A Multimodal Framework for Embodied Cognition in Oral Explanations

Dr. Jason W Morphew, Purdue University – West Lafayette

Dr. Jason Morphew is an assistant professor at Purdue University in Engineering Education and serves as the director of undergraduate curriculum and advanced learning technologies for SCALE, and is affiliated with the INSPIRE research institute for Pre-College Engineering and the Center for Advancing the Teaching and Learning of STEM. Dr. Morphew's research focuses on the application of principles of learning derived from cognitive science and the learning sciences to the design and evaluation of learning environments and technologies that enhance learning, interest, and engagement in STEM. Morphew earned his Ph.D. from the University of Illinois Urbana-Champaign in Educational Psychology from the Cognitive Science of Teaching and Learning Division where he examined self-regulated learning of engineering students enrolled in introductory physics courses.

Mr. Amirreza Mehrabi, Purdue Engineering Education

Amirreza Mehrabi is an accomplished scholar at Purdue University specializing in machine learning and AI-aided modeling and Engineering Education. As an interdisciplinary researcher, he designs and deploys intelligent platforms that tackle complex decision-making challenges, integrating advanced techniques such as transformer models, reinforcement learning, and statistical inference. His work spans across adaptive testing, cognitive fatigue detection by ML, agentic AI agents tutors, and multimodal analytics, leading to innovative solutions with measurable real-world impact. Amirreza's expertise is reflected in his custom architectures, model optimization research, and psychometrics—combining educational measurement science with deep learning. Amirreza's collaborative spirit, technical mastery, and commitment to transparent, bias-aware modeling foster a vibrant culture of innovation.

Junior Anthony Bennett, Purdue University – West Lafayette (College of Engineering)

Junior Anthony Bennett is a Graduate Research Assistant and Lynn Fellow at Purdue University, West Lafayette, Indiana, USA. He is pursuing an Interdisciplinary Ph.D. program in Engineering Education majoring in Ecological Sciences and Engineering (ESE). His research focus is the 'Impact of Extended Reality (XR) Technologies on Learning'. He worked for over a decade in higher education and held multiple positions of responsibilities including Lecturer of Mathematics, Engineering Physics and Industrial Engineering Core Courses; He had served as Program Leader, Academic Advisor, Applied Sciences and Engineering Cooperative Education Coordinator, Program Coordinator & Chair of Infrastructure Committee in higher education. He earned a Master of Science in Manufacturing Engineering Systems from Western Illinois and was recognized for being an outstanding graduate student. He earned a Bachelor of Education in TVET Industrial Technology – Electrical from the University of Technology, Jamaica. He is a Certified Manufacturing Engineer (CMfgE) with the Society for Manufacturing Engineers (SME). He is the immediate past president of the ESE Interdisciplinary Graduate Program Graduate Student Organization and recent co-chair of the 18th Annual 2024 ESE Symposium at Purdue University. He is the 2025 recipient of the School of Engineering Education (ENE) Outstanding Service/Leadership Award at Purdue University, West Lafayette.

Aashvi Majmundar, Purdue University – West Lafayette (College of Engineering)

A Multimodal Framework of embodied cognition via gesture–speech dynamics

Jason W. Morphew[1] Amirreza Mehrabi[1,2]
Junior Anthony Bennett[1,3] Aashvi Miten Majmundar[4]

[1]School of Engineering Education, Purdue University
[2]Dept. of Computer and Electronic Engineering, Purdue University
[3]Ecological Sciences and Engineering, Purdue University
[4]School of Aeronautics and Astronautics, Purdue University

jmorphew@purdue.edu

Abstract

This study investigates how students' statistical reasoning emerges via embodied cognition and language. According to embodied cognition theory, gesture is treated as part of the cognitive process that shapes how ideas are organized during oral explanation. An analytic framework integrates computer-vision tracking of gesture via k-Nearest Neighbors with Large Language Model (LLM). Across interviews, students often attempted to show what they meant through gestures. McNeill's taxonomy classifies these gestures into iconic, referencing imagined spaces, and metaphorical movements. Our expert and computer analysis indicates that when students expressed high confidence, their gestures were larger, temporally stable, and spatially consistent. Episodes in which gesture and speech were closely coordinated tended to be associated with more coherent conceptual talk, whereas moments of divergence often coincided with partial or developing ideas. Rather than providing interpretation on its own, the framework makes visible how embodied action and verbal reasoning jointly contribute to students' explanations.

Introduction

Assessment practices typically rely on the assumption that speech is the dominant mode for communicating understanding, but learners frequently communicate conceptual information via embodied actions, especially hand gestures, that are not redundant with speech (Alibali & Goldin-Meadow, 1993; Church & Goldin-Meadow, 1986). Gesture serves not only as a facilitator of learning abstract and complex concepts (Junokas et al., 2018; Lindgren et al., 2019, 2022), but also as an observable indicator of how knowledge is structured at a given moment (Tscholl et al., 2021). Gestures are not auxiliary from the perspective of embodied cognition, but they are part of

the cognitive process (Alibali & Nathan, 2012; Bennett et al., 2024; Hostetter & Alibali, 2008; Lakoff, 2012; Weisberg & Newcombe, 2017). Explanations assessed only through speech can fail to capture implicit or transitional dimensions of understanding (Nathan, 2012) and verbal fluency can be mistaken for conceptual knowledge – putting learners with language-related challenges at a disadvantage (Mainela-Arnold et al., 2014). This is especially true in engineering contexts where inherent conceptual explanations naturally lend themselves to gestures that reveal how learners organize conceptually (Alibali & Goldin-Meadow, 1993; Church & Goldin-Meadow, 1986; Evans & Rubin, 1979; Goldin-Meadow et al., 1993; McNeill, 1985; Perry et al., 1988).

