The TME Framework: Multimodal Learner Modeling for Active Listening Skills in Collaborative Problem Solving

Xiaomeng Huang(✉), Xavier Ochoa, and Dani Hiterer

New York University, 370 Jay Street, New York City, NY, USA
{xiaomeng.huang,xavier.ochoa,dh3949}@nyu.edu

Abstract. Active listening is an important social–emotional skill for productive and inclusive collaboration, yet it remains difficult to assess in small-group collaborative problem-solving (CPS) activities. Teachers cannot continuously observe multiple groups simultaneously, while existing assessment approaches rely either on self-report—which is prone to bias—or on manual coding, which is labor-intensive and difficult to scale. We present the Theory-Measurement-Evaluation (TME) framework, a theory-based and explainable approach for modeling active listening in CPS grounded in the HURIER active listening framework. TME comprises three interlocking models. First, a **Theory Model** operationalizes HURIER into eight constructs, each translated into observable behavioral indicators. Second, a **Measurement Model** automatically extracts construct-aligned multimodal behavioral traces from audio and video and situates them within layered collaboration context derived from idea-evolution analysis across multiple time scales (micro-moments, collaboration phases, and whole-conversation dynamics). Building on evidence generated by the Measurement Model, we introduce an **Evaluation Model** that applies Bayesian Skill Tracing to estimate time-varying construct expression while preserving uncertainty. This produces an interpretable **Multimodal Learner Model** that combines quantitative metrics, exemplary behaviors, and descriptive narratives, and can be used by both human stakeholders and intelligent agents for downstream applications such as providing feedback or personalized learning. We evaluate the faithfulness and usefulness of the learner model outputs through expert ratings. Results suggest that TME can assess active listening skills from multimodal CPS data in a scalable and interpretable way.

Keywords: Active Listening · Collaborative Problem Solving · Multimodal Learner Modeling

1 Introduction

Small-group collaborative problem solving (CPS) is widely used in classrooms, yet groups often struggle to make progress when they fail to establish and

E. G. Blanchard et al. (Eds.): AIED 2026, LNAI 16581, pp. 416–430, 2027.
https://doi.org/10.1007/978-3-032-29744-0_28

maintain a shared problem space [3]. Constructing this shared space requires more than contributing ideas; it depends on how learners attend to, interpret, and responsively build on one another's contributions to negotiate and update common ground in real time. Accordingly, establishing and maintaining shared understanding is emphasized as a core CPS competency across multiple frameworks [10,13]. Because shared understanding is achieved through interaction, it relies on effective communication—and effective communication begins with active listening [6].

The National Communication Association defines active listening as "the process of receiving, constructing meaning from, and responding to spoken and/or nonverbal messages" [17]. This definition emphasizes not only hearing, understanding, and remembering, but also interpreting, evaluating, and responding (HURIER model; [6]). In CPS, these skills help learners interpret peers' intentions, make space for diverse voices, and align perspectives toward shared understanding. Despite its importance, active listening is rarely taught or coached in classrooms because it is difficult to observe and assess in real time. Teachers typically cannot monitor subtle listening behaviors across multiple groups or translate observations into actionable feedback [16]. Existing measurement approaches are also ill-suited to CPS: active listening is often assessed through self-report surveys or manual behavioral coding [17], which are respectively vulnerable to bias and difficult to scale. Moreover, CPS is multi-party and multimodal, and listening behaviors are interactionally situated—the same behavior can signal different meanings depending on what the group is doing in the moment (e.g., a nod during idea sharing versus negotiation) [3,10]. Together, these challenges reveal a gap: we lack scalable, theory-grounded, context-aware approaches for assessing active listening in small-group CPS.

To address this gap, we propose the Theory-Measurement-Evaluation (TME) framework for generating a multimodal learner model of active listening skills in CPS–a learner representation that uses multimodal evidence (e.g., audio/video) as input and produces multimodal, evidence-linked, quantitative and qualitative outputs. Following a theory-to-measurement perspective [12], TME operationalizes the HURIER framework into eight constructs with observable behavioral indicators [6], extracts aligned verbal and nonverbal traces from audio and video, and situates these behaviors within layered collaboration context through Idea Tracing Analysis (ITA) [11]. TME then applies an uncertainty-aware Bayesian Skill Tracing model, inspired by Bayesian Knowledge Tracing but adapted to CPS by integrating bidirectional behavioral evidence and abstaining when evidence is insufficient, to produce interpretable construct estimates suitable for feedback-oriented use.

