Studies in Educational Evaluation 71 (2021) 101090

Contents lists available at ScienceDirect

Studies in Educational Evaluation

journal homepage: www.elsevier.com/locate/stueduc

The dynamic relationship between response processes and self-regulation in critical thinking assessments

Heidi Hyytinen [a,*], Jani Ursin [b], Kaisa Silvennoinen [b], Katri Kleemola [a,b], Auli Toom [a]

[a] Centre for University Teaching and Learning, Faculty of Educational Sciences, The University of Helsinki, Finland
[b] Finnish Institute for Educational Research, The University of Jyväskylä, Finland

ARTICLE INFO

Keywords:
Performance-based assessment
Critical thinking
Response processes
Self-regulation
Undergraduate students
Qualitative analysis

ABSTRACT

Our aim was to explore higher education students' response and self-regulatory processes plus the relationship between these, as evidenced in two types of performance-based critical thinking tasks included in the *Collegiate Learning Assessment (CLA+) International* instrument. The data collection consisted of 20 cognitive laboratories. The data were analyzed using a qualitative approach. The tasks were found to trigger different response and self-regulatory processes. Overall, the performance task evoked more holistic processes than the selected-response questions, in which students' processes were more question-oriented. The results also indicated the entanglement of students' response and self-regulation processes. Three self-regulation groups were identified. Students with *versatile* self-regulation skills were able to complete the task thoroughly, whereas students with *moderate* self-regulation skills faced challenges in monitoring and evaluating their performance. Students who were *lacking* in self-regulation struggled both with the task as a whole and their own progress. Implications for higher education are discussed.

1. Introduction

There is a growing need to investigate the quality of higher education students' critical thinking skills, since these have been shown to be important for students' learning and study success in higher education, and in working life (Arum & Roksa, 2011; Tuononen et al., 2019). Critical thinking is conceptualized as a demanding and multifaceted capacity covering a combination of a set of cognitive skills and the affective dispositions to use these skills (Hyytinen, Toom, & Shavelson, 2019; Ennis, 2015; Facione, 1990; Halpern, 2014). It involves a purposeful and self-regulatory act of thinking, which results in interpretation, analysis, evaluation, inference, and found explanation (Ennis, 2015; Halpern, 2014). Despite a widespread consensus on the importance of learning critical thinking, there is evidence that higher education students differ significantly in their ability to think critically (Arum & Roksa, 2011; Badcock et al., 2010; Evens et al., 2013).

Research has frequently relied on self-reporting methods in assessing students' critical thinking (Shavelson, 2010; Zlatkin-Troitschanskaia et al., 2015). However, due to the breadth and complexity of the construct of critical thinking, the reliability of self-report methods has been questioned (Karabenick et al., 2007; Pekrun, 2020; Zlatkin-Troitschanskaia et al., 2015). Hence, there has been an increasing emphasis on performance-based assessments, namely the complex assessment tasks that evoke authentic performance covering aspects of critical thinking through situations that resemble the real world (Shavelson et al., 2018; Zahner & Ciolfi, 2018). Performance-based assessments allow students to demonstrate their knowledge and skills measured in the assessment task (Shavelson, 2010). Such assessments can be used to provide empirically-based feedback to students, and they form a data source for the development of critical thinking throughout degree program curricula.

In order to provide adequate and valid feedback, it is necessary to identify the response processes triggered by performance-based assessments, and to obtain empirical evidence on the constructs assessed (Ercikan & Oliveri, 2016; Ercikan & Pellegrino, 2017). To be valid, it is essential that the assessments should tap into the intended constructs (American Educational Research Association, 2014; Ercikan & Pellegrino, 2017; Messick, 1995). Previous methodological research on performance-based assessment has focused on students' responses and test scores (e.g., Hyytinen, Nissinen, Ursin, Toom, & Lindblom-Ylänne, 2015; Gu et al., 2018; Zlatkin-Troitschanskaia et al., 2019) or on experts' perceptions concerning the usefulness of the test instruments (e.g., Beck, 2020; Zlatkin-Troitschanskaia et al., 2018). These studies have mostly

* Corresponding author at: Centre for University Teaching and Learning, University of Helsinki, 00014, University of Helsinki, Finland.
E-mail address: heidi.m.hyytinen@helsinki.fi (H. Hyytinen).

