A Pilot Study Investigating Stakeholder Subjectivity in Collaborative Dialog Analysis

Ananya Ganesh[1,2], Martha Palmer[1], Katharina von der Wense[1]
[1]University of Colorado, [2]University of Wisconsin–Madison
Correspondence: aganesh27@wisc.edu

Abstract

Qualitative research in education relies on"ground truth" codes or labels generated by having a trained or expert coder code observations in data such as student dialog. Although rigorous validity checks are a part of the coding process, there is limited research investigating how and to what extent, this notion of the ground truth is influenced by inherent task subjectivity. This paper presents a pilot study of task subjectivity centered around the phenomenon of verbal off-task behavior. The context for this study is real-world small-group collaborative conversations among three to five students in a middle-school science classroom. To investigate how stakeholders such as teachers and students show subjectivity in approaching this task, we recruit five teachers from the Prolific online platform, and five students from local middle and high schools as annotators of off-task speech. We show that teachers, students, and expert coders differ in their perception of off-task speech, with some of these differences being systematic. Drawing upon recent research in machine learning and natural language processing, we then outline the potential benefits of collecting and modeling a *range* of codes that explicitly represent the subjective perspectives of a diverse set of coders.

Keywords: perspectivist approaches for social good, NLP for education, stakeholder subjectivity

1. Introduction and Background

Collaborative learning – the process through which multiple students come together to interactively learn a common topic – has been recognized as a critical component of modern classroom environments (Graesser et al., 2020). Apart from facilitating the socio-cognitive development of learners (Vygotsky, 1978), specific skills imparted to students through the collaborative learning process include argumentation (Osborne, 2010), discussion and negotiation (Roschelle, 1992), and shared knowledge construction (Bereiter, 2002). Towards understanding collaborative learning, learning analytics research has underscored the importance of students' dialogic interactions. However, manual qualitative analysis of student dialog poses a heavy demand on the time and labor of qualitative researchers, and is thus constrained by speed and scale. Consequently, there has been much interest in the development of computational models both to analyze collaborative dialog (Rosé et al., 2008; Yin et al., 2025) and to improve collaboration through automated interventions (D'Mello et al., 2024) based on such analysis.

Data-driven analysis and modeling of collaborative dialog is typically centered around "ground truth" observations of behaviors of interest. Examples of relevant behaviors include displays of collaborative problem solving skills such as *negotiation* with team members (Sun et al., 2020). Human coders or annotators are trained to identify such observations, typically from video data, but also from textual transcripts and audio recordings. The coding process comprises multiple stages of validation where inter-rater reliability is measured and the agreement between raters is iteratively improved through discussions and by refining the codebook based on nuances observed in the data (Reitman et al., 2023). The motivating factor in this process is the need to minimize disagreement, i.e., codes are thought to be more reliable when they are independently produced by multiple coders. Moreover, when the goal is to use the generated codes to train automated classifiers, the individual codes or labels may undergo further aggregation such as by selecting a single label through majority or ensemble voting.

However, several papers in machine learning (ML) and natural language processing (NLP) have argued that disagreements in annotation should be "embraced", that is, used in downstream analysis orin classification models, rather than pruned (Reidsma and op den Akker, 2008; Plank et al., 2014b; Alm, 2011). Arguments include better data utilization (Plank, 2022) (i.e., not discarding human feedback particularly with already scarce datasets), improved estimates of predictive uncertainty (Khurana et al.), the possibility of multiple valid answers (Alm, 2011), as well as the need to model diverse perspectives, since individual experiences may affect the way that text may be interpreted (Prabhakaran et al., 2021). Modeling diverse perspectives is also considered a step towards robust and reliable models that minimize bias (Kirk et al., 2024).

