Feedback on Feedback: Comparing Classic Natural Language Processing and Generative AI to Evaluate Peer Feedback

Stephen Hutt
University of Denver
stephen.hutt@du.edu

Allison Depiro
CueThink
adepiro@cuethink.com

Joann Wang
CueThink
jwang@cuethink.com

Samuel Rhodes
Georgia Southern University
srhodes@georgiasouthern.edu

Ryan S. Baker
University of Pennsylvania
rybaker@upenn.edu

Grayson Hieb
University of Denver
grayson.hieb@du.edu

Sheela Sethuraman
CueThink
sheela@cuethink.com

Jaclyn Ocumpaugh
University of Pennsylvania
ojaclyn@upenn.edu

Caitlin Mills
University of Minnesota
cmills@umn.edu

ABSTRACT

Peer feedback can be a powerful tool as it presents learning opportunities for both the learner receiving feedback as well as the learner providing feedback. Despite its utility, it can be difficult to implement effectively, particularly for younger learners, who are often novices at providing feedback. It can be difficult for students to learn what constitutes "good" feedback – particularly in open-ended problem-solving contexts. To address this gap, we investigate both classical natural language processing techniques and large language models, specifically ChatGPT, as potential approaches to devise an automated detector of feedback quality (including both student progress towards goals and next steps needed). Our findings indicate that the classical detectors are highly accurate and, through feature analysis, we elucidate the pivotal elements influencing its decision process. We find that ChatGPT is less accurate than classical NLP but illustrate the potential of ChatGPT in evaluating feedback, by generating explanations for ratings, along with scores. We discuss how the detector can be used for automated feedback evaluation and to better scaffold peer feedback for younger learners.

CCS CONCEPTS

• **Human-centered computing**; • **Applied computing** → Education; • **Computing methodologies** → Artificial Intelligence; Natural language processing; Machine learning;

KEYWORDS

Peer Feedback, Language Analytics, Natural Language Processing, Generative AI, Large Language Models

ACM Reference Format:
Stephen Hutt, Allison Depiro, Joann Wang, Samuel Rhodes, Ryan S. Baker, Grayson Hieb, Sheela Sethuraman, Jaclyn Ocumpaugh, and Caitlin Mills. 2024. Feedback on Feedback: Comparing Classic Natural Language Processing and Generative AI to Evaluate Peer Feedback. In *The 14th Learning Analytics and Knowledge Conference (LAK '24), March 18–22, 2024, Kyoto, Japan.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3636555.3636850

LAK '24, March 18–22, 2024, Kyoto, Japan
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1618-8/24/03.
https://doi.org/10.1145/3636555.3636850

1 INTRODUCTION

Feedback is essential to successful learning and instruction. A considerable amount of scholarly work has considered what makes feedback effective and how it can be improved [47]. Feedback should be a learner-centered process [37] that helps students enhance their learning skills and encourages them to be active learners [46]. Much of the recent feedback literature has focused on instructors providing feedback to students [27, 43], or learning systems automatically providing feedback [29]. There is also a growing body of work considering peer feedback, where learners provide feedback to other learners. The effect of peer feedback is two-fold: 1) learners receive feedback from their peers, which they can use as much as they would use feedback from an instructor [28], and 2) learners provide feedback to others, developing their metacognitive skills in evaluating their own work as well as important communication and teamwork skills [7].

In order to provide feedback, the learner must first use their own internalized set of standards to evaluate the quality of their peers' work. This process allows students to develop a better understanding of the assessment process and criteria and, in turn, can improve self-assessment skills [38, 39], which are critical to many self-regulated learning practices [1, 45]. Some studies have shown that *giving* feedback is just as effective, or more so, for improving a learner's understanding and communication of a topic than receiving feedback (see review in [17]). Given the benefits of both giving and receiving feedback, incorporating peer-based feedback into learning technologies could significantly improve learning, while also allowing for increased scalability. However, it is not yet clear how that technology might teach students to deliver "good" feedback – that is, feedback that is beneficial to both the provider and, ultimately, to the recipient. This may be especially true for younger learners, who may be novices at providing feedback and may not yet know the elements of "good" feedback.

The question is then how to give feedback about "good" peer feedback, such that students are not only mastering content, but also developing skills for providing helpful feedback to their peers. In this paper, we take a step towards automated "feedback on feedback" within an existing learning platform with a peer review element. Helpful "feedback on feedback" would likely need to be automated and given in real-time so that students can revise their feedback in a timely fashion. This goal would, however, first require us to make automated assessments about a peer's feedback in the moment– an objective we tackle in the current paper.

In the current study, we hand-coded peer feedback comments from an online mathematics platform, where students provide comments on their peers' problem-solving approach. We then compared two methods to provide ratings of peer feedback that could, in turn, inform automated feedback. First, we used "classic" natural language processing (including tokenization and parts of speech tagging) and supervised machine learning to develop an automated coding process for student comments. We then interrogated the model through feature analysis to derive the most important features of successful comments in order to provide more actionable insights for either students or teachers.

Second, we explore advances to large language models (specifically ChatGPT 4) for the same task. Whereas the "classic" approach dissects student comments at a granular level, identifying patterns in successful feedback, ChatGPT offers the possibility of a more holistic approach, where instead of breaking comments down, it assimilates information in its entirety and provides ratings aligned with a defined (by us) rubric. To translate ratings to actionable feedback for this model (as feature analysis is not possible), we include in the ChatGPT prompt a request to include reasoning for its rating, providing an interpretable message that can inform student feedback. Finally, to check for the generalizability of our models to score peer feedback, we applied both models to a held-out test set (i.e., unseen data) and compared automated scores to external measures to test for evidence of convergent validity, including measures of executive function, metacognition, and anxiety.

2 RELATED WORK

2.1 Peer Feedback

Providing peer feedback offers a number of potential learning opportunities for the learner providing the feedback [36]. In interviews with 15 learners, Ertmer and colleagues (2007) show that through providing peer feedback, students reflected more critically and thoroughly on their own work. Students in this study also reported that providing feedback to their peers also strengthened their internal representations of the problem. Similarly, a study with 143 Computer Science undergraduates [24] showed that students engaging in peer feedback developed analytical skills that, though initially applied to their peers' work, were then also applied to their own work. Though students often have strong feelings about feedback (i.e., liking or disliking it), they may not know the difference between effective and ineffective feedback [20]. As such, methods to guide learners in providing effective peer feedback could support both learners involved in the interaction.