Advances in multimodal sensing now make it possible to systematically capture and align speech with embodied action. In this study, we introduce a multimodal framework that integrates computer-vision tracking of gesture kinematics with lLarge Language Model (LLM) analysis of spoken discourse to jointly examine verbal explanations and embodied enactments during conceptual reasoning. Taking foundational concepts of engineering statistics as an illustrative domain, we investigate to what extent conceptual understanding is manifested in the coordination of speech and gesture (Son et al., 2018). We argue that embodied action is meaningful, often-overlooked evidence of understanding that addresses the persistent underrepresentation of constructs in explanation-focused research and assessment (Nathan, 2012). Guided by this perspective, we investigate:

- **RQ1.** How do engineering students express conceptual understanding across verbal explanation and embodied action when explaining foundational statistics concepts?
- **RQ2.** What is the increase in diagnostic information about conceptual understanding that can be obtained from the analysis of speech synced with embodied enactment, relative to the analysis of speech alone?

Literature Review

Embodied cognition does not see the mind as a symbolic act but as a combination of perception and action while we think (Barsalou, 2008; Wilson, 2002). Thinking and reasoning trigger sensorimotor actions, like encoding spatial, relational, or structural information (Barsalou, 2008; Hostetter & Alibali, 2008). This sensorimotor activity can support learning and be used as evidence of learning if it is synchronized with the concept (Alibali & Nathan, 2012).

The gesture-speech unity hypothesis states that speech and gesture have a common origin with two separate but coordinated outputs (McNeill, 1992). Gestures often have specific conceptual content. In education we can see it in several response behaviour such as tracing a slope in the air when describing a formula, and may occur prior to the spoken words, hinting at representations ready for action before they are linguistically stable (Church & Goldin-Meadow, 1986; Goldin-Meadow, 2003; Goldin-Meadow et al., 1993). However, not all hand movements are conceptual. McNeill's taxonomy distinguishes between representational gestures and non-representational ones (McNeill, 1992). McNeill theory view gestures as evidence, which can be studied in terms of form, target relevance and recurrence (McNeill, 1992).

Gesture–Speech Configuration

McNeill argues that gesture and thought are bidirectionally linked: gesture expresses mental representations, but it also actively shapes how ideas are organized during speech (Alibali & Goldin-Meadow, 1993; Church & Goldin-Meadow, 1986; Perry et al., 1992). Within McNeill's framework, gesture–speech concordance serves as an indicator of conceptual representation, where the alignment between gesture and speech can reflect either accurate or transitional understanding (McNeill, 1992; Perry et al., 1992). Mismatches between gesture and speech have been shown to signal transitional or weakly held conceptions that are susceptible to change (Perry et al., 1988, 1992), a pattern that has been extended beyond children to learners in more advanced conceptual domains.

In particular, studies support the McNeill framework in mathematics, science, and statistics, gestures express implicit ideas before they are verbalized (Alibali & Nathan, 2012; Broaders et al., 2007; Goldin-Meadow et al., 1993; Herrera & Riggs, 2013; Ping et al., 2021; Rueckert et al., 2017; Son et al., 2018; Zhang et al., 2025) and reduce cognitive load by offloading relational structure onto space (Glenberg, 2008; Wilson, 2002; Zhang et al., 2025), and in collaborative settings help coordinate joint attention and shared representations. Thus, gesture capture and analysis provides support for inferences about conceptual understanding beyond what can be identified with speech alone (Sung et al., 2023). This is especially true in engineering contexts, where conceptual explanation naturally lends itself to gesture, revealing how learners spatially organize abstract ideas (Son et al., 2018).

Gestures must be aligned with semantically meaningful units in speech (Hostetter & Alibali, 2008; McNeill, 1992). Temporal units, alignment criteria and inference thresholds have to be explicitly defined (Blikstein & Worsley, 2016; Di Mitri et al., 2018; Stoll et al., 2018). Speech and gesture together provide more information than speech alone (Alibali & Nathan, 2012; Ping et al., 2021; Rueckert et al., 2017; Son et al., 2018) and mitigate the biases of verbal fluency, thus increasing construct validity and equity (Mainela-Arnold et al., 2014; Messick, 1995). Recent advances enable the detection, classification, and semantically align gesture and speech at scale due to recent advances in visual language models, computer vision, and motion detection (Mehrabi & Morphew, 2025; Mehrabi et al., 2024).

Methodology

Data Collection and Gesture Library

We collected synchronized video and audio recordings of two undergraduate engineering students during semi-structured oral interviews focused on linear regression and descriptive statistics (Mehrabi et al., 2024). Interview questions were focused on conceptual understanding of the statistical concepts rather than computation or procedural knowledge (Appendix B). In addition, the questions were intentionally open-ended to encourage spontaneous explanations rather than memorized answers, allowing students to express understanding through words and/or gestures. For example, participants were asked what the mean (or median, variance, etc.) of a data set represented and how they would interpret these statistics (Giannakos et al., 2022). Follow-up questions were asked to clarify or explore conceptual understanding. Some of the follow-up

questions included ”How does skewness relate to data distribution and central tendency?” ”What is your understanding of variance, and how would you apply it in real-world contexts?” and ”What does correlation mean, and how would you evaluate it between two variables?” (Appendix B)

To systematically analyze participant gesture, and make gesture interpretation educationally meaningful, we first constructed a domain-specific gesture library (Table 1). This library catalogs the most frequently observed hand movements that students spontaneously produce when explaining statistical ideas (e.g., tracing a regression line, sketching a bell curve, or pointing out an outlier). Each gesture is explicitly linked to a statistical concept, providing a shared interpretive framework across all analyses. Keeping this library small and conceptually grounded ensures that detected gestures remain interpretable by instructors and researchers, rather than becoming opaque machine-learning categories.

