Analytical Techniques for Developing Argumentative Writing in STEM: A Pilot Study

Patricia Marybelle Davies, *Member, IEEE*, Rebecca Jane Passonneau, Smaranda Muresan, and Yanjun Gao

***Abstract*—*Contribution:* Demonstrates how to use experiential learning (EL) to improve argumentative writing. Presents the design and development of a natural language processing (NLP) application for aiding instructors in providing feedback on student essays. Discusses how EL combined with automated support provides an analytical approach to improving written-communication skills.**

***Background:* High-quality, timely, feedback is an effective way to improve students' writing. However, large class sizes and limited instructor backgrounds often make formative feedback impossible. Recent trends, including lowering entry requirements, have added to these challenges. Assistive technologies for implementing inclusive education provide viable solutions.**

***Research Questions:* 1) How and why can EL be used to develop argumentative writing skills in university STEM students? 2) How can technologies be developed to support using EL in teaching writing? and 3) How might the holistic impact of using such analytic techniques be evaluated?**

***Methodology:* Participants in an EL project were assigned two essays in sequence. They were given instructions on making good arguments and shown how to use an analytic rubric to maximize their scores. The essays were hand scored by tutors who provided scores for each dimension of the rubric. Subsequently, the content and argumentation of the essays were analyzed using NLP techniques to obtain independent scores. Qualitative data were also collected.**

***Findings:* The project produced transformative writing experiences for the participants. It showed how analytical techniques help improve writing skills and how relevant automated instructor assistance can be developed using NLP technologies.**

***Index Terms*—Argumentation, content annotation, experiential learning (EL), higher education (HE), natural language processing (NLP), rubric reliability, STEM, written communication.**

Manuscript received 7 February 2021; revised 2 August 2021; accepted 16 September 2021. Date of publication 11 October 2021; date of current version 15 August 2022. This work was supported by the National Science Foundation under Award 1847853 and Award 1847842. *(Corresponding author: Patricia Marybelle Davies.)*

Patricia Marybelle Davies is with the College of Sciences and Human Studies, Prince Mohammad Bin Fahd University, Al Khobar 31952, Saudi Arabia (e-mail: pdavies@pmu.edu.sa).

Rebecca Jane Passonneau and Yanjun Gao are with the Department of Computer Science and Engineering, Penn State University, Pennsylvania, PA 16802 USA (e-mail: rjp49@cse.psu.edu; yug125@cse.psu.edu).

Smaranda Muresan is with the Department of Computer Science, Columbia University, New York, NY 10027 USA (e-mail: smara@columbia.edu).

Digital Object Identifier 10.1109/TE.2021.3116202

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 adequate writing skills for work in a professional setting [2], [3]. In fact, the problem starts earlier. As Gibbs [4] argues, many students in the U.K. leave secondary school without proficiency in reading, writing, and communication. In the U.S., student achievement in reading and writing from the National Assessment of Educational Progress points to a serious crisis in writing instruction: most students do not achieve grade proficiency [5]. Furthermore, even those secondary graduates with good language skills may run a risk of diminishing them once at university because there are too few opportunities to write essays and for getting good-quality feedback on writing assignments in STEM courses.

The research reported here demonstrates how analytical techniques could be used to help STEM students develop an understanding of and the skills for writing good arguments. An experiential learning (EL) project, which involved designing, delivering, and assessing two argumentative essays for first-year undergraduates, is presented. Three researchers, including the course instructor, collaborated on 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 the preliminary investigation, a pilot study involving 141 computer science and engineering first-year students at a U.K. university was conducted. For these students, there is sometimes a lack of contextual coherence due to the dearth of opportunities for systematic inquiry in writing assignments. This is indicative of the fact that traditional approaches to teaching writing differ from those used in STEM courses with laboratory work in which students have the opportunity to investigate hypotheses through actions and activities. Such explorations enable the development of knowledge through internal and external discourses; for example, by watching videos or through discussions with their peers.

EL involves immersing students in educational activities and then encouraging them to reflect on the experience and develop new ways of thinking. It is a constructivist process that includes cycles of inquiry and reflection, a typical approach in STEM fields. Students can work individually, as part of a group, and with or without the guidance of a facilitator. The ways in which HE institutions organize curriculum, integrate technology, and infuse other resources to improve student outcomes have garnered scrutiny in recent decades [6]. Now that universities are being held accountable for the quality of their teaching through instruments such as the U.K. National Student Survey [7], the multidimensional nature of student achievement has assumed a new urgency. Central to students' developing good writing skills is constructive commentary that they can use as a scaffold. Consequently, of concern is the availability of technologies for assisting instructors, especially those with large classes, to provide high-quality, timely, and consistent feedback to guide students as they experiment with writing to develop their written communication skills. The EL project involved providing students with two tasks, each asking them to analyze source material critically prior to writing an argumentative essay.

This article hopes to provide guidance for practitioners seeking to integrate EL in university writing courses for STEM students, as well as information for researchers concerned with using NLP for building applications to support writing instruction. The EL project was set up to explore the following.

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

Following this introduction—Section I, the remainder of the article is organized as follows. Section II provides background information on EL and previous work on technology for aiding writing instruction. The methodology reported in Section III is divided into two main parts. First, there is the design of the teaching aspects of the EL project and related rubrics, plus the delivery and assessment of the essays. The second part discusses development and use of the NLP application for supporting writing instruction. Section IV discusses the impact the project had on students and what might constitute a full appraisal of the benefits of learning technologies that promote experiential education. Finally, Section V concludes the report and offers ideas on future directions.