https://doi.org/10.1016/j.stueduc.2021.101090
Received 18 September 2020; Received in revised form 23 August 2021; Accepted 25 September 2021
Available online 6 October 2021

been quantitative in nature (Kleemola, Hyytinen, & Toom, 2021; Davey et al., 2015). However, little is known about the characteristics of the response processes triggered by different types of performance-based assessments (Cook et al., 2014; Ercikan & Pellegrino, 2017; Leighton, 2017a). Thus, there is a need for qualitative research on cognitive validity that goes beyond traditional psychometric analyses of response patterns, as noted by Ercikan & Oliveri (2016).

In performance-based tasks, responding involves not only certain cognitive skills, but also an ability to interpret the task, and to intentionally adapt and monitor strategies for solving the task (Hyytinen & Toom, 2019; Zimmerman & Campillo, 2003). Previous research has shown that self-regulation plays a significant role when one is working on complex tasks (Beckman et al., 2021; Winne, 2018). Even though self-regulation has been found to be crucial for critical thinking and for the performance of complex tasks (Halpern, 2014; Lau, 2015), there have been few reports about research investigating the self-regulation skills applied in such tasks. This indicates a need to explore students' response and self-regulatory processes as they occur during assessment, and how these processes are related to each other.

This study focused on the response and self-regulation processes triggered by one instrument, namely the *Collegiate Learning Assessment (CLA+) International for Higher Education.* The CLA + International is a performance-based assessment which measures critical thinking (i.e., analysis, problem solving, reasoning, critical reading and evaluation, written communication). It consists of (a) a performance task and (b) a set of document-based selected-response questions (Zahner & Ciolfi, 2018). The study shed additional light on the validity of the CLA + International (cf. Aloisi & Callaghan, 2018). Investigation of the response processes provided insights into the problematic combination of different task types (Kleemola et al., 2021; Zahner & Ciolfi, 2018). By identifying the qualitative variation in response and self-regulatory processes that occur during task completion, and how the task characteristics are associated with these processes, this study was able to outline the strengths of the different tasks, and also the modifications needed to develop further performance-based assessments of critical thinking. The findings provide important insights into the nature of critical thinking and its connections with self-regulation skills, with a potential for achieving more targeted instructions and tasks within higher education.

2. Theoretical framework

2.1. Performance-based assessments for exploring students' critical thinking

Critical thinking is conceptualized as a purposeful, goal directed self-regulatory judgment about what to believe and do in a certain situation (Hyytinen et al., 2019; Ennis, 2015; Halpern, 2014). It is a combination of complex cognitive skills (i.e., analysis, reasoning, problem solving, argumentation and evaluation of information) together with relevant affective dispositions (i.e., open-minded, flexible in evaluation) to use these skills (Braun et al., 2020; Ennis, 2015; Facione, 1990; Halpern, 2014). The definition of critical thinking highlights that self-regulation is an essential part of critical thinking; it is a function that guides this complex thinking process (Halpern, 2014; Lau, 2015). Critical thinking makes it possible to assess, evaluate, synthesize, and interpret relevant knowledge associated with a situation, and to apply that knowledge to solve a problem, to decide on a course of action, to find an answer to a given question, or to reach a well-reasoned conclusion (Hyytinen & Toom, 2019; Ennis, 2015; Shavelson et al., 2018).

Indirect methods and materials, such as self-reports of learning, have often been applied in investigations of critical thinking (Shavelson, 2010). However, due to the complex multidimensional nature of the phenomenon, critical thinking is extremely difficult to capture solely via indirect measurements (Karabenick et al., 2007; Pekrun, 2020; Zlatkin-Troitschanskaia et al., 2015). To draw more valid inferences on students' performance, it is necessary to use measurements that address and assess the students' actual abilities (Davey et al., 2015; Ercikan & Oliveri, 2016; Ercikan & Pellegrino, 2017; Ercikan & Por, 2020; McClelland, 1973). This is the rationale for a comprehensive performance-based approach that would cover critical thinking in an authentic manner (Braun et al., 2020; Shavelson et al., 2018).

