Analytical Techniques for Developing Argumentative Writing in STEM

Patricia Davies [1], Yanjun Gao [2], Smaranda Muresan [2], and Rebecca Passonneau [2]

[1]Prince Mohammad Bin Fahd University
[2]Affiliation not available

October 30, 2023

Abstract

This article demonstrates how experiential learning could be used to develop argumentative essay writing skills in STEM students. Written feedback, when delivered in a timely manner, is an effective way of advancing students' understanding of the writing process. Unfortunately, large class sizes and the limited backgrounds of instructors do not always make formative feedback possible. STEM students are especially disadvantaged since approaches to teaching written communication tend to differ from the trial-and-error strategies compatible with many STEM areas. An experiential learning approach to writing instruction can have a positive impact on developing writing skills in STEM learners. Implementing algorithms for providing STEM students with immediate, dependable, formative feedback is expected to improve their performance in writing. This paper discusses an experiential learning project for teaching argumentative writing was delivered to computer science and engineering freshmen. Also discussed are automated analysis of content and argumentation in the essays, using NLP methods.

Analytical Techniques for Developing Argumentative Writing in STEM

Patricia M. Davies, *Member, IEEE,* Rebecca J. Passonneau, Smaranda Muresan, and Yanjun Gao

***Abstract*—*Contribution:* This article demonstrates how experiential learning could be used to develop argumentative essay-writing skills in STEM students. It illustrates the design, implementation and evaluation of an Experiential Learning project for undergraduate Computer Science and Engineering students, and discusses the development of a natural language processing application, designed to aid instructors in providing students with high-quality, prompt, formative feedback on writing tasks.**

***Background:* Written feedback, when delivered in a timely manner, is an effective way of advancing students' understanding of the writing process. Unfortunately, large class sizes and the limited backgrounds of instructors do not always make formative feedback possible. STEM students are especially disadvantaged since approaches to teaching written communication tend to differ from the trial-and-error strategies compatible with many STEM areas.**

***Intended Outcomes:* An experiential learning approach to writing instruction can have a positive impact on developing writing skills in STEM learners. Implementing algorithms for providing STEM students with immediate, dependable, formative feedback is expected to improve their performance in writing.**

***Application Design:* An experiential learning project for teaching argumentative writing was delivered to computer science and engineering freshmen. The structure of the project is described: the teaching approach, essay assignments, the rubric used for grading the essays, and its reliability. Also discussed are automated analysis of content and argumentation in the essays.**

***Findings:* The project was successful in producing a transformative writing experience for computer science and engineering students. It demonstrates ways of incorporating experiential education to help STEM students develop strategies for good essay writing.**

***Index Terms*—argumentation, content annotation, experiential learning, higher education, natural language processing, rubric reliability, STEM, writing.**

I. Introduction

GOOD writing, especially making effective arguments, demonstrates excellent critical thinking skills [1]. Writing as a means of communicating knowledge is a necessity in Higher Education (HE). Yet students enrolled in STEM programs worldwide often have little opportunity to develop and practice writing during their college years. Several studies have shown that graduates in computer science and engineering seldom have the required writing skills needed for work in a professional setting [2], [3]. Gibbs [4] argues that many students leave secondary school without proficiency in reading, writing and communication. Furthermore, even those who have good language skills run a risk of losing them while studying at university because there are too few opportunities to write essays and for getting good-quality feedback on writing assignments.

This paper discusses an Experiential Learning project, which included designing, delivering and assessing two argumentative writing assignments for 141 freshmen studying computer science or computer engineering. Three researchers, including the course instructor, collaborate to investigate the development of a Natural Language Processing (NLP) application designed to assist instructors with providing students with prompt, reliable feedback on argumentative writing assignments. As part of their preliminary investigation, the project was planned for a freshman Academic Skills course at a university in the U.K. For these students there is sometimes a lack of contextual coherence due to the absence of opportunities for systematic inquiry when it comes to writing assignments. This is indicative of the fact that approaches to teaching writing traditionally differ from those used in STEM courses, in which students are assigned associated laboratory work allowing them to investigate hypotheses through actions and activities. Such explorations enable the development of knowledge through internal and external discourse; for example, by watching videos or through discussions with their peers.

Experiential learning involves immersing students in educational activities, and then encouraging them to reflect on the experience and develop new ways of thinking. Experiential learning is a constructivist process allowing the learner to expand their ideas through a process of inquiry and reflection, as is usually done in STEM. Students can work individually, as part of a group, or under the guidance of a facilitator. The ways in which HE institutions organise curriculum, integrate technology and infuse other resources to improve student outcomes have garnered scrutiny in recent decades [5]. Now that universities are being held accountable for the quality of their teaching through instruments such as the U.K. National Student Survey [6], the multidimensional nature of getting students to achieve desired outcomes has assumed a new urgency. Central to attaining the best academic results for students is constructive commentary that they can use as a scaffold. One concern is developing technology to assist instructors, especially those with large classes, in providing

Manuscript received *** This work is supported by the National Science Foundation under Grant NSF IIS 1847842.

P. Davies is with the College of Sciences and Human Studies, Prince Mohammad Bin Fahd University, Al Khobar, Kingdom of Saudi Arabia (e-mail: pdavies@pmu.edu.sa).

R. Passonneau and Y. Gao are with the Department of Computer Science and Engineering, Penn State University, Pennsylvania, USA (e-mail: rjp49@cse.psu.edu; yug125@cse.psu.edu).

S. Muresan is Department of Computer Science, Columbia University, New York, USA (e-mail: smara@columbia.edu).

high-quality, timely and consistent feedback to guide students as they experiment with writing to develop their written communication skills. The project involved providing students with two tasks, each asking them to analyze source material critically prior to writing an argumentative essay.

John Dewey [7] presents the idea that the process of learning should involve a cycle of doing and reflection to produce an awareness of the problem at hand, formulating a response, experiencing the consequences and finally modifying or confirming a proposed solution. Such a process of transformation involves concrete experiences as opposed to abstract conceptualization. More recently, Vygotsky [8] has been credited for providing the foundations for experiential learning. He contends that knowing, understanding and thinking all happen within a sociocultural context. His arguments are expanded by the American scholar Kolb [9] through an exploration of processing information via concrete experiences. The resulting experiential learning model involves a cycle of observation, formulation, testing and experiencing. In other words, we do something, experience its consequences, take action in response to these and then repeat the process, this time with a more developed understanding of what the process involves.

This paper provides guidance for practitioners seeking to integrate experiential learning in writing courses for STEM students, as well as information for researchers concerned with using NLP for building applications to support writing instruction. The project was set up to explore the following:

1) How and to what extent can experiential learning be used to develop argumentative writing skills of students enrolled in STEM programs?
2) How can technologies that help promote experiential learning in argumentative writing be developed?
3) How can the holistic impact of using such an approach be evaluated?

The next two sections of the paper address the first two of these questions. Each part begins with a review of related literature. Section IV discusses the impact the project had on students and what might constitute a full appraisal of the impact of learning technologies that promote experiential education. Section V concludes the paper.

II. Experiential Learning Contexts

A. Making a Case for Experiential Learning

This literature review section discusses the multifaceted interpretations of experiential learning, ranging from on-the-job training to engaging in and reflecting on work. It provides evidence-based arguments in favour of learning-by-doing.

