Using Learner Trace Data to Understand Metacognitive Processes in Writing from Multiple Sources

Mladen Raković
mladen.rakovic@monash.edu
Centre for Learning Analytics, Monash University
Clayton, Victoria, Australia

Yizhou Fan
yizhou.fan@ed.ac.uk
School of Informatics, The University of Edinburgh
United Kingdom

Joep van der Graaf
j.vandergraaf@pwo.ru.nl
Radboud University
Netherlands

Shaveen Singh
Shaveen.Singh@monash.edu
Centre for Learning Analytics, Monash University
Australia

Jonathan Kilgour
Jonathan.Kilgour@ed.ac.uk
School of Informatics, The University of Edinburgh
United Kingdom

Lyn Lim
lyn.lim@tum.de
Technical University of Munich
Germany

Johanna Moore
jmoore@staffmail.ed.ac.uk
School of Informatics, The University of Edinburgh
United Kingdom

Maria Bannert
maria.bannert@tum.de
Technical University of Munich
Germany

Inge Molenaar
inge.molenaar@ru.nl
Radboud University
Netherlands

Dragan Gašević
Dragan.Gasevic@monash.edu
Centre for Learning Analytics, Monash University
Australia

ABSTRACT

Writing from multiple sources is a commonly administered learning task across educational levels and disciplines. In this task, learners are instructed to comprehend information from source documents and integrate it into a coherent written composition to fulfil the assignment requirements. Even though educationally potent, multi-source writing tasks are considered challenging to many learners, in particular because many learners underuse monitoring and control, critical metacognitive processes for productive engagement in multi-source writing. To understand these processes, we conducted a laboratory study involving 44 university students. They engaged in multi-source writing task hosted in digital learning environment. Adding to previous research, we unobtrusively measured metacognitive processes using learners' trace data collected via multiple data channels and in both writing and reading space of the multi-source writing task. We further investigated how these processes affect the quality of a written product, i.e., essay score. In the analysis, we utilised both automatically and human-generated essay score. The rating performance of the essay scoring algorithm was comparable to that of human raters. Our results largely support the theoretical assumptions that engagement in metacognitive monitoring and control benefits the quality of written product. Moreover, our results can inform the development of analytics-based tools that support student writing by making use of trace data and automated essay scoring.

CCS CONCEPTS

• **Applied computing** → Education; • **Information systems** → Data mining.

KEYWORDS

writing from multiple sources, reading, monitoring, semantic similarity

ACM Reference Format:
Mladen Raković, Yizhou Fan, Joep van der Graaf, Shaveen Singh, Jonathan Kilgour, Lyn Lim, Johanna Moore, Maria Bannert, Inge Molenaar, and Dragan Gašević. 2022. Using Learner Trace Data to Understand Metacognitive Processes in Writing from Multiple Sources. In *LAK22: 12th International Learning Analytics and Knowledge Conference (LAK22), March 21–25, 2022, Online, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3506860.3506876

LAK22, March 21–25, 2022, Online, USA

ACM ISBN 978-1-4503-9573-1/22/03…$15.00
https://doi.org/10.1145/3506860.3506876

1 INTRODUCTION

Writing from multiple sources (or multi-source writing) is a common learning task assigned to many students enrolled in different

courses and levels of education. In these tasks, students are required to comprehend the information from multiple source texts that are often provided by a course instructor, and integrate this information into a written composition to fulfil task and genre requirements, e.g., write an argumentative essay arguing against the use of palm oil in cooking or write a literature review informing next steps in a class research project. Many educators thus administer multi-source writing tasks aiming to provide their students with opportunities to practice creating new knowledge from available sources of information and producing texts that effectively communicate new knowledge, e.g., an essay convincing readers to pursue an argument that palm oil should be abandoned in cooking.

However, despite its educational potential [16], multi-source writing is considered a highly demanding task as it involves different cognitive and metacognitive processes spanning reading comprehension, writing production, and self-regulation [3, 17, 19, 24, 46]. To succeed in this task, it is typically not enough for a learner to merely collect chunks of information across source documents and structure those chunks into the essay draft. Productive multi-source writers also metacognitively monitor and control their work throughout the writing session, i.e., they judge whether the evolving essay draft aligns to task and genre requirements (monitoring for task and goals, [19, 24]), and whether the source information was properly handled in the essay draft (monitoring for comprehension, [17]). As well, productive multi-source writers evaluate the discrepancies between the evolving essay draft and the essay they intend to produce in the task, and revise the essay to improve its quality (metacognitive control, [19]). Many learners, however, not fully engage in metacognitive monitoring and control when writing from multiple sources [9, 29, 33, 46] which, in turn, limits their work in the task and, ultimately, prevents them from producing high quality written compositions. Consequently, many multi-source writers need external support to boost their metacognitive processing and succeed in this task [39].

To tailor on-scale and timely support to those learners, it is therefore important to understand (1) how metacognitive processes unfold in a multi-source writing session and (2) how these processes relate to the quality of the evolving written product. To date, researchers have mainly measured metacognitive processes in multi-source writing by utilising self-report questionnaires administered after the writing session, e.g., [31, 45, 54], and collecting digital traces of writing processes during the session, e.g., [9, 10, 61]. These prior studies have documented important theorised relationships of metacognition with the application of genre knowledge, revision behaviours, and task performance. However, most studies on multi-source writing to date have exclusively analysed data about learners' writing behaviours, i.e., data collected while learners compose their essays in text editors, and rarely included data about learners' reading behaviours, i.e., data collected while learners operate on source texts (but see [56]). Second, even though researchers have documented that the quality of the evolving written product, including source-based essays, can be reliably scored in an automatic way [5, 51, 52, 63, 64], promising to inform at-scale support to multi-source writers as they progress in the task, researchers have yet to explore the association of automatic essay scores with traces of learners' metacognitive processes that occur during the task, and, subsequently, with learners' performance in the task as scored by humans.

To fill the above gaps in the literature and obtain a more comprehensive account of metacognitive processes and performance in multi-source writing, we conducted a lab study involving university students. The students engaged in a multi-source writing task using an experimental technology-enhanced learning environment (TEL). As students were using the TEL, we collected their behavioural trace data from multiple channels (i.e., navigational logs, keystrokes and eye-fixations), combined and analysed those data to detect metacognitive monitoring and control, and deepen the understanding of these processes as they emerge in both writing and reading space of the task. In particular, we aimed at investigating how metacognitive processes affect the quality of a written product represented as the essay score. To score the student essays observed in our study, we developed a new algorithm that measures specific essay characteristics, expected products of writer's engagement in metacognitive processing, e.g., the high internal coherence of the essay is expected to be a product of extensive essay revisions (i.e., metacognitive control) a writer enacted during the task. We thus investigated the association between traced measures of metacognition and automatically generated essay score. As well, we investigated the relationship between automatically generated essay score and the essay score generated by human raters, to validate the newly proposed automated scoring approach.

Our results indicate that increased engagement in metacognitive monitoring and control relate to an increase in essay scores. For that reason, these metacognitive processes should be promoted in challenging multi-source writing tasks, particularly for learners at risk of submitting low-quality essays. The results further demonstrate a strong and positive correlation between the automatically generated and human-generated essay score, indicating that the automatic essay scoring algorithm developed in our study can be used to score student essays with considerable reliability and thus identify low-quality essay drafts. Taken together, our results inform the development of analytics-based tools that support student writing by making use of trace data and automated essay scoring, a key contribution of our study.

2 RELATED WORK

2.1 Metacognitive monitoring and control as essential processes in multi-source writing

Different cognitive and metacognitive processes interweave in multi-source writing as writers strategically cycle between comprehending information from source documents and integrating it into a coherent written composition to fulfil the assignment requirements. For this reason, writing from multiple sources is considered a self-regulatory activity [18, 60]. Raković and Winne [46] synthesised research on multiple source comprehension, writing production and self-regulated learning (SRL). Informed by Winne and Hadwin's four-stage model of SRL [59], the authors broadly classified processes in multi-source writing as setting rhetorical goals, comprehending source information, producing written composition, and metacognitively monitoring and controlling work on the task [7, 17, 19, 23, 40, 43, 48, 49, 53]. We use this classification as a theoretical framework to guide the present study.