Table 1: Gesture library for linear regression and descriptive statistics.

Gesture Name	**Meaning**	**Hands Involved**
Straight Line Gesture	Represents a linear function or regression line	One or Two
Slope Gesture	Indicates positive or negative slope of a line	One
Dot in Air	Represents a single data point	One
Scatter Plot Gesture	Represents multiple data points in a scatter plot	One or Two
Coordinate Plane Gesture	Represents the x–y coordinate plane	Two
Normal Distribution Gesture	Represents a bell-shaped distribution	Two
Mean / Median Gesture	Indicates central tendency of a distribution	One
Outlier Gesture	Highlights a data point distant from a cluster	One
Correlation Gesture	Represents positive or negative correlation between variables	Two
Grouping Gesture	Indicates clustering or grouping of data points	Two

Multimodal Processing Pipeline

Our pipeline separates (i) perceptual gesture detection from (ii) semantic interpretation, allowing us to analyze embodied expression independently of language-model inference.

Gesture detection and temporal segmentation. We apply the MediaPipe library that tracks the index fingertip in each frame. The fingertips make a contentious trajectory $\mathbf{p}_t \in \mathbb{R}^2$ with $L = 16$ frames.

Normalization and feature construction. Each motion window is centered at its centroid and isotropically scaled by its maximum radial distance, removing variation due to camera placement, student posture, and drawing scale while preserving geometric structure and direction (e.g., increasing versus decreasing slope). Normalized coordinates are vectorized to form fixed-length feature representations for classification.

Robustness via data augmentation (training only). To model natural variation in how students enact the same representational form, training samples are augmented using small rotations,

horizontal reflections, mild scaling, and nonlinear temporal warping. These transformations preserve gesture identity while improving robustness to execution variability without introducing new gesture categories.

Gesture classification, smoothing, and episode formation. Each time window of motion is classified into a gesture type using a k-nearest neighbors (KNN) classifier. KNN classifier assign gesture labels based on trajectory similarity. While the new gestures may be noisy, KNN resolves the issue with normalization. Contiguous frames with stable labels are merged into gesture episodes, each defined by an onset time, offset time, and detector-assigned gesture form (Table 1).

Gesture Training and Duration Estimation Multiple sample videos and training videos were recorded and collected for every gesture and included in the gesture library, such as those for a vertical line, horizontal box, SSE, wave, or infinity gesture (Table 1). These videos were organized and processed by libraries of OpenCV and MediaPipe to feed into the KNN algorithm. The noise was significantly reduced with the help of prediction smoothing throughout the frames. Then, labels were merged to complete gesture episodes.

Episode-level confidence. How the gesture classifier favors one hand's movement form over all others for each moment of hand motion, and we average this value over the duration of a gesture episode. Perceptual confidence for each time window is defined as the normalized dominance of the most probable gesture class relative to the total classification score across all classes. Episode-level confidence computed as the average of these window-level confidence values. This measure reflects the visual stability and not the semantic correctness of the accompanying speech. A conservative threshold (HIGH_CONF; confidence at or above 0.75) is applied to restrict subsequent analyses to episodes with stable and reliable visual patterns.

Speech transcription and multimodal alignment. Audio is transcribed using automatic speech recognition to produce time-stamped utterances. Each gesture episode is aligned to the spoken segment whose temporal midpoint is closest to the episode midpoint, capturing natural synchrony between representational movement and verbal explanation without requiring forced-alignment models.

Multimodal embedding and semantic retrieval. Each aligned gesture–speech pair constitutes a multimodal explanatory unit. These units are embedded using a sentence-transformer model and indexed for semantic retrieval, enabling comparison across participants and retrieval of conceptually similar explanations.

The pipeline yields three feature representations for assessment modeling: speech-only, gesture-only, and combined multimodal features. These representations are used to predict observable indicators of conceptual understanding during the oral exam, enabling direct evaluation of whether gesture contributes diagnostic value beyond speech alone. Full algorithmic details and parameter settings are provided in Appendix A. Figure 1 summarizes the workflow.