II. Background

A. Experiential Learning Contexts

The construct of EL or learning-by-doing refers to the work expounded by Dewey [8]. For him, the ultimate purpose of an experience is to reawaken the senses, to uncover ideas that might have been missed, and to validate what is being studied. More recently, Vygotsky [9] have been credited for providing the foundations for EL. His claims are that thinking, understanding, and knowing happen within a context based on social and cultural factors. The American scholar Kolb [10] expands on these ideas to produce an EL model involving a cycle of observing, formulating, testing, and experiencing. In other words, 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. EL can, therefore, be seen as an analytical process that allows students to develop enduring understandings.

HE institutions worldwide use EL to develop in students 21st century skills and competencies—empathy, resilience, and collaboration [11]—needed to prepare them better 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. However, such activities are often viewed as extracurricular or co-curricular, separate yet in conjunction with what is presented in textbooks. For a variety of reasons, EL is seldom used in lecture rooms. The EL project is distinctive in that it is situated within the lectures of a university course.

B. Higher Education Contexts

Seeking to tap into the potential of HE to transform individuals, local communities, and society at large, the U.K. Government introduced a *Widening Participation in HE* agenda in 2014 [12]. One measure of an HE institution's performance in this scheme is the proportion of students from disadvantaged backgrounds they admit each year. Ensuring this metric is kept high has given rise to a variety of nontraditional approaches to recruiting students. These include running on-campus Bridge Programs and Taster Days, University Ambassadors visiting schools, and lowering entry requirements. Widening participation initiatives generally focus on the aspirations of applicants rather than their academic preparedness. Additionally, the migration toward a hybrid part-public, part-private funding system in the U.K. has prompted many universities to intensify their campaign for even more students [13]. Once admitted, these students become the responsibility of those who teach. It is instructors, often with no additional training, who must ensure that these students progress to graduation.

A primary concern is assessing these students. The move away from HE students being assessed solely on final examinations happened over the past five decades [14]. Nowadays, overall student assessment is determined by their performance on a series of tasks spread out over the term. Assignments are given in almost every field through projects, dissertations, and theses. The increase in student numbers, added to a lack of STEM instructors' knowledge of scoring essays, makes providing formative feedback on writing tasks challenging. As a result, students miss out on receiving information that could be integral to improving their writing skills. Scholars have argued that being able to receive comments and suggestions that help students revise their writing deepens understandings [15]; when received in a timely and systematic manner, formative feedback could have a lasting impact. Concerns

about the inadequacies of feedback on HE assessment tasks are well documented via the U.K. National Student Survey and the Australian Course Experience Questionnaire [16]. Assistive technologies, including NLP applications for implementing inclusive education by providing prompt feedback on essay drafts, offer new opportunities for improving students' writing performance.

C. Experiential Learning Project Context

A large proportion of the participants in the pilot study were admitted under a widening participation initiative. The essays assigned were two of the five tasks they had to complete in a writing course. These were to be done individually, whereas the other three assignments involved working in teams. The challenges posed by grading individual student work are amplified by reductions in per capita funding. A cohort of 215 students enrolled in academic skills is typically assigned one course leader and six tutors. Previously, each student had his or her essay marked by the tutor responsible for their section of the course. Grading was done by hand and without a rubric to facilitate consistency. The lack of a congruous process often resulted in wide variations in the scores across sections. One repercussion of this approach was that many papers had to be regraded by the course leader as part of the moderation process. Such concerns have prompted many instructors to explore automated scoring as an alternative.

D. Previous Work on Technology for Writing Instruction

A comprehensive review of research on instruction indicates 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. There is little work, however, on whether reliability of assessment can be achieved in classroom settings.

Existing literature discussed above together with the present study provide evidence that reliable classroom assessment is difficult to achieve. This section highlights the promise of combining human and automated assessment to ease instructors' assessment burden and improve reliability.

The time involved in traditional methods for assessing writing and the difficulties discussed in the previous section provide strong motivation for technological support for writing instruction. While there is little work on reliability of analytic rubrics in classroom settings, there is research on the reliability of so-called constructed response questions (open-ended prompts), where students provide short answers in a few sentences. Studies of rater cognition and validity for constructed response questions find that automated methods tend to be more consistent than human raters, and that the variability across humans observed in both TAs and raters is difficult to overcome [17]. A similar case has been made in a study that compares human and machine scoring of constructed responses to assess science teachers' pedagogical content knowledge [18].

Reviews of recent papers summarizing automated writing assessment tools and studies indicate that these tools can improve student engagement and support for peer collaboration [19], and provide feedback to generate students' revision. The main drawback noted in [19] is that instructors found it challenging to integrate technology in the writing curriculum. A review of 44 tools for supporting academic writing instruction showed that most tools rely on automated writing evaluation (AWE) techniques used across a variety of subject domains and genres [20]. AWE tools have been defined as consisting of two components, one for scoring and one for feedback [21]. Apart from pointing to the need to address languages other than English, the authors of the tool review found promising results in 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 builds on many distinct automated technologies and tools that can support analytic or holistic rubrics, including Coh-Metrix [22], C-rater-ML [23], G-rubric [24], Coh-Viz [25], and PEG [26].

