Metacognition and Learning (2022) 17:949–987
https://doi.org/10.1007/s11409-022-09291-1

# Towards investigating the validity of measurement of self-regulated learning based on trace data

**Yizhou Fan[1] · Joep van der Graaf[2] · Lyn Lim[3] · Mladen Raković[4] · Shaveen Singh[4] · Jonathan Kilgour[1] · Johanna Moore[1] · Inge Molenaar[2] · Maria Bannert[3] · Dragan Gašević[1,4]**

Received: 28 May 2021 / Accepted: 21 January 2022 /Published online: 4 May 2022


## Abstract

Contemporary research that looks at self-regulated learning (SRL) as processes of learning events derived from trace data has attracted increasing interest over the past decade. However, limited research has been conducted that looks into the validity of trace-based measurement protocols. In order to fill this gap in the literature, we propose a novel validation approach that combines **theory-driven** and **data-driven** perspectives to increase the validity of interpretations of SRL processes extracted from trace-data. The main contribution of this approach consists of three alignments between trace data and think aloud data to improve measurement validity. In addition, we define the match rate between SRL processes extracted from trace data and think aloud as a quantitative indicator together with other three indicators (sensitivity, specificity and trace coverage), to evaluate the "degree" of validity. We tested this validation approach in a laboratory study that involved 44 learners who learned individually about the topic of artificial intelligence in education with the use of a technology-enhanced learning environment for 45 minutes. Following this new validation approach, we achieved an improved match rate between SRL processes extracted from trace-data and think aloud data (training set: 54.24%; testing set: 55.09%) compared to the match rate before applying the validation approach (training set: 38.97%; test set: 34.54%). By considering think aloud data as "reference point", this improvement of the match rate quantified the extent to which validity can be improved by using our validation approach. In conclusion, the novel validation approach presented in this study used both empirical evidence from think aloud data and rationale from our theoretical framework of SRL, which now, allows testing and improvement of the validity of trace-based SRL measurements.

**Keywords** Self-regulated learning · Measurement validity · Trace data · Think aloud data

✉ Yizhou Fan
yizhou.fan@ed.ac.uk

Extended author information available on the last page of the article.

## Introduction

Self-regulated learning (SRL) skills are considered critical to ensure productive lifelong learning (Klug et al., 2011) and success in different learning tasks (Maldonado-Mahauad et al., 2018b; Bannert & Reimann, 2012). To date, researchers have proposed several theoretical models to describe SRL processes (e.g., Winne and Hadwin (1998), Pintrich (2004), and Zimmerman (2000)). In these models, learners are commonly considered as agents (Haggard & Tsakiris, 2009) who decide which learning processes they should engage in to accomplish their goals for learning (Dunlosky & Thiede, 2013).

Moreover, these models broadly agree that SRL is a cyclic process that unfolds over the three major phases: preparation, enactment and appraisal. For example, the model proposed by Zimmerman (2000) posits that self-regulation includes learners' "self-generated thoughts, feelings, and actions that are planned and cyclically adapted to the attainment of personal goals" and, according to this model, each learning cycle consists of forethought, performance and self-reflection phase. Pintrich (2000) proposed a similar cyclical structure of SRL processes including four basic phases (1) forethought, planning and activation; (2) monitoring; (3) control; and (4) reaction and reflection. The Winne and Hadwin (1998) COPES model describes SRL as a dynamic, loosely sequential and cyclical set of skills where learning processes unfold over four general stages: understanding task requirements, developing plans and setting learning goals, enacting learning tactics to accomplish goals for the task, and adapting learning approaches to future learning tasks. These four phases encompass different cognitive and metacognitive processes, including orientation, monitoring and evaluation. Even though different theorists have contributed to the construction of SRL models differently, the consensus is that in order to understand SRL, it is essential that researchers validly measure SRL processes described in these theories.

Methods commonly used to measure SRL processes to date include self-report surveys (Pintrich & et al. 1991), think aloud protocols (Bannert, 2007; Azevedo et al., 2005; Greene et al., 2008, 2009), trace-based measurement (Kinnebrew et al., 2014; Siadaty et al., 2016c; Saint et al., 2020b; Fan et al., 2021), and more recently the use of new data channels such as eye-tracking (Taub et al., 2017; Mudrick et al., 2019; Fan et al., 2020). In the case of self-report surveys, learners are usually provided with the questionnaires in the Likert scale form (e.g., Motivated Strategies for Learning Questionnaire (MSLQ) Pintrich & et al. 1991) and asked to rate their use of different learning processes, e.g., metacognitive monitoring or elaboration. However, since the introduction of self-report surveys in SRL research, researchers have shown that self-reports are not a suitable approach to reveal actual SRL processes (e.g., Veenman, 2007). Hence, think aloud methods have been introduced to improve the measurement of SRL. Per this approach, a learner verbalises their thoughts during the task and, as a result, researchers collect rich verbal protocols. To unveil SRL processes from these protocols, several coding schemes have been developed based on different theoretical frameworks and have been applied in different learning context, e.g., (Bannert, 2007; Molenaar et al., 2011; Azevedo et al., 2005; Greene et al., 2008, 2009). For example, Bannert's (2007) scheme operationalises (1) cognitive learning processes (e.g. read, encode, elaborate) that learners enact to accomplish a learning task and (2) metacognitive processes (e.g. orientate, plan, monitor and evaluate) that learners enact to monitor and control learning and motivation, as per (Winne & Handwin 1998, 2007); Greene & Azevedo's (2009) scheme operationalises five macro-level SRL processes such as planning,

monitoring and strategy use, which are further subdivided into 35 micro-level SRL processes such as judgement of learning, self-questioning and content evaluation, as per (Azevedo et al., 2004, 2005; Greene et al., 2008). Researchers have applied the think aloud coding schemes in a group of studies, e.g., to distinguish between macro- and micro-level SRL processes (Azevedo et al., 2005, Greene et al., 2008, 2009) and to examine the relationships between learners' SRL processes and their mental models in a complex learning system (Greene & Azevedo, 2009). The literature documents that think aloud data can provide a more valid account of observed SRL processes than the data collected via self-report surveys, and also that think aloud data are a stronger predictor of learning achievement than the self-report data (Bannert, 2007, 2013; Veenman, 2007; Greene & Azevedo, 2010).

As another promising approach to measure SRL, researchers have recognised trace-based methods that rely upon learner trace data captured in technology-enhanced learning environments (TELs) (Winne, 2010; Siadaty et al., 2016c). Trace data, for instance, can unobtrusively record instances of cognition and metacognition in authentic learning environments, and thus operationalize "what learners do as they do it" (Winne, 2010, p. 275). Importantly, by using dynamic trace-based measurements, researchers can avoid methodological shortcomings of retrospective measurement approaches where learners, e.g., self-report on what they believe they did or provide distorted picture of cognitive and metacognitive processes they engaged in (Winne, 2010, p. 275). Trace data that learners generate are thus temporally proximal to decisions learners make during a study session and generally reflect learning events more completely and with less bias than self-report surveys (Gasevic et al., 2017).

Several researchers have proposed and utilised trace-based protocols to measure SRL processes (Siadaty et al., 2016a, b; Saint et al., 2020a, b, 2021; Fan et al., 2021). In these studies, researchers investigated SRL processes as patterns or sequences of events. For example, a learner's transition from accessing a practice assignment to accessing a course content is labelled as "reflection", one of the key SRL processes (Saint et al., 2020a). As suggested by Winne (2014), these trace-based methods rest on a common assumption that patterns of the observed events complement self-reports and provide a deeper insight into how SRL is constituted (Winne, 2014, p. 234). Even though the use of trace-based measuring approaches to study SRL in real-time is becoming more popular among educational researchers, it is critical to ensure that the researchers' interpretations grounded in trace data are valid (Winne, 2020).

Validity has been commonly recognised as a central construct in educational measurement. Samuel Messick defined validity as the "integrated evaluative judgement of the degree to which empirical evidence and theoretical rationale supports the adequacy and appropriateness of inferences and actions based on test scores" (Messick, 1987, p. 1). We reviewed the previous trace-based SRL research via lenses of this definition and found that even though several studies have proposed protocols and analytical methods for measuring SRL based on trace data (Siadaty et al., 2016a, b; Saint et al., 2020a), the validity of trace-based measurements and interpretations of the corresponding results have rarely been examined or *empirically documented* which, consequently, offers no grounding for valid inferences Winne (2020). Moreover, this limitation may fundamentally influence research studies utilising trace-based approaches. Therefore, in this paper, we propose a validation approach that allows for investigating the validity of trace-based SRL measurements.

## Background

In this section, we first operationally define the concept of validity examined in the present study. Next, we review prior research that investigated SRL using trace data. In the reviewed studies, we identified the challenges with the validity of SRL measurement, including the issues related to examining SRL as events. After that, we review the approaches that researchers have previously used to improve the validity of trace-based SRL measurement. Finally, we identify existing gaps in the literature and pose research questions to guide the present study.

### The validity concept

Traditionally, there are three commonly considered types of validity in the literature: **content validity**, **criterion-related validity**, and **construct validity** (Messick, 1987). According to Messick 1987, content validity is determined by domain experts who judge whether the content of a test is representative of a behaviour that the test attempts to measure. Criterion-related validity, on the other hand, is determined as an empirical relationship between the test scores and predefined benchmark scores for a behaviour (Messick, 1987). Last, to determine construct validity, researchers bring evidence to support the interpretation of test scores (Messick, 1987, p. 10). Messick pointed out "since content- and criterion-related evidence contribute to score meaning or interpretation, they have come to be recognised as aspects of construct validity" (Messick, 1987, p. 16). Following this notion, we studied construct validity in the present study, i.e., the terms validity and construct validity were used interchangeably in this study. Specifically, we relied upon Winne's (Winne & Perry, 2000) definition of construct validity operationalised in the context of SRL measurement: construct validity of instruments or protocols represents a set of concerns about whether the measurement methods, as they are operationally defined, represent the SRL processes researchers intend to measure and not other phenomena.

### Validity challenges in researching SRL as events

SRL is characterised as an aptitude and an event (Winne & Perry, 2000). An aptitude describes "a relatively enduring attribute of a person that predicts future behaviour" (Winne & Perry, 2000, p. 534). For example, it might be expected that a learner will behave differently when studying for a summative (e.g., exam) than when studying for a formative assessment (e.g., post-lecture quiz) if the learner reported that they adapt their learning tactics to the circumstances of assignments. More importantly, this prediction is considered valid regardless whether this question was asked a week or a year before the task, and regardless of the educational context the assignment was administered in (e.g., course subject) (Winne & Perry, 2000). In contrast, events are defined as the "very actions learners perform, rather than descriptions of those actions or of mental states that actions generate" (Winne, 2010, p. 269). The notion of SRL as an event entertains the analysis of temporal sequences and patterns [1] as a new approach to measuring SRL. The dynamic SRL processes represented as events thus can be identified by using data mining techniques or by stipulation based on hypothesis (Winne & Perry, 2000, 2014; Bannert et al., 2014). This marks an

[1] in this study, we define action patterns which can be interpreted as **SRL processes** according to theoretical models of SRL

important shift in the theoretical views on SRL, with possible new implications for measurement. However, researchers have documented important challenges related to measuring SRL processes as events (Winne & Perry, 2000), e.g., mapping raw trace data to theoretically meaningful SRL processes. New methods and protocols still need to be developed to harness the research potential of trace data and more comprehensively characterise learners' SRL processes, including learning tactics and strategies that learners enact to accomplish their learning goals.

Several previous studies have adopted **data-driven approaches** to discover latent SRL processes and to track these processes as they unfold over time (Boroujeni & Dillenbourg, 2019). In these studies, researchers mapped raw trace data to learning actions or events, and then applied data analytic techniques (e.g., process mining) to identify SRL processes from those actions and events. This approach has been shown useful in identifying frequent patterns of learning actions and events, and interpreting these patterns as SRL processes and learning strategies. For example, Maldonado-Mahauad et al. (2018a) identified six distinct patterns of actions that learners enact in a massive open online course. The authors then matched these action patterns to theory-informed SRL processes or learning strategies, and identified three clusters of learners relative to their use of SRL processes and strategies. For instance, the authors interpreted the pattern of actions "Complete video lecture -> Attempt assessment" as the *Evaluation* strategy. However, there are still many patterns generated in trace data that cannot be associated to SRL processes whereas, on the other hand, some of the patterns can be associated to multiple SRL processes. For example, the "highlight text during reading" pattern is considered indicative of the *Metacognitive monitoring* process Winne (2019). However, this pattern can also be interpreted as a cognitive process of *Comprehending reading materials*. Therefore, the validity of the interpretation of SRL processes unveiled using **data-driven approaches** can often be questioned. Equally importantly, researchers using trace-based data-driven approaches have typically been able to discover a limited number of high-frequency patterns from trace data (e.g., six patterns reported in Maldonado-Mahauad et al., 2018a), while leaving less frequent patterns unexamined. As a consequence, all the processes theorised in SRL have not been comprehensively measured as of yet. To remedy this challenge, we posit theory-driven approach is needed, i.e., researchers may begin their analytical procedures by looking at SRL theoretical assumptions, and then use these assumptions to guide systematic identification of SRL processes from trace data.

Specifically, in a **theory-driven approach**, a pattern of fine-grained actions can be operationally defined and mapped to SRL processes based on a theoretical model of SRL (Siadaty et al., 2016c). To date, several trace-based SRL measurement protocols have been proposed and applied to identify theoretically supported SRL processes from trace data (Siadaty et al., 2016b, c; Saint et al., 2020a; Fan et al., 2020). For example, the protocol proposed in (Fan et al., 2020) included 27 different cognitive and metacognitive SRL processes, e.g., "opening the planner tool to review learning goals and plans" is considered *Planning*, a metacognitive SRL process (Winne & Hadwin, 1998). Even though theory-driven trace-based measurement protocols can provide a valuable means to measuring SRL processes at-scale, it is often challenging to unambiguously infer SRL processes from those protocols. In other words, there are challenges related to construct validity of trace-based measurement protocols. For example, it may be unclear whether the aforementioned pattern "opening the planner tool to review learning goals and plans" should be interpreted as *Planning* or *Orientation* process? Or, should this pattern be considered *Monitoring* if it appears

later in a learning session? In this study, we make a step forward towards addressing such a construct validity challenge.