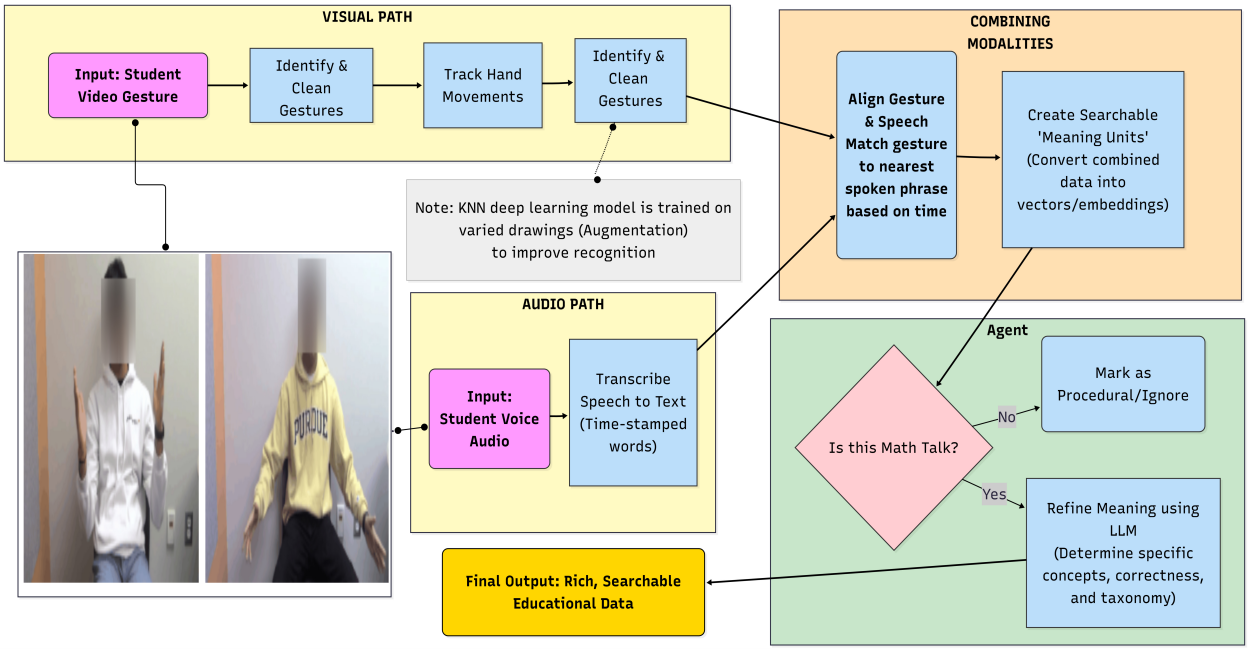


Figure 1: Overview of the multimodal pipeline, separating perceptual gesture detection from semantic interpretation.

Meaning Agent for Episode-Level Semantic Annotation

We designed a hybrid agent to assign episode-level semantic fields used in analysis, including statistical concept labels, gesture–speech relation, and McNeill taxonomy. This agent was responsible for connecting the sentences to the gesture, considering the timeline. A correct gesture would be one that exhibits an accurate representation of the nature of the statistical term that the gesture is intended to illustrate through its alignment with speech. An incorrect gesture may either incorrectly depict the spatial aspect of the statistical concept or be improperly aligned with the verbal description. When math-context evidence is present, a local instruction-tuned model (FLAN-T5) is prompted to refine gesture form and to label correctness, gesture–speech relation (aligned/partial/misaligned/none), and McNeill category (iconic/metaphoric/deictic/beat/other) under a forced JSON schema that must include short verbatim evidence quotes. Outputs are parsed using wrapped-JSON constraints with a single retry. Full prompts, gating rules, and fallback logic are documented in Appendix B.

Results

Validity and Robustness of the High-Signal Subset

Most of the gestures for both students occurred during math talk and met the confidence threshold (see Table 2). Of the 127 and 58 total gestures for Greg and Fred respectively, the pipeline retained 92 and 36. Confidence values remained tightly above threshold, indicating that the retained gestures represent real, stable movement rather than noise. Table 3 indicates that in this filtered set, 97.3% of Fred's and 93.9% of Greg's gestures were explanatory. This means that a

gesture that has passed both filters—math context and high confidence—is almost certainly doing conceptual work, not just taking up conversational space. Together these two tables confirm that the high-signal subset is robust and ready for more analysis. According to Fig. 2, confidence values for retained episodes are tightly clustered around the threshold, not at ceiling, suggesting visually stable movement and not incidental motion. The form-labeling model with an accuracy of 0.85 and a Macro-F1 of 0.84 on the held-out split (Table 4) also accredits this filtering process, meaning that the form labels assigned to each episode can be trusted.

Table 2: Gesture quality of two students.

Student	Total	Math	HighConf	Expl.	% Math	% HighConf	% Expl.
Fred	58	37	58	36	0.638	1.00	0.621
Greg	127	98	127	92	0.772	1.00	0.724

Table 3: Explanatory coupling within math-context high-confidence gestures.

Student	Math+HighConf	Explanatory	Rate
Fred	37	36	0.973
Greg	98	92	0.939

Table 4: Summary of form-label reliability on a held-out split.

Metric	Accuracy	Macro-F1	Weighted-F1
Held-out performance	0.85	0.84	0.85

Notes. Labels are produced by a model trained on the gesture library and evaluated on a held-out split; no Greg/Fred episodes appear in training or validation. Full per-class precision/recall/F1 are reported in Table 7.

Embodied Structure of High-Signal Gesture Episodes

As shown in Table 5, square/box made up 61.1% of Fred's episodes and 54.3% of Greg's, with straight line as the second most common form for both, and idiosyncratic or unknown forms remaining small for both students. Fig. 3(a) provides visual confirmation of this concentration. The fact that two students converged on the same narrow set of forms suggests these are not personal habits but a shared spatial vocabulary for communicating statistical ideas—consistent with spatially stable expression under high epistemic confidence. Gestures are not evenly distributed across the discourse but rather cluster around specific statistical concepts (Fig. 3(b)). Both students gestured most on *mean*, but while Fred's range of concepts was narrow and focused, Greg's extended to *slope*, *regression*, and *standard deviation* as well. This difference reflects a meaningful distinction in how much of the statistical content each student chose to ground spatially. Since concepts are assigned only when episodes are temporally aligned with explanatory speech, these distributions characterize the conceptual landscape of gestured explanation, not speech alone. As seen in Table 6, Fred's square/box appeared with *mean* in 15 of his episodes, while Greg's square/box was spread over *slope* (14), *mean* (13), and *SD* (5).

Table 5: Distribution of gesture forms (high-confidence explanatory episodes). Percentages are within-student (Fred N=36, Greg N=92).