This paper makes three contributions: (1) a theory-driven, context-aware measurement approach that operationalizes multimodal behaviors as situated evidence for active listening in CPS; (2) an uncertainty-preserving Bayesian tracing approach for estimating time-varying construct expression from heterogeneous, bidirectional evidence streams; and (3) an evidence-centered learner model representation that unifies construct-level metrics, linked behavioral

exemplars, and theory-aligned narrative summaries to support actionable feedback. Given the novelty of this approach, the main research question of this study is: *How faithful and useful are the outputs of the TME-based multimodal learner model of active listening skills in CPS?*

2 Related Work

In this section, we review research on active listening measurement and automated learner modeling. We synthesize these strands and highlight the need for context-aware multimodal learner modeling of active listening skills in small-group CPS.

2.1 Measurement of Active Listening Skills

Communication researchers have developed and validated multiple instruments for assessing active listening, including the Active Listening Observation Scale (ALOS) [9], the Active-Empathic Listening Scale (AELS) [4], the Active Listening Attitude Scale (ALAS) [8], and the HURIER Listening Skills Measurement Scale [6]. These instruments operationalize active listening in complementary ways: for example, ALAS emphasizes attitudes and perceived conversational opportunity, whereas HURIER specifies listening as a set of behavior-oriented processes. Because our goal is to model listening from observable interaction traces, including both verbal and nonverbal behaviors, we use HURIER as the guiding framework. Beyond self-report, active listening has also been studied using observational methods, including manual coding of verbal and nonverbal behaviors [17]. While such approaches capture listening-as-enacted, they are labor-intensive and difficult to scale.

These measurement challenges are amplified in small-group CPS. Self-report depends on learners' self-perceptions and retrospective recall, which can be unreliable in cognitively demanding, fast-moving collaboration, and manual coding is impractical for large-scale classroom use. Moreover, CPS is multi-party and unfolds over time as common ground is continuously negotiated [3,10], so the meaning of a listening behavior is often context-dependent. Together, these limitations motivate scalable, context-sensitive multimodal approaches that treat listening behaviors as situated evidence rather than decontextualized signals.

2.2 Automated Learner Modeling

Automated learner modeling is a foundational topic in AIED, aiming to infer learners' latent knowledge, skills, or dispositions from observable traces and update those inferences as new evidence becomes available [14]. The design of a learner model is driven by its intended use (e.g., personalization, diagnosis, reflection, feedback) [1], which in turn shapes three key choices: *(1) inputs* (what evidence is observed), *(2) inference model* (how evidence is mapped to latent states over time), and *(3) outputs* (what is reported and to whom).

A substantial body of work, especially in intelligent tutoring systems, illustrates common learner-model design choices: *inputs* are typically structured traces such as item responses or time and hint use, *inference* relies on probabilistic estimators of time-varying state with uncertainty (e.g., Bayesian Knowledge Tracing, Performance Factors Analysis, and IRT variants [1,14]) or learned sequence models (e.g., Deep Knowledge Tracing [15]), and *outputs* are mastery/skill profiles used for diagnosis and personalization. In parallel, open learner models are typically designed for reflection and feedback, and thus emphasize interpretable representations that expose learner state and the supporting evidence to users [7].

These traditions provide useful foundations but do not directly address the requirements of modeling collaboration skills such as active listening in CPS. Collaborative behavioral traces are not correctness events: they are multimodal, often non-binary, and may provide varying degrees of positive or negative evidence for a construct, with diagnosticity shaped by interaction context. To the best of our knowledge, no learner models have targeted active listening as the primary construct, especially in CPS. This motivates learner modeling approaches that (i) integrate bidirectional behavioral evidence with uncertainty-aware tracing and (ii) produce evidence-linked outputs that are interpretable and usable for feedback.

2.3 Summary of Gaps

Taken together, prior work provides strong foundations but leaves three gaps for modeling active listening in small-group CPS. First, a **measurement gap**: we lack theory-driven, context-aware methods for constructing multimodal listening evidence from verbal and nonverbal interaction traces. Second, an **inference gap**: we lack uncertainty-aware tracing approaches that integrate bidirectional behavioral evidence rather than correctness events. Third, an **output gap**: we lack evidence-centered learner model representations that are actionable for supporting collaboration skill development, such as active listening. These gaps motivate the TME framework proposed in this paper.

3 Study Context and Dataset

We conducted an in-person lab study in summer 2024 with IRB approval from New York University. Thirty participants (students and local professionals; median age = 27; 20 female, 9 male, 1 nonbinary) were assigned to nine teams of 3–4 members. Teams engaged in a CPS activity in which they designed a community park by discussing and deciding which facilities to include. Each participant received a handout assigning a persona with unique interests, potential conflicts, and information that only they possessed (see supplementary materials[1]). This task design was chosen to elicit active listening behaviors in a controlled CPS

[1] https://osf.io/skfx4/overview?view_only=fd869f49ad7f497e9bc512ac0be37a2c.

setting. By introducing information asymmetry and interdependent roles, successful task completion requires participants to attend to, interpret, and respond to others' contributions.