Subjective perspectives have been shown to depend on the socio-demographic backgrounds of annotators and their lived experiences. Waseem et al. (Waseem, 2016) show how hate speech annotations done by expert annotators who are

Proceedings of NLPerspectives @ LREC 2026, pages 136–144

activists differ from those done by crowdworkers. Age (Diaz et al., 2018) and gender (Biester et al., 2022) have also been shown to contribute to variance in judgments. Much of this work on subjectivity focuses on analyzing variations in third-party annotators such as crowdworkers, who typically have no direct involvement in the task being studied. Some exceptions include Arora et al. (Arora et al., 2020) – who recruit female journalists on Twitter who have been targets of abuse as annotators for hate speech data, and Patton et al. (Patton et al., 2019) – who show that members from groups discussed in tweets about gang activity annotate psycho-social attributes differently than social work students. In this work, we propose to study subjectivity from the perspective of *stakeholders* as they will be beneficiaries of similar automated systems based on data-driven models and as they tend to generate speech very similar to that observed in our data on a regular basis (i.e., classroom dialog).

To the best of our knowledge, there is no work that discusses annotator subjectivity as applied to collaborative dialog analysis or towards tools for instructional support. However, labels solicited from students through self-reports have been used successfully to model student affect (Broekens and Brinkman, 2013). (Zambrano et al., 2024) examined how supervised classifiers using self-reported labels vary from those using classroom observations, finding that both labels are useful in modeling different components of affect. Moreover, in learning sciences and education research, some work highlights how an ethnographic perspective should be adopted when coding discourse (Hennessy et al., 2020), which i) takes into account the socio-cultural setting where the conversations take place; ii) treats the observer (or annotator) as another source of influence on the "knowledge" that is produced during coding (Haraway, 1988; Gee and Green, 1998). Hennessy et al. (Hennessy et al., 2020) further discuss how coding schemes may be hard to adhere to for a coder who was not involved in the development of the scheme itself. Despite this observation, they discuss the importance of measuring reliability with multiple coders in order to share schemes for general use.

Given how achieving high reliability goes hand-in-hand with a rigorous annotation process where disagreements are discussed, the primary contribution of our study will be an analysis of disagreements resulting from the subjective perspectives of annotators in annotating classroom dialog. Specifically, we study the task of detecting students' *off-task* speech. Off-task behavior during collaborative learning has been the subject of study by learning scientists from multiple perspectives. Some work, such as Sabourin et al. (Sabourin et al., 2011) discusses the negative effect of off-task behavior on learning; more recent work by Langer-Osuna et al. (Langer-Osuna et al., 2018) discusses the impact of off-task speech on equity by serving as a mechanism for marginalized students to make bids for participation when their voices are neglected. Computational modeling of off-task speech has thus focused on both enabling further qualitative analysis (Ganesh et al., 2023) and on detection with the goal of adaptively scaffolding collaborative learning (Carpenter et al., 2020).

Judging whether an utterance is on-task or off-task naturally requires some subjectivity on the part of the labeler. Since data-driven tools for supporting collaborative learning will ultimately be used by the target audience of students and teachers, our main contribution is a pilot study focused on understanding how such potential stakeholders compare in their perceptions of off-task speech. We investigate the research question: *do teachers and students differ in their annotations when instructed to annotate classroom dialog for verbal off-task behaviors, and if so, in what way?* As our second contribution, we also discuss the potential benefits of explicitly representing subjectivity when developing applications for studying or supporting collaborative learning.

2. Methods

2.1. Data

We use the dataset described in Southwell, et al. (Southwell et al., 2022), which was shared with us for research purposes. The dataset consists of five-minute long transcripts of small-group discussions, collected from a middle-school science classroom in the United States. The subject of instruction is a curriculum unit called Sensor Immersion (SI) focused on "programmable sensor technology". Student interactions are recorded through desk-top mics, and manually transcribed and anonymized, yielding 27 transcripts with 1680 student utterances in total. The entire dataset was then labeled by trained annotators for whether or not each utterance is on-task, for the purposes of a separate standalone study on detecting off-task student speech. Each transcript was double-annotated, i.e., labeled by two annotators who have extensive experience in linguistic annotation tasks. The labels provided by the annotators included *on-task*, *off-task*, and *undecidable* given the context. When labeling an utterance, the annotators look at the entire text transcript to obtain contextual information, but due to data protection restrictions, the annotators do not have access to audio or video recordings. The annotators also have access to all the curriculum materials contained in the sensor immersion curriculum unit. Disagreements between annotators were

adjudicated using a third annotator, and inter-rater agreement was 0.647 (as measured by Cohen's kappa), indicating substantial agreement. The resulting labeled dataset is highly skewed, as the natural occurrence of off-task speech is lower than on-task speech, making up only a fourth of the dataset.