Fred (N=36)			**Greg** (N=92)		
Form	Count	%	Form	Count	%
square/box	22	61.1	square/box	50	54.3
straight line	7	19.4	straight line	16	17.4
beat/rhythm	2	5.6	beat/rhythm	11	12.0
curve/arc	1	2.8	curve/arc	5	5.4
unknown	4	11.1	circle/loop	3	3.3
			pointing/dot	3	3.3
			bell curve	2	2.2
			unknown	2	2.2

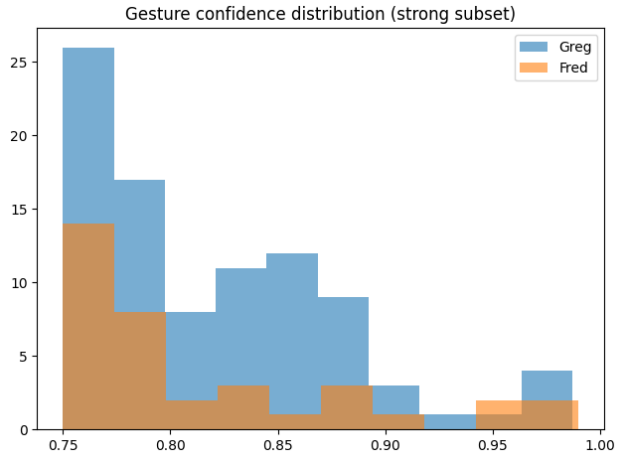


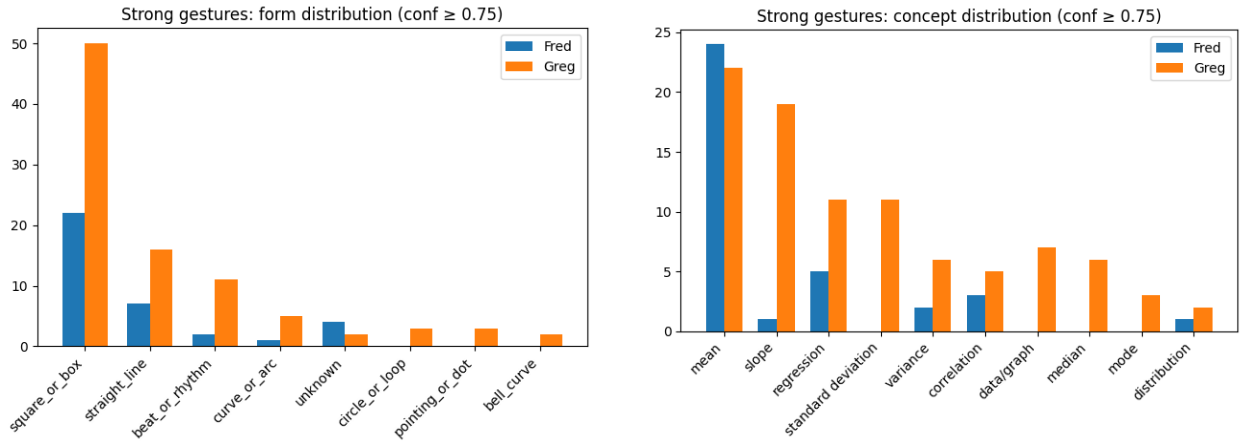
Figure 2: Gesture confidence distributions (aligned with high-confidence explanatory gestures).

Discussion and Conclusion

The results demonstrate that gesture is not simply a support for speech but rather that it acts as a stable spatial system for organizing conceptual relations during reasoning. Students did reuse a compact set of gesture forms consistently aligned to specific statistical concepts . This suggests that learners recruit durable spatial schemas to structure abstract ideas like center, variation and trend . These gestures occurred precisely at the moments when relational structure needed to be expressed, consistent with accounts of cognition as distributed across modalities and unfolding dynamically in time rather than through speech alone. Methodologically, this framework provides a systematic and interpretable way to incorporate embodied evidence at scale (Blikstein & Worsley, 2016; Di Mitri et al., 2018). Most prior approaches rely on manual coding or coarse aggregation, obscuring the moment-to-moment coordination between gesture and speech (Hostetter & Alibali, 2008; McNeill, 1992). Validating episode-level units and extracting stable form concept mappings captures how explanatory meaning is enacted through coordinated gesture speech activity (Alibali & Nathan, 2012; Son et al., 2018). This moves beyond

Table 6: Form×Concept counts for high-confidence explanatory episodes.

Fred			Greg		
Form	Concept	Count	Form	Concept	Count
square/box	mean	15	square/box	slope	14
straight line	mean	5	square/box	mean	13
			straight line	regression	5
			square/box	SD	5
			straight line	mean	4



(a) (b)

Figure 3: Form distributions (a) and concept distribution (b) of Greg and Fred.

speech-based assessment by enabling representation-level diagnostics (Perry et al., 1988, 1992). High-confidence mathematical gestures overwhelmingly expressed explanatory content without requiring additional elicitation from students (Broaders et al., 2007). These findings have implications on online learning environments, where instructors lack access to embodied cues that naturally occur when they are in an offline setting. Digital platforms with gesture recognition integrated will help assist student understanding and fill the gap that verbal explanations leave.

Practical application

This work has two outcomes. First, it provides a means of quantifying the typically qualitative gesture analysis. It does this through a reproducible, automated pipeline that could even see application beyond this study. Second, it adds a new dimension to the assessment of students in engineering and STEM contexts where explanation is already expected—assessment can be made not only through language but also whether the gestures of a student are meaningful and aligned with the underlying statistical concept being explained. The full pipeline is released publicly at https://github.com/amehrabi67/OralExams.git for future use by instructors and researchers. Users upload gesture recordings and information about expected gesture types. The output is automated gesture detection results for whatever classroom or assessment contexts they were working

with.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Awards No. 2322015 and No. 2142317. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Alibali, M. W., & Goldin-Meadow, S. (1993). Gesture-speech mismatch and mechanisms of learning: What the hands reveal about a child's state of mind. *Cognitive Psychology*, *25*, 468–523. https://doi.org/10.1006/cogp.1993.1012

Alibali, M. W., & Nathan, M. J. (2012). Embodiment in mathematics performance and instruction: Evidence from mathematical discourse. *Developmental Review*, *32*(4), 249–286.