1) Deweyan Viewpoint: The construct of learning-by-doing refers to the work expounded by Dewey in his review of educational philosophy—*Experience and Education*. One of his distinguishing suggestions is that learning is developed from within and is based on ideas formed by performing certain actions. Experience, he argues, is the basis of understanding. He furthers this by explaining that an experience comprising certain qualities, such as uniqueness and wholeness, can expand human perception and increases one's value of what is being experienced [10]. For him the ultimate purpose of an experience is to reawaken the senses, to see differently [11] ideas that might have been missed, and to validate that being studied. In describing transformative experiences, Dewey discusses educational possibilities rather than actualities, which raises rather than answers questions about how such experiences could be fostered in teaching [12]. Nevertheless, the theoretical underpinning of experiential learning developed over fifty years ago is now being implemented in the public and private sectors to foster experiential innovation in industry [13], and in HE through activities such as internships and fieldwork.

2) Modern Perspectives: HE institutions worldwide use experiential learning to develop in students 21st century skills and competencies, including empathy, resilience and collaboration [14], to better prepare them for an unpredictable world. Through volunteering, internships and field studies involving local and global communities, some online, students are expected to harness communal traits that enable them to become more integrated and better connected as human beings. Although underutilized [15], such community partnerships are seen as beneficial in preparing students for a life of work.

Experiential learning is a strategy used to integrate active, structured, and meaningful reflection into teacher preparation programs [16]. The aim is to develop in teachers additional knowledge and skills in readiness for the challenging situations they might encounter in schools. Learning by experience is central also to nursing education; for example, where students spend half of their studies doing hands-on practice [17]. Roakes and Norris-Tirrell [18] argue that practical situations provide uncertainties and complexities that cannot be replicated in the classroom. Thus, as with students of engineering, high importance is placed on the cyclical process of experiential learning to help connect textbook theories with real-life.

The traditional standardised testing approach prevalent in K-12 settings leaves teachers little room for experience-based learning. Studies connecting instructional practices with policies centred on accountability and rankings [19] highlight the pressures put on teachers by large class sizes and a lack of time to complete the syllabus, stifling their desires to make learning interesting and engaging for students. Consequently, teaching to the test becomes the only option available to them.

The disconnect between work expectations of young employees and their academic training is underscored even more during university. Teacher-dominated instruction remains the primary mode of teaching in HE notwithstanding existing research showing that experiential learning promotes teamwork [20] and develops critical-thinking skills [21]. Instead, attempts are made to enhance students' chances of gaining employment by including employability in the curriculum and providing initiatives, such as inviting alumni and other professional speakers, to help students garner some understanding of pursuing a career. Unfortunately, these approaches involve telling students what to do instead of showing them how. Another approach to preparing students for employment is through work-based learning, often available at some point during their college career. However, there is little agreement within institutions and indeed across countries [22] on how and when to implement student internships in the curriculum.

B. An Experiential Learning Approach

1) The Setting and Assignments: In the project there are around 200 first-year computer science and engineering students at a public university in the U.K. At the beginning of their studies these students are required to complete *Academic Skills*, a semester-long freshman course designed to enhance academic writing. The course aims to develop in students proficiency in writing and communications skills necessary for success in college and for future employment. The university attracts students from surrounding towns and cities, and is recognised for widening participation in HE. Thus, the project participants come from a wide range of socioeconomic and academic backgrounds. Most study full-time, but a small number are part-time students. The learning outcomes of the course are drawn from the UNESCO [23] definition of literacy, which centers on ensuring that students are able to "identify, understand, interpret, create, communicate and compute, using printed and written materials, as well as ... to solve problems in an increasingly technological and information-rich environment". The course leader is supported by six tutors. Course meetings are scheduled over six hours per week; on two days, each with an hour-long lecture followed by a two-hour hands-on workshop. A highly structured course design, based on interactive lessons and hands-on practice, is used to help deepen students' understanding of the content.

Academic Skills is run by the university each year, but this particular year the project took it over. The assessment included two argumentative essays described below. Students were required, first, to analyze critically reading material provided and summarize it in 150 to 250 words; next, to write a short argumentative essay (300 to 500 words), based on the reading, in response to a given prompt. They could choose one of the following three topics for the first essay.

- Autonomous Vehicles (AV): *will these change how we travel today?*
- Cryptocurrencies (Crypto): *are they the currencies of the future?*
- Cybercrime (Cyber): *will education and investment provide the solution?*

The areas of specialization of these students include cybersecurity, information technology and computer engineering. Thus for the first essay they had the opportunity to choose a topic they were already familiar with or interested in. For the second essay, all students had to respond to the same question—*Should artificial intelligence be used in teaching and learning?* This second essay task limited students to making arguments based only on material from two articles they were given. The two essays were designed to be developmental exercises, in that the second assignment was more difficult.

2) Rubric - Motivation and Design: Writing scales arose in the early 20th century to compare performance of schools and teachers [24], and only later were they developed within classroom contexts to provide guidance for students. It has been shown that analytic rubrics, where scores are assigned to distinct dimensions, have greater reliability than holistic rubrics [25]. Still, many studies highlight as problematic inconsistency among raters and scoring professionals [26] in applying analytic rubrics, due to lack of training or familiarity. Despite these debates, analytic rubrics can serve as an instructional tool to improve students' writing quality [27]. An analytic rubric was developed and used both for instruction and grading.

The rubric was designed through a collaborative process by the three researchers working on the project. The instructor, who is one of the researchers, has a background in educational technology whereas the other two specialize in applying NLP to educational data. Part of the investigation involved understanding how the rubric supports instruction in argumentative writing. The rubric contained explicit descriptions of performance characteristics, each corresponding to a point on a rating scale. Table I shows the four dimensions of the rubric, the dimension weights and sub-dimensions with the points assigned. Design of the rubric was guided by Ferretti's well-known argument rubric [28] and the Source-based Argument Scoring Attributes (AWC) [29]. Research has shown that the range of a rubric scale is important because it affects reliability and ability to make meaningful distinctions; more than seven levels lead to cognitive difficulty, and fewer levels produce sharper classification.

TABLE I
RUBRIC DIMENSIONS, WEIGHTS AND SUB-DIMENSIONS

Dimensions	Weights	Sub-Dimensions (with points assigned)
Content	3/7	quality, coverage, coherence (0 - 5 each)
Argument	2/7	claims, support, counterargument (0 - 10)
Conventions	1/7	lexis and grammar (0 - 5)
Referencing	1/7	sources and citations (0 - 5)

Timely feedback, as a means of supporting student learning, has long been advocated in the assessment literature. It provides a learner-centered approach in which, from a social-constructivist standpoint, students to learn from one another. A key component of authentic assessment, rubrics provide descriptive feedback and can also be used for self-assessment as a criterion of written work. It was therefore important that students were given the rubric together with the first assignment. The rubric was also used to provide formative feedback on the first essay before the second essay was assigned.