2.2 Automated Feedback Evaluation

Automated evaluation of feedback has been considered by a number of scholars, often with older students [22, 32]. This process typically has involved using natural language processing techniques (similar to approaches in the current research) to extract features from text feedback. These features can then be used in conjunction with human coding or supervised machine learning, to provide insights into feedback content [22] and quality [4, 32].

Thematic analysis of feedback is typically most useful to the person receiving the feedback; by providing a summary of feedback, the receiver of the feedback can identify broad areas that need addressing. Analysis of feedback content, however, can break it down into its component parts and provide more useful information for the person providing the feedback. For example, [4] use random forest classifiers to identify the presence of good feedback elements within larger blocks of text from an instructor (feedback to students). Features were extracted primarily using the Linguistic Inquiry and Word Count (LIWC) [35] and Coh-metrix [15] tools. A follow-up study extended this work [3] using a similar approach. In both studies, one classifier was trained per feedback component (a total of seven components) that could then be interpreted for overall feedback quality and areas of improvement.

Work by Osakwe and colleagues [32] also used the text mining tools LIWC and Coh-Metrix to investigate instructor feedback quality across datasets that included feedback in two different languages (Portuguese and English). This study showed good classification of feedback elements within language (e.g., a model training on English feedback, performed well on data of the same language). The study also investigated the transferability of the NLP features between languages; however, models did not generalize between languages, demonstrating the importance of context, and a potential limitation of this approach. To date, much of the work on automated evaluation of feedback quality has focused on instructor feedback rather than peer feedback. There has also been limited work considering automated feedback evaluation (either instructor or peer) in mathematics [33]. One exception to this is [50], which uses bag of words models, sentence embeddings and parts of speech tagging to detect individual components of effective feedback in mathematics. Their results indicated that sentence embeddings were most successful for predicting elements of good peer feedback, but this work did not provide an "overall" evaluation.

The recent developments (and increase in public availability) of Large Language Models (LLMs) such as ChatGPT [31] has allowed for the automation of coding of text. Though not yet applied to student peer feedback in published work, LLMs have already been shown in recent results to outperform humans in coding of text data [14]. Due to their large training corpus, intense training process, and highly complex underlying mechanisms (GPT 4 currently claims over 1.7 trillion internal parameters), LLMs have the potential to outperform more specialized methods that are trained on much smaller corpora. Similarly, LLMs can discern semantic relationships among words and concepts, capturing complex linguistic patterns that are difficult for human beings to identify rationally through regular expressions [48]. Furthermore, ChatGPT has an advantage over other automated detection methods in that it is designed to interact with human beings, rather than simply deliver a code or response. This ability has the potential to make the output of an LLM-based automated feedback evaluation process more interpretable to both students and teachers without the need for complex feature analysis and subsequent interpretation. To our knowledge, no one has yet attempted to derive learning analytics regarding peer feedback quality using LLMs – an objective we tackle in the current paper.

2.3 Components of Effective Feedback

The literature identifies good feedback as that which allows the receiver to improve their self-regulation [30], and further describes it based on three characteristics (Informativeness, Polarity, and Timelines), which we examine below with an eye toward how it can be automatically extracted from text data via natural language processing techniques.

Two highly context-dependent elements of feedback are polarity and timeliness [41]. Polarity refers to the tone of the feedback, whether it is reinforcing good practice, or correcting a mistake. Timeliness, meanwhile, refers to when the feedback is received, for example, in the moment, end of class, or the following day. Both of these components vary by context, and there is no clear connection between these constructs and feedback quality (i.e., not all in-the-moment feedback is good, depending on the student and context). These elements are thus challenging for automated evaluation from text alone. Positive and negative words can be analyzed and extracted through parts of speech tagging (including negated positive words), but how that translates to feedback quality is less clear. For example, students' motivation level might impact how they respond to either a corrective or reinforcing comment (polarity) [18], which cannot be extracted from the text. Additionally, assessing the timing of feedback requires data beyond just text data and information regarding prior knowledge and the task context [13, 17].

Informativeness refers to the extent to which the comments contain both feed-back and feed-forward components [11, 17] and can be easier to asses from text data alone. Feed-back elements communicate to a learner their progress toward a certain goal or expected outcome [30]. These elements can also reinforce successful strategies or methods used towards the goal. Meanwhile, feed-forward elements direct the learner on possible future behaviors [17]. This may include suggesting strategies or alternate approaches. By combining information on current progress and future directions, feedback comments are informative to the learner and support them in progressing towards the objective or goal. In this paper, we focus on elements of informativeness in our coding scheme and detectors.

2.4 Current Study and Novelty

This study includes data from 203 6-8 graders using CueThink, an online learning environment designed to develop student problem-solving skills in middle school mathematics. As students use the system, they interact with their peers, providing feedback on each other's solutions. Using some of the components of effective feedback described above, we develop a coding scheme for evaluating peer feedback comments in a mathematics problem-solving setting. We then answer the following three research questions:

RQ1. How effective is the combination of natural language processing and supervised machine learning for automated feedback evaluation for younger learners (6th-8th grade)?

RQ2. How do large language models, such as ChatGPT compare for evaluating peer feedback and providing feedback on feedback?

RQ3. How do model outputs correlate with external measures? Through answering these research questions, we tackle the problem of providing automated "feedback on feedback" for younger learners who are likely novices at providing feedback to others. The work serves as an initial step towards providing scaffolds and advice for learners as they produce peer feedback and develop their own self-regulated learning skills.

3 DATA COLLECTION AND CODING

In this section, we describe the learning environment used in this work, along with the data collection and data coding process. The codes derived are then used as ground truth values in the machine learning process discussed in the following section.

3.1 Online Learning Platform

CueThink [6] is a digital learning application that focuses on enhancing middle school math problem-solving skills, encouraging students to engage in self-regulated learning and develop math language to communicate problem-solving processes. The platform asks students both to solve a math problem and to create a shareable screen-cast video that provides the student's solution and demonstrates their problem-solving process. CueThink structures a problem into a Thinklet, a process that includes four phases—Explore, Plan, Solve, and Review—that closely align with Winne & Hadwin's model of SRL [45]. A full description of the CueThink phases can be seen in [49]

Once students have completed the problem-solving process and recorded their solution, their video is shared with their class for peer review (following the "review" phase, where students review their own work). In this process, teachers and peers annotate both the textual responses and video, often asking the student for their underlying reasoning or why the student picked specific methods. These annotations are then sent back to the video's author for possible revision. These annotations are the focus of the current analyses.