AWE technology to support revision based on providing feedback that automatically applies an analytic rubric shows promise across multiple subject areas. These include second language learners' persuasive essays [27], college students' physics lab reports [28], and middle school students in english language arts (ELA) classes [29]. In [27], 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 [28] investigated the ability to predict peer reviews of lab reports in a study with 41 students, where eight Coh-Metrix indices had modest but significant correlations with the human scores. However, the other work that attempted to replicate findings from the use of Coh-Metrix found a different set of Coh-Metrix features to be predictive from those identified in the previous work [30]. The PEG tool provides scores for six dimensions of writing quality (e.g., idea development, style, and word choice), each on a 5-point scale, based on NLP and machine learning techniques. A few recent studies of PEG for writing revision in ELA classes found PEG to reduce teacher effort, and to lead to improved scores on standardized tests [29], [31]. Other machine learning methods for predicting scores on analytic rubrics have also been investigated for college level essays from second language (L2) learners [32] and middle school argument writing [33].

Rubric-free methods for AWE have also been investigated. G-rubric [24], [34] is a modification of latent semantic analysis (LSA) [35], 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 the 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 [34], and with business students in a MOOC [36].

Concept maps are another rubric-free feedback method. Concept maps, a visualization tool that has been used in education for decades [37], are graphs that depict explanatory

knowledge, where nodes represent concepts and edges represent relations between them. Sung *et al.* [38] compared four conditions of feedback for sixth graders in the U.S. writing summaries over six weeks: 1) no feedback; 2) LSA-based visualization; 3) concept maps; and 4) LSA plus concept maps. Students who received feedback all improved between pretest and posttest, and students with concept-map feedback outperformed students with the other two conditions. The Coh-Viz tool automatically creates concept maps for individual sentences, similar to subject-predicate-object graphs [39], and has been tested with students studying education in a German university. Students' revisions based on concept map conditions showed 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 to students [27], [40], [41] and that feedback can support student revision [31], [34], [36], which can, in turn, lead to improved writing skills [29], [38]. Machine learning methods as used in PEG, C-rater-ML, G-Rubric, and the method reported in [27] generalize better than Coh-Metrix alone, although Coh-Metrix provides useful features for the machine learning approach used in [27]. Lancho *et al.* [36] suggested that automated support could also be integrated with human grading to improve the consistency and reliability of summative assessment.

III. Methodology

A. Experiential Learning Teaching Approach

1) Setting—Participants and Assignments: Participants in the project comprise around 200 first-year computer science and engineering students at a public university in the U.K. They are required to complete an academic skills module in their first year. The semester-long course aims to develop in students the proficiency in writing and communication skills necessary for success in college and future employment. The university attracts students from surrounding towns and cities. 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 [42] 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 $\cdots$ to solve problems in an increasingly technological and information-rich environment." The course leader is supported by six tutors. Classes are scheduled over 6 h per week, with an hour-long lecture followed by a 2-h hands-on workshop on each of two days. A highly interactive workshop design was used to help deepen students' understanding of the lectures.

The EL project included two argumentative essays described below. Each student chose one of the following three topics for the first essay. First, they had to analyze critically reading material provided and summarize it in 150–250 words. Next, they were asked to write a short argumentative essay (300–500 words), based on their reading, addressing one of the following prompts.

TABLE I
Rubric Dimensions, Weights, and Subdimensions

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)

1) *Autonomous Vehicles (AV):* Will these change how we travel today?
2) *Cryptocurrencies (Crypto):* Are they the currencies of the future?
3) *Cybercrime (Cyber):* Will education and investment provide the solution?

The areas of specialization of the participants 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 were asked to discuss the same question—*Should artificial intelligence be used in teaching and learning?* They were limited to making arguments based only on material from two articles they were given. The two essays were designed to be developmental exercises with the second assignment extending the writing skills the students had developed in the first essay.

2) Rubric—Motivation and Design: Writing scales arose in the early 20th century to compare performance of schools and teachers [43], 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 [44]. Still, many studies highlight as problematic inconsistency among raters and scoring professionals [45] when applying analytic rubrics, due to the lack of training or familiarity. Despite these debates, analytic rubrics can serve as an instructional tool to improve students' writing quality [46]. 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 subdimensions with the points assigned. Design of the rubric was guided by Ferretti's well-known argument rubric [47] and the source-based argument scoring attributes (AWC) [48]. The 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.

Timely instructor feedback has long been advocated in the assessment literature as a means of supporting student

learning. It provides a learner-centered approach in which, from a social-constructivist standpoint, students can scaffold their learning. A rubric that provides descriptive feedback is a key component of authentic assessment. It can also be used for self-assessment as a criterion of written work. It was, therefore, important that the 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 research-based Universal Design for Learning framework [49] guided the formulation of the assignments. The framework provides a structure for developing curriculum—learning outcomes, instructional methods, and assessment. It is composed of three main ideas: 1) providing engagement; 2) encouraging alternate forms of representation; and 3) facilitating action and expression. Students were first introduced to the four elements of writing argumentative essays: 1) engaging with the prompt; 2) formulating a claim; 3) developing arguments and counterarguments; and 4) concluding the essay [50]. Students spent the tutorial following the lecture exploring these components. Black [50] advocated 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 created something.

The rubric became an instructional tool used to explain the purpose of the assignment. Six model essays on each of the three topics were written by sophomores who were more experienced writers and had previously done well in the course. These 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 own 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 went through the writing process again.

4) Reliability of the Rubric: The use of rubrics for writing has engendered debate about their reliability and purpose. Educational intervention studies apply rubrics whose reliability is usually quite good. For example, Graham and Perin [51], in a metaanalysis of educational interventions, excluded interventions with reliability lower than 0.60. Yet a large body of research, including [52], 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 overreliance by teachers on the rubric as authority. Turley and Gallagher [43] argued 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 on comparing 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.