The rhetorical goals writers set for a multi-source writing task are guided by task requirements [17], e.g., requirements for genre, topic, format and length. At this stage, writers develop a mental representation of the task [48] which can include goals, planned procedures for approaching those goals and standards for monitoring (e.g., what information in sources should be judged relevant). Writers then initially read available sources to identify relevant information and work out topical connections among informational chunks scattered across the source documents. To facilitate source comprehension, writers often create additional resources, e.g., notes, highlighted selections and diagrams [4]. While drafting a composition, writers monitor whether the text they produced aligns with task requirements and goals (monitoring for task and goals; [19, 24]) and also whether they properly handled source information in their composition drafts (monitoring for comprehension; [17]). To this end, writers enact monitoring strategies, e.g., re-reading and reviewing [19]. Metacognitive monitoring often co-occurs with metacognitive control, i.e., writers modify their compositions to improve text quality, e.g., by editing and revising [19, 53]. Accordingly, metacognitive monitoring and control are considered central processes for writers to successfully navigate a complex and dynamic landscape of a multi-source writing task, ensure sufficient progress in the task and produce a composition that is of a high quality and in line with task requirements [6, 19, 24, 46].

Many learners, however, underuse metacognitive monitoring and control while writing from multiple sources [9, 29, 33, 46]. For example, they rarely review task instructions and highlights previously created in source documents, or perform insufficient text revisions to improve draft quality [9, 29]. These challenges hinder learners' performance on the task, preventing them from developing genre-appropriate writing skills, a major goal for which multi-source writing tasks are assigned. Many learners, therefore, need external support to succeed in those tasks. To effectively intervene and help multi-source writers engage in more productive metacognitive processing, it is critical to understand how these processes have been enacted throughout the task and how they affected the quality of the essay draft (i.e., evolving written product).

2.2 Measuring metacognitive processes in writing

In a limited number of studies, researchers have measured metacognitive processes in multi-source writing to understand how they predict writing performance. Undergraduate students participating in the study presented in [45] reported on different strategies they use in argumentative writing. The authors found that high-performing students monitored for task conditions (e.g., time arrangement or organisation), engaged in content checking and revising of a developing composition. Similarly, Karlen [31] revealed positive correlations among the university students' self-reported metacognitive strategy use and the scores those students received on an academic writing assignment. For instance, the high-achieving students deemed it useful to monitor their processes while writing (e.g., by double-checking whether the text they produced aligns with the assignment requirements) and to revise their drafts to improve content. Sun et al. [54] developed a questionnaire to elicit undergraduates' metacognitive experiences in a writing task and found that students who enacted metacognitive judgements throughout the task (e.g., judgements of effort and time spent on task) boosted their writing performance. Even though these findings illuminated important connections between metacognitive processing and performance in writing assignments [31, 45, 54], challenges with using self-report approaches to measuring those processes remain [57]. For instance, given the dynamics and complexity of processes in a multi-source writing task, a learner may often not be able to recall all the details about their engagement with the task, even right after completing the task.

With recent technological advances of research tools that can record fine-grained trace data generated in digital learning environments, new possibilities have emerged to unobtrusively study traces of writing processes as they dynamically unfold in authentic task settings. In particular, navigation logs, keyboard strokes, screen recordings and eye-tracking data have demonstrated a potential to more precisely and more comprehensively characterise metacognitive processes in writing [30, 38, 56]. For that reason, educational researchers, including those who study reading and writing, have begun increasingly utilising these data channels. The multi-source writing studies that utilised trace data have so far generated important findings unveiling how metacognitive processes affect the application of genre knowledge in writing [61], how automated feedback and scaffolding in a digital writing environment foster students' metacognitive processing during revision [10, 35], and what types of revisions can be associated with different students' characteristics (e.g., undergraduates and English as a second language) [2, 9].

The prior research, however, primarily focused on analysing learners' writing behaviours, i.e., how learners compose essays in their text editors, and rarely included learners' reading behaviours in those analyses, i.e., how learners operate on source texts. As a more recent exception, Vandermeulen et al. [56] studied the effects of process-related feedback tailored to students working on a multi-source synthesis writing task. To this end, the authors utilised not only trace data about text production (e.g., the total number of revisions modelled as text deletions/insertions and the mean duration of the revision bursts), but also data about reading comprehension (e.g., time spent on different digital sources measured at different points throughout the task). Writing processes automatically detected in this way were mirrored to learners to encourage them to reflect on different processes in their writing (e.g., use of time, revision and use of sources) and to compare those processes to the same kinds of processes enacted by high-achieving learners. This approach, in turn, helped learners improve the quality of their written products.

Sought to add to this line of research, in the current study, we collected the student trace data from multiple channels (i.e., navigational logs, keystrokes and eye-fixations) and combined and analysed data from these channels to deepen the understanding of monitoring and control, metacognitive processes that interplay in multi-source writing [19, 46, 53]. Importantly, we observed these processes as they occurred in both writing and reading space of the task. We also explored how the observed metacognitive processes affected essay scores. In this way, we laid a foundation for

future and innovative writing analytics solutions that provide on-scale support to learners, helping them engage in more productive multi-source writing and develop a quality written product.

2.3 Automatic scoring of essays developed from multiple sources

Current literature reports on several systems that are developed to automatically score source-based essays, e.g., [21, 28, 58, 63, 64]. These systems have demonstrated the overall ability to reliably measure different characteristics of student essays, e.g., use of source information [21, 64] and the quality of scientific explanations present in the essays [58]. However, despite these encouraging results, to our knowledge, no scoring model has been proposed in prior research to automatically score the essay by measuring its characteristics relative to metacognitive processes theorised to lead to the creation of those essay characteristics. In other words, it can be expected that learners who extensively monitored their multi-source compositions for task requirements/goals and for comprehension, and who extensively revised compositions during the task, would produce essays having a high semantic overlap with source texts and high compliance with the other requirements for the task (e.g., internal essay coherence and word count), which would ultimately lead to a high score generated by the algorithm. Consequently, as this scoring algorithm has yet to be developed, the link between automatic essay scores it produces and (1) traced measures of metacognition and (2) human-generated essay scores (to examine external validity) has yet to be studied.

To address this gap in the literature, we developed a new essay scoring algorithm that computes the coverage of source topics and internal text coherence in student essays. To this end, we harnessed the potential of the semantic analytical approaches in Natural Language Processing (NLP). So far, researchers have utilised different analytical approaches to examine semantic characteristics of written compositions relative to their corresponding source documents. Broadly, these approaches are based on (1) keyword matching/regular expressions and (2) high-dimensional semantic spaces (for an overview see [40]). The approaches based on keyword matching and regular expressions are typically customised to a particular task, i.e., keywords and linguistics patterns are predefined to capture semantics in a response. The approaches based on semantic spaces, i.e., high-dimensional representations of words in a form of vectors, on the other hand, are considered more robust to different tasks and compositions, as these approaches can reflect different meanings of an individual word relative to other words in a data set. The semantic space, therefore, can be used to underlie the comparison of different linguistic elements in the essay (e.g., paragraphs, sentences or words) to each other and/or to source documents, and compute the semantic overlap for each pair of the observed elements [40].

Researchers have demonstrated that text similarity measures obtained using semantic space approaches, such as latent semantic analysis (LSA, [37]) and word2vec [41], can be moderately-to-strongly related to those obtained by human scorers, e.g., the quality of sources' summary [11, 55] and the quality of internal text coherence [12, 14]. Given the documented promises of the semantic space approaches, we selected the word2vec algorithm to compute the coverage of source topics and internal text coherence in the essays (i.e., the product characteristics that reflect writer's metacognitive processing), and thus automatically obtain the essay score. Details about the algorithm implementation are provided in the Method section.