Performance-based critical thinking assessments simulate real-life decision-making situations, and require students to make and justify their decisions by utilizing the available evidence (Shavelson, 2010; Shavelson et al., 2019). In the best case, performance assessments constitute real learning situations for students. They arouse students' interest in the task, engage them in the task, and invite them to use higher order thinking skills and intellectual capacities in solving the task (Kane et al., 2005; Shavelson, 2010; Zlatkin-Troitschanskaia et al., 2015). Thus, through performance-based critical thinking assessments, students can demonstrate what they know and are able to do in tasks that simulate authentic problem-solving situations (Hyytinen & Toom, 2019; Shavelson et al., 2019). In completing the tasks, the students assess, evaluate, synthesize, elaborate, and use evidence from multiple knowledge sources instead of merely accumulating facts.

Performance-based assessment is an umbrella term covering a variety of task types, such as performance tasks (PTs) and selected-response questions (SRQs), in addition to brief constructed-response items. According to Messick (1994, p.3), these different performance-based assessments form "a continuum representing different degrees of structure versus openness in the allowable responses." This means that in SRQs, students analyze a question and materials, and using these as the basis, select correct answers from a list of options. The challenge in SRQs is that guessing or eliminating incorrect options without using critical thinking skills is always possible (Aloisi & Callaghan, 2018; Hyytinen et al., 2015). Moreover, SRQs are often designed so that they focus a single skill (Shavelson, 2010). PTs, for their part, are open-ended tasks in which students need to analyze, evaluate, and synthesize complex information, and further, to provide reasoned explanations, in written form (Davey et al., 2015; Shavelson, 2010; Shavelson et al., 2018). Thus, rather than focusing on the individual components of critical thinking skills, PTs require the integration of several skills in a holistic manner, for example, by analyzing a range of documents and completing a task based on these documents (Shavelson et al., 2019; Zlatkin-Troitschanskaia et al., 2019).

We know from the previous studies that PT and SRQs highlight different aspects of critical thinking (Aloisi & Callaghan, 2018; Hyytinen et al., 2015; Messick, 1994; Perie, 2020; Shavelson, 2010). To take an example, through SRQs we cannot assess students' ability to build arguments as these questions primarily guide the students to select the best option among the responses offered to them. Therefore, SRQs do not necessary invite students to utilize higher order thinking skills (Braun et al., 2020). Within these two types of performance-based critical thinking assessment (PTs and SRQs), SRQs have been the dominate assessment format in higher education (Zlatkin-Troitschanskaia et al., 2019). Earlier work on validity of performance-based critical thinking assessments has often been quantitative focusing on statistical issues and psychometric characteristics of assessment (Kleemola et al., 2021; Davey et al., 2015; Perie, 2020; Zlatkin-Troitschanskaia et al., 2019). So far, very little attention has been paid to response processes activated by PT and SRQs intending to assess critical thinking. In the present study, applying a qualitative analysis, we focused on students' response and self-regulation processes in the PT and in the SRQs used in the CLA + International instrument.

2.2. Response processes in performance-based assessments

In examining the validity of a performance-based assessment, it is necessary to investigate the characteristics of the response processes triggered by the assessments, particularly when the object of interest is a complex construct such as critical thinking (Ercikan & Oliveri, 2016; Ercikan & Pellegrino, 2017). Response processes refer to the processes of

acquiring and operating with knowledge (Winne, 2018), plus the strategies and behaviors that students use in undergoing the assessment (Ercikan & Pellegrino, 2017). They are considered to provide important potential evidence of validity in performance-based tasks (American Educational Research Association, 2014; Ercikan & Pellegrino, 2017)

To gain a deep understanding of response processes in a performance-based assessment, several aspects need to be considered. Investigations of response processes involve not only students' response strategies, but also their interpretation of tasks, the skills they use, their effort and engagement, and the knowledge and resources they use while performing the task (Ercikan & Pellegrino, 2017). Response processes arise from students' reciprocal interaction with the task (Brückner & Pellegrino, 2016) and thus, both student and task characteristics need to be considered (Hyytinen & Toom, 2019). For instance, different task types aiming to measure the same construct (i.e., PT that invites students to generate their answer versus SRQs that ask students to define and select) may tap into different response processes (Ercikan et al., 2015; Perie, 2020; Shavelson, 2010), or the content of task may be interpreted differently due to the complex concepts and terminology by subsections of the intended population (American Educational Research Association, 2014; Karabenick et al., 2007; Mislevy et al., 2013).

