Scaling Reflection Prompts in Large Classrooms via Mobile Interfaces and Natural Language Processing

Xiangmin Fan[1], Wencan Luo[1], Muhsin Menekse[2], Diane Litman[1], Jingtao Wang[1]
[1]Department of Computer Science and LRDC, University of Pittsburgh, Pittsburgh, PA 15260, USA
{xiangmin, wencan, litman, jingtaow}@cs.pitt.edu
[2]School of Engineering Education, Purdue University, West Lafayette, IN 47907, USA
menekse@purdue.edu

ABSTRACT
We present the iterative design, prototype, and evaluation of CourseMIRROR (Mobile In-situ Reflections and Review with Optimized Rubrics), an intelligent mobile learning system that uses natural language processing (NLP) techniques to enhance instructor-student interactions in large classrooms. CourseMIRROR enables streamlined and scaffolded *reflection prompts* by 1) reminding and collecting students' in-situ written reflections after each lecture; 2) continuously monitoring the quality of a student's reflection at composition time and generating helpful feedback to scaffold reflection writing; 3) summarizing the reflections and presenting the most significant ones to both instructors and students. Through a combination of a 60-participant lab study and eight semester-long deployments involving 317 students, we found that the reflection and feedback cycle enabled by CourseMIRROR is beneficial to both instructors and students. Furthermore, the reflection quality feedback feature can encourage students to compose more specific and higher-quality reflections, and the algorithms in CourseMIRROR are both robust to cold start and scalable to STEM courses in diverse topics.

Author Keywords
Mobile Learning, Reflection Prompts, Collaborative Learning, Natural Language Processing.

ACM Classification Keywords
H5.2. Information interfaces and presentation (e.g., HCI): User Interfaces.

INTRODUCTION
The degree and quality of interactions between students and instructors are critical factors for students' engagement, retention, and learning outcomes [39]. However, such interactions are limited in large classrooms (e.g., undergraduate level introductory STEM courses) and online courses. It is safe to predict that the issue of class size will only get worse due to enrollment increase (e.g., undergraduate enrollment increased by 46% from 1990 to 2013 [4]) and educational budget cuts [33].

In recent years, researchers in education have discovered the feasibility and effectiveness of *"reflection prompts"* [7] (a.k.a. *"muddy cards"* [34] or *"one-minute papers"* [24]) to improve both teaching and learning across multiple disciplines. In a typical deployment of *reflection prompts*, students are given index cards at the end of each lecture and are encouraged to reflect on what was confusing in the lecture. After collecting responses from students, the instructor summarizes the student reflections, identifies major misunderstandings, and plans follow-up actions, such as providing feedback in the following lectures, and tailoring the teaching plan in the future. Previous studies in different domains [1, 5, 7, 23, 32] consistently confirmed that reflective activities could benefit students by enhancing their retention and comprehension in learning.

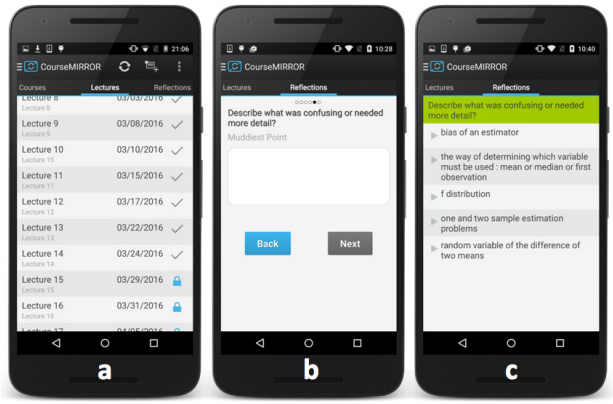

Figure 1. CourseMIRROR interfaces. a) lecture list; b) a sample reflection prompt; c) reflection summary page.

Despite the simple workflow and the encouraging efficacy, there are at least three key challenges when deploying *reflection prompts* in large classrooms. First, it is tedious and time consuming to remind and collect students' reflective responses after each lecture. Second, it is also time consuming for instructors to summarize and make sense of the raw response data [34]. Third, as highlighted by Fan et al [21], it is difficult to maintain students'

IUI 2017, March 13-16, 2017, Limassol, Cyprus
 ISBN 978-1-4503-4348-0/17/03…$15.00
DOI: http://dx.doi.org/10.1145/3025171.3025204

sustained motivation to compose concrete, specific and pedagogically valuable reflections through multiple months.

In this paper, we present the iterative design, prototype, and evaluation of CourseMIRROR[1] (Mobile In-situ Reflections and Review with Optimized Rubrics, Figure 1), a mobile learning system that uses natural language processing (NLP) techniques to enhance large classroom instructor-student interactions via streamlined and scaffolded *reflection prompts*. CourseMIRROR can 1) remind students to submit in-situ written reflections after each lecture, and collect such reflections in a scalable manner; 2) continuously monitor the quality of the reflection in composition time and generate engaging and helpful feedback to scaffold reflection writing; 3) summarize the gist of reflections and present the most significant ones to both instructors and students. Through a combination of a 60-participant lab study and eight semester-long deployments involving 317 students, we found that the reflection and feedback cycle enabled by CourseMIRROR are beneficial to both instructors and students.

Specifically, this paper makes the following contributions:

- We present CourseMIRROR, a scalable mobile learning system that uses NLP techniques to facilitate the collection and use of high quality responses to *reflection prompts* in large classrooms.
- We show that the interactive reflection quality feedback feature can scaffold students to write concrete and specific reflections. Our algorithms are scalable to courses in diverse topics and robust to cold start.
- We find students were willing to submit reflections via CourseMIRROR in a timely manner.
- We share our insights and lessons learned from eight semester-long deployments.

RELATED WORK

Reflections in Learning

Reflection is a key component of self-regulated learning [10]. It is a fundamental learning activity in which people *"recapture their experience, think about it, mull it over and evaluate it"* [8]. Previous research illustrated the value of learners' reflection on what they had done, processed or engaged in [1, 7], as well as on their confusing (i.e. *muddy*) points [32]. Studies also suggested that reflection could benefit students by helping them identify the misconceptions in their current beliefs [13, 29] and enhance their retention and comprehension in learning [1], even without external feedback [43]. Williams and colleagues [43] found that prompting and encouraging students to *explain* abnormal corollaries (e.g. people receiving lower *absolute* grades in exam A could have higher *relative* performance than those in exam B) were more effective than asking students to describe a concept.

Traditional implementations of *reflection prompts* via *muddy cards* [34] and *one-minute-papers* [24] can face scalability problems in large classrooms. As reported by Mosteller [34], it took an instructor 30-45 minutes to summarize reflections from a 50-student class. Moreover, recklessly composing *any* reflection is insufficient for effective learning—the quality also matters. Menekse et al [32] related the characteristics (e.g., the details included and the cognitive processes identified) of students' daily reflections to Chi's iCAP framework [14] (i.e. *passive*, *active*, *constructive* and *interactive* learning activities). By analyzing and coding the reflections based on a quality rubric (Figure 2), Menekse and colleagues [32] observed a significant positive correlation between the quality of reflections (i.e. *none*, *vague*, *general* and *specific*) and the learning gains. CourseMIRROR goes beyond a mobile implementation of *reflection prompts* by facilitating and scaffolding the composition and dissemination of reflection prompts via intelligent user interfaces.