3.2 Developing the Codebook

Our coding scheme was developed using 116 peer-review comments from 87 students. The code development process then followed a recursive, iterative process [44], including conceptualization of codes, generation of codes, refinement of the first coding system, generation of the first codebook, continued revision and feedback, coding implementation, and continued revision of the codes [12].

We initially created a four-tiered code for feedback robustness – combining elements of "feed-back" and "feed-forward" [17, 30] along with elements of informativeness, to create one measure of feedback that can be used in this method [17]. In designing one coherent metric from the literature regarding feedback, we could not include all components, therefore we prioritized the elements we deemed to be most relevant to peer feedback. "Feed-up" elements were excluded as these relate to students' understanding of the overall learning goal; peers are not typically expected to provide feedback on this (instead, instructors do). Feedback tiers ranged from tier 1, which was the least robust, and used for non-helpful, vague, or generic comments that could be applied in any feedback situation, to tier 4, the most robust category with evidence of peer-to-peer learning. Three coders then coded the same 60 annotations individually and came together to discuss agreements. During this process, we observed high variance in Tier 2, with two distinct groups, with different levels of specificity. We chose to split the tier to capture this difference, yielding our final coding scheme with 5 tiers. A further round of coding agreement was performed with the updated codebook, and example annotations were added. The final list of codes, along with examples, are in Table 1.

3.3 External Measures

Though our primary goal is to develop automated detectors relative to our code book, we also evaluated the convergent validity of our detector (i.e., does it correlate with other constructs that we would expect it to?). This was done through comparison to external measures, including survey measures of math anxiety [2], executive functions [42], metacognitive awareness [40], and beliefs on problem-solving [21]. For example, for purposes of convergent validity, we would expect that students who give better feedback (according to the detector) would be more likely to have lower math anxiety scores and higher metacognitive beliefs.

The modified Abbreviated Math Anxiety Scale [2] uses a two-factor structure, which results in two subscales learning math anxiety (Learning subscale), and math evaluation anxiety (Evaluation subscale) [2, 19]. The scale was developed for math learners between 8 and 13 years old (i.e., overlapping our research sample) and includes a 9-item self-report of math anxiety that are averaged to produce one final scale for analysis.

Adaptive Cognitive Evaluation (ACE) is a series of 15 measures of executive function, implemented in an interactive game environment. Because the environment is adaptive, it can be used for a wide variety of ages and is recommended for any age group from 7 up. Students completed a subset of the 15 measures, in the online testing environment. Once logged in, students were allowed to progress at their own pace, as with the other measurements.

The Junior Metacognitive Awareness Inventory (JrMAI) is a measure of metacognitive and cognitive strategies applied by learners [40] that was developed for students in grades 6 through 12. The current study uses a 9-item abbreviated version of the JrMAI that was validated with student in grades 6 -8 [16]. The measure includes 5 items identified as regulation of cognition and 4 items identified as knowledge of cognition.

The Indiana Mathematics Belief Scales [21] prompts students regarding their beliefs about mathematics and mathematical problem-solving over a 36-item survey. The measure is divided into six subscales. In this work, we administered 3 of the 6 subscales. Specifically, we administered shortened versions of three subscales, which measure (a) student beliefs that they can solve time consuming math problems (5 items), (b) student beliefs that effort increases ability in mathematics (3 items), and (c) student beliefs that math is useful in their real lives, respectively (3 items). The shortened versions were validated using middle school students.

Each of these measures were combined into a pre-test that was delivered in three sections. Each section was delivered by the students' math teacher and at times convenient with other classroom instruction. Teachers were given a two-week window to complete the tests. The three sections could be delivered in any order, and over multiple class periods (e.g., sections 1 in one class, and sections 2 and 3 in the next class).

Once the pre-test was complete, students were provided with access to the learning platform. Their interaction with the platform was integrated into their regular classroom instruction, with their teacher assigning problems for them to complete. Students used the learning platform for the majority of the academic year. However, this was alongside a variety of other instruction, so on average, students completed 3 problems during the year. Each problem took an average of 1.8 hours. A post-test (identical to the pre-test) was also conducted following students' use of the platform, but this data is not analyzed in the current work.

4 SUPERVISED MACHINE LEARNING – RQ1

We next developed supervised machine learning models using the same peer feedback comments used to develop the codebook (N=116), the first approach we investigate in this paper. This process aimed to develop automated evaluation of the peer feedback comments, using the human codes from our coding scheme described above as the ground truth. We first tokenized each peer-review comment and then extracted features for each comment using the nltk package in python [25]. We then recorded the length of the comment with and without stop words (i.e., function words like "and," "the," "for," etc.) as defined by the nltk package. These two measures gave an impression of the length of the comment; by removing stop words, we also gain an approximate measure of the number of content words. We next counted the number of "starter phrases" (sentence scaffolds provided by the learning platform) used in the comment. Finally, we generated features using nltk's 32 tags for parts of speech, removing seven that were not found in our data to exclude zero variance features. Our final feature set was a total of 28 features.

We used the scikit-learn library [34] to implement commonly used regressors: Bayesian ridge regression, linear regression, XGBoost (via the XGBoost library [5]), Huber regression, and random forest regression. Hyperparameters were tuned on the training set only using the cross-validated grid search method provided by scikit-learn, where appropriate. As Spearman correlations (the metric used below) do not have a predefined chance value (e.g., unlike AUC ROC), we also derived a Chance baseline using the Dummy classifier in scikit-learn, which simply makes random predictions based on the prior probabilities observed in the training data. For instance, if 20% of the training data are labeled as '1', this classifier will predict '1' with a 20% probability, and so on for other values. This model undergoes the same training process as other models, and its performance is assessed using the same evaluation techniques applied to other classifiers.

All models were trained using 4-fold student-level cross-validation (multiple data points from the same student are in the same fold) and repeated for ten iterations, each with a new random seed. This type of cross-validation promotes generalizability to new students. Multiple iterations were performed to verify the stability of the results. For evaluation, predictions were pooled across folds, and then results were averaged across iterations.