For the purpose of the pilot study, we scale down the dataset to four transcripts owing to time and cost constraints. Based on the viability of the pilot, we plan to extend the study on a larger dataset and more tasks in future work. The four transcripts we select have a good representation of both on-task and off-task utterances. We discard any unclear utterances that are transcribed only as `[inaudible]` or `[noise]`. We are left with 399 content-ful utterances in total, and the corresponding label distribution is shown in Figure 1. Any student names in the utterances are redacted, and the transcripts are completely anonymized.

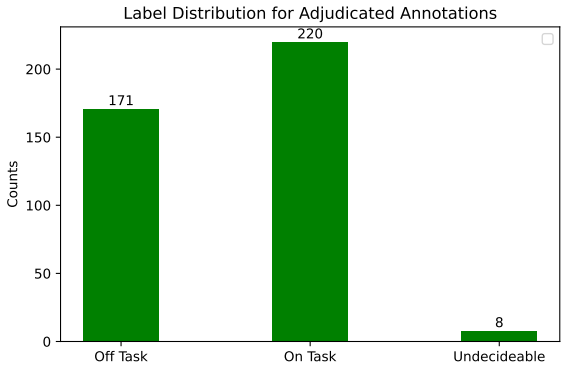

Figure 1: Distribution of the three labels for the on-/off-task classification task for the four transcripts used in the pilot study. The labels are from the adjudicated dataset after double-annotation by experts.

2.2. Participants

As mentioned above, our goal is to study annotations contributed by two types of stakeholders, namely students and teachers. Previous work in NLP that provide datasets with a range of annotations vary in the number of annotators that they include, with many using three (Plank et al., 2014b,a; Arora et al., 2020), some using three to five (Demszky et al., 2020), and some going up to 641 annotators (Sap et al., 2022). Based upon these references, we recruit five annotators each for the *student* and *teacher* group for a total of ten annotators. We share an annotation guideline with the participants that lists examples of on-task and off-task utterances and gives them instructions on how to make use of contextual information to handle ambiguous or edge cases. This annotation guideline is largely the same as that used by the expert annotators; however, unlike the expert annotators who were given access to the entire curriculum, in the interest of time, we provide a brief description of the topics seen in the transcripts and explicitly list keywords like micro:bit, sensor, Makecode, etc., in the annotation guideline.

Student recruitment: Since the conversations that we model are all from middle-school classrooms, we recruit student annotators from a similar age range. We advertise the study in community hubs around Boulder, Colorado and the surrounding areas and on social media. Our five participants are all between 12-14, with four of them from Boulder, and one from the California Bay Area. We collect annotations from each of the participating students through a single 90 minute one-on-one Zoom study. During the study, we share the annotation guideline containing the task description and labels, and ask them to code a small sample of five utterances while explaining their reasoning. We correct any misconceptions at this stage and then collect annotations for all transcripts, framing the task as answering *yes/no/I don't know* to the question *"Is the given utterance on-task?"*. We do not interfere or help the students when they start annotating the transcripts. At the end of the study, each participant is paid $37.5 (at $25 per hour) for their time. We note here that while the participants are students, and represent the perspectives of stakeholders of interest to us, they are not the same students whose speech was originally recorded in the data.

Teacher recruitment: To collect teacher annotations, we use a crowd-sourcing tool called Prolific[1] where domain experts who are verifiably employed in specific professions can be recruited. We recruit middle and high school teachers located in the US, and our five participants are all middle-aged teachers. We share the same annotation guideline that is shown to the students, and use the same five examples as a filter condition, i.e., if a candidate answers those questions incorrectly, their submissions are not accepted. Unlike the study with the students, the teachers submit their annotations entirely offline, although they have the option to contact us if needed. Compensation is again $25/hr per person.

2.3. Evaluation

Our focus at the evaluation stage is to compute agreement metrics that shed some light on whether teachers and students differ noticeably in their responses. We therefore report intra-group agreement, e.g., between all five students, as well as

[1] https://www.prolific.com/

inter-group agreement, e.g., between students and teachers.

