

# Leveraging Process-Action Epistemic Network Analysis to Illuminate Student Self-Regulated Learning with a Socratic Chatbot

Joel Weijia Lai<sup>1</sup>, Wei Qiu<sup>2</sup>, Maung Thway<sup>3</sup>, Lei Zhang<sup>4</sup>, Nurabidah Binti Jamil<sup>5</sup>, Chit Lin Su<sup>6</sup>, Samuel S.H. Ng<sup>7</sup>, Fun Siong Lim<sup>8\*</sup>

## Abstract

The growing use of generative AI (GenAI) has sparked discussions regarding integrating these tools into educational settings to enrich the learning experience of teachers and students. Self-regulated learning (SRL) research is pivotal in addressing this inquiry. One prevalent manifestation of GenAI is the large-language model (LLM) chatbot, enabling users to seek information and assistance. This paper aims to showcase how data on student interaction with a chatbot can be used in learning analytics to gain insights into SRL. This is achieved by adapting existing SRL frameworks to comprehend 34 students' interaction with an educational Socratic chatbot for a statistics class at the introductory undergraduate level. Chatbot conversations from students are categorized into learning actions and processes using the framework's process-action library. Thereafter, we analyze this data through ordered epistemic network analysis, furnishing valuable insights into how different students interact with the chatbot. Our findings reveal that higher-scoring students engage more frequently in reflective and evaluative activities, while lower-scoring students focus on searching for answers. Furthermore, students should shift from structured problem-solving, such as solving classroom questions, to questioning fundamental concepts with the chatbot and soliciting more examples to improve their learning gains.

## Notes for Practice

- This study introduces a novel method to categorize and analyze student interactions with a chatbot, producing data that can be studied. This data offers insights into how different student ability levels and learning gains affect their interaction behaviours with the chatbot.
- Large-language model (LLM) chatbots have shown potential in supporting self-regulated learning (SRL), focusing on learning actions and processes.
- Tailoring chatbot prompts based on student performance can help scaffold lower-performing students toward more complex interactions and ensure that higher-performing students continue to engage in SRL actions.
- Generalization of results calls for validation through larger, more diverse studies; the development of other forms of learning chatbots; and the exploration of chatbots' roles in various learning contexts.

## Keywords

Self-regulated learning, learning analytics, epistemic network analysis, ordered network analysis, generative AI, chatbot.

**Submitted:** 03/07/2024 — **Accepted:** 02/02/2025 — **Published:** 15/03/2025

<sup>1</sup> Email: [joellai@ntu.edu.sg](mailto:joellai@ntu.edu.sg) Address: Institute for Pedagogical Innovation, Research and Excellence (InsPIRE), Nanyang Technological University, 50 Nanyang Avenue, Singapore 639798, Singapore. ORCID iD: <https://orcid.org/0000-0002-5619-2051>

<sup>2</sup> Email: [qiuwei@ntu.edu.sg](mailto:qiuwei@ntu.edu.sg) Address: InsPIRE, Nanyang Technological University, Singapore. ORCID iD: <https://orcid.org/0000-0003-4030-9718>

<sup>3</sup> Email: [maung.thway@ntu.edu.sg](mailto:maung.thway@ntu.edu.sg) Address: InsPIRE, Nanyang Technological University, Singapore

<sup>4</sup> Email: [lei.zhang@ntu.edu.sg](mailto:lei.zhang@ntu.edu.sg) Address: InsPIRE, Nanyang Technological University, Singapore

<sup>5</sup> Email: [nurabidah.jamil@ntu.edu.sg](mailto:nurabidah.jamil@ntu.edu.sg) Address: InsPIRE, Nanyang Technological University, Singapore

<sup>6</sup> Email: [lschit@ntu.edu.sg](mailto:lschit@ntu.edu.sg) Address: InsPIRE, Nanyang Technological University, Singapore

<sup>7</sup> Email: [samuel.ng@ntu.edu.sg](mailto:samuel.ng@ntu.edu.sg) Address: InsPIRE, Nanyang Technological University, Singapore

<sup>8\*</sup> Corresponding Author Email: [lim.fun.siong@ntu.edu.sg](mailto:lim.fun.siong@ntu.edu.sg) Address: InsPIRE, Nanyang Technological University, Singapore. ORCID iD: <https://orcid.org/0000-0001-8887-6047>

## 1. Introduction

Learning analytics is “the measurement, collection, analysis, and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs” (Long & Siemens, 2011). The objective is to gain insights into learning processes and environments to optimize educational outcomes. Drawing upon methodologies from educational data mining, machine learning, and psychometrics, learning analytics leverages big data to enhance educational practices (Siemens, 2013). The increasing availability of substantial data in education and the demand for personalized learning solutions fuel the adoption of learning analytics. Its applications span various domains in teaching and learning (Tsai et al., 2019), including personalized learning, early prediction and intervention systems (Qiu, Khong, et al., 2024), and curriculum enhancement (Ferguson, 2012). Learning analytics focuses on understanding how students engage in the learning process. Self-regulated learning (SRL) as a theoretical framework posits that learners actively manage their learning environments, actions, and behaviours to achieve specific learning objectives (Zimmerman, 2008, 2015). This concept underscores the significance of learners’ active participation in their learning journey, including interactions with learning materials, adaptation of strategies to meet goals, and accountability for outcomes (Dunlosky & Ariel, 2011). While various theoretical models offer diverse perspectives on SRL, they commonly emphasize processes such as goal-setting, strategic planning, and evaluation for effective learning (Pintrich, 2000; Zimmerman & Schunk, 2011; Panadero, 2017). Further research on the adaptability of these strategies in different learning contexts would enrich the discourse on SRL.

In particular, with the rise in the use of generative AI (GenAI), this work seeks to understand how SRL can be examined in the learning context of GenAI chatbot usage. The teaching and learning landscape is evolving rapidly in a brittle, anxious, nonlinear, and incomprehensible (BANI) world (Bushuyev et al., 2023). BANI describes the unique traits of our modern environment, such as the swift transformations in the education space. This evolution is fuelled by rapid advancements in GenAI, which has the potential to revolutionize education as students increasingly interact, often independently, with GenAI tools that support learning in novel ways. These tools influence how students decide on their learning environment and self-regulate their learning, raising questions about how to harness GenAI to enhance students’ motivation, engagement, and academic performance. This study seeks to address this challenge by investigating how student interaction with GenAI can serve as a lens to understand SRL in a rapidly changing learning environment.

## 2. Literature Review

SRL yields significant psychological and socio-cognitive benefits that enhance learners’ and instructors’ experiences. Empirical studies have shown that SRL positively influences academic achievement, motivation, and self-perception among students (Zimmerman, 1989; She et al., 2023). For learners, adopting SRL fosters the development of self-efficacy through active progress monitoring, goal-setting, and feedback integration, thereby refining and enhancing learning strategies. This success cultivates metacognitive awareness, goal orientation, and emotional resilience (Daniela, 2015; Clemons & Hopkins, 2020; Stoeger & Ziegler, 2010; Williamson, 2015). Equipping students with these SRL skills provides essential psychological tools for lifelong learning, personal development, and success across various domains. Additionally, by assuming greater responsibility for their learning, students enable educators to focus on guidance rather than control, optimizing instructional effectiveness (Chung, 2000; Zhang et al., 2022). This shift reduces the teaching burden, fostering a supportive educational environment that enhances student engagement, autonomy, and motivation, thereby enhancing instructors’ professional satisfaction and effectiveness (Brenner, 2022). The relevance of SRL becomes even more pronounced in distance learning, such as during the COVID-19 pandemic, where learners and educators must adapt to unprecedented digitalization and distance learning challenges through other forms of emerging educational technologies such as virtual and augmented labs (Edisherashvili et al., 2022; Biwer et al., 2021; Cheong et al., 2023).

Zimmerman has significantly advanced the field of SRL with three distinct models. His first model, created with Pons, identifies 15 self-regulation strategies, aiding in recognizing SRL traits in learners and their interaction with technologies such as chatbots (Zimmerman & Pons, 1986). The second model, the triadic analysis of SRL, examines the relationships among environmental, behavioural, and personal factors, drawing on Bandura’s theories of social cognition and reciprocal determinism (Zimmerman, 1989; Bandura, 2002). The third and final model, Zimmerman’s cyclical phases model, builds on his earlier work and investigates metacognitive and motivational dynamics across three phases: forethought, performance, and self-reflection, emphasizing processes like task analysis and self-motivation (Zimmerman & Schunk, 2011; Zimmerman & Campillo, 2003). Pintrich is well known for his extensive SRL framework, which enhances understanding of the link between SRL and motivation. He created the Motivated Strategies for Learning Questionnaire (MSLQ), and his model outlines four sequential phases—forethought, monitoring, control, and reflection—covering cognition, motivation, behaviour, and context. This comprehensive approach supports the application of the model across various settings (Pintrich, 2000). Winne and Hadwin focus on the metacognitive components of SRL, proposing a model where learners actively manage their learning processes. Rooted in information processing theory, their model depicts SRL through four recursive phases: task definition, goal-setting, enactment of strategies, and metacognitive adaptation. This model, which includes task conditions, operations, products,

evaluations, and standards, is particularly relevant in computer-supported educational environments (Winne & Hadwin, 1998; Winne, 2011; Panadero et al., 2015; Panadero, 2017).

Recent advances in SRL research conceptualize it as a dynamic, adaptive process that unfolds in response to situational demands rather than following a fixed cyclical model. This shift reflects a broader understanding that self-regulation is not static but rather flexible and sensitive to changes in the learning environment (Bakhtiar & Hadwin, 2020; Järvelä & Bannert, 2021). Bakhtiar and Hadwin, for example, emphasize the dynamic interplay between various modes of regulation in response to motivational challenges, demonstrating how learners adjust their regulatory behaviours in real time to sustain engagement and collaboration. Building on this, Järvelä and Bannert argue that SRL can be observed at different temporal levels, from moment-to-moment behavioural adjustments to longer-term changes in learners' beliefs and mindsets (Järvelä & Bannert, 2021). Such a perspective underscores the adaptive nature of SRL, where students continually monitor and respond to immediate situational demands.