2.4 The present study

Inspired by prior research, we posited the following five hypotheses to guide our study: **H1:** Increase in the duration of metacognitive monitoring will be positively related to automatically generated essay score; **H2:** Increase in the duration of metacognitive control (essay revisions) will be positively related to automatically generated essay score; **H3:** Automatically generated essay score will be positively related to human-generated essay score; **H4:** There will be a positive indirect effect of metacognitive monitoring on the human-generated essay score through the automatically generated essay score and **H5:** There will be a positive indirect effect of metacognitive control on the human-generated essay score through the automatically generated essay score. Specifically, hypotheses H1, H2, H4 and H5 aimed at testing the previous theoretical propositions that engagement in metacognitive monitoring and control boosts learner's performance in multi-source writing tasks [6, 19, 24, 46]. These hypotheses aimed to provide insights into the yet-to-be-examined link between traced measures of metacognition and essay performance. Hypothesis H3 aimed at testing the external validity of the new essay scoring algorithm that measures specific essay characteristics resulting from writers engagement in metacognitive processing. Since the scoring algorithm is based on the semantic space approach previously documented to perform comparably well relative to human ratings (e.g., [11, 12, 14, 21, 55]), we expected the algorithm would generate valid essay scores compared to human raters. This hypothesis sought to provide insights into the yet-to-be-examined link between the essay scores generated using our proposed algorithmic approach and the essay scores generated by human raters. The hypothesized relations are represented in the path model in Figure 1.

In addition, we examined how learners' (1) prior knowledge, (2) time they took to initially read the task instructions page and scoring rubric, (3) time they took to engage in planning, and (4) time they took to initially study the source documents would associate with the automatic essay score (**research question 1 — RQ1**).

3 METHOD

3.1 Participants and context

A total of 46 students enrolled in a Dutch public university participated in the laboratory study. The participants' mean age was 21.70 years (SD=2.99 years). Due to technical issues during data collection, data generated by 2 participants were removed from the analysis and the analysis was conducted with 44 participants (34 women and 10 men). Of these, 39 participants were undergraduate and 5 were graduate students declaring different majors, e.g., psychology, communication science. Dutch was considered the native language of all the participants. The participants were paid 20 euro for participating.

The study session included a pre-test, a brief training on how to use the learning environment used in the study, a 45-minute

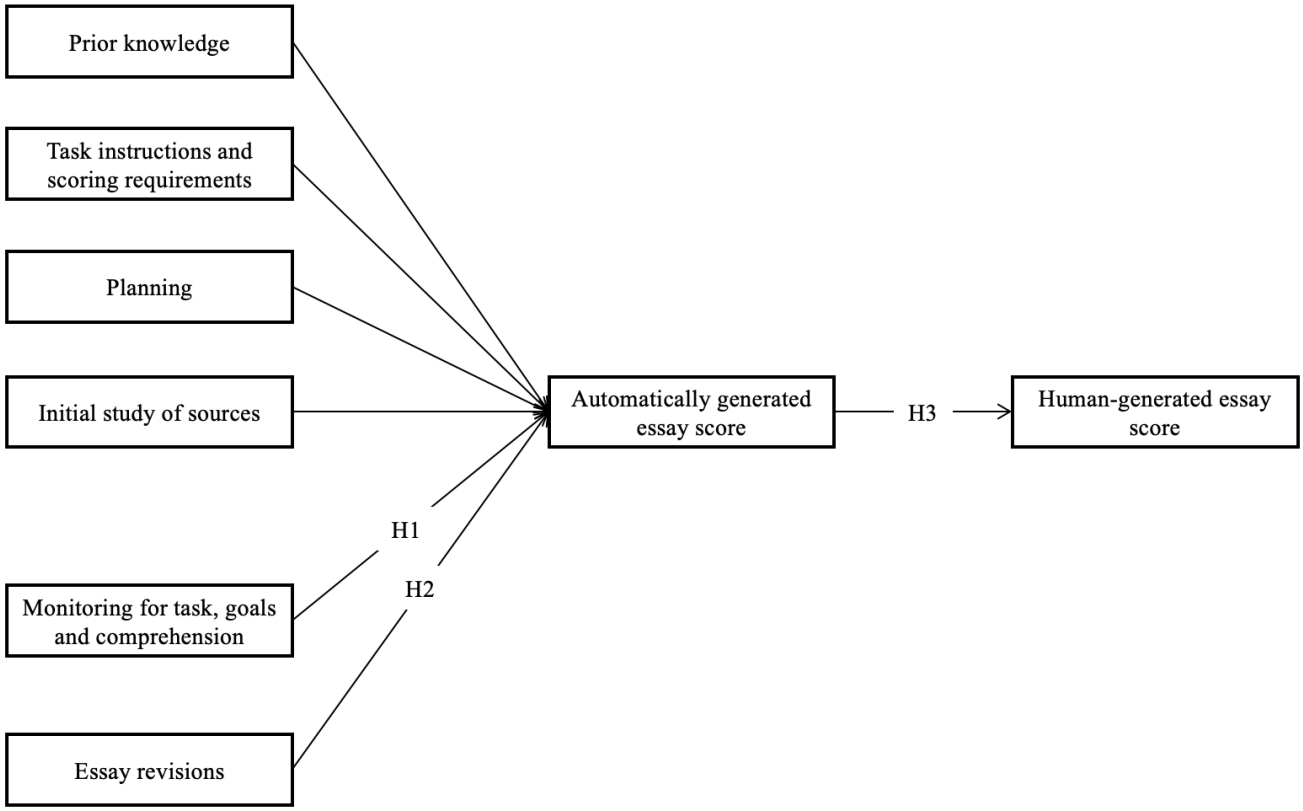

Figure 1: The Hypothesized Path Model. Metacognitive monitoring and control (essay revisions) are hypothesised to be positively related to essay scores (H1, H2, H4, H5); automatically generated essay score is hypothesised to be positively related to human-generated essay score (H3); the variables examined in RQ1 (prior knowledge, time to initially read the task instructions/scoring rubric, time to plan, and time to initially study the source documents) are expected to be positively related to automatically generated essay score

multi-source writing session and a post-test. During the writing session, the participants were given the opportunity to study the pre-selected source materials written in Dutch and hosted in the learning environment, i.e., the participants operated on the closed-documents-set [40]. The materials spanned the three topics 1) artificial intelligence (AI), e.g., describing general concept and key approaches to AI (8 texts with the average length of 260 words), 2) differentiation in education, e.g., describing how instruction is adapted to different students' needs (3 texts with the average length of 349 words), and 3) scaffolding, e.g., providing external support to students (5 texts with the average length of 368 words). The participants were required to integrate the three topics into the 300-400 words vision essay that describes learning in schools in 2035.

3.2 Learning environment

A technology-enhanced learning environment was developed for this study. The environment was web-based and included a catalogue of learning materials and navigation area on the left; a reading and essay writing zone in the middle; and instrumentation tools on the right (Figure 2). In the navigation area, the learners could navigate source texts for the essay, but they could also navigate back to task instructions and the essay scoring rubric at any point during the session. In the essay writing zone, the participants were composing their essays. They could begin writing at any point after the 45-minute session has started. The instrumentation tools included the tools for annotating and planning. The annotation tool afforded the learners the opportunity to highlight parts of a text in the reading zone and create notes or search for/edit highlights and notes they created earlier. The learners could also use the planner tool to plan their writing session (e.g., "read the source texts first, then start writing...") and the timer tool to monitor time left for the task. The environment recorded learner trace data in a form of timestamped navigational logs and keystrokes. In addition, we recorded the learners' eye fixations on the screen (e.g., navigation zone or reading zone) using the eye-tracking hardware. The collected trace data were stored in the local PHP-server.