Various methods can be used to study response processes, the aim being to pinpoint processes relevant to the construct at hand (Nichols & Huff, 2017). When the aim of the task is to assess the quality of students' critical thinking, the assumption is that the assessment will trigger response processes representing the construct of critical thinking (American Educational Research Association, 2014; Brückner & Pellegrino, 2016; Kane et al., 2005). In exploring the validity of performance-based assessments, expert reviews are frequently used, even though they have been found to be inconsistent with the cognitive evidence (Cook et al., 2014; Ercikan et al., 2015). Few studies have analyzed response processes in ways that would lead to a profound understanding of the mechanisms of the assessment (Cook et al., 2014; Ercikan & Pellegrino, 2017; Leighton, 2017a). This is especially important if one is seeking to adapt an assessment to a culture and language that differs from its original version (Solano-Flores & Chia, 2017).

2.3. Self-regulation skills are essential in responding to complex tasks

As indicated above, it is not only cognitive skills that are important to bear in mind in assessing critical thinking; in fact, self-regulation also needs to be considered (Halpern, 2014; Lau, 2015). Self-regulation refers to an intentional and adaptive process that allows students to plan, adapt, and monitor their thoughts, emotions, and behaviors to the demands of the task (Beckman et al., 2021; Schunk & Greene, 2018; Zimmerman, 2002; Zimmerman & Campillo, 2003). Several studies have reported considerable variety in the self-regulation skills of university students (Beckman et al., 2021; Räisänen et al., 2016), and self-regulation has also been shown to vary in different learning contexts (Saariaho, Pyhältö, Toom, Pietarinen, & Soini, 2016).

Previous research has shown that students use self-regulation processes to monitor their levels of understanding, and adapt their learning processes in an ongoing manner to achieve set goals (Schunk & Greene, 2018; Zimmerman & Campillo, 2003). Self-regulation has been found to take place across three phases, which are (a) task analysis, goal setting, and strategic planning before taking any action, (b) using a range of methods, monitoring, and observing learning during the performance/action, and (c) evaluating and reflecting during and/or after the action (Usher & Schunk, 2018; Zimmerman, 2002). This process has been characterized as being cyclical in nature (Schunk & Greene, 2018; Usher & Schunk, 2018); moreover, the three phases are not linear; rather, they are tightly intertwined, with students going back and forth between the phases (Pintrich, 2004).

Students with well-developed self-regulation skills are able to analyze the task ahead by considering what will be required for successful action, dividing complex tasks into manageable components, setting goals, and identifying the approaches and strategies that will be needed to accomplish the task (Beckman et al., 2021; Toering et al., 2012; Zimmerman & Campillo, 2003). While performing the task, the individual needs to be able to monitor and observe cognition, behaviors, and emotions within a given context, and to modify and adapt strategies according to the demands of the task. During and after the performance phase, self-regulation skills are needed to assess, evaluate, and review one's efforts, strategies, and behaviors to decide whether their performance met the goals. This prompts a new behavioral cycle (Usher & Schunk, 2018). The more complex a task, the higher is the cognitive and metacognitive effort needed to complete the task (Beckman et al., 2021; Hoyle & Dent, 2018; Winne, 2018). As an example, students with well-developed self-regulation skills are in a better position to perform the complex open-ended task than the students who have problems in self-regulation (Beckman et al., 2021).

Even though self-regulation skills have been found to be crucial for critical thinking (Halpern, 2014; Lau, 2015) and for completion of complex tasks, such as open-ended tasks (Beckman et al., 2021; Zimmerman & Campillo, 2003), there has been surprisingly little empirical research into self-regulation and its interconnections with the response processes of students when they read, interpret, and formulate solutions to performance-based critical thinking tasks. The present study sought to obtain data that would address this research gap.

2.4. Aims of the study

The purpose of this study was to examine the variety in students' response processes and in their self-regulation skills, plus the relationships between these aspects, in two types of performance-based assessment tasks in CLA + International, utilizing cognitive laboratory data (Leighton, 2017a, 2017b). The following research questions were set:

1 What response processes do PT and SRQs evoke among students?
2 To what extent do the phases of self-regulation emerge in completing PT and SRQs?
3 To what extent do the phases of self-regulation associate with response processes in the context of the assessment tasks?