Table 1: Final Annotation Coding Scheme

Tier	Description
1 – Least Robust	These annotations are generic and lack constructive feedback for the author. Tier 1 annotations could be applied to any thinklet as they do not connect to specific components of the thinklet. Example: "it was good"
2 – Somewhat Robust	These annotations lack specific details and provide little or no elaboration aside from the mathematics. Authors may attempt to connect to the work in the thinklet, but the connection is vague and not clearly explained. Example: "I hadn't thought of the way you found the answer. Although while you were explaining it, it became clear."
3 – Robust	These annotations attempt to elaborate on a specific piece of the thinklet or problem related to the problem-solving process. Authors may attempt to connect to the work in the thinklet, with specificity. Example: "I agree with your answer but maybe next time make the numbers that are supposed to be negative, negative in the equation to make it more clear."
4 – More Robust	These annotations also identify specific components of the thinklet, or problem related to the problem-solving process. Authors may connect specific components of the work in the thinklet but lack recommendations for next steps. Example: "My strategy is like yours because I put the information in almost the exact same way. I think I just switched the places of the two withdrawals."
5 – Most Robust	These annotations identify relevant components of the problem-solving process in a more elaborate fashion while also providing helpful feedback that is likely to support peer-to-peer learning. Example: "I respectfully disagree with you on the last pieces of your math as you had added a positive with a negative. While you should have added -30 with the -83 and gotten -113 then subtractive that with the positive 76 and gotten -37."

Table 2: Automated Detection Results – Spearman's rho. Correlation with ground truth.

Regressor	Rho
Chance	*.16*
Linear Regression	.85
Huber Regression	.75
Bayesian Ridge Regression	.82
Random Forest Regression	.91
XGBoost Regressor	.89

4.1 Model Evaluation

In order to address research question 1, we compare model accuracy by computing the correlation between the model predictions and human codes, used here as "ground truth" (described above in Table 1).

We used the Spearman correlation coefficient (i.e., Spearman rho) since the true labels are on an ordinal scale and the model predictions are continuous. All results reported are from the test folds (reported in Table 2). All results are above the chance baseline, with Random Forest Regression providing the best detector. The high correlation magnitude (rho=.9) with the ground truth signifies not only the supervised machine learning detector's very high performance but also indicates a robust relationship between the feedback robustness and the input features derived.

We investigated the models to understand how features related to predictions of feedback robustness, using SHapley Additive exPlanations (SHAP) values [26] as implemented in the shap library in Python. Table 3 lists the Shapley values with the largest impact on predictions.

5 LARGE LANGUAGE MODEL – RQ2

Large language models, such as OpenAI's GPT series, have advanced the field of natural language processing by demonstrating a capacity to comprehend and generate human-like text. These models are trained on significantly more text than any of the models trained in section 4, and have significantly more parameters. The features extracted in section 4, explicitly dissected the individual elements of the peer feedback to inform the model. In contrast, large language models present a more generalized approach to text-based information, allowing for a more holistic approach to evaluating feedback, that may more accurately emulate an instructor's approach. However, as general models of language, they also do not have the advantage of being designed for the specific purpose of evaluating feedback.

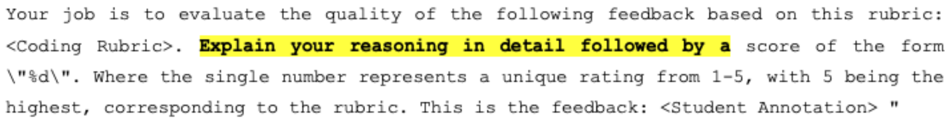
Recognizing the potential of these models in understanding and processing language, we employed ChatGPT via its API to examine its utility in classifying student feedback statements against a specified rubric, thus answering our second research question. This section outlines our methodology, the challenges we encountered, and the results derived from using ChatGPT as a detector of student peer feedback robustness.

5.1 Using Large Language Models as a Detector

In order to prompt the LLM, we combined the use of a rubric with a the same data used in section 4. The rubric in Table 1 was presented to ChatGPT along with 3 examples for each rating (within the prompt). The examples that were provided were not part of the

Table 3: High Shapley Value Variables for best performing model (Random Forest)

Predictor	Directionality	Predictor	Directionality
Digits	Positive	Unrecognized Words	Negative
Plural Nouns	Positive	Verb base form	Negative
Length of Annotation	Positive	Adverbs	Negative
Preposition/Subordinating	Positive	Modal	Negative
Determiner	Positive		
Adjective	Positive		
Verb 3rd person	Positive		



```
Your job is to evaluate the quality of the following feedback based on this rubric:
<Coding Rubric>. Explain your reasoning in detail followed by a score of the form
\"%d\". Where the single number represents a unique rating from 1-5, with 5 being the
highest, corresponding to the rubric. This is the feedback: <Student Annotation> "
```

Figure 1: ChatGPT Prompt for evaluating student peer feedback. The full rubric is not shown to avoid repeating Table 1. Yellow highlighted and bolded section was included in the run with explanations only; in the no explanation version, it is replaced with "Please provide a".

data used for evaluation. This rubric used the same language that had been given to the human coders who had previously human coded ground truth labels for this data, independently from GPT.

The prompt was iteratively refined (on unlabeled example data) in order to ensure that GPT provided ratings that were appropriate for the context and accurate to the request. For example, ChatGPT did not appear to view the rubric as restrictive. Even though the rubric clearly had only five tiers, when the explicit instruction to "rate from 1-5" was omitted from the prompt, GPT would generate extraneous categories for its labeling system. The final evaluations were performed on the entire data set, and no cross-validation was performed for the purpose of prompt engineering or fine-tuning.

We tested two final versions of the prompt (see Figure 1), one that included the phrase "Explain your reasoning in detail" and one that did not. Work has shown that the reasonings from ChatGPT when used as a classifier, may be of use in educational settings [48]. Though more expensive (due to the increased number of tokens), this can potentially result in more explainable decision-making that could further inform students.

Previous research has shown that even when setting the random seed and trying to account for random elements, it is likely that ChatGPT will not be deterministic between iterations [9]. To address this, we ran each prompt over 5 separate iterations of the entire dataset, gaining five sets of ratings for each peer feedback instance to allow to assess variability and stability of the approach.

For this initial experiment, each annotation to code was delivered to ChatGPT through a separate API session, thus limiting the amount that the system can learn from previous codes. This was done to limit confounds or ordering effects that may arise from delivering the annotations sequentially.

5.2 Results

Results for both versions of the ChatGPT detector (with and without requiring explanations in the prompt) are shown in Table 4. To allow comparability to earlier results, we again evaluate the model with a Spearman correlation between the detector values and the human codes used as ground truth. While these detectors still outperform chance (see Table 2), they only outperform one of the classifiers trained in section 3, and perform much more poorly than the best detectors from that section (.91, .89).

We note that the version of the prompt that requested explanations along with the rating had greater stability between iterations and a higher overall correlation with human-coded ground truth values. Though this difference in correlation is small, there was also a wider range of correlations between single iterations for the version without explanations (rho = .64-.78) than the version with explanations (rho = .74-.78). We performed a paired t-test on the standard deviations across runs for each annotation between the two versions of the prompt, finding a significant difference between the two distributions, t(116)=2.66, $p = .008$, implying that by including the request for explanations in detail as part of the prompt, we significantly improved the stability of the detector. One hypothesis for this may be that ChatGPT is evaluating itself for global cohesion [31], though as the exact workings of the model and response mechanism are not known, this remains simply conjecture.