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology*, *59*, 617–645.

Bennett, J., Morphew, J., & McColgan, M. (2024). Embodied learning with gesture representation in an immersive technology environment in stem education.

Blikstein, P., & Worsley, M. (2016). Multimodal learning analytics and education data mining: Using computational technologies to measure complex learning tasks. *Journal of Learning Analytics*, *3*(2), 220–238.

Broaders, S. C., Cook, S. W., Mitchell, Z., & Goldin-Meadow, S. (2007). Making children gesture brings out implicit knowledge and leads to learning. *Journal of Experimental Psychology: General*, *136*(4), 539.

Church, R. B., & Goldin-Meadow, S. (1986). The mismatch between gesture and speech as an index of transitional knowledge. *Cognition*, *23*, 43–71. https://doi.org/10.1016/0010-0277(86)90053-3

Di Mitri, D., Schneider, J., Specht, M., & Drachsler, H. (2018). From signals to knowledge: A conceptual model for multimodal learning analytics. *Journal of Computer Assisted Learning*, *34*(4), 338–349. https://doi.org/10.1111/jcal.12288

Evans, M. A., & Rubin, K. H. (1979). Hand gestures as a function of communal and egocentric speech in preschool children. *Journal of Genetic Psychology*, *135*(2), 209–216.

Giannakos, M., Spikol, D., Di Mitri, D., Sharma, K., Ochoa, X., & Hammad, R. (Eds.). (2022). *The multimodal learning analytics handbook*. Springer. https://doi.org/10.1007/978-3-031-08076-0

Glenberg, A. M. (2008). Embodiment for education. In *Handbook of cognitive science* (pp. 355–372). Elsevier.

Goldin-Meadow, S. (2003). *Hearing gesture: How our hands help us think*. Harvard University Press.

Goldin-Meadow, S., Alibali, M. W., & Church, R. B. (1993). Transitions in learning: Evidence from gesture-speech mismatch. *Developmental Psychology*, *29*(2), 279–289.

Herrera, N. M., & Riggs, E. M. (2013). Visual attention, perception, and learning: The role of gesture in learning science. *Journal of Geoscience Education*, *61*(4), 358–371.

Hostetter, A. B., & Alibali, M. W. (2008). Visible embodiment: Gestures as simulated action. *Psychonomic Bulletin & Review*, *15*(3), 495–514.

Junokas, M. J., Lindgren, R., Kang, J., & Morphew, J. W. (2018). Enhancing multimodal learning through personalized gesture recognition. *Journal of Computer Assisted Learning*, *34*(4), 350–357.

Lakoff, G. (2012). Explaining embodied cognition results. *Topics in cognitive science*, *4*(4), 773–785.

Lindgren, R., Morphew, J., Kang, J., & Junokas, M. (2019). An embodied cyberlearning platform for gestural interaction with cross-cutting science concepts. *Mind, Brain, and Education*, *13*(1), 53–61.

Lindgren, R., Morphew, J. W., Kang, J., Planey, J., & Mestre, J. P. (2022). Learning and transfer effects of embodied simulations targeting crosscutting concepts in science. *Journal of Educational Psychology*, *114*(3), 462.

Mainela-Arnold, E., Alibali, M. W., Hostetter, A. B., & Evans, J. L. (2014). Gesture–speech integration in children with specific language impairment. *International Journal of Language & Communication Disorders*, *49*(6), 761–770. https://doi.org/10.1111/1460-6984.12115

McNeill, D. (1985). So you think gestures are nonverbal? *Psychological review*, *92*(3), 350.

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. University of Chicago Press.

Mehrabi, A., & Morphew, J. (2025). Uncovering the cognitive roots of misconceptions in physics education for engineering students through transitional diagnostic models. *2025 ASEE Annual Conference & Exposition*.

Mehrabi, A., Morphew, J. W., & Quezada, B. S. (2024). Enhancing performance factor analysis through skill profile and item similarity integration via an attention mechanism of artificial intelligence. *Frontiers in Education*, *9*, 1454319.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741.

Nathan, M. J. (2012). Rethinking formalisms in formal education. *Educational Psychologist*, *47*(2), 125–148.

Perry, M., Church, R. B., & Goldin-Meadow, S. (1988). Learners' use of gesture in the production of explanations. *Child Development*, *59*(4), 881–892.

Perry, M., Church, R. B., & Goldin-Meadow, S. (1992). Is gesture-speech mismatch a general index of transitional knowledge? *Cognitive Development*, *7*, 109–122.

Ping, R., Church, R. B., Decatur, M.-A., Larson, S. W., Zinchenko, E., & Goldin-Meadow, S. (2021). Unpacking the gestures of chemistry learners: What the hands tell us about correct and incorrect conceptions of stereochemistry. *Discourse Processes*, *58*(3), 213–232. https://doi.org/10.1080/0163853X.2020.1839343

Rueckert, L., Church, R. B., Avila, A., & Trejo, T. (2017). Gesture enhances learning of a complex statistical concept. *Cognitive Research: Principles and Implications*, *2*(1), 2. https://doi.org/10.1186/s41235-016-0036-1

Son, J. Y., Ramos, P., DeWolf, M., Loftus, W., & Stigler, J. W. (2018). Exploring the practicing-connections hypothesis: Using gesture to support coordination of ideas in

understanding a complex statistical concept. *Cognitive Research: Principles and Implications*, *3*, 1–13. https://doi.org/10.1186/s41235-017-0085-0

Stoll, J., Edwards, C., & Humphrey, S. (2018). Multimodal sensing of student behavior in learning environments. *Journal of Educational Technology Systems*, *47*(1), 102–115.