### Approaches to improving construct validity

Almost any information concerning a particular measurement process can contribute to the understanding of the construct validity of that measurement (Messick, 1987). However, this contribution becomes stronger if the information is evaluated using the **theoretical rationale** (Messick, 1987). In the context of event-based SRL measurement, limited research has been conducted to generate theory-informed evidence of SRL processes and thus improve the construct validity of SRL measurement. For example, the authors of the two groups of studies, Maldonado-Mahauad et al. (2018a, b) and Siadaty et al. (2016b, c), and Saint et al. (2020b), studied SRL based on Winne and Hadwin's (Winne & Hadwin, 1998) and Zimmerman's (2000) models, respectively, to determine what trace data should be gathered to measure SRL processes and how findings based on these data should be interpreted (Winne, 2020). Saint et al. (2020b) operationally defined SRL processes and then extracted those processes from trace data based on a theoretical model of SRL, e.g., the authors defined the *Reflection* process following Zimmerman's (2000) SRL theoretical model, and then identified this process from the pattern of traced actions "attempt the assessment and then access course content or watch video lecture" (Saint et al., 2020b, p. 7).

Another approach to improving the validity of SRL measurement is related to how learners perceive learning tasks and the learning environment. It is critical that researchers ensure that learners are trained or prepared to self-regulate their learning in a certain context (Winne, 1982, 2010). Several standards for validating a researcher's interpretation of events recorded by trace data should be ensured. For instance, learners should remain alert to whether conditions for cognition or SRL are present in the learning environment, and fully understand which particular cognitive operations they should apply when specific conditions are perceived in learning (Winne, 2010). Learners should also be capable and motivated to carry out the underlying cognitive operations when the context is right (Winne, 2010). For example, if a learner fails to realise that a planner tool is available in the learning environment, then the researcher's interpretation around SRL processes extracted through the analysis of events recorded in trace data about the learner's use of the planner tool may be invalid. Therefore, many studies (Beheshitha et al., 2015; Munshi et al., 2018; Siadaty et al., 2016b; Kinnebrew et al., 2013) emphasise the training of learners, especially to ensure that they are familiar with their learning environment.

Efforts related to theoretical grounding and training of learners are necessary but not sufficient to assure validity. Validity is fundamentally dependent on empirical evidence (Messick, 1987; Winne, 2020). Following this direction, a small number of studies emphasised the validity of research by focusing on SRL processes that reflect a certain degree of difference on learning performance or affective states. In an early study, Aleven et al. (2006) developed a metacognition model to measure learners' different help-seeking processes when they use a cognitive tutor, and then validated this model by examining the correlations between learners' help-seeking processes and post-test learning gain (when controlling for the pre-test) (Aleven et al., 2006). The comparison of learners' help-seeking models against learners' interactions and learning gains provided empirical evidence on (1) the validity of the model which was "generally on track" and (2) the adjustments that were needed for large-scale application (Aleven et al., 2006). In a more recent study, Munshi et al. (2018) focused on SRL processes that triggered major differences in the number of instances of affective states observed in high-performance versus low-performance learner

groups (Munshi et al., 2018). In this way, they were able to improve validity by focusing on specific processes that proved relevant in describing high- versus low-performers. Another approach to gauging validity is to triangulate the measurement between different data channels, which use several methods to collect data about the same event (Winne, 2020). However, triangulation across measurement protocols is very infrequent (Winne & Perry, 2000; Azevedo & Gašević, 2019), and has only recently received attention (Taub et al., 2016; Azevedo & Gašević, 2019; Reimann, 2019). According to our literature review, only a small number of studies such as Azevedo et al. (2009, 2010) used a combination of trace data and think aloud to understand the nature of learners' deployment of SRL processes (Azevedo & Witherspoon, 2009). However, even these studies have not systematically matched all dimensions across different measurement protocols; for example, by examining the SRL processes detected from trace data by aligning them with those detected with think aloud data.

In order to address the above open challenges and provide new empirical evidence about the validity of trace-based SRL measurement, we propose a novel validation approach in the study reported in this paper. As addressed by several previous studies, think aloud protocols represent a primary method for capturing, analysing, and classifying SRL processes (Veenman, 2007; Bannert, 2007; Azevedo et al., 2010). Although think aloud protocols have limitations (Young, 2005), the interpretations based on think aloud codes are still considered a more valid measurement of observed SRL, and can be used as a stronger predictor of learning achievement than the self-report survey data (Bannert, 2007, 2013; Veenman, 2007; Greene & Azevedo, 2010). More importantly, in order to quantitatively evaluate and improve the validity of trace-based SRL measurement, we need to adopt one measurement as the "reference point", and think aloud method is currently the best option to be considered as such. Therefore, we considered think aloud as "reference point" in our validation approach, to provide data-driven evidence about SRL and to validate inferences drawn from trace data regarding SRL processes. We would also like emphasise that think aloud as "reference point" is not an "absolute truth". Therefore, we combined the theoretical rationales based on SRL theory and the empirical evidence based on think aloud to help us improve the validity of trace-based SRL measurement. We further discuss the importance of theory and the limitations of think aloud in the discussion section. On the basis of these considerations, we set the following research questions:

**RQ1** How does a theory-driven trace-based measurement of SRL perform in terms of validity when considering think aloud as "reference point"?

**RQ2** How can the validity of trace-based SRL measurement be improved based on both the theoretical rationale (i.e., theory-driven) and empirical evidence (i.e., data-driven)?

**RQ3** To what extent can the validity of trace-based SRL measurement be improved?

By answering these research questions, we aim to provide novel methodological contributions to the existing body of research on SRL. The systematic validation approach proposed in this paper can support trace-based SRL measurement with both theoretical rationale (theory-driven perspective) and empirical evidence (data-driven perspective). In this study, we used a theoretical framework to map the action patterns to SRL processes which provide the theoretical rationale of measuring SRL from trace data. At the same time, we analysed the alignment between measurement results from trace data and think aloud data as a way to produce empirical evidence that can improve the validity of trace-based SRL measurement. In order to improve the reproducibility of our validation approach, in

the method section, we provide a detailed description of how our approach was designed, how the SRL processes were operationally defined, and how the measurement validity was verified by using alignment of trace data and think aloud data.

## Methods

### Research design and the learning environment

The laboratory study was conducted at a university in the Netherlands and involved 44 learners with an average age of 21 years (SD = 3 years) as participants [2]. Of these learners, 39 were undergraduate students and 5 graduate students from very diverse majors (e.g., psychology and communication science). The study used a pre-post design with a 45-minute learning session during which participants were asked to study three topics: (1) artificial intelligence (the basic concepts of artificial intelligence), (2) differentiation in the classroom (the concept of differentiation explains how teachers can deal with differences between learners, and the idea of adaptive learning), and (3) scaffolding (as an essential way to support learners during learning). The learning task was to integrate the three topics into an essay (300-400 words) that describes learning in school in 2035.

A technology-enhanced learning environment (TEL) developed for this study, contained a catalogue and navigation zone on the left, reading and writing zone in the middle, instrumentation tool zone on the right and other tools such as search and timer (see Fig. 1). A more detailed introduction about these instrumentation tools and how learners use these tools could be found in van der Graaf et al. (2021). Learners could use a) the navigation zone to check general instructions and the rubric for scoring essays and b) the catalogue zone or search tool on the left to navigate through learning materials. Learners could also use the planner tool to allocate time and the timer tool to check the time countdown of 45 minutes. While reading, learners could highlight some keywords or sentences or take notes about their learning, and they could also create labels or search for their highlights and notes. The size of the reading and writing zone could be adjusted, for example, a half-size reading zone and half-size writing zone (as shown in Fig. 1), or full-size reading zone (with the minimised writing window) or full-size writing zone (with the minimised reading window).

The following setup in the lab was used: an internet-capable computer with a keyboard and mouse to collect trace data, and one webcam and one microphone to collect think aloud data. Before the study began, the experimenter introduced the study requirements to the participants, asked the participants to complete the pre-test, familiarised them with the learning environment, and led a training to familiarise participants with the think aloud procedure. In the process of the study, the experimenter ensured the learners continuously kept thinking aloud by providing prompts if there was a long period of silence and they answered procedural questions posed by learners, but they did not interfere with the learner's learning process. After the participants finished the whole learning session, the experimenter asked them to complete the post-test and transfer-test, as learning outcomes. Trace data in this study included navigation logs, keyboard strokes and mouse traces (click and scroll), which were obtained via a local PHP-server. The audio recordings of participants' think aloud were used for coding SRL processes.

[2]There was originally 45 participants. However, we found out that the first language of one participant was not Dutch and we excluded this participant from the dataset

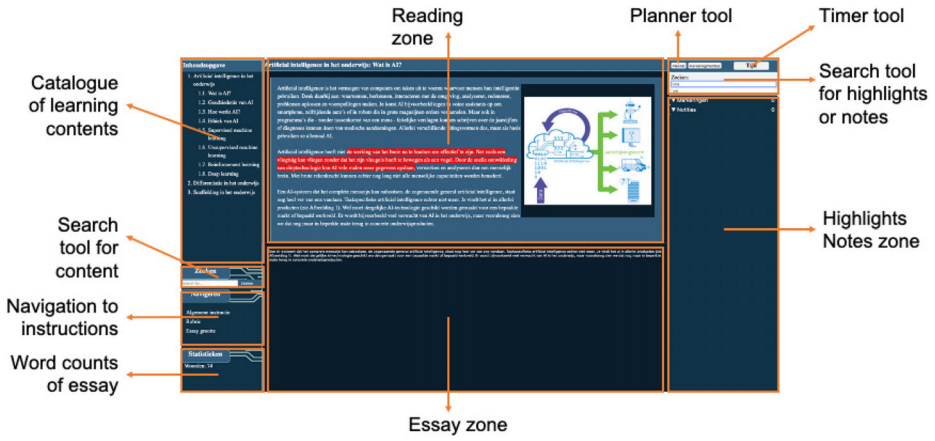


**Fig. 1** Learning environment and different functional zones

## Validation approach

In order to explain the overall validation approach in this study, we created a schematic diagram (see Fig. 2) which contains two main perspectives (which can also be understood as two sub-level approaches): a **theory-driven perspective** to build a **theory-driven process library** (Version 1) and a **data-driven perspective** to obtain a **data-driven process library** (Version 2). Finally, based on these two versions of the process libraries, we constructed an **improved process library** (Version 3), and then evaluated the validity of the improved process library based on the alignment with think aloud data. The proposed validation approach and the terminology in Fig. 2 are explained in detail in the following five steps.

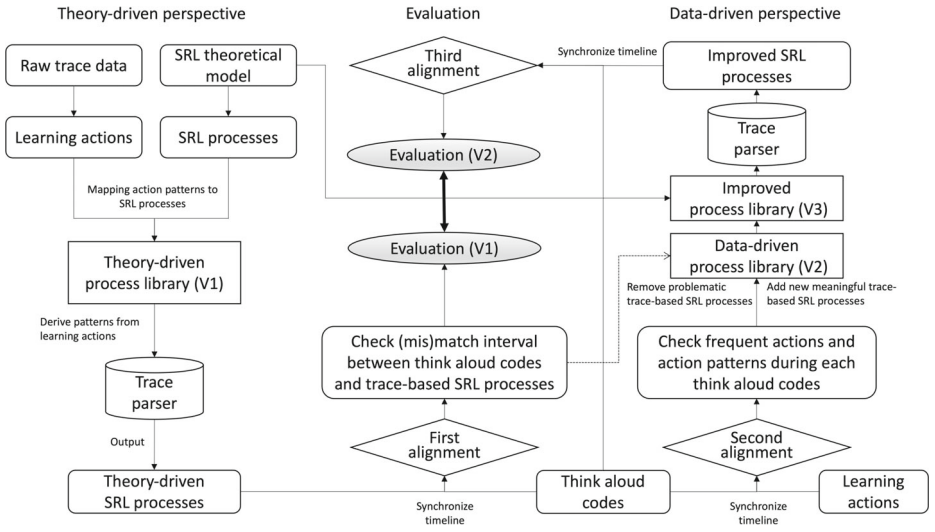


**Fig. 2** Validation approach of trace-based SRL measurement

### Step 1: Theory-driven perspective: theory-driven process library (version 1)

We first followed a **theory-driven perspective**, starting from labelling raw trace data into meaningful **learning actions** (as shown in Fig. 2). In order to achieve this, we built an action library which provides the descriptions for a list of learning actions. For example, learners' keyboard strokes within the essay zone were labelled as the "WRITE_ESSAY" action; and their keyboard strokes within the note tool were labelled as the "NOTE_EDITING" action.