3. Materials and methods

3.1. Context of the study

As of 2020, the Finnish higher education system consists of 13 research-intensive universities and 22 universities of applied sciences. The basic task of the research-intensive universities is to engage in scientific research and to provide the highest level of education on that basis. Universities of applied sciences are regional higher education institutions whose activities relate to their connection with working life and regional development. In 2019, research-intensive universities had around 153 000 students, while universities of applied sciences had around 140 000 students (Vipunen, 2021). Higher education institutions in Finland are autonomous actors, responsible for the content of their education and research, and for the development of their own activities. In both types of higher education institutions, students can complete bachelor's degrees and master's degrees, but only research-intensive universities can award third-cycle postgraduate degrees. Each higher education institution decides on which students it admits, and on the criteria for admission. Higher education leading to a degree is tuition-free, with the exception of students coming from countries that are not members of the European Union or the European Economic Area. For the present study, two higher education institutions were selected: one university of applied sciences, and one research-intensive university. Both institutions are large multidisciplinary institutions within the metropolitan area of Finland.

3.2. Participants

The study was conducted with 20 first- and second-year students (12 female, 8 male) drawn by purposeful sampling from two Finnish higher education institutions in spring 2019. The participants represented both genders, and both official languages in Finland (Finnish and Swedish). Their age varied from 22 to 55 years, with a mean of 29 years. They were slightly more often female and older than average students. According to Vipunen (2021), 54 % of higher education students in Finland are females and 46 % are males. Mean age of higher education students is 28 years. The participants were studying in programs representing various disciplines (see Table 1).

The aim of the recruiting process was to select a representative sample of key informants representing various fields of study. In both participating institutions, an open invitation was sent for students to participate in the study. Out of the volunteers, the researchers selected 20 participants to represent as many fields of study as possible. Participation in the research was voluntary, and informed consent was obtained from the participants. All participants were informed about the nature, duration, and purpose of research. The cognitive lab procedures were conducted in such a way as to respect the participants' requests, without asking personally sensitive questions. The lab events were recorded only if the participants gave permission. The data were analyzed so that individual participants could not be identified. Non-monetary incentives (i.e., cinema tickets) were given to the participants.

3.3. The CLA + International

The study used the Collegiate Learning Assessment International (CLA + International) instrument. CLA + International has been developed by the US-based Council for Aid to Education (CAE) (Klein et al., 2007; Shavelson, 2010; Zahner & Ciolfi, 2018). Outside the United States (USA), CLA + International has been adapted for use in Italy, the United Kingdom (UK), and Germany (Shavelson et al., 2018; Zahner & Ciolfi, 2018; Zlatkin-Troitschanskaia et al., 2018). CLA + International is a performance-based assessment that measures critical thinking skills at the tertiary level. It is administered online and includes one PT and 25 SRQs. The PT requires students to generate a written response to a given real-life scenario. The prompt expects students to integrate analytical reasoning, problem solving, argumentation, and written-communication skills, consulting materials in a Document Library, and using these to formulate a response. The source documents in this study included five documents: a blog, a podcast transcript, a research memorandum, a newspaper article, and an infographic. When students write a response to the scenario, they need to justify their decision/recommendation and provide reasons and evidence against any opposing argument(s). Students have 60 min to respond to this open-ended question (for an example of a PT see Tremblay et al., 2012, 220–236).

Each SRQ presents students with four options (here termed *items*) and asks them to choose a single answer. The SRQs are arranged within three separate sets, each focusing on a different skill area. The first measures scientific and quantitative reasoning, the second measures critical reading and evaluation, and the third measures students' ability to identify logical fallacies and questionable assumptions. As in the PT, each SRQ set is document-based. Students are instructed to use the source documents (either one or two per set) when preparing their responses. Students have 30 min to complete the SRQs.