Azevedo and colleagues (Azevedo et al., 2004, 2007; Greene & Azevedo, 2009) present a foundational framework for analyzing SRL through a dual-level approach that examines both micro- and macro-level processes. At the micro level, their framework captures specific, observable actions, such as help-seeking, monitoring, and information-seeking, primarily through think-aloud protocols. At the macro level, these actions are aggregated into broader SRL processes like planning, monitoring, and reflecting, offering a comprehensive view of how learners adapt to task demands over time. A key strength of their framework is its focus on temporal adaptation, emphasizing how learners adjust strategies in response to task difficulty and evolving goals. This framework for SRL is beneficial as it enables researchers and educators to understand how individual SRL processes contribute to learning complex subjects in hypermedia settings. These temporal and adaptive processes are particularly relevant for understanding interactions with generative AI tools, which present unique opportunities to examine how learners adjust their regulatory strategies in response to interactive, responsive technologies. Furthermore, recent literature highlights the potential of AI to support personalized, hybrid human-AI learning environments. Molenaar discusses the personalization of learning through AI, noting that AI can enhance SRL by facilitating real-time adjustments tailored to individual needs (Molenaar, 2021).

GenAI technologies facilitate prompt-driven, contextually relevant content generation across various computing platforms. They produce content formats similar to human-created materials, supporting diverse educational initiatives. Platforms like TutorAI and NOLEJ exemplify GenAI applications in swiftly generating interactive educational content (Dilmegani, 2023). These innovations enable personalized learning experiences, including customized study plans and real-time feedback, surpassing the efficiency and scalability of traditional educational methods. GenAI has been particularly effective in teaching functional skills such as coding and writing (Li et al., 2023; Hartley et al., 2024). Its ability to deliver immediate, tailored feedback significantly enhances the learning process by addressing each student's strengths and weaknesses. Additionally, it aids educators in designing lessons and content that cater to diverse learner profiles and needs, facilitating more effective and personalized educational experiences without requiring human instructors or peers to be present.

It should be noted that systematic reviews of tutoring bots, before the emergence of these GenAI-based chatbots, had also emphasized their potential to personalize learning experiences and provide immediate feedback. Kuhail and colleagues (2022) suggest how a structured approach could enhance the growing adoption of chatbots in educational settings, highlighting their pivotal role in delivering personalized learning experiences and instant support. Another systematic review, by Okonkwo and Ade-Ibijola (2021), further underscores the importance of chatbots in providing personalized learning and immediate support. Educational chatbots serve as knowledge repositories, facilitating information retrieval, concept review, and interactive practice sessions, bolstering the SRL processes. Moreover, they help students articulate learning goals, break them down into manageable steps, and set realistic timelines, promoting structured learning approaches and goal-oriented behaviours. While there is room for improvement in chatbots' provision of motivational and emotional support, their extensive use in education underscores the need for research focused on learning analytics to assess their effectiveness in promoting SRL (Wollny et al., 2021; Chocarro et al., 2021; Meng & Dai, 2021).

Despite GenAI being an emerging and disruptive technology in education, there remains a gap in how student interaction with GenAI can be used to investigate SRL and consequently drive learning analytics strategies to uncover and enhance students' motivation and performance. Current literature on SRL largely addresses traditional educational settings or other forms of digital interaction. Still, it does not adequately account for the unique affordances posed by GenAI (e.g., contextualized real-time interaction, conversational learning support, and tracing and analyzing learner behaviour). While GenAI's role in education is expanding, research exploring how student interactions with these tools may impact SRL processes is scarce. This study aims to address this gap, providing insights into how GenAI could be used to support SRL and inform targeted learning analytics strategies. This work addresses this gap by answering the following research questions: (1) How do we process conversations with educational chatbots from the perspective of SRL? (2) What are the similarities and differences in the interaction behaviours of students in (a) ability and (b) learning gain groups?

To answer these research questions, we examine the existing frameworks for SRL in Section 3, choosing a framework

that best illustrates the concept of applying learning analytics in GenAI for SRL. In the same section, we outline the method used to classify chatbot conversations according to the chosen framework. We apply this to 34 students' interaction with an in-house educational Socratic chatbot built for a statistics class taught at the introductory undergraduate level. This will help us to answer the second research question, with results and discussion presented in Section 4, before concluding in Section 5. By examining how students interact with GenAI and analyzing their SRL strategies in this context, we aim to reveal critical insights for educators, academic leaders, learning designers, and students. This study not only fills a gap in the literature but also provides actionable findings that can help design learning analytics tools to support SRL and improve student motivation and performance.

### 3. Methodology

#### 3.1 Chatbot Development

Unlike conventional chatbots that provide direct answers, a Socratic chatbot employs a question-based approach that encourages students to explore their own understanding and reasoning. This aligns with the principles of the Socratic method, which uses guided questions to foster deeper learning and critical thinking—a key aspect of SRL. Through discussions with university faculty, we learned that there is a growing demand for chatbots that not only provide guidance but also encourage students to arrive at answers independently. Faculty indicated that such a design could better support SRL by prompting students to engage in reflection and self-assessment, rather than relying solely on AI for immediate solutions.

Based on this feedback, we developed a Socratic chatbot that aligns with these educational priorities, supporting students in building critical SRL competencies. Our choice of a Socratic chatbot was driven by its potential to enhance SRL processes. Socratic questioning encourages students to actively engage in the learning process, promoting self-regulation through self-questioning, goal-setting, and reflection. In contrast to answer-based chatbots, this approach helps students cultivate independent learning habits, thereby potentially enhancing their motivation and overall academic performance.

The chatbot is designed and developed with Streamlit, an open-source, user-friendly Python web framework that seamlessly combines the capabilities of both front end and back end. When a student interacts with the chatbot, an input query from the student is received and preprocessed before being sent to the large language model (LLM) to generate a human language response. A default system prompt is included with the input query to guide the LLM for the desired response, serving as an initial instruction.

The chatbot leverages a knowledge base that stores relevant data resources to enrich the context. This knowledge base provides additional information to support the LLM in generating relevant and accurate responses. This knowledge base has been preprocessed to create a vector store, enabling efficient knowledge retrieval. During the generation of responses, the chatbot also considers the student's conversation history, considering up to 40 previous conversation exchanges with the chatbot.

At the core of the chatbot lies GPT-4, a large multimodal model that generates the actual responses. These responses, the processed system prompt, and the input query are stored in JavaScript object notation (JSON) file format. Each response is associated with a unique identifier assigned to an individual student for further analysis on refining the custom chatbot to achieve a better user experience in learning. The front-end chat interface is shown in Figure 1. The chatbot's primary role is summarized in the following key metaprompt:

*As an AI tutor, your role is to guide the student to understand the Statistics course materials by breaking down the student's questions in a step-by-step manner. You need to break down the student's main question into several steps and ask follow-up questions one at a time to guide the student. Make sure that your questions help the student to understand the original question one step at a time.*

The Socratic chatbot is designed to ask open-ended follow-up questions that will help the student to think critically. The student needs to figure the answer out on their own. Thus, the chatbot was instructed not to answer the students' questions if they directly asked about a work or example problem.

#### 3.2 Research Methodology

This study was conducted in two parts, involving an experimental investigation with a prototype Socratic chatbot and an analysis using an SRL framework. The experimental study comprised 34 undergraduates enrolled in a three-week pilot class on introductory statistics. Before the interaction with the chatbot, a pre-course test was administered to establish a baseline measure of students' knowledge of statistics. The class adopts a blended learning approach, where the first week is onsite, while the following two weeks are conducted online. Both the onsite and online sessions were conducted asynchronously. Throughout the three weeks, student participants were required to watch related online videos and engage with a prototype Socratic chatbot embedded with class-specific content. The Socratic chatbot was further prompt-engineered to respond to students' questions with questions in a scaffolded manner (Qiu, Su, et al., 2024). The participants interacted and inquired about statistical concepts

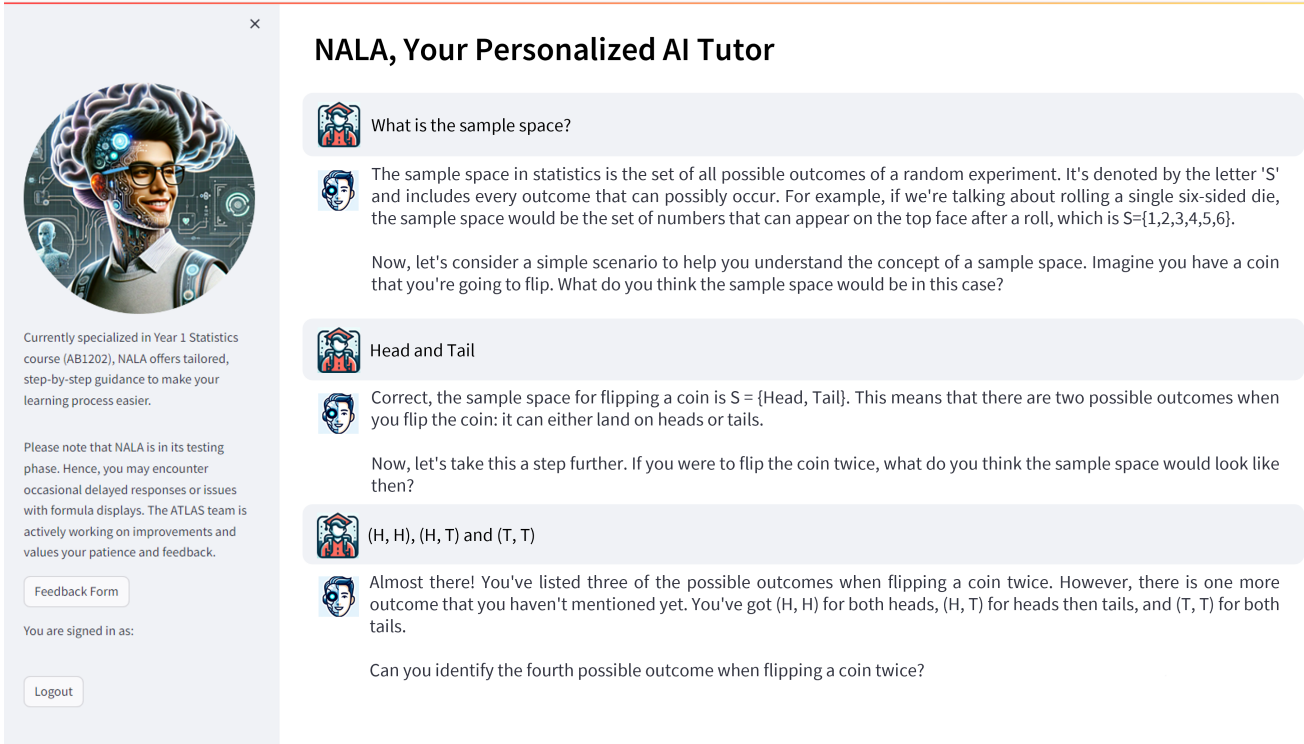


Figure 1. Chat interface that students will interact with for each session.

at their own pace. No human tutor was involved in this self-regulated class. In other words, the intervention of this pilot class is the Socratic statistics chatbot. After three weeks, a post-course quiz was administered to evaluate the knowledge gains. This quiz covers the same concepts as the pre-test, comprising different questions to measure learning progression. The timeline of the experimental study is captured in Figure 2.