Sung, H., Kim, D., Swart, M. I., & Nathan, M. J. (2023). Multimodal behavior analysis: Two patterns of collaborative construction of embodied knowledge. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *45*.

Tscholl, M., Morphew, J., & Lindgren, R. (2021). Inferences on enacted understanding: Using immersive technologies to assess intuitive physical science knowledge. *Information and Learning Sciences*, *122*(7/8), 503–524.

Weisberg, S. M., & Newcombe, N. S. (2017). Embodied cognition and stem learning: Overview of a topical collection in cr: Pi. *Cognitive research: principles and implications*, *2*(1), 38.

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin & Review*, *9*(4), 625–636.

Zhang, I., Son, J. Y., & Stigler, J. W. (2025). Toward a cognitive developmental theory of embodied learning in stem domains. *Educational Psychologist*, 1–21.

A Multimodal Gesture Processing and Meaning Inference Pipeline

A.1 Overview

This appendix documents the full multimodal pipeline used to extract, classify, align, and interpret gesture episodes in synchronized classroom-style video and audio data. The pipeline consists of six stages: (i) gesture trajectory acquisition, (ii) normalization and feature construction, (iii) data augmentation, (iv) gesture classification and temporal smoothing, (v) speech transcription and multimodal alignment, and (vi) semantic embedding and retrieval.

A.2 Gesture Trajectory Acquisition

Video streams were processed using OpenCV for frame decoding and MediaPipe for hand landmark detection. At each frame t, the two-dimensional coordinates of the index fingertip were extracted, yielding a raw trajectory

$$\mathbf{p}_t \in \mathbb{R}^2, \quad t = 1, 2, \ldots, T,$$

sampled at the video frame rate f_{ps}.

To capture short, interpretable motion primitives while maintaining temporal continuity, trajectories were segmented into overlapping windows of fixed length $L = 16$ frames using a constant stride. Each window is represented as

$$\mathbf{P}_t = \{\mathbf{p}_t, \mathbf{p}_{t+1}, \ldots, \mathbf{p}_{t+L-1}\}.$$

A.3 Normalization and Feature Representation

To ensure invariance to camera placement, drawing location, and gesture scale, each trajectory window was spatially normalized. The centroid of each window was computed as

$$\bar{\mathbf{p}} = \frac{1}{L} \sum_{i=1}^{L} \mathbf{p}_i,$$

and all points were translated by subtracting this centroid. The trajectory was then scaled by the maximum Euclidean distance from the centroid,

$$r = \max_i \|\mathbf{p}_i - \bar{\mathbf{p}}\|_2,$$

producing normalized coordinates

$$\hat{\mathbf{p}}_i = \frac{\mathbf{p}_i - \bar{\mathbf{p}}}{r + \varepsilon},$$

where ε is a small constant added for numerical stability.

The normalized trajectory was vectorized by concatenating coordinates across time, yielding a $2L$-dimensional feature vector

$$\boldsymbol{\phi}(\mathbf{P}) \in \mathbb{R}^{32}.$$

A.4 Data Augmentation

To improve robustness to variation in gesture execution, geometric and temporal data augmentation was applied during training. Augmentations included small random rotations about the centroid, horizontal reflections, mild isotropic scaling, and nonlinear temporal warping. Rotations were implemented as

$$\mathbf{p}'_i = \bar{\mathbf{p}} + \mathbf{R}(\theta)(\mathbf{p}_i - \bar{\mathbf{p}}),$$

where $\mathbf{R}(\theta)$ is a 2D rotation matrix and θ is sampled from a restricted range. All augmented trajectories were resampled to the fixed length L.

A.5 Gesture Classification and Temporal Smoothing

Gesture classification was performed using a k-nearest neighbors (KNN) classifier with $k = 7$ and inverse-distance weighting. For a test feature vector $\mathbf{x}$, the class score for class c was computed as

$$s_c(\mathbf{x}) = \sum_{j \in \mathcal{N}_k(\mathbf{x})} \frac{\mathbb{1}[y_j = c]}{\|\mathbf{x} - \mathbf{x}_j\|_2 + \delta},$$

where $\mathcal{N}_k(\mathbf{x})$ denotes the set of k nearest neighbors in feature space and δ is a small constant to avoid division by zero.

Frame-level predictions were temporally smoothed using majority voting over a sliding window of $M = 9$ frames. Gesture episodes were constructed by merging contiguous frames with identical labels, with short gaps bridged and spurious segments removed based on minimum duration constraints.

A.6 Speech Transcription and Multimodal Alignment

Audio streams were transcribed using an automatic speech recognition system, producing time-stamped segments

$$\{(s_i, e_i, T_i)\}_{i=1}^{N},$$

where s_i and e_i denote start and end times and T_i denotes the transcript.

Each gesture episode with temporal bounds (t_s, t_e) was aligned to the speech segment whose midpoint was temporally closest:

$$i^\star = \arg\min_i \left| \frac{t_s + t_e}{2} - \frac{s_i + e_i}{2} \right|.$$

The aligned transcript window was used as linguistic evidence for interpreting the gesture episode.

A.7 Semantic Embedding and Retrieval

Gesture-speech embedded by a sentence-transformer model

$$\mathbf{z}_j = E(d_j) \in \mathbb{R}^{384}.$$

All embeddings were ℓ_2-normalized and indexed using FAISS for approximate nearest-neighbor search.

Given a query text q with embedding $\mathbf{q} = E(q)$, retrieval was performed by minimizing Euclidean distance:

$$\text{Retrieve}(q) = \arg\min_j \|\mathbf{q} - \mathbf{z}_j\|_2.$$

A.8 Episode-level confidence.