Groups sat around a square table equipped with a 360° camera system. Video captured both panoramic and individual views (15 fps), and smart pens logged writing activity. Four researchers unobtrusively observed each session. Collaborative sessions ranged from 19–41 minutes (M = 29.9), totaling 4.5 h of recorded interaction. Audio was transcribed and speaker-diarized using WhisperX [2], yielding 3,204 utterances. Three researchers verified transcript content and speaker diarization against video. Videos were rendered in two formats: a panoramic group view and individual close-up views to support behavioral annotation.

4 Theory-Measurement-Evaluation (TME) Framework

We propose the Theory-Measurement-Evaluation (TME) framework for modeling active listening in small-group CPS. As shown in Fig. 1, the framework takes multimodal interaction data as input and produces an interpretable multimodal learner model. TME proceeds in three stages: the *Theory Model* operationalizes active listening constructs and specifies behavioral indicators; the *Measurement Model* extracts and contextualizes multimodal behavioral traces aligned to these indicators; and the *Evaluation Model* aggregates evidence over time to estimate construct expression with uncertainty. We describe each component below.

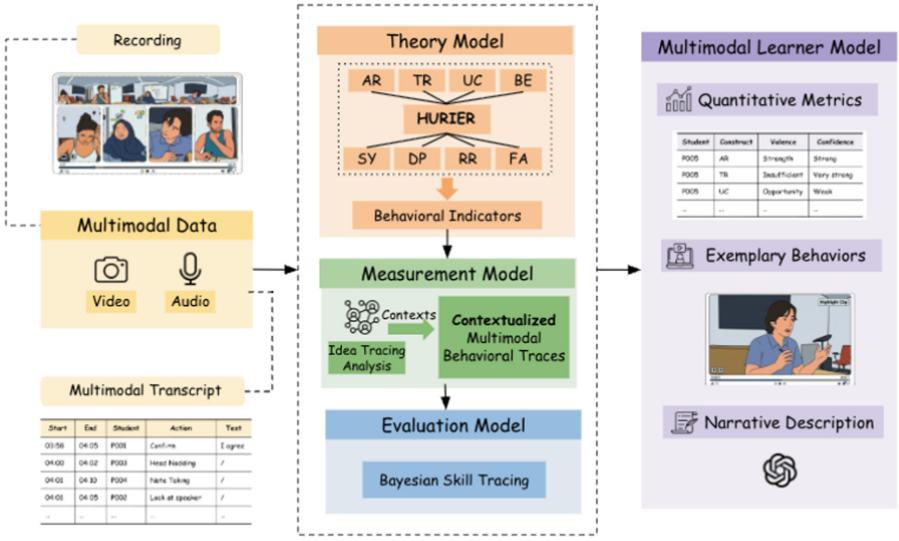

Fig. 1. Overview of the Theory–Measurement–Evaluation (TME) Framework.

4.1 Theory Model: HURIER Operationalization and Behavioral Indicators

The HURIER framework was primarily used in dyadic communication contexts. To apply it to small-group CPS, we operationalized HURIER into observable constructs and behavioral indicators suitable for multimodal measurement. We began with the 35 HURIER-based self-report items [6] and retained 28 items that could plausibly be evidenced through observable interaction behaviors. We excluded items that rely primarily on internal states or private judgments that are difficult to infer from audio/video without self-report (e.g., items requiring self-reflection). Rather than using the retained items as independent units, we grouped them into eight sub-constructs. This grouping was guided by (1) conceptual similarity among items and (2) the overlap in their observable behavioral indicators. In multimodal data, multiple items often manifest through the same or highly similar behavioral indicators.

This process yielded eight constructs that capture distinct but complementary components of active listening in CPS: (1) **Attention regulation (AR)**: maintains attention to the ongoing discussion and minimizes distractions; shows reception/orientation signals. (2) **Tracking and remembering (TR)**: tracks the flow of ideas and recalls prior contributions. (3) **Understanding checks and confirmation (UC)**: checks understanding, clarifies meaning, and confirms shared understanding. (4) **Broad engagement across peers (BE)**: engages with multiple peers' ideas rather than focusing on a single partner or only one's own contributions. (5) **Synthesis and other-oriented modification (SY)**: produces synthesized ideas that build on peers' contributions (beyond self-elaboration). (6) **Evaluation depth (DP)**: evaluates ideas using constraints/criteria (e.g., feasibility, evidence, tradeoffs), not only agreement/disagreement. (7) **Respectful and responsible responding (RR)**: responds constructively during disagreement (e.g., reasoned, repair-oriented) and avoids derailing interruptions. (8) **Facilitation of an environment for sharing (FA)**: invites participation and creates space for others to contribute.

For each construct, we produced a *construct card* that specifies: (a) a construct definition, (b) linked HURIER items, (c) observable behavioral indicators grounded in the item content and in CPS interaction patterns, and (d) *prime contexts* in which the construct is most diagnostic. Indicators include verbal actions (e.g., asking for clarification, paraphrasing, confirming), nonverbal cues (e.g., nodding, orienting gaze toward the speaker), and, when appropriate, derived interaction metrics (e.g., the proportion of contributions that engage with peers' ideas).