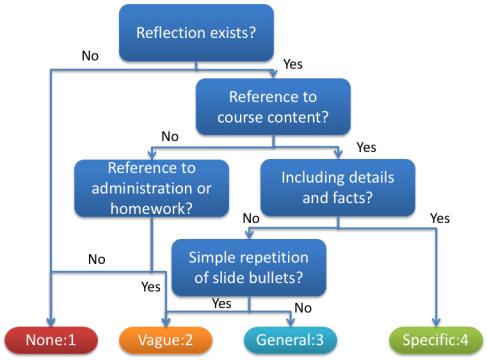

Figure 2. Rubric of reflection quality [32].

Computerized Reflection and Feedback Collection

Instructors in traditional classrooms can leverage audience response systems (ARSes, a.k.a. *"clickers"*) [11, 15] to collect real-time responses from students. However, ARSes are designed for multiple choice questions (MCQ) or True/False questions rather than open-ended reflections. Moreover, the hardware requirements and cost issues could prevent the widespread adoption of such systems.

Researchers also proposed various analytic techniques [25, 26, 27, 41, 44] to gain insights into student activities in MOOCs and flipped classrooms by analyzing artifacts generated in the learning process. For example, instructors can infer confusions and misconceptions of students by monitoring online discussion forums [25, 41], analyzing students' interaction logs [27], embedding and reviewing in-video exercises [26], and detecting students' cognitive states by mining their physiological signals [44].

Mudslide by Glassman et al [23] allows students to spatially anchor their confusions as circular "*muddy points*" directly on lecture slides and visualizes the aggregated

[1] Mobile apps for Android and iOS platforms and a mobile HTML5 optimized web version are available for free at http://www.coursemirror.com

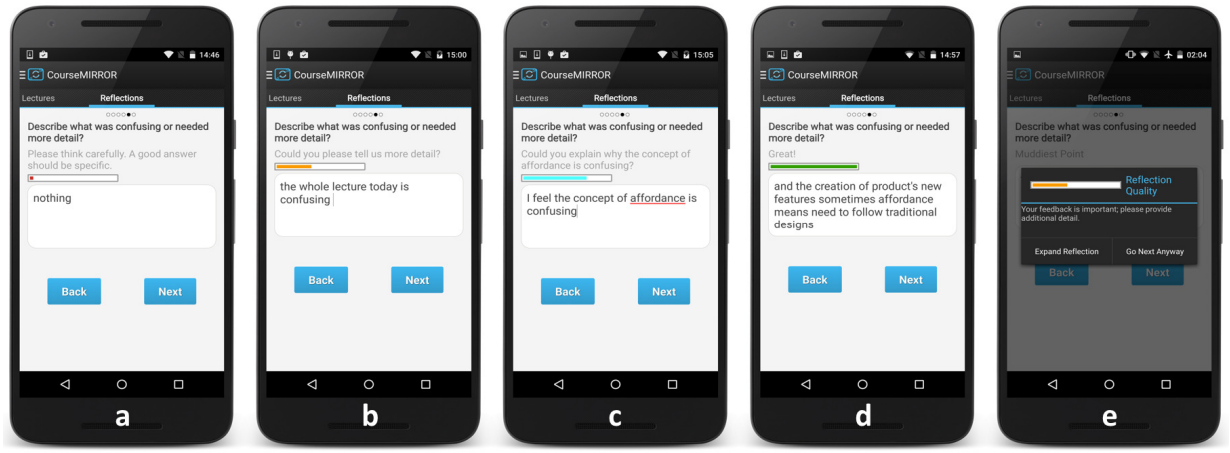

Figure 3. Reflection writing interfaces with quality feedback. a, b, c, d) instant feedback (IF, appear constantly at composition time); e) latent feedback (LF, appear in a dialog box after a submission attempt).

annotations to instructors. Although both Mudslide and CourseMIRROR can scale the *"muddy cards"* workflow [34], there are major differences between the two systems beyond target platforms (i.e. PCs vs. mobile). First, Mudslide is optimized for video watching in online courses and flipped classrooms, whereas CourseMIRROR reminds and collects students' reflections *in-situ* in traditional large classrooms. Second, Mudslide relies on lecture slides to localize confusions of students spatially. In comparison, CourseMIRROR distributes *open-ended* prompts and leverages interactive *scaffolding* to help students to compose high quality reflections in *natural language*. Third, CourseMIRROR uses text summarization algorithms to capture the gist of student responses while Mudslide leverages point cloud style visualizations to help instructors quickly locate confusions in the lecture slides.

Mobile Survey and Experience Sampling Methods

Through a study with 1,500 U.S. panelists, researchers found that mobile phone participants were willing to provide short responses to open-ended questions [42]. Multiple research projects also confirmed that mobile phones can be viable and comparable devices for short and optimized surveys [9, 20].

Reflection collection is also relevant to the Experience Sampling Method (ESM) [28] and Diary Studies [12] in HCI. Although systems such as Momento [12] and MyExperience [22] support event-contingent ESM via either SMS or context-activated polling, they were not designed and optimized in educational settings.

DESIGN OF COURSEMIRROR

There are four major design goals for CourseMIRROR:

G1: Provide students a convenient and efficient way to compose and submit reflection responses *in-situ*.

G2: Encourage and help students to create specific and pedagogically valuable reflections.

G3: Facilitate instructors to make sense of students' written reflections efficiently in large classrooms.

G4: Assist students to read their classmates' reflections for peer learning.

CourseMIRROR is designed as a mobile app to fulfill **G1**. A recent survey [17] indicates 92% of undergraduates in the U.S. own smart phones. The *instant on*, *always connected* abilities of mobile devices could allow students to compose and submit reflections efficiently. Further, CourseMIRROR sends automatic, lecture time-triggered push notifications to collect students' reflections *in-situ*.

In order to fulfill **G2**, CourseMIRROR continuously monitors the quality of the reflection at composition time and generates engaging and helpful feedback to scaffold reflection writing (Figure 3). This design was inspired by recent research findings that providing context-sensitive feedback on students' self-explanations could help them construct better explanations when using intelligent tutoring systems (e.g., in Cognitive Tutor [1] and SE-COACH [16]).

To achieve **G3**, CourseMIRROR runs customized automatic text summarization algorithms on the server to generate a summary after each lecture (Figure 1.c). We hypothesize that relevant and coherent summaries can help instructors quickly identify students' confusion and misunderstandings. To realize **G4**, CourseMIRROR also allows students to share and access the summaries with their classmates. We hypothesize that reading reflection summaries can benefit students by letting them revisit and reevaluate the learning contents from different perspectives.

Text Summarization Algorithm

We explored word level, phrase level and sentence level summarization techniques and chose phrase level summarization after pilot tests. We found phrases are easy to read and browse just like keywords, and can fit better on small devices than sentences. Phrase level summarization also provides more coverage than sentence level summarization under a given length limit.