3) Essays - Assigning and Grading: A Universal Design for Learning framework guided the formulation of the assignments. This framework provides a structure for developing curriculum—learning outcomes, instructional methods, and assessment—and is composed of three main ideas: provide engagement, provide representation, and provide action and expression. Students were taught how to write good essays in two lectures. The first of these focused on the following four elements of writing argumentative essays: engaging with the prompt, formulating a claim, developing arguments and counterarguments, and concluding the essay [30]. Students spent the tutorial following the lecture honing in on each of these components. In [30] Black advocates that argumentative writing should be considered as an aesthetically pleasing art form and that on completion of the work, authors should have the satisfaction of knowing that they have *made* something".

The rubric was an instructional tool for explaining the purpose of the assignment. Six model essays on the each of the three topics were written by sophomores who had formed part of a Wise Crowd. The exemplars were used to highlight how students could maximize their scores. The first assignment was scored by three course tutors, with each person scoring essays with the same title; the number of students per topic was capped for even distribution. All tutors were trained to use the rubric consistently.

Once scored, examples from the first essays were used to point out avoidable mistakes. As a more learner-centered approach the students were encouraged to reflect on their scored essays. Taking this first attempt at argumentative writing as a point of departure, the second essay was assigned. The project participants had experienced the consequences of their first attempt, reflected on the feedback received and now has to go through the process again.

4) Reliability of the Rubric: The use of writing rubrics has engendered debate about their reliability and purpose. Educational intervention studies apply rubrics whose reliability is usually quite good. For example, Graham and Perin [31] in a meta-analysis of educational interventions exclude interventions with reliability lower than 0.60. Yet a large body of research, including [32], has documented how trained raters can exhibit different levels of severity on analytic rubric categories. There has also been skepticism about applying rubrics to classroom grading, due to subjectivity in interpreting rubric criteria and over-reliance by teachers on the rubric as authority. Turley and Gallagher [24] argue that the debate should not be about whether rubrics are good or bad, but about how to use them. They discuss how the interpretation of a rubric depends, in part, on developing a community of users who understand the language of the rubric criteria. However, very little work has been done to compare how rubrics are used for instruction with how they are used in scoring, or to examine the difference between their reliable use and inconsistent classroom use. The present work investigates these problems.

Although having an analytic rubric for both instruction and grading is beneficial for students, it is difficult to apply an analytic rubric reliably in the context of large numbers of students. This motivates the view that development of algorithms to support the application of a rubric is an important goal. Development of automated methods is facilitated by creation of training data for a specific rubric, consisting of a large number of examples where the rubric has been applied.

Two advanced undergraduates were trained to use the rubric over a period of seven weeks. Subsequently, each of them spent 10 hours per week re-scoring half the essays written by students for the first assignment. Their training included understanding the structure of argument writing and completing both essay assignments; see Table II.

TABLE II
SEVEN-WEEK RATER TRAINING

Week	Activity
1	Virtual meeting to review argument writing: assignment #1 and rubric #1
2	Raters write individual essays: one on AV, and one on Crypto or Cyber; then each rater applies rubric to essays written by the other rater
3	Virtual meeting to review raters' essays and assessments
4	Both raters assess the same three Crypto essays
5	Virtual meeting on their first round of assessment centering on discussion between raters and with all three researchers
6	Each rater assesses the same three additional Crypto essays
7	Feedback on the second round of assessment, with some discussion on assignment #2 and rubric #2

Pearson correlation on the content and argument components of the rubric was used to assess rater agreement. Their correlations with each other and with the assigned grades varied widely, from negative correlation to high correlation, after the raters applied the rubric to the first sample. This improved following the second round of three essays; the correlation between the raters was perfect on two, and poor on the third. After discussing these results, the raters were allowed to proceed independently with applying the rubric to the remaining essays.

To check reliability, each rater scored 28 essays per week for three weeks and 31 in the fourth week. Ten randomly selected essays were assigned to both raters for continued monitoring of their reliability. The correlations for the content and argument dimensions on the ten essays were generally high. They ranged from one low outlier of -0.52 to 1, with an average of 0.75, or 0.89 after dropping the outlier. The reliable raters had lower correlations with the assigned grades for the three-essay assignment, with averages ρ equal to 0.72, 0.63 and 0.59, respectively.

The reliability study shows that the rubric can be applied very reliably by specially trained raters and with moderate to low reliability in uncertain classroom contexts. It indicates the difficulty of using a fine-grained rubric in large classes, where teaching assistants do the grading, where students want to see their grades quickly, and where timely and specific feedback is beneficial.

III. TECHNOLOGY FOR EXPERIENTIAL LEARNING

A. Previous Work on Technology for Writing Instruction

A comprehensive review of research on instruction revealed that writing skills develop best given a formative assessment cycle. This involves successive stages of instruction to target specific learning goals, followed by assessments for which instructors provide feedback to help students scaffold their learning. Reliable and valid assessment is seen to be important as part of instruction. The time involved in regular assessments of writing, and the difficulties in assessing writing discussed in the previous section, both provide strong motivation for technological support for writing instruction.

A recent review of the impact of technology on writing instruction found the main strength to be increased student engagement in writing assignments, and support for peer collaboration [33]. The main drawback was that instructors found it challenging to integrate technology into the writing curriculum. A concurrent review compared forty-four tools intended to support academic writing instruction, the majority of which concern automated writing evaluation (AWE) [34]. Most of these tools focus on college-level English L1, or L1

combined with L2 learners, and are not tied to a specific domain or genre. Apart from pointing to the need to address languages other than English, the authors conclude with recommendations that align with several of their research goals; in particular, feedback linked to writing goals and genres, and to strategy instruction, meaning techniques for planning and revising text in general, or specific kinds of text such as persuasive writing. AWE has been used to generate feedback to support students' revisions. It also supports analytic or holistic rubrics, or content maps using specific tools and techniques, including Coh-Metrix [35], C-rater-ML [36], G-rubric [37], Coh-Viz [38], and PEG [39].

Automated support for revision feedback using analytic rubrics has been applied to second language learners' persuasive essays [40], college students' physics lab reports [41], and middle school students in English language arts (ELA) classes [42]. Liu *et al.* [40] developed machine-learned models by training on pairs of student sentences and teacher comments from a previously collected dataset of L2 learner essays. A comparison of teacher versus automated feedback for 104 students found that the automated feedback led to the same kinds of improvements between first and second drafts on four of seven classes. Park and Cho [41] investigated whether Coh-Metrix indices could predict peer reviews of lab reports in a study with 41 students. Eight out of fifty-four Coh-Metrix indices had modest but significant correlations with the human scores on the final drafts. Perin and Lauterbach [43] apply Coh-Metrix to community college students' summaries to predict four dimensions of an analytic rubric, and found a different set of Coh-Metrix features to be predictive than those identified in previous work. Wilson and Czik [44] conducted a study with eight 8th grade ELA classes where four classes received teacher feedback alone, and four received a combination of teacher and automated feedback from PEG [39]. PEG provides scores for six dimensions of writing quality (e.g., idea development, style, word choice), each on a 5-point scale. It combines natural language processing and machine learning techniques, using more than 500 variables to predict essay ratings assigned by expert raters. Results indicated that teachers gave more feedback on higher-level writing skills to students in the combined condition, and that reliance on PEG saved one-third to half the time it took to provide feedback without PEG. In a somewhat larger study, middle school students in ELA classes using PEG had more positive writing self-efficacy and higher scores on the state ELA test than a control group [42]. Other machine learning methods for predicting scores using all or some of the analytic rubric dimensions have also been investigated for college level L2 essays [45] and middle school argument writing [46].