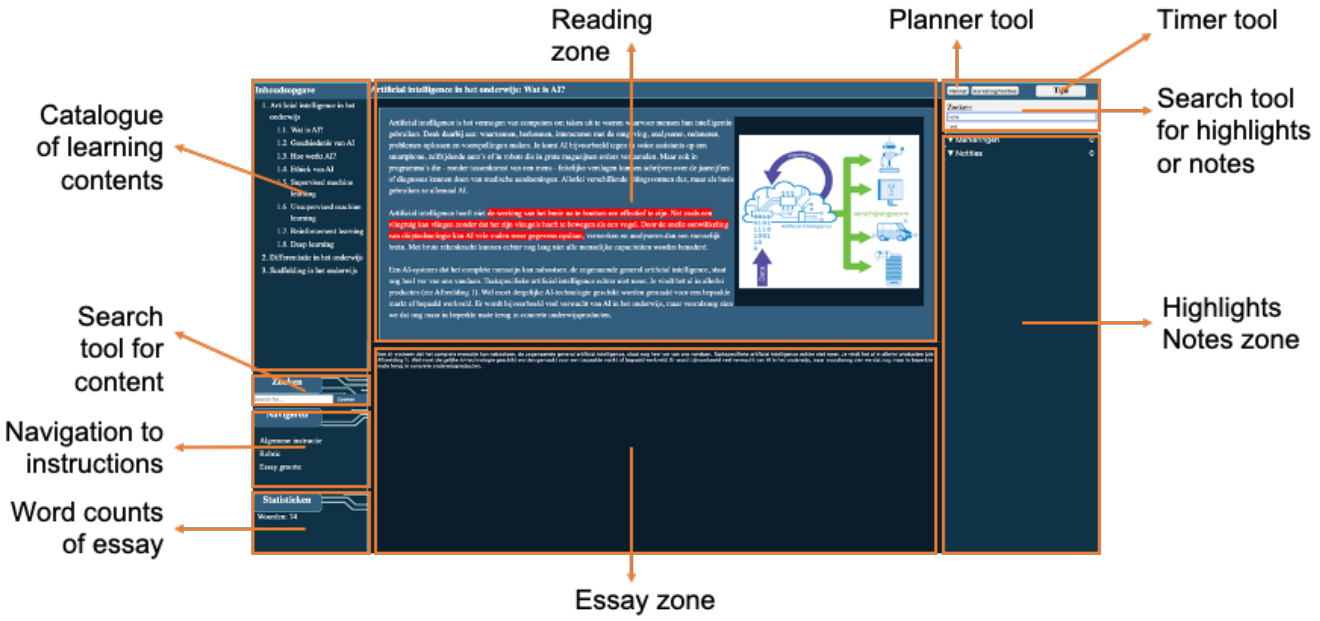

Figure 2: Learning environment with functional zones

3.3 Measurement

3.3.1 Prior knowledge score. We obtained the prior knowledge score from the pretest administered before the learning session. The pretest included 30 multiple-choice items and tested for recall, understanding, application and evaluation [36] of concepts in artificial intelligence, differentiation and scaffolding. The prior knowledge score was calculated as a proportion of correctly answered items.

3.3.2 Time on the instructions and rubric page. At the outset of the learning session, the learners were provided with the task instructions and the scoring rubric details. We measured the time each student spent on these pages before they proceeded further with the task.

3.3.3 Planning. We measured planning as the total time learners spent using the planner tool to create or edit their plans for the writing task. The plans, in turn, could help learners to more efficiently engage in the task.

3.3.4 Studying, monitoring and revising behaviours. During the learning session, the learners studied source texts hosted in the TEL and developed their essay responses in the TEL essay zone. We collected a substantial amount of timestamped trace data generated in the environment, including navigation logs, keyboard strokes and eye-tracking data[1] (eye fixations). These data were used to model reading, monitoring and revising (i.e., control) behaviours. We modelled learner behaviours as time variables, i.e., we computed duration of each behaviour (e.g., creating and editing note, reviewing highlights, revising the text in the essay draft), aggregated over the whole session.

3.3.5 Essay score. Two human raters scored the 44 essays created in the writing sessions. The raters were introduced the scoring scheme, trained together by scoring 5 randomly selected essays and then scored the remaining 39 essays separately (20-19 split), achieving a high inter-rater reliability of Fleiss-Cohen *kappa*=.89. Each essay was scored out of 21 points based on the seven-part scoring scheme: coverage of each of the three topics (parts 1-3), integration of topics in the essay (4), word count (5), vision (6) and originality (7). Maximum of 3 points was assigned per part. In the present study, we adopted the essay score based on parts 1-5 (out of 15 points). We did not include in our analyses the scores for vision and originality, as those constructs appeared to be hard to reliably measure automatically. For instance, what writers envision to happen in the future is often influenced by their past and current social experiences [15] and, since data about participants' social experiences were not available to our research team, we elected to not include these scoring components in our analysis. To ensure the version of the essay score we used does not excessively deviate from the score that included all the scoring components, we computed the correlation between the two scores and found it positive, strong and statistically significant $r(42)=0.91$ ($p<0.001$) [8], that is, the scores obtained from the five-part scoring scheme reliably reflected the scores obtained from the seven-part scoring scheme. Therefore, we proceeded with using the five-part-based essay scores in our analysis.

[1]The eye tracking data was collected using the Tobii TX300, a screen-based eye tracking device with a sampling rate of 300 Hz. The device was calibrated relative to 9 points on the screen. The Tobii I-VT Filter was used to categorize learners' eye fixations. The timestamps for eye-tracking and log data were stored locally, allowing synchronisation of the data sources.

Table 1: Descriptive Statistics and Correlations for Major Study Variables

Variable	M	SD	Skew	Kurtosis	1	2	3	4	5	6	7	8
1. PK	54.24	13.16	-.92	.56	1.00	-.15	-.12	-.22	.05	.11	.39*	.29
2. Initial time on I/R	46.59	40.47	.92	-.06	-.15	1.00	-.14	-.08	-.01	.08	-.29	-.09
3. Planning	16.76	46.54	3.44	12.39	-.12	-.14	1.00	.03	.10	.12	.12	.18
4. SA	504.12	288.43	.69	.71	-.22	-.08	.03	1.00	-.33*	-.09	-.28	-.42**
5. Monitoring	711.39	363.05	.33	-.38	.05	-.01	.10	-.33*	1.00	-.11	.37*	.40**
6. Revisions	10.53	10.16	1.33	1.29	.11	.08	.12	-.09	-.11	1.00	.22	.24
7. Automatic score	4.16	.33	-.86	-.05	.39*	-.29	.12	-.28	.37*	.22	1.00	.72***
8. Human score	11.30	3.14	-1.25	1.49	.29	-.09	.18	-.42**	.40**	.24	.72***	1.00

Note. **PK**: prior knowledge (percentage of correctly answered items); **Initial time on I/R**: initial time on instructions and scoring rubric page (sec); **Planning**: total time in planning tool (sec); **SA**: total duration of studying activity (sec); **Monitoring**: total duration of monitoring activity (sec); **Revisions**: total duration of essay revisions (sec); **Automatic score**: essay score obtained automatically (points, out of 5.2); **Human score**: essay score obtained by human assessors (points, out of 15). **Significance codes**: *p<.05, **p<.01, ***p<.001

3.4 Data Preparation

3.4.1 Initial study of sources. We operationalised the activity of initial study of sources as a total time a learner spent (1) initially reading source documents and (2) creating/editing notes and highlights based on source documents. The initial reading was operationalised as the first-time reading of a source text page, i.e., a time span between the point when a learner opened the page for the first time and the point they left the page (e.g., moved to another page), as detected in navigation logs. The initial reading time on each page was computed as a total duration of the eye gaze fixated at the reading zone of that page, as indicated by eye-tracking data. In other words, the time learner's eyes were fixated somewhere else (e.g., at the page navigation zone of the TEL or at the lab wall) was not included in the computation. The gaze duration values were then summed across all the pages to obtain the total initial reading time. Further, the time a learner spent creating/editing highlights and notes was computed as a total duration of these activities across the learning session, as indicated by the timestamped navigational logs. Finally, we obtained the measure of the initial study of sources for a learner by adding together the total initial reading time and the total time for creating/editing highlights and notes.