We then interpreted action patterns as **SRL processes** based on our theoretical framework. As shown in Table 1, our theoretical framework contained four main categories and eight subcategories. A collection of these subcategories of SRL processes constituted the

**Table 1** Categories and subcategories of the theoretical framework that were used to code the data collected in this study

| Main categories | Subcategories | Codes | Definitions |
|---|---|---|---|
| Metacognition | Orientation | MC.O | Orientation on the learning-related activities; on prior knowledge; on the task and feeling about the task. Reading of general instructions and rubric. |
| | Planning | MC.P | Planning of the learning process by arranging activities and determining strategies. Proceeding to the next topic. |
| | Monitoring | MC.M | Monitoring and checking the learning process; checking of progress according to instruction or plan. |
| | Evaluation | MC.E | Evaluation of the learning process; checking of content-wise correctness of learning activities. Saying that one owns work is correct. |
| Low_Cognition | First-reading | LC.F | Reading information from the texts and superficial describing of pictorial representations. |
| | Re-reading | LC.R | Rereading of information in the text or figures. |
| High_Cognition | Elaboration/ Organisation | HC.E/O | Elaborate by connecting content related comments and concepts; reasoning and association. Organising of content by creating an overview; write down information point by point; summarising; adding information generated by oneself; and editing information by rephrasing or integrating information with prior knowledge. |
| Other | Motivational/ Procedural | Other | Learner's positive or negative expressions about the task, situation, or the ability; Learner ask researcher whether they can begin working on the task |

**Table 2** The action library used for labelling learning actions based on trace data

| Action Labels | Action descriptions |
|---|---|
| GENERAL_INSTRUCTION | Learners read or re-read general instructions and learning goals |
| RUBRIC | Learners read or re-read the rubric for essay writing |
| RELEVANT_READING | Learners read and learn learning content for the first time |
| RELEVANT_RE-READING | Learners re-read and review for learning content which they have read before |
| IRRELEVANT_READING | Learners read the pages which are not relevant to the learning goal or writing task |
| IRRELEVANT_RE-READING | Learners re-read the pages which are not relevant to the Learning goal or writing task |
| NAVIGATION | Learners navigate through pages or scroll at catalogue zone |
| WRITE_ESSAY | Learners write, edit or stay in the essay zone |
| COPY_PASTE | Learners copy and paste some content from reading materials into the essay or notes |
| NOTE_EDITING | Learners create, delete, edit or label the notes |
| NOTE_READING | Learners click to open and read or re-read the notes |
| HIGHLIGHT_EDITING | Learners create, delete or edit the highlights |
| HIGHLIGHT_READING | Learners click to open and read or re-read the highlights |
| HIGHLIGHT_LABELLING | Learners create tags for highlights |
| TIMER | Learners click to check timer during learning |
| SEARCH_CONTENT | Learners use the search tool on the left to search learning contents |
| SEARCH_HIGHLIGHT_NOTE | Learners use the search tool on the right to search notes or highlights |
| PLANNER | Learners click to open planner tool, and create or edit their plans |
| OFF_TASK | Learners do not have any action for a relatively long time (5 minutes in this study) |

More details about the labelling process of the action library can be found in the Appendix

process library. A detailed description of the theoretical framework, action library (Table 2) and process library along with examples can be found in the Appendix. The Metacognition category included the subcategories: *Orientation*, *Planning*, *Monitoring*, and *Evaluation*; the Cognition in our study was divided into Low_Cognition and High_Cognition and, in this way, distinguished between low-level (*First-reading* and *Re-reading*) and high-level (*Elaboration/Organisation*) cognitive processes; and the Other category included *Motivational* and *Procedural* issues.

Building on the previous literature (Siadaty et al., 2016b, c; Saint et al., 2020a, b; Kizilcec et al., 2017), we conducted multiple rounds of in-depth discussion to construct the **theory-driven process library** that is based on our theoretical framework. Those involved in the discussions included the researcher who developed this theoretical framework and the think aloud coding scheme, the researcher who ran the lab study who also coded think aloud data and who was familiar with the learning task, the designer of the learning environment used in the study, and an experienced researcher who was familiar with the extraction of SRL processes from trace data. The team members discussed whether and to what extent the SRL

processes reflected the categories from our theoretical framework, how the SRL processes were extracted, what was the length of the SRL processes (e.g., a two-step pattern as action A to action B, or a three-step pattern as action A to action B to action C), and what are the possible interpretations of patterns.

For example, during our discussions, an action pattern "GENERAL_INSTRUCTION -> PLANNER" was proposed by one researcher in an attempt to map this pattern to the *Planning* process under the *Metacognition* category in our theoretical framework. However, another researcher added their observation that the timing of checking the planner tool affected the interpretation of this pattern. For example, learners could have opened the planner tool to plan at the beginning of their learning sessions (i.e., this is indicative of the *Planning* process). They could have also checked the general instructions page and their plans while reading and writing, and this same pattern ("GENERAL_INSTRUCTION -> PLANNER") could be interpreted as the *Monitoring* process. Therefore, we created an extraction rule for this SRL process which considered "GENERAL_INSTRUCTION -> PLANNER" in the first third of the learning session (first 15 minutes) as *Planning* but in the last two thirds of the learning session as *Monitoring*[3]. These discussions, therefore, resulted in a theory-driven process library (version 1).

Based on the action library and the theory-driven process library, we built the trace parser (Fig. 3), which enabled us to extract SRL processes as action patterns from learning actions. This approach was originally proposed by Siadaty and colleagues (Siadaty et al., 2016c), which has been shown as an effective way to extract SRL processes from trace data (Saint et al., 2020b). In this **theory-driven perspective**, we effectively characterised learners' SRL processes in the context of our theoretical framework. More details about the theory-driven process library (version 1) can be found in the Appendix.

### Step 2: Empirical evaluation of the theory-driven process library against think aloud data (first alignment)

After obtaining the theory-driven SRL processes as our output, we conducted the first round of alignment between trace-based SRL processes and think aloud codes[4].

To collect think aloud data, we recorded learners' utterances during the learning sessions. These utterances were then segmented, transcribed and coded by well-trained coders based on a previously developed coding scheme (Bannert, 2007; Molenaar et al., 2011). For example, learners' utterances during their reading of the instruction page, such as, "these three topics are important in the task", which were coded as *Orientation* (MC.O) based on our coding scheme.

Figure 4 visualises the alignment of these two measurement results based on a synchronised timeline, and displays five states of alignment. These five states are: (1) Match, that

[3]This is not to say that we assumed the *Planning* process did not occur after 15 minutes. Instead, we interpreted this specific pattern relative to its temporal occurrence during the session, i.e, ("GENERAL_INSTRUCTION -> PLANNER") as *Planning* in the first 15 minutes and as *Monitoring* in the last 30 minutes. The results based on think aloud data also supported this interpretation (details are provided in the Results section).

[4]It is worth pointing out that **think aloud codes** also represented SRL processes, that is, both SRL processes extracted from trace data and coded based on think aloud data, were all SRL processes which reflected the categories from our theoretical framework. However, in order to distinguish and simplify the expression of these two concepts, below, we use **SRL processes** to refer to SRL processes extracted from trace data and **think aloud codes** to refer to SRL processes extracted from think aloud data.

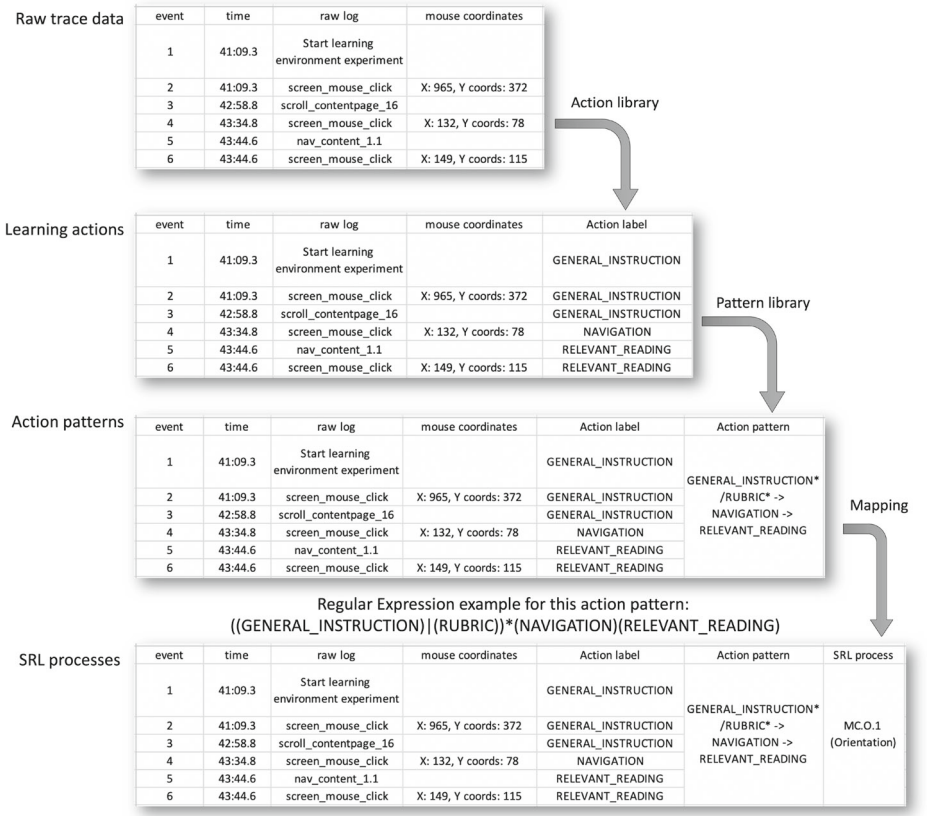


**Fig. 3** The process followed by the trace parser to extract SRL processes

is same codes/processes were extracted based on both trace and think aloud data, for example both trace data and think aloud data measured *Orientation* (MC.O) for the same time slot (see the first green box in Fig. 4); (2) Mismatch, that is different codes/processes were extracted based on trace and think aloud data, for example trace data measured *Elaboration/Organisation* (HC.E/O) but think aloud measured *First-reading* (LC.F) for the same time slot (see the first yellow box in Fig. 4); (3) Only trace, that is no think aloud data was collected and coded in this time period, only processes in trace data were collected (see the dark gray box in Fig. 4); (4) Only think aloud, that is no SRL processes were extracted in trace data in this time period (see the light gray box in Fig. 4); (5) No processes, that during this time period, there were neither SRL processes extracted in trace data nor codes assigned to think aloud data (see the black box in Fig. 4).

In this study, we used four quantitative indicators to evaluate the agreement between the two methods and determine the necessity for further validity improvement of the trace-based SRL measurement if think aloud data is used as reference point. These four quantitative indicators are: (1) Sensitivity (True Positive Rate) based on frequency; (2) Specificity (True Negative Rate) based on frequency; (3) Match rate based on duration; and (4) Trace coverage based on duration. The definitions of these four indicators can be found in Table 3.

**Table 3** The definitions of the four quantitative indicators

| Indicators | Bases on | Definitions |
|---|---|---|
| Sensitivity | Frequency | The proportion of occasions when reference point (think aloud) measured one SRL process (e.g., MC.O) and the trace data also measured the same SRL process (e.g., MC.O): $Sensitivity = \frac{Trace\ data\ measured\ A}{When\ think\ aloud\ measured\ A}$ |
| Specificity | Frequency | The proportion of occasions when trace data measured one specific SRL process (e.g., MC.O) and the reference point (think aloud) also measured it as that SRL process (e.g., MC.O): $Specificity = \frac{Think\ aloud\ measured\ A}{When\ trace\ data\ measured\ A}$ |
| Match rate | Duration | The ratio between the matched duration (e.g., two methods measured MC.O at the same time) and the sum of both matched and mismatch duration: $Match\ rate = \frac{Match\ duration}{Match\ duration + Mismatch\ duration}$ |
| Trace coverage | Duration | The ratio between the duration of the trace-based SRL processes and the full task duration $Trace\ coverage = \frac{Trace-based\ SRL\ processes}{Whole\ task\ duration}$ |

Sensitivity and Specificity are two indicators that are often used to describe the accuracy of a test which indicates the presence or absence of a condition, in comparison to the "gold standard" or the "reference point". However, in addition to the frequency-based evaluation, we also conducted the duration-based evaluation. To that end, in this study we defined **Match rate**[5] and **Trace coverage** as the other two indicators to describe the measurement results more comprehensively. The reason we introduce the match rate is that we believe that for very fine-grained data channels (such as think aloud), only using frequency-based statistics may create an incomplete view. For example, some SRL processes or think aloud codes appear very frequently but they overall occupied a very short time. Therefore, from a temporal point of view, we value the match rate which can provide an overall assessment of the validity of the trace-based measurement results when considering think aloud as reference point. We also defined the **Trace coverage** as another indicator to evaluate how big a proportion of the whole task duration can be interpreted into SRL processes using trace data only. This last indicator was introduced here to avoid excessive sacrifice of coverage to improve measurement accuracy. For example, if we only measured a small proportion of the whole learning process, even if the validity of the measurement could be guaranteed, the measurement result could not be used to reflect the self-regulation process of the whole task. Therefore, the ideal outcome of our validation approach should be to increase the first three indicators (sensitivity, specificity and match rate) while ensuring that the last indicator (trace coverage) does not fluctuate too much.

In this step, in order to examine the reasons behind the mismatches in the theory-driven process library, we also measured the duration of all mismatch pairs between SRL processes and think aloud codes. This analysis enabled us to extract invalid or over-interpretations and improve our process library into a more valid measurement protocol with the help of think aloud data.

[5]Note that the "No processes", "Only trace" and "Only think aloud"(see Fig. 4) were not included in the calculation of the match rate.

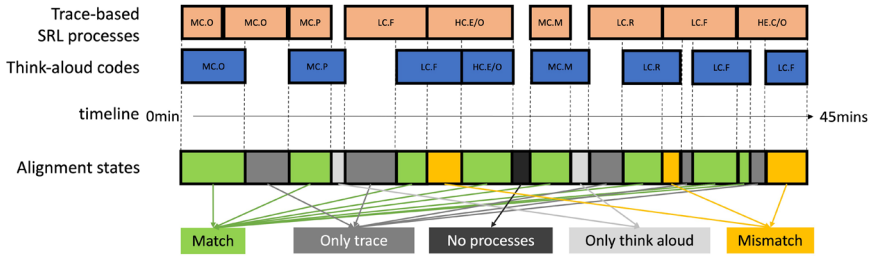


**Fig. 4** Schematic diagram of first alignment

### Step 3: Data-driven perspective: data-driven process library (version 2) guided by a training set of think aloud data (second alignment)

After the first alignment had been performed and the four indicators calculated (Evaluation V1 in Fig. 2), the third step of the whole validation approach started, which is referred to as a **data-driven perspective**. In this step, we did not utilise the theory-driven SRL processes; instead, we conducted a second alignment between think aloud data and **learning actions**. As shown in Fig. 5, we first merged the same consecutive learning actions and think aloud, because the granularity of think aloud data was much finer than that of the learning actions in trace data (e.g., there were many short intervals during a series of reading codes, but there was only one long reading action labelled from trace data). Then, we segmented the learning actions based on the events encapsulated by think aloud codes (see Fig. 5). In this way, we created a short session for each think aloud code, and assigned a code-based session ID to the time slot of each code (e.g., P1_code2_01 is the session ID of time slot of think aloud code 2 in Fig. 5). All code-based sessions could be grouped by the corresponding think aloud codes, which formed code-based **session-groups**; for example, all short sessions which were segmented and coded as *Monitoring* (MC.M) were considered the MC.M session-group. By analysing and identifying common learning actions and action patterns in the sessions which belonged to different session-groups corresponding to each type of code, we were able to extract and summarise mapping between SRL processes extracted from trace data and codes applied to think aloud data. For example, if the action pattern "label 1 -> label 2 -> label 3" (see Fig. 5) frequently appeared in the MC.M session-group and was rarely extracted in the other session-groups, then we would conclude that this action pattern matched the *Monitoring* code. Therefore, this action pattern extracted from trace data could be interpreted as the SRL process *Monitoring*.