As an example, Synthesis & Other-Oriented Modification (SY) is derived from HURIER items (1, 6, 27, 34) and captures the extent to which a learner generates synthesized or modified ideas that build on peers' contributions rather than solely elaborating their own. SY is most diagnostic during idea *exploration*, *evaluation*, and *merging* micro-moments, and during phases where ideas are actively negotiated (divergent, mixed, and convergent phases). Evidence for SY combines (i) a primary metric: the ratio of modified ideas that build on oth-

ers' contributions to all modified ideas; with (ii) secondary behavioral indicators reflecting synthesis moves (e.g., modifying/synthesizing peers' ideas, explicitly connecting or integrating contributions). Repeated re-statement of one's own ideas without incorporating peer input is treated as negative evidence. Construct cards, operational details for the metric bands and indicator mappings are reported in the supplementary materials.

4.2 Measurement Model: Context-Aware Multimodal Evidence Extraction

The Measurement Model transforms raw multimodal recordings into *contextualized evidence units* aligned with the active listening constructs defined in the Theory Model. Given synchronized audio, video, and pen traces, we (1) extract multimodal behavioral traces corresponding to construct indicators, (2) infer collaboration context at multiple timescales, and (3) time-align behaviors and context to produce contextualized multimodal transcript for the Evaluation Model.

Behavior Extraction. Guided by HURIER, we defined 19 verbal action types (e.g., asking for clarification, paraphrasing, inviting participation). A codebook was developed iteratively by three coders (20% double-coded with adjudication) and then used to build an LLM-based labeling pipeline (ChatGPT 5.0 API), with prompt and context-window refinement validated against the human-coded subset. We employed iterative prompt engineering, including structured task instructions, context-window design, and example-based prompting. Prompts were refined through multiple iterations by comparing model outputs against the human-coded subset, and we incorporated rationale generation (i.e., requiring the model to justify its labels) to improve consistency. The final labeling pipeline was evaluated against the human-coded subset, achieving substantial agreement (above 70% accuracy). The finalized pipeline was then applied to the remaining unlabeled data. For nonverbal behaviors, we annotated gaze orientation, head nods, and writing activity in Datavyu; disagreements were rare and resolved through group adjudication. While prior work in AIED has demonstrated the feasibility of automatically detecting nonverbal behaviors using computer vision methods, we intentionally used manual annotation in this study to ensure high reliability and construct validity of the behavioral measures. Given that these nonverbal indicators serve as inputs to the measurement model, annotation noise could directly affect downstream inferences. We therefore prioritized high-quality human annotations to establish a reliable foundation for validating our construct operationalization. We view the integration of automated nonverbal behavior detection as an important direction for future work once the measurement framework is further validated.

Context Extraction. To interpret behaviors as interactionally situated evidence, we applied Idea Tracing Analysis (ITA) [11], which automatically extracts collaboration context by tracing group idea evolution (e.g., ideas being introduced,

expanded, evaluated, accepted, or refuted). ITA produced context labels at multiple timescales, including micro-moments (e.g., introduction, exploration, evaluation, merging, stance-taking), collaboration phases (e.g., ideation, divergence, convergence, mixed), and conversation-level indices (e.g., engagement with others' ideas, discussion depth, discussion balance). These contexts were time-aligned with behavioral traces to form evidence units used for context-aware construct estimation downstream.

4.3 Evaluation Model: Uncertainty-Aware Bayesian Skill Tracing

The Evaluation Model aggregates the construct-indexed evidence streams produced by the Measurement Model to estimate each learner's latent proficiency for each active listening construct, while preserving uncertainty. The key idea is to treat behavior-in-context evidence as weighted pseudo-observations that update a probabilistic belief state. This design is inspired by Bayesian student modeling and Bayesian Knowledge Tracing, but adapted to collaborative discourse where observations can provide varying degrees of positive or negative evidence rather than binary correctness events.

Evidence Encoding. For each learner s, construct c, and evidence unit t, extracted behavior-in-context is mapped to one of five evidence states: strong positive (S^+), weak positive (W^+), ambiguous (AMB), weak negative (W^-), or strong negative (S^-). These states contribute pseudo-count evidence with fixed weights: $S^+ \mapsto 1.0$, $W^+ \mapsto 0.5$ to positive evidence; $S^- \mapsto 1.0$, $W^- \mapsto 0.5$ to negative evidence; AMB$\mapsto$ 0. To reflect context-dependent diagnosticity, we scale evidence by a construct-specific multiplier $m_{c,t} \geq 1$ when the unit falls in the prime context defined in the construct card. Aggregating over units yields total weighted evidence $P_{s,c}$ (positive) and $N_{s,c}$ (negative), with evidence mass $m_{s,c} = P_{s,c} + N_{s,c}$.