Figure 2. Timeline for experimental study comprising a pre-test and a post-test, covering 12 topics in an asynchronous onsite-online blended learning format.

In the second part of the study, the collected chat logs were meticulously analyzed by tagging each student’s input using categories derived from an established SRL framework. This tagging helped identify the types of SRL actions students engaged in during their interactions with the chatbot. A process map and ordered epistemic networks were then constructed to visualize conversation flow and correlate these interactions with quantitative measures such as time spent, number of interactions, and

pre- and post-test scores.

### 3.3 SRL Framework

Various frameworks have been developed to articulate conceptual and practical applications of SRL, reflecting diverse perspectives on how learners can effectively manage and direct their learning activities. Four influential SRL frameworks, developed by Zimmerman and Campillo (2003), Pintrich (2000), Winne and Hadwin (1998) and Azevedo and colleagues (2007), contribute unique insights and tools for implementing SRL in educational settings. These models highlight the processes involved in SRL and align closely with its benefits, providing a structured approach to enhancing learner autonomy and effectiveness.

For our study, we use the process-action framework model (Lai, 2024a). It is developed by considering the metacognitive perspective (called *processes*) from Winne and Hadwin’s model and the four phases of SRL by Pintrich and Zimmerman’s subprocesses (called *actions*), as well as adopting the hierarchical structure by Azevedo and collaborators. This concise framework comprises four processes—defining, seeking, engaging, and reflecting—and nine actions—problem identification, goal-setting, information search, content selection, material review, material organization, performance rehearsal, task evaluation, and self-evaluation, facilitating easy implementation for learning with GenAI. This framework is suitable for our study as it contains actions and processes that can be directly interpreted from the content of their chat with the chatbot without having participants report their SRL actions or processes. Consequently, we can tag each line of conversation with an associated process and action to the chatbot. Table 1 summarizes these processes and the description and example conversations to describe the process-action.

**Table 1.** Process-action framework containing the definitions and examples of conversations associated with the process-action. Adapted under Creative Commons CC BY 4.0 license.

Process	Action	Code	Process-Action Description	Generic Conversation Examples
Defining	Identification	D.I	Conversation lines indicate the identification of problems or provide a chatbot with the problem.	These are important ...; I am given a problem ...
	Goal-set	D.G	Conversation lines indicate the setting of educational goals or sub-goals.	I need to achieve ...; I want to learn/know ...; Create a checklist ...
Seeking	Search	S.S	Conversation lines indicate efforts to secure information from a chatbot.	Tell me about ...; Please define ...; What is ...
	Select	S.SL	Conversation lines indicate efforts to record important information or results.	What are the key points ...; Extract the main idea of ...; Please recommend some resources ...
Engaging	Review	E.RV	Conversation lines indicate efforts to re-engage with curated material or check for correctness.	This does not seem to be correct ...; Does this mean that ...; The answer is ...
	Organize	E.O	Conversation lines indicate rearrangement of materials by category or in chunks. Extracting course materials and definitions.	Can you help me to summarize ...; Put this in a list/table ...; Tell me step by step ...
	Rehearse	E.RH	Conversation lines indicate efforts to memorize material or put it into practice. Asking for examples.	Can you help me fill the gaps ...; Please generate questions on ...; I want to spend more time on ...
Reflecting	Task Evaluation	R.ET	Conversation lines indicate evaluations of their work quality. Conversation lines reflect the completion of tasks.	What is the next thing ...; I learned (topic); How is this related to (another topic) ...
	Self-Evaluation	R.ES	Conversation lines indicating evaluations of motivation. Conversation lines show an indication of (not) understanding.	Can you tell me where I am lacking ...; I am not getting it ...; Yes, I understand ...

### 3.4 Inter-Annotator Agreement and Reliability for Process-Action Tag

Inter-annotator agreement (IAA) is adopted in various fields, such as natural language processing, information retrieval, and content analysis, where multiple annotators independently label or code data (Artstein & Poesio, 2008; Artstein, 2017). The primary objective of performing IAA is to determine the level of agreement among multiple annotators when tagging categorical

data. Specifically, in tagging each conversation line with the process-action framework, we employed three independent annotators to tag every conversation line between the student and the chatbot. A chief annotator is appointed to reconcile any disagreements. We use these tags and data collected to provide insights into learning analytics through network analysis.

Krippendorff’s alpha is a statistical measure used to assess the reliability of agreement among annotators assigning classifications to a set of items (Krippendorff, 1995). This measure is particularly valuable in content analysis and data classification tasks, accommodating varying numbers of annotators and handling missing data efficiently, thus offering a flexible tool for assessing rater consistency in diverse research settings. Krippendorff’s alpha is given by

$$\alpha = 1 - \frac{D_0}{D_e}, \tag{1}$$

where  $D_0$  is the observed disagreement among annotators and  $D_e$  is the expected disagreement under random assignment.  $\alpha$  evaluates the extent of agreement beyond what could be expected by chance, expressed on a scale from  $-1$ , denoting complete disagreement, to  $1$ , denoting complete agreement, with  $0$  indicating agreement equivalent to chance. Lower observed disagreement than the expected disagreement yields a higher alpha value, signifying stronger reliability. We estimate the confidence interval of  $\alpha$  using bootstrap resampling, a statistical technique to approximate the sampling distribution of  $\alpha$  and hence estimate its confidence interval.

### 3.5 Learning Analytics through Network Analysis

Epistemic network analysis (ENA) is a valuable tool within the landscape of SRL and learning analytics, particularly for studying knowledge construction processes (Shaffer et al., 2009). By quantitatively analyzing the structure of connected learning and reasoning within annotated data, ENA enables researchers to identify key transitions in the learning cycle, assess the effectiveness of the educational tools, and inform instructional practices aimed at fostering conducive SRL environments. In essence, ENA’s ability to uncover hidden patterns and structures within annotated data offers a novel lens through which to explore the dynamics of knowledge building and its evolution in longitudinal studies.

Ordered network analysis (ONA) is a technique for modelling the structure of directed connections in data. ONA extends the theoretical and analytical advantages of ENA to account for the order of events by producing directed weighted networks rather than undirected models (Tan et al., 2023). Like ENA, ONA takes coded data as input, identifies and measures connections among coded items, and visualizes the structure of connections in a metric space that enables both statistical and visual comparison of networks. However, ONA models the order in which codes appear in the data, enabling analysis of phenomena in which the order of events is hypothesized to be important. ONA models the connections among codes by quantifying the co-occurrence of codes within conversations, producing a directed weighted network of co-occurrences, along with associated visualizations for each unit of analysis in the data. In our context, these units correspond to a sequence of process-action codes associated with students’ conversation history. The mathematical formalism of ENA and ONA is not discussed in this article, but Bowman and colleagues (2021) provide a comprehensive guide on the mathematical foundations of such networks.

Next, we partitioned the students into ability and gains groups, as described in Table 2. The first subgroup (A1) contains students who score at most 50 (out of 100) on the pre-course test, and the second subgroup (A2) contains students who score above 50 on the pre-course test. For the gains group, we partitioned the students into three. The first subgroup (G1) contains students whose improvement from the pre- to the post-course test was at most 15. The second subgroup (G2) contains students with improvements between 15 (exclusive) and 35 (inclusive), while the third subgroup (G3) contains students with improvements greater than 35. The ENA WebTool (Wisconsin Center for Education Research, 2024) was developed by the Wisconsin Center for Education Research at the University of Wisconsin-Madison and is used in the present study to plot the ordered networks.

**Table 2.** Partition of participants into ability groups and gains groups.

Participants	<i>N</i>	(%)
	<b>34</b>	<b>100.0</b>
<i>Ability</i>		
A1: pre-course test score $\leq 50$	16	47.1
A2: pre-course test score $> 50$	18	52.9
<i>Gains</i>		
G1: $\Delta \leq 15$	11	32.4
G2: $15 < \Delta \leq 35$	11	32.4
G3: $35 < \Delta$	12	35.3

$\Delta$ : post-course test score – pre-course test score

### 3.6 Comparative Analysis

For our analysis, we focused on student conversations as our primary data source. The lines of analysis included students’ conversation histories with the chatbot during the experimental phase. The units of analysis were defined as each individual student, categorized by Ability or Gains Group, to facilitate comparisons based on student grouping. Conversations were selected based on these groupings, allowing us to analyze network structures associated with each specific group. We used an infinite window size (stanza model) for our analysis. This window size allowed us to capture the complete conversational context without imposing artificial boundaries, providing a holistic view of the interaction patterns within each conversation. In this work, we study specific patterns of back-and-forth exchanges between the student and the chatbot, such as the sequence of questions and answers at high resolution (Siebert-Evenstone et al., 2017). By detailing these parameters, we aimed to enhance the interpretive clarity of our findings and ensure the replicability of our analysis.

We conducted planned comparisons based on student ability or gains groups to analyze differences between groups. Specifically, we compared network structures and interaction patterns between these groups to understand how student ability levels or gains impact network outcomes.

Lastly, given that our data did not meet the assumptions for parametric testing, we used the Mann–Whitney U test to assess statistical differences between groups (Mann & Whitney, 1947). The Mann–Whitney U test is a non-parametric test suitable for comparing two independent groups, making it appropriate for our data. We report *p*-values for each comparison, noting significance levels where applicable, to highlight statistically meaningful differences in network interactions between groups.

## 4. Results and Discussion

The distribution of pre-course and post-course test scores is reported in Figure 3. Although not the focus of this present study, a paired *t*-test was conducted to compare the pre-course test scores with the post-course test scores. The mean difference between the pre-course and post-course test scores is approximately 27.21, with a standard deviation of 21.18. The analysis revealed a statistically significant improvement from pre- to post-course test scores ( $p < 0.001, d \approx 1.28$ ). In total, we have 16 and 18 students classified under A1 and A2, respectively. Additionally, 11, 11, and 12 students were classified in G1, G2, and G3, respectively.

