MEASURING MEANINGFUL CONTRIBUTION IN GROUP DISCUSSION

Nanzheng Xie, Ryuichi Ikeda, Suk Min Hwang

University of California, Berkeley
Berkeley, CA 94720, USA

ABSTRACT

Traditional measures of participation in group discussions rely on surface-level indicators such as speaking frequency, obscuring the qualitative value of individual utterances. This paper proposes an NLP-based framework for measuring meaningful contribution that emphasizes how utterances advance collective problem solving rather than how often participants speak. We introduce the Meaningful Contribution Score (MCS), an utterance-level, five-dimensional measure spanning semantic and interactional constructs (relevance, novelty, enablement, affect, and decision proximity). Focusing on construct validity, we validate MCS on the GAP corpus using a new human-annotated subset and a three-way comparison between human ratings, heuristic MCS components, and LLM-based judgments. Results show strong alignment for semantic relevance, partial alignment for novelty and enablement, and persistent unreliability for affect, highlighting where lightweight heuristics suffice and where context-aware models are needed for interpretable assessment of contribution.

KEYWORDS

Meaningful Contribution, Group Discussion, Construct Validity, Discourse Modeling, Sentence Embeddings

1. INTRODUCTION

Group discussion is a central mechanism through which people learn, coordinate, and solve problems together, yet our ability to quantify the quality of an individual contribution remains limited (Rosé et al., 2008). Instructors and researchers often rely on coarse or subjective judgments to determine whether a speaker "moved the discussion forward," leaving open the question of how such contributions can be measured in a systematic and reproducible way. Although recent NLP work has begun to characterize conversational dynamics, the field still lacks validated, utterance-level constructs for measuring how individual contributions advance collaborative problem solving (Lee et al., 2018).

We address this gap by studying Meaningful Contribution in small-group problem-solving discussions. Rather than treating contribution as a single dimension, we conceptualize it as a multidimensional phenomenon grounded in both semantic content and conversational function. Building on

David C. Wyld et al. (Eds): ACSTY, MLSC, SVC, AIBD, ITCSS, ADCOM, NATP, SOFE – 2026
pp. 237-249, 2026. CS & IT - CSCP 2026 DOI: 10.5121/csit.2026.160419

prior work, we introduce the **Meaningful Contribution Score (MCS)**, a five-dimensional framework intended to capture distinct ways in which an utterance can help a group make progress: Relevance, Novelty, Enablement, Affect, and Decision Proximity. A formal definition of each component and its computational implementation is provided in Section 4.

While our preliminary experiments explored links between MCS and downstream group outcomes (e.g., influence, agreement improvement), we identified that predictive correlations provide only indirect evidence of validity and are difficult to interpret given small sample sizes. Before relying on MCS as a predictor, it is essential to first establish construct validity: determining whether each automated component genuinely aligns with human judgments of the phenomenon it is intended to measure.

In this paper, we conduct the first systematic validation of MCS. We compare its five components against two independent reference points: (a) **human annotations** provided by two trained coders using a shared codebook, and (b) **LLM-based annotations** generated by Claude 3.5 Sonnet. This three-way comparison—human-human reliability, MCS-human alignment, and AI-human alignment—allows us to rigorously assess which dimensions of MCS capture psychologically meaningful constructs, and where heuristic approaches diverge from context-aware models.

Our contributions are threefold:

1. A transparent computational formulation of MCS integrating semantic embeddings and functional heuristics;
2. A new human-annotated dataset covering six meetings and five conversational constructs; and
3. A comprehensive construct-validity evaluation, revealing that while semantic measures (e.g., Relevance) align well with human judgment, interactional and affective dimensions benefit significantly from the contextual reasoning capabilities of LLMs over static heuristics.

2. RELATED WORK

Research on conversational quality spans computational linguistics, discourse analysis, and the learning sciences. Prior work has proposed a wide range of frameworks for characterizing how speakers contribute to collective problem solving, often emphasizing semantic content, interactional function, or affective tone.

2.1. Semantic Contribution

A first line of work focuses on semantic contribution. Traditional measures of topical cohesion rely on lexical overlap across speakers, as in Group Communication Analysis (GCA) (Dowell et al., 2018b). While effective for tracking surface-level cohesion, these lexical measures often miss deeper semantic connections. Advances in sentence embeddings (Reimers and Gurevych, 2019) enable finer-grained assessment of topical alignment and informational novelty, providing a richer basis for modeling Relevance and Novelty in multi-party dialogue. Our work builds on this by applying SBERT to capture semantic contribution beyond exact keyword matching.