3.4.2 Monitoring. We operationalised a learner monitoring activity relative to two types of monitoring theorised to occur in multi-source writing: monitoring for task requirements and goals [19] and monitoring for comprehension [17]. Monitoring for task and goals was computed as a total time a learner spent (1) re-reading the task instructions, (2) re-reading the scoring rubric page, and (3) looking at the timer while working on the task. Monitoring for comprehension, on the other hand, was computed as a total time a learner spent (1) re-reading source pages, (2) reviewing previously created highlights and notes and (3) searching for previously created highlights and notes. Finally, we obtained the monitoring activity for a learner by adding together the total time they spent monitoring for task/goals and the total time they spent monitoring for comprehension.

3.4.3 Essay Revision. We computed an essay revision as a total time a learner spent revising the essay draft. The revision activity is recorded whenever a learner modified the existing text in the essay writing zone, as indicated by coordinates of learners' eye positions, and by coordinates and a timestamp of keystrokes collected during writing.

3.4.4 Automatically generated essay score. We automatically computed the essay score as a sum of the following three indices: (1) overall topics coverage, (2) integration of topics in the essay and (3) word count. To obtain the overall topics coverage for the essay, we computed a coverage for each of the three topics separately and added those values together. The coverage for one individual topic was computed as a semantic similarity between the essay and a set of source texts associated with that topic, e.g., the 8 texts about AI. To compute the semantic similarity, we utilised the word2vec algorithm [41] with the vector space that contained 500 thousand unique words, each represented as a 300-dimensional vector. The algorithm was trained on a large corpus of news and media texts written in Dutch and implemented in Python using the state-of-the-art natural language processing system spaCy [27]. Further, using this same word vector space and the spaCy model, we computed the integration of topics in each essay. To this end, we calculated the semantic similarity between each pair of the sentences in the essay and then calculated the mean semantic similarity across all the pairs. Last, the word count score was computed as total number of words in the essay normalised by 300, because 300 was a minimum number of words required in this task. As the maximum number of words required for the essay was 400, the word count score for all the essays that exceeded 400 words was set to 1.33 (or 400 divided by 300) to avoid inflated estimates of word count score due to the essays that were longer than required. Finally, we added together the overall topics coverage (out of 3 points), integration of topics in the essay (out of 1 point) and word count scores (out of 1.2 points) to obtain automatically generated essay score (out of 5.2 points).

3.5 Data analysis

3.5.1 Initial data screening. During the initial data screening we identified two essays that contained nearly 19% and 29% of the content that was directly copied from source texts. We elected to remove those submissions from further analysis and our final analytical sample included 42 learners. Further, we computed the

Table 2: Hypothesized Direct and Indirect Effects

Hypothesis	Estimate	p-value
H1 (Monitoring predicting automatically generated essay score)	.49	<.001
H2 (Control predicting automatically generated essay)	.31	.03
H3 (Automatically generated score predicting human-generated score)	.41	<.001
H4 (Monitoring predicting human-generated score through automatically generated score)	.20	.001
H5 (Control predicting human-generated score through automatically generated score)	.13	.04

absolute Pearson correlations among predictors in our study, i.e., prior knowledge, initial time on the instructions/scoring rubric page, planning time, studying activity time, monitoring and control, to test whether the effects of predictors on the essay score can be considered independent, i.e., to test for potential multicollinearity issues in the model. As the absolute values of all the correlation coefficients we obtained were well beyond 0.70, the commonly accepted threshold for multicollinearity [62], we found no multicollinearity issues in the data set and proceeded with the analysis by simultaneously entering all the predictors in the model. The descriptive statistics and correlations for major study variables are presented in Table 2.

3.5.2 Testing the hypotheses. To test our hypotheses, we computed the hypothesised path model using the lavaan [47] package implemented in R. Since Shapiro-Wilk's normality test indicated non-normal distributions in all but the studying activity variable, we applied robust maximum likelihood estimation with Satorra-Bentler corrections [50] to mitigate the normality issue when fitting the path model. Per suggestions provided in [34], we computed the following test statistics and fit indices to estimate a goodness of data-model fit (1) chi-square with its degrees of freedom and p-value, (2) Comparative Fit Index (CFI), (3) Root Mean Square Error of Approximation (RMSEA), and (4) Standardized Root Mean Square Residual (SRMR).

4 RESULTS

The indices in our hypothesised model indicated a good data-model fit χ^2 = 1.55 (df = 6), p = .96; CFI = 1.00, RMSEA = .00, SRMR = .02. The final path model with standardised coefficients is presented in Figure 3.

The relationship between metacognitive monitoring and automatically measured essay score was statistically significant and positive (β=0.49, p<0.001), supporting hypothesis H1. Moreover, after controlling for other predictors of the automatic essay score (i.e., prior knowledge, initial time on task instructions/scoring rubric, planning, initial studying of sources, and control), this finding indicates that learners who spent more time monitoring for task requirements, scoring criteria, and comprehension of source materials were more likely to achieve a higher score automatically obtained by the scoring algorithm. The time a learner spent revising their essay (i.e., metacognitive control) was statistically significantly and positively related to automatically generated essay score (β=0.31, p=0.03), supporting hypothesis H2. In other words, after controlling for other predictors of the automatic essay score (i.e., prior knowledge, initial time on task instructions/scoring rubric, initial studying of sources, and monitoring), this finding indicates that learners who invested more time in revising their drafts throughout the learning session were more likely to achieve a higher score automatically obtained by the scoring algorithm. The relationship between automatically and human-generated essay score was statistically significant and positive (β=0.41, p<0.001), supporting hypothesis H3 and indicating that the automatic scoring approach implemented in this study can predict the human-generated score with considerable reliability. The indirect path from metacognitive monitoring to human-generated essay score through the automatically generated essay score was positive and statistically significant, (β=0.20, p=0.001), supporting hypothesis H4. The indirect path from metacognitive control to human-generated essay score through the automatically generated essay score was positive and statistically significant, (β=0.13, p=0.04), supporting hypothesis H5.

The results obtained for RQ1 indicate that prior knowledge was statistically significantly related to automatically generated essay score (β=0.50, p=0.01), and, indirectly, to human-generated essay score (β=0.21, p=0.01), through the automatically generated essay score. Initial time on task instructions and scoring requirements was statistically significantly related to automatically generated essay score (β=-0.36, p=0.03), and, indirectly, to human-generated essay score (β=-0.14, p=0.02), through the automatically generated essay score. The relationships between time on planning and initial studying of sources, and automatically generated essay score, were not statistically significant (β=0.07, p=0.59 and β=-0.14, p=0.59, respectively).