5.3 Explanations

Though a traditional "feature analysis" is not possible with ChatGPT, one advantage of using an LLM is that it can be asked to explain its reasoning (see prompt above). We stored all of the given explanations from each run of the detector. Example reasoning statements are shown in Table 5.

We note anecdotally that the content of ChatGPT's feedback does not follow a strict formula, and there is considerable variance between iterations in the wording of the feedback provided. In some cases, the feedback quotes whole or individual elements of the student annotation, in others, no quotation is made.

Table 4: Results from ChatGPT Detectors of Peer Feedback Robustness

Measure	ChatGPT – No Explanations	ChatGPT – With Explanations
Chance Baseline Correlation	*.16*	*.16*
Correlation between averaged rating (across iterations) and Ground Truth	.76	.79
Minimum Correlation with Ground Truth (Single Iteration)	.64	.74
Maximum Correlation with Ground Truth (Single Iteration)	.78	.78
Number of Instances with 5 identical ratings	64/116	90/116
Average Standard Deviation Across Instances	.21	.13

Table 5: Example of rating explanation provided by ChatGPT for ratings of peer feedback

Rating Tier	Example Explanation Provided by ChatGPT
1 – Least Robust	The feedback "that is a really good answer" falls into the Tier 1 category. It is a generic comment that does not provide any specific or constructive feedback to the author. It does not elaborate on any part of the problem-solving process or provide any recommendations for improvement. Therefore, the score is "1".
2 – Somewhat Robust	The feedback provided is somewhat specific, as it suggests a way to improve the problem-solving process by separating addition and subtraction equations. However, it lacks a clear connection to the work in the thinklet and does not provide any elaboration on the mathematics involved. Therefore, this feedback falls under Tier 2 - Somewhat Robust.
3 – Robust	This feedback falls into the **Tier 3 - Robust** category. The feedback is specific to the thinklet, mentioning a particular step in the problem-solving process (adding two deposits together). However, it lacks any recommendations for next steps or further elaboration on the problem-solving process. Therefore, it doesn't reach the level of Tier 4 or 5. So, the score is "3".
4 – More Robust	The feedback provided falls into the Tier 4 category. The author of the feedback is connecting to the work in the thinklet with specificity, stating that their strategy is similar to the one used in the thinklet. However, the feedback lacks recommendations for next steps or any further elaboration on the problem-solving process. Therefore, the score is "4".
5 – Most Robust	The feedback provided is a Tier 5 - Most Robust. The author of the feedback has elaborated on a specific piece of the problem-solving process, pointing out that the original author was finding the median, not the mean. They then go on to explain how to correctly find the mean, demonstrating a clear understanding of the problem and providing constructive, specific feedback for the original author. This feedback promotes peer-to-peer learning and is therefore categorized as Tier 5.

Though these explanations are likely not good enough to be presented directly to students without a human in the loop or further refinement, they do present an easier interpretation than a feature analysis as to why an automated code may have been given. This increased interpretability is a clear advantage of the model as the future application is to provide information to students or instructors via the learning platform. These explanations present a starting point for teachers to understand why a student's rating might be lower and to translate to direct and meaningful "feedback on feedback". It should be noted that the prompt did not specify that the reasoning be constructive or aimed at a target audience (we discuss this limitation in the future work section).

6 APPLYING TO NEW DATA – RQ3

We next applied the best-performing classic NLP model from section 4 to the remainder of the peer feedback data (444 peer feedback comments from 124 students). In order to avoid confounding the analysis with training data, we consider only unseen student feedback (comments not used in training process) in this analysis. Using the same feature preprocessing as above, we extracted features from the remaining peer comments. These features were then used as the input to the best-performing model trained above, providing a score for each peer review comment. For the GPT model, the process remained identical to that discussed in section 5. As it both performed better (correlation to ground truth) and had higher stability, we used the version of the prompt that asked for detailed explanations alongside the ratings.

6.1 Correlating with External Measures

We next considered how the peer review annotation scores related to other measures in our dataset to investigate the model's convergent validity. As survey measures were recorded at the student level (i.e., not repeated measures), we averaged annotation scores by student (the annotator) to produce one value per student. These were then correlated with the survey measures (see Table 2), a Benjamini–Hochberg post-hoc correction for multiple comparisons was applied (see Table 6). We note that the correlations are similar in both magnitude and direction between the supervised approach

Table 6: Spearman Correlations of Average Annotation Scores and External Measures

Survey Measures	Correlation with Supervised Machine Learning Detector	Correlation with ChatGPT Detector
Jr Metacognitive Awareness Inventory	.282**	.337***
Math Anxiety Scale	-.271**	-.188*
Indiana Math – Belief that students can solve complex Problems	.234**	.181*
Indiana Math – Belief that effort increases math ability	.245**	.223*
Indiana Math – Belief that math is useful in daily life	.378***	.314***
ACE Task Switch	-0.032	-.066
ACE Flanker Task	-0.17	-.194*

and the GPT detector: students with higher anxiety scores were less likely to give effective feedback, whereas students with high metacognition scores were more likely to give effective feedback. We saw no significant relationship for the ACE task switch task; however, there was a significant negative correlation for the Flanker task (typically used to assess executive function components), implying that students who perform better on the Flanker task may provide worse peer feedback. However, this effect was only significant for the ChatGPT detector and requires further study to examine what elements of the Flanker task may be predictive of peer feedback quality, or vice versa. We note significant correlations with all of the Indiana Mathematics Belief scales included for both types of detectors. Of particular note is that students with a greater belief that mathematics has value in the real world are more likely to leave higher quality feedback, perhaps indicating a motivation component to effective feedback.

Finally, we analyzed the scores for changes over the course of the student's interaction with the learning platform (multiple weeks) using mixed-effects linear regressions with students as the random intercept. We did not observe significant differences from this analysis in either the predictions from the supervised machine learning detector (p=.348) or the ChatGPT detector (p=.174), indicating that peer review comments did not improve (as measured by our detectors) with increased usage of the platform.

6.2 Comparison of Supervised Machine Learning to ChatGPT Detector

This paper has considered two different methodologies for analyzing student peer feedback comments: supervised machine learning and an approach based on ChatGPT. As with many learning analytics analyses, the goal is not just to obtain good model performance, but ultimately to use that measurement in a way that can support a students' educational goals, put simply, inform "feedback on feedback," it is through these two lenses that we now compare the two approaches.