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

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, as discussed in detail in the following section. The 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 third-year undergraduates were trained to use the rubric over a period of seven weeks, by which point they had achieved measurably reliable performance in applying the rubric. Subsequently, each of them spent 10 h per week rescoring 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).

Pearson correlation [53] on the content and argument components (ArgC) of the rubric was used to assess rater agreement. After the raters applied the rubric to the first sample, their correlations with each other and with the assigned grades varied widely, from negative correlation to high correlation. This difference improved for 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 ($n = 10$), or 0.89 after dropping the one outlier ($n = 9$). The high average correlation of 0.75 indicates raters are consistent with one another, and dropping the one outlier shows that apart from an exceptional case of relatively low correlation, the raters had very high average correlation. The reliable raters

TABLE III
EXAMPLE OF A 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

had lower correlations, however, with the assigned grades for the three-essay assignment, with averages ρ equal to 0.72, 0.63 and 0.59, respectively. This shows that rater training leads to greater consistency in application of the rubric, and therefore, room for improvement for in-class grading.

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.

B. Technology for Supporting Experiential Learning

1) 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 [51]. 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 in the source texts being summarized. Many similar methods have been utilized in educational psychology, including expert consensus [54], ranking of propositional units in source texts [55], and successive elimination of less important propositional units in source texts [56]. 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 [57], [58]. Ideas expressed in more of the expert wise crowd summaries have higher weights. Content scores for a summary consist of summing the weights of the ideas expressed, and then normalizing, as explained below. The previous work on this method found that four to five expert wise crowd summaries are sufficient to ensure that the probability of misranking a pair of summaries is below 0.1 [58].

Table III illustrates a summary content unit (SCU) from five expert wise crowd summaries of the AV article. Four of the five summaries expressed the idea that the use of public transportation might decrease with increased reliance on autonomous vehicles. Although the idea is expressed in different ways, all four expert summaries express the same idea clearly. The 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. The 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 them. 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.

In the 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 ($\rho = 0.88$) [54]. The method itself has been found to be highly reliable with respect to manual identification of content units, correlations of expert wise crowd scores from different annotators, and ranking of automated summarizers [59]. Originally, this method was applied through manual annotation procedure (see next paragraph). An automated approach to the assessment and feedback step was developed [57] 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 [60]. PyrEval was tested on summaries from the AV 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.

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 150–250 words, and then to construct an argument addressing one of the prompts. Fig. 1 illustrates the workflow of the manual annotation through the use of two annotation tools: 1) DUCView and 2) SEAView. The reference summaries are annotated first to identify the SCUs and to create a pyr(amid) file (a list of SCUs derived from reference summaries is referred to as a pyramid). The pyr file serves as a content model that is used to assess student summaries. After an annotator matches the propositions, the student expresses to the weighted SCUs, and DUCView can export a single pan file for each student summary. The new aspect of content annotation that has been added is for the argument part of a student's essay. In this step, which uses SEAView, elementary discourse units (EDUs) are annotated [61], [62]; these are effectively individual clauses or propositions stored in the 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

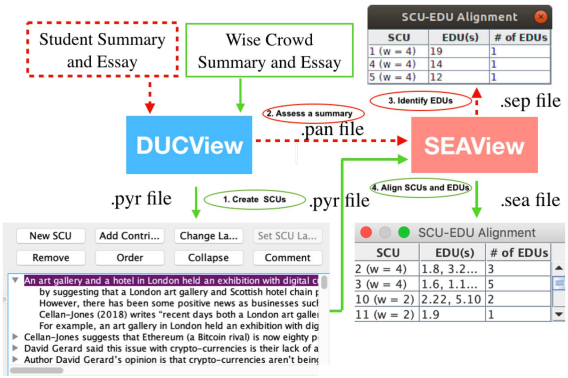

Fig. 1. Workflow diagram for content annotation, which use DUCView and SEAView. The green box and arrows indicate the flow of the expert wise crowd summaries to annotate the SCUs, and the box and arrows in dashed red lines show the flow of a student's essay, divided into summary and argument. The file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

the same content is expressed as in a reference argument. On the other hand, of interest is how much of what the students' 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 [60], 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. In a previous study [63], the instructor found the content scores and justifications to be very useful in understanding cases where the tool gave lower scores to written work compared to those given by the tutors. On inspection, this difference appeared to be due to some tutors assessing the writing fluency rather than the content of the essays. 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 of whether summarization skills might provide insight into how well students use external sources in their arguments.

The 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 section describes the analysis of students' arguments.

2) 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 [64]: *main claim*, *claim*, and *premise/evidence* as ArgC, and *support* and *attack* as argument relations (ArgR). The advantage of a simple annotation scheme is twofold: 1) more reliable human annotation and 2) 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 [65].

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 set of argumentative features introduced by Ghosh *et al.* [66] was used on the annotated essays to test whether they correlate with the argument dimension scores used in the reliability study. The features are grouped in three categories: 1) features related to ArgC, such as the proportion of argumentative sentences—those which contain a main claim, claims and/or premise—and the number of ArgC in an essay; 2) features related to 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 [66] for more details).

While Ghosh *et al.* [66] showed that these features correlate with the holistic essay score (low, medium, and high) when applied to persuasive essays from standardized tests for teaching English as a second language (TOEFL), 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 and evaluated using quadratic-weighted kappa (QWK) against the human scores. QWK is a standard metric used for essay scoring [66], [67]. 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 [68].