2.2. Functional Roles and Interactional Moves

A second line of research examines functional roles. Studies of sociocognitive roles in group problem solving (Dowell et al., 2018a) identify behaviors such as coordinating, elaborating, or challenging. These models highlight the importance of interactional cues—questions, invitations, confirmations—for interpreting how utterances facilitate group progress. Such work motivates our treatment of Enablement as a measurable conversational function. However, prior computational approaches often rely on rigid linguistic heuristics (e.g., counting question marks), which we critically evaluate against human and LLM judgments.

2.3. Affect and Decision Dynamics

A third area addresses affect and decision dynamics. Emotional tone influences cohesion, trust, and decision quality in small-group settings (Slater et al., 2017). In the Group Affect and Performance (GAP) corpus (Braley and Murray, 2018), utterances are annotated with affective and decision-relevant markers. This motivates our inclusion of Affect and Decision Proximity. Crucially, however, applying general-purpose sentiment analysis to task-oriented discussions remains challenging, as procedural confusion can often be misclassified as negative sentiment—a validity gap we investigate in this study.

2.4. Validating Automatic Measures with Humans and LLMs

Finally, NLP research has increasingly emphasized the need for validating automatic measures against human judgments. Work on evaluation metrics for summarization and dialogue systems has shown that automatic proxies often diverge from human assessments unless explicitly validated. Furthermore, recent work explores using Large Language Models (LLMs) as surrogates for human annotators (Chiang and Lee, 2023). This motivates our three-way validation strategy: rather than assuming that heuristic or embedding-based measures capture meaningful contribution, we directly compare MCS components with both human annotations and LLM-based judgments to establish **construct validity**.

3. Data

Our analyses use the **Group Affect and Performance (GAP) Corpus**, which contains 28 transcribed small-group discussions recorded at the University of the Fraser Valley (Braley and Murray, 2018). Each group completes the well-studied Winter Survival Task: participants first rank a set of survival items individually, then engage in a structured discussion to reach a collective ranking. The corpus includes time-aligned transcripts, speaker identities, and multiple individual- and group-level outcome measures such as influence, satisfaction, and agreement improvement. These features make the dataset suitable for studying both conversation dynamics and collaborative decision making.

In the midterm stage of this project, we computed the Meaningful Contribution Score (MCS) for every utterance across all 28 meetings and conducted exploratory analyses relating MCS to downstream outcomes. These corpus-wide computations serve as the foundation of our operationalization of MCS.

For the construct validity analysis conducted in this final paper, we additionally required fine-grained human judgments. To make this feasible while ensuring diversity of interaction patterns, we selected a subset of six meetings (Groups 4, 7, 11, 17, 18, and 26). These groups span the full range of the corpus's Agreement Improvement Index (AII), ensuring that both low- and high-performing discussions are represented. Two trained annotators labeled every utterance in these six meetings according to the five MCS components using a shared codebook. The resulting dataset provides the basis for our human-human reliability estimates and our evaluation of MCS and LLM-based annotations. This two-tier data design allows us to (a) retain the breadth of the full GAP corpus for computing and motivating the MCS framework, while (b) enabling detailed construct validity analyses on a carefully chosen, human-annotated subset.

4. Method

4.1. Meaningful Contribution Score (MCS)

The Meaningful Contribution Score (MCS) provides an utterance-level measure of how a turn contributes to group progress. It is defined as the product of a semantic contribution term (g_{sem}) and a contextual weighting term (f_{ctx}):

$$MCS(u_i) = g_{sem}(u_i) \times f_{ctx}(u_i) \tag{1}$$

The semantic term captures informational content, while the contextual term adjusts importance based on timing and affect. We describe the five components below.

Relevance. Relevance measures topical alignment between utterance embedding u_i and the task description vector t via cosine similarity:

$$Relevance(u_i) = \frac{u_i \cdot t}{||u_i|| \, ||t||} \tag{2}$$

Novelty. Novelty captures the introduction of new information, defined as the inverse of the maximum similarity to prior utterances:

$$Novelty(u_i) = 1 - \max_{j<i} \cos(u_i, u_j) \tag{3}$$

Enablement. Enablement reflects facilitative intent. Following Dowell et al. (2018a), we adopt a minimal heuristic baseline: an utterance receives a score of 1 if it contains a question or inclusive pronoun (e.g., "we"), and 0 otherwise.