CourseMIRROR utilizes the text summarization algorithm proposed by Luo et al. [31], which was specifically

designed for the purpose of summarizing reflective responses from students. The algorithm emphasizes both the *representative* (high-frequency reflections) and the *diversity* of the students (who wrote the reflections). It consists of three steps. First, use a syntax parser to generate candidate noun phrases since the knowledge concepts are usually referred as noun phrases. Second, cluster the candidate phrases into groups via the K-Medoids algorithm based on similarities of the semantic meaning. The algorithm measures semantic similarity between phrases via Latent Semantic Analysis [19]. With relevant clustering, the algorithm addresses the lexical variety problem (e.g., students use different words *"bicycle parts"* and *"bike elements"* for the same meaning). Third, select the most representative phrase in each cluster via a graph-based ranking model (i.e. LexRank). The selected phrases are then re-ranked by the number of students who mentioned the phrases. Phrases mentioned by more students should receive more attention from the instructor. This algorithm was evaluated on an engineering course corpus provided by [32], and achieved a significantly better performance in terms of ROUGE scores than a variety of other algorithms, such as MEAD, LexRank, and MMR.

Interactive Reflection Quality Feedback

In two pilot deployments of an early version of CourseMIRROR, Fan et al [21] found that some students began to submit brief and trivial reflections (e.g., *"none"*, *"N/A"*, *"all good"*) after months of extended use. Such reflections were neither informative for instructors nor beneficial for learning. Meanwhile, the length of reflections decreased significantly over time (12.3 words in the first half of the semester vs. 9.9 words in the second half of the semester [21]). Such findings highlight the challenges in 1) maintaining the *sustained motivation* for students throughout a semester; and 2) encouraging students to compose *high quality* reflections. Similar problems also existed in traditional intelligent tutoring systems, e.g., Aleven and colleagues [2] observed that students provided very few explanations and even fewer good explanations when using an intelligent tutor that only prompted for explanations.

We have designed and implemented a novel quality feedback feature (Figure 3) in CourseMIRROR to address, at least in part, these two challenges. When a student is composing a reflection, CourseMIRROR continuously monitors the quality of the reflection and generates encouraging and informative feedback to scaffold the reflection writing process. The feedback is provided via a color-coded progress bar and improvement suggestions in natural language. The progress bar (Figure 3.a-3.d, above the reflection edit box) creates a visual of the quality of the current reflection in composition. A full progress bar indicates that the reflection is specific and detailed. This metaphor could inform students of how close they are to creating high-quality reflections. The improvement suggestions in natural language are also shown above the progress bar. Such suggestions give students specific, easy-to-follow instructions on *how* to improve the quality of their current reflection. This design is in part inspired by findings on providing feedback in intelligent tutoring systems [1, 16] and peer review systems [35]. Researchers found that *context-sensitive* feedback can help students construct better explanations to their solutions, even when the feedback is very simple (e.g., the correctness of the explanations [1]). Previous study also suggested that providing feedback regarding the presence of solutions to students could help them generate more comments with solutions in peer reviews [35].

We explored two different timings to deliver quality feedback by designing both an *instant feedback* (**IF**) feature and a *latent feedback* (**LF**) feature. Instant feedback (Figure 3.a-3.d) is always visible to students during the composition process. Latent feedback (Figure 3.e) appears in a dialog box after clicking *"next"* or *"submit"* button. Students can choose either to go back and revise the draft or to submit the reflections after receiving the latent feedback.

Reflection Quality Prediction

CourseMIRROR extended the classifier-based approach proposed by Luo [30] to predict reflection quality based on the rubric in Figure 2. The original quality classifier [30] uses a Support Vector Machine (SVM) with linear kernel. Features include unigram (i.e. whether a word is present), word count, and part-of-speech (e.g., whether a proper-noun is present). These features are also widely used in other NLP tasks including automatic text scoring [37] and text classification [38], and are proven to be effective. The classifier was trained on previous student reflection datasets [32] containing 1,257 reflections and the corresponding expert-rated quality scores.

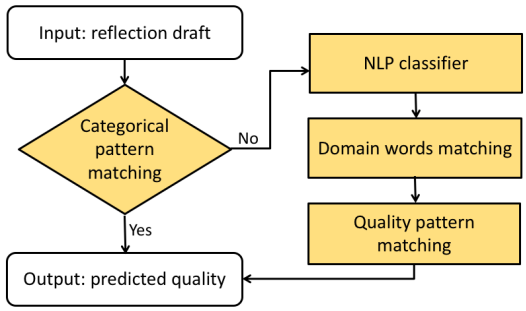

Figure 4. Workflow of reflection quality prediction.

Although the quality classifier above can achieve good accuracies on pre-collected reflection corpora, the classification accuracies drop significantly when classifying reflections from new courses with very different vocabulary and learning topics when compared with the training courses. The domain miss-match problem (i.e. cold start) is commonly acknowledged in various natural language processing applications, such as text classification [18], sentiment classification [36], and part-of-speech tagging [3]. In practice, it would be impossible to collect and

annotate reflections for each new course, and then train a course-dependent quality classifier.

To address this challenge, CourseMIRROR uses a combination of a statistical NLP classifier and three complementary *pattern matching* techniques (Figure 4) to achieve high accuracy and more relevant reflection quality prediction in a course-independent manner. The three pattern matching techniques include *domain words matching*, *categorical patterns*, and *quality patterns*.

Domain words matching is based on an exhaustive list of domain words extracted from the lecture slides[2]. It is a reasonable assumption that reflections with domain words are at least on-topic and relevant. Thus it introduces domain knowledge for the quality prediction.

Categorical patterns are the frequently appeared exemplar patterns in each quality category. For example, *"N/A"*, *"nothing"*, and *"all good"* are categorical patterns of "none (1)" reflections while a simple repetition of a slide title is a categorical pattern of "vague (2)" reflections.

Quality patterns include *abstract* phrase and word level signals for both high and low quality reflections. They are independent from specific course topics. For instance, starting with *"what/how/why"* and ending with *"?"* typically indicates that the input is a concrete question, which is a sign of high quality reflections. In comparison, the words *"everything"* or *"the whole lecture"* usually lead to vague expressions, and thus they are signs of low quality reflections.

Category	Examples	Action
Categorical Patterns	*"nothing", "N/A", "all good"*	Output the category
Quality Patterns (positive or negative)	*"...what/how…?" "…relationship between…"*	Add as a new entry in Eq 1 (3rd component)
Domain Words	*"..affordance..", "…p value…"*	Add as a new entry in Eq 1 (2nd component)

Table 1. Examples of pattern matching.

By analyzing the expert-annotated student reflection dataset [32], two researchers iteratively generated a total of 15 *categorical patterns* and a total of 33 *quality patterns*. Table 1 shows some sample patterns.