Both the Supervised Machine Learning models and the ChatGPT-based detector considerably surpassed the chance baseline. This outcome implies that rather than capturing mere randomness or noise, both methodologies discern genuine patterns or signals from the text they evaluate, signaling their capability to genuinely process the feedback relative to the expert-defined rubric. Considering accuracy (as measured by Spearman correlations with ground truth) the supervised machine learning models, for the most part, outperformed the ChatGPT detector. This heightened accuracy may be attributed to the rigorous training processes and the capability to tune model parameters and select the most relevant features.

On the other hand, the ChatGPT detector presents an advantage in terms of adaptability. There is the potential to make minor changes in the rubric without changing the detector process (or re-training a model). The foundation of large language models in a comprehensive knowledge base and a more generalized grasp of language provides ChatGPT with greater flexibility with regard to variations or modifications in feedback rubrics that the specialized models trained in section 3 may not have. Moreover, the ChatGPT detector's inherent understanding of context allows it to potentially deliver more actionable feedback without necessitating human design efforts. This characteristic can translate into generating insights that are immediately comprehensible to teachers and students, removing the need for additional interpretation methods. In essence, it can pave the way for a more efficient feedback loop, aiding educators and learners in rapidly acting upon the suggestions provided.

In summation, the accuracy advantage displayed by Supervised Machine Learning models is counterbalanced by the adaptability and directness of feedback offered by ChatGPT. As educational paradigms continue to evolve, there is merit in leveraging the strengths of both models for analytics, aiming to provide students with a comprehensive and dynamic feedback mechanism.

7 DISCUSSION AND CONCLUSIONS

Peer feedback offers learning opportunities for the learner receiving feedback, but also (and perhaps more so) for the learner providing the feedback. With known benefits to the development of self-regulated learning skills, including reflection, analysis, and problem-solving [36], peer feedback offers a valuable learning opportunity for those providing feedback. However, younger learners are typically novices in this process and require instruction as to what constitutes effective feedback. This paper presents two methods for automated evaluation of peer feedback, with the long-term goal of developing adaptive scaffolding for students. In the remainder of this section, we discuss our main findings, consider potential applications, and discuss limitations and future work.

7.1 Main Findings

We developed a five-tier coding system for student peer feedback, measuring the robustness of comments in a mathematics problem-solving environment. We extend previous work using natural language processing to automatically evaluate feedback messages from instructors to evaluate peer feedback comments relative to our five-tier coding process. We consider a two-pronged approach, one that uses supervised machine learning (RQ1), and another that leverages an existing large language model, ChatGPT (RQ2).

For the supervised machine learning (RQ1), we use relatively simple descriptive features (i.e., length, parts of speech) to develop an automated model with high accuracy when compared to human-coded ground truth examples. We also observed acceptable performance for the ChatGPT detector (RQ2), though it was overall less accurate when compared to the supervised machine learning approach. The ChatGPT-based detectors also had significantly less variance in performance between iterations when a request for a detailed explanation for result was included in the prompt. Notably, such explanations also have the potential to improve interpretability of the predictions for both students and teachers and form the basis for automated feedback.

We applied both models to previously unseen data (RQ3) and assessed convergent and external validity through the relationship between generated codes and external survey measures collected from the same students. We observed strong convergence: students who value mathematics more or who have higher metacognitive scores are giving better feedback to their peers (as measured by our detectors). Our results also indicated that students' scores are not improving over time, and thus, students may not currently be developing their feedback skills in the platform. Given the benefit to a learner of providing feedback [8, 36], it is clear that learners should receive more support or instruction in producing high-quality feedback for their peers.

7.2 Application

There are two principal applications of this work, real-time support of students, and reporting to teacher dashboards. For direct student support, the model developed in this work could be used to provide real-time scaffolding to learners as they write feedback comments. For example, students providing peer feedback that would be categorized as "1-least robust," might be encouraged to increase the length of their annotation, or include more adjectives, Personalized feedback would ideally be based on how the feature importance measures of the classic models match up with the feature values for the student's feedback. For example, length of annotation was a positive predictor of robustness, so when a student is about to submit a short annotation, they could be encouraged to expand upon their ideas to provide more constructive feedback. Through simulation, the detector can also be used to see how an adapted version of the student's feedback would be evaluated. Such simulations could then be shown to the student as a way to concretely demonstrate how they could improve, along with detail as to why. For the ChatGPT-based detector, students could be shown the explanation of why their feedback was rated a particular way, and this explanation could also identify areas of improvement relative to the rubric for the student to consider next time.

With regard to providing data to teachers, summative data on the quality of peer comments could be presented to teachers. Using feature analysis similar to what was done here, the model could also identify trends in the students' comments. Such trends could be communicated to teachers to aid them in their instructional design or provide students with class-level feedback. Feedback for individual students is also possible though likely will require some human editing or interpretation from domain experts (in this case teachers).

7.3 Limitations and Future Work

The features used in the supervised machine learning approach are fairly basic, compared to what has been used in other work, but already achieve very good correlation. Though it is encouraging that we can achieve such positive results from a limited feature set, future work should consider expanding this feature set to provide a deeper insight into components of successful peer assessment. One example of this might be sentiment analysis [23], which could provide more detail regarding the polarity of feedback comments (and could be done both with classical ML and ChatGPT). It should be noted, however, that this analysis may be harder to turn into real-time scaffolding for learners leaving comments since, as the complexity of the features increases, so too does the complexity of the interpretation. In addition, more complex NLP approaches are often at their most effective with larger amounts of source material than the short student response that we use in this work.

As with feature engineering, prompt engineering could further refine the ChatGPT based approach. As GPT continues to be developed and adapted, prompts may also need to evolve – many researchers have reported that prompts can cease to function correctly when OpenAI updates ChatGPT. It should also be considered that the ChatGPT approach leveraged the OpenAI API, which charges users a fee for usage. The supervised machine learning models can be run at much lower cost, making them more accessible to a wider audience, and feasible to be integrated into a platform in the long term. However, this limitation of ChatGPT could be addressed by switching to an open source LLM, when one of equivalent quality becomes available. It is also of note that we did not do any cross validated prompt engineering within the 'training' process. As methods with LLMs improve, one avenue of future work may be to refine this approach with micro-level fine-tuning within the detection process.