Compared with analytic rubrics, there are fewer studies investigating automated support for holistic rubrics. A significant exception involves application of the ETS C-rater technology to facilitate teacher feedback on middle school students' revisions of short answers to science questions [47], [48]. In Gerard *et al.* [47], C-rater was adapted to analyze short answers in tests given in two sixth grade science units, so that the teacher could intervene more efficiently to strengthen student collaboration. This work extended a previous study involving seventh graders [48]. Other applications that investigate holistic rubrics include assessment of English proficiency in writing from sources [49], analysis of features of good writing in college-level students [50], [43], and analysis of science argumentation of high school seniors [51]. More recent work on similar short answer tasks compared three kinds of machine learning models, and found that pre-trained transformer models performed better than RNNs or support-vector machines [52].

Rubric-free methods have also been investigated. G-rubric [53], [37], is a modification of latent semantic analysis (LSA) [54], a method to create numeric vector representations of the meanings of words, where the number of vector dimensions is up to the investigator. G-Rubric converts LSA vector space with latent dimensions of meaning to a new vector space with semantic grounding i (e.g., 300), a fixed number of relevant concepts. It has been used to give college students iterative feedback during revision of source-based summaries [53], and with business students in a MOOC [55]. Concept maps are another rubric-free feedback method. Concept maps, a visualization used in education for decades [56], are graphs that depict explanatory knowledge, where nodes represent concepts and edges represent relations between them. Sung *et al.* [57] compared four conditions of feedback for sixth graders writing summaries over six weeks: none, LSA-based visualization, concept maps, and LSA plus concept maps. Students who received feedback all improved between pre- and post-test, and students with concept-map feedback outperformed the other two conditions. Coh-Viz tool automatically creates concept maps for individual sentences, similar to subject-predicate-object graphs [58], and has been tested with students studying education in a German university. Students' revisions based on concept map conditions had significantly greater improvements in cohesion over a baseline.

To summarize, studies show automated analysis can support formative assessment during writing instruction by helping the instructor to provide prompt feedback [40], [47], [48], to students while revising their drafts [44], [53], [55], which can lead to improved writing skills [42], [57]. Machine learning methods as used in PEG, C-rater-ML, G-Rubric and [40] generalize better than Coh-Metrix alone, although Coh-Metrix provides useful features for the machine learning approach used in [40]. Santamaría Lancho *et al.* [55] suggest that automated support could also be integrated with human grading to improve the consistency and reliability of summative assessment.

B. Content Annotation and Analysis: the Wise Crowd Method

Writing summaries of source texts has been found to be among the best instructional tools to develop students' reading and writing skills for conceptual knowledge [31]. Their use as a pedagogical tool requires a method to assess the conceptual quality of a summary, which in turn rests on the identification of the main ideas of the source texts being summarized. Many similar methods have been utilized in educational psychology, including expert consensus [59], ranking of propositional units in source texts [60], and successive elimination of less

important propositional units in source texts [61]. All these methods elicit explicit judgements of propositions. The present study relies on exploiting the notion of a wise crowd of experts who summarize the same sources [62]. Ideas that are expressed in more of the wise crowd summaries have higher weight. Table III illustrates a summary content unit (SCU) from five wise crowd summaries of the Autonomous Vehicles article. Four of the five summaries expressed the idea that use of public transportation might decrease with increased reliance on autonomous vehicles. Although the idea is expressed in different ways, all four expert summaries clearly express the same idea. Ideas in student summaries that match an SCU are credited with the corresponding SCU weight. Given five reference summaries, SCU weights can range from 1 to 5. Ideas in student summaries that do not match an SCU are assigned a weight of 0. The score assigned to a student summary normalizes the total sum of the weights of their ideas by the number of ideas in the student's summary, and by the average number of ideas in a reference summary. Thus a summary gets a higher score if the ideas expressed by the student match more of the high weighted SCUs, and if the student expressed a good proportion of the high weighted SCUs. The scores can then be explained to the student or instructor in terms of the overlap of ideas in the student's summary with the full repertoire of SCUs for a given text.

TABLE III
EXAMPLE OF A SUMMARY CONTENT UNIT (SCU): FOUR OF FIVE REFERENCE SUMMARIES, IDENTIFIED HERE AS A, B, C, AND D, EXPRESSED THE SAME IDEA (PROPOSITION) ABOUT THE POTENTIAL NEGATIVE IMPACT OF DRIVERLESS VEHICLES ON PUBLIC TRANSPORTATION.

Weight = 4	Label = Driverless vehicles will likely reduce reliance on public transport
Reference ID	Text
A	One of the main points is the displacement of public transport
B	the rise of autonomous vehicles will disrupt the current standing of public transport
C	even more people would switch from public transport
D	it could also have a negative impact upon the public transport systems

In previous work, it was shown that the wise crowd method for identifying important ideas in source texts, ranking them, and using the resulting ranked list to assess student summaries, correlates very well with a main-ideas-rubric used in an educational intervention (ρ=0.88) [59]. The method itself has been found to be highly reliable given four or five reference summaries [63]. Originally this method was applied through manual annotation procedure (see next paragraph). An automated approach to the assessment and feedback step has been developed [62] and, more recently, a fully automated approach called PyrEval that identifies and ranks the SCUs from a set of reference summaries, then uses the weighted SCUs to assess new summaries [64], was developed. PyrEval was tested on summaries from the Autonomous Vehicles and Cryptocurrency topics. Also described here is an extension to this annotation linking the SCUs from the summaries to propositions in the argument portion of a student essay.

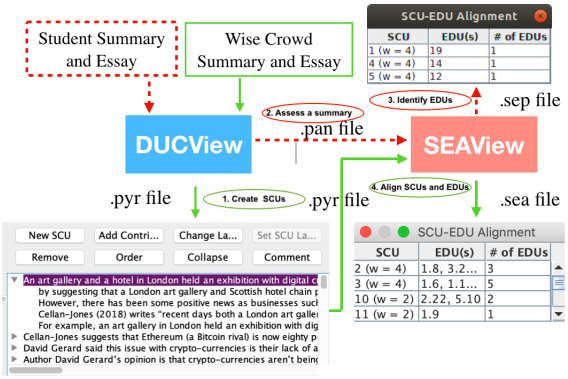

Fig. 1. Workflow diagram for content annotation, which use DUCView and SEAView. The green box and arrows indicate the flow of the wise crowd summaries to annotate the SCUs, and the box and arrows in dashed red lines show the flow of a student's essay, divided into summary and argument. The file extensions (pan, sep, etc.) indicate the underlying XML format, to differentiate the stages of annotation.