In order to find more SRL processes, we applied a process mining technique to analyse the session-groups corresponding to each type of think aloud code. We used a process mining toolkit called DISCO which is based on the fuzzy miner algorithm (Günther & Rozinat, 2012) to further process maps that visualised the most dominant action patterns (extracted from trace data) within one session-group. The process maps generated for each session-group and statistics of the frequency of patterns enabled us to identify meaningful SRL processes which aligned with think aloud.

As shown in the validation approach (Fig. 2), this second alignment and results based on process mining enabled us to (1) add more meaningful action patterns that can be interpreted as new SRL processes; and (2) find more appropriate interpretations for action patterns. In addition, the first alignment also helped us identify and remove problematic SRL processes which were largely mismatched with think aloud codes. These two alignments provided

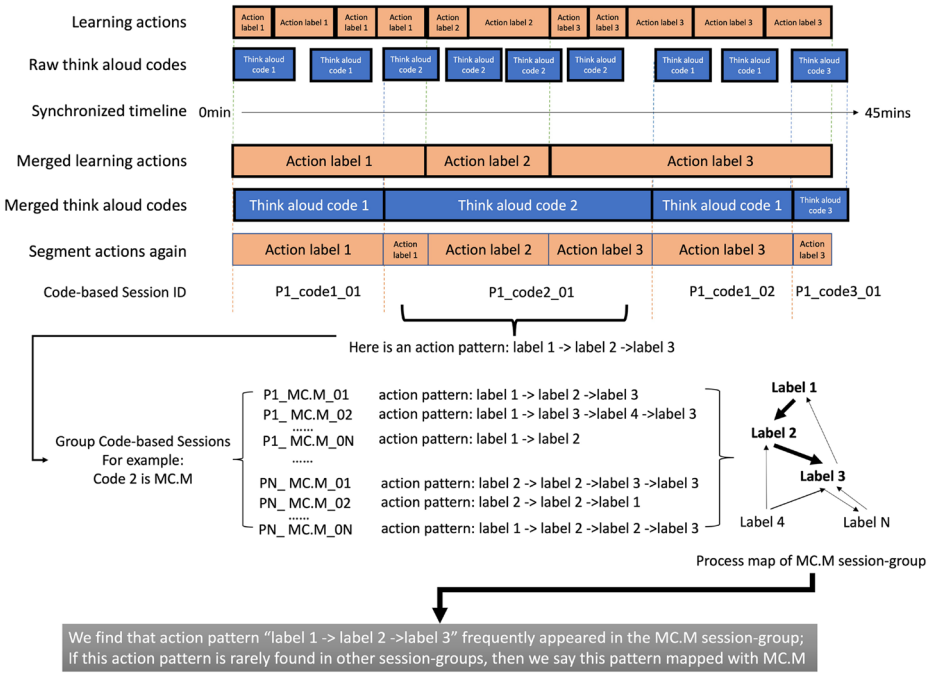


**Fig. 5** Schematic diagram of the second alignment

empirical evidence to the valid interpretation and helped us constructed the **data-driven process library**.

### Step 4: Combining the empirical evidence and the theoretical rationale: improved process library (version 3)

It is worth noting that, after constructing the **data-driven process library**, we once again introduced theoretical rationale to help us make final decisions on the interpretation of controversial action patterns. This step is shown in Fig. 2, in which two arrows pointing to the **improved process library** are the **data-driven process library** and the **SRL theoretical model** (Table 1). This is because, although we regarded the think aloud as the reference point in this research, we did not want to completely "rely on" or "blindly follow" the think aloud data and break away from our SRL theoretical model when interpreting the trace data.

For example, when the think aloud data failed to give a specific inference on a certain action pattern, i.e., this pattern evenly distributed across two or multiple session-groups, but this action pattern was still meaningful from the perspective of our theoretical framework, then we interpreted this pattern based on the theoretical rationale. In other cases, the interpretations of certain patterns guided by the think aloud could also conflict with definitions of certain SRL processes based on our theoretical model. Under these circumstances, we also made comprehensive judgements by combining the empirical evidence and the theoretical rationale. More specific examples are teased out in the results and discussion sections. Based on the **improved process library** we constructed in this step, we parsed the learning

actions again (as shown in the upper right part of Fig. 2) and obtained the improved SRL processes as output.

### Step 5: Empirical evaluation of the improved process library against think aloud data in both training set and test set (third alignment)

These improved SRL processes were then aligned with the coded think aloud data for the third time. The third alignment (see upper middle of Fig. 2) was performed in the same way as the first alignment (which is shown in Fig. 4) and we calculated the second set of four indicators (Evaluation V2 in Fig. 2) by following the same calculation methods as already defined. By comparing these four indicators before (V1) and after (V2) the validation, we evaluated whether and to what extent we improved the validity of the **theory-driven process library**. This entire validation approach could be iterated for a second round if the improvement of the measurement did not reach a satisfactory level.

### Training set and test set

Even though this study was not a typical prediction-based study, our modelling approach (here, in particular, the validation approach) could have also led to a model of SRL processes that could be too closely fitted to our limited set of data and could fail to fit unseen data. For example, the match rate between the **improved SRL processes** and **think aloud codes** may have reached a satisfactory level, because many infrequent action patterns in different session-groups were overly interpreted as SRL processes. If these infrequent action patterns were noisy in this dataset but not found in an unseen dataset, the model we built would fail to reach a high match rate with think aloud codes (reference point) again.

In order to avoid this problem of over-fitting, we randomly selected 32 participants from the sample of the 44 participants as the training set to improve the validity of our trace-based SRL measurement. We then used the remaining 12 participants[6] as the test set to verify the performance of the **improved process library**. The four indicators (V1 and V2) of both the training set and test set are reported in the Results section.

### Think aloud coding as "reference point"

Our study is unique in considering coded think aloud data as the reference point to measure SRL processes and relating the think aloud data to trace data. Therefore in our study, we paid great attention to improving the quality of the coding process of think aloud data, providing as rich and reliable information as possible for measuring SRL. First, there was a short training session in which the experimenter explained how to think aloud and the participants had a chance to practice thinking aloud. Second, two coders used the ELAN software (Aguera et al., 2011) to code the utterances of each of the participants. We also examined the inter-rater reliability for coding of think aloud data between coders, which reached acceptable inter-rater reliability: $k$ = .53-.65 ($k_{max}$ = .81-.82).

[6]In this study, we originally selected 70% (32) of 45 participants as training set and the other 30% (13) participants as test set. However, one participant in the test set was excluded after the analysis was done. This is because we afterwards found out that the participant's first language was not Dutch which may have influenced the quality of their think aloud data. Therefore, we now only show the results based on data from 12 participants as the test set.

## Results

The results section is organised according to the five steps of our validation approach, as described above.

### The theory-driven process library (version 1)

Following the **theory-driven perspective**, we constructed the theory-driven process library (Table 4) and generated the theory-driven SRL processes based on our theoretical framework. The overall duration for each theory-driven SRL process is shown in Table 5. In this table, we also calculated the duration of each code from think aloud data, for the sake of comparison. As shown in Table 5, the overall duration for each theory-driven SRL process and think aloud code were quite different. A total of 41.66% of the duration of all learning sessions did not have any code assigned based on the coding of think aloud data; however, only 3.19% were not coded (no process) based on trace data. A total of 50.42% of the duration of all learning sessions were extracted as instances of the *HC.E/O* process based on trace data (theory-driven SRL process); only 15.42% were coded as *HC.E/O* based on think aloud data. We also extracted a smaller percentage of instances of the *LC.F* and *MC.P* process in the theory-driven SRL process results, compared to the think aloud codes.

### Empirical evaluation of the theory-driven process library

In order to answer the first research question, we conducted the first alignment between the theory-driven SRL processes and think aloud codes by following the approach shown in Fig. 4. In Fig. 6, we use one learner as an example, to display the results of the first alignment between the theory-driven SRL processes and think aloud codes for this selected learner. The first and second tracks show SRL processes and think aloud codes on the same timeline. For example, the learner started their learning with *MC.O* (orange) then moved on to *LC.F* (green) or *HC.E/O* (blue). The third track shows five alignment states of alignment given in Fig. 4, which includes match (green), mismatch (orange), only trace (dark grey), only think aloud (light grey), and No processes from both data channels (black). The match rate (green part/(green part + orange part) in Fig. 6) was used as a quantified indicator together with the other three indicators (sensitivity, specificity and trace coverage) to evaluate the agreement between the two methods of measuring this learner's SRL. The first alignment we conducted for 32 participants in the training set provided the baseline to evaluate and further improve the validity of the trace-based SRL measurement with the coded think aloud data that were used as the reference point. Table 6 shows the evaluation result based on all four indicators[7].

As shown in Table 6, we only achieved a match rate of 38.97% (median) between the theory-driven SRL processes and think aloud codes at the first alignment. The sensitivity and specificity are also low based on this first alignment. These results indicate there was a need and room for improvement of the validity of the trace-based SRL measurement as compared to the reference point.

[7]The sensitivity and specificity in the table are the overall sensitivity and specificity of 32 participants and of all codes. Take sensitivity as an example, we first calculate sensitivity for a single code (e.g., MC.O) of a single learner (e.g., participant NO.5), then we calculated the average sensitivity of all seven codes for the learner, then we calculated the average of the average sensitivities of all 32 participants as the overall sensitivity which is shown in Table 6.

**Table 4** The theory-driven process library for detection of SRL processes from action labels

| Code | No. | Processes |
|---|---|---|
| MC.O | MC.O.1 | GENERAL_INSTRUCTION* /RUBRIC* -> NAVIGATION -> RELEVANT_READING* |
| | MC.O.2 | GENERAL_INSTRUCTION /RUBRIC -> GENERAL _INSTRUCTION/RUBRIC |
| | MC.O.3 | GENERAL_INSTRUCTION /RUBRIC <-> HIGHLIGHT_EDITING/NAVIGATION |
| | MC.O.4 | GENERAL_INSTRUCTION */RUBRIC* |
| MC.P | MC.P.1 | PLANNER -> NAVIGATION -> RELEVANT_READING |
| | MC.P.2 | GENERAL_INSTRUCTION /RUBRIC <-> PLANNER (during first 15mins) |
| | MC.P.3 | PLANNER (during first 15mins) |
| | MC.P.4 | SEARCH_CONTENT* |
| MC.E | MC.E.1 | IRRELEVANT_(RE-)READING* -> GENERAL_INSTRUCTION -> RELEVANT_(RE-)READING* |
| MC.M | MC.M.1 | GENERAL_INSTRUCTION* -> WRITE_ESSAY -> WRITE_ESSAY |
| | MC.M.2 | NAVIGATION <-> NOTE_READING |
| | MC.M.3 | GENERAL_INSTRUCTION <-> PLANNER (after the first 15mins) |
| | MC.M.4 | WRITE_ESSAY <-> PLANNER |
| | MC.M.5 | TIMER* |
| | MC.M.6 | PLANNER* (after the first 15mins) |
| LC.F | LC.F.1 | (IR)RELEVANT_READING* -> HIGHLIGHT_EDITING/NOTE_EDITING -> (IR)RELEVANT_READING* |
| | LC.F.2 | (IR)RELEVANT_READING* -> NAVIGATION -> (IR)RELEVANT_READING* |
| | LC.F.3 | RELEVANT_READING* -> IRRELEVANT_READING -> IRRELEVANT_READING* |
| | LC.F.4 | (IR)RELEVANT_READING* <-> HIGHLIGHT_EDITING/ NOTE_EDITING/HIGHLIGHT _READING/ NOTE_READING |

**Table 4** (continued)

| Code | No. | Processes |
|---|---|---|
| | LC.F.5 | (IR)RELEVANT_READING* <-> (IR)RELEVANT_READING* |
| | LC.F.6 | IRRELEVANT_READING* |
| LC.R | LC.R.1 | RELEVANT_RE-READING* |
| | LC.R.2 | IRRELEVANT_RE-READING* |
| | LC.R.3 | (IR)RELEVANT_RE-READING* <-> HIGHLIGHT_EDITING/ NOTE_EDITING/HIGHLIGHT _READING/ NOTE_READING |
| HC.E/O | HC.E/O.1 | (IR)RELEVANT_RE-READING -> WRITE_ESSAY -> WRITE_ESSAY* |
| | HC.E/O.2 | RUBRIC* -> WRITE_ESSAY -> WRITE_ESSAY* |
| | HC.E/O.3 | WRITE_ESSAY -> WRITE_ESSAY* |
| | HC.E/O.4 | WRITE_ESSAY <-> HIGHLIGHT_READING/NOTE_READING |
| | HC.E/O.5 | GENERAL_INSTRUCTION/RUBRIC <-> WRITE_ESSAY |
| | HC.E/O.6 | NAVIGATION <-> RELEVANT_RE-READING/WRITE_ESSAY |
| | HC.E/O.7 | HIGHLIGHT_LABELLING* |
| | HC.E/O.8 | NOTE_EDITING* |
| | HC.E/O.9 | HIGHLIGHT_READING/NOTE_READING* |

Legend: "->" means a transition from action A to action B; "<->" means a transition from action A to action B or the other way around; "()" means optional; "*" means one or more consecutive instances of the same action; "/" means either action A or action B

**Table 5** The distribution of time on each SRL process extracted from trace data and think aloud code

| SRL processes | Codes | Theory-driven SRL processes | Think aloud codes |
|---|---|---|---|
| Orientation | MC.O | 4.35% | 2.40% |
| Planning | MC.P | 0.57% | 2.88% |
| Monitoring | MC.M | 14.70% | 5.56% |
| Evaluation | MC.E | 2.65% | 0.11% |
| First-reading | LC.F | 19.98% | 22.37% |
| Re-reading | LC.R | 4.14% | 0.87% |
| Elaboration/Organisation | HC.E/O | 50.42% | 15.42% |
| Other | Other | 0.00% | 8.73% |
| No codes/No processes | Not-coded | 3.19% | 41.66% |

In order to examine the reasons behind the mismatches in the theory-driven process library, we also measured the duration of all mismatch pairs between SRL processes and think aloud codes. A total of 41.69% of all mismatch time was caused by the mismatch between *HC.E/O* and *LC.F*. For instance in Fig. 6, some *LC.F* (green parts) in the coded think aloud data were extracted from trace data as *HC.E/O* (blue parts) in the theory-driven SRL processes. This largest mismatch between LC.F and HC.E/O can be understood from two aspects: (1) these two processes occupied the largest proportion of time during the whole learning session and therefore, there was a high probability that a mismatch would be found between these processes; (2) several action patterns that occurred frequently and accounted for a large amount of time were incorrectly interpreted or over interpreted. For example, based on our initial discussion, we interpreted "making highlights during reading" as a High_Cognition process which indicates learners organising the reading materials by using different tags in the highlight tool. This was a pattern that often occurred and took a long time during the reading process, and therefore, caused mismatches that occupied a large proportion of time. However, this interpretation was not supported by think aloud data where most of the time slots of this process were coded as reading (i.e., *LC.F*) in think aloud. Because in most cases, learners were simply using the highlight tool to process the learning content just by visually distinguishing certain keywords or sentences. This situation should be interpreted as a reading pattern (LC.F), although it required more cognitive input than "just reading". Therefore, we changed our interpretation of this pattern from HC.E/O to LC.F in the **improved process library**.