In this work, we have considered only considered one learning environment. Though our results are promising, it is not clear how well the findings will generalize across platforms. Future work should thus consider a similar approach for feedback across multiple learning environments to evaluate the robustness of this approach.

7.4 Concluding Remarks

In this study, we used data from 203 middle school students (6-8 graders) engaged in the CueThink online learning environment, to explore methods of analyzing student peer feedback. In addressing RQ1, we showed that with a combination of simplistic NLP approaches and supervised machine learning, we could detect feedback robustness for young learners significantly above chance levels. We further showed that we could use ChatGPT to provide similar classifications, and that it also provided plain text explanations for the classification, though with lower accuracy than classic NLP (RQ2). Lastly, we show convergent validity by applying the detectors to new data and comparing to external measures (RQ3). Overall, this study shows the feasibility of automating feedback mechanisms for young learners who are novices at providing feedback and provides the groundwork for providing automated "feedback on feedback".

ACKNOWLEDGMENTS

The research reported here was supported by the EF+Math Program of the Advanced Education Research and Development Program (AERDF) through funds provided to the University of Pennsylvania, University of Minnesota, and CueThink. The opinions expressed are those of the authors and do not represent views of the EF+Math Program or AERDF.

REFERENCES

[1] Monique Boekaerts. 1991. Subjective competence, appraisals and self-assessment. *Learning and Instruction* 1, 1 (January 1991), 1–17. https://doi.org/10.1016/0959-4752(91)90016-2

[2] Emma Carey, Francesca Hill, Amy Devine, and Dénes Sz Hucs. 2017. The modified abbreviated math anxiety scale: A valid and reliable instrument for use with children. *Frontiers in psychology* 8, (2017), 11.

[3] Anderson Pinheiro Cavalcanti, Arthur Diego, Rafael Ferreira Mello, Katerina Mangaroska, André Nascimento, Fred Freitas, and Dragan Gašević. 2020. How good is my feedback? a content analysis of written feedback. In *Proceedings of the tenth international conference on learning analytics & knowledge*, 2020. 428–437. .

[4] Anderson Pinheiro Cavalcanti, Rafael Ferreira Leite de Mello, Vitor Rolim, Máverick André, Fred Freitas, and Dragan Gaševic. 2019. An analysis of the use of good feedback practices in online learning courses. In *2019 IEEE 19th international conference on advanced learning technologies (ICALT)*, 2019. IEEE, 153–157. .

[5] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (*KDD '16*), 2016. ACM, New York, NY, USA, 785–794. . https://doi.org/10.1145/2939672.2939785

[6] CueThink. 2022. CueThink. Retrieved from https://www.cuethink.com/howitworks

[7] Magda B. L. Donia, Merce Mach, Tom A. O'Neill, and Stéphane Brutus. 2022. Student satisfaction with use of an online peer feedback system. *Assessment & Evaluation in Higher Education* 47, 2 (2022), 269–283. https://doi.org/10.1080/02602938.2021.1912286

[8] Kit S Double, Joshua A McGrane, and Therese N Hopfenbeck. 2020. The impact of peer assessment on academic performance: A meta-analysis of control group studies. *Educational Psychology Review* 32, 2 (2020), 481–509.

[9] Richard H. Epstein and Franklin Dexter. 2023. Variability in Large Language Models' Responses to Medical Licensing and Certification Examinations. Comment on "How Does ChatGPT Perform on the United States Medical Licensing Examination? The Implications of Large Language Models for Medical Education and Knowledge Assessment." *JMIR Medical Education* 9, 1 (July 2023), e48305. https://doi.org/10.2196/48305

[10] Peggy A. Ertmer, Jennifer C. Richardson, Brian Belland, Denise Camin, Patrick Connolly, Glen Coulthard, Kimfong Lei, and Christopher Mong. 2007. Using Peer Feedback to Enhance the Quality of Student Online Postings: An Exploratory Study. *Journal of Computer-Mediated Communication* 12, 2 (January 2007), 412–433. https://doi.org/10.1111/j.1083-6101.2007.00331.x

[11] Douglas Fisher and Nancy Frey. 2009. Feed up, back, forward. *Educational Leadership* 67, 3 (2009), 20–25.

[12] Sheila M Fram. 2013. The constant comparative analysis method outside of grounded theory. *Qualitative Report* 18, (2013), 1.

[13] Graham Gibbs. 2006. How assessment frames student learning. In *Innovative assessment in higher education*. Routledge, 43–56.

[14] Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences* 120, 30 (July 2023). https://doi.org/10.1073/pnas.2305016120

[15] Arthur C. Graesser, Danielle S. McNamara, Max M. Louwerse, and Zhiqiang Cai. 2004. Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers* 36, 2 (May 2004), 193–202. https://doi.org/10.3758/BF03195564

[16] Gutierrez de Blume. Under Review. Metacognitive Awareness among Middle School Adolescents: Development and Validation of a Shortened Version of the MAI, Jr. (Under Review).

[17] John Hattie and Helen Timperley. 2007. The power of feedback. *Review of Educational Research* 77, 1 (2007), 81–112. https://doi.org/10.3102/003465430298487

[18] Ilona E. de Hooge, Marcel Zeelenberg, and Seger M. Breugelmans. 2007. Moral sentiments and cooperation: Differential influences of shame and guilt. *Cognition and Emotion* 21, 5 (August 2007), 1025–1042. https://doi.org/10.1080/02699930600980874

[19] Derek R. Hopko, Rajan Mahadevan, Robert L. Bare, and Melissa K. Hunt. 2003. The Abbreviated Math Anxiety Scale (AMAS): Construction, Validity, and Reliability. *Assessment* 10, 2 (2003), 178–182. https://doi.org/10.1177/1073191103010002008

[20] Sophie Ioannou-Georgiou and Pavlos Pavlou. 2003. *Assessing young learners*. Oxford University Press.

[21] Peter Kloosterman and Frances K. Stage. 1992. Measuring Beliefs About Mathematical Problem Solving. *School Science and Mathematics* 92, 3 (1992), 109–115. https://doi.org/10.1111/j.1949-8594.1992.tb12154.x

[22] Angela Lee and Tong Ming Lim. 2016. Mining opinions from university students' feedback using text analytics. *Information Technology in Industry* 4, 1 (2016).

[23] Bing Liu. 2015. *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge University Press.

[24] Eric Zhi-Feng Liu, Sunny SJ Lin, Chi-Huang Chiu, and Shyan-Ming Yuan. 2001. Web-based peer review: The learner as both adapter and reviewer. *IEEE Transactions on education* 44, 3 (2001), 246–251.

