https://www.tcpres.com/products/understanding-english-language-variation-in-u-s-schools_9780807751480
- [6] Dorottya Demszky and Heather Hill. 2023. The NCTE transcripts: A dataset of elementary math classroom transcripts. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, Ekaterina Kochmar, Jill Burstein, Andrea Horbach, Ronja Laarmann-Quante, Nitin Madnani, Anaïs Tack, Victoria Yaneva, Zheng Yuan, and Torsten Zesch (Eds.). Association for Computational Linguistics, Toronto, Canada, 528–538. doi:10.18653/v1/2023.bea-1.44
- [7] Ove Gunnar Drageset. 2015. Different types of student comments in the mathematics classroom. *The Journal of Mathematical Behavior* 38 (June 2015), 29–40. doi:10.1016/j.jmathb.2015.01.003
- [8] Kirstin Erath, Jenni Ingram, Judit Moschkovich, and Susanne Prediger. 2021. Designing and enacting instruction that enhances language for mathematics learning: a review of the state of development and research. *ZDM – Mathematics Education* 53, 2 (May 2021), 245–262. doi:10.1007/s11858-020-01213-2
- [9] Kirstin Erath and Susanne Prediger. 2021. Quality dimensions for activation and participation in language-responsive mathematics classrooms. In *Classroom Research on Mathematics and Language* (1 ed.), Núria Planas, Candia Morgan, and Marcus Schütte (Eds.). Routledge, Abingdon, Oxon ; New York, NY : Routledge, 2021., 167–183. doi:10.4324/9780429260889-12
- [10] Indigo Esmonde and Jennifer M. Langer-Osuna. 2013. Power in numbers: Student participation in mathematical discussions in heterogeneous spaces. *Journal for Research in Mathematics Education* 44, 1 (Jan. 2013), 288–315. doi:10.5951/jresmetheduc.44.1.0288
- [11] Nelson Flores. 2024. Producing deficiency and erasing colonialism in the bilingual education act. In *Becoming the System: A Raciolinguistic Genealogy of Bilingual Education in the Post-Civil Rights Era*. Oxford University Press. <https://doi.org/10.1093/oso/9780197516812.003.0005> Type: 10.1093/oso/9780197516812.003.0005.
- [12] Nelson Flores and Jonathan Rosa. 2015. Undoing appropriateness: Raciolinguistic ideologies and language diversity in education. *Harvard Educational Review* 85, 2 (June 2015), 149–171. doi:10.17763/0017-8055.85.2.149
- [13] Liza Gak. 2024. Reflecting on the relational: Youth-centered approaches to living with technology. In *Companion Publication of the 2024 Conference on Computer-Supported Cooperative Work and Social Computing (CSCW Companion '24)*. Association for Computing Machinery, New York, NY, USA, 27–30. doi:10.1145/3678884.3682047
- [14] Ofelia García and Li Wei. 2014. *Translanguaging*. Palgrave Macmillan UK, London. doi:10.1057/9781137385765
- [15] Maisie L. Gholson, , and Darrius D. Robinson. 2019. Restoring mathematics identities of Black learners: A curricular approach. *Theory Into Practice* 58, 4 (Oct. 2019), 347–358. doi:10.1080/00405841.2019.1626620 _eprint: <https://doi.org/10.1080/00405841.2019.1626620>.
- [16] Norma Gonzalez, Luis C. Moll, and Cathy Amanti. 2009. *Funds of knowledge: Theorizing practices in households, communities, and classrooms*. Routledge.
- [17] Kris D. Gutiérrez and Marjorie Faulstich Orellana. 2006. AT LAST: The "problem" of english learners: constructing genres of difference. *Research in the Teaching of English* 40, 4 (May 2006), 502–507. doi:10.58680/rte20065110
- [18] Emma Harvey, Allison Koencke, and Rene F. Kizilcec. 2025. "Don't forget the teachers": Towards an educator-centered understanding of harms from large language models in education. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, 1–19. doi:10.1145/3706598.3713210
- [19] Jane H. Hill. 1998. Language, race, and white public space. *American Anthropologist* 100, 3 (Sept. 1998), 680–689. doi:10.1525/aa.1998.100.3.680
- [20] Christina (Stina) Krist, Nessorine Machaka, Dan Voss, Nitasha Mathayas, Susan Kelly, and Soo-Yean Shim. 2023. Teacher noticing for supporting students' epistemic agency in science sensemaking discussions. *Journal of Science Teacher Education* 34, 8 (Nov. 2023), 799–819. doi:10.1080/1046560X.2022.2155355 _eprint: <https://doi.org/10.1080/1046560X.2022.2155355>.
- [21] Jennifer Langer-Osuna, Jen Munson, Emma Gargroetzi, Immanuel Williams, and Rosa Chavez. 2020. "So what are we working on?": how student authority relations shift during collaborative mathematics activity. *Educational Studies in Mathematics* 104, 3 (July 2020), 333–349. doi:10.1007/s10649-020-09962-3
- [22] Jennifer M. Langer-Osuna, Emma Gargroetzi, Jen Munson, and Rosa Chavez. 2020. Exploring the role of off-task activity on students' collaborative dynamics. *Journal of Educational Psychology* 112, 3 (April 2020), 514–532. doi:10.1037/edu0000464
- [23] Ramón Antonio Martínez. 2013. Reading the world in Spanglish: Hybrid language practices and ideological contestation in a sixth-grade English language arts classroom. *Linguistics and Education* 24, 3 (2013), 276–288. doi:10.1016/j.linged.2013.03.007
- [24] Ramón Antonio Martínez, Danny C. Martinez, and P. Zitlali Morales. 2022. Black lives matter versus Castañeda v. Pickard: a utopian vision of who counts as bilingual (and who matters in bilingual education). *Language Policy* 21, 3 (Sept. 2022), 427–449. doi:10.1007/s10993-021-09610-3
- [25] Ramón Antonio Martínez and Alexander Feliciano Mejia. 2020. Looking closely and listening carefully: A sociocultural approach to understanding the complexity of Latina/o/x students' everyday language. *Theory Into Practice* 59, 1 (Jan. 2020), 53–63. doi:10.1080/00405841.2019.1665414 _eprint: <https://doi.org/10.1080/00405841.2019.1665414>.
- [26] Ramón A. Martínez and P. Zitlali Morales. 2014. Puras groserías?: Rethinking the role of profanity and graphic humor in Latin@ students' bilingual wordplay. *Anthropology & Education Quarterly* (Jan. 2014). doi:10.1111/AEQ.12074
- [27] Nitasha Mathayas and Christina Krist. 2026. Re-Indexing epistemic responsibility: A grammatical analysis of how a teacher made space for students' epistemic agency. *Journal of Research in Science Teaching* 63, 1 (2026), 62–82. doi:10.1002/tea.70023 _eprint:

- <https://onlinelibrary.wiley.com/doi/pdf/10.1002/tea.70023>.
- [28] Jennifer Meyer, Thorben Jansen, Ronja Schiller, Lucas W. Liebenow, Marlene Steinbach, Andrea Horbach, and Johanna Fleckenstein. 2024. Using LLMs to bring evidence-based feedback into the classroom: AI-generated feedback increases secondary students' text revision, motivation, and positive emotions. *Computers and Education: Artificial Intelligence* 6 (June 2024), 100199. doi:10.1016/j.caeai.2023.100199
 - [29] Hector Morales, Jr. and Joseph DiNapoli. 2025. The underlife of a mathematics classroom: Latinx bilinguals navigating the official and unofficial spaces. *Journal of Research in Mathematics Education* 14, 2 (June 2025), 115–138. doi:10.17583/redimat.16567
 - [30] Luis Morales-Navarro, Daniel J. Noh, and Yasmin Kafai. 2025. Building babyGPTs: Youth engaging in data practices and ethical considerations through the construction of generative language models. In *Proceedings of the 24th Interaction Design and Children*. Association for Computing Machinery, New York, NY, USA, 1021–1026. <https://dl.acm.org/doi/10.1145/3713043.3731525>
 - [31] Maria Conchita A. Navarro and Orit Shaer. 2022. Re-imagining systems in the realm of immigration in higher education through participatory design. In *Companion Publication of the 2022 Conference on Computer-Supported Cooperative Work and Social Computing (CSCW Companion '22)*. Association for Computing Machinery, Taiwan, 76–79. <https://dl.acm.org/doi/10.1145/3500868.3559457>
 - [32] Nickolaus Alexander Ortiz. 2024. Lessons in paradise: envisioning a Black liberatory mathematics education. *Educational Studies in Mathematics* 116, 3 (July 2024), 539–550. doi:10.1007/s10649-023-10263-8
 - [33] Nickolaus Alexander Ortiz and Dalitso Ruwe. 2021. Black English and mathematics education: A critical look at culturally sustaining pedagogy. *Teachers College Record* 123, 10 (2021), 185–212. doi:10.1177/01614681211058978 _eprint: <https://doi.org/10.1177/01614681211058978>
 - [34] Sehee Park, Nia Nixon, Sidney D'Mello, Danielle Shariff, and Jaeyoon Choi. 2025. Understanding collaborative learning processes and outcomes through student discourse dynamics. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference (LAK '25)*. Association for Computing Machinery, New York, NY, USA, 938–943. doi:10.1145/3706468.3706547
 - [35] Luis E. Poza. 2018. The language of ciencia: translanguaging and learning in a bilingual science classroom. *International Journal of Bilingual Education and Bilingualism* 21, 1 (Jan. 2018), 1–19. doi:10.1080/13670050.2015.1125849 _eprint: <https://doi.org/10.1080/13670050.2015.1125849>
 - [36] Samuel L. Pugh, Arjun Rao, Angela E.B. Stewart, and Sidney K. D'Mello. 2022. Do speech-based collaboration analytics generalize across task contexts?. In *LAK22: 12th International Learning Analytics and Knowledge Conference (LAK22)*. Association for Computing Machinery, New York, NY, USA, 208–218. doi:10.1145/3506860.3506894
 - [37] Samuel L. Pugh, Shree Krishna Subburaj, Arjun Ramesh Rao, Angela E B Stewart, Jessica Andrews-Todd, and Sidney K D'Mello. 2021. Say what? Automatic modeling of collaborative problem solving skills from student speech in the wild. (2021).
 - [38] Jason G. Reitman, Charis Clevenger, Quinton Beck-White, Amanda Howard, Sierra Rose, Jacob Elick, Julianna Harris, Peter Foltz, and Sidney K. D'Mello. 2023. A multi-theoretic analysis of collaborative discourse: A step towards AI-facilitated student collaborations. In *Artificial Intelligence in Education*, Ning Wang, Genaro Rebollo-Mendez, Noboru Matsuda, Olga C. Santos, and Vania Dimitrova (Eds.). Springer Nature Switzerland, Cham, 577–589. doi:10.1007/978-3-031-36272-9_47
 - [39] Jonathan Rosa. 2018. *Looking like a language, sounding like a race: Raciolinguistic ideologies and the learning of latinidad*. Oxford University Press. <https://ccsre.stanford.edu/publications/looking-language-sounding-race-raciolinguistic-ideologies-and-learning-latinidad>