The analysis of the content of the students' essays asks how well the automated method of summary content analysis replicates the manual method, and how the automated method could support feedback on the rubric, either to help students revise the essay as a whole, or to give the instructor an overview of students' grasp of the content and their ability to draw on it to support their arguments. Recall that the assignments first asked students to summarize the source text or texts in 150 to 250 words, then to construct an argument addressing one of the prompts. Fig. 1 illustrates the work flow of the manual annotation. The reference summaries are annotated first to identify the SCUs, to create a pyr file (a list of SCUs derived from reference summaries is referred to as a pyramid). The pyr file is used to assess student summaries, with one pan file per student summary; in this step the propositions the student expresses are matched to the weighted SCUs. The new aspect of content annotation that has been added is for the argument part of a student's essay. Thus, Elementary Discourse Units (EDUs) are annotated [65], [66]; these are essentially individual clauses or propositions (sep file). In contrast to a summary of a source text, the quality of a student's argument is not expected to depend on how much of the same content is expressed as in a reference argument. On the other hand, of interest is how much of what they summarized from a source appears in their arguments, and what sort of content they use to frame their arguments. The last annotation step therefore involves matching the EDUs in a student's argument to the SCUs.

The automated wise crowd method performs fairly well on these summaries, as described in [64], with a Pearson correlation of 0.66 on the Autonomous Vehicle summaries when comparing the manual and automated summary content assessment, and a Pearson correlation of 0.72 for the Cryptocurrency summaries. Previously, the instructor found the content scores and justifications to be very useful, including cases where the tool gave low scores to written work that, on reflection, were scored favorably based on the writing fluency rather than the content [67]. For the present study, neither

the manual nor automated content scores on the summaries correlate well with the content dimensions of the rubric. This is because the rubric content dimensions of quality and coherence relate to the essay as a whole not to the summaries alone, and the students consulted other sources of their choosing to find evidence to support their arguments. Clearly, the content assessment of the students' summaries reflects their reading skills, which suggests further investigation into whether summarization skills might provide insight into how well students use external sources in their arguments.

Ongoing work on the content analysis of the students' essays includes investigation of the essays on the third topic (Cybercrime), and analysis of the relationship between the content and argumentation, particularly with regard to the overall structure of the students' essays. The next subsection describes the analysis of students' arguments.

C. Argumentation Annotation and Automated Analysis

Effective argumentative writing presents a claim, considers evidence in support of and against the claim, and demonstrates how the pros outweigh the cons. The project aimed to test whether argumentation features derived from coarse-grained argumentative discourse structure correlate well with the 6-point scale rubric that rate the *quality* of the argument. To do this, the first step was to label the argumentative part of the 37 Cryptocurrency essays using an annotation scheme generally used in argument mining [68]: *main claim*, *claim* and *premise/evidence* as argument components, and *support* and *attack* as argument relations. The advantage of a simple annotation scheme is two-fold: more reliable human annotation, and better performance of automatic methods to detect the argument structure. Two expert annotators with background in linguistics and argumentation performed the annotation that resulted in a gold standard set of 36 main claims, 559 claims, 277 premises, 560 support relations and 101 attack relations. A *proposition* was considered as the unit of annotation, given that premises are frequently propositions that conflate multiple clauses and sometimes even sentences [69].

The set of argumentative features introduced by Ghosh *et al.* [70] were used on the annotated essays to test whether they correlate with the argument quality scores obtained in the reliability study. The features are grouped in three categories: 1) features related to *argument components* (ArgC) such as the proportion of argumentative sentences (i.e., contain a main claim, claims and/or premise) and the number of argument components in an essay; 2) features related to *argument relations* (ArgR) such as the number of supported and unsupported claims and the number of attack relations (counterarguments); and 3) features related to the *typology of argument structure* (Str)—the number of argument chains and argument trees (see [70] for more details).

While Ghosh *et al.* [70] showed that these features correlate with the holistic essay score (low, medium and high) when applied to TOEFL persuasive essays, this study aims to test the effectiveness of these argumentation features in predicting the argument quality scores (scale of 0-5) obtained in the reliability study. Logistic Regression (LR) learners were used to evaluate them using quadratic-weighted kappa (QWK) against the human scores. QWK has been used for essay scoring [71], [70]. Table IV reports the results from a 5-fold cross validation setting for the three argumentation feature groups and their combination. The baseline feature is the essay length in sentences, since it has been shown to be highly correlated with essay scores [72].

TABLE IV
CORRELATION OF LR (5 FOLD CROSS-VALIDATION) WITH ARGUMENT QUALITY SCORES.

Features	Correlations
baseline	0.15
ArgC	0.27
ArgR	0.35
Str	0.17
baseline + ArgC	0.21
baseline + ArgR	0.26
baseline + STr	0.33
ArgC + ArgR + Str	0.41
baseline + ArgC + ArgR + Str	0.26

The best correlation is obtained when using all the argumentative features (ArgC+ArgR+Str), while ArgR is the best performing individual feature group. Moreover, all argumentation features outperform the baseline. Also, the *argument tree* feature correlates with high scoring essays, which is not surprising as these features capture the complexity of a well-written argument. In addition, top-scoring essays (with score 5) have a higher number of *attack* relations to the main claim, showing that these essays contain counterarguments, which is an aspect in the rubric. The number of claims supporting the main claim was negatively correlated with low scoring essays since students who received a low score, although forming arguments, failed to link them to their main claim. Similar to the work of Ghosh *et al.* [70], the number of supported claims correlate negatively with lower scoring essays, which show that students who receive low scores do not provide evidence for their claims. Another interesting observation from this analysis is that in the best essays (score 5) the ratio of argumentative sentences to total number of sentences was higher than for essays with a score of 4, whereas essays with a score of 4 were generally longer than the essays with a score of 5. That could also explain why the baseline feature (essay length) performed so poorly, since length alone is not indicative of argument quality.

The correlations scores were lower than the ones reported by Ghosh *et al.* [70], a finding that could have several explanations: 1) the number of essays is smaller, 37 compared to 107; 2) a 6-point scale rather than a 3-point one was used; and 3) the scale used reflects the argument quality and not an overall score. Looking at argument structure alone might not be enough; instead, both the structure and the semantics of arguments need to be examined in order to predict the argument quality more reliably. [73]. This approach will be pursued in future work.

IV. EVALUATING THE EXPERIENTIAL LEARNING PROJECT

Foregoing discussions have addressed the first two research questions. Central to these arguments is the importance of

providing timely, formative feedback to enhance students' understanding of what is expected in argumentative writing. Rubrics provide an avenue for doing this, and may be used in a manner that integrates the feedback into instruction so that students begin to view their grades as more than just a score. However, the experiential aspects of the learning cycle hinges on the premise that the writing activities students engage in must also include reflection.

Expectancy-value theory [74] can be used to assess the holistic impact of experiential learning. This framework depicts learners' motivation as based on their expectancy of success and the value attributed to a given task. Students with low self-efficacy, typically find understanding and acting on formative feedback difficult. As a result, they tend not to engage in reflection. Motivation and engagement are intrinsic to all learning. The latter is currently receiving much attention in HE because students' opinions now play a principal role in rating teaching and learning in tertiary education. Consequently, utilizing innovative teaching techniques to produce positive academic outcomes for students is no longer an option; it gives impetus to experiential learning.

Students were asked to complete a questionnaire about the rubric after receiving feedback to the first of the two essay assignments. First, students were asked the following initial questions requiring Yes/ No responses.

1) Did you get the mark you were expecting on Argumentative Essay 1?
2) Did you use the rubric?