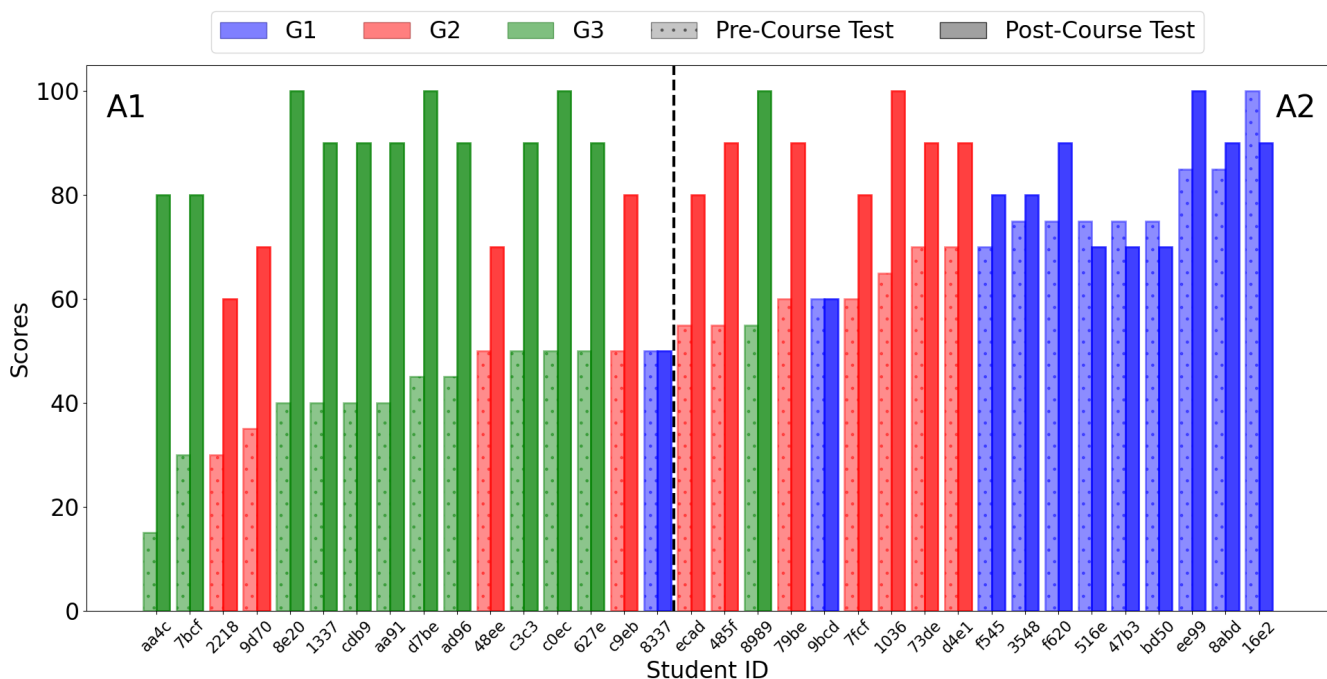


Figure 3. Score distribution and the number of students in each group.

### 4.1 Inter-Annotator Agreement and Reliability

After conducting the tagging process, we assessed the inter-annotator reliability of the categorical data using Krippendorff’s alpha, yielding a value of  $\alpha = 0.693$ . The nominal level of measurement was chosen based on the nature of the data, with categories assigned to the content of conversation exchanges. We employed bootstrapping techniques to ascertain the robustness and reliability of our coding agreement estimate. Bootstrapping is particularly useful when calculating confidence intervals

for statistics like  $\alpha$  that lack a straightforward analytical solution for their variance. We generated 1000 bootstrap samples by resampling the coded data with replacement. For each sample,  $\alpha$  was recalculated to estimate the sampling distribution of the measure. This process allowed us to derive a 95% confidence interval of (0.619,0.973). This confidence interval suggests that we are 95% confident that the true value of Krippendorff’s alpha for our dataset falls between 0.619 and 0.973. This interval provides valuable information about the precision of our estimate and helps us interpret the reliability of the tagging agreement.

### 4.2 Conversation Data Tagging

The distribution of tagged codes and ordered code-pairs is summarized in Table 3. Of the 6716 conversation lines, we coded 6689, where the remaining 27 were lines unrelated to SRL. Referencing the same table, we only counted conversation lines by the same student on the same day for the ordered pairs, which depict how students transit from one action to another.

**Table 3.** Summary table of coded conversation lines ( $N = 6689$ ), including the frequency of each code and the number of ordered pairs, reading the row first.

	D.I	D.G	S.S	S.SL	E.RV	E.O	E.RH	R.ET	R.ES
N	454	17	1399	23	3871	43	210	181	491
D.I	57	0	42	1	313	1	1	2	28
D.G	1	1	3	1	7	0	2	0	2
S.S	45	5	535	6	541	7	65	34	119
S.SL	3	0	9	4	5	0	1	1	0
E.RV	245	4	437	4	2650	16	80	103	259
E.O	2	0	9	0	13	8	2	3	4
E.RH	3	0	42	1	110	2	25	7	12
R.ET	24	1	86	0	15	2	11	2	3
R.ES	29	1	115	4	209	3	16	29	63

Initially, our network models included all the codes: D.I, D.G, S.S, S.SL, E.RV, E.O, E.RH, R.ET, and R.ES. The model normalizes the networks for all units of analysis, so including all these codes consistently resulted in no statistical differences in our analyses of both the ability and gains groups. This lack of difference is due to the disproportionate frequency of code co-occurrence, as shown in Table 3. The Socratic chatbot was designed to elicit replies from students through educational scaffolding, leading to a skewed distribution favouring the E.RV and S.S codes. The high occurrence of E.RV is because the chatbot breaks down problems into multiple parts, prompting students to reply to verify their answers. Additionally, there are numerous instances of S.S because students frequently use the chatbot to ask questions. Upon reflection, this could also be because students are not trained in prompt engineering to get the response that they desire. Hence, they are engaging in back-and-forth conversations with the chatbot without eventually arriving at the response that they desire, thus not completing the entire SRL process, but being stuck in E.RV and S.S actions. A snippet of a conversation that exhibits this is shown in Figure 4.

Thus, we remove multiple occurrences of the same code in succession to truncate the data and count only code changes. For example, previously, if we had a student with the process-action tagging D.I→D.I→D.I→S.S→S.S→D.I→D.I, we would have counted this as  $N = 5$  for D.I and  $N = 2$  for S.S; 3 for D.I→D.I; and 1 each for D.I→S.S, S.S→S.S, and S.S→D.I. Under this new data truncation, we will simplify this to D.I→S.S→D.I. The new summary of coded data is in Table 4. Subsequent analyses are conducted on this truncated dataset.

**Table 4.** Summary table of coded data after truncation ( $N = 3372$ ), including the frequency of each code and the number of ordered pairs, reading the row first.

	D.I	D.G	S.S	S.SL	E.RV	E.O	E.RH	R.ET	R.ES
N	400	16	861	19	1224	35	210	179	428
D.I	0	0	42	1	311	1	1	2	28
D.G	1	0	3	1	7	0	2	0	2
S.S	47	5	0	6	538	7	65	34	118
S.L	3	0	9	0	5	0	1	1	0
E.RV	250	4	435	4	0	16	79	103	259
E.O	2	0	9	0	13	0	2	3	4
E.RH	3	0	42	1	110	2	0	7	12
R.ET	24	1	87	0	15	3	11	0	3
R.ES	30	1	116	4	209	3	17	29	0

User	Message	Code
Student	what is the difference between expectation and variance?	S.S
Chatbot	Expectation and variance are two fundamental concepts in statistics that describe different characteristics of a probability distribution.  Let's start with expectation. Can you explain, in your own words, what the expectation of a random variable tells us about the distribution?	-
Student	it tells us the mean and standard variance of the distribution	E.RV
Chatbot	Actually, the expectation of a random variable, often denoted as $E(X)$ , is the theoretical long-run average value of the variable if the experiment were repeated an infinite number of times. It's also known as the mean of the distribution. The expectation is not about the standard variance; rather, it's a central measure that gives us an idea of the central tendency of the data.  Now, let's talk about variance. Variance, denoted as $Var(X)$ or $\sigma^2$ , measures the spread or dispersion of the random variable around the mean (expectation). It is the average of the squared differences from the Mean.  To help clarify the difference, could you tell me what squaring the differences from the mean (as done in the variance calculation) achieves compared to just taking the differences (as in the expectation)?	-
Student	yes	S.S
Chatbot	Great! Since you understand that adding or subtracting a constant to each value in a dataset changes the mean but not the variance, let's consider another operation: multiplying or dividing each value in the dataset by a constant.  How do you think multiplying each value in a dataset by a constant affects the variance?	-
Student	it increases the variance by the constant multiplied	E.RV
Chatbot	That's right, when you multiply each value in a dataset by a constant, the variance is affected by the square of that constant. If the constant is 'c', then the new variance will be:  $Var(cX) = c^2 * Var(X)$  This is because when you multiply each data point by 'c', the deviations from the mean are also multiplied by 'c', and when you square these deviations, the factor becomes $c^2$ .  Now, can you think of how this property of variance might affect the standard deviation when you multiply each data point by a constant?	-
Student	variance is a non negative value so by squaring the mean, it tells us how far the value is from the mean to determine the skewness of the data	E.RV

**Figure 4.** An example of coded conversations between a student and the chatbot. The chatbot is designed to ask only one question at a time by breaking down the main question into scaffolding questions to guide students in answering their questions. Grammatical errors are intentionally left unedited in this example.

Since our analysis used an infinite stanza window, which considers connections across the entire conversation rather than within a limited moving window, this setup minimizes the risk of spurious connections due to data truncation, as each conversational line is analyzed in the context of the entire discourse rather than being influenced by adjacent lines alone.