For each sliding window feature vector $\mathbf{x}_t$, the KNN classifier produces distance-weighted class scores

$$s_c(\mathbf{x}_t) = \sum_{j \in \mathcal{N}_k(\mathbf{x}_t)} \frac{\mathbb{1}[y_j = c]}{\|\mathbf{x}_t - \mathbf{x}_j\|_2 + \delta},$$

and predicts $\hat{y}_t = \arg\max_c s_c(\mathbf{x}_t)$. We define the window-level confidence as the normalized dominance of the winning class,

$$c_t = \frac{\max_c s_c(\mathbf{x}_t)}{\sum_{c'} s_{c'}(\mathbf{x}_t)} \in (0, 1].$$

After temporal smoothing of labels, contiguous windows with a stable label are merged into gesture episodes e with index set $\mathcal{T}_e$. Episode confidence is the average window confidence within the episode,

$$\mathrm{Conf}(e) = \frac{1}{|\mathcal{T}_e|} \sum_{t \in \mathcal{T}_e} c_t.$$

We apply a conservative episode-level threshold (HIGH_CONF: $\mathrm{Conf}(e) \geq 0.75$) to restrict analysis to visually stable gesture episodes. Importantly, this confidence quantifies perceptual certainty in gesture *form* (derived from the vision classifier), not the semantic correctness of speech.

Table 7: Gesture form labeling performance on a held-out split (independent of Greg/Fred episodes).

Form	**Precision**	**Recall**	**F1**
Straight Line	0.80	0.88	0.84
Curved Line	0.84	0.83	0.83
Slope Gesture	0.71	0.69	0.70
Horizontal Box	0.77	0.85	0.81
Circle Gesture	0.87	0.88	0.88
Normal Distribution	0.83	0.72	0.77
Infinity Gesture	0.95	0.94	0.94
Wave Gesture	0.77	0.73	0.75
Vertical Box	0.80	0.84	0.82
SSE	0.98	0.93	0.95
DATA (pointing)	0.94	0.87	0.90

B Meaning Agent: Implementation Details

B.1 Inputs and Episode Unit

The Meaning Agent operates on episode-level files generated by the multimodal alignment stage (e.g., `Greg_highconf_gesture_transcript.csv`). Each record corresponds to a single gesture episode and includes participant identifier, gesture identifier, detector-assigned form label, episode start and end times, mean detection confidence, and a time-bounded transcript evidence window represented both as line-formatted text and as a compact fallback string.

B.2 Evidence Selection and Length Controls

For each episode, the agent preferentially selects the line-formatted transcript window as evidence and falls back to the compact representation when the formatted text does not meet a minimum character threshold. Evidence length is capped to control prompt size and stabilize model behavior.

B.3 Context Gating: Administrative vs. Mathematical Discourse

To prevent semantic inference on procedural or administrative speech, a two-stage gating mechanism is applied. First, an administrative detector identifies segments related to consent, recording procedures, or interview logistics using compiled pattern rules. Second, a mathematics and statistics keyword detector identifies domain-relevant

discourse using a curated lexicon (e.g., *mean, slope, regression, variance, standard deviation, correlation, distribution, residuals, data points*).

An episode is marked `gate_math=True` only if at least one mathematical keyword is present and no administrative flag is triggered. The number of keyword matches (`math_kw_hits`) is recorded for auditing.

B.4 Rule-Based Priors

Independently of language model inference, two priors are computed for each episode:

1. **Concept prior:** up to two canonical statistical concepts inferred from transcript keywords.
2. **Form prior:** a normalized gesture-form category inferred from the detector label (e.g., mapping "Vertical Box" and "SSE" to `square_or_box`, and "DATA" to `pointing_or_dot`).

These priors serve as soft cues rather than ground truth and are retained for interpretability and audit.

B.5 Local LLM Adjudication under a Forced Schema

When `gate_math=True`, a local instruction-tuned language model (FLAN-T5) is prompted to produce a single JSON object under a closed schema. The model must return: (a) refined gesture form, (b) up to two speech concepts, (c) speech correctness (correct / mostly_correct / incorrect), (d) gesture–speech relation (aligned / partial / misaligned / none), (e) McNeill gesture taxonomy (iconic / metaphoric / deictic / beat / other), (f) up to two short verbatim evidence quotes, and (g) a brief rationale referencing at least one quote.

Requiring explicit quotation enforces evidence coupling and reduces unsupported inferences.

B.6 Robust Parsing, Retry Strategy, and Fallback

Model outputs are first parsed by searching for JSON enclosed within `<json>...</json>` tags; if absent, brace-delimited JSON extraction is attempted. If parsing fails, the agent retries once using a shorter and more constrained prompt. If parsing fails again, a structured rule-based fallback is returned using the form prior and concept prior, with conservative defaults for relation and taxonomy. This design prevents downstream aggregation from collapsing into null categories due to formatting failures. All raw model outputs are retained for auditability.

B.7 Outputs Used in Analysis

The agent generates one enriched episode-level file per participant (e.g., `Greg_gesture_meaning_llm.csv`) containing gating flags, priors, language-model adjudications, evidence quotes, and rationales. All "strong" summary statistics reported in the main analysis are computed exclusively on the math-context subset (`gate_math=True`), ensuring that coupling rates and form–concept distributions reflect mathematical explanation rather than procedural discourse.

B.8 Interview questions

How would you explain the concept of central tendency to a person who is struggling with statistics?

How do you view the relationship between data distribution and skewness as it relates to central tendency?

What comes to your mind when you think of the concept of variance?

Can you give some real-world examples of variance?

Tell me, how would you connect the variability of data set to skewness?

How do you explain the concept of regression?

What does the slope of a regression line mean to you?

How do you report or interpret the quality of a regression model?

What does it mean when we talk about correlation in statistics?