Those who had used the rubric were questioned further:

3) When was the rubric used? *before starting the assignment, while doing it, after completing it; or some combination of all of these*
4) What was it used for? *to understand the requirements, as a guide, for checking; or some combination of all of these*
5) Do you feel the rubric helped you achieve your desired score? If yes, explain how.

Out of the 84 respondents, almost two-third of them (63%) reported using the rubric in one or more of the ways suggested above. Thirty-four percent of these students believed that the passing score they received was due to having access to the rubric. Others said they probably would have achieved the same score without using the rubric, with only 11% suggesting that the rubric did not help them at all. The Wise Crowd were also questioned about their experiences with using the rubric. In addition to saying it helped them understand different aspects of the writing process, one said that it made "very clear what the expectations were". Another suggested that without the rubric, he "wouldn't have known exactly what was expected". What was even more striking is that more than 65% of students who attempted both essays scored the same or a higher mark on the second essay. There is limited evidence to suggest that the feedback from the first assignment aided their performance on the second essay. Instead, what seems more likely is that the second essay allowed students to master the experiential learning approach they had been exposed to in the first assignment.

V. Conclusion

Discussions in this paper have centered on how argumentative writing instruction for STEM students could be aligned with the learn-by-doing approaches used in these fields. In sum, experiential learning provides an alternative to simply telling these students what is expected of them in college assignments. Furthering Dewey's conjecture—experience as reawakening—it is suggested that conceptual understanding of writing arguments could be fostered by engaging students in transformative experiences that allow them to confirm and extend their ideas. Part of this process involves providing them with a rubric for instruction and assessment. The reliability study showed that rubrics can be applied reliably outside classroom contexts but classroom grading tends to be less reliable, which can be attributed to time pressures and lack of training. The study also provides a benchmark for training and testing the algorithms being developed ultimately to support instructors or raters. Previous work has shown that automated methods for applying analytic rubrics can reduce the demands on instructors' time, and can be used fruitfully to support students in revising their written work.

Source-based writing draws on reading comprehension as well as on writing skills, which are skills that support each other, but which require different kinds of instruction [75]. The automated analysis of the summaries that students included in their essays shows that the automated summary analysis performs well. It could therefore be used by instructors to provide feedback on students' understanding of sources. The same features that have proved useful for automated analysis of argument in previous work [76] are shown to be the most predictive of the feature sets used here as well. In the context of freshman writing courses, especially for STEM students, Work on integrating automated assessment of argumentation and subject matter content is already in progress. The availability of 21st century educational technologies now make it possible to support new pedagogical approaches through automation; at least for transparency, uniformity and competence. Although automated scoring is still subject to much debate, what is being advocated here is automation with human intervention. Automation needs to be designed with student-centred learning in mind.

Acknowledgment

The authors would like to thank the two NSF-funded REUs, Connor Heaton and Annie Qin Sui, who participated in the reliability studies and other project activities.

References

[1] S. Cottrell, *Critical thinking skills: Effective analysis, argument and reflection*. Macmillan International Higher Education, 2017.
[2] D. F. Radcliffe, "Innovation as a meta attribute for graduate engineers," *International Journal of Engineering Education*, vol. 21, no. 2, pp. 194–199, 2005.
[3] C. S. Nair, A. Patil, and P. Mertova, "Re-engineering graduate skills: a case study," *European journal of engineering education*, vol. 34, no. 2, pp. 131–139, 2009.
[4] N. Gibb, *Reading: the next steps: supporting higher standards in schools*. Department for Education, 2015.

[5] L. A. Schindler, G. J. Burkholder, O. A. Morad, and C. Marsh, "Computer-based technology and student engagement: a critical review of the literature," *International Journal of Educational Technology in Higher Education*, vol. 14, no. 1, pp. 1–28, 2017.

[6] J. T. Richardson, J. B. Slater, and J. Wilson, "The national student survey: development, findings and implications," *Studies in Higher Education*, vol. 32, no. 5, pp. 557–580, 2007.

[7] J. Dewey, *Experience and education*. New York: Macmillan Company, 1938.

[8] L. S. Vygotsky, *Mind in society: The development of higher psychological processes*. Harvard University Press, 1980.

[9] D. A. Kolb, *Experiential learning: experience as the source of learning and development*. Prentice-Hall, Inc., 1984.

[10] J. Dewey, *Experience and nature*. Courier Corporation, 1958, vol. 471.

[11] P. W. Jackson, *John Dewey and the lessons of art*. Yale University Press, 2000.

[12] J. Dewey, *Human nature and conduct*. Southern Illinois University Press Carbondale, IL, 1988.

[13] S. Parisi, V. Rognoli, and M. Sonneveld, "Material tinkering. an inspirational approach for experiential learning and envisioning in product design education," *The Design Journal*, vol. 20, no. sup1, pp. S1167–S1184, 2017.

[14] G. Harfitt, "Community-based experiential learning in teacher education," in *Oxford Research Encyclopedia of Education*. Oxford University Press, 2019.

[15] A. Rosenstein, C. Sweeney, and R. Gupta, "Cross-disciplinary faculty perspectives on experiential learning," *Contemporary Issues in Education Research (CIER)*, vol. 5, no. 3, pp. 139–144, 2012.

[16] C. Girvan, C. Conneely, and B. Tangney, "Extending experiential learning in teacher professional development," *Teaching and teacher education*, vol. 58, pp. 129–139, 2016.

[17] L. Ruholl and R. Boyajian, "The senior wellness project: focus on experiential learning," *Teaching and Learning in Nursing*, vol. 2, no. 3, pp. 72–79, 2007.

[18] S. L. Roakes and D. Norris-Tirrell, "Community service learning in planning education: A framework for course development," *Journal of Planning Education and Research*, vol. 20, no. 1, pp. 100–110, 2000.

[19] K. J. Anderson, "Science education and test-based accountability: Reviewing their relationship and exploring implications for future policy," *Science Education*, vol. 96, no. 1, pp. 104–129, 2012.

[20] K. Obenchain and B. Ives, "Experiential education in the classroom and academic outcomes: For those who want it all," *Journal of Experiential Education*, vol. 29, no. 1, pp. 61–77, 2006.

[21] W. F. Heinrich, G. B. Habron, H. L. Johnson, and L. Goralnik, "Critical thinking assessment across four sustainability-related experiential learning settings," *Journal of Experiential Education*, vol. 38, no. 4, pp. 373–393, 2015.

[22] M. Klein and F. Weiss, "Is forcing them worth the effort? benefits of mandatory internships for graduates from diverse family backgrounds at labour market entry," *Studies in Higher Education*, vol. 36, no. 8, pp. 969–987, 2011.

[23] UNESCO, "Recommendation on adult learning and education," 2016, Paris, UNESCO. [Online]. Available: https://unesdoc.unesco.org/ark:/48223/pf0000245179

[24] E. D. Turley and C. W. Gallagher, "On the uses of rubrics: Reframing the great rubric debate," *The English Journal*, vol. 97, no. 4, pp. 87–92, 2008.

[25] A. Jonsson and G. Svingby, "The use of scoring rubrics: Reliability, validity and educational consequences," *Educational Research Review*, vol. 2, no. 2, pp. 130–144, 2007.

[26] J. Trace, V. Meier, and G. Janssen, "I can see that: Developing shared rubric category interpretations through score negotiation," *Assessing Writing*, vol. 30, pp. 32–43, 2016.