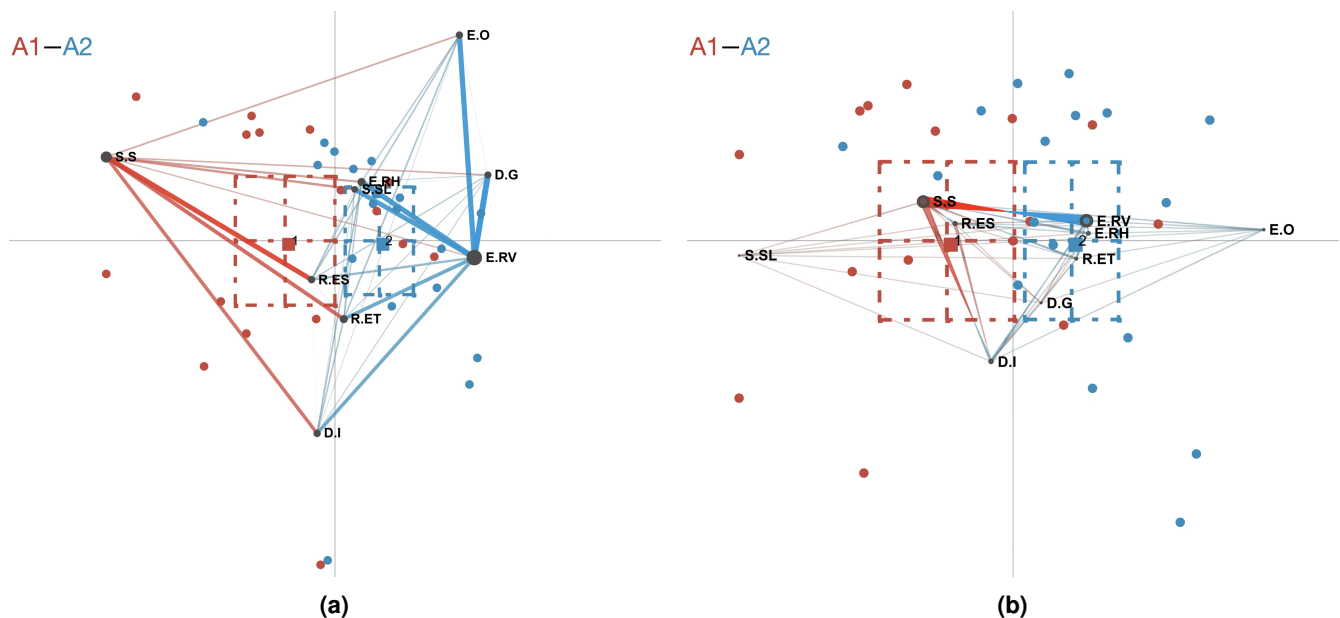
#### 4.2.1 Interaction Behaviours by Student Ability

For comparison, we included all the codes in the construction and analysis of the ordered network. Network nodes correspond to the codes, and edges reflect the relative frequency of co-occurrence between two codes of students in the same group. The ONA model first normalized the networks across all units of analysis before applying dimensionality reduction. Singular value decomposition (SVD) was used for this purpose, identifying the two orthogonal dimensions that captured the greatest variance in the data. The units were then plotted based on these two dimensions. To determine the positions of the nodes in the network graph, an optimization routine was employed to minimize the difference between the plotted points and their respective network centroids, ensuring accurate representation of the relationships within the data. Because of this co-registration of network graphs and projected space, the positions of the network graph nodes can be used to interpret the dimensions of the projected space and explain the positions of plotted points in the space. We performed both ENA and ONA for ability groups.

Using ENA, our model, shown in Figure 5a, had co-registration correlations of 0.99 (Pearson) and 1.00 (Spearman) for the X-axis (SVD1) and co-registration correlations of 0.99 (Pearson) and 1.00 (Spearman) for the Y-axis (SVD2). Along the X-axis, a Mann–Whitney U test showed that A1 (Median = 0.11,  $N = 16$ ) was statistically significantly different from A2 (Median = -0.08,  $N = 18$ ). The test revealed a statistically significant difference in behaviour between the two groups ( $U = 227.00, p < 0.01$ ). The effect size ( $r = -0.58$ ) indicates a large effect, suggesting that the difference in chatbot usage behaviour between groups A1 and A2 is substantial. There was no statistical difference along the Y-axis. Our ONA model, shown in Figure 5b, had co-registration correlations of 0.90 (Pearson) and 0.89 (Spearman) for the X-axis (SVD1) and co-registration correlations of 0.99 (Pearson) and 0.99 (Spearman) for the Y-axis (SVD2), which indicates that our model is a good fit. Along the X-axis, a Mann–Whitney U test showed that A1 (Median = 0.17,  $N = 16$ ) was statistically significantly different from A2 (Median = -0.11,  $N = 18$ ). The test revealed a statistically significant difference in behaviour between the two groups ( $U = 62.00, p < 0.01$ ). The effect size ( $r = 0.57$ ) indicates a large effect, suggesting that the difference in chatbot usage behaviour between groups A1 and A2 is substantial. There was no statistical difference along the Y-axis.

We discuss key observations by examining the comparison epistemic network in Figure 5a. First, we observe that the nodes D.I, S.SL, and E.RH and both R actions are centrally located (along the X-axis) and connected to other process-action codes in networks A1 and A2. This suggests that both groups of students keenly adopt these actions. This is facilitated by the chatbot, which is programmed to check in on students' progress on the task and suggest going through examples by asking questions

such as “How would [concept] help you tackle more complex probability problems?” or “Can you recall [method/concept] for [a simpler problem]?” Such questioning by the chatbot gives students a marker to consolidate their understanding and scaffold their learning by breaking complex problems and tasks into manageable parts.



**Figure 5.** (a) Epistemic and (b) ordered networks comparing students in groups A1 (red) and A2 (blue). This network is calculated by subtracting the weight of each connection in A2 from the corresponding connection in A1. For (a), the frequency of co-occurrence in each group is determined by the thickness of the edges. For (b), the frequency and direction of the edges are determined by the thickness at the ends of the edges. For example, the directed edge between E.RV and S.S shows that A2 engaged in S.S→E.RV more frequently than students in A1, who engaged in S.S→E.RV and E.RV→S.S.

Keeping the process of truncation in mind, we also observe that students in group A2 tend to use co-occurrence interactions like D.G, E.RV, and E.O more frequently, albeit D.G and E.O at a much lower frequency, indicating a more complex approach to scoping what they would like to learn, checking for correctness of answers or understanding, and organizing study materials or online materials through the chatbot. On the contrary, students in A1 more frequently engage in the single action of S.S, with high co-occurrence with D.I and both R actions, indicating a tendency to search for concepts from problem identification and a lower tendency to engage with the materials provided by the chatbot.

For this pilot class, the topics covered were basic, which could explain the lack of interaction with the action S.S by students in the higher-ability group, A2. Nevertheless, this comparison network remains a valuable aspect of our analysis, as it provides critical insights that can inform the customization of chatbot behaviour and responses to better support diverse student needs. For instance, providing more scaffolded prompts for students who score lower could gradually build their ability to engage in more complex interactions, such as D.G with E.RV and E.O, which require students to set personal goals and review and organize information, respectively. The epistemic network highlights distinct interaction patterns between higher- and lower-scoring students. Higher-scoring students engage more in goal-setting and engaging with material, while lower-scoring students focus more on seeking information. These insights can inform targeted educational strategies to foster deeper engagement and improve learning outcomes across all student groups.

The ordered network, illustrated in Figure 5b, further sheds light on the nuanced differences between ability groups. We observe that students in the higher-ability group are equally likely to have this back-and-forth engagement of review and search for materials (S.S↔E.RV). In stark contrast, students with lower initial ability are twice as likely to review and then search (E.RV→S.S) for more material than to search and then review (S.S→E.RV). This means that students in the lower-ability group tend to exhibit a more reactive learning behaviour, where they engage in review first before seeking additional materials, potentially as a way to reinforce their understanding before moving forward. This pattern suggests that lower-ability students may approach learning with a more cautious strategy, focusing on consolidating their knowledge before exploring new content. Conversely, the balanced back-and-forth engagement observed in higher-ability students indicates a more dynamic and exploratory learning approach, where reviewing and searching occur in tandem, reflecting greater confidence and adaptability in their learning processes. This distinction highlights the importance of tailoring educational strategies to support diverse learning behaviours across ability groups.

#### 4.2.2 Interaction Behaviours by Learning Gains

We plotted ordered networks to investigate the interaction behaviours between learning gains groups (ENAs returned inconclusive results). The results are shown in Figure 6. Our model had co-registration correlations of 0.98 (Pearson) and 0.98 (Spearman) for the  $X$ -axis (SVD1) and co-registration correlations of 0.98 (Pearson) and 0.96 (Spearman) for the  $Y$ -axis (SVD2), which indicates that our model is a good fit. Comparing students in G1 and G2 along the  $X$ -axis (SVD1), a Mann–Whitney  $U$  test showed that G1 (Median = 0.13,  $N = 11$ ) was statistically significantly different from G2 (Median =  $-0.20$ ,  $N = 11$ ), with  $U = 109.00$ ,  $p < 0.01$ , and  $r = -0.80$ . Additionally, G2 was statistically significantly different from G3 (Median = 0.15,  $N = 12$ ), with  $U = 21.00$ ,  $p < 0.01$ , and  $r = 0.68$ . There is no mutual statistical difference between the three groups along the  $Y$ -axis (SVD2), and no statistical difference between G1 and G3 along the  $X$ -axis.

When comparing gains, E.RH is the most central node, meaning it is adopted almost equally by students of all three groups. Observing the placement of the nodes and distance to the median of each group, we can conclude that students in G1 engaged predominantly in R.ES, students in G2 engaged mainly in D.G with weaker association with D.I, and students in G3 engaged mainly with S.S. The other actions, R.ET and E.RV, are equally adopted by G1 and G2. Lastly, E.O and S.SL are not found to be strongly associated with any of the three groups.