5 DISCUSSION

Metacognitive monitoring and control have been theorised as critical processes that multi-source writers should enact to ensure productive engagement and success in this demanding task [19, 46]. However, a few researchers have attempted to empirically document how these metacognitive processes unfold in both reading and writing spaces of the task and how they affect the quality of the written product. To bridge this gap, we gathered trace data reflecting learners' reading and writing behaviours as they used the study TEL in the multi-source writing session. Informed by prior research [17, 19, 24, 53], we mapped those data to corresponding monitoring and control processes. We examined the effects of these constructs on the quality of vision essays learners produced during the experimental session. The essays were scored automatically, by applying the state-of-the-art NLP approach word2vec [41], and also by human assessors. Our results largely support (1) the theoretical propositions that engagement in metacognitive monitoring and control benefits the quality of written product [19, 60] and (2) the

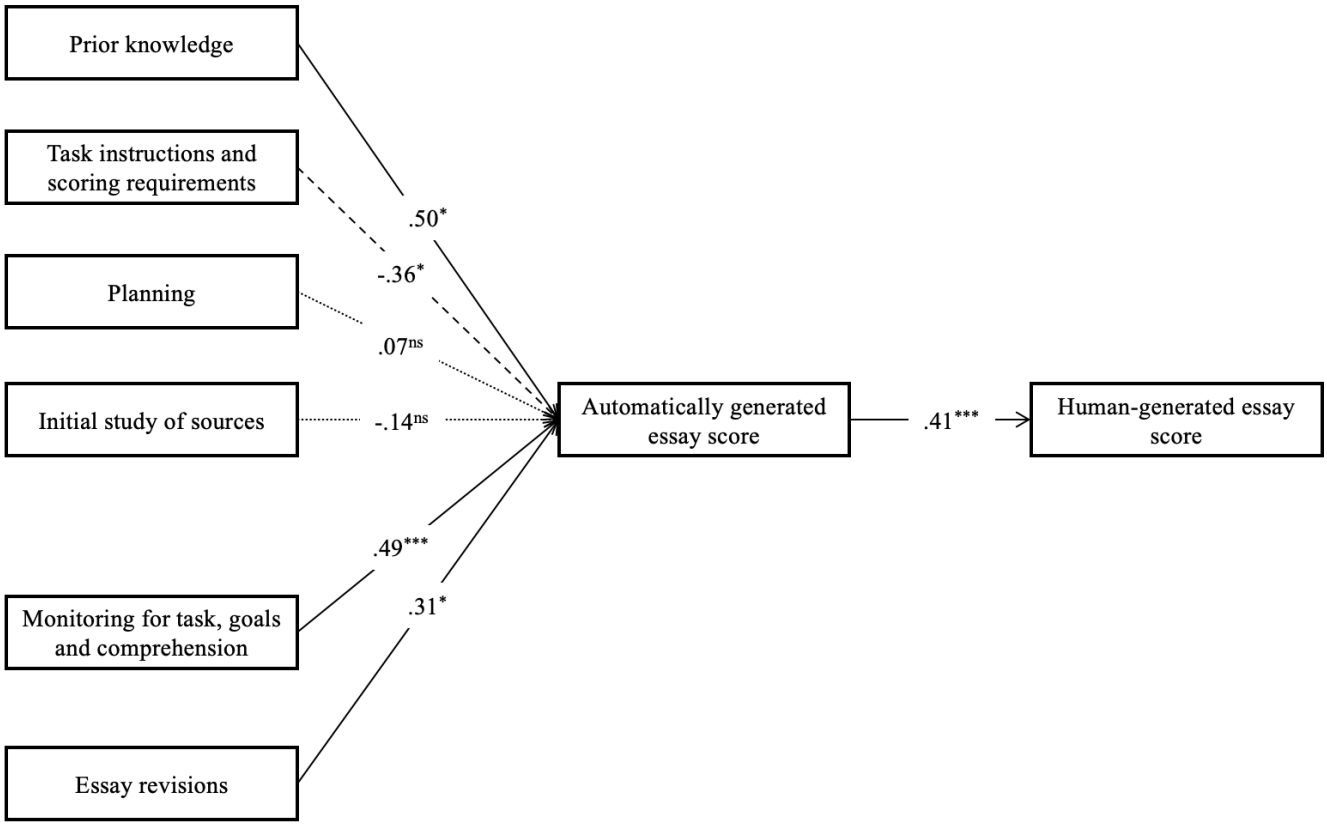

Figure 3: The Final Path Model. As expected, the statistically significant and positive relationships were detected between metacognitive monitoring and automatically generated essay score (H1), between essay revisions (i.e., metacognitive control) and automatically generated essay score (H2), and between automatically and human-generated essay score (H3). The statistically significant and positive indirect paths were detected from metacognitive monitoring (H4) and from essay revisions (H5) to human-generated essay score through the automatically generated essay score.

expectations based on prior research that the performance of the essay scoring algorithm is comparable to that of human assessors.

5.1 Theoretical implications

We obtained empirical evidence of writing from multiple sources as a self-regulatory and recursive learning activity [65]. In particular, our results conform to prior research showing that self-regulated multi-source writers tend to improve the quality of their initial task understanding, reading and writing efforts through the prolonged engagement in metacognitive monitoring and text revisions [60, 65], which ultimately leads to a quality written product. More specifically, the positive and statistically significant effects of the duration of monitoring activities on the essay scores documented in our study (H1 and H4) indicate that high-performing writers productively self-regulated their writing by maintaining their own awareness of requirements and purposes of a written product [26], and by judging whether they handled the source information properly [17], as opposed to simply reading each source document once and translating chunks of information into the essay. Moreover, the positive and statistically significant effects of the duration of revision activities (i.e., metacognitive control) on the essay scores (H2 and H5) align with prior research suggesting that productive multi-source writers tend to judge what they have already written to identify the discrepancies between the product in its current state and the product they intend to create [60], and then modify the product to improve its quality [25].

Further, we detected the additional two statistically significant predictors of the essay score in our model (RQ1). First, the positive relationship between the writer's prior knowledge of writing topics and the essay score substantiate previous theoretical positions and empirical findings (cf.[13]) that the writer's prior knowledge is an important internal precondition influencing performance in writing tasks. Moreover, the writer's initial representation of content (i.e., content schema, [44]) can facilitate assimilation and comprehension of new information throughout the task. For instance, a writer possessing a rich content schema about scaffolding and differentiation topics at the outset of the task may be able to more efficiently identify new connections between these topics in the source texts which, in turn, benefits the quality of composing [3]. Lack of prior knowledge of writing topics, on the other hand, may

hinder the writer's ability to compose from sources [44]. However, unlike process-based variables such as metacognitive monitoring and control, prior knowledge is rather a static variable that cannot be intervened upon during the task. This further amplifies the need for supporting multi-source writer's processes as they unfold during the task, a critical initiative to compensate for low prior knowledge.

Second, the model results indicated the negative relationship between the time that writers initially spent reading task/scoring requirements and the essay score. This finding appears to be a bit counter intuitive, given previously established positions that task understanding generally promotes academic performance in self-regulatory learning contexts (e.g., [20, 42]). We speculate this might be due to the limited time frame provided to learners to complete the task in our experiment. For instance, the learners who spent more time reading the task instructions and scoring requirements at the beginning of the experiment may have ended up having less time available later in the session to comprehend sufficient source information and develop a quality essay. Therefore, the limited time for the task could have prevented the learners from making full use of writing processes, as previously noted in [29]. This speculation should be addressed in the future field studies where learners will be provided with the longer time frames for the task.

Last, we found no statistically significant relationships between automatically generated essay score and (1) time learners spent planning their work and (2) time learners spent initially studying source documents (RQ1). These associations should be further investigated in the future field studies.

5.2 Practical implications

As our findings indicated the external validity of trace measures of metacognition by being able to predict scores of student written products, we demonstrated that the use of trace data can be a promising approach to measuring metacognition in multi-source writing. Moreover, the measures developed in this study can be used to benefit writer-facing analytics-based tools that support writing progress and also instructor-facing tools that empower instructors to support their students.

For instance, given the strong correlation and positive relationship that we found between automatically and human-generated essay score (H3), we expect that the automatic scoring algorithm, proposed in the current study, can be used to reliably estimate essay scores on-the-fly. The scores obtained in this way can signal which learners are likely to produce a low quality essay. Those learners, in turn, will need to alter their learning behaviours to succeed in a multi-source writing task, i.e., the characteristics of a written product will determine whether a learner is flagged for support or not. However, previous research has cautioned that the holistic score does not provide a sufficient information about a quality of the learner's learning behaviours in a writing task [22, 32], making it hard to determine the right support for a learner in need. To address this issue, future writing analytics tools could track/analyse writing processes by harnessing learner trace data from multiple channels, as demonstrated in our study. In this way, the details about learner engagement prior to submitting a draft can be utilised to inform appropriate support, e.g., a computer-based scaffold reminding a learner to review task instructions, as learner trace data revealed no such behaviours earlier in the session.