Figure 4 illustrates the overall workflow of the reflection quality prediction algorithms in CourseMIRROR. When a reflection matches any *categorical pattern* during runtime, the algorithm directly outputs the corresponding quality score without invoking the NLP classifier. This branch reduces both the computational power and network bandwidth. Otherwise, the NLP classifier first predicts the reflection quality, then the predicted quality score is adjusted according to the results of *domain words matching* and *quality pattern matching*, according to **eq. 1** below:

$$Q = q + \alpha * DW + \sum_{p_i \in QP} p_{i_{weight}} \quad \textbf{(eq. 1)}$$

Here q represents the classifier-predicted quality, *DW* is the number of matched *Domain Words*, α is the weight (i.e. 0.5), *QP* is the set of matched *Quality Patterns*, and p_{i_weight} is the weight (range from -1 to 1) of the particular pattern p_i.

The three complementary patterns are implemented as database tables of regular expressions on the server side. In addition to global patterns, CourseMIRROR also allows instructors to define and customize course-specific patterns and improvement suggestions.

Improvement Suggestions (Hints) Generation

CourseMIRROR provides encouraging and specific improvement suggestions based on the predicted quality (i.e. none, vague, general, specific) and the actual contents of the reflection. For example, when a student writes a "none" reflection, the system asks her to *"think carefully and start by naming a concept that is difficult to understand"*. When a student writes a "general" reflection, the system asks her to *"be more specific and tell us why you feel confused"*. CourseMIRROR pre-loads multiple hand-crafted sentences as candidate suggestions for each category, and randomly selects one from the corresponding group to maintain the feedback diversity.

By supporting the *capture group* feature in regular expressions, CourseMIRROR can detect, extract specific concepts (e.g. *affordance*) in reflections and refer to them in the improvement suggestions. For example, when CourseMIRROR detects that the input pattern is *"[X] is confusing"* (where *[X]* is a concept in the lecture), it then generates the hint *"please explain *why* [X] is confusing"*. In this way the system could generate more relevant and specific feedback based on the semantic meaning or the structure of the input.

LAB STUDY

Study Design

We conducted a 60-participant lab study to investigate the usability and efficacy of the interactive reflection quality feedback feature. We applied a between-subjects design with three conditions: *No-Feedback* (**NF**), *Latent-Feedback* (**LF**), and *Instant-Feedback* (**IF**). Under **NF** condition, participants write reflection without any feedback from CourseMIRROR. In comparison, CourseMIRROR provides both quality feedback and textual hints under both **LF** and **IF** conditions. During the study, participants watched 3 short lecture videos (7-10 minutes each) from the "Model Thinking" course by Prof. Scott Page in University of Michigan [40]. After finishing each lecture, participants responded to the following reflective questions on CourseMIRROR:

[2] Although CourseMIRROR maintains a course-dependent domain word list for each course, the NLP classifier in CourseMIRROR no longer requires course-specific training.

- Learning Point: "What have you learned in today's class?
- Muddy Point: "What was confusing in today's class?

At the end of the study, we conducted semi-structured interviews to solicit participants' subjective feedback on the interactive quality feedback design. We aimed to gain further understanding about how the feedback on reflection quality was perceived and digested and how it affected the writing process.

Participants and Apparatus

We recruited 60 participants (25 female) between 19 and 36 years of age (mean=27) from a local university, who were randomly assigned to the three conditions. The study lasted for around 60 minutes, and each participant received $10 for their time.

The participants watched the lecture videos on an Apple iMAC, with a 1.6GHz dual-core Intel Core-i5 processor, 8 GB RAM, and a 21.5-inch display. We used a Samsung Galaxy Note 3 smartphone with a 5.7-inch display running Android 5.0 for the CourseMIRROR mobile client.

Experimental Results

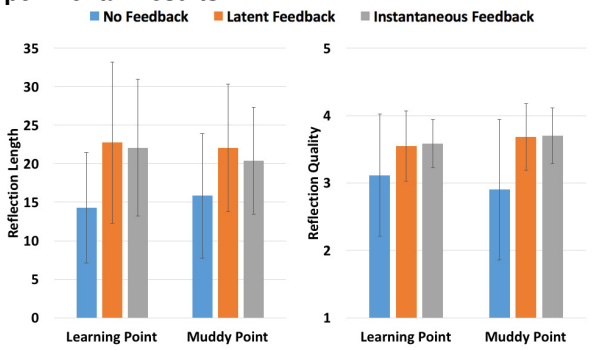

Figure 5. Reflection length and quality by reflection question. Error bars show one standard deviation.

Quality feedback can help participants create longer and higher-quality reflections.

We chose word count as the first quantitative metric to understand reflections collected. As shown in previous research such as Experience Sampling Method (ESM) [45] and creative editing [6], word count is a good marker of writing quality because it correlates indirectly with the number of details included. As shown in Figure 5, left, the average reflection length was 15.07, 22.40, and 21.24 words for the **NF**, **LF**, and **IF** condition respectively. Analysis of variance results showed that there was a significant difference ($F(2, 57)=5.64$, $p<0.01$) in reflection length. Pairwise mean comparison (t-tests) showed that the reflection length between **NF** and **IF** ($p<0.05$), **NF** and **LF** ($p<0.01$) were significantly different. There was no significant difference in reflection length between **IF** and **LF** ($p=0.62$). Question type (learning point, muddy point) did not exhibit a significant effect on reflection length ($F(1, 57)=0.07$, $p=0.79$).

We recruited two raters to give independent quality ratings of the reflections based on the rubric in [32] (Figure 2). The agreement between the two raters was high (percent agreement: 85.0%; Cohen's kappa: 0.72; Quadratic Weighted Kappa[3]: 0.91). Disagreements were settled by discussions between the two raters after the independent coding sessions. As shown in Figure 5, right, the average reflection quality was 3.01, 3.62, and 3.64 for the **NF**, **LF**, and **IF** condition respectively. Analysis of variance results showed that there was a significant difference ($F(2, 57)=12.63$, $p<0.001$) in reflection quality. Pairwise mean comparison (t-tests) showed that the reflection quality of **IF** was significantly higher than **NF** ($p<0.001$), the reflection quality of **LF** was significantly higher than **NF** ($p<0.001$). There was no significant difference in reflection quality between **IF** and **LF** ($p=0.22$). Question type did not exhibit a significant effect on reflection quality ($F(1, 57)=3.34$, $p=0.073$) either.

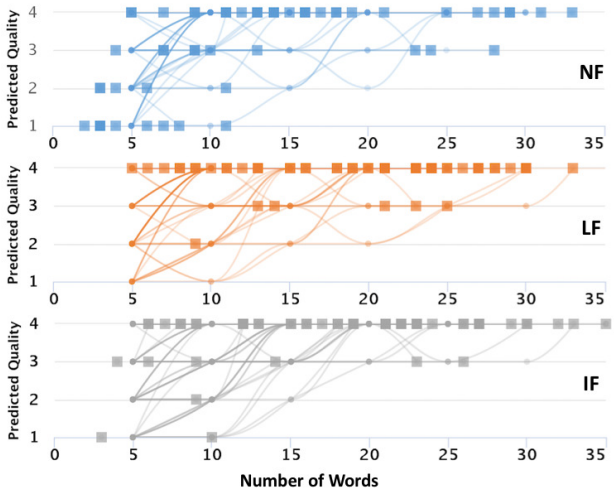

Figure 6. Predicted reflection quality by writing progress (i.e. words completed). Small dots denote the predicted quality at corresponding length. Square symbols represent submission attempts by learners.