Decision Proximity. This component assigns higher weight to utterances near decision moments d_j, defined as an inverse-distance decay:

$$DecisionProx(u_i) = \frac{1}{1 + \min_j \Delta t(u_i, d_j)} \tag{4}$$

Affect. Affect captures emotional intensity. Using a pretrained sentiment classifier, we define a threshold-based activation:

$$Affect(u_i) = \begin{cases} 1 & \text{if } |Sentiment(u_i)| > \tau \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

Aggregation. The semantic component sums the content dimensions (weighted by $\alpha_k = 1$), while the contextual term modulates them:

$$g_{sem}(u_i) = \alpha_1 Rel(u_i) + \alpha_2 Nov(u_i) + \alpha_3 En(u_i) \tag{6}$$

$$f_{ctx}(u_i) = (1 + \beta_1 DecProx(u_i)) \times (1 + \beta_2 Affect(u_i)) \tag{7}$$

Each participant's overall MCS is the mean of their utterance scores.

4.2. Human Annotation of MCS Components

To assess construct validity, two trained annotators independently labeled every utterance in six selected meetings (Groups 4, 7, 11, 17, 18, 26) along the five MCS dimensions. Annotations were assigned on a 3-point Likert scale using a shared codebook developed from prior literature. The codebook emphasizes intuitive judgments of each construct rather than surface lexical cues. Annotators viewed utterances in context; unassessable items were treated as missing. No adjudication was performed, allowing us to estimate inter-annotator reliability and use the labels as a ground-truth baseline.

4.3. LLM-Based Annotation

As a second reference point, we annotated the same utterances using Claude 3.5 Sonnet. The model was provided with the identical codebook used by human annotators and prompted to assign a 1-3 score based on the conversational context. Unlike the continuous algorithmic MCS, the LLM produces discrete ratings based on interpretive reasoning. This provides a complementary baseline to evaluate whether current LLMs align better with human intuition than heuristic measures.

4.4. Agreement and Validity Metrics

We evaluate reliability and validity using four complementary metrics across the three sources (Human, MCS, LLM). To assess surface-level consistency between annotators, we compute **Percent Agreement**. For rigorous inter-annotator reliability robust to missing data and ordinal scales, we report **Krippendorff's** α. To evaluate the alignment between continuous MCS scores and human Likert judgments, we compute **Pearson's r** (linear correspondence) and **Mean Absolute Error (MAE)** (magnitude of divergence). These metrics allow us to systematically compare human-human reliability against machine-human alignment.

5. RESULTS

Our primary objective is to evaluate the construct validity of the five MCS components. Establishing construct validity requires demonstrating that (i) humans can apply each construct consistently, (ii) the automatic MCS measure captures the same underlying construct humans intend to rate, and (iii) LLM-based annotations converge with human judgments where appropriate. We therefore present four analyses: human-human reliability, MCS-human alignment, LLM-human agreement, and three-way convergence among all methods. We additionally perform a qualitative error analysis to understand the sources of divergence. We conclude with a re-assessment of midterm outcome analyses in light of these validity results.

5.1. Human-Human Reliability

Table 1 reports inter-annotator agreement across the five MCS components, using four metrics: percent exact agreement, Krippendorff's α for ordinal labels, Pearson correlation, and sample size.

Three patterns are clear. First, Relevance is the only construct with meaningful human-human correlation ($r = 0.41$), even though exact agreement is modest. Second, Novelty and Enablement have high raw agreement but low α due to skewed label distributions, indicating partial but inconsistent shared understanding. Finally, Affect and Decision Proximity have extremely low

Table 1: Human-human reliability across MCS components. "Agree" = percent exact agreement; α = Krippendorff's coefficient; r = Pearson correlation.

Label	**n**	**Agree**	α	r
Relevance	1688	0.31	0.04	0.41
Novelty	1689	0.81	0.18	0.25
Enablement	1689	0.74	0.08	0.15
Affect	1689	0.31	0.01	-0.01
Decision Proximity	1689	0.48	0.05	0.08

reliability ($\alpha \approx 0$), suggesting that humans themselves do not apply these constructs consistently at the utterance level. This sets an upper bound on the construct validity any automatic method can achieve for these components.

5.2. Construct Validity of MCS

We next compare the continuous MCS scores with human ground-truth (GT) labels. GT is defined as the rounded mean of the two human labels when both are present. Table 2 reports correlations and Mean Absolute Error (MAE).