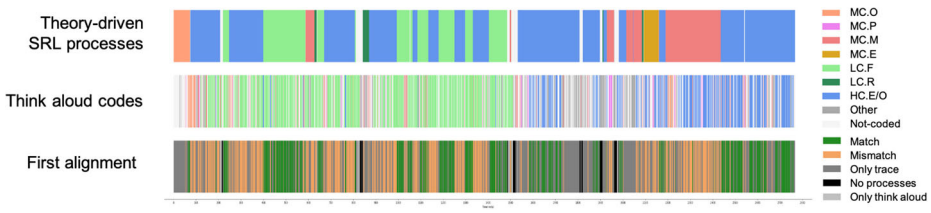


**Fig. 6** Alignment between the theory-driven SRL processes extracted from trace data and coded think aloud data (one learner as an example); Legend: MC.O – Orientation; MC.P – Planning; MC.E – Evaluation; MC.M – Monitoring; LC.F – First-reading; LC.R – Re-reading; HC.E/O – Elaboration and Organisation; Other – other think aloud codes; Un-coded – No think aloud codes or SRL processes being extracted

**Table 6** Empirical evaluation (V1) of the theory-driven process library against think aloud data (first alignment results)

| Indicators | Based on | Median | 25th | 75th |
|---|---|---|---|---|
| Sensitivity | Frequency | 21.60% | 17.30% | 25.62% |
| Specificity | Frequency | 21.88% | 19.67% | 26.99% |
| Match rate | Duration | 38.97% | 34.61% | 47.20% |
| Trace coverage | Duration | 97.43% | 96.16% | 98.66% |

Another 22.74% of all mismatches were caused by the mismatch between *MC.M* and *HC.E/O*. For instance, in Fig. 6, some *MC.M* (red parts) in the theory-driven SRL processes extracted from trace data were coded as *HC.E/O* (blue parts) in think aloud data. This is mainly caused by one specific action pattern (write essay after checking the instruction page) that we over-interpreted as MC.M in the **theory-driven process library**. We found this interpretation invalid because according to think aloud data, learners, in most cases, were elaborating in the essay window as required or guided by the instruction page. Therefore, we changed our interpretation of this pattern from MC.M into HC.E/O in the **improved process library**.

Based on this analysis of mismatches, we found several interpretations invalid in the **theory-driven process library**. This previous step enabled us to extract invalid or over-interpretations and to improve our process library to support a more valid measurement protocol with the help of think aloud data. However, this step could not guide us to find more SRL processes in trace data that would match the codes of think aloud data. Therefore, we opted to use the **data-driven perspective** which aligned codes used for the analysis of think aloud data and learning actions extracted from trace data to detect more SRL processes.

### The data-driven process library (version 2)

As addressed in Fig. 2, the **data-driven perspective** of our validation approach started with the alignment of codes assigned to think aloud data and learning actions from trace data. In Figs. 7 and 8, we refer to two session-groups (*LC.F* and *MC.O*) as examples to explain how we identified data-driven SRL processes with process mining. For instance, in the first example *LC.F*, we found two three-step patterns frequently appearing in the *First-reading* sessions (see the left part of the process map of *LC.F* in Fig. 7). This suggested that action patterns such as "RELEVANT_READING -> HIGHLIGHT_EDITING -> RELEVANT_READING" could be a typical reading process (making highlights while reading) during SRL. Figure 8 shows the process map of all *MC.O* sessions as another example. This example helped us find frequent SRL processes such as "GENERAL_INSTRUCTION/RUBRIC -> NAVIGATION -> RELEVANT_READING" (after reading the task instruction, the learners navigated to some content pages). These findings enabled us to interpret patterns as valid SRL processes which were supported by using the think aloud data as reference point. The full collection of process maps for all sub-categories of think aloud codes is given in the Appendix.

In addition to three-step patterns, we use two-step patterns which would start with HIGHLIGHT_EDITING as an example, to show which think aloud codes were coded during these patterns (see Table 7). For example, as shown in Table 7: 75.00% of "HIGHLIGHT_EDITING -> GENERAL_INSTRUCTION" and 85.72% "HIGHLIGHT_EDITING -> RUBRIC" were coded as *MC.O* based on think aloud data. This

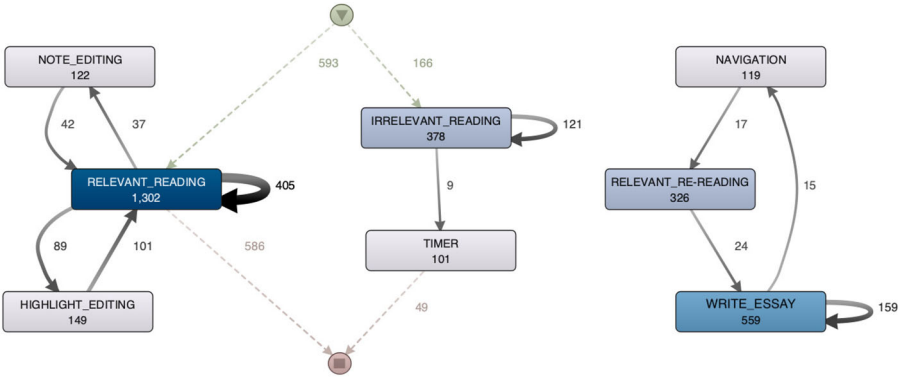


**Fig. 7** Example 1: process map of all *First-reading, LC.F* sessions ; The numbers are the frequency of actions or transitions of actions in this session group

indicates that these two patterns extracted from trace data were largely mapped to the code *MC.O*, and therefore can be interpreted as *MC.O* processes; 69.12% of "HIGHLIGHT_EDITING -> RELEVANT_READING" and 73.33% of "HIGHLIGHT_EDITING -> IRRELEVANT_READING" were coded as *LC.F* based on think aloud data. This indicates these two patterns extracted from trace data were largely mapped to the code *LC.F*, and therefore, can be interpreted as *LC.F* processes.

Following the **data-driven perspective**, we were able to extract many new SRL processes from trace data which could be mapped to codes used for the analysis of think aloud data. The output of the data-driven approach was the **data-driven process**

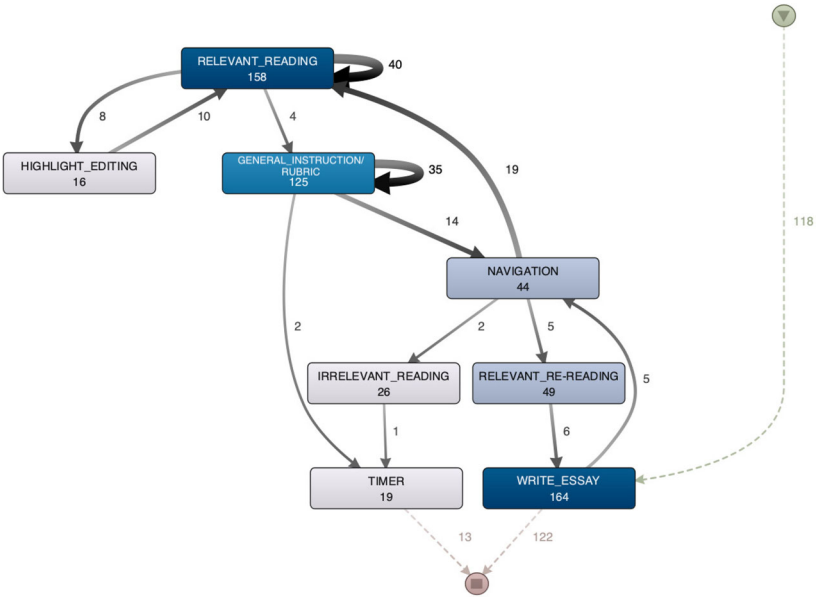


**Fig. 8** Example 2: process map of all *Orientation, MC.O* sessions; The numbers are the frequency of actions or transitions of actions in this session group

**Table 7** Two-step patterns that started with HIGHLIGHT_EDITING and their distribution under the think aloud code from Table 1

| HIGHLIGHT_EDITING to | HC.E/O | LC.F | LC.R | MC.E | MC.M | MC.O | MC.P |
|---|---|---|---|---|---|---|---|
| GENERAL_INSTRUCTION | 25.00% (1) | - | - | - | - | 75.00% (3) | - |
| RUBRIC | - | 14.28% (1) | - | - | - | 85.72% (6) | - |
| RELEVANT_READING | 4.11% (6) | 69.18% (101) | 2.74% (4) | - | 15.07% (22) | 4.11% (6) | 4.79% (7) |
| RELEVANT_RE-READING | 5.41% (2) | 59.46% (22) | - | - | 16.22% (6) | 2.70% (1) | 16.22% (6) |
| IRRELEVANT_READING | 13.33% (11) | 73.33% (11) | 6.67% (1) | - | 6.67% (1) | - | - |
| IRRELEVANT_RE-READING | - | - | - | - | 50.00% (2) | - | 50.00% (2) |

**Table 8** Distribution of SRL processes under each think aloud code (in percentage) (part 1)

| SRL processes | HC.E/O | LC.F | LC.R | MC.E | MC.M | MC.O | MC.P |
|---|---|---|---|---|---|---|---|
| GENERAL_INSTRUCTION*/ RUBRIC* -> (NAVIGATION) -> RELEVANT_READING | 8.28 | 8.11 | 0.56 | 0.00 | 12.50 | 59.49 | 11.06 |
| PLANNER -> (NAVIGATION) -> RELEVANT_READING | 7.82 | 25.78 | 0.00 | 0.00 | 17.48 | 16.39 | 32.53 |
| IRRELEVANT _READING/IRRELEVANT _RE-READING -> (NAVIGATION) -> GENERAL_INSTRUCTION*/RUBRIC* -> (NAVIGATION) -> (RELEVANT_READING/RELEVANT_RE-READING) | 11.85 | 31.73 | 0.00 | 16.35 | 6.56 | 31.20 | 1.42 |
| (IR)RELEVANT_READING -> HIGHLIGHT_EDITING /NOTE_EDITING/ HIGHLIGHT_LABELLING -> (IR) RELEVANT_READING | 15.48 | 70.11 | 0.93 | 0.19 | 7.19 | 1.43 | 4.68 |
| (IR)RELEVANT_READING -> NAVIGATION -> (IR)RELEVANT_READING | 13.24 | 65.93 | 0.65 | 0.22 | 11.76 | 3.42 | 4.79 |
| RELEVANT_READING -> IRRELEVANT_READING -> IRRELEVANT_READING | 7.86 | 74.02 | 1.47 | 0.38 | 8.11 | 3.29 | 4.86 |
| (IR)RELEVANT_RE-READING -> (NAVIGATION) -> WRITE_ESSAY -> WRITE_ESSAY* | 57.02 | 15.25 | 3.27 | 0.14 | 16.04 | 1.66 | 6.61 |
| GENERAL_INSTRUCTION /RUBRIC -> (NAVIGATION) -> WRITE_ESSAY -> WRITE_ESSAY* | 51.82 | 7.55 | 0.76 | 0.15 | 13.83 | 16.62 | 9.28 |
| GENERAL_INSTRUCTION /RUBRIC -> (NAVIGATION) -> GENERAL_INSTRUCTION/ RUBRIC | 21.71 | 23.70 | 1.44 | 0.11 | 12.91 | 31.60 | 8.54 |
| GENERAL_INSTRUCTION /RUBRIC <-> HIGHLIGHT_EDITING/ NOTE_EDITING/ NAVIGATION | 13.61 | 25.92 | 0.24 | 0.00 | 13.58 | 38.51 | 8.14 |
| GENERAL_INSTRUCTION /RUBRIC <-> PLANNER* (during first 15mins) | 0.00 | 0.00 | 0.00 | 0.00 | 18.71 | 61.71 | 19.57 |
| NAVIGATION <-> NOTE_READING | 50.42 | 2.52 | 0.00 | 0.00 | 21.41 | 18.36 | 7.28 |

**library**, and all the data-driven SRL processes extracted from trace data can be found in Tables 8 and 9. We also calculated the distribution of these SRL processes under different think aloud codes. For example, 59.49% of the first SRL process in Table 8 ("GENERAL_INSTRUCTION*/RUBRIC* -> (NAVIGATION) -> RELEVANT_READING")