The adaptation and translation of the CLA + followed the International Translation Committee's (ICT) guideline for translating and adapting tests (International Test Commission, 2018). The CLA + International used in this study was translated into Finnish and Swedish, which are the two main official languages of Finland. The adaptation and translation of both language versions of the tasks included four phases. Firstly, the test instruments were translated from English into the target language. Then, two translators (who had knowledge of English-speaking cultures but whose mother tongue was the primary language of the target culture) independently checked and confirmed the translations. Subsequently, the research team reconciled and verified the revisions. The translations were then pretested in cognitive labs, with final modifications embedded as necessary. The use of cognitive labs with think-alouds made it possible to ensure that the translation and adaptation process had not altered the meaning or difficulty of the task (Leighton, 2019).

In analyzing an assessment, one must consider the sources of its validity (American Educational Research Association, 2014). The present study adds to this understanding. So far, there has been little research on the validity of CLA + International (Aloisi & Callaghan, 2018). Studies on its predecessor, CLA, indicate that the results of the assessment correlate with other assessments with similar goals (Arum & Roksa, 2011; Klein et al., 2007; Pascarella et al., 2011). However, studies on similarly constructed assessments indicate potential problems pertaining to CLA + International, notably regarding the combination of PT and SRQ. The different task types capture different skills, and combining them may cause incoherence in the interpretation of the assessment (Aloisi & Callaghan, 2018; Hyytinen et al., 2015; Kleemola et al., 2021; Davey et al., 2015). Furthermore, while PT is often considered capable of capturing the complex nature of critical thinking (Shavelson, 2010; Shavelson et al., 2019), it favors students with strong writing skills, and may be correspondingly less able to capture thinking *per se* (Aloisi & Callaghan, 2018). In addition, the construct, and especially the definition of critical thinking underpinning CLA+, has been labeled as obscure (Aloisi & Callaghan, 2018).

Table 1
Characteristics of the Participants (RIU = Research-Intensive University, UAS = University of Applied Sciences).

Code of student	Gender	Age	Discipline	Type of HE
ID1	male	26	Arts	RIU
ID2	female	23	Arts	RIU
ID3	male	29	Science	RIU
ID4	female	28	Science	RIU
ID5	male	26	Arts	RIU
ID6	female	22	Social science	RIU
ID7	male	30	Social science	RIU
ID8	female	22	Social science	RIU
ID9	female	22	Science	RIU
ID10	female	22	Social science	RIU
ID11	male	23	Science	RIU
ID12	female	55	Health/welfare	UAS
ID13	female	37	Health/welfare	UAS
ID14	female	48	Health/welfare	UAS
ID15	female	22	Services	UAS
ID16	male	36	Engineering	UAS
ID17	male	34	Arts	UAS
ID18	female	26	Arts	UAS
ID19	female	30	Arts	UAS
ID20	male	22	Science	RIU

3.4. Data collection

The data collection consisted of a sample of 20 cognitive lab events with think-alouds and follow-up interviews (Leighton, 2017a, 2017b, 2019). This method can be used to investigate differences between participants in response processes and self-regulation that may influence test performance (Leighton, 2017b; van Someren et al., 1994). The think-aloud method makes it also possible to gather authentic data on a participant's ongoing thinking processes while they are working on a task (Leighton, 2019).

The data were collected individually from all the participants in three phases. In Phase 1, the purpose of the cognitive lab was explained to the participants, and they were trained to think aloud via some small exercise tasks, such as "please tell me how many doors there are in your home". The participants were also asked to fill in an informed consent form, and were told that they could withdraw from the study at any

stage. In Phase 2, the participants proceeded to the actual think-aloud phase of the study; at this point they first completed the PT and thereafter 25 SRQs in a secured online environment. The procedure applied a neutral form of the think-aloud protocol (Leighton, 2017b; van Someren et al., 1994), in which the participants were not interrupted while they were performing the PT and SRQs. The participants were asked to utter every thought that occurred to them as they worked through the task, uninterrupted and unedited. If the participant became silent for a long time, the investigator reminded them to "please keep talking" or "please remember to verbalize" (Leighton, 2017b, 2019). In Phase 3, the participants were interviewed individually concerning their thinking and response processes during the tasks. This included general questions posed to all participants, and also tailored questions arising out of the think-alouds. After ten cognitive labs we learned that data saturation (i.e., new data repeated what was expressed in previous data collection situations) was attained (Saunders et al., 2018). Each cognitive lab event was video- and audio-recorded, and transcribed verbatim. The first- and second-named authors collected the data.