Bayesian Tracing and Outputs. We model construct expression as a latent parameter $\theta_{s,c} \in [0, 1]$ with a Beta prior:

$$\theta_{s,c} \sim \text{Beta}(\alpha_0, \beta_0). \tag{1}$$

Given aggregated evidence, the posterior is:

$$\theta_{s,c} \mid \mathcal{E}_{s,c} \sim \text{Beta}(\alpha_0 + P_{s,c},\ \beta_0 + N_{s,c}), \qquad \hat{p}_{s,c} = \frac{\alpha_0 + P_{s,c}}{\alpha_0 + \beta_0 + P_{s,c} + N_{s,c}}. \tag{2}$$

We discretize $\hat{p}_{s,c}$ into five ordinal **valence** categories (growth area $\rightarrow$ clear strength; thresholds in the supplementary material). We compute **confidence** from evidence mass using a saturating mapping:

$$\text{conf}_{s,c} = 1 - \exp\left(-\frac{m_{s,c}}{\tau}\right), \tag{3}$$

and discretize $\text{conf}_{s,c}$ into five ordinal confidence levels (very weak $\rightarrow$ very strong; bins in the supplementary material). Finally, to avoid over-interpretation under sparse data, we label a construct as *insufficient evidence* when $m_{s,c} < m_{\min}$ and do not report directional valence.

4.4 Multimodal Learner Model

The output of the TME framework is a multimodal learner model intended to support evaluation and downstream feedback. It comprises three components: (1) quantitative construct estimates, (2) exemplary interaction moments linked to those estimates, and (3) natural-language summaries that contextualize the evidence.

Quantitative Metrics. For each learner–construct pair (s, c), the model outputs a continuous construct estimate $\hat{p}_{s,c}$ (or a metric-based valence label where applicable) together with an uncertainty-aware confidence score and ordinal confidence level derived from evidence mass. For interpretability, we also report an ordinal valence label (e.g., *strength* vs. *opportunity*). For example, Participant 003's *Attention Regulation* received $\hat{p} = 0.94$ (*clear strength*) with confidence $= 0.99$ (*very strong*), whereas *Broad Engagement Across Peers* received $\hat{p} = 0.06$ (*opportunity*) with confidence $= 0.75$ (*strong*).

Exemplary Moments. To ground quantitative estimates in observable behavior, we select 3–4 exemplary moments per participant. Moments are sampled to cover both strengths and growth areas (typically 1–2 of each), prioritizing higher-confidence constructs and including at least one insufficient evidence case when applicable. Including insufficient-evidence cases is important because the absence of observable listening behaviors can itself be diagnostically and instructionally meaningful (e.g., limited engagement with peers). For each selected construct, we identify the most diagnostic action (i.e., the highest-weighted indicator contributing to the estimate) and use its timestamp to retrieve a short video clip. To preserve interactional context, each clip includes a 10-second buffer before and after the focal action.

Narrative Descriptions. Finally, we translate the structured learner-model outputs into brief narrative descriptions intended for educators and downstream intelligent agents. Using an LLM (ChatGPT 5.2; prompts in the supplementary material), we generate summaries that (a) reflect the quantitative metrics, (b) cite linked exemplars as supporting evidence, and (c) highlight actionable feedback opportunities consistent with the target constructs. Each narrative description includes four elements: the group context, strengths, growth areas, and exemplary moments. We present one generated narrative as an example:

> This discussion appeared deep overall but unbalanced in participation. The student showed strong attention regulation by staying oriented to the group's immediate needs and responding with clarification when questions arose. Moment: 02:00–02:10 (Clip: 01:50–02:20) reflects the student's present, on-task engagement through answering questions and providing needed clarification.
> The student also demonstrated understanding checks that supported shared meaning-making. Moment: 14:02–14:30 (Clip: 13:52–14:40) shows the student asking for clarification, a concrete move that helps prevent the group from building on misaligned assumptions. This type of checking supports continuity in a

high-depth discussion by confirming what others mean before the conversation advances.
A next growth area is making synthesis more explicit and portable for the group. Moment: 10:10–10:20 (Clip: 10:00–10:30) shows the student starting to expand others' ideas and establish common ground, and the student could strengthen this by clearly linking multiple contributions and stating a combined takeaway. Facilitation for sharing was not clearly observed in the available evidence, so additional, observable inclusion moves would help the student support more balanced participation in future discussions.