To gain further understanding of the impact of feedback type on the reflection composition process, we plotted the predicted quality scores for each submission (Figure 6) and average performance (Figure 7) when reflection was in composition at different lengths. Please note that the quality feedback was invisible to participants in **NF** condition and was also not visible to participants in **LF** condition before a submission attempt. From both Figure 6 and Figure 7, we can clearly observe that participants in **NF** condition tended to submit reflections early, with lower quality when compared with participants in **LF** and **IF** condition. In the **LF** condition, due to the lack of quality feedback before a

[3] Since the quality scores are ordered, incorrect predictions have different costs (e.g., predicting "3" as "1" is more severe than predicting "3" as "2"). Therefore, we also report Quadratic Weighted Kappa.

submission attempt, it was more common to encounter *decreases* in predicted quality in the middle of composition when compared with the **IF** condition (Figure 6). Overall, 60% of the reflections in **NF**, 88.3% of the reflections in **LF**, and 85% of the reflections in **IF** received the highest quality rating (Specific:4, Figure 2) when participant finalized their compositions. At the same time, participants in **NF** condition submitted more vague:2 or none:1 reflections (13.3%) than those in **IF** (3.3%) and **LF** (0%) condition.

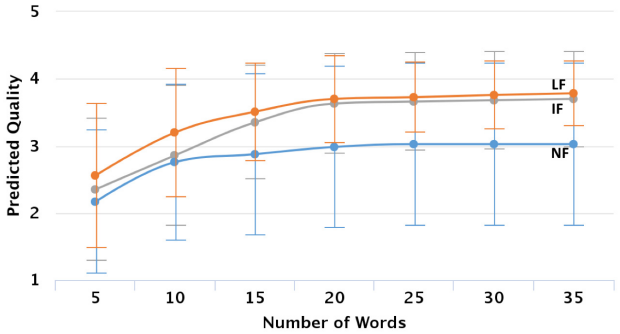

Figure 7. Average predicted quality scores by condition and writing progress (i.e. words completed).

Qualitative results on instant quality feedback (IF)

Participants in **IF** reported that the progress bar made them feel *"mental pressure" [S20]* and *"obligated to fulfill the bar" [S19]* while writing reflections— *"this feature act like a supervisor that stared at me to force me to do a better work" [S1]*. At the same time, they got *"the feeling of achievement" [S4]* and were highly encouraged when they saw their progresses:

- *"It gives you a hint about how is your feedback's quality and it feels like a reward to gain full credit for feedback." [S14]*
- *"it is pretty satisfying to see the bars filling up—it is quite encouraging" [S1]*

Participants also reported that the improvement suggestions in natural language were helpful in guiding them to create deeper reflections.

- *"It tells you specifically what you should improve on" [S9]*
- *"At first I just wrote some topic words, but I saw the quality is low and it asked me to illustrate why the concept is confusing. This can definitely make me think deeper." [S1]*

To our surprise, two participants reported that sometimes the progress bar metaphor could be discouraging—they stopped thinking and writing immediately or shortly after the progress bar was fully filled. They believed that it was the *"desired amount" [S8]* when the bar was full:

- *"Originally I had 4 sentences to write. After writing 2 the progress bar is full and it told me the reflection is great, so I stopped right there." [S13]*
- *"the system said that the reflections were good enough" [S28].*

This suggested that we need to be careful when using conclusive feedback, e.g., fully filled progress bars, textual hints such as "great reflection", etc.

Quantitative results on latent quality feedback (LF)

In **LF** condition, when participants clicked the *"submit"* button, they saw the system feedback and were able to choose to revise the reflection. Therefore, we attribute the reflection quality improvement and length increase (compared with **NF**) to participants' revisions after they saw the feedback.

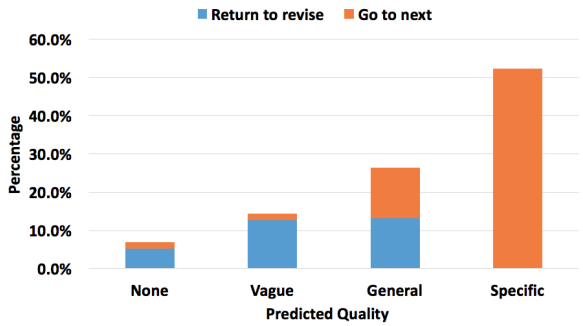

Figure 8. Participants' reactions by reflection quality.

We first compare the system-predicted reflection quality between the first drafts and the final submissions in **LF** condition. The average quality of the first draft is 3.11 (σ =1.02). In comparison, the average quality of the submitted draft is 3.78 (σ=0.48). In total they viewed the latent feedback panel for 174 times, among which they chose to go back and revise the reflection for 54 (31.0%) times. Figure 8 shows the participants reactions (i.e. go back to revise, go to next/submit without revision) when they saw the system feedback. Among the 54 revisions, 49 (90.7%) revisions lead to better reflections (Table 2).

Qualitative results on latent quality feedback (LF)

There were 29 occasions when participants chose to submit reflection even though they did not get the *"perfect reflection"* feedback from the system. Reasons include:

- *"I don't think I can write more when I go back." [S21]*
- *"I think I've provided enough details, even though CourseMIRROR still asked me to provide more detail." [S33]*
- *"I was very confused about the '[no] free lunch theorem' and I knew nothing about it so I cannot further illustrate why it is confusing." [S35]*

Tradeoffs between IF and LF

Although the quantitative analysis did not show a significant difference between **IF** and **LF**, we discovered qualitative differences through observations and interviews with participants.

Real-time vs. Attention. Participants expressed their preference on **IF** for its visibility in real-time. At the same time, they reported that they mainly focused on the quality feedback (i.e. the progress bar). Two participants reported that they totally ignored the improvement suggestions and another participant only followed the improvement suggestions when he *"tried hard but still can't fill the bar" [S3]*. By comparison, participants in **LF** reported that they paid sufficient attention to both the progress bar and the textual suggestions. We attribute this to the *intrusive* nature of latent feedback (via a dialogue box), which drew more attention by pausing the composition process.

		After			
		None	Vague	General	Specific
Before	None	0	**1.85%**	**7.41%**	**7.41%**
	Vague	0	1.85%	**20.37%**	**18.52%**
	General	0	0	7.41%	**35.19%**
	Specific	0	0	0	0

Table 2. The distribution of system-predicted quality changes after revision.

However, the **LF** can frustrate participants for delayed information:

- *"The system should tell me what is the expected reflection at the beginning rather than after I spend time thinking and writing the reflection." [S40]*

Pattern matching improves the accuracy of quality prediction

In order to assess the efficacy of *pattern matching* in improving the quality prediction accuracy, we conducted an off-line comparison between using the classifier only and using the combinations of the classifier and pattern matching (Table 3). The gold standard quality scores were human annotations.

Method	Percent	Kappa	QWKappa
Classifier Only	58.3%	0.28	0.67
Classifier+All Pattern Matching	**77.2%**	**0.52**	**0.83**
Classifier+Domain Word List	73.3%	0.46	0.76
Classifier+Quality Patterns	71.7%	0.44	0.80
Classifier+Categorical Patterns	58.9%	0.30	0.70

Table 3. Accuracies of quality prediction algorithms.