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.* [66], the number of supported claims correlates 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. That could also explain why the baseline feature (essay length) performed so poorly, since length alone is not indicative of argument quality.

The correlations were lower than the ones reported by Ghosh *et al.* [66], a finding that could have several explanations: 1) the number of essays was smaller, 37 compared with 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 [69]. Combining the work on content annotation and argument annotation achieves this goal, and is planned for future work.

IV. Discussion

A. How and Why EL Can Be Used to Develop Writing

Foregoing discussions have highlighted the importance of providing timely, formative feedback to enhance students' understanding of what is expected in argumentative writing assignments. Rubrics provide an avenue for this and can be used to integrate commentary into students' grades to give them meaning beyond just a numerical value. Also, presented above are methods for using EL to facilitate the development of good writing skills in university STEM students. Reasons why EL advances student writing go beyond clarifying expectation. Prompt, consistent feedback also provides assessment standards and benchmarks, promotes independent learning and self-regulation, and raises student aspirations [70]. Thereby, it becomes possible to impact the broader range of students with differing abilities and capacities who are now entering HE.

B. How the Holistic Impact of EL Can Be Assessed

The experiential aspects of the learning cycle hinge on the premise that the writing activities students engage in include reflection. The expectancy-value framework [71] depicts the motivations of learners as being based on their expectancy of success and the value credited to an assigned task. Thus, it can be used to assess the holistic impact of EL. 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 both intrinsic to all learning. They are 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 simply an option, rather it is now essential to HE and gives new impetus to EL.

C. Evaluating the EL Project

After receiving feedback on the first of two essay assignments, the students were asked to complete a questionnaire about the role of the rubric. First, they were asked the following 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.

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

Out of the 84 respondents, almost two-thirds of them (63%) reported using the rubric in one or more of the ways suggested above, and of these, 34% 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. 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 EL approach they had been exposed to in the first assignment. The expert 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 "would not have known exactly what was expected." These responses confirm research showing that rubrics could help students to comprehend tacit knowledge embedded in assessment.

D. Developing Technology to Support EL

Automated analysis of student writing has great potential given the demands on instructors to provide feedback and the difficulty of reliable classroom grading for large classes. The ongoing work on assessing the content of student writing includes applying the wise crowd method for feedback to students and teachers during classroom instruction in middle school science courses. It is noted that designing the prompts

for student writing tasks needs to proceed hand-in-hand with refinements to the automated assessment so that the content of students' writing is more focused on the concepts associated with the learning objectives. Furthermore, designing automated feedback that is useful to students is an additional step beyond identifying which ideas the student needs to articulate more fully. This shows the need for deep collaboration between educators and technology developers.

While providing feedback on students' ability to construct an argument (claim, premises, support, and attack relations) is important, ongoing work by the authors points to the additional need to highlight whether or not the arguments have reasoning flaws. Three areas of need relating to source-based argumentative essays were identified: 1) detect whether or not the evidence provided contradicts the source texts; 2) detect the argument schemes used (e.g., analogy, from example, causal); and 3) generate the unstated premise (enthymeme) that links a claim to the evidence provided. While the first problem helps provide feedback whether students misunderstood the source-text material, the latter two help highlight better whether or not reasoning flaws (fallacies) occurred. To tackle this, joint modeling of content and argumentation will be paramount.

V. Conclusion

This article has discussed how and why argumentative writing instruction for STEM students should be aligned with the learn-by-doing approaches used in these fields. It argues that EL provides an alternative to simply telling these students what is expected of them in writing assignments. Furthering Dewey's conjecture—experience as reawakening—it suggests that conceptual understanding of writing arguments could be fostered by engaging students in transformative experiences that allow them to confirm and extend their ideas. A part of this process involves providing them with a rubric for instruction and assessment. The reliability study shows 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 provides a benchmark for training and testing the algorithms being developed ultimately to support instructors or raters. Also shown is that the automated method for scoring summary content closely replicates hand scoring and supports feedback given via the rubric. This supports previous work demonstrating that automated methods for applying analytic rubrics can reduce the demands on instructors' time, and be used fruitfully to aid students in revising written work. The article goes further to suggest that the structure and semantics of arguments together could provide a more reliable prediction of the argument quality.

Source-based writing draws on reading comprehension as well as on writing skills, competencies which complement each other but require different kinds of instruction [72]. The results of the automated analysis of the student summaries show that the automated summary analysis performs well and 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 [73] are shown to be the most predictive of the feature sets used here, in the context of undergraduate writing courses designed for STEM students. The work on integrating automated assessment of argumentation and subject matter content is already in progress.

Teaching and assessment vary greatly across HEIs. As these institutions seek better ways of engaging students, the need for more personalized forms of assessment continues to gain relevance. The availability of 21st century educational technologies makes 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 that is designed with student-centered 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. Thanks also to the reviewers for the careful review and helpful feedback.