**Table 9** Distribution of SRL processes under each think aloud code (in percentage) (part 2)

| SRL processes | HC.E/O | LC.F | LC.R | MC.E | MC.M | MC.O | MC.P |
|---|---|---|---|---|---|---|---|
| GENERAL_INSTRUCTION /RUBRIC <-> PLANNER* (after the first 15mins) | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| WRITE_ESSAY <-> PLANNER* | 10.55 | 3.31 | 0.00 | 0.00 | 49.24 | 36.13 | 0.78 |
| (IR)RELEVANT _READING <-> HIGHLIGHT _EDITING /NOTE_EDITING/ HIGHLIGHT_READING/ NOTE_READING | 15.89 | 69.09 | 1.78 | 0.02 | 5.85 | 2.76 | 4.60 |
| (IR)RELEVANT_READING <-> (IR)RELEVANT_READING | 11.49 | 68.72 | 0.73 | 0.29 | 8.67 | 3.91 | 6.18 |
| WRITE_ESSAY -> WRITE_ESSAY* | 52.30 | 25.62 | 2.29 | 0.29 | 12.70 | 1.98 | 4.82 |
| WRITE_ESSAY <-> HIGHLIGHT _READING/NOTE_READING | 21.26 | 71.27 | 3.55 | 0.00 | 0.92 | 0.00 | 2.99 |
| GENERAL_INSTRUCTION */RUBRIC* | 16.76 | 25.77 | 1.61 | 1.11 | 28.03 | 17.95 | 8.76 |
| PLANNER* (during first 15mins) | 5.83 | 0.00 | 0.00 | 0.00 | 9.45 | 40.10 | 44.62 |
| SEARCH_CONTENT* | 5.96 | 46.13 | 4.63 | 0.00 | 17.17 | 0.00 | 26.12 |
| TIMER* | 58.78 | 19.40 | 2.04 | 0.12 | 13.29 | 3.03 | 3.33 |
| PLANNER* (after the first 15mins) | 50.47 | 0.00 | 0.00 | 0.00 | 49.53 | 0.00 | 0.00 |
| SEARCH_HIGHLIGHT_NOTE* | 32.22 | 10.26 | 0.00 | 0.00 | 28.34 | 14.38 | 14.80 |
| HIGHLIGHT_READING /NOTE_READING* | 45.94 | 31.16 | 1.40 | 0.00 | 13.45 | 1.53 | 6.52 |
| IRRELEVANT_READING* | 16.42 | 62.10 | 2.01 | 0.00 | 12.54 | 2.41 | 4.52 |
| RELEVANT_READING* | 16.42 | 65.76 | 1.85 | 0.05 | 8.58 | 2.40 | 4.93 |
| RELEVANT_RE-READING* | 24.01 | 50.57 | 3.77 | 0.43 | 12.86 | 2.75 | 5.61 |
| IRRELEVANT_RE-READING* | 29.80 | 43.76 | 7.19 | 0.00 | 12.89 | 3.18 | 3.18 |
| HIGHLIGHT_LABELLING* | 6.51 | 80.17 | 0.00 | 0.00 | 9.06 | 0.00 | 4.26 |
| NOTE_EDITING* | 28.88 | 42.78 | 0.71 | 0.00 | 16.30 | 4.67 | 6.65 |

Legend: "->" means a transition from learning action A to learning action B; "<->" means a transition from learning action A to learning action B or the other way around; "()" means this learning action is optional; "*" means one or more consecutive instances of the same learning action; "/" means either learning action A or learning action B; "(IR)RELEVANT_READING" means RELEVANT_READING or IRRELEVANT_READING, and (IR)RELEVANT_RE-READING means RELEVANT_RE-READING or IRRELEVANT_RE-READING

was extracted during the *MC.O* (think aloud code) session-group, which was higher than in the other session-groups. This three-step pattern was also proposed in our **theory-driven process library** as an *Orientation* process. Therefore, the interpretation of this three-step pattern was supported by both evidence from think aloud data and our theoretical hypothesis.

**Table 10** The improved process library for detection of SRL processes from action labels

| Code | No. | Processes |
|---|---|---|
| MC.O | MC.O.1 | GENERAL_INSTRUCTION */RUBRIC* -> NAVIGATION -> RELEVANT_READING |
| | MC.O.2 | GENERAL_INSTRUCTION /RUBRIC -> GENERAL_INSTRUCTION /RUBRIC |
| | MC.O.3 | GENERAL_INSTRUCTION /RUBRIC <-> HIGHLIGHT_EDITING/NOTE_EDITING/ NAVIGATION |
| | MC.O.4 | GENERAL_INSTRUCTION */RUBRIC* |
| MC.P | MC.P.1 | PLANNER -> NAVIGATION -> RELEVANT_READING |
| | MC.P.2 | GENERAL_INSTRUCTION /RUBRIC <-> PLANNER* (during first 15mins) |
| | MC.P.3 | PLANNER* (during first 15mins) |
| | MC.P.4 | SEARCH_CONTENT* |
| MC.E | MC.E.1 | IRRELEVANT_(RE-)READING -> (NAVIGATION) -> GENERAL_INSTRUCTION*/ RUBRIC* -> (NAVIGATION) -> RELEVANT_(RE-)READING |
| MC.M | MC.M.1 | NAVIGATION <-> NOTE_READING |
| | MC.M.2 | GENERAL_INSTRUCTION /RUBRIC <-> PLANNER* (after the first 15mins) |
| | MC.M.3 | WRITE_ESSAY <-> PLANNER* |
| | MC.M.4 | TIMER* |
| | MC.M.5 | PLANNER* (after the first 15mins) |
| | MC.M.6 | SEARCH_HIGHLIGHT_NOTE* |
| | MC.M.7 | HIGHLIGHT_READING/NOTE_READING* |
| LC.F | LC.F.1 | (IR)RELEVANT_READING -> HIGHLIGHT_EDITING/NOTE_EDITING -> (IR)RELEVANT_READING |
| | LC.F.2 | (IR)RELEVANT_READING -> NAVIGATION -> (IR)RELEVANT_READING |
| | LC.F.3 | RELEVANT_READING -> IRRELEVANT_READING -> IRRELEVANT_READING |
| | LC.F.4 | (IR)RELEVANT_READING <-> HIGHLIGHT_EDITING/NOTE_EDITING/ HIGHLIGHT_READING/NOTE_READING |

**Table 10** (continued)

| Code | No. | Processes |
|---|---|---|
| | LC.F.5 | (IR)RELEVANT_READING <-> (IR)RELEVANT_READING |
| | LC.F.6 | IRRELEVANT_READING* |
| | LC.F.7 | RELEVANT_READING* |
| LC.R | LC.R.1 | RELEVANT_RE-READING* |
| | LC.R.2 | IRRELEVANT_RE-READING* |
| HC.E/O | HC.E/O.1 | (IR)RELEVANT_RE-READING -> (NAVIGATION) -> WRITE_ESSAY |
| | HC.E/O.2 | GENERAL_INSTRUCTION* /RUBRIC* -> (NAVIGATION) -> WRITE_ESSAY |
| | HC.E/O.3 | WRITE_ESSAY -> WRITE_ESSAY |
| | HC.E/O.4 | WRITE_ESSAY <-> HIGHLIGHT_READING/NOTE_READING |
| | HC.E/O.5 | HIGHLIGHT_LABELLING* |
| | HC.E/O.6 | NOTE_EDITING* |

Legend: "->" means a transition from action A to action B; "<->" means a transition from action A to action B or the other way around; "()" means optional; "*" means one or more consecutive instances of the same action; "/" means either action A or action B

### The improved process library (version 3)

Not all SRL processes from the **data-driven process library** were consistent with our theoretical framework. For example, most of the TIMER actions in trace data (which was also a one-step process) co-occurred (almost 80%) with the *HC.E/O, Elaboration and Organisation* or *LC.F, First-reading* codes in think aloud data. However, based on our theoretical framework, check timer is a meta-cognitive process (mostly *MC.M, Monitoring*) which is captured by learners checking the timer to monitor time left to complete the task. When interpretations of certain action or action patterns guided by the think aloud data conflicted with definitions of certain SRL processes based on our theoretical model (as in the above example), we made comprehensive judgements by combining empirical evidence and theoretical rationale. For example, we found that many learners would not verbally express their monitoring for time (e.g., "now I want to check time left") when they quickly clicked on the timer while they were reading or writing. Therefore, in this step, we comprehensively considered the theoretical assumptions and the think aloud data, and opted for the theoretical assumptions to interpret this action (here, timer) into an SRL process (here, monitoring), although this would sacrifice a certain degree of matching rate with think aloud. In this case, we stuck with our original interpretation that "check timer" is *Monitoring*, which was not strictly driven by the think aloud codes.

After a thorough consideration of each action or action pattern, comparing and combining the empirical evidence from think aloud protocols with assumptions from our theoretical framework, we constructed the **improved process library** (see Table 10).

### Empirical evaluation of the improved process library

In order to evaluate the **improved process library**, we conducted the third alignment between the improved SRL processes extracted from trace data and codes assigned to think aloud data. We used one learner from our sample (the same learner as the one shown in Fig. 6) as an example, and created Fig. 9 as the equivalent of Fig. 6 to illustrate the third alignment results for this learner. The upper half of Fig. 9 shows the first alignment result based on the **theory-driven process library** and the lower half of Fig. 9 shows the third alignment result based on the **improved process library**. The green part of alignment results indicates the measurement results were matched between the analysis of trace data and the results of the coding of think aloud data; and the orange part represents mismatches. As shown in Fig. 9, the match rate between the improved SRL processes extracted from trace data and think aloud codes was higher than the match rate based on the theory-driven SRL processes extracted from trace data (larger green part and smaller orange part).

The third alignment we conducted for learners in the training set (32 participants) provided the results for the second set of the four indicators, and by comparing them with the first set of the four indicators, we were able to evaluate our trace-based SRL measurement protocol and our validation approach. As shown in the left part of Fig. 10, the sensitivity, specificity and match rate for the training set all improved, and the trace coverage remained at a high level (more than 90%). Taking the match rate indicator as an example, the median match rate for the training set improved from 38.97% (based on the theory-driven SRL processes) to 54.24% (based on the improved SRL processes). A series of Mann-Whitney U tests were conducted to determine if our validation approach led to a significance difference in match rates, sensitivity and specificity, and the results showed that our validation approach statistically significantly improved these three indicators in the training set which are: 1) match rate ($U = 203.00$, $r = .27$, $p < .0001$); 2) sensitivity ($U = 107.00$, $r = .10$, $p < .0001$); and 3) specificity ($U = 87.00$, $r = .14$, $p < .0001$). For 25% of the participants in the training set, we achieved a relatively high match rate (higher than 65%), which means the measurement results were highly consistent between the two data channels for these learners. When using the coded think aloud data as the reference point, the improvement of sensitivity, specificity, and match rate indicate that the improved SRL processes extracted

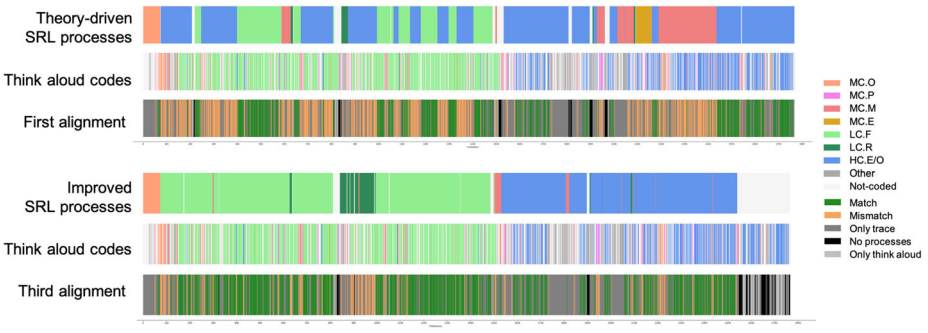


**Fig. 9** The first alignment (upper three rows) and the third alignment (lower three rows), using one learner as an example; Legend: MC.O – Orientation; MC.P – Planning; MC.E – Evaluation; MC.M – Monitoring; LC.F – First-reading; LC.R – Re-reading; HC.E/O – Elaboration and Organisation; Other – Other think aloud codes; Not-coded – No think aloud codes or SRL processes were extracted

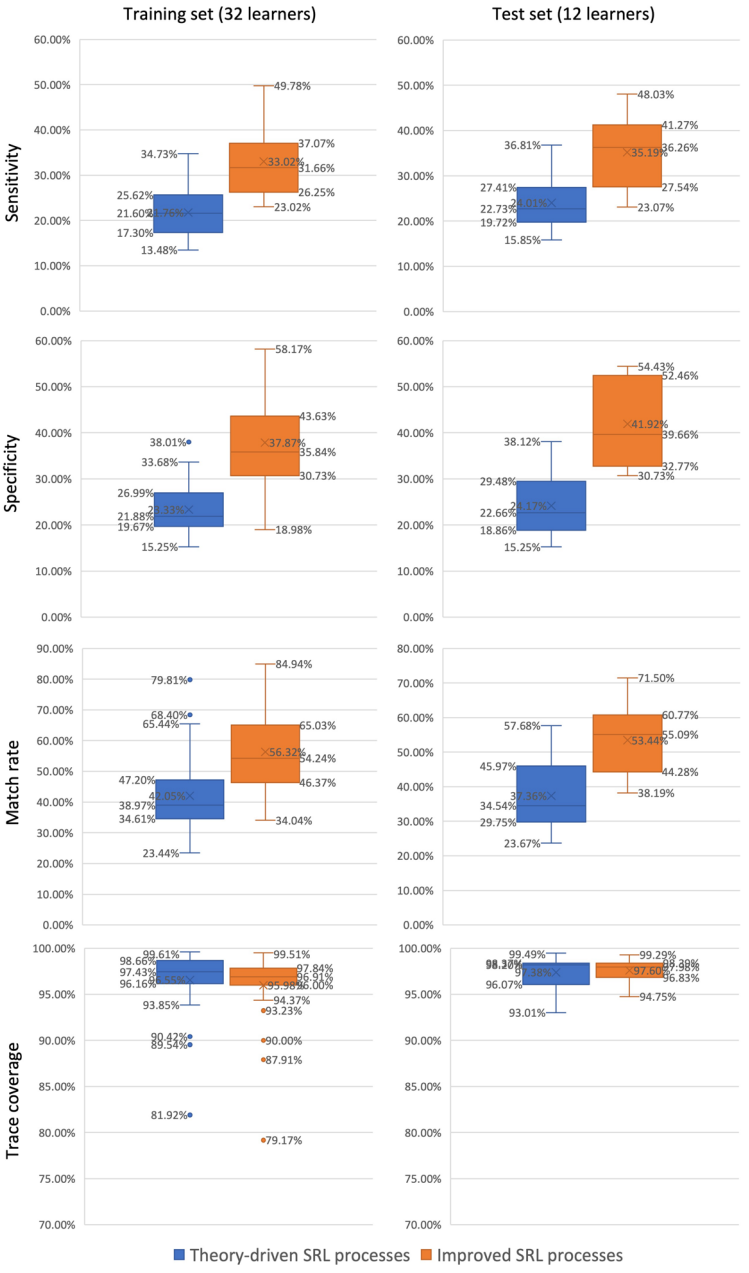


**Fig. 10** Empirical evaluation of validity improvement of the trace-based SRL measurement results against the think aloud data

from trace data were more valid in comparison to the theory-driven SRL processes extracted directly from trace data.