The classifier (i.e. SVM) used in the study was trained on previous student reflection datasets [32] containing 1,257 reflections and the experts' quality ratings. It is worth mentioning that the domain of the course (i.e. data modeling) in this study is different with the domain of the training course (i.e. material science and engineering).

On average the *domain word matching*, *quality pattern matching*, and *categorical pattern matching* are triggered by 1.12 (σ=0.89), 1.41 (σ=1.06), 0.05 (σ=0.2) times, respectively, for each reflection. The results in Table 3 confirm that integrating pattern matching could enhance the quality prediction accuracy, and mediate the domain miss-match problem.

IN THE WILD DEPLOYMENTS

CourseMIRROR has been deployed in eight courses[4] in two universities as of September 2016, involving a total of six instructors and 317 students. Most of the courses were undergraduate level STEM courses, such as Basic Physics, Data Structures, and Statistics for Industrial Engineering.

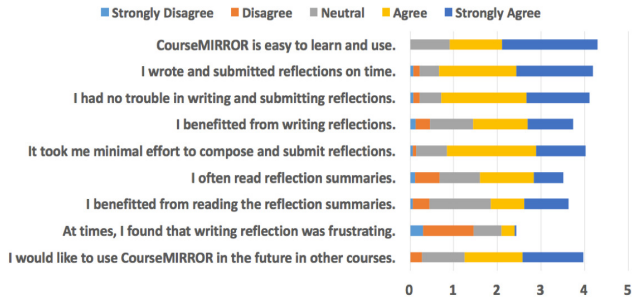

Figure 9. Subjective ratings on a 5-Point Likert scale.

Overall, students reported positive experiences with CourseMIRROR (Figure 9). Ratings were measured on a 5-point Likert scale (1 = *strongly disagree*, 5 = *strongly agree*). Students thought CourseMIRROR was easy to use μ=4.30 (σ =0.80) and would like to use CourseMIRROR in future courses μ=3.96 (σ=0.94).

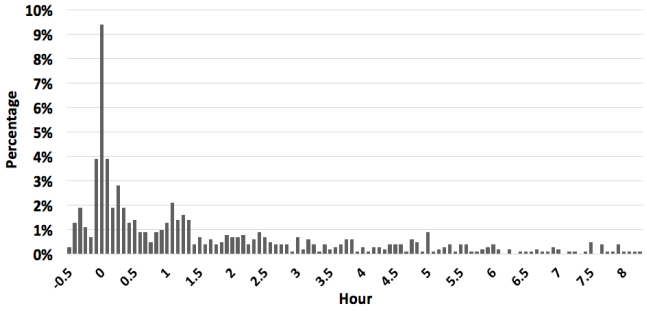

Figure 10. The histogram of response time (hour).

Finding 1: *Students were willing to submit reflections in a timely manner.*

In total we collected 3,855 reflections from the eight deployments. The average response rate was 53.1% (σ =0.16). This rate is encouraging considering that there was a significant portion of quiet and shy students who rarely asked questions or seek for help actively in each lecture. We further analyzed the submission time of the reflections. We found 48.3% of the reflections were submitted within two hours after the end of each lecture. 9.2% of the reflections were submitted before the end of the lectures (Figure 10). We attribute the timely reflection submission in part to the novelty and efficacy of our CourseMIRROR mobile client.

[4] Fan et al [21] is a non–archival publication (i.e. extended abstract) reporting preliminary findings from two pilot deployments.

Finding 2: *Students benefitted from the reflection and feedback cycle enabled by CourseMIRROR.*

The average response to question *"I benefitted from writing reflections"* was 3.74 (σ =1.08) (Figure 9). Students reported that the benefits were two fold. Firstly, composing reflections in CourseMIRROR enhanced their retention by encouraging them to revisit what they learned:

- *"It's not a long time to learn which subjects aren't understood by the class."*
- *"I can think about what I learned and what I didn't understand."*

Secondly, the timely instructor feedback enabled by CourseMIRROR helped students clear up their confusions:

- *"Because our prof used those reflections and cleared the muddy points."*
- *"Especially, when our instructor started to solve more examples on class, I saw this benefit in a more concrete way."*

Finding 3: *Reflection summaries allowed instructors to understand students' difficulties efficiently.*

All the instructors responded positively to CourseMIRROR according to post-study questionnaires and interviews. Instructors reported that they regularly read the reflection summaries generated by CourseMIRROR, e.g., one instructor reported that she *"never skip the summary"* while another instructor reported that he *"tried to look at every summary"*. The time needed to understand the summary for each lecture was minimal, ranging from *"definitely less than 5 minutes" to "5-10 minutes"*. In comparison, an instructor spent 30-45 minutes summarizing the responses from a 50-student class in traditional paper-based deployments [34]. The automatic text summarization was promising— e.g., instructors can *"get an idea of the issues some students are having trouble with"* by reading the summaries, and *"clarify/go over some topics that indicated as problematic"* in future lectures.

Finding 4: *Students enjoyed reading summaries of reflections from their classmates.*

The average subjective ratings of *"I often read reflection summaries"* and *"I benefitted from reading reflection summaries"* on a 5-point Likert scale (Figure 9) are 3.51 (σ =1.09) and 3.63 (σ =1.06), respectively. They reported that seeing their classmates' reflections could broaden their views and allow them to reevaluate from different perspectives (e.g., *"I feel I was also confused about other people's muddiest points after I see the summary"*). At the same time, realizing that other students having the same confusion could reduce their frustration and enhance their confidence (e.g., *"Good to see other people were also confused, I know it's not my problem and relaxed"*).

Finding 5: *The quality feedback feature can help students compose higher-quality reflections in real-world settings.*

After finishing the lab study on the reflection quality feedback feature, we integrated the updated CourseMIRROR client with instant quality feedback to the Data Structures course (29 lectures in total, 40 CS undergraduate students enrolled) in a local university in Spring 2016. The feature was enabled in an app update made in the middle of the semester. 12 students updated the app (2 started from lecture 19, the other 10 started from lecture 20). The following analysis focuses on the reflections generated by the 12 students who used both versions (i.e. with/without quality feedback). Specifically, we compare their reflections submitted from lecture 9-18 (i.e. without quality feedback) and lecture 20-29 (with quality feedback) to see whether the feedback could help students generate more specific reflections (Table 4). To measure reflection quality, the same two raters (who rated the reflection corpus in the lab study) rated the reflections collected from the 12 students with the same rubrics ([32], Figure 2). Their independent ratings achieved high agreement (percent agreement: 88.5%; Cohen's kappa: 0.73; Quadratic Weighted Kappa: 0.93). They discussed on the disagreements to achieve consensus.

	W/O Interactive Feedback	W/ Interactive Feedback
Total # of Reflections	86	79
Average Length	7.5	12.5
Reflection Quality	Average: 2.8	Average: 3.4
None(1)	21 (24.4%)	6 (7.6%)
Vague(2)	7 (8.1%)	5 (6.3%)
General(3)	27 (31.4%)	19 (24.1%)
Specific(4)	31 (36.0%)	49 (62.0%)

Table 4. Distribution of reflection quality.