Table 2: Correlation and MAE between MCS and human ground-truth (GT).

Label	**n**	$r(MCS, GT)$	**MAE**
Relevance	1198	0.50	2.12
Novelty	1199	0.23	0.87
Enablement	1199	0.30	0.89
Affect	1199	0.02	1.78
Decision Proximity	1199	0.29	1.09

Key findings:

- **Relevance is a validated construct.** Its MCS-GT correlation ($r = 0.50$) is as high as human-human correlation, demonstrating that the embedding-based relevance measure captures what humans intend to rate.
- **Novelty, Enablement, and Decision Proximity show partial alignment.** Correlations in the 0.2-0.3 range indicate that the current heuristics capture some aspects of the intended constructs but miss more nuanced cues.
- **Affect shows no meaningful alignment with human judgment.** Near-zero correlation and high MAE, combined with low human reliability, indicate that the current sentiment-based MCS implementation does not reflect the affective judgments humans attempt to make.

5.3. LLM-Human Agreement

Table 3 reports agreement between the LLM's discrete 1-3 scores and human GT.

Insights:

- **Relevance:** LLM matches human reliability and is only slightly weaker than MCS.

Table 3: Agreement between LLM annotations and human GT labels.

Label	**n**	κ	$r(AI, GT)$	**MAE**
Relevance	1688	0.32	0.36	0.44
Novelty	1689	0.08	0.22	0.66
Enablement	1689	0.15	0.45	0.27
Affect	1689	0.06	0.06	0.42
Decision Proximity	1689	0.11	0.15	0.50

- **Enablement:** LLM aligns better with humans ($r = 0.45$) than MCS does ($r = 0.30$), suggesting that models infer pragmatic intent better than surface heuristics.
- **Novelty / Decision Proximity:** Modest correlations similar to MCS, indicating partial but incomplete construct capture.
- **Affect:** Extremely weak agreement, mirroring poor human reliability.

5.4. Three-Way Comparison: Human, MCS, and LLM

Table 4 shows pairwise correlations among all three systems.

Table 4: Pairwise correlations among human GT, MCS, and LLM labels. ($n = 1198$)

Label	$r(GT, MCS)$	$r(GT, AI)$	$r(MCS, AI)$
Relevance	0.50	0.35	0.29
Novelty	0.23	0.23	0.35
Enablement	0.30	0.39	0.40
Affect	0.02	0.08	0.23
Decision Proximity	0.29	0.14	-0.04

Takeaways:

- **Relevance is the most coherent construct.** All three methods converge to the same underlying signal.
- **Enablement:** LLM surpasses MCS in human alignment and correlates strongly with MCS, suggesting that richer operationalizations should replace simple lexical heuristics.
- **Affect:** Low convergence across all pairs confirms that the current affective operationalization is not theoretically or empirically sound.

5.5. Qualitative Error Analysis

To understand the quantitative discrepancies reported above, we conducted a qualitative inspection of utterances where the MCS diverged significantly from human and LLM judgments. Table 5 presents representative examples of these divergences.

The qualitative data reveals distinct failure modes. For Enablement, the heuristic approach proved brittle, missing explicit invitations for input (e.g., "Which one was yours?") that the LLM suc-

Table 5: Qualitative examples of divergence between Human, LLM, and MCS annotations. Scores are normalized where appropriate (MCS Enablement is binary 0/1; MCS Affect is sentiment score -1 to 1).

Construct	**Utterance**	**Human**	**LLM**	**MCS**	**Error Analysis**
Enablement	"Which one was yours?"	2.0	3.0	0.0	Heuristic Brittleness: The MCS heuristic failed to register this simple question, likely due to rigid keyword matching. The LLM correctly identified the facilitative intent.
Relevance	"No, I just think it's - I think it's fat."	2.5	2.0	-0.08	Context Blindness: The speaker refers to shortening as "fat." While relevant in context, the static embedding scored it low due to semantic distance from the "survival" topic vector.
Affect	"Oh, we have two sixes."	2.0	1.0	-0.99	Model Mismatch: The sentiment model flagged a procedural observation (checking numbers) as extremely negative, illustrating the noise introduced by generic sentiment tools.

cessfully captured. For Relevance, embedding-based scoring failed when relevance relied on dialogue history (e.g., implicit coreference to "shortening") rather than explicit keywords. Finally, the Affect component suffered from domain mismatch: the underlying sentiment model frequently misinterpreted procedural confusion or neutral observations as high-intensity negative sentiment (e.g., scoring "two sixes" as -0.99), thereby degrading the construct's validity.