3.5. Data analysis

The data were examined with two goals in mind: (a) to understand the kinds of response processes triggered by the different task types, and (b) to explore participants' self-regulation processes during completion of the tasks, using a qualitative abductive approach. Abduction refers to an analytical process that produces a new interpretation of the phenomenon studied combining data observations from different materials and the theoretical understanding based on previous studies (Timmermans & Tavory, 2012). Both methodological (i.e., the use of several data collection methods) and investigator triangulation (i.e., all authors participated in data analysis) was used (Denzin, 2012). The non-linear data analysis consisted of four overlapping phases. In the first phase, the verbatim transcriptions of the cognitive lab data were read through. Content logs were created for each video by the third-named author, with precise descriptions and summaries of events systematically recorded within the logs. Content logs made it possible to combine both verbal and non-verbal response data, and to encompass all participant inputs (i.e., assertions, behavior, and actions within the think-aloud process). The content logs externalized and visualized participants' thinking processes and behaviors associated with the test constructs, and their progress in the tasks (see Table 2). The logs also included information on the sequencing, timing, and variety of the participants' response behaviors and actions (Oranje et al., 2017).

The second phase of the analysis was that of coding. The coding features were negotiated jointly by the first- and fifth-named authors. They first coded the entire dataset. The dataset consisted of transcribed think-alouds and interviews, content logs, and participants' responses. We analyzed these different materials simultaneously following the course of the test situation several times. This phase was guided by theory, as the coding of the data involved response processes (Ercikan & Pellegrino, 2017) and self-regulation (e.g., Beckman et al., 2021; Schunk & Greene, 2018; Zimmerman, 2002). The coding related to the response processes focused on (a) the effort and time the students put into completing the task, i.e., the activity needed to proceed and progress in the tasks; (b) how the students read and interpreted the documents and questions, i.e., the details and comprehensive understanding of the information and documents; (c) the strategies the students used in responding to the tasks, i.e., the steps they took in solving the tasks in relation to the instructions and goals of the tasks; (d) the knowledge and resources they used to carry out the tasks, i.e., the sources utilized by students, and how they combined these; (e) the skills used by students in completing the tasks in relation to the instructions and goals of the tasks (Ercikan & Pellegrino, 2017).

The coding related to students' self-regulation concentrated on three main phases of self-regulation (Beckman et al., 2021; Schunk & Greene, 2018; Zimmerman, 2002), namely:

Table 2
An Extract from a Content Log.

Time	Duration	Code of student: ID2
0:00:00–0:00:27	0:00:27	Logging into the test plus the privacy notice Glances through the privacy notice and asks how to move on. Asks the same also at the summary of the test. Glances through the summary of the test.
0:00:27–0:08:14	0:07:47	Performance task
0:00:27–0:01:08	0:00:41	Reads and glances through the general instructions for the performance task.
0:01:08–0:08:14	0:07:06	Moves to the actual performance task. First, quickly reads the task instruction and some of the documents. Moves directly to writing the answer, does not plan it beforehand. Browses the documents. Concentrates on the infographic. Using that as a basis, says that the physical activity habits of the residents should be improved. Does not substantiate the answer more precisely, compare the information in the documents, or evaluate the reliability of their content aloud. Completes the answer in no more than eight minutes and moves on to the SRQ items. Written response: *The physical activities of Reagan's inhabitants should be improved. The inhabitants must be told about a healthy diet. The education level must be improved.*
0:08:14–0:59:52	0:51:38	SRQs
0:08:14–0:10:39	0:02:25	Moves to the SRQ section. Browses through the SRQ instructions. Asks for help on how to move on. Glances at the first question and items, then the document provided for the first SRQ set. Moves back to the first question and items. Then identifies the relevant section from the source document.
0:10:39–0:15:32	0:04:53	Compares the items to the document. Thinks aloud which item (A–D) would most weaken the main claim of the document. Says that option A could be true based on the document, hence A is *not* the right answer. Selects option D. Moves to the second question. — —

1 task-analysis, planning, and setting goals (entailing identification of the characteristics of the task, forethought, anticipating the next steps and progress, and what would be expected in terms of completion of the task);
2 self-monitoring the progress in the task, controlling one's own thinking and behavior in the situation, and using different strategies;
3 evaluation, meaning reflection on the task situation, plus evaluation of one's own process and progress, and success in completing the task.