In order to compute inter-group agreement, we perform label aggregation at the group level. For every utterance, we take the mode among all labels from a group (five in this case) and use that to represent the final group judgment, e.g., if four teachers answer *on-task*, and one answers *off-task*, the "teacher" label is considered to be on-task. We also use this aggregate to compare against our original expert-annotated labels.

We report percentage agreement for all sets of annotations. Since we have a group of five annotations, we report two percentages: i) *full agreement*, where all five annotators in a group agree on the label and ii) *all-but-one agreement*, where all but one of the annotators agree. The second metric provides some leeway for one of the annotators being a slight outlier, and has been reported in prior work as well (Demszky et al., 2020).

In addition to percentage agreement (which gives some insight into accuracy), we also report the Fleiss kappa to measure agreement between all five annotators in a group (henceforth denoted by κ). The Fleiss κ (Fleiss, 1971) is found to be more suitable than the Cohen's κ when the number of annotators is more than two. κ is from a scale of 0 to 1. The goal of the κ score is to shed light on whether the agreement is due to random chance or if it is due to a reliable overlap between the raters. While the interpretation of the score is always contextual due to the number of labels/raters, we follow the guideline of (Landis and Koch, 1977) to judge whether agreement is poor or good. They suggest that $\kappa > 0.61$ indicates substantial agreement, $0.61 < \kappa > 0.41$ indicates moderate agreement, and $0.41 < \kappa > 0.21$ indicates fair agreement.

Finally, we also compute and report statistics regarding the distribution of labels in each group. For a group, we compute this cumulatively: given five sets of annotations for four transcripts (total length 399), we combine all the labels such that we have a pool of 399 * 5 labels (i.e., 1995). The label distribution that arises from this will tell us about all the group members' judgments.

3. Results

3.1. Intra-Group Agreement

Table 1 shows the agreement between the five teachers and between the five students who participate in our study. Looking at the percentage agreements for teachers, we see that on slightly less than half the utterances, all the group members assign the same label to the utterance. When we look at only whether four of them agree, the agreement jumps to three-quarters of the dataset. The κ score of 0.419 indicates moderate agreement.

The agreement between students is slightly higher, with about half of the utterances receiving complete agreement, and almost 80% of utterances having at least four of the students providing the same annotation. Similar trends are indicated by the numerical value of κ, which, at 0.522 is higher for the student group compared to 0.419 for the teacher group. However, according to the guideline mentioned above, both these κ values indicate moderate agreement.

Table 1 also shows the agreement between experts on the subset of four transcripts used here. The κ value of 0.519 indicates moderate agreement which shows that this specific subset has lower agreement even among the experts than the entire dataset (on which agreement was substantial). However, since the number of expert annotators is fewer than either the teacher or student group, we do not directly compare the rates of expert agreement to the other two groups.

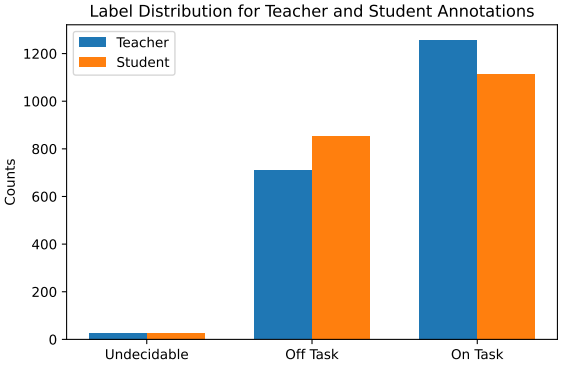

Figure 2: Distribution of labels across all four transcripts for all five annotators, shown for both the teacher group and the student group.

Next, we look at the label distribution on the cumulative labels of all annotators in a group, shown in Figure 2. We see that both groups use the *undecidable* label at equivalent rates. However, the student group uses the *off-task* label much more than the teacher group, indicating that students judge that utterances are *off-task* more often. while teachers may consider them to be *on-task*.

3.2. Inter-Group Agreement

As described in Section 2.3, we report inter-group agreement on the aggregated labels in Table 2. For each utterance, the teacher label represents the majority label out of all five teachers' individual labels, and similarly for the students. The expert label in this case refers to the adjudicated label from the dataset.