References

[1] S. Cottrell, *Critical Thinking Skills: Effective Analysis, Argument and Reflection*. London, U.K.: Macmillan Int. High. Educ., 2017.
[2] D. F. Radcliffe, "Innovation as a meta-attribute for graduate engineers," *Int. J. Eng. Educ.*, vol. 21, no. 2, pp. 194–199, 2005.
[3] C. S. Nair, A. Patil, and P. Mertova, "Re-engineering graduate skills—A case study," *Eur. J. Eng. Educ.*, vol. 34, no. 2, pp. 131–139, 2009.
[4] N. Gibb, *Reading: The Next Steps: Supporting Higher Standards in Schools*. London, U.K.: Dept. Educ., 2015.
[5] D. Reilly, D. L. Neumann, and G. Andrews, "Gender differences in reading and writing achievement: Evidence from the national assessment of educational progress (NAEP)," *Amer. Psychol.*, vol. 74, no. 4, pp. 445–458, 2019.
[6] L. A. Schindler, G. J. Burkholder, O. A. Morad, and C. Marsh, "Computer-based technology and student engagement: A critical review of the literature," *Int. J. Educ. Technol. High. Educ.*, vol. 14, no. 1, pp. 1–28, 2017.
[7] J. T. E. Richardson, J. B. Slater, and J. Wilson, "The national student survey: Development, findings and implications," *Stud. High. Educ.*, vol. 32, no. 5, pp. 557–580, 2007.
[8] J. Dewey, "Experience and education," *Educ. Forum*, vol. 50, no. 3, pp. 241–252, 1986.
[9] L. S. Vygotsky, *Mind in Society: The Development of Higher Psychological Processes*. Cambridge, MA, USA: Harvard Univ. Press, 1980.
[10] D. A. Kolb, *Experiential Learning: Experience as the Source of Learning and Development*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1984.
[11] G. Harfitt, "Community-based experiential learning in teacher education," in *Oxford Research Encyclopedia of Education*. Oxford, U.K.: Oxford Univ. Press, 2019.
[12] Department for Business, Innovation and Skills. *National Strategy for Access and Student Success*. Accessed: Sep. 23, 2020. [Online]. Available: https://www.gov.uk/government/publications/national-strategy-for-access-and-student-success
[13] S. Marginson, "Global trends in higher education financing: The United Kingdom," *Int. J. Educ. Develop.*, vol. 58, pp. 26–36, Jan. 2018.
[14] J. Heywood, *Assessment in Higher Education: Student Learning, Teaching, Programmes and Institutions*, vol. 56. London, U.K.: Jessica Kingsley Publ., 2000.
[15] J. C. Archer, "State of the science in health professional education: Effective feedback," *Med. Educ.*, vol. 44, no. 1, pp. 101–108, 2010.
[16] D. Boud and E. Molloy, "Rethinking models of feedback for learning: The challenge of design," *Assess. Eval. High. Educ.*, vol. 38, no. 6, pp. 698–712, 2013.
[17] I. I. Bejar, "Rater cognition: Implications for validity," *Educ. Meas. Issues Pract.*, vol. 31, no. 3, pp. 2–9, 2012.

[18] X. Zhai, K. C. Haudek, M. A. M. Stuhlsatz, and C. Wilson, "Evaluation of construct-irrelevant variance yielded by machine and human scoring of a science teacher PCK constructed response assessment," *Stud. Educ. Eval.*, vol. 67, Dec. 2020, Art. no. 100916.

[19] C. Williams and S. Beam, "Technology and writing: Review of research," *Comput. Educ.*, vol. 128, pp. 227–242, Jan. 2019.

[20] C. Strobl *et al.*, "Digital support for academic writing: A review of technologies and pedagogies," *Comput. Educ.*, vol. 131, pp. 33–48, Apr. 2019.

[21] D. Grimes and M. Warschauer, "Utility in a fallible tool: A multi-site case study of automated writing evaluation," *J. Technol. Learn. Assess.*, vol. 8, no. 6, pp. 4–44, 2010.

[22] A. C. Graesser, D. S. McNamara, M. M. Louwerse, and Z. Cai, "Coh-metrix: Analysis of text on cohesion and language," *Behav. Res. Methods Instrum. Comput.*, vol. 36, pp. 193–202, May 2004.

[23] M. Heilman and N. Madnani, "ETS: Domain adaptation and stacking for short answer scoring," in *Proc. 7th Int. Workshop Semantic Eval. (SemEval)*, Atlanta, GA, USA, Jun. 2013, pp. 275–279. [Online]. Available: https://www.aclweb.org/anthology/S13-2046

[24] R. Olmos, G. Jorge-Botana, J. A. León, and I. Escudero, "Transforming selected concepts into dimensions in latent semantic analysis," *Discourse Process.*, vol. 51, nos. 5–6, pp. 494–510, 2014.

[25] A. Lachner, C. Burkhart, and M. Núckles, "Mind the gap! automated concept map feedback supports students in writing cohesive explanations," *J. Exp. Psychol. Appl.*, vol. 23, no. 1, pp. 29–46, 2017.

[26] 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, USA: Lawrence Erlbaum Assoc. Publ., 2003, pp. 43–54.

[27] M. Liu, Y. Li, W. Xu, and L. Liu, "Automated essay feedback generation and its impact on revision," *IEEE Trans. Learn. Technol.*, vol. 10, no. 4, pp. 502–513, Oct./Dec. 2017.

[28] J. Park and K. Cho, "Toward the integration of peer reviewing and computational linguistics approaches," *J. Educ. Comput. Res.*, vol. 55, no. 1, pp. 123–144, 2016.

[29] J. Wilson and R. Roscoe, "Automated writing evaluation and feedback: Multiple metrics of efficacy," *J. Educ. Comput. Res.*, vol. 58, no. 1, pp. 87–125, 2019.

[30] D. Perin and M. Lauterbach, "Assessing text-based writing of low-skilled college students," *Int. J. Artif. Intell. Educ.*, vol. 28, no. 1, pp. 56–78, 2018.