5 Evaluation Methodology of Multimodal Learner Model

Evaluating learner models for enacted social skills differs from cognitive knowledge tracing because there is rarely a single ground-truth target comparable to the correctness labels typically used in knowledge tracing. Because our goal is to support feedback, we adopt an open learner model perspective and evaluate outputs in terms of transparency and actionability rather than predictive accuracy alone [7]. Moreover, judgments of listening behaviors are perspective-dependent and show limited convergence across raters, suggesting that no single "clean" objective measure is available for supervision [5]. Accordingly, rather than comparing construct estimates against a single set of human scores, we evaluate our learner-model outputs by (i) *faithfulness*—whether each claim is supported by the linked evidence, and (ii) *usefulness*—whether the output can support actionable feedback by a teacher or downstream agent.

To assess the faithfulness and usefulness of the multimodal learner model outputs (construct estimates, exemplary moments, and narrative descriptions), we developed 5-point ordinal rating scales. Faithfulness captures the extent to which linked evidence supports the model's claims (1 = not supported; 5 = fully supported with clear evidence). Usefulness captures the extent to which an output could support actionable instructional feedback (1 = not actionable; 5 = clearly actionable). The full rating scheme is provided in the supplementary material.

To cover both strong inferences and potential failure modes, we adopted a stratified sampling approach to select constructs for evaluation. For each participant, we selected four constructs representing distinct confidence and evidence conditions: (1) the highest-confidence strongest claim, (2) a high-confidence claim in the opposite direction when available, (3) the lowest-confidence non-insufficient claim, and (4) one construct labeled as insufficient evidence. This strategy ensured coverage of confident, contrasting, uncertain, and insufficient-evidence cases. For exemplary moments, we rated all selected moments (3–4 per participant). We also rated one narrative description per participant (30 total).

Two researchers who directly observed and assessed the participants independently rated an overlapping subset comprising 30% of items for each output type. Because ratings were concentrated toward the high end of the scale, we report inter-rater reliability using Gwet's AC2 with quadratic weights. Agreement was high across all outputs: construct scores (Faithfulness AC2= 0.952,

95% CI [0.924, 0.980], $N = 44$; Usefulness AC2= 0.839, 95% CI [0.770, 0.908], $N = 44$), exemplary moments (Faithfulness AC2= 0.937, 95% CI [0.889, 0.984], $N = 39$; Usefulness AC2= 0.863, 95% CI [0.794, 0.931], $N = 39$), and generated descriptions (Faithfulness AC2= 0.987, 95% CI [0.966, 1.000], $N = 11$; Usefulness AC2= 0.978, 95% CI [0.951, 1.000], $N = 11$). After establishing reliability on the overlap subset, the remaining items were divided evenly and rated by a single researcher.

6 Results

In this section, we report results addressing the research question: we first present the multimodal learner model produced by the TME framework and then evaluate the faithfulness and usefulness of its outputs.

6.1 Quantitative Metrics Evaluation

For construct-level estimates, Faithfulness averaged 4.59 (SD = 0.57) with 95.8% of ratings ≥ 4, and Usefulness averaged 4.50 (SD = 0.55) with 97.5% ≥ 4. We inspected the small number of cases rated 3 on faithfulness or usefulness to characterize failure modes. These cases were concentrated in the *Respectful & Responsible Response (RR)* construct, where the model produced weaker or insufficient-evidence judgments that raters viewed as overly conservative.

Qualitative review of the corresponding videos suggests a context-driven explanation: the group reached consensus on strategy early, and the remaining discussion proceeded smoothly with limited disagreement or negotiation. As a result, the interaction contained few opportunities for the RR construct to surface through its most diagnostic behaviors (e.g., reasoned disagreement, repair moves, or constructive responses under tension). Participants still exhibited some evaluative talk and occasional elaboration on others' ideas, which generated non-zero evidence and prevented a purely "insufficient" classification, but the overall evidence mass remained low. Notably, the model's confidence for these RR estimates was also weak/insufficient, consistent with limited diagnostic evidence rather than a strong negative judgment.

This failure mode highlights a limitation of construct estimation for behaviors that are opportunity-dependent: when a collaboration episode lacks the interactional conditions that elicit a construct (here, negotiation under disagreement), both human raters and models may differ in whether the appropriate conclusion is "insufficient evidence" versus "neutral/adequate." Future refinements could incorporate explicit opportunity modeling (e.g., detecting whether disagreement/negotiation opportunities occurred) before interpreting RR evidence.

6.2 Exemplary Moments Evaluation

For exemplary moments, Faithfulness averaged 4.56 (SD=0.83), with 89.4% ≥ 4 and 5.7% ≤ 2; Usefulness averaged 4.40 (SD=0.85), with 90.1% ≥ 4 and 5.7%

≤ 2. Inspection of low-rated cases revealed two recurring failure modes that reflect a tension between *evidential diagnosticity* (what best supports estimation) and *instructional salience* (what works as a coachable exemplar).