We see that the majority label chosen from the student group does have a very high overlap with

Metric	Teacher Agreement	Student Agreement	Expert Agreement
Full agreement	44.61%	51.13%	72.43%
All-but-one agreement	75.94%	79.70%	-
Fleiss κ	0.419	0.522	0.519

Table 1: Percentage agreement and Fleiss κ within the five teachers and within the five students. The last column reports agreement between the two expert annotators on the four transcripts alone.

Group	% Ag.	κ
Teacher–Student	91.73%	0.819
Teacher–Expert	86.72%	0.731
Student–Expert	89.47%	0.795
Teacher–Student–Expert	83.96%	0.782

Table 2: Inter-group agreement between student annotations, teacher annotations, and the expert annotations. We report both percentage agreement and Fleiss κ scores. Results are across every utterance in the dataset.

the majority label chosen from the teacher group – 91.73% of all utterances have the same majority labels, indicating "almost perfect" agreement with a κ value of 0.819 (Landis and Koch, 1977). Looking at how the majority labels from both groups compare to the expert annotator's labels, we note that the students agree with our expert annotators slightly more than the teachers do: the agreement between the student label and the expert label is 89.47% with $\kappa = 0.795$, and the agreement between the teacher label and the expert label is 86.72 with $\kappa = 0.731$. However both these κ values indicate substantial agreement. Overall, when we consider the agreement between the teacher, student and adjudicated label, we get a percentage agreement of 83.96% where $\kappa = 0.782$, once again indicating substantial agreement.

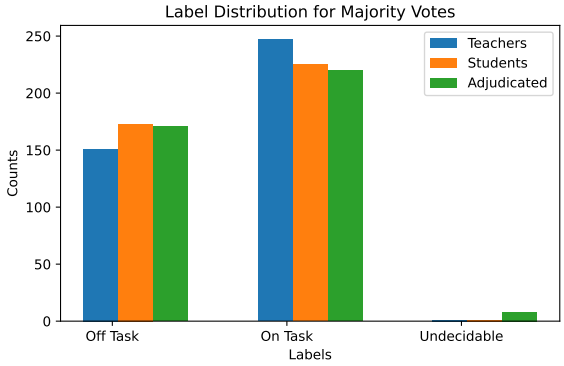

Figure 3: Label distributions of teacher, student and the expert labels. The teacher and student labels are given by the respective majority vote, and the expert label is given by the adjudicated label.

Figure 3 shows the majority label distribution of all three annotator groups, namely the teachers, students, and the expert annotators. We report results on every utterance in the dataset after group-level aggregation: the teacher and student group's label are their respective majority labels, and the expert labels are the adjudicated labels from the dataset. We first observe that the expert annotators resort to using the *undecidable* label more than the teachers or the students. We believe this could be explained by two reasons: i) Since there are only two expert annotators, the adjudicated label could be *undecidable* when only two people choose it. However, for the group annotations, the label could only be *undecidable* if a majority of the five choose it, typically at least three. ii) The expert annotators spend a longer time working with the transcripts, as they annotate 27 transcripts as compared to only four for the student and teacher groups. Thus they may be applying finer-grained judgments when distinguishing between the labels and are more sensitive to the highly ambiguous cases where a decision cannot be made given the information available.

Next, we observe that apart from the *undecidable* label, the label distribution of the student majority label shares more similarities with the expert annotator's label rather than the teacher. In comparison to the expert annotator and the students, the teachers have more of a tendency to label utterances as on-task rather than off-task. This is a surprising result as anecdotal evidence indicated that teachers may be 'stricter' or be more likely to conclude that speech is off-task. Although prior research hasn't investigated subjectivity in judging off-task classroom speech, some research has shown that experienced teachers exhibit more awareness of off-task behavior in classrooms than novice or "student" teachers (Wolff et al., 2017; Shinoda et al., 2021). Based on these findings, if the student can be considered as a less experienced judge, we would again expect the teachers to identify more utterances as off-task than on-task, which is not the case here.

Teacher–Student: On-Task–Off-Task	Teacher–Student: Off-Task–On-Task
"I think she's color blind."	"Oh my god, oh my god. oh my god. oopsie daisy."
"Oh that's just dope. I love that color."	"Dang it."
"Well whose fault is that?"	"Do do do do. Can't touch this. [noise]"

Table 3: Examples of utterances where students and teachers disagree. In the first column, the teacher label is on-task but the student label is off-task. In the second column, the opposite occurs.