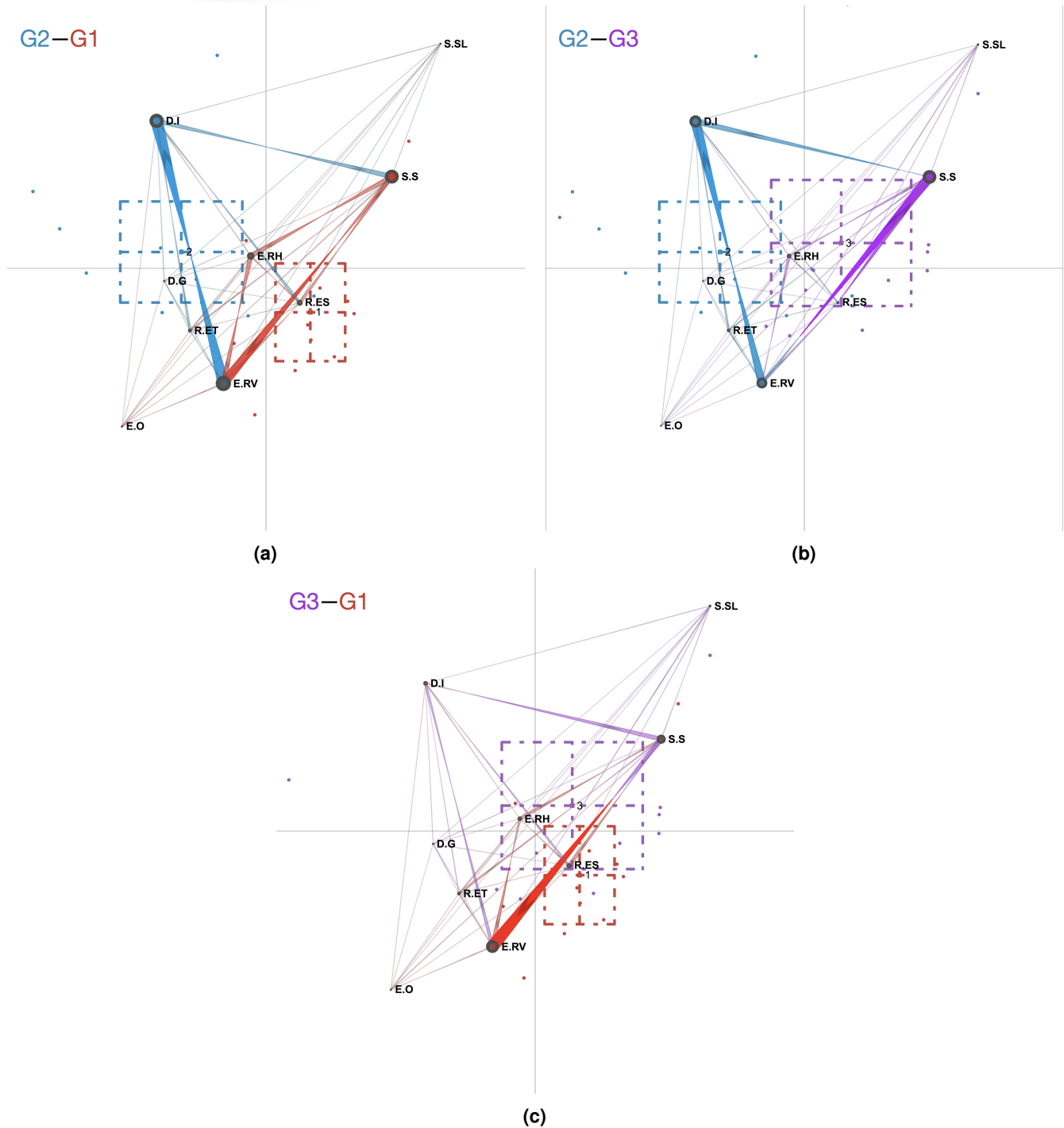
As reported previously, the collective average behaviour of students in G1 and G3 is statistically indistinguishable. One of the factors that visually distinguishes the two, as seen in Figure 6c, is that students in G1 engage in S.S→E.RV more than those in G3. However, given that S.S→E.RV is the transition with the highest frequency, this observation cannot be conclusive. The outcome of not being able to distinguish G1 and G3 students is counterintuitive—a deeper dive into the metaprompt of the Socratic chatbot could suggest reasons for this phenomenon:

*As you engage with the student, please take note of phrases indicating uncertainty . . . You will secretly count the total number of times the student has expressed uncertainty and when the total number is 4, you will only respond by telling the student to revisit certain key foundational topics to enhance [their] understanding and make [their] sessions more productive. You will also offer recommendations on relevant topics [they] should review, and provide a brief explanation of how these concepts tie back to [their] original question and at the end, you will ask the student whether [they] would like to go through any of the foundational topics with you.*

We observe that the metaprompt feedback mechanism is not adaptive. The chatbot does not provide differentiated feedback based on the student's initial baseline ability but rather a standard suggestion to review foundational topics, i.e., engaging in the action S.S↔E.RV. The recommendation to review foundational topics is a one-size-fits-all approach. This might not be necessary for high-performing students, while for low-performing students, it could be essential. Thus, as students in A1 use the chatbot, they could be directed to foundational concepts, which might contribute to their improvement.

Of the 12 students in G3, 11 of them are also classified in group A1, meaning they have the greatest potential for learning gains through interaction with the chatbot. Furthermore, it is worth noting that 10 out of the 11 students in G1 are also in A2, of which eight scored at least 75 for the pre-course test. Thus, this could indicate that while both groups of students adopt effective learning methods that allow students in G3 to have large learning gains, students in G1 who already have high baseline knowledge have less room for measurable improvement. Thus, while it seems counterintuitive that both G1 and G3 have adopted the same behaviour, closer inspection of students' pre-course test scores reveals limitations imposed on students in G1. Quantitatively, an analysis of covariance (ANCOVA) (Wright, 2020) was conducted to examine the impact of gain groups G3 and G1 on post-test scores, while controlling for pre-test scores as a covariate. The covariate, pre-test scores, was significantly related to post-test scores,  $F(1, 20) = 29.35$ ,  $p < .001$ . This indicates that baseline ability, as measured by pre-test performance, had a substantial effect on post-test outcomes. Participants with higher pre-test scores tended to perform better on the post-test, irrespective of their gain group. This finding underscores the importance of adjusting for initial differences in ability when evaluating the effectiveness of interventions. After accounting for pre-test scores, the main effect of gain group on post-test scores remained statistically significant,  $F(1, 20) = 49.88$ ,  $p < .001$ . This demonstrates that, even after controlling for baseline ability, participants in G3 outperformed those in the low-gain group G1 on the post-test. The large  $F$ -value suggests a strong and robust effect of the intervention on learning gains, independent of initial ability differences. These findings highlight two key insights: First, baseline ability significantly influences post-test performance, reaffirming the necessity of accounting for individual differences in ability when assessing intervention efficacy. Second, the substantial difference between gain groups, even after adjusting for pre-test scores, indicates that the intervention associated with the high-gain group is particularly effective at enhancing learning outcomes.

Having postulated that the learning behaviour adopted by G1 and G3 is effective, we now look at students in G2 and ask how to move them into G3. Notably, all students in G2 have the potential of improvement seen in G3. Thus, what is it about the learning behaviour of students in G3 (and by extension G1) that we want to see in students in G2? To answer this question, we compare the respective ordered networks between G1 and G2 and between G2 and G3. Immediately, we see that students in G1 and G3 engaged more in S.S↔E.RV. These actions are characteristic of students who repeatedly ask the chatbot basic



**Figure 6.** Ordered networks comparing students in (a) groups G2 and G1, (b) groups G2 and G3, and (c) groups G3 and G1. These graphs are calculated by subtracting the weight of each ordered connection in the second group from the corresponding ordered connection in the first group. The frequency and direction of the edge are determined by the thickness at the ends of the edge. For example, the directed edge between E.RV and S.S in (c) shows that students in G1 engaged in S.S→E.RV more frequently than students in G3, who engaged in S.S→E.RV and E.RV→S.S.

conceptual questions and directly reply to the chatbot’s question. We further observe that E.RH features more prominently among students in G1 and G3 than in G2. On the other hand, students in G2 engaged more in D.I↔E.RV. When analyzing student responses associated with D.I, students in G2 tend to copy and paste questions from a question sheet into the chatbot and progress from there. Students in G2 are also less inclined to ask for examples. Thus, it is plausible to conclude that moving

from G2 to G3, where learning gains are larger, requires the student to ask more fundamental questions and seek more examples rather than focus on complex problem-solving ones. For more effective questioning, the chatbot can be designed to encourage fundamental questioning by showing examples of well-structured basic questions. The Socratic chatbot can also be designed to be modular, where students start with basic definitions and gradually progress to more complex applications, ensuring that students master each level before moving on.

### 4.3 Discussion

Framing SRL as an adaptive, context-sensitive process provides a valuable lens for interpreting our findings. Our results demonstrate that students engaging with the Socratic chatbot frequently sought and reviewed information, illustrating dynamic and situationally driven regulatory behaviours. These findings align with prior research of Järvelä and colleagues (2021), who stressed the complex and non-linear nature of SRL in authentic contexts. Students do not need to go through the SRL process methodically or in a cyclic manner to experience knowledge gains. Nonetheless, given that the effectiveness of the chatbot relies on students' self-regulation, the framework would be useful beyond Socratic chatbots in understanding how students use chatbots in general.

By examining SRL in real-time interactions, we observed that students shifted between phases of regulation based on situational needs, revealing insights into both their immediate regulatory strategies and broader, evolving patterns of engagement. This perspective enriches our understanding of how GenAI can uniquely influence and support the fluid, context-driven nature of SRL, highlighting the potential of such technology to foster self-regulatory behaviours within complex, dynamic learning environments. Our observation of students demonstrating a more fluid and dynamic approach to self-regulation instead of following a rigid sequence diverges from previous theoretical models. It also highlights the complexity and variability inherent in real-world learning environments and aligns better with more recent advances in our understanding of SRL.

In the case of the Socratic chatbot, its design to break down complex problems into simpler steps and scaffold the learning for students resulted in a heavy emphasis on seeking and engaging with the learning material and topics. This finding, corroborated by earlier studies by Tegos and colleagues (2016), has important implications for educational practices and interventions designed to support SRL in an age of GenAI where GenAI chatbots can scaffold SRL processes in real time by dynamically adapting their prompts to the learner's inputs. It underscores the need for flexible and adaptable theories and strategies to accommodate the diverse and non-linear ways students manage their learning.

Conversely, while our findings suggest that the Socratic chatbot may support SRL by encouraging reflective and self-directed learning behaviours, it is important to consider the limitation posed by the lack of a control condition. Without a comparison group, we cannot definitively conclude that these outcomes are solely due to the chatbot interaction. This limitation aligns with challenges noted by Greene and Azevedo (2009), where further research is needed to establish causal effects in the complex interplay between learners and AI tools. Regardless, it is important to note that while all students may use the Socratic chatbot, not all engage in SRL actions that lead to meaningful learning gains. Additionally, simply providing access to pedagogically engineered chatbots may not be sufficient to ensure improved learning outcomes (Fan et al., 2025; Zhai et al., 2024). To address this, learning analytics could play a crucial role in identifying how students are interacting with the chatbot and pinpointing areas where they may need additional guidance (Chihab et al., 2023). By analyzing these interactions, the chatbot itself can guide students to ask more conceptual questions or explore concrete examples that deepen their understanding. This approach aligns with the earlier point about leveraging learning analytics to provide personalized, data-driven support, ultimately enhancing the effectiveness of the Socratic chatbot in promoting SRL and learning outcomes.

Instructors can use the data generated from chatbot interactions to gain valuable insights into students' learning processes and behaviours, thereby advancing the science of learning. By analyzing patterns in the types of questions asked, the frequency of interactions, and the progression of responses, instructors can identify common misconceptions, knowledge gaps, and areas where students tend to struggle. This data can inform the development of more targeted instructional strategies, such as revising course materials, creating supplementary resources, or adjusting the pacing of lessons to better align with student needs. Furthermore, data from chatbot interactions can help instructors design more effective interventions by identifying which Socratic prompts or scaffolding techniques are most successful at fostering deeper understanding and SRL. Ultimately, this feedback loop enables a more evidence-based approach to teaching, allowing educators to refine their practices and contribute to a more nuanced understanding of how students learn in different contexts.