Low-salience Evidence. In several cases, the highest-weight indicators were subtle behaviors that are meaningful as evidence in aggregate but weak as standalone exemplars (e.g., brief head nods or short gaze shifts contributing to attention regulation). These clips were often rated as faithful because the behavior was observable and aligned with the construct, yet less useful because the action was fleeting, ambiguous, and difficult to coach reliably. This suggests that selecting exemplars solely by evidence weight can surface signals that support inference but do not function well as persuasive instructional examples.

Form–function Mismatch. A second pattern involved moments whose surface form matched a construct, but whose function depended on immediate local outcomes. For example, an utterance like "Is that fair?" can be coded as facilitating participation, yet may not create genuine space if the speaker does not pause, yield the floor, or if peers do not take up the invitation. In these cases, usefulness ratings decreased because stakeholders judged the clip as a less authentic strategy, even when the coded behavior was present. This points to the need for moment selection criteria that incorporate short-horizon interactional consequences (e.g., floor transfer or peer uptake) in addition to phase-level context.

6.3 Narrative Description Evaluation

For generated descriptions, Faithfulness averaged 4.73 (SD=0.52), with 96.7% of ratings ≥ 4, and Usefulness averaged 4.83 (SD=0.38), with 100% ≥ 4. Overall, experts rated the narratives highly on both dimensions. This pattern is consistent with our design: narratives are generated from structured, evidence-linked inputs: the construct estimates and selected exemplary moments. We also prioritize higher-confidence constructs and clips when producing descriptions. As a result, narrative quality largely reflects the quality of upstream evidence selection rather than unconstrained free-text generation.

The small number of lower narrative faithfulness ratings co-occurred with cases where the underlying construct estimate or selected moment was rated lower in faithfulness, suggesting that narrative errors were primarily inherited from upstream evidence rather than introduced by the narrative generation step itself. These results indicate that evidence-linked narrative summaries can improve interpretability and perceived instructional usefulness, provided that the underlying evidence selection is reliable.

7 Discussion

Regarding our research question, expert ratings suggest that the TME-based multimodal learner model produces outputs that are generally faithful to linked

evidence and useful for feedback in small-group CPS, with the strongest performance for narrative summaries and high-confidence construct estimates. Importantly, the failure cases point to a substantive measurement insight: some active listening constructs are *opportunity-dependent* and some evidence is *diagnostic but not instructionally salient*. For example, respectful/responsible responding may not surface when groups reach early consensus and rarely negotiate disagreement, and subtle nonverbal cues can support inference in aggregate but function poorly as standalone teaching exemplars. These findings suggest that future iterations should more explicitly model local interaction outcomes (e.g., whether invitations yield peer uptake or floor transfer) and construct-specific opportunities before interpreting low-evidence cases as learner deficits.

TME is a promising framework because it targets three gaps in prior work: a **measurement gap**, an **inference gap**, and an **output gap**. Addressing the **measurement gap**, TME builds on HURIER [6] and the view that CPS success depends on maintaining shared understanding [3,10] by operationalizing active listening as construct-aligned behavioral indicators and interpreting multimodal behaviors as *situated evidence* within evolving collaboration contexts, responding to concerns about decontextualized CPS features and limited interpretability [12]. Addressing the **inference gap**, our uncertainty-aware Bayesian Skill Tracing extends learner-state estimation beyond correctness events by integrating bidirectional behavioral evidence and abstaining under insufficient evidence, producing estimates that are interpretable and calibrated for feedback-oriented use. Addressing the **output gap**, TME links construct estimates to exemplary moments and theory-aligned narratives, aligning with open learner model goals of transparency and actionability [7]. Together, expert ratings indicate that these evidence-linked outputs are both inspectable and instructionally meaningful, complementing prior measurement approaches that rely on self-report or labor-intensive observation [17].

8 Conclusion

This paper introduces the Theory–Measurement–Evaluation (TME) framework for modeling active listening in small-group CPS from multimodal interaction traces. By linking theory-grounded constructs to contextualized behavioral evidence and uncertainty-aware tracing, TME produces evidence-linked learner-model outputs (construct estimates, exemplars, and narratives) that expert raters judged to be generally faithful and useful for feedback.

This work has limitations. Evidence rules, prime-context specifications, and weighting choices are design decisions that may require calibration across tasks, populations, and classroom conditions, and our evaluation is based on a relatively small lab dataset. Future work should validate robustness in authentic classrooms and test downstream impact—whether feedback informed by the learner model improves active listening and collaboration outcomes at scale. Deployment of multimodal learner models also raises important ethical and privacy considerations. The use of audio, video, and interaction data requires careful

attention to informed consent, data minimization, and secure storage, particularly in classroom settings. To preserve privacy, for example, outputs of TME can be designed to limit access to the individuals who generated the data or to authorized stakeholders, thereby reducing unnecessary exposure of sensitive behavioral information.

Although demonstrated in the context of active listening, TME represents a generalizable framework for constructing multimodal learner models. It could be extended to other learning constructs that are operationalized through multimodal behavioral indicators, with corresponding adaptations to theory, measurement, and evaluation models.