3.3. Qualitative Analysis of Teacher/Student Labels

To get a closer understanding of stakeholder's subjectivity, we examine the pattern of disagreements between the majority labels of the teachers and the students. We first reiterate that agreement between teachers and students was almost perfect, with 91% of the utterances being labeled similarly by the teachers and students. The remaining 9% comprises 35 utterances. Of these, 28 of the utterances (80%) are cases where the teacher labels an utterance as on-task and the student labels the utterance as off-task, and only 13 utterances are cases where the teacher label is off-task and the student label is on-task.

Table 3 shows some examples of utterances where disagreements occur. The utterances where teachers judge an utterance as on-task but students do not are instances of students engaging in a minor aside in the middle of a problem-solving interaction, such as admiring the color of an artifact they produced on screen. Interestingly, the comment "I think she's color blind" does not seem to be made rudely; the surrounding context reveals that the student is indeed color blind and needs help distinguishing wires, but a student annotator may have interpreted it to be a disparaging comment, whereas the teacher annotators interpret it differently. The cases where students label utterances as on-task but teachers do not, include minor swear words such as "dang it" as well as effusive or repetitive.

4. Discussion

The research question that we investigated through this study was to investigate if teachers and students differ in their labeling of classroom dialog, and if so, in what ways. The high agreement (κ > 0.8) that we observed when comparing majority labels indicates that there is no statistically significant difference in teacher and student annotations. However, we do find that i) students show higher in-group agreement than teachers ii) students and expert annotators recognize off-task utterances with a higher frequency than teachers. Moreover, the disagreements appear to be *systematic* and not random, since there is a pattern of teachers being more lenient with minor asides, while students strictly judge these utterances as off-task.

Applications of stakeholder perspectives in classrooms The first implication of our study is on the notion of a "ground truth", particularly when the ground truth data comes from technologists who are building the tools and not from the stakeholders who may be using them. In this case, if an automated off-task utterance classifier was to be a part of a teacher-facing real-time learning analytics dashboard to support small-group discussions, the teacher's preferences may differ from the classifier's learned representations, resulting in false positives. This is also the case if an underlying model is expected to serve multiple stakeholders or applications, such as providing analytics to both teachers and qualitative researchers. Collecting a range of annotations that go beyond experts and include contributions from stakeholders could therefore be useful in modeling different preferences for different application.

Implications on reliability and trust The other benefit in using a range of annotations is from a reliability perspective when deploying machine learning models in the high-stakes environments of classrooms. Discriminative classifiers' estimate of the likelihood of each class can be obtained as a probability score, referred to as the model's uncertainty. In designing dashboards or in instructional tools that intervene based on a students' detected state, these probabilities are used to decide a threshold before an action is taken. However, research has shown that machine learning models, particularly, deep neural-network based models are not well-calibrated and do not produce uncertainty estimates that accurately reflect the true likelihood of an event (Ovadia et al., 2019). Typically, this manifests as overconfidence, especially towards the majority label or class. The calibration of models can be improved by using a range of labels (Prabhakaran et al., 2021; Khurana et al.). One mechanism is through multi-annotator models, as demonstrated by Davani et al. (Davani et al., 2022): by treating each individual annotator's data as a separate task, they construct a multi-tasking model that outputs multiple predictions, which can then be aggregated. The resulting uncertainty estimate

is shown to be a better measure of disagreements between annotators than if the labels were aggregated prior to model training.

Through this pilot study, we showed that subjectivity, specifically from the perspective of stakeholders, *is* a factor when coding classroom dialog. We advocate for the collection and release of a wide range of annotator-level labels to facilitate the creation of reliable models that incorporate stakeholder judgments. In the future, we will extend this study to a larger dataset as well as investigate different behaviors, such as CPS skills.

5. Limitations

The central goal of our study is to examine stakeholder subjectivity in annotating classroom dialog. However, despite our pilot study showing some differences in the way students and teachers perceive off-task dialog, the strength of our conclusions is limited by scope as we only look at five students and five teachers. A broader study that includes a larger pool of participants is essential before we can make strong recommendations for the design of learning analytics systems or interventions based on the phenomenon of stakeholder subjectivity. Moreover, for greater ecological validity, both the teachers and students must share context (such as being from the same school) with the students who generate the utterances represented in our data – which is a setting that we were unable to demonstrate in our study due to unavailability of the original classroom participants.