Future research could build on our study by incorporating the coding of chatbot responses to gain deeper insights into the interplay between GenAI and learner behaviour. While our current work focuses on the learner's actions, analyzing the chatbot's responses could provide valuable information about how specific types of AI feedback stimulate particular SRL actions and processes. This approach could help identify the most effective chatbot response strategies for promoting self-regulation and guiding learners through the processes and actions of their tasks. Such findings would be instrumental in developing more sophisticated AI systems tailored to support SRL by dynamically adapting to learner needs. Recognizing the potential of this avenue, we have included this as part of our planned future research to further advance the understanding and efficacy of GenAI in educational settings.

## 5. Conclusion

In conclusion, this study set out to investigate two questions regarding using a Socratic chatbot in a blended learning environment for a pilot class on introductory statistics. The first question asked how researchers can process student conversations with educational chatbots, and the second examined the interaction behaviours among students of different ability levels and learning gains, using the data collected from the first research question to drive learning analytics.

The first research question is answered by using our novel method of annotating interactions within the SRL framework. While  $\alpha = 0.693$ , with 95% confidence interval of (0.619, 0.973), indicates moderate reliability, the coding process involved three annotators and a chief annotator, and conclusions drawn have limited generalizability, possibly requiring further validation in the context of the investigation. Regardless, this value of  $\alpha$  is acceptable for this exploratory study. This work lays the framework for studying how students interacting with GenAI can drive learning analytics. Potential extension of this work could include alternative annotation agreement frameworks and improved reliability with improved methodology, such as having more annotators.

To address the second research question, we analyzed process-action codes and observed that students predominantly engaged in actions such as reviewing, searching, and self-evaluating. These findings indicate that the chatbot successfully facilitated an interactive learning environment where students responded actively, asked questions, and reflected on their understanding, promoting deeper cognitive processing. Using ENA, we compared two groups of students based on their baseline ability and observed distinct interaction patterns. Higher-scoring students demonstrated a greater co-occurrence frequency between actions within the process of engaging with the material, whereas lower-scoring students focused more on searching for answers with reflective activities. It is important to remember that groups A1 and A2 were separated based on their pre-course test scores, not on whether they met specific learning outcomes, which can be achieved only through assessment. The observed behavioural distinctions, more cautious for lower-ability students and exploratory for higher-ability students, highlight the need to customize chatbot prompts. Such tailoring can scaffold lower-ability students toward engaging in more complex, reflective interactions while encouraging higher-ability students to perform critical SRL actions. Furthermore, our analysis reveals that students with large learning gains and those with high baseline knowledge but smaller gains use the same effective learning method of engaging in the Socratic method. The primary difference lies in their pre-course scores, which limit the latter's measurable improvement. We further discovered that for students to improve their learning gains, they should shift from structured problem-solving, such as solving classroom questions, to fundamental questioning and soliciting more examples.

The Socratic chatbot can facilitate this by prompting foundational questions that encourage critical thinking and deep exploration of concepts. By leveraging learning analytics, the chatbot can track user progress, analyze patterns in responses, and identify areas where the user may need additional support or challenge. This data-driven insight allows the chatbot to dynamically adapt its Socratic questioning style, tailoring follow-up questions to the user's specific needs and guiding them through structured, scaffolded learning pathways. These pathways start with fundamental definitions and gradually lead to more complex applications, fostering a personalized and reflective learning experience. Through this combination of Socratic dialogue and analytics, the chatbot not only stimulates intellectual curiosity but also ensures that learning is both targeted and effective.

Overall, this research demonstrates the potential of a Socratic chatbot to tailor educational support to diverse student needs, fostering deeper engagement and improved learning outcomes and revealing how the engagements could be improved through the lens of SRL. Future research should focus on refining these methodologies and expanding their application to diverse data types, like temporal data and longitudinal studies, on different chatbots, learning contexts, and subject areas. For example, because the Socratic chatbot was designed to prompt students with questions specific to the learning content, it does not guide students explicitly toward SRL process-actions. Future research into chatbots that guide students toward these processes and actions would build on the work in this paper. Furthermore, the role of chatbots in collaborative learning can be examined through the lens of social and opinion dynamics, which influence learning processes and may lead to counterintuitive phenomena (Lim et al., 2023; Lai, 2024b; Lai & Cheong, 2024). By continually adapting and enhancing educational technologies, we can better support students' learning journeys and contribute to more effective and personalized education while challenging existing educational theories in light of the adoption of GenAI in education.

## Declaration of Conflicting Interest

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## Funding

The authors declared no financial support for the research, authorship, and/or publication of this article.

## References

- Artstein, R. (2017). Inter-annotator agreement. In N. Ide & J. Pustejovsky (Eds.), *Handbook of linguistic annotation* (pp. 297–313). Springer Netherlands. [https://doi.org/10.1007/978-94-024-0881-2\\_11](https://doi.org/10.1007/978-94-024-0881-2_11)
- Artstein, R., & Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4), 555–596. <https://doi.org/10.1162/coli.07-034-r2>
- Azevedo, R., Guthrie, J. T., & Seibert, D. (2004). The role of self-regulated learning in fostering students' conceptual understanding of complex systems with hypermedia. *Journal of Educational Computing Research*, 30(1–2), 87–111. <https://doi.org/10.2190/dvwx-gm1t-6thq-5wc7>
- Azevedo, R., Moos, D. C., Greene, J. A., Winters, F. I., & Cromley, J. G. (2007). Why is externally-facilitated regulated learning more effective than self-regulated learning with hypermedia? *Educational Technology Research and Development*, 56(1), 45–72. <https://doi.org/10.1007/s11423-007-9067-0>
- Bakhtiar, A., & Hadwin, A. (2020). Dynamic interplay between modes of regulation during motivationally challenging episodes in collaboration. *Frontline Learning Research*, 8(2), 1–34. <https://doi.org/10.14786/flr.v8i2.561>
- Bandura, A. (2002). Social foundations of thought and action. In D. F. Marks (Ed.), *The health psychology reader* (pp. 94–106). SAGE Publications Ltd. <https://doi.org/10.4135/9781446221129.n6>
- Biwer, F., Wiradhany, W., oude Egbrink, M., Hospers, H., Wasenitz, S., Jansen, W., & de Bruin, A. (2021). Changes and adaptations: How university students self-regulate their online learning during the COVID-19 pandemic. *Frontiers in Psychology*, 12, 642593. <https://doi.org/10.3389/fpsyg.2021.642593>
- Bowman, D., Swiecki, Z., Cai, Z., Wang, Y., Eagan, B., Linderoth, J., & Shaffer, D. W. (2021). The mathematical foundations of epistemic network analysis. In A. R. Ruis & S. B. Lee (Eds.), *Advances in quantitative ethnography* (pp. 91–105). Springer International Publishing. [https://doi.org/10.1007/978-3-030-67788-6\\_7](https://doi.org/10.1007/978-3-030-67788-6_7)
- Brenner, C. A. (2022). Self-regulated learning, self-determination theory and teacher candidates' development of competency-based teaching practices. *Smart Learning Environments*, 9(3). <https://doi.org/10.1186/s40561-021-00184-5>
- Bushuyev, S., Bushuyeva, N., Murzabekova, S., & Khussainova, M. (2023). Innovative development of educational systems in the BANI environment. *Scientific Journal of Astana IT University*, 14(14), 104–115. <https://doi.org/10.37943/14ynsz2227>
- Cheong, K. H., Lai, J. W., Yap, J. H., Cheong, G. S. W., Budiman, S. V., Ortiz, O., Mishra, A., & Yeo, D. J. (2023). Utilizing Google Cardboard virtual reality for visualization in multivariable calculus. *IEEE Access*, 11, 75398–75406. <https://doi.org/10.1109/access.2023.3281753>
- Chihab, L., El Mhouthi, A., Massar, M., & Hamdane, K. (2023). Learning analytics and big data: Huge potential to improve online education. In S. Motahhir & B. Bossoufi (Eds.), *Digital technologies and applications. ICDTA 2023. Lecture notes in networks and systems* (pp. 405–411, Vol. 668). Springer Nature Switzerland. [https://doi.org/10.1007/978-3-031-29857-8\\_41](https://doi.org/10.1007/978-3-031-29857-8_41)
- Chocarro, R., Cortiñas, M., & Marcos-Matás, G. (2021). Teachers' attitudes towards chatbots in education: A technology acceptance model approach considering the effect of social language, bot proactiveness, and users' characteristics. *Educational Studies*, 49(2), 295–313. <https://doi.org/10.1080/03055698.2020.1850426>
- Chung, M.-K. (2000). The development of self-regulated learning. *Asia Pacific Education Review*, 1(1), 55–66. <https://doi.org/10.1007/bf03026146>
- Clemons, M. L., & Hopkins, T. (2020). Facilitating success: Using self-regulated learning and servant leadership in the college classroom. *Journal of Teaching and Learning with Technology*, 9(1), 127–132. <https://doi.org/10.14434/jotlt.v9i1.29176>
- Daniela, P. (2015). The relationship between self-regulation, motivation and performance at secondary school students. *Procedia—Social and Behavioral Sciences*, 191, 2549–2553. <https://doi.org/10.1016/j.sbspro.2015.04.410>
- Dilmegani, C. (2023). Top 6 use cases of generative AI in education [Accessed: 13 June 2023]. <https://research.aimultiple.com/generative-ai-in-education/>
- Dunlosky, J., & Ariel, R. (2011). Chapter four—self-regulated learning and the allocation of study time. In B. H. Ross (Ed.), *Advances in research and theory* (pp. 103–140, Vol. 54). Academic Press. <https://doi.org/10.1016/B978-0-12-385527-5.00004-8>
- Edisherashvili, N., Saks, K., Pedaste, M., & Leijen, Ä. (2022). Supporting self-regulated learning in distance learning contexts at higher education level: Systematic literature review. *Frontiers in Psychology*, 12, 792422. <https://doi.org/10.3389/fpsyg.2021.792422>
- Fan, Y., Tang, L., Le, H., Shen, K., Tan, S., Zhao, Y., Shen, Y., Li, X., & Gašević, D. (2025). Beware of metacognitive laziness: Effects of generative artificial intelligence on learning motivation, processes, and performance. *British Journal of Educational Technology*, 56(2), 489–530. <https://doi.org/10.1111/bjet.13544>