[25] Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. *arXiv preprint cs/0205028* (2002).

[26] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems*, 2017. 4768–4777. .

[27] Katherine McEldoon, Jessica Yarbro, Samuel Downs, Matthew Ventura, Faby Gagné, Seth Corrigan, Devon Skerritt, and Ruth Lahti. 2020. Instructor Feedback Practices in Undergraduate Writing at Scale. (2020).

[28] Kamila Misiejuk, Barbara Wasson, and Kjetil Egelandsdal. 2021. Using learning analytics to understand student perceptions of peer feedback. *Computers in Human Behavior* 117, (2021), 106658. https://doi.org/10.1016/j.chb.2020.106658

[29] Mara Negrut and Kihyun Ryoo. 2021. Designing Effective Automated Feedback for Modeling Tools. In *Proceedings of the 15th International Conference of the Learning Sciences-ICLS 2021.*, 2021. International Society of the Learning Sciences. .

[30] David J Nicol and Debra Macfarlane-Dick. 2006. Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in higher education* 31, 2 (2006), 199–218.

[31] OpenAI. 2023. GPT-4 Technical Report.

[32] Ikenna Osakwe, Guanliang Chen, Alex Whitelock-Wainwright, Dragan Gašević, Anderson Pinheiro Cavalcanti, and Rafael Ferreira Mello. 2022. Towards automated content analysis of educational feedback: A multi-language study. *Computers and Education: Artificial Intelligence* 3, (January 2022), 100059. https://doi.org/10.1016/j.caeai.2022.100059

[33] Melissa Patchan, Karen Rambo-Hernandez, Brianna Dietz, and Kennedy Hathaway. 2018. Assessing the Validity of Peer Feedback in a Sixth Grade Mathematics Class. In *13th International Conference of the Learning Sciences (ICLS)*, July 2018. International Society of the Learning Sciences. . Retrieved December 1, 2023 from https://repository.isls.org//handle/1/847

[34] Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Mattieu Perrot, and Edouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12, (2011), 2825–2830. https://doi.org/10.1007/s13398-014-0173-7.2

[35] James W. Pennebaker, Martha E. Francis, and Roger J. Booth. 2001. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates* 71, 2001 (2001), 2001.

[36] Esther van Popta, Marijke Kral, Gino Camp, Rob L. Martens, and P. Robert-Jan Simons. 2017. Exploring the value of peer feedback in online learning for the provider. *Educational Research Review* 20, (2017), 24–34. https://doi.org/10.1016/j.edurev.2016.10.003

[37] Tracii Ryan, Michael Henderson, Kris Ryan, and Gregor Kennedy. 2021. Designing learner-centred text-based feedback: a rapid review and qualitative synthesis. *Assessment & Evaluation in Higher Education* 46, 6 (August 2021), 894–912. https://doi.org/10.1080/02602938.2020.1828819

[38] D. Royce Sadler. 2009. Indeterminacy in the use of preset criteria for assessment and grading. *Assessment & Evaluation in Higher Education* 34, 2 (2009), 159–179. https://doi.org/10.1080/02602930801956059

[39] D. Royce Sadler. 2010. Beyond feedback: developing student capability in complex appraisal. *Assessment & Evaluation in Higher Education* 35, 5 (2010), 535–550. https://doi.org/10.1080/02602930903541015

[40] Rayne A Sperling, Bruce C Howard, Lee Ann Miller, and Cheryl Murphy. 2002. Measures of Children's Knowledge and Regulation of Cognition. *Contemporary Educational Psychology* 27, 1 (January 2002), 51–79. https://doi.org/10.1006/ceps.2001.1091

[41] Lesa A. Stern and Amanda Solomon. 2006. Effective faculty feedback: The road less traveled. *Assessing Writing* 11, 1 (January 2006), 22–41. https://doi.org/10.1016/j.asw.2005.12.001

[42] University of California San Francisco. 2021. Adaptive Cognitive Evaluation Explorer. Retrieved from https://neuroscape.ucsf.edu/researchers-ace/

[43] Martin Van Boekel, Shelby Weisen, and Ashley Hufnagle. 2021. Feedback in the Wild: Discrepancies Between Academics' and Students' Views on the Intended Purpose and Desired Type of Feedback. In *Proceedings of the 15th International Conference of the Learning Sciences-ICLS 2021.*, 2021. International Society of the Learning Sciences. .

[44] Cynthia Weston, Terry Gandell, Jacinthe Beauchamp, Lynn McAlpine, Carol Wiseman, and Cathy Beauchamp. 2001. Analyzing interview data: The development and evolution of a coding system. *Qualitative Sociology* 24, 3 (2001), 381–400. https://doi.org/10.1023/A:1010690908200

[45] Philip H Winne and A.F. Hadwin. 1998. Studying as Self-Regulated Learning. *Metacognition in Educational Theory and Practice* (1998), 277–304.

[46] Naomi Winstone, David Boud, Phillip Dawson, and Marion Heron. 2022. From feedback-as-information to feedback-as-process: a linguistic analysis of the feedback literature. *Assessment & Evaluation in Higher Education* 47, 2 (2022), 213–230. https://doi.org/10.1080/02602938.2021.1902467

[47] Benedikt Wisniewski, Klaus Zierer, and John Hattie. 2020. The Power of Feedback Revisited: A Meta-Analysis of Educational Feedback Research. *Frontiers in Psychology* 10, (2020). https://doi.org/10.3389/fpsyg.2019.03087

[48] Andres Felipe Zambrano, Xiner Liu, Amanda Barany, Ryan S. Baker, Juhan Kim, and Nidhi Nasiar. 2023. From nCoder to ChatGPT: From Automated Coding to Refining Human Coding. In *Advances in Quantitative Ethnography* (*Communications in Computer and Information Science*), 2023, Cham. Springer Nature Switzerland, Cham, 470–485. . https://doi.org/10.1007/978-3-031-47014-1_32

[49] J. Zhang, J. M. A. L. Andres, Stephen Hutt, R. S. Baker, J. Ocumpaugh, C. Mills, J. Brooks, S. Sethuraman, and T. Young. 2022. Detecting SMART Model Cognitive Operations in Mathematical Problem-Solving Process. In *Proceedings of the International Conference on Educational Data Mining*, 2022. .

[50] Jiayi Zhang, R. S. Baker, J M Alexandra L Andres, Stephen Hutt, and Sheela Sethuaman. 2023. Automated Multi-Dimensional Analysis of Peer Feedback in Middle School Mathematics. In *Proceedings of the International Conference on Computer Supported Collaborative Learning.*, 2023.