6. Acknowledgments

We thank the reviewers for their time and helpful feedback. We also thank all the student participants and teacher participants who provided their responses on this study. This study was approved by the University of Colorado's Institutional Review Board under protocol #25-0253. This research was supported by the NSF National AI Institute for Student-AI Teaming (iSAT) under grant DRL 2019805 and under grant DRL 1920510. The opinions expressed are those of the authors and do not represent views of the NSF.

7. Bibliographical References

Cecilia Ovesdotter Alm. 2011. Subjective natural language problems: Motivations, applications, characterizations, and implications. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 107–112.

Ishaan Arora, Julia Guo, Sarah Ita Levitan, Susan McGregor, and Julia Hirschberg. 2020. A novel methodology for developing automatic harassment classifiers for Twitter. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 7–15, Online. Association for Computational Linguistics.

Carl Bereiter. 2002. *Education and mind in the Knowledge Age.* Education and mind in the Knowledge Age. Lawrence Erlbaum Associates Publishers, Mahwah, NJ, US. Pages: xiii, 526.

Laura Biester, Vanita Sharma, Ashkan Kazemi, Naihao Deng, Steven Wilson, and Rada Mihalcea. 2022. Analyzing the effects of annotator gender across nlp tasks. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP@ LREC2022*, pages 10–19.

Joost Broekens and Willem-Paul Brinkman. 2013. Affectbutton: A method for reliable and valid affective self-report. *International Journal of Human-Computer Studies*, 71(6):641–667.

Dan Carpenter, Andrew Emerson, Bradford W. Mott, Asmalina Saleh, Krista D. Glazewski, Cindy E. Hmelo-Silver, and James C. Lester. 2020. Detecting off-task behavior from student dialogue in game-based collaborative learning. In *Artificial Intelligence in Education*, pages 55–66. Springer.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.

Mark Diaz, Isaac Johnson, Amanda Lazar, Anne Marie Piper, and Darren Gergle. 2018. Addressing age-related bias in sentiment analysis. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–14, New York, NY, USA. Association for Computing Machinery.

Sidney K D'Mello, Nicholas Duran, Amanda Michaels, and Angela EB Stewart. 2024. Improving collaborative problem-solving skills via automated feedback and scaffolding: a quasi-experimental study with cpscoach 2.0. *User modeling and user-adapted interaction*, 34(4):1087–1125.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Ananya Ganesh, Michael Alan Chang, Rachel Dickler, Michael Regan, Jon Cai, Kristin Wright-Bettner, James Pustejovsky, James Martin, Jeff Flanigan, Martha Palmer, et al. 2023. Navigating wanderland: Highlighting off-task discussions in classrooms. In *International Conference on Artificial Intelligence in Education*, pages 727–732. Springer.

James Paul Gee and Judith L Green. 1998. Chapter 4: Discourse analysis, learning, and social practice: A methodological study. *Review of research in education*, 23(1):119–169.

Arthur C Graesser, Samuel Greiff, Matthias Stadler, and Keith T Shubeck. 2020. Collaboration in the 21st century: The theory, assessment, and teaching of collaborative problem solving.

Donna Haraway. 1988. Situated knowledges: The science question in feminism and the privilege of partial perspective. *Feminist Studies*, 14(3):575–599.

Sara Hennessy, Christine Howe, Neil Mercer, and Maria Vrikki. 2020. Coding classroom dialogue: Methodological considerations for researchers. *Learning, Culture and Social Interaction*, 25:100404.

Urja Khurana, Eric Nalisnick, Antske Fokkens, and Swabha Swayamdipta. Crowd-calibrator: Can annotator disagreement inform calibration in subjective tasks? In *First Conference on Language Modeling*.

Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, et al. 2024. The prism alignment project: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. *arXiv preprint arXiv:2404.16019*.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Jennifer Langer-Osuna, Emma Gargroetzi, Rosa Chavez, and Jen Munson. 2018. Rethinking loafers: Understanding the productive functions of off-task talk during collaborative mathematics problem-solving. International Society of the Learning Sciences.