[27] T. H. Sundeen, "Instructional rubrics: Effects of presentation options on writing quality," *Assessing Writing*, vol. 21, pp. 74–88, 2014.

[28] R. P. Ferretti, C. A. MacArthur, and N. S. Dowdy, "The effects of an elaborated goal on the persuasive writing of students with learning disabilities and their normally achieving peers." *Journal of Educational Psychology*, vol. 92, no. 4, p. 694, 2000.

[29] H. A. Gallagher, N. Arshan, and K. Woodworth, "Impact of the national writing project's college-ready writers program in high-need rural districts," *Journal of Research on Educational Effectiveness*, vol. 10, no. 3, pp. 570–595, 2017.

[30] S. Black, *Crack the Essay: Secrets of Argumentative Writing Revealed*. Gramercy House Publishing, 2018.

[31] S. Graham and D. Perin, "A meta-analysis of writing instruction for adolescent students," *Journal of Educational Psychology*, vol. 99, p. 445–476, 2007.

[32] T. Eckes, *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessment*. Frankfurt, Germany: Peter Lang, 2011.

[33] C. Williams and S. Beam, "Technology and writing: Review of research," *Computers & Education*, vol. 128, pp. 227 – 242, 2019.

[34] C. Strobl, E. Ailhaud, K. Benetos, A. Devitt, O. Kruse, A. Proske, and C. Rapp, "Digital support for academic writing: A review of technologies and pedagogies," *Computers & Education*, vol. 131, pp. 33 – 48, 2019.

[35] A. C. Graesser, D. S. McNamara, M. M. Louwerse, and Z. Cai, "Coh-metrix: Analysis of text on cohesion and language," *Behavior Research Methods, Instruments, & Computers*, vol. 36, pp. 193–202, 2004.

[36] M. Heilman and N. Madnani, "ETS: Domain adaptation and stacking for short answer scoring," in *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013)*. Atlanta, Georgia, USA: Association for Computational Linguistics, June 2013, pp. 275–279. [Online]. Available: https://www.aclweb.org/anthology/S13-2046

[37] R. Olmos, G. Jorge-Botana, J. A. León, and I. Escudero, "Transforming selected concepts into dimensions in latent semantic analysis," *Discourse Processes*, vol. 51, no. 5-6, pp. 494–510, 2014.

[38] A. Lachner, C. Burkhart, and M. Núckles, "Mind the gap! automated concept map feedback supports students in writing cohesive explanations," *Journal of Experimental Psychology: Applied*, vol. 23, no. 1, pp. 29–46, 2017.

[39] E. B. Page, "Project essay grade: PEG," in *Automated essay scoring: A cross-disciplinary perspective*, M. D. Shermis and J. C. Burstein, Eds. Mahwah, NJ: Lawrence Erlbaum Associates Publishers, 2003, pp. 43–54.

[40] M. Liu, Y. Li, W. Xu, and L. Liu, "Automated essay feedback generation and its impact on revision," *IEEE Transactions on Learning Technologies*, vol. 10, no. 4, pp. 502–513, 2017.

[41] J. Park and K. Cho, "Toward the integration of peer reviewing and computational linguistics approaches," *Journal of Educational Computing Research*, vol. 55, no. 1, pp. 123–144, 2016.

[42] J. Wilson and R. Roscoe, "Automated writing evaluation and feedback: Multiple metrics of efficacy," *Journal of Educational Computing Research*, vol. 58, no. 1, pp. 87–125, 2019.

[43] D. Perin and M. Lauterbach, "Assessing text-based writing of low-skilled college students," *International Journal of Artificial Intelligence in Education*, vol. 28, no. 1, pp. 56–78, 2018.

[44] J. Wilson and A. Czik, "Automated essay evaluation software in english language arts classrooms: Effects on teacher feedback, student motivation, and writing quality," *Computers & Education*, vol. 100, pp. 94 – 109, 2016.

[45] T. Sladoljev-Agejev and J. Šnajder, "Using analytic scoring rubrics in the automatic assessment of college-level summary writing tasks in L2," in *Proceedings of the Eighth International Joint Conference on Natural Language Processing (IJCNLP) (Volume 2: Short Papers)*. Taipei, Taiwan: Asian Federation of Natural Language Processing, 2017, pp. 181–186. [Online]. Available: https://www.aclweb.org/anthology/I17-2031

[46] Z. Rahimi, D. Litman, R. Correnti, E. Wang, and L. Matsumura, "Assessing students' use of evidence and organization in response-to-text writing: Using natural language processing for rubric-based automated scoring," *International Journal of Artificial Intelligence in Educatio*, vol. 27, no. 4, pp. 694–728, 2017.

[47] L. Gerard, A. Kidron, and M. Linn, "Guiding collaborative revision of science explanations," *International Journal of Computer-Supported Collaborative Learning*, vol. 14, pp. 291–324, 2019.

[48] L. F. Gerard and M. C. Linn, "Using automated scores of student essays to support teacher guidance in classroom inquiry," *Journal of Science Teacher Education*, vol. 27, no. 1, pp. 111–129, 2016.

[49] B. Beigman Klebanov, N. Madnani, J. Burstein, and S. Somasundaran, "Content importance models for scoring writing from sources," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Baltimore, Maryland: Association for Computational Linguistics, June 2014, pp. 247–252.

[50] D. S. McNamara, S. A. Crossley, and P. M. McCarthy, "Linguistic features of writing quality," *Written Communication*, vol. 27, no. 1, pp. 57–86, 2010.

[51] M. Zhu, O. L. Liu, and H.-S. Lee, "The effect of automated feedback on revision behavior and learning gains in formative assessment of scientific argument writing," *Computers & Education*, vol. 143, p. 103668, 2020.

[52] B. Riordan, S. Bichler, A. Bradford, J. K. Chen, K. Wiley, L. Gerard, and M. C. Linn, "An empirical investigation of neural methods for content