References

1. Abyaa, A., Khalidi Idrissi, M., Bennani, S.: Learner modelling: systematic review of the literature from the last 5 years. Educ. Tech. Res. Dev. **67**(5), 1105–1143 (2019). https://doi.org/10.1007/s11423-018-09644-1, http://link.springer.com/10.1007/s11423-018-09644-1
2. Bain, M., Huh, J., Han, T., Zisserman, A.: WhisperX: time-accurate speech transcription of long-form audio (2023). https://doi.org/10.48550/arXiv.2303.00747, http://arxiv.org/abs/2303.00747, arXiv:2303.00747 [cs]
3. Barron, B.: When smart groups fail. J. Learn. Sci. **12**(3), 307–359 (2003). https://doi.org/10.1207/S15327809JLS1203_1
4. Bodie, G.D.: The active-empathic listening scale (AELS): conceptualization and evidence of validity within the interpersonal domain. Commun. Quarterly **59**(3), 277–295 (2011). https://doi.org/10.1080/01463373.2011.583495, http://www.tandfonline.com/doi/abs/10.1080/01463373.2011.583495
5. Bodie, G.D., Jones, S.M., Vickery, A.J., Hatcher, L., Cannava, K.: Examining the construct validity of enacted support: a multitrait-multimethod analysis of three perspectives for judging immediacy and listening behaviors. Commun. Monographs **81**(4), 495–523 (2014). https://doi.org/10.1080/03637751.2014.957223, http://www.tandfonline.com/doi/abs/10.1080/03637751.2014.957223
6. Brownell, J.: Listening: Attitudes, Principles, and Skills. Routledge, New York, 7 edn. (2023). https://doi.org/10.4324/9781003316794, https://www.taylorfrancis.com/books/9781003316794
7. Bull, S., Kay, J.: Open learner models. In: Nkambou, R., Bourdeau, J., Mizoguchi, R. (eds.) Advances in Intelligent Tutoring Systems, pp. 301–322. Springer, Berlin, Heidelberg (2010). https://doi.org/10.1007/978-3-642-14363-2_15
8. Bybee, B., Frost, J.: Active listening attitude scale (ALAS). In: The Sourcebook of Listening Research, pp. 167–173. John Wiley & Sons, Ltd (2017). https://doi.org/10.1002/9781119102991.ch9, https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119102991.ch9
9. Fassaert, T., Van Dulmen, S., Schellevis, F., Bensing, J.: Active listening in medical consultations: development of the active listening observation scale (ALOS-global). Patient Educ. Couns. **68**(3), 258–264 (2007). https://doi.org/10.1016/j.pec.2007.06.011, https://linkinghub.elsevier.com/retrieve/pii/S0738399107002637
10. Graesser, A.C., Fiore, S.M., Greiff, S., Andrews-Todd, J., Foltz, P.W., Hesse, F.W.: Advancing the science of collaborative problem solving. Psychologic. Sci. Publ. Int. **19**(2), 59–92 (2018). https://doi.org/10.1177/1529100618808244

11. Huang, X., Ochoa, X., Gopalakrishnan, M., Zhang, S.: Idea tracing analysis (ITA): a temporal-structural methodology for modeling collaborative discourse and contextualizing multimodal interactions. In: Proceedings of the International Conference on Computer Supported Collaborative Learning (CSCL) (2026)
12. Huang, X., Ochoa, X.: Charting the development of collaboration skills through collaborative learning analytics systems. J. Learn. Anal. **12**(1), 338–366 (2025). https://doi.org/10.18608/jla.2025.8523, https://learning-analytics.info/index.php/JLA/article/view/8523
13. OECD: PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematic, Financial Literacy and Collaborative Problem Solving. PISA, OECD (2017). https://doi.org/10.1787/9789264281820-en, https://www.oecd-ilibrary.org/education/pisa-2015-assessment-and-analytical-framework_9789264281820-en
14. Pelánek, R.: Bayesian knowledge tracing, logistic models, and beyond: an overview of learner modeling techniques. User Model. User-Adap. Inter. **27**(3), 313–350 (2017). https://doi.org/10.1007/s11257-017-9193-2
15. Piech, C., et al.: Deep knowledge tracing. In: Advances in Neural Information Processing Systems. vol. 28. Curran Associates, Inc. (2015). https://proceedings.neurips.cc/paper/2015/hash/bac9162b47c56fc8a4d2a519803d51b3-Abstract.html
16. Worsley, M., Ochoa, X.: Towards collaboration literacy development through multimodal learning analytics. In: Companion Proceedings 10th International Conference on Learning Analytics & Knowledge (LAK20), p. 12 (2020)
17. Worthington, D.L., Bodie, G. (eds.): The sourcebook of listening research: methodology and measures. John Wiley & Sons Inc, Hoboken, NJ (2017)