Jonathan Osborne. 2010. Arguing to learn in science: The role of collaborative, critical discourse. *science*, 328(5977):463–466.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. 2019. *Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift*. Curran Associates Inc., Red Hook, NY, USA.

Desmond Patton, Philipp Blandfort, William Frey, Michael Gaskell, and Svebor Karaman. 2019. Annotating social media data from vulnerable populations: Evaluating disagreement between domain experts and graduate student annotators.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014a. Learning part-of-speech taggers with inter-annotator agreement loss. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 742–751, Gothenburg, Sweden. Association for Computational Linguistics.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014b. Linguistically debatable or just plain wrong? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 507–511, Baltimore, Maryland. Association for Computational Linguistics.

Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Díaz. 2021. On releasing annotator-level labels and information in datasets. In *Proceedings of the Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 133–138.

Dennis Reidsma and Rieks op den Akker. 2008. Exploiting 'subjective' annotations. In *Coling 2008: Proceedings of the workshop on Human Judgements in Computational Linguistics*, pages 8–16, Manchester, UK. Coling 2008 Organizing Committee.

Jason G Reitman, Charis Clevenger, Quinton Beck-White, Amanda Howard, Sierra Rose, Jacob Elick, Julianna Harris, Peter Foltz, and Sidney K D'Mello. 2023. A multi-theoretic analysis of collaborative discourse: A step towards ai-facilitated

student collaborations. In *International Conference on Artificial Intelligence in Education*, pages 577–589. Springer.

Jeremy Roschelle. 1992. Learning by collaborating: Convergent conceptual change. *The journal of the learning sciences*, 2(3):235–276.

Carolyn Rosé, Yi-Chia Wang, Yue Cui, Jaime Arguello, Karsten Stegmann, Armin Weinberger, and Frank Fischer. 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.

Jennifer Sabourin, Jonathan P. Rowe, Bradford W. Mott, and James C. Lester. 2011. When Off-Task is On-Task: The Affective Role of Off-Task Behavior in Narrative-Centered Learning Environments. In *Artificial Intelligence in Education*, pages 534–536, Berlin, Heidelberg. Springer Berlin Heidelberg.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.

Hirofumi Shinoda, Tsuyoshi Yamamoto, and Kyoko Imai-Matsumura. 2021. Teachers' visual processing of children's off-task behaviors in class: A comparison between teachers and student teachers. *PLoS One*, 16(11):e0259410.

R. Southwell, S. Pugh, E.M. Perkoff, C. Clevenger, J. Bush, and S. D'Mello. 2022. Challenges and feasibility of automatic speech recognition for modeling student collaborative discourse in classrooms. In *Proceedings of the 15th International Conference on Educational Data Mining*. International Educational Data Mining Society.

Chen Sun, Valerie J Shute, Angela Stewart, Jade Yonehiro, Nicholas Duran, and Sidney D'Mello. 2020. Towards a generalized competency model of collaborative problem solving. *Computers & Education*, 143:103672.

Lev S Vygotsky. 1978. *Mind in society: The development of higher psychological processes*, volume 86. Harvard university press.

Zeerak Waseem. 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science*, pages 138–142.

Charlotte E Wolff, Halszka Jarodzka, and Henny PA Boshuizen. 2017. See and tell: Differences between expert and novice teachers' interpretations of problematic classroom management events. *Teaching and teacher education*, 66:295–308.

Stella Xin Yin, Zhengyuan Liu, Dion Hoe-Lian Goh, Choon Lang Quek, and Nancy F. Chen. 2025. Scaling up collaborative dialogue analysis: An ai-driven approach to understanding dialogue patterns in computational thinking education. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, LAK '25, page 47–57, New York, NY, USA. Association for Computing Machinery.

Andres Felipe Zambrano, Nidhi Nasiar, Jaclyn Ocumpaugh, Alex Goslen, Jiayi Zhang, Jonathan Rowe, Jordan Esiason, Jessica Vandenberg, and Stephen Hutt. 2024. Says who? how different ground truth measures of emotion impact student affective modeling. In *Proceedings of the 17th International Conference on Educational Data Mining*, pages 211–223.