We also found a very similar improvement in the testing set, which was based on the data from the remaining 12 participants (right part of Fig. 10). The sensitivity, specificity and

match rate for the testing set also improved to almost the same extent as the training set, and the trace coverage remained at the same high level. The median match rate for the testing set improved from 34.54% (based on the theory-driven SRL processes) to 55.09% (based on the improved SRL processes). A series of Mann-Whitney U tests were also performed on the testing set, and the results showed that all three indicators were statistically significantly improved: 1) match rate ($U = 20.50$, $r = .38$, $p = .0032$); 2) sensitivity ($U = 17.00$, $r = .12$, $p = .0008$); and 3) specificity ($U = 6.00$, $r = .17$, $p < .0001$).

## Discussion

Our study reports on a novel method for evaluating and investigating the validity of trace-based SRL measurement, based on theory and think aloud protocols. Here, we discuss the value of this study from three aspects: why the validation process is necessary (RQ1), how to validate the trace-based SRL measurement (RQ2) and to what extent can we improve the validity (RQ3).

### RQ1: The necessity of validation process

Our first alignment enabled us to evaluate how a theory-driven trace-based measurement of SRL performs before validation. The first set of sensitivity, specificity and match rate between theory-driven SRL processes extracted from trace data and codes used to analyse think aloud data showed that there was room for validity improvement, relative to the original interpretations based on trace data. This finding exposes some limitations in measurement protocols that have previously been proposed (Siadaty et al., 2016b, c, Saint et al., 2020a, 2021; Fan et al., 2020, 2021) but have not been tested for validity. The relative low sensitivity, specificity and match rate of first alignment pointed out **why** it is problematic to merely rely on the measurement protocol that was constructed based on theory and brainstorming only. The validity challenge in trace-based SRL measurement identified in this and previous studies (Saint et al., 2020a; Winne, 2020, 2014) may encourage more researchers to consider validity in the development of trace-based SRL measurement protocols.

### RQ2: The combination of theory-driven and data-driven approaches

As reviewed in Background Section, a majority of previous studies measured SRL processes using either a data-driven approach (e.g., Maldonado-Mahauad et al., 2018a), or a theory-driven approach (e.g., Siadaty et al., 2016b). To our knowledge, only a few studies extracted SRL processes by applying both approaches (via comparison or combination) (Biswas et al., 2014; Boroujeni & Dillenbourg, 2019), but without systematically validating the measurement results. Here, we first constructed the **theory-driven process library** with the theoretical rationale that underpinned our SRL framework. Then, we generated the **data-driven process library** with the empirical evidence obtained from the analysis of think aloud data. Finally, we combined empirical evidence from think aloud protocols with assumptions from theoretical framework, and constructed the **improved process library**. Our validation approach established a useful methodological pathway and demonstrated the potential to significantly improve the validity of trace-based SRL measurements, which answered RQ2.

To our knowledge, our study is the first to develop a validation approach by considering think aloud data as "reference point". We based this approach on the findings from several previous studies, which concluded that think aloud protocols are more suitable than questionnaires for capturing, analysing, and classifying SRL processes (Veenman, 2007; Bannert, 2007; Azevedo et al., 2010). In our study, three alignments between trace and think aloud data provided the basis for discovering and evaluating our process libraries. However, there are several shortcomings that are associated with think aloud protocols. Firstly, audio recording during learning in an authentic learning setting can be challenging. Secondly, coding think aloud data is very time-consuming and is heavily reliant on well-trained coders. Finally, think aloud protocols can have significant limitations regarding what experiences are sampled and how heterogeneity is averaged in a response (Winne, 2020).

As discussed by Winne (2019), when considering the origins of different tools or instruments and their properties when measuring SRL, "ground truth" can be elusive. Data generated by instruments must be verified (Winne, 2019), and this also applies to data obtained using a think aloud protocol. A previous study highlighted three issues in relation to think aloud protocols including reactivity, verbal acuity, and validity (Young, 2005). The reactivity issue is the challenge for participants to undertake learning and verbalise its mechanisms at the same time (Young, 2005). The verbal acuity issue is the ability of learners to articulate the mechanisms in the form of useful data (Young, 2005). The validity issue is the intrinsic veracity of the verbal utterances and their inferential value (Young, 2005). For instance, in our study, a learner could have navigated back to the instruction page to evaluate whether their current reading was relevant to the task or not; however, this learner may have not been able to articulate this as *Evaluation*. Instead, the verbal utterances of this learner could be coded as *Re-reading* of the instruction page.

In order to improve the "reference point" (i.e., think aloud data), we used previously established and validated methods and paid great attention when collecting and coding think aloud data. For example, we trained the participants of our study to familiarise themselves with the requirements of think aloud before the study, and we also prompted them during the learning process to think aloud to avoid long silences, and we also trained the coders to achieve acceptable coding reliability. However, regardless of paying great attention to its quality, using a think aloud protocol alone will not achieve an absolutely complete or accurate reflection of the whole SRL process. For example, the time slots when learners checked the TIMER were frequently coded as lower-cognition and higher-cognition codes based on think aloud data. However, based on our theoretical framework and the scheme used for coding think aloud data, such events should be mapped under the category of *Monitoring* (meta-cognition). This mismatch was mainly caused by the learners' inability to think aloud about all their SRL processes; this is an example of the reactivity and verbal acuity issues raised in Young (2005). Given these challenges related to think aloud methods, our approach benefited from including theoretical framework to support the analysis.

As stated earlier in this paper, validity is fundamentally dependent on evidence (Messick, 1987). Evidence, in Messick's view (1987), includes both facts from data and a theoretical rationale. The theoretical framework was used in both **theory-driven** and **data-driven** approaches, and enabled us to make sense of the second alignment results and to extract the improved SRL processes. The example about how to interpret "learners' checking on the timer" (in Results Section) is a good demonstration of how we combined the empirical evidence and theoretical rationale when making final interpretations. In summary, both

facts (think aloud codes) and theoretical rationale (our SRL framework) played essential and indispensable roles in improving the validity of the trace-based SRL measurement.

### RQ3: The improvement of validity of trace-based SRL measurements

It is important to note that validity is a matter of degree, not a matter of binary "all or none" (Messick, 1987). Validity always refers to the degree to which empirical evidence and theoretical rationale support the adequacy and appropriateness of interpretations (Messick, 1987). Therefore, in this study, we defined sensitivity, specificity, match rate and trace coverage between SRL processes from trace data and codes used for coding think aloud data as four quantitative indicators to evaluate the "degree" of validity. By comparing the first and the second sets of these four indicators, we achieved a certain degree of improvement for both the training and testing sets (as shown in Fig. 10). By considering think aloud as the "reference point", this improvement of the sensitivity, specificity, and match rate quantified the extent to which validity can be improved by using our validation approach, which answered RQ3.

The effect sizes we obtained in the Mann-Whitney U tests range from small (lowest r = .10) to medium (highest r = .38) (Cohen, 1992). An important question remains as to what match rate (or sensitivity and specificity) ensures an acceptable level for trace-based SRL measurement. First, the mismatches generated from trace data compared to think aloud codes can provide additional and useful information for the trace-based SRL measurement; therefore, it is not reasonable to aim for an "as-high-as-possible" match rate for the validation approach. Second, the limitations of trace data (especially if researchers use navigational logs only) resulted in a certain ceiling for match rate which related to the granularity issue of the trace data. This granularity issue is also related to the sampling rate issue, which is a very important aspect to consider when aligning different data channels. For instance, without eye-tracking data, when a learner's mouse cursor was located in the writing zone and stopped typing for a short while, it was difficult to be certain whether the learner's attention was still within the essay zone or had already been shifted to the reading zone or note taking zone (see Fig. 1) to read more materials or their own notes. If the trace data is not fine-grained enough, these alternate reading and writing behaviours are difficult to be captured, so there will inevitably be a certain degree of mismatch when matching trace data with think aloud data. Therefore, a match rate from 50%-60% might already be an ideal state for trace-based SRL measurement as compared to coded think aloud data. In order to test whether we can further increase the match rate after reaching an approximately 55% match rate, we iteratively performed a second round of the validation approach. We found that the match rate was already difficult to be further improved. However, new trace data channels such as eye-tracking may improve the granularity and the validity of the SRL measurement. To conclude, the body of evidence derived from multi-data channels should not be used to check if each channel can be an alternative for another one, but rather multiple channels should be used as supplements to each other (Winne, 2010) in order to develop SRL research further.

### Implications for related research

In addition to answering the "why" (RQ1), "how to" (RQ2) and "to what extent" (RQ3) questions, the value of this study also lies in raising the "when" questions for validity.

Researchers should pay attention **when** they should examine validity in measuring SRL from trace data. As generally agreed, SRL is characterised as highly contextual (Winne, 2014). We propose that the measurement of SRL is also highly contextual, and thus, one fundamental question remains as to whether the meaning of a measure is context-specific or whether it can be generally applied across contexts (Messick, 1987). The generalizability of trace-based measurement protocol, which includes the action library and the process library, is an important perspective to be considered and discussed in future studies. For example, the SRL process library in this study has a certain degree of generalizability and it can be adjusted and used in other contexts or learning environments; however, researchers should also make appropriate adjustments according to their own learning environment, available learning tools, and types of accessible data.

Whenever a measurement protocol is used in a new context (e.g., a new learning environment, and new task designs), researchers should re-test the validity of their choices that underpin the measurement protocol. For instance, when a new learning environment layout is re-designed, previously valid processes may become undetectable and new SRL processes are most likely shaped. Researchers could use the validation approach (or part of it) proposed in this paper to examine and investigate the validity of their measurement protocol. For example, that may involve collection of think aloud data on a small scale and in an experimental setting to investigate the validity of their trace-based measurement protocol, and then use the optimised trace-based measurement protocol to a large-scale and field setting without think aloud. Once more trace-based measurement protocols that developed in different contexts (such as Siadaty et al., 2016a, b, Saint et al., 2020a, b, 2021, Fan et al., 2020, 2021) are validated using the approach we proposed, researchers in the field of SRL can review the similarity of all the interpretations and start thinking about generally recognised principles in measuring SRL using trace data.

Another implication for related research is that our findings revealed the significance of further triangulation of SRL measurement based on trace data and think aloud data. Although our study used think aloud data as the "reference point" to investigate the trace-based SRL measurement, especially for the interpretation of SRL processes extracted from trace data, it does not mean that mismatches between trace and think aloud equalled to errors in the trace-based SRL measurement. On the contrary, trace data can also provide information that is difficult to capture or code in think aloud data, for instance the TIMER example. Once the trace-based measurement protocols are validated, they could and should be combined with think aloud protocols to measure and describe "a fuller picture" of SRL (Winne, 2010). This potential approach will greatly deepen our understanding of the complex self-regulation process of learners in the real context.

## Conclusion, limitations and future works

In the field of SRL, a view is generally shared that focusing on the analysis of events of SRL processes can sharpen the theory of SRL, and thus potentially elevate the levels of achievement and satisfaction for learners (Winne, 2014). With increasing application of trace data in measuring SRL as events, specific considerations should be given to the validity of conclusions made based on trace data (Winne, 2020). In this study, we have proposed a novel validation approach to evaluating and investigating the validity of trace-based SRL

measurement protocols. Our validation approach includes a **theory-driven perspective** and a **data-driven perspective**, using both empirical evidence from think aloud data and rationale from our theoretical framework of SRL to construct an **improved process library**. More importantly, our results showed that measuring and interpreting SRL from trace data is a very promising method which deserves more attention and practical application.

The findings in this study need to be interpreted with a few limitations in mind. First, we only conducted the study based on a single theoretical framework, using a single dataset which was collected from a specific learning environment. Future studies should test the generalizability of our validation approach using other datasets which were collected using different learning environments, especially using other trace-based measurement protocols and think aloud coding schemes which based on different theoretical framework. For example, the think aloud coding scheme proposed by Greene and Azevedo (2009) was developed based on a different theoretical framework from ours, and therefore, researchers hold different positions and understandings when interpreting the similar learning events. For instance, Greene & Azevedo consider "*Re-reading*" as part of learners' "*Strategy use*", but we classified "*Re-reading*" as "*Low_Cognition*" in this present study. Future research using more diverse database, learning environment and theoretical framework will deepen our understanding of SRL measurement validity, and could test the generalizability of our validation approach. Second, more fine-grained data channels, such as eye-tracking data, are not included in the scope of present study which might limited the validity degree of our trace-based measurement protocol. For instance, if the dwell time between two keyboard strokes is long, it was difficult to determine whether the learner was still conceiving the essay or went back reading material without using eye-tracking data. Future research should focus on integrating new data channels and analysing to what extent can these data channels further improve the validity of the trace-based measurement of SRL. Thirdly, several of our interpretations about action patterns were also limited by the analysis techniques we used. For example, without the use of natural language processing, it is very difficult to distinguish whether learners simple copy and paste information or create an in-depth elaboration in their notes, therefore, it was also difficult to validly interpret different types of patterns in their note usage.

From a methodological point of view, our study calls on future studies to challenge, test, or optimise our validation method, or propose new methods to investigate the validity of trace-based SRL measurement. Future research such as combining evidence based on multi-channel data to achieve cross-validation, or integrating both self-report data and trace data to study SRL warrants further attention (Winne, 2010).

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/s11409-022-09291-1.

**Funding** This study was funded through Call 5 of the Open Research Area (ORA) which is jointly supported by Deutsche Forschungsgemeinschaft (BA20144/10-1), Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO464.18.104), and the Economic and Social Research Council of the United Kingdom (ES/S015701/1).

## Declarations

**Research involving human participants and/or animals** Yes, we collected the learning trace data and think aloud data of human participants. We collected, stored and analysed these data in accordance with ethical approvals.

**Ethics approval (include appropriate approvals or waivers)** The ethical committee in the Faculty of Social Sciences of Radboud University approved the present research.

**Consent to participate (include appropriate statements)** We informed participants about the aim of our study and they were given the opportunity to ask questions, after which they gave active consent to collect data.