6 LIMITATIONS AND FUTURE DIRECTIONS

We observed a few major limitations to our study that may need attention in future research. First, the sample of learners in the study did not provide us with a sufficient statistical power to create and analyse process variables at finer granularity, e.g., variables representing monitoring for task, monitoring for comprehension, re-reading of highlights and re-reading of notes, separately. A future study should be conducted with a larger sample of learners to explore viability of using those variables to further improve measurement of processes in multi-source writing. Moreover, the inferences about metacognitive processes based on duration should be validated and expanded upon in future studies by including the analysis of temporal and sequential characteristics of those processes. Next, even though the automatically generated essay score explained a considerable amount of variance in the human-generated score (r=.71, R^2=.50), relative to common field standards (e.g., [1]), there still remains a portion of unexplained variance. The future versions of the scoring algorithm thus should be improved by enriching the training corpus with additional texts that approximate writing style of post-secondary students. We also note that all the participants in our study were native speakers. Since prior research has documented that learners with different language proficiency levels can achieve different outcomes in the same writing tasks [45], this study should be replicated and findings evaluated with learners of different language proficiency levels. Last, as the eye-tracking devices are not available to most of the students who will use our learning platform outside the lab, it may be challenging to observe the eye-tracking data channel in real-world teaching environments. The viability of using web-based eye trackers to this end should be explored in future research.

7 CONCLUSION

Many educators around the world regard writing from multiple sources as a useful instructional task that provides opportunities for learners to engage in reading and writing activities within a particular disciplinary genre, a potential benefit for their composing skills. Even though educationally potent, multi-source writing tasks are still challenging for many learners who seek to harvest knowledge from source documents and integrate this knowledge into a coherent composition to address task and genre requirements. To succeed, multi-source writers need to engage in metacognitive monitoring and control during the task. Many learners, however, underuse these processes. To provide appropriate support to those learners, it is important to understand how metacognitive processes unfold in a writing task and how these processes predict the essay quality. Expanding upon the prior research, we collected trace data that multi-source writers generated in the experimental writing session, and analysed how measures extracted from those data relate to essay performance. Path model results supported our hypotheses, i.e., engagement in metacognitive monitoring and control benefits the quality of a written product. Moreover, our results indicate that the essay scoring algorithm applied in this study can be used to estimate essay scores with considerable reliability and thus identify

writers at risk of producing low quality drafts, i.e., writers who may need to increase their metacognitive processing to succeed in a multi-source writing task.

ACKNOWLEDGMENTS

This study is a result of the FLoRA research project funded by DFG (Germany), NWO (The Netherlands), and ESRC (United Kingdom) as part of the Open Research Area (ORA) BA20144/10-1, NWO464.18.104, ES/S015701/1.

REFERENCES

[1] Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics* 34 (2008), 555–596.
[2] Khaled Barkaoui. 2016. What and when second-language learners revise when responding to timed writing tasks on the computer: The roles of task type, second language proficiency, and keyboarding skills. *The Modern Language Journal* 100, 1 (2016), 320–340.
[3] Carl Bereiter and Marlene Scardamalia. 1987. The psychology of written composition. (1987).
[4] Jason LG Braasch, Jean-François Rouet, Nicolas Vibert, and M Anne Britt. 2012. Readers' use of source information in text comprehension. *Memory & cognition* 40, 3 (2012), 450–465.
[5] Brent Bridgeman, Catherine Trapani, and Yigal Attali. 2012. Comparison of human and machine scoring of essays: Differences by gender, ethnicity, and country. *Applied Measurement in Education* 25, 1 (2012), 27–40.
[6] John T Bruer. 1993. *Schools for thought*. Vol. 120. Cambridge, MA: MIT Press.
[7] Byeong-Young Cho, Peter Afflerbach, and Hyeju Han. 2018. Strategic processing in accessing, comprehending, and using multiple sources online. In *Handbook of multiple source use*. Routledge, 133–150.
[8] Jacob Cohen. 1988. Set correlation and contingency tables. *Applied psychological measurement* 12, 4 (1988), 425–434.
[9] Rianne Conijn, Emily Dux Speltz, Menno Van Zaanen, Luuk Van Waes, and Evgeny Chukharev-Hudilainen. 2020. A process-oriented dataset of revisions during writing. In *Proceedings of the 12th Language Resources and Evaluation Conference*. 363–368.
[10] Elena Cotos, Sarah Huffman, and Stephanie Link. 2020. Understanding Graduate Writers' Interaction with and Impact of the Research Writing Tutor during Revision. *Journal of Writing Research* 12, 1 (2020).
[11] Scott A Crossley, Minkyung Kim, Laura Allen, and Danielle McNamara. 2019. Automated summarization evaluation (ASE) using natural language processing tools. In *International Conference on Artificial Intelligence in Education*. Springer, 84–95.
[12] Scott A Crossley, Kristopher Kyle, and Mihai Dascalu. 2019. The Tool for the Automatic Analysis of Cohesion 2.0: Integrating semantic similarity and text overlap. *Behavior research methods* 51, 1 (2019), 14–27.
[13] Alister Cumming, Conttia Lai, and Hyeyoon Cho. 2016. Students' writing from sources for academic purposes: A synthesis of recent research. *Journal of English for Academic purposes* 23 (2016), 47–58.
[14] Saad M Darwish and Sherine Kh Mohamed. 2019. Automated essay evaluation based on fusion of fuzzy ontology and latent semantic analysis. In *International Conference on Advanced Machine Learning Technologies and Applications*. Springer, 566–575.
[15] Patricia Freitag Ericsson. 2006. The meaning of meaning. *Machine scoring of human essays: Truth or consequences* (2006), 28–37.
[16] Linda Flower. 1989. Cognition, context, and theory building. *College composition and communication* 40, 3 (1989), 282–311.
[17] Linda Flower, Victoria Stein, John Ackerman, Margaret J Kantz, Kathleen McCormick, Wayne C Peck, et al. 1990. *Reading-to-write: Exploring a cognitive and social process*. Oxford University Press on Demand.
[18] Jeffrey Alan Greene, Dana Z Copeland, Victor M Deekens, and Rebekah Freed. 2018. Self-regulated learning processes and multiple source use in and out of school. In *Handbook of multiple source use*. Routledge, 320–338.
[19] Douglas J Hacker, Matt C Keener, and John C Kircher. 2009. Writing is applied metacognition. In *Handbook of metacognition in education*. Routledge, 166–184.
[20] Allyson F Hadwin. 2006. Do your students really understand your assignment. *LTC Currents Newsletter, II (3)* (2006), 1–9.
[21] Peter Hastings, Simon Hughes, Joseph P Magliano, Susan R Goldman, and Kimberly Lawless. 2012. Assessing the use of multiple sources in student essays. *Behavior Research Methods* 44, 3 (2012), 622–633.
[22] Richard H Haswell. 2006. Automatons and automated scoring: Drudges, black boxes, and dei ex machina. *Machine scoring of student essays: Truth and consequences* (2006), 57–78.
[23] John R Hayes. 1992. Planning in writing: The cognition of a constructive process. *A rhetoric of doing: Essays on written discourse in honor of James L. Kinneavy* (1992), 181.
[24] John R Hayes. 2000. Understanding Cognition and Affect in Writing. *Perspectives on writing: Research, theory, and practice* (2000), 6.
[25] John R Hayes. 2004. What triggers revision? In *Revision cognitive and instructional processes*. Springer, 9–20.
[26] John R Hayes. 2012. Evidence from language bursts, revision, and transcription for translation and its relation to other writing processes. (2012).
[27] Matthew Honnibal and Mark Johnson. 2015. An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 conference on empirical methods in natural language processing*. 1373–1378.
[28] Simon Hughes, Peter Hastings, Mary Anne Britt, Patricia Wallace, and Dylan Blaum. 2015. Machine learning for holistic evaluation of scientific essays. In *International Conference on Artificial Intelligence in Education*. Springer, 165–175.
[29] Heidi Hyytinen, Erika Löfström, and Sari Lindblom-Ylänne. 2017. Challenges in argumentation and paraphrasing among beginning students in educational sciences. *Scandinavian Journal of Educational Research* 61, 4 (2017), 411–429.
[30] Roger Johansson, Åsa Wengelin, Victoria Johansson, and Kenneth Holmqvist. 2010. Looking at the keyboard or the monitor: relationship with text production processes. *Reading and writing* 23, 7 (2010), 835–851.
[31] Yves Karlen. 2017. The development of a new instrument to assess metacognitive strategy knowledge about academic writing and its relation to self-regulated writing and writing performance. *Journal of Writing Research* 9, 1 (2017).
[32] Ronald T Kellogg and Alison P Whiteford. 2009. Training advanced writing skills: The case for deliberate practice. *Educational Psychologist* 44, 4 (2009), 250–266.
[33] Perry D Klein and Pietro Boscolo. 2016. Trends in research on writing as a learning activity. *Journal of writing research* 7, 3 (2016), 311–350.
[34] Rex B Kline. 2016. Principles and practice of structural equation modeling. (2016).
[35] Simon Knight, Antonette Shibani, Sophie Abel, Andrew Gibson, Philippa Ryan, Nicole Sutton, Raechel Wight, Cherie Lucas, Agnes Sandor, Kirsty Kitto, et al. 2020. AcaWriter: A learning analytics tool for formative feedback on academic writing. *Journal of Writing Research* 12, 1 (2020), 141–186.
[36] David R Krathwohl. 2002. A revision of Bloom's taxonomy: An overview. *Theory into practice* 41, 4 (2002), 212–218.
[37] Thomas K Landauer, Peter W Foltz, and Darrell Laham. 1998. An introduction to latent semantic analysis. *Discourse processes* 25, 2-3 (1998), 259–284.
[38] Mariëlle Leijten and Luuk Van Waes. 2013. Keystroke logging in writing research: Using Inputlog to analyze and visualize writing processes. *Written Communication* 30, 3 (2013), 358–392.
[39] María Luna, Ruth Villalón, Mar Mateos, and Elena Martín. 2020. Improving university argumentative writing through online training. *Journal of Writing Research* (2020).
[40] Joseph P Magliano, Peter Hastings, Kristopher Kopp, Dylan Blaum, and Simon Hughes. 2018. Computer-Based Assessment Of Essays Based On Multiple Documents: Evaluating the Use of Sources. In *Handbook of multiple source use*. Routledge, 502–526.
[41] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.
[42] Mika Oshige. 2009. *Exploring task understanding in self-regulated learning: Task understanding as a predictor of academic success in undergraduate students*. Ph.D. Dissertation.
[43] Charles A Perfetti, Jean-François Rouet, and M Anne Britt. 1999. Toward a theory of documents representation. *The construction of mental representations during reading* 88108 (1999).
[44] Dolores Perin, Alla Keselman, and Melissa Monopoli. 2003. The academic writing of community college remedial students: Text and learner variables. *Higher Education* 45, 1 (2003), 19–42.
[45] Limin Qin and Lawrence Jun Zhang. 2019. English as a foreign language writers' metacognitive strategy knowledge of writing and their writing performance in multimedia environments. *Journal of Writing Research* 11, 2 (2019).
[46] Mladen Raković and Philip H Winne. in press. SR-WMS: A Typology of Self-Regulation in Writing from Multiple Sources. In *Social and Emotional Learning: An Inclusive Learning Analytics Perspective*. in press.
[47] Y Rossel. 2012. Lavaan: na R package for structural equation modeling. Jounal of Statistical Sofware, 48 (2), 1-36.
[48] Jean-Francois Rouet and M Anne Britt. 2011. Relevance processes in multiple document comprehension. *Text relevance and learning from text* (2011), 19–52.
[49] Jean-François Rouet, M Anne Britt, and Amanda M Durik. 2017. RESOLV: Readers' representation of reading contexts and tasks. *Educational Psychologist* 52, 3 (2017), 200–215.
[50] Albert Satorra. 1992. Asymptotic robust inferences in the analysis of mean and covariance structures. *Sociological methodology* (1992), 249–278.
[51] Mark D Shermis. 2014. State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. *Assessing Writing* 20 (2014), 53–76.