5.6. Relation to Outcome-Based Analyses

The midterm report tested whether MCS predicts the Agreement Improvement Index (AII). Linear regression explained little variance ($R^2 = 0.018, p = 0.22$), while a nonlinear GAM indicated possible roles for Relevance and Enablement (pseudo-$R^2 = 0.57$). The present validity results explain why: the dimensions showing the strongest theoretical and empirical coherence (Relevance and Enablement) are also the ones that exhibited meaningful relationships with AII. Conversely, constructs with poor reliability (Affect, Decision Proximity) should not be expected to predict outcomes. Thus, outcome analyses should be interpreted as exploratory and subordinate to internal construct validity.

6. DISCUSSION

Our goal was to assess whether the five components of the Meaningful Contribution Score (MCS) behave as coherent constructs when compared to human and LLM judgments. The results reveal a mixed picture: while semantic alignment is solvable with current embeddings, interactional and affective dimensions require more sophisticated modeling than previously assumed.

6.1. Validated Components: Relevance

Relevance emerges as the most robust dimension. Human annotators show the highest mutual consistency on this construct, and the embedding-based MCS implementation correlates with human judgments ($r = 0.50$) at roughly the ceiling implied by human-human agreement. This supports treating topical alignment as a central and reliably measurable aspect of meaningful contribution using sentence embeddings.

6.2. The Case for LLM-based Measurement

Enablement appears to be a meaningful construct, but its current heuristic operationalization is limited. As shown in the qualitative analysis (Table 5), the keyword-based MCS misses context-dependent facilitative moves (e.g., "Which one was yours?") that LLMs and humans correctly identify. The strong correlation between LLM and human annotations ($r = 0.45$) suggests that facilitative behavior is best captured not by surface cues, but by models capable of pragmatic inference. Future iterations of MCS should likely replace heuristic counters with LLM-based classifiers for this dimension.

6.3. The Failure of Affect in Task Contexts

Affect performs poorly on all metrics. The qualitative analysis highlights the root cause: generic sentiment models misclassify procedural task talk (e.g., "we have two sixes") as negative emotion. Affective stance in collaborative decision-making is subtle and often neutral; applying off-the-shelf sentiment tools introduces significant noise. In future frameworks, we may consider exclude utterance-level affect or fine-tune models specifically on decision-making dialogues.

6.4. Limitations

Our findings should be interpreted in light of several limitations. First, our construct validity analysis relies on a small sample of six meetings ($N = 1,689$ utterances) from a single task domain (Winter Survival Task). While we prioritized depth of annotation over breadth, results may vary in open-ended or creative discussions. Second, while we employed two trained annotators, the low reliability for Affect suggests that the construct itself may be inherently ambiguous at the utterance level. Finally, our MCS implementation used fixed weights (Eq. 6); learning these weights from data could potentially improve alignment, though this risks overfitting given the limited dataset size.

6.5. Conclusion and Implications

This study provides the first systematic validity test of the Meaningful Contribution Score. We demonstrate that "contribution" is not a monolithic metric but a composite of semantic and functional signals. Our results suggest a path toward a hybrid evaluation framework: using efficient embedding models for semantic dimensions (Relevance, Novelty) while leveraging the reasoning capabilities of LLMs for pragmatic dimensions (Enablement). By validating these components against human judgment, we establish a grounded foundation for automated feedback systems that can support more effective group collaboration.

ACKNOWLEDGEMENTS

The authors would like to thank Professor David Banman and Naitian Zhou for their support and guidance throughout this research journey.

REFERENCES

[1] Matthew Braley and Gabriel Murray. 2018. The group affect and performance (gap) corpus. In *Proceedings of the Group Interaction: Frontiers in Technology (GIFT'18)*, pages 1-9. ACM.

[2] Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607-15631. Association for Computational Linguistics.

[3] N. M. Dowell, T. Nixon, and A. C. Graesser. 2018a. A computational linguistics approach for detecting sociocognitive roles in multi-party interactions. Retrieved from ResearchGate.

[4] Nia M Dowell, Trevor Nixon, and Arthur C Graesser. 2018b. Group communication analysis: A computational linguistics approach for detecting sociocognitive roles in multiparty interactions. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, pages 396-405.