- Ferguson, R. (2012). Learning analytics: Drivers, developments and challenges. *International Journal of Technology Enhanced Learning*, 4(5/6), 304–317. <https://doi.org/10.1504/ijtel.2012.051816>
- Greene, J. A., & Azevedo, R. (2009). A macro-level analysis of SRL processes and their relations to the acquisition of a sophisticated mental model of a complex system. *Contemporary Educational Psychology*, 34(1), 18–29. <https://doi.org/10.1016/j.cedpsych.2008.05.006>
- Hartley, K., Hayak, M., & Ko, U. H. (2024). Artificial intelligence supporting independent student learning: An evaluative case study of ChatGPT and learning to code. *Education Sciences*, 14(2), 120. <https://doi.org/10.3390/educsci14020120>
- Järvelä, S., & Bannert, M. (2021). Temporal and adaptive processes of regulated learning—What can multimodal data tell? *Learning and Instruction*, 72, 101268. <https://doi.org/10.1016/j.learninstruc.2019.101268>
- Krippendorff, K. (1995). On the reliability of unitizing continuous data. *Sociological Methodology*, 25, 47–76. <https://doi.org/10.2307/271061>
- Kuhail, M. A., Alturki, N., Alramlawi, S., & Alhejori, K. (2022). Interacting with educational chatbots: A systematic review. *Education and Information Technologies*, 28(1), 973–1018. <https://doi.org/10.1007/s10639-022-11177-3>
- Lai, J. W. (2024a). Adapting self-regulated learning in an age of generative artificial intelligence chatbots. *Future Internet*, 16(6), 218. <https://doi.org/10.3390/fi16060218>
- Lai, J. W. (2024b). Dynamic opinion formation in networks: A multi-issue and evidence-based approach. *Computers*, 13(8), 190. <https://doi.org/10.3390/computers13080190>
- Lai, J. W., & Cheong, K. H. (2024). Winning with losses: The surprising success of negative strategies in social interaction behavior. *Physical Review Letters*, 133(16), 167401. <https://doi.org/10.1103/physrevlett.133.167401>
- Li, Y., Sha, L., Yan, L., Lin, J., Raković, M., Galbraith, K., Lyons, K., Gašević, D., & Chen, G. (2023). Can large language models write reflectively. *Computers and Education: Artificial Intelligence*, 4, 100140. <https://doi.org/10.1016/j.caeai.2023.100140>
- Lim, Z. Y., Munshi, T., Yap, J. H., Mishra, A., Lai, J. W., Yeo, D. J., & Cheong, K. H. (2023). Using opinion dynamics to identify groups for targeted intervention in lifelong learning: A case study of SkillsFuture in Singapore. *International Journal of Modern Physics B*, 38(20), 2450257. <https://doi.org/10.1142/s0217979224502576>
- Long, P., & Siemens, G. (2011). Penetrating the fog: Analytics in learning and education. *EDUCAUSE Review*, 46(5), 31–40. <https://er.educause.edu/-/media/files/article-downloads/erm1151.pdf>
- Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1), 50–60. <https://doi.org/10.1214/aoms/1177730491>
- Meng, J., & Dai, Y. (2021). Emotional support from AI chatbots: Should a supportive partner self-disclose or not? *Journal of Computer-Mediated Communication*, 26(4), 207–222. <https://doi.org/10.1093/jcmc/zmab005>
- Molenaar, I. (2021). Personalisation of learning: Towards hybrid human-AI learning technologies. In A. Schleicher (Ed.), *OECD digital education outlook 2021*. OECD. <https://doi.org/10.1787/2cc25e37-en>
- Okonkwo, C. W., & Ade-Ibijola, A. (2021). Chatbots applications in education: A systematic review. *Computers and Education: Artificial Intelligence*, 2, 100033. <https://doi.org/10.1016/j.caeai.2021.100033>
- Panadero, E. (2017). A review of self-regulated learning: Six models and four directions for research. *Frontiers in Psychology*, 8, 422. <https://doi.org/10.3389/fpsyg.2017.00422>
- Panadero, E., Klug, J., & Järvelä, S. (2015). Third wave of measurement in the self-regulated learning field: When measurement and intervention come hand in hand. *Scandinavian Journal of Educational Research*, 60(6), 723–735. <https://doi.org/10.1080/00313831.2015.1066436>
- Pintrich, P. R. (2000). Chapter 14—the role of goal orientation in self-regulated learning. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 451–502). Academic Press. <https://doi.org/10.1016/b978-012109890-2/50043-3>
- Qiu, W., Khong, A. W. H., Supraja, S., & Tang, W. (2024). A dual-mode grade prediction architecture for identifying at-risk students. *IEEE Transactions on Learning Technologies*, 17, 803–814. <https://doi.org/10.1109/tlt.2023.3333029>
- Qiu, W., Su, C. L., Jamil, N. B., Ng, S. S., Chen, C.-M., & Lim, F. S. (2024). “I am here to guide you”: A detailed examination of late 2023 Gen-AI tutors capabilities in stepwise tutoring in an undergraduate statistics course. In *Proceedings of the 18th International Technology, Education and Development Conference (INTED2024)*, 4–6 March 2024, Valencia, Spain (pp. 3761–3770). IATED. <https://doi.org/10.21125/inted.2024.0984>
- Shaffer, D. W., Hatfield, D., Svarovsky, G. N., Nash, P., Nulty, A., Bagley, E., Frank, K., Rupp, A. A., & Mislevy, R. (2009). Epistemic network analysis: A prototype for 21st-century assessment of learning. *International Journal of Learning and Media*, 1(2), 33–53. <https://doi.org/10.1162/ijlm.2009.0013>
- She, C., Liang, Q., Jiang, W., & Xing, Q. (2023). Learning adaptability facilitates self-regulated learning at school: The chain mediating roles of academic motivation and self-management. *Frontiers in Psychology*, 14, 1162072. <https://doi.org/10.3389/fpsyg.2023.1162072>

- Siebert-Evenstone, A. L., Arastoopour Irgens, G., Collier, W., Swiecki, Z., Ruis, A. R., & Williamson Shaffer, D. (2017). In search of conversational grain size: Modeling semantic structure using moving stanza windows. *Journal of Learning Analytics*, 4(3), 123–139. <https://doi.org/10.18608/jla.2017.43.7>
- Siemens, G. (2013). Learning analytics: The emergence of a discipline. *American Behavioral Scientist*, 57(10), 1380–1400. <https://doi.org/10.1177/0002764213498851>
- Stoeger, H., & Ziegler, A. (2010). Do pupils with differing cognitive abilities benefit similarly from a self-regulated learning training program? *Gifted Education International*, 26(1), 110–123. <https://doi.org/10.1177/026142941002600113>
- Tan, Y., Ruis, A., Marquart, C., Cai, Z., Knowles, M., & Shaffer, D. (2023). Ordered network analysis. In C. Damşa & A. Barany (Eds.), *Advances in quantitative ethnography. ICQE 2022. Communications in computer and information science* (pp. 108–120, Vol. 1785). Springer. [https://doi.org/10.1007/978-3-031-31726-2\\_8](https://doi.org/10.1007/978-3-031-31726-2_8)
- Tegos, S., Demetriadis, S., Papadopoulos, P. M., & Weinberger, A. (2016). Conversational agents for academically productive talk: A comparison of directed and undirected agent interventions. *International Journal of Computer-Supported Collaborative Learning*, 11(4), 417–440. <https://doi.org/10.1007/s11412-016-9246-2>
- Tsai, Y.-S., Perrotta, C., & Gašević, D. (2019). Empowering learners with personalised learning approaches? Agency, equity and transparency in the context of learning analytics. *Assessment & Evaluation in Higher Education*, 45(4), 554–567. <https://doi.org/10.1080/02602938.2019.1676396>
- Williamson, G. (2015). Self-regulated learning: An overview of metacognition, motivation and behaviour. *Journal of Initial Teacher Inquiry*, 1, 25–27. <https://doi.org/10.26021/851>
- Winne, P. H. (2011). A cognitive and metacognitive analysis of self-regulated learning. In B. J. Zimmerman & D. H. Schunk (Eds.), *Handbook of self-regulation of learning and performance* (pp. 15–32). Routledge. <https://doi.org/10.4324/9780203839010>
- Winne, P. H., & Hadwin, A. F. (1998). Studying as self-regulated learning. In D. Hacker, J. Dunlosky, & A. Graesser (Eds.), *Metacognition in educational theory and practice* (pp. 277–304). Taylor & Francis. <https://psycnet.apa.org/record/1998-07283-011>
- Wisconsin Center for Education Research. (2024). ENA [Accessed: 13 April 2024]. <https://www.epistemicnetwork.org/>
- Wollny, S., Schneider, J., Mitri, D. D., Weidlich, J., Rittberger, M., & Drachler, H. (2021). Are we there yet? A systematic literature review on chatbots in education. *Frontiers in Artificial Intelligence*, 4, 654924. <https://doi.org/10.3389/frai.2021.654924>
- Wright, D. B. (2020). Gain scores, ANCOVA, and propensity matching procedures for evaluating treatments in education. *Open Education Studies*, 2(1), 45–65. <https://doi.org/10.1515/edu-2020-0107>
- Zhai, C., Wibowo, S., & Li, L. D. (2024). The effects of over-reliance on AI dialogue systems on students' cognitive abilities: A systematic review. *Smart Learning Environments*, 11(1). <https://doi.org/10.1186/s40561-024-00316-7>
- Zhang, Y., Tian, Y., Yao, L., Duan, C., Sun, X., & Niu, G. (2022). Individual differences matter in the effect of teaching presence on perceived learning: From the social cognitive perspective of self-regulated learning. *Computers & Education*, 179, 104427. <https://doi.org/10.1016/j.compedu.2021.104427>
- Zimmerman, B. J. (1989). A social cognitive view of self-regulated academic learning. *Journal of Educational Psychology*, 81(3), 329–339. <https://doi.org/10.1037/0022-0663.81.3.329>
- Zimmerman, B. J. (2008). Investigating self-regulation and motivation: Historical background, methodological developments, and future prospects. *American Educational Research Journal*, 45(1), 166–183. <https://doi.org/10.3102/0002831207312909>
- Zimmerman, B. J. (2015). Self-regulated learning: Theories, measures, and outcomes. In J. D. Wright (Ed.), *International encyclopedia of the social & behavioral sciences* (2nd edition, pp. 541–546). Elsevier. <https://doi.org/10.1016/B978-0-08-097086-8.26060-1>
- Zimmerman, B. J., & Campillo, M. (2003). Motivating self-regulated problem solvers. In J. E. Davidson & R. J. Sternberg (Eds.), *The psychology of problem solving* (pp. 233–262). Cambridge University Press. <https://doi.org/10.1017/CBO9780511615771.009>
- Zimmerman, B. J., & Pons, M. M. (1986). Development of a structured interview for assessing student use of self-regulated learning strategies. *American Educational Research Journal*, 23(4), 614–628. <https://doi.org/10.3102/00028312023004614>
- Zimmerman, B. J., & Schunk, D. H. (Eds.). (2011). *Handbook of self-regulation of learning and performance*. Routledge. <https://doi.org/10.4324/9780203839010>