Students composed significantly longer (12.5 vs. 7.5, $p<0.001$) reflections with interactive feedback (Table 4). At the same time, the reflection quality was also significantly higher (3.4 vs. 2.8, $p<0.05$) with interactive feedback. Considering the quality feedback feature was updated in the middle of the semester, this is not a tightly controlled comparative study. However, considering that the reflection length and quality *decreased* over time in previous studies without the quality feedback feature, this result is still promising. It implied that the interactive feedback motivated students write higher-quality reflections in a sustainable manner. We plan to conduct larger scale and controlled deployment in the future to verify this finding.

Students reported positive experiences with the interactive quality feedback. On a 5-point Likert scale, they reported that the interactive feedback and suggestions were relevant to their reflections (μ = 4.29, σ = 0.82). They also reported that the interactive feedback helped them think deeper and compose more specific reflections (μ = 4.6, σ = 0.68). Sample comments include:

- *"The new app was helpful most of the time, especially when I only gave a general idea, it pushed me to think deeper about what I'm interested or confused, and be able to find the specific point."*

- *"I really think more carefully about the lesson when writing reflections using the updated version."*

Students also reported that the interactive feedback helped them learn how to do deep reflection and the ability could last:

- *"After the first 1-2 times, I knew what is a desired reflection and I can write a 'perfect reflection' without reading the suggestions then."*

Although the overall reflection quality improved, there were still 7.6% non-substantive reflections submitted. The major complaint on the interactive feedback feature was the lack of diversity. One student reported that *"the pattern of the suggestions seems to be fixed"*. In the future, we plan to significantly increase the diversity of the feedback to avoid boredom, e.g., by changing the presentation or by integrating more pattern matching templates to make the feedback more specific to the input.

We also discovered minor gaming behaviors by analyzing the user interaction log. For example, one student originally wrote *"N/A"* and got the quality feedback as "none" reflection. After that, the student tried to get a higher score by rephrasing the reflection, such as *"no muddy point"*, *"all clear and no muddy point"*, and finally submitted as *"everything is confusing"*. In the future, we need to detect such gaming behavior in real time and provide scaffolds explicitly designed for gaming behaviors, e.g., by prompting the student to explain the *why* behind a concept mentioned in the lecture.

Finding 6: *Active integration to the curriculum is essential.*

There was no mandatory requirement for students to participate in any of the deployments. We also explicitly informed students who opted-in that they were free to quit at any time. Although we observed high response rates in most deployments, we cannot claim that CourseMIRROR could always work in every condition. For example, the response rate (24.8%) in the Basic Physics class in Spring 2015 was significantly lower than other deployments (e.g., 56.7% in Statistics for Industrial Engineers, 57.7% in Mobile Interface Design). We attribute the low response rate to the weak integration to the course curriculum.

First, there were no course incentives (i.e. extra credit) provided in the physics deployment. Surprisingly, according to past experiences in deployments, course incentives (as low as one extra point in class participations) worked better than monetary incentive (e.g., as much as $30 for semester-long participation). Thus we encouraged, but did not require, instructors to provide some extra credit for participation in later deployments.

Second, the instructor did not refer to CourseMIRROR in class after he announced the deployment in the first lecture. We found that it was more effective for the instructor to explicit acknowledge the source of clarifications, i.e. CourseMIRROR, in the reflection and feedback cycle.

LIMITATIONS AND FUTURE WORK

Although both lab studies and in-the-wild deployments show the benefits of CourseMIRROR in facilitating and scaling reflection prompts, it is still necessary to improve and deploy CourseMIRROR in even larger scale, more diversified courses in the near future. More importantly, we plan to conduct large scale class deployment with *control groups* (ideally 40 or more students per condition) to further verify the educational value of CourseMIRROR in different contexts (e.g. What would be the best practices for deploying CourseMIRROR? Whether and to what extent CourseMIRROR combine with other instructional interventions synergistically?).

Another interesting future work is to enable *personalized learning* by analyzing reflections collected via collaborative filtering algorithms. Potential opportunities include recommending relevant learning materials (e.g., MOOC videos) and exercises, as well as establishing the connection and collaboration among peers with complementary skills.

While reading the summaries generated by CourseMIRROR can help instructors understand students' difficulties and misconceptions, there still exists opportunities to facilitate instructors to convert summaries to *actions* and *resources* in the follow-up lectures. We plan to explore techniques (e.g. instructor-side visualizations, revision tracking, and improvement suggestions) to scaffold instructors to carter the upcoming teaching activities according to reflections from students.

CONCLUSION

We presented the iterative design, prototype, and evaluation of CourseMIRROR, an intelligent mobile learning system that uses NLP techniques to enhance large classroom instructor-student interactions via streamlined and scaffolded *reflection prompts*. CourseMIRROR reminds students to compose their reflections directly on their mobile devices *in-situ* after each lecture. CourseMIRROR also scaffolds students to compose high quality reflections and facilitates both instructors and students to identify major points of confusion in a lecture via customized natural language processing algorithms. We conducted both controlled lab studies and eight semester-long deployments to evaluate the efficacy of CourseMIRROR. Overall we show that the reflection and feedback cycle enabled by CourseMIRROR is scalable and beneficial to both instructors and students.

ACKNOWLEDGEMENTS

We thank Xiang Xiao, Phuong Pham, Wei Guo, Carrie Demmans Epp, Gangzheng Tong, Zhuogu Yu and the anonymous reviewers for the help and constructive feedback. This research is in-part supported by an RDF from the Learning Research and Development Center (LRDC) at the University of Pittsburgh.