[5] S. P. Lee, M. Perez, B. Burgess, and M. Worsley. 2018. Utilizing natural language processing (nlp) to evaluate engagement in project-based learning. In *Proceedings of the International Conference of the Learning Sciences (ICLS)*. International Society of the Learning Sciences.

[6] Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*.

[7] Carolyn Penstein Rosé, Yi-Chia Wang, Yue Cui, Jaime Arguello, and John Hopcroft. 2008. Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learning. *International Journal of Computer-Supported Collaborative Learning*, 3(3):237-271.

[8] S. Slater, J. Ocumpaugh, R. S. Baker, and S. Karumbaiah. 2017. Using natural language processing tools to develop complex models of student engagement. In *Proceedings of the Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 236-242. IEEE.

ANNEX A. HUMAN ANNOTATION CODEBOOK

The following codebook was used by human annotators to score utterances on a 3-point Likert scale across five dimensions.

1. Relevance

Definition: The extent to which the utterance contributes to the task goal or ongoing topic of discussion.
What to look for:

- Does it address the survival task?
- Does it relate to the item being discussed?
- Does it move the conversation forward?

Scr	Description	Examples
1	Off-topic, irrelevant, or social.	"By the way, what time is it?"
2	Vaguely related; general opinions.	"Hmm, I'm not sure, maybe it could matter."
3	Directly engages with task/item.	"The map should be higher because it helps us navigate."

Common pitfalls: Do not reward long utterances; only topicality matters. Neutral fillers are Low.

2. Novelty

Definition: Degree to which the utterance introduces new information not previously mentioned.
What to look for:

- Is this idea new relative to recent turns?
- Is it repetition or genuine addition?

Scr	Description	Examples
1	Repetition/agreement; no new content.	"Yeah, same as you said."
2	Minor extension or variation.	"Maybe it also helps a bit with signaling."
3	Clearly new idea or shift in thinking.	"The mirror is critical because aircraft can spot us faster."

Common pitfalls: New wording $\neq$ new content. Check last few utterances.

3. Enablement

Definition: How much the utterance facilitates, invites, or coordinates others participation.
What to look for:

- Does it invite others to speak?
- Does it clarify roles or structure the task?

Scr	Description	Examples
1	Self-contained; no attempt to involve others.	"I think the rope is useless."
2	Mildly collaborative; soft prompts.	"I guess... what do you think?"
3	Strong facilitation: invitations, coordination.	"Let's compare each item-who wants to start?"

Common pitfalls: Question form alone $\neq$ enablement. Short acknowledgments are Low.

4. Affect (Sentiment)

Definition: The emotional orientation: negative, neutral, or positive.
What to look for:

- Criticism or frustration?
- Approval or enthusiasm?

Scr	Description	Examples
1	Negative tone.	"No, that won't work at all."
2	Neutral, factual.	"The map is number five."
3	Positive tone.	"Yeah, that makes perfect sense!"

Common pitfalls: Judge valence only, not intensity.

5. Decision Proximity

Definition: Extent to which utterance contributes to finalizing a decision.
What to look for:

- Helping the group choose?
- Proposal or agreement?

Scr	Description	Examples
1	Not decision-related.	"Interesting point."
2	Indirect support.	"That might matter if rescue takes long."
3	Direct proposals / decisions.	"Let's put the lighter at #1."

Common pitfalls: Opinions are not always decision-related.

AUTHORS

Nanzheng Xie

Nanzheng Xie is a Master of Information Management and Systems (MIMS) candidate at the University of California, Berkeley. He returned to academia in 2025 after five years of professional experience spanning cybersecurity, data analytics, and critical infrastructure operations. His research interests focus on artificial intelligence, applied cybersecurity, and the analysis of conversational dynamics in socio-technical systems.

Ryuichi Ikeda

Ryuichi Ikeda is an MIMS candidate at the University of California, Berkeley, specializing in data science, generative AI, and product management. Prior to his current role, he served as a Business Lead at Mitsui & Co., focusing on global digital transformation (DX) and new ventures. His current research explores the application of machine learning and AI to optimize group interactions.

Suk Min Hwang

Suk Min Hwang is a graduate student at the University of California, Berkeley. With a professional background as a software engineer, his academic work primarily explores Human-Computer Interaction (HCI) research. He is actively involved in studying the intersection of technology and user experience, specifically focusing on collaborative problem-solving and conversational dynamics.

©2026 By AIRCC Publishing Corporation. This article is published under the Creative Commons Attribution (CC BY) license.