[52] Mark D Shermis, Jill Burstein, Norbert Elliot, Shayne Miel, and Peter W Foltz. 2016. Automated writing evaluation: An expanding body of knowledge. (2016).
[53] Lisa K Son and Bennett L Schwartz. 2002. The relation between metacognitive monitoring and control. (2002).
[54] Qiyu Sun, Lawrence Jun Zhang, and Susan Carter. 2021. Investigating Students' Metacognitive Experiences: Insights From the English as a Foreign Language Learners' Writing Metacognitive Experiences Questionnaire (EFLLWMEQ). *Frontiers in Psychology* (2021), 3779.
[55] Yao-Ting Sung, Chia-Ning Liao, Tao-Hsing Chang, Chia-Lin Chen, and Kuo-En Chang. 2016. The effect of online summary assessment and feedback system on the summary writing on 6th graders: The LSA-based technique. *Computers & Education* 95 (2016), 1–18.
[56] Nina Vandermeulen, Mariëlle Leijten, and Luuk Van Waes. 2020. Reporting Writing Process Feedback in the Classroom Using Keystroke Logging Data to Reflect on Writing Processes. *Journal of Writing Research* 12, 1 (2020), 109–139.
[57] Marcel VJ Veenman. 2007. The assessment and instruction of self-regulation in computer-based environments: a discussion. *Metacognition and Learning* 2, 2-3 (2007), 177–183.
[58] Jennifer Wiley, Peter Hastings, Dylan Blaum, Allison J Jaeger, Simon Hughes, Patricia Wallace, Thomas D Griffin, and M Anne Britt. 2017. Different approaches to assessing the quality of explanations following a multiple-document inquiry activity in science. *International Journal of Artificial Intelligence in Education* 27, 4 (2017), 758–790.
[59] Philip H Winne and Allyson F Hadwin. 1998. Studying as self-regulated engagement in learning. in metacognition in educational theory and practice. *Metacognition in Educational Theory and Practice* (1998), 277–304.
[60] Anke Wischgoll. 2016. Combined training of one cognitive and one metacognitive strategy improves academic writing skills. *Frontiers in psychology* 7 (2016), 187.
[61] Hui-Chin Yeh. 2015. Facilitating metacognitive processes of academic genre-based writing using an online writing system. *Computer Assisted Language Learning* 28, 6 (2015), 479–498.
[62] Wonsuk Yoo, Robert Mayberry, Sejong Bae, Karan Singh, Qinghua Peter He, and James W Lillard Jr. 2014. A study of effects of multicollinearity in the multivariable analysis. *International journal of applied science and technology* 4, 5 (2014), 9.
[63] Haoran Zhang and Diane Litman. 2020. Automated Topical Component Extraction Using Neural Network Attention Scores from Source-based Essay Scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 8569–8584.
[64] Haoran Zhang, Ahmed Magooda, Diane Litman, Richard Correnti, Elaine Wang, LC Matsmura, Emily Howe, and Rafael Quintana. 2019. eRevise: Using natural language processing to provide formative feedback on text evidence usage in student writing. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 9619–9625.
[65] BJ Zimmerman and A Kitsantas. 2007. A writer's discipline: The development of self-regulatory skill. P. Boscolo & S. Hidi (Eds.), Writing and motivation (pp. 51-69).