REFERENCES

1. Vincent Aleven, and Kenneth R. Koedinger. An effective metacognitive strategy: Learning by doing and explaining with a computer-based Cognitive Tutor. *Cognitive Science* 26, no. 2 (2002): 147-179.
2. Vincent Aleven, Octav Popescu, and Kenneth R. Koedinger. Towards tutorial dialog to support self-explanation: Adding natural language understanding to a cognitive tutor. In *Proceedings of Artificial Intelligence in Education*, pp. 246-255. 2001.
3. Rie Kubota Ando, and Tong Zhang. A high-performance semi-supervised learning method for text chunking. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pp. 1-9. Association for Computational Linguistics, 2005.
4. Susan Aud, Sidney Wilkinson-Flicker, Thomas Nachazel, and Allison Dziuba. *The condition of education 2013*. Government Printing Office, 2013.
5. John R. Baird, Peter J. Fensham, Richard F. Gunstone, and Richard T. White. The importance of reflection in improving science teaching and learning. *Journal of research in Science Teaching* 28, no. 2 (1991): 163-182.
6. Joshua E. Blumenstock. Size matters: word count as a measure of quality on wikipedia. In *Proceeding of the 17th international conference on World Wide Web - WWW* '08, ACM Press (2008), 1095–1096.
7. David Boud, Rosemary Keogh, and David Walker. *Reflection: Turning experience into learning*. Routledge 2013.
8. David Boud, Rosemary Keogh, and David Walker. Promoting reflection in learning: A Model. *Boundaries of adult learning* 1 (2013): 32.
9. Trent D. Buskirk and Charles Andrus. Online surveys aren't just for computers anymore! Exploring potential mode effects between smartphone vs. computer-based online surveys. *AAPOR Annual Conference*. 2012.
10. Deborah L. Butler, and Philip H. Winne. Feedback and self-regulated learning: A theoretical synthesis. *Review of educational research* 65.3 (1995): 245-281.
11. Jane E. Caldwell. Clickers in the large classroom: Current research and best-practice tips. *CBE-Life Sciences Education* 6.1 (2007): 9-20.
12. Scott Carter, Jennifer Mankoff, and Jeffrey Heer. Momento: support for situated ubicomp experimentation. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 125-134. ACM, 2007.
13. Michelene TH Chi, Nicholas De Leeuw, Mei-Hung Chiu, and Christian LaVancher. Eliciting self-explanations improves understanding. *Cognitive science* 18.3 (1994): 439-477.
14. Michelene TH Chi. Active-constructive-interactive: A conceptual framework for differentiating learning activities. *Topics in Cognitive Science* 1.1 (2009): 73-105.
15. Linda J. Collins. Livening up the classroom: Using audience response systems to promote active learning. *Medical reference services quarterly* 26.1 (2007): 81-88.
16. Cristina Conati, and Kurt Vanlehn. Toward computer-based support of meta-cognitive skills: A computational framework to coach self-explanation. *International Journal of Artificial Intelligence in Education* 11 (2000): 389-415.
17. Eden Dahlstrom, Christopher Brooks, Susan Grajek, and Jamie Reeves. ECAR study of undergraduate students and information technology, 2015. *Educause Center for Applied Research*.
18. Wenyuan Dai, Gui-Rong Xue, Qiang Yang, and Yong Yu. Co-clustering based classification for out-of-domain documents. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 210-219. ACM, 2007.
19. Susan T Dumais. Latent semantic analysis. *Annual review of information science and technology* 38, no. 1 (2004): 188-230.
20. Marika de Bruijne, and Arnaud Wijnant. Comparing survey results obtained via mobile devices and computers: an experiment with a mobile web survey on a heterogeneous group of mobile devices versus a computer-assisted web survey. *Social Science Computer Review* (2013): 0894439313483976.
21. Xiangmin Fan, Wencan Luo, Muhsin Menekse, Diane Litman, and Jingtao Wang. CourseMIRROR: Enhancing large classroom instructor-student interactions via mobile interfaces and natural language processing. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 2015.
22. Jon Froehlich, Mike Y. Chen, Sunny Consolvo, Beverly Harrison, and James A. Landay. MyExperience: a system for in situ tracing and capturing of user feedback on mobile phones, In *Proceedings of the 5th international conference on Mobile systems, applications and services*, pp. 57-70. ACM, 2007.
23. Elena L. Glassman, Juho Kim, Andrés Monroy-Hernández, and Meredith Ringel Morris. Mudslide: A spatially anchored census of student confusion for online lecture videos. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 2015.
24. William S. Harwood. The one-minute paper. *Journal of Chemical Education* 73.3 (1996): 229.

25. Karen Kear*. Peer learning using asynchronous discussion systems in distance education. *Open Learning: The Journal of Open, Distance and e-Learning*, 19.2 (2004)., 151-164.

26. Juho Kim, Elena L. Glassman, Andrés Monroy-Hernández, and Meredith Ringel Morris. RIMES: Embedding interactive multimedia exercises in lecture videos. *In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pp. 1535-1544. ACM, 2015.

27. Juho Kim, Philip J. Guo, Carrie J. Cai, Shang-Wen Daniel Li, Krzysztof Z. Gajos, and Robert C. Miller. Data-driven interaction techniques for improving navigation of educational videos. *Proceedings of the 27th annual ACM symposium on User interface software and technology*. ACM, 2014.

28. Reed Larson, and Mihaly Csikszentmihalyi, The experience sampling method. *New Directions for Methodology of Social & Behavioral Science*, 1983.

29. Tania Lombrozo. The structure and function of explanations. *Trends in cognitive sciences* 10.10 (2006): 464-470.

30. Wencan Luo, and Diane Litman. Determining the Quality of a Student Reflective Response. *The Twenty-Ninth International FLAIRS Conference*. 2016.

31. Wencan Luo, and Diane Litman. Summarizing Student Responses to Reflection Prompts. In *Proceedings of EMNLP*, 2015.

32. Muhsin Menekse, Glenda Stump, Stephen Krause, and Michelene Chi. The effectiveness of students' daily reflections on learning in engineering context. In *118th ASEE Annual Conference and Exposition*. 2011.

33. Michael Mitchell, and Michael Leachman. Years of cuts threaten to put college out of reach for more students. *Center on Budget and Policy Priorities* (2015): 1-26.

34. Frederick Mosteller. The 'Muddiest Point in the Lecture' as a feedback device. On *Teaching and Learning: The Journal of the Harvard-Danforth Center* 3 (1989): 10-21.

35. Huy Nguyen, Wenting Xiong, and Diane Litman. Instant Feedback for Increasing the Presence of Solutions in Peer Reviews. In *Proceedings Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations (NAACL-HLT)*, pp. 6-10

36. Sinno Jialin Pan, Xiaochuan Ni, Jian-Tao Sun, Qiang Yang, and Zheng Chen. Cross-domain sentiment classification via spectral feature alignment. In *Proceedings of the 19th international conference on World wide web*, pp. 751-760. ACM, 2010.

37. Zahra Rahimi, Diane J. Litman, Richard Correnti, Lindsay Clare Matsumura, Elaine Wang, and Zahid Kisa. Automatic scoring of an analytical response-to-text assessment. In *International Conference on Intelligent Tutoring Systems*, pp. 601-610. Springer International Publishing, 2014.

38. François Rousseau, Emmanouil Kiagias, and Michalis Vazirgiannis. 2015. Text categorization as a graph classification problem. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, 1702–1712.

39. Helen Quinn, Heidi Schweingruber, and Thomas Keller, eds. *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. National Academies Press, 2012.

40. Scott Page. Model Thinking. https://www.coursera.org/learn/model-thinking

41. Matthew JW Thomas. Learning within incoherent structures: The space of online discussion forums. *Journal of Computer Assisted Learning* 18.3 (2002): 351-366.

42. Tom Wells, Justin T. Bailey, and Michael W. Link. Comparison of smartphone and online computer survey administration. *Social Science Computer Review* 32.2 (2014): 238-255.

43. Joseph Jay Williams, Tania Lombrozo, Anne Hsu, Bernd Huber, and Juho Kim. Revising Learner Misconceptions Without Feedback: Prompting for Reflection on Anomalies. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 2016.

44. Xiang Xiao and Jingtao Wang. Towards Attentive, Bi-directional MOOC Learning on Mobile Devices. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 2015.

45. Zhen Yue, Eden Litt, Carrie J. Cai, Jeff Stern, Kathy K. Baxter, Zhiwei Guan, Nikhil Sharma, and Guangqiang George Zhang. Photographing information needs: the role of photos in experience sampling method-style research. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1545-1554. ACM, 2014.