 - [40] Jaemarie Solyst, Cindy Peng, Wesley Hanwen Deng, Praneetha Pratapa, Jessica Hammer, Amy Ogan, Jason Hong, and Motahhare Eslami. 2025. Investigating youth AI auditing. doi:10.48550/arXiv.2502.18576 arXiv:2502.18576 [cs] version: 1.
 - [41] Jaemarie Solyst, Shixian Xie, Ellia Yang, Angela E.B. Stewart, Motahhare Eslami, Jessica Hammer, and Amy Ogan. 2023. "I would like to design": Black girls analyzing and ideating fair and accountable AI. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3544548.3581378
 - [42] Angela E.B. Stewart and Stephen Hutt. 2026. Beyond the numbers: Socio-Cultural context as a frame for learning analytics. In *Proceedings of the LAK26: 16th International Learning Analytics and Knowledge Conference (LAK '26)*. Association for Computing Machinery, New York, NY, USA, 852–858. doi:10.1145/3785022.3785076
 - [43] Tiera Tanksley, Angela D. R. Smith, Saloni Sharma, and Earl W Huff. 2025. "Ethics is not neutral": Understanding ethical and responsible AI Design from the lenses of Black youth. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, 1–20. doi:10.1145/3706598.3713510
 - [44] Sepehr Vakili and Maxine McKinney de Royston. 2022. Youth as philosophers of technology. *Mind, Culture, and Activity* 29, 4 (Oct. 2022), 336–355. doi:10.1080/10749039.2022.2066134 _eprint: <https://doi.org/10.1080/10749039.2022.2066134>
 - [45] Noreen Webb, Megan L. Franke, Marsha Ing, and Jacqueline Wong. 2014. Engaging with others' mathematical ideas: Interrelationships among student participation, teachers' instructional practices, and learning | Request PDF. *International Journal of Educational Research* (Dec. 2014). doi:10.1016/j.ijer.2013.02.001
 - [46] Li Wei and Ofelia Garcia. 2022. Not a first language but one repertoire: Translanguaging as a decolonizing project. *RELC Journal* 53, 2 (Aug. 2022), 313–324. doi:10.1177/00336882221092841
 - [47] Cathery Yeh, Daniel Lee Reinholz, Hakeoung Hannah Lee, and Mariah Moschetti. 2025. Beyond verbal: A methodological approach to highlighting students' embodied participation in mathematics classroom. *Educational Researcher* 54, 2 (March 2025), 103–110. doi:10.3102/0013189X241310169

A Research Methods

A.1 Student Interviews

Table 6. Overview of Participant Retrospection and Member Check Protocol

Phase	Questions / Prompts
Participant Retrospection Introduction	I want to talk about some of the moments when you were talking in your math class and how you felt and what you were thinking during those moments. Let's read this example together.
Retrospection Questions	"Can you tell me more about this and what you were doing here?" "Could you tell me how you were feeling during this moment?"
Member Check Introduction	One of the things we want to be able to learn from looking at how students talk in math is how to help teachers notice the brilliant and interesting conversations you all are having. As part of our research to understand the ways students talk in math class we are using Large Language Models (think of something like ChatGPT) to see how they pick up on the way you all talk to each other. So we can give the model a written version of your conversation and ask it to look for when a student asked a question, added on to what someone else said, or shared an idea about math. Here are some examples of what this model picked up from your conversation. Thinking about what you remember happening in this moment, what do you think about what the model noticed?
Member Check Questions	"If you agree, why do you agree?" "If you disagree, why do you disagree?" "Would you change anything or add on to what the model noticed?" "Was your intent captured correctly? Are there any things that the LLM is missing?" "Is there anything else you wish your teacher would notice when you are talking during math class?"
Closing	"Is there anything else you think I should know about math talk?"