[31] J. Wilson and A. Czik, "Automated essay evaluation software in English language arts classrooms: Effects on teacher feedback, student motivation, and writing quality," *Comput. Educ.*, vol. 100, pp. 94–109, Sep. 2016.

[32] T. Sladoljev-Agejev and J. Šnajder, "Using analytic scoring rubrics in the automatic assessment of college-level summary writing tasks in L2," in *Proc. 8th Int. Joint Conf. Nat. Lang. Process. (IJCNLP)* vol. 2. Taipei, Taiwan, 2017, pp. 181–186. [Online]. Available: https://www.aclweb.org/anthology/I17-2031

[33] Z. Rahimi, D. Litman, R. Correnti, E. Wang, and L. C. Matsumura, "Assessing students' use of evidence and organization in response-to-text writing: Using natural language processing for rubric-based automated scoring," *Int. J. Artif. Intell. Educ.*, vol. 27, no. 4, pp. 694–728, 2017.

[34] 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," *Inf. Process. Manage.*, vol. 52, no. 3, pp. 359–373, 2016.

[35] 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," *Psychol. Rev.*, vol. 104, no. 2, pp. 211–240, 1997.

[36] M. S. Lancho, M. Hernández, Á. S.-E. Paniagua, J. M. L. Encabo, and G. de Jorge-Botana, "Using semantic technologies for formative assessment and scoring in large courses and MOOCs," *J. Interact. Media Educ.*, vol. 2018, no. 1, p. 12, 2018, doi: 10.5334/jime.468.

[37] J. D. Novak and A. J. Cañas, "Theoretical origins of concept maps, how to construct them, and uses in education," *Reflecting Educ.*, vol. 3, no. 1, pp. 29–42, 2007.

[38] 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," *Comput. Educ.*, vol. 95, pp. 1–18, Apr. 2016.

[39] A. Lachner, C. Burkhart, and M. Núckles, "Formative computer-based feedback in the university classroom: Specific concept maps scaffold students' writing," *Comput. Human Behav.*, vol. 72, pp. 459–469, Jul. 2017.

[40] L. Gerard, A. Kidron, and M. Linn, "Guiding collaborative revision of science explanations," *Int. J. Comput. Supported Collab. Learn.*, vol. 14, pp. 291–324, May 2019.

[41] L. F. Gerard and M. C. Linn, "Using automated scores of student essays to support teacher guidance in classroom inquiry," *J. Sci. Teach. Educ.*, vol. 27, no. 1, pp. 111–129, 2016.

[42] UNESCO. *Recommendation on Adult Learning and Education*. Accessed: Aug. 15, 2019. [Online]. Available: https://unesdoc.unesco.org/ark:/48223/pf0000245179

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

[44] A. Jonsson and G. Svingby, "The use of scoring rubrics: Reliability, validity and educational consequences," *Educ. Res. Rev.*, vol. 2, no. 2, pp. 130–144, 2007.

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

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

[47] 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," *J. Educ. Psychol.*, vol. 92, no. 4, pp. 694–702, 2000.

[48] H. A. Gallagher, N. Arshan, and K. Woodworth, "Impact of the national writing project's college-ready writers program in high-need rural districts," *J. Res. Educ. Effect.*, vol. 10, no. 3, pp. 570–595, 2017.

[49] M. J. Capp, "The effectiveness of universal design for learning: A meta-analysis of literature between 2013 and 2016," *Int. J. Inclusive Educ.*, vol. 21, no. 8, pp. 791–807, 2017.

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

[51] S. Graham and D. Perin, "A meta-analysis of writing instruction for adolescent students," *J. Educ. Psychol.*, vol. 99, pp. 445–476, Aug. 2007.

[52] T. Eckes, *Introduction to Many-Facet Rasch Measurement: Analyzing and Evaluating Rater-Mediated Assessment*. Frankfurt, Germany: Peter Lang, 2011.

[53] D. Freedman, R. Pisani, and R. Purves, *Statistics*, 4th ed. New York, NY, USA: W. W. Norton, 2007.

[54] D. Perin, R. H. Bork, S. T. Peverly, and L. H. Mason, "A contextualized curricular supplement for developmental reading and writing," *J. Coll. Read. Learn.*, vol. 43, no. 2, pp. 8–38, 2013.

[55] A. L. Brown and J. D. Day, "Macrorules for summarizing texts: The development of expertise," *J. Verb. Learn. Verb. Behav.*, vol. 22, pp. 1–14, Feb. 1983.

[56] R. E. Johnson, "Recall of prose as a function of the structural importance of the linguistic units," *J. Verb. Learn. Verb. Behav.*, vol. 9, no. 1, pp. 12–20, 1970.

[57] R. J. Passonneau, A. Poddar, G. Gite, A. Krivokapic, Q. Yang, and D. Perin, "Wise crowd content assessment and educational rubrics," *Int. J. Artif. Intell. Educ.*, vol. 28, no. 1, pp. 29–55, 2018.

[58] A. Nenkova, R. Passonneau, and K. McKeown, "The pyramid method: Incorporating human content selection variation in summarization evaluation," *ACM Trans. Speech Lang. Process.*, vol. 4, no. 2, p. 4, May 2007.

[59] R. J. Passonneau, "Formal and functional assessment of the pyramid method for summary content evaluation," *Nat. Lang. Eng.*, vol. 16, no. 2, pp. 107–131, Apr. 2010.