**Consent for publication (include appropriate statements)** We have the consent from the ethical committee, the participants and all co-authors to publish this paper.

**Conflict of Interests** The authors declare that there is no conflict of interest that could be perceived as prejudicing the impartiality of the research reported.



## References

Aguera, P.-E., Jerbi, K., Caclin, A., & Bertrand, O. (2011). *ELAN: a software package for analysis and visualization of MEG, EEG, and LFP Signals*. ISSN: 1687–5265 Pages: e158970 Publisher: Hindawi Volume: 2011.

Aleven, V., McLaren, B., Roll, I., & Koedinger, K. (2006). Toward meta-cognitive tutoring: a model of help seeking with a cognitive tutor. *International Journal of Artificial Intelligence in Education*, *16*(2), 101–128.

Azevedo, R., Cromley, J. G., & Seibert, D. (2004). Does adaptive scaffolding facilitate students' ability to regulate their learning with hypermedia?. *Contemporary Educational Psychology*, *29*(3), 344–370.

Azevedo, R., Cromley, J. G., Winters, F. I., Moos, D. C., & Greene, J. A. (2005). Adaptive human scaffolding facilitates adolescents' self-regulated learning with hypermedia. *Instructional Science*, *33*(5-6), 381–412.

Azevedo, R., & Gašević, D. (2019). Analyzing multimodal multichannel data about self-regulated learning with advanced learning technologies: issues and challenges. *Computers in Human Behavior*, *96*, 207–210.

Azevedo, R., Moos, D. C., Johnson, A., & Chauncey, A. (2010). Measuring cognitive and metacognitive regulatory processes during hypermedia learning: issues and challenges. *Educational Psychologist*, *45*(4), 210–223. Publisher: Routledge _eprint: https://doi.org/10.1080/00461520.2010.515934.

Azevedo, R., & Witherspoon, A. (2009). Self-regulated learning with hypermedia. In *Handbook of metacognition in education* (pp. 331–351). Routledge.

Bannert, M. (2007). *Metakognition beim lernen mit hypermedien*. Waxmann Verlag.

Bannert, M., & Mengelkamp, C. (2013). Scaffolding hypermedia learning through metacognitive prompts. In R. Azevedo, & V. Aleven (Eds.) *International Handbook of Metacognition and Learning Technologies, Springer International Handbooks of Education* (pp. 171–186). New York: Springer.

Bannert, M., & Reimann, P. (2012). Supporting self-regulated hypermedia learning through prompts. *Instructional Science*, *40*(1), 193–211.

Bannert, M., Reimann, P., & Sonnenberg, C. (2014). Process mining techniques for analysing patterns and strategies in students' self-regulated learning. *Metacognition and Learning*, *9*(2), 161–185.

Beheshitha, S. S., Gašević, D., & Hatala, M. (2015). A process mining approach to linking the study of aptitude and event facets of self-regulated learning. In *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge - LAK '15* (pp. 265–269). Poughkeepsie: ACM Press.

Biswas, G., Segedy, J. R., & Kinnebrew, J. S. (2014). "A Combined Theory-and Data-Driven Approach for Interpreting Learners' Metacognitive Behaviors in Open-Ended Tutoring Environments. *Design recommendations for intelligent tutoring systems*, *2*, 135.

Boroujeni, M. S., & Dillenbourg, P. (2019). Discovery and temporal analysis of MOOC study patterns. *Journal of Learning Analytics*, *6*(1), 16–33.

Cohen, J. (1992). A power primer. *Psychological bulletin*, *112*(1), 155. Publisher: American Psychological Association.

Dunlosky, J., & Thiede, K. W. (2013). Four cornerstones of calibration research: Why understanding students' judgments can improve their achievement. *Learning and Instruction*, *24*, 58–61.

Fan, Y., Lim, L., van der Graaf, J., Kilgour, J., Engelmann, K., Bannert, M., Molenaar, I., Moore, J., & Gasevic, D. (2020). Measuring Micro-Level Self-Regulated Learning Processes with Enhanced Log Data and Eye Tracking Data.pdf. In *Companion Proceedings 10th International Conference on Learning Analytics & Knowledge (LAK20)* (pp. 433–436).

Fan, Y., Matcha, W., Uzir, N. A., Wang, Q., & Gašević, D. (2021). Learning analytics to reveal links between learning design and self-regulated learning. *International Journal of Artificial Intelligence in Education*.

Fan, Y., Saint, J., Singh, S., Jovanovic, J., & Gašević, D. (2021). A_learning analytic approach to unveiling self-regulatory processes in learning tactics. In *LAK21: 11th International Learning Analytics and Knowledge Conference, LAK21* (pp. 184–195). New York, NY: Association for Computing Machinery.

Gasevic, D., Jovanovic, J., Pardo, A., & Dawson, S. (2017). Detecting Learning Strategies with Analytics: Links with Self-reported Measures and Academic Performance. *Journal of Learning Analytics*, *4*(2), 113–128–113–128.

Günther, C. W., & Rozinat, A. (2012). Disco: discover your processes. *BPM (Demos)*, *940*, 40–44. Publisher: Citeseer.

Greene, J. A., & Azevedo, R. (2010). The Measurement of Learners' Self-Regulated Cognitive and Metacognitive Processes While Using Computer-Based Learning Environments. *Educational Psychologist*, *45*(4), 203–209.

Greene, J. A., Moos, D. C., Azevedo, R., & Winters, F. I. (2008). Exploring differences between gifted and grade-level students' use of self-regulatory learning processes with hypermedia. *Computers & Education*, *50*(3), 1069–1083.

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.

Haggard, P., & Tsakiris, M. (2009). The experience of agency: feelings, judgments, and responsibility. *Current Directions in Psychological Science*, *18*(4), 242–246.

Kinnebrew, J. S., Biswas, G., Sulcer, B., & Taylor, R. S. (2013). Investigating self-regulated learning in teachable agent environments. In R. Azevedo, & V. Aleven (Eds.) *International Handbook of Metacognition and Learning Technologies, Springer International Handbooks of Education* (pp. 451–470). New York: Springer.

Kinnebrew, J. S., Segedy, J. R., & Biswas, G. (2014). Analyzing the temporal evolution of students' behaviors in open-ended learning environments. *Metacognition and Learning*, *9*(2), 187–215.

Kizilcec, R. F., Pérez-Sanagustín, M., & Maldonado, J. J. (2017). Self-regulated learning strategies predict learner behavior and goal attainment in Massive Open Online Courses. *Computers & Education*, *104*, 18–33.

Klug, J., Ogrin, S., Keller, S., Ihringer, A., & Schmitz, B. (2011). A plea for self-regulated learning as a process: Modelling, measuring and intervening.

Maldonado-Mahauad, J., Pérez-Sanagustín, M., Kizilcec, R. F., Morales, N., & Munoz-Gama, J. (2018a). Mining theory-based patterns from Big data: Identifying self-regulated learning strategies in Massive Open Online Courses. *Computers in Human Behavior*, *80*, 179–196.

Maldonado-Mahauad, J., Pérez-Sanagustín, M., Moreno-Marcos, P. M., Alario-Hoyos, C., Mu noz-Merino, P. J., & Delgado-Kloos, C. (2018b). Predicting learners' success in a self-paced MOOC through sequence patterns of self-regulated learning. In V. Pammer-Schindler, M. Pérez-Sanagustín, H. Drachsler, R. Elferink, & M. Scheffel (Eds.) *Lifelong Technology-Enhanced Learning, Lecture Notes in Computer Science* (pp. 355–369). Cham: Springer International Publishing.

Messick, S. (1987). Validity. *ETS research report series*, *1987*(2), i–208. _eprint: https://doi.org/10.1002/j.2330-8516.1987.tb00244.x.

Molenaar, I., van Boxtel, C. A. M., & Sleegers, P. J. C. (2011). Metacognitive scaffolding in an innovative learning arrangement. *Instructional Science*, *39*(6), 785–803. Publisher: Springer.

Mudrick, N. V., Azevedo, R., & Taub, M. (2019). Integrating metacognitive judgments and eye movements using sequential pattern mining to understand processes underlying multimedia learning. *Computers in Human Behavior*, *96*, 223–234.

Munshi, A., Rajendran, R., Ocumpaugh, J., Biswas, G., Baker, R. S., & Paquette, L. (2018). Modeling learners' cognitive and affective states to scaffold SRL in open-ended learning environments. In *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization - UMAP '18* (pp. 131–138). Singapore: ACM Press.

Pintrich, P. R., et al. (1991). A manual for the use of the motivated strategies for learning questionnaire (MSLQ).

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). San Diego: Academic Press.

Pintrich, P. R. (2004). A Conceptual Framework for Assessing Motivation and Self-Regulated Learning in College Students. *Educational Psychology Review*, *16*(4), 385–407.

Reimann, P. (2019). Methodological progress in the study of self-regulated learning enables theory advancement. *Learning and Instruction, 101269*.

Saint, J., Fan, Y., Singh, S., Gasevic, D., & Pardo, A. (2021). Using process mining to analyse self-regulated learning: a systematic analysis of four algorithms. In *LAK21: 11th International Learning Analytics and Knowledge Conference, LAK21* (pp. 333–343). New York: Association for Computing Machinery.

Saint, J., Gašević, D., Matcha, W., Ahmad Uzir, N. A., & Pardo, A. (2020a). Combining analytic methods to unlock sequential and temporal patterns of self-regulated learning. In *Proceedings of the 10th International Conference on Learning Analytics & Knowledge* (p. 10). Frankfurt.

Saint, J., Whitelock-Wainwright, A., Gasevic, D., & Pardo, A. (2020b). Trace-SRL: a framework for analysis of micro-level processes of self-regulated learning from trace data. *IEEE Transactions on Learning Technologies*, 1–1. Conference Name: IEEE Transactions on Learning Technologies.

Siadaty, M., Gašević, D., & Hatala, M. (2016a). Associations between technological scaffolding and micro-level processes of self-regulated learning: A workplace study. *Computers in Human Behavior*, *55*, 1007–1019.

Siadaty, M., Gašević, D., & Hatala, M. (2016b). Measuring the impact of technological scaffolding interventions on micro-level processes of self-regulated workplace learning. *Computers in Human Behavior*, *59*, 469–482.

Siadaty, M., Gasevic, D., & Hatala, M. (2016c). Trace-based micro-analytic measurement of self-regulated learning processes. *Journal of Learning Analytics*, *3*(1), 183–214–183–214.

Taub, M., Mudrick, N. V., Azevedo, R., Millar, G. C., Rowe, J., & Lester, J. (2016). Using multi-level modeling with eye-tracking data to predict metacognitive monitoring and self-regulated learning with crystal island. In A. Micarelli, J. Stamper, & K. Panourgia (Eds.) *Intelligent Tutoring Systems, Lecture Notes in Computer Science* (pp. 240–246). Cham: Springer International Publishing.

Taub, M., Mudrick, N. V., Azevedo, R., Millar, G. C., Rowe, J., & Lester, J. (2017). Using multi-channel data with multi-level modeling to assess in-game performance during gameplay with Crystal Island. *Computers in Human Behavior*, *76*, 641–655.

van der Graaf, J., Lim, L., Fan, Y., Kilgour, J., Moore, J., Bannert, M., Gasevic, D., & Molenaar, I. (2021). Do instrumentation tools capture self-regulated learning? In *LAK21: 11th International Learning Analytics and Knowledge Conference, LAK21* (pp. 438–448). New York: Association for Computing Machinery.

Veenman, M. V. J. (2007). The assessment and instruction of self-regulation in computer-based environments: a discussion. *Metacognition and Learning*, *2*(2), 177–183.

Winne, P. H. (1982). Minimizing the black box problem to enhance the validity of theories about instructional effects. *Instructional Science*, *11*(1), 13–28.

Winne, P. H. (2010). Improving measurements of self-regulated learning. *Educational Psychologist*, *45*(4), 267–276.

Winne, P. H. (2014). Issues in researching self-regulated learning as patterns of events. *Metacognition and Learning*, *9*(2), 229–237.

Winne, P. H. (2019). Paradigmatic dimensions of instrumentation and analytic methods in research on self-regulated learning. *Computers in Human Behavior*, *96*, 285–289.

Winne, P. H. (2020). Construct and consequential validity for learning analytics based on trace data. *Computers and Human Behavior*, (In press).

Winne, P. H., & Hadwin, A. F. (1998). Studying as self-regulated engagement in learning. in metacognition in educational theory and practice. *Metacognition in Educational Theory and Practice*, 277–304.

Winne, P. H., & Hadwin, A. F. (2007). The weave of motivation and self-regulated learning. In *Motivation and Self-Regulated Learning*. Routledge. Num Pages: 18.

Winne, P. H., & Perry, N. E. (2000). Chapter 16 - measuring self-regulated learning. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.) *Handbook of Self-Regulation* (pp. 531–566). San Diego: Academic Press.

Young, K. A. (2005). Direct from the source: The value of 'think-aloud' data in understanding learning.

Zimmerman, B. J. (2000). Chapter 2 - attaining self-regulation: a social cognitive perspective. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.) *Handbook of Self-Regulation* (pp. 13–39). San Diego: Academic Press.



## Affiliations

**Yizhou Fan[1] · Joep van der Graaf[2] · Lyn Lim[3] · Mladen Raković[4] · Shaveen Singh[4] · Jonathan Kilgour[1] · Johanna Moore[1] · Inge Molenaar[2] · Maria Bannert[3] · Dragan Gašević[1,4]**

Joep van der Graaf
j.vandergraaf@pwo.ru.nl

Lyn Lim
lyn.lim@tum.de

Mladen Raković
mladen.rakovic@monash.edu

Shaveen Singh
shaveen.singh1@monash.edu

Jonathan Kilgour
jonathan@inf.ed.ac.uk

Johanna Moore
jmoore@staffmail.ed.ac.uk

Inge Molenaar
i.molenaar@pwo.ru.nl

Maria Bannert
maria.bannert@tum.de

Dragan Gašević
dragan.gasevic@monash.edu

1 School of Informatics, University of Edinburgh, Edinburgh EH8 9AB, UK

2 Behavioural Science Institute, Radboud University, Nijmegen 6500 HE, The Netherlands

3 TUM School of Education, Technical University of Munich, Munich 80333, Germany

4 Faculty of Information Technology, Monash University, Clayton, Victoria 3800, Australia