scoring of science explanations," in *Proceedings of the 15th Workshop on Innovative Use of NLP for Building Educational Applications*, 2020, pp. 135–144.
[53] R. Olmos, G. Jorge-Botana, J. M. Luzón, J. I. Martin-Cordero, and J. A. León, "Transforming LSA space dimensions into a rubric for an automatic assessment and feedback system," *Information Processing & Management*, vol. 52, no. 3, pp. 359–373, 2016.
[54] T. K. Landauer and S. T. Dumais, "A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge," *Psychological Review*, vol. 104, no. 2, pp. 211–240, 1997.
[55] M. Santamaría Lancho, M. Hernández, Á. Sánchez-Elvira Paniagua, J. M. Luzón Encabo, and G. de Jorge-Botana, "Using semantic technologies for formative assessment and scoring in large courses and MOOCs," *Journal of Interactive Media in Education*, vol. 2018, no. 1, 2018.
[56] J. D. Novak and A. J. Cañas, "Theoretical origins of concept maps, how to construct them, and uses in education," *Reflecting Education*, vol. 3, no. 1, pp. 29–42, 2007.
[57] Y.-T. Sung, C.-N. Liao, T.-H. Chang, C.-L. Chen, and K.-E. Chang, "The effect of online summary assessment and feedback system on the summary writing on 6th graders: The lsa-based technique," *Computers & Education*, vol. 95, pp. 1 – 18, 2016.
[58] A. Lachner, C. Burkhart, and M. Núckles, "Formative computer-based feedback in the university classroom: Specific concept maps scaffold students' writing," *Computers in Human Behavior*, vol. 72, pp. 459 – 469, 2017.
[59] D. Perin, R. H. Bork, S. T. Peverly, and L. H. Mason, "A contextualized curricular supplement for developmental reading and writing," *Journal of College Reading and Learning*, vol. 43, no. 2, pp. 8–38, 2013.
[60] A. L. Brown and J. D. Day, "Macrorules for summarizing texts: The development of expertise," *Journal of Verbal Learning and Verbal Behavior*, vol. 22, pp. 1–14, 1983.
[61] R. E. Johnson, "Recall of prose as a function of the structural importance of the linguistic units," *Journal of Verbal Learning and Verbal Behavior*, vol. 9, no. 1, pp. 12–20, 1970.
[62] R. J. Passonneau, A. Poddar, G. Gite, A. Krivokapic, Q. Yang, and D. Perin, "Wise crowd content assessment and educational rubrics," *International Journal of Artificial Intelligence in Education*, vol. 28, no. 1, pp. 29–55, 2018.
[63] R. J. Passonneau, "Formal and functional assessment of the pyramid method for summary content evaluation," *Natural Language Engineering*, vol. 16, no. 2, pp. 107–131, Apr. 2010.
[64] Y. Gao, C. Sun, and R. J. Passonneau, "Automated pyramid summarization evaluation," in *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 404–418. [Online]. Available: https://www.aclweb.org/anthology/K19-1038
[65] W. C. Mann and S. A. Thompson, "Rhetorical structure theory: Toward a functional theory of text organization," *Text-Interdisciplinary Journal for the Study of Discourse*, vol. 8, no. 3, pp. 243–281, 1988.
[66] D. Marcu, "The rhetorical parsing, summarization, and generation of natural language texts," Ph.D. dissertation, University of Toronto, Canada, 1997.
[67] Y. Gao, P. M. Davies, and R. J. Passonneau, "Automated content analysis: A case study of computer science student summaries," in *Proceedings of the 13th Workshop on Innovative Use of NLP for Building Educational Applications*. New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 264–272. [Online]. Available: https://www.aclweb.org/anthology/W18-0531
[68] C. Stab and I. Gurevych, "Annotating argument components and relations in persuasive essays," in *Proceedings of the 25th International Conference on Computational Linguistics (COLING 2014)*, 2014, pp. 1501–1510.
[69] C. Hidey, E. Musi, A. Hwang, S. Muresan, and K. McKeown, "Analyzing the semantic types of claims and premises in an online persuasive forum," in *Proceedings of the 4th Workshop on Argument Mining*, September 2017.
[70] D. Ghosh, A. Khanam, Y. Han, and S. Muresan, "Coarse-grained argumentation features for scoring persuasive essays," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 2016, pp. 549–554.
[71] N. Farra, S. Somasundaran, and J. Burstein, "Scoring persuasive essays using opinions and their targets," in *Proceedings of the 10th Workshop on Innovative Use of NLP for Building Educational Applications*. Denver, Colorado: Association for Computational Linguistics, June 2015, pp. 64–74. [Online]. Available: http://www.aclweb.org/anthology/W15-0608
[72] M. Chodorow and J. Burstein, "Beyond essay length: Evaluating e-rater®'s performance on toefl® essays," *ETS Research Report Series*, vol. 2004, no. 1, pp. i–38, 2004.
[73] B. B. Klebanov, C. Stab, J. Burstein, Y. Song, B. Gyawali, and I. Gurevych, "Argumentation: Content, structure, and relationship with essay quality," in *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, 2016, pp. 70–75.
[74] A. Wigfield, "Expectancy-value theory of achievement motivation: A developmental perspective," *Educational psychology review*, vol. 6, no. 1, pp. 49–78, 1994.
[75] J. Fitzgerald and T. Shanahan, "Reading and writing relations and their development," *Educational Psychologist*, vol. 35, no. 1, pp. 39–50, 2000.
[76] Y. Gao, A. Driban, B. X. McManus, E. Musi, P. Davies, S. Muresan, and R. J. Passonneau, "Rubric reliability and annotation of content and argument in source-based argument essays," in *Proceedings of the 14th Workshop on Innovative Use of NLP for Building Educational Applications*, 2019, pp. 507–518.

Patricia Marybelle Davies (Member, IEEE) received a Doctorate in Educational Technology from the University of Manchester, U.K. in 2013. She also holds Master's degrees in Mathematics, from the UC Berkeley, and in Educational Technology, from Columbia University. Her research critically examines educational technology applications for advancing student learning. She is an Associate Professor in the College of Science and Human Studies at Prince Mohammad bin Fahd University in Saudi Arabia. From 2002 to 2014 she was Head of Computer Science at an American curriculum college in the UK. She is a Fellow of the Higher Education Academy, has membership of various other professional bodies including the IEEE and the British Computer Society. She is Co-Convenor for the Educational Technology SIG of the British Educational Research Association. In 2016, 2017 and 2018, she was awarded Google Educator Grants to provide professional development for GCSE and A-Level teachers of computing and computer science.

Rebecca Jane Passonneau received her Ph.D. in 1985 from the University of Chicago Department of Linguistics. Her main area of research is natural language processing, which she has pursued at many academic and industry research labs. She joined the Department of Computer Science and Engineering at Pennsylvania State University in 2016. Her work is reported in over 120 publications in journals and refereed conference proceedings, and has been funded through nearly 30 sponsored projects, from 14 sources, including government agencies, corporate sponsors, corporate gifts, and foundations. She is a member of many professional organizations, is currently on the editorial board of the International Journal of Artificial Intelligence in Education, and other academic journals.

Smaranda Muresan received her PhD in Computer Science from Columbia University in 2006. She is a Research Scientist at the Data Science Institute at Columbia University. Her research expertise is natural language processing (NLP), focusing on argument mining and persuasion, figurative language understanding and generation, and NLP applications in education and public health. From 2008 to 2013, she was a faculty member in the School of Communication and Information at Rutgers University where she co-founded the Laboratory for the Study of Applied Language Technologies and Society. There she received a Distinguished Achievements in Research Award. Dr. Muresan is a board member of the North American Association for Computational Linguistics (2020-2021). She co-organized Workshops on Argument Mining and Figurative Language Processing at NAACL/ACL. She is the co-chair of the series of New York Academy of Sciences' Symposia on Natural Language, Dialog and Speech. Her research, funded by NSF, DARPA and IARPA, has led to over 80 publications.

Yanjun Gao is currently a PhD candidate in Computer Science and Engineering at Pennsylvania State University. Since 2017, she has been a Research Assistant in NLP Lab, Pennsylvania State University. Her research focuses on developing AI applications for education using NLP techniques, including proposition identification, semantic representation for discourse, and summarization evaluation. She recently served on the committee for the China National Conference on Chinese Computational Linguistics (CCL, 2020) and 1st Workshop in NLP Evaluation and Comparison (Eval4NLP, 2020).