[60] Y. Gao, C. Sun, and R. J. Passonneau, "Automated pyramid summarization evaluation," in *Proc. 23rd Conf. Comput. Nat. Lang. Learn. (CoNLL)*, Hong Kong, Nov. 2019, pp. 404–418. [Online]. Available: https://www.aclweb.org/anthology/K19-1038

[61] W. C. Mann and S. A. Thompson, "Rhetorical structure theory: Toward a functional theory of text organization," *Text Interdiscipl. J. Study Discourse*, vol. 8, no. 3, pp. 243–281, 1988.

[62] D. Marcu, "The rhetorical parsing, summarization, and generation of natural language texts," Ph.D. dissertation, Dept. Comput. Sci., Univ. Toronto, Toronto, ON, Canada, 1997.

[63] Y. Gao, P. M. Davies, and R. J. Passonneau, "Automated content analysis: A case study of computer science student summaries," in *Proc. 13th Workshop Innovat. Use NLP Build. Educ. Appl.*, New Orleans, LA, USA, Jun. 2018, pp. 264–272. [Online]. Available: https://www.aclweb.org/anthology/W18-0531

[64] C. Stab and I. Gurevych, "Annotating argument components and relations in persuasive essays," in *Proc. 25th Int. Conf. Comput. Linguist. (COLING)*, 2014, pp. 1501–1510.

[65] 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 *Proc. 4th Workshop Argument Min.*, Sep. 2017, pp. 11–21.
[66] D. Ghosh, A. Khanam, Y. Han, and S. Muresan, "Coarse-grained argumentation features for scoring persuasive essays," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguist.*, 2016, pp. 549–554.
[67] N. Farra, S. Somasundaran, and J. Burstein, "Scoring persuasive essays using opinions and their targets," in *Proc. 10th Workshop Innovat. Use NLP Build. Educ. Appl.*, Denver, CO, USA, Jun. 2015, pp. 64–74. [Online]. Available: http://www.aclweb.org/anthology/W15-0608
[68] M. Chodorow and J. Burstein, "Beyond essay length: Evaluating e-rater®'s performance on toefl® essays," ETS Res. Rep. Series, Princeton, NJ, USA, Rep. RR-04-04, 2004.
[69] B. B. Klebanov, C. Stab, J. Burstein, Y. Song, B. Gyawali, and I. Gurevych, "Argumentation: Content, structure, and relationship with essay quality," in *Proc. 3rd Workshop Argument Min. (ArgMining)*, 2016, pp. 70–75.
[70] C. Brooks, A. Carroll, R. M. Gillies, and J. Hattie, "A matrix of feedback for learning," *Aust. J. Teach. Educ.*, vol. 44, no. 4, p. 2, 2019.
[71] A. Wigfield, "Expectancy-value theory of achievement motivation: A developmental perspective," *Educ. Psychol. Rev.*, vol. 6, no. 1, pp. 49–78, 1994.
[72] J. Fitzgerald and T. Shanahan, "Reading and writing relations and their development," *Educ. Psychol.*, vol. 35, no. 1, pp. 39–50, 2000.
[73] Y. Gao *et al.*, "Rubric reliability and annotation of content and argument in source-based argument essays," in *Proc. 14th Workshop Innovat. Use NLP Build. Educ. Appl.*, 2019, pp. 507–518. [Online]. Available: https://www.aclweb.org/anthology/W19-4452

Patricia Marybelle Davies (Member, IEEE) received the Dr.Edu. degree in educational technology from the University of Manchester, Manchester, U.K., in 2013.

She is an Associate Professor with the College of Science and Human Studies, Prince Mohammad bin Fahd University, Al Khobar, Saudi Arabia. Her research critically examines educational technology applications for advancing student learning.

Dr. Davies was awarded the Google Educator Grants from 2016 to 2018. She is a Fellow of the Higher Education Academy. She is Co-Convenor for the Educational Technology SIG of the British Educational Research Association and has membership of other professional bodies.

Rebecca Jane Passonneau received the Dr.Phil. degree from the Department of Linguistics, University of Chicago, Chicago, IL, USA, in 1985.

She joined the Department of Computer Science and Engineering with Penn State University, Pennsylvania, PA, USA, 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. Her main area of research is natural language processing, which she has pursued at many academic and industry research labs.

Dr. Passonneau is a member of many professional organizations, and is currently on the editorial board of the *International Journal of Artificial Intelligence in Education*, and other academic journals.

Smaranda Muresan received the Dr.Phil. degree in computer science from Columbia University, Columbia University, New York, NY, USA, in 2006.

She is a Research Scientist with the Data Science Institute, Columbia University. From 2008 to 2013, she was a Faculty Member with the School of Communication and Information, Rutgers University, New Brunswick, NJ, USA, where she co-founded the Laboratory for the Study of Applied Language Technologies and Society and also she received the Distinguished Achievements in Research Award. Her research, funded by NSF, DARPA, and IARPA, has led to over 80 publications. 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.

Dr. Muresan is a Board Member of the North American Association for Computational Linguistics from 2020 to 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.

Yanjun Gao received the Dr.Phil. degree from the Department of Computer Science and Engineering, Penn State University, Pennsylvania, PA, USA, in 2021.

She is currently a Postdoctoral Research Associate with the University of Wisconsin-Madison, Madison, WI, USA. From 2017 to 2021, she was a Research Assistant with NLP Lab, Penn State University. Her research focuses on developing AI applications for education using NLP techniques, including proposition identification, semantic representation for discourse, and summarization evaluation.

Dr. Gao 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).