These different aspects were searched for and coded systematically and simultaneously across the entire data set and within the data items, with reference to the content logs, the transcribed think-alouds, and the interviews with each person. Coding was continued until no new aspects emerged from data (i.e. the point at which saturation was obtained). The trustworthiness of the coding was checked by the third- and fourth-named authors, who carefully followed up on the coding process, and recoded the dataset. The final coding was jointly discussed and negotiated by all the authors until consensus was reached to ensure the reliability of the findings.

Thereafter, for each student, a written account was made, containing brief descriptions that touched on the qualities of the response processes and the self-regulation of learning. After that, we explored all the data to identify the similarities and differences in each student's dataset, giving possibilities for further insights into the nature of the coded material.

During the analysis, we found that the ways in which the phases of self-regulation emerged and how extensively they were related to each other varied among the students. Based on this variation, the students were placed into three groups. The groups were then further distinguished by exploring the differences and similarities between them, in such a way that each group included a particular combination of self-regulation phases and response processes that was sufficiently distinct

from the other groups. This fourth phase of analysis was conducted by the first and fifth-named authors, but the detailed final description and interpretation was discussed between all five authors.

During the first three phases, the basic unit of analysis was the individual student. In the fourth phase, the groups formed the units. The phases are visualized in Fig. 1. All the data extracts presented in this article have been translated from Finnish and Swedish into English.

4. Findings

4.1. Response processes triggered by the tasks

In line with Ercikan and Pellegrino (2017), we understood students' response processes as encompassing (a) the effort and time put into completing a task, (b) reading and interpretation of the documents and questions, (c) response strategies, (d) knowledge and resources, and (e) the skills used to carry out the tasks. The analysis revealed that PT and SRQs activated different processes. The main characteristics of the response processes triggered by the tasks are summarized in Table 3. Below, we consider these in more detail, elaborating each aspect of response processes separately.

4.1.1. Effort and time put into completing the tasks

Students perceived both the PT and the SRQs as demanding, albeit interesting. In general, both task types maintained the students' effort and interest throughout. While most students (n = 15) put maximum effort into completing both tasks, two students did not invest in the tasks at all. There were also three students who put their effort mainly into the SRQs rather than the PT. In the SRQs, six students appeared to lose concentration while answering. Most of the students (n = 14) were unable to complete all the SRQs within the allotted time. By contrast, only three students were unable to complete the PT in the given time.

4.1.2. Reading and interpretation of the documents and questions

In the PT, the students read the instructions and the documents. They related the information from the various sources to the instructions, and compared the information provided by the sources. This was particularly the case when they started the task, but also when they progressed further. By contrast, in the SRQs, the students focused on the content of the source documents in relation to the questions. Thus, their reading was more fragmented than for the PT.

Table 3
The Main Characteristics of the Response Processes Triggered by the Tasks.

	Performance task	Selected-response questions
Effort and time – how the students in the sample performed	Mostly maintained throughout the task	Mostly maintained throughout the task
		Some lost their concentration while completing the task
	Most students completed the task in the given time	Most students did not complete the task in the given time
Reading and interpretation (task requirements)	Comprehensive, covering the whole task	Question-oriented
Response strategies (task requirements)	Holistic and thorough: taking notes, structuring the answer, evaluating the answer	Question-oriented and outcome-oriented: identifying details, defining and selecting the response
	Varied	Less variation
Knowledge and resources (task requirements)	Comprehensive use of knowledge presented in the documents	Use of knowledge required in the question
	Comparison of knowledge provided in the source documents	Use of prior knowledge
	Comparison with prior knowledge	
Skills and task completion (task requirements, and overall nature of the task)	Using and combining a variety of skills	Limited repertoire of skills
		Question-oriented
	Holistic	
	Varying in skills and cognitive demands	Less variation

4.1.3. Response strategies

The students' strategies varied in the PT. Many students (n = 12) read the instructions carefully, made notes while reading and interpreting the documents, and structured their answer accordingly. Some of these students (n = 8) also evaluated their response in relation to the documents and instructions while writing their response. In contrast, some students (n = 6) started to answer while reading or browsing the documents. Two participants generated their answer without reading all the documents. In responding to the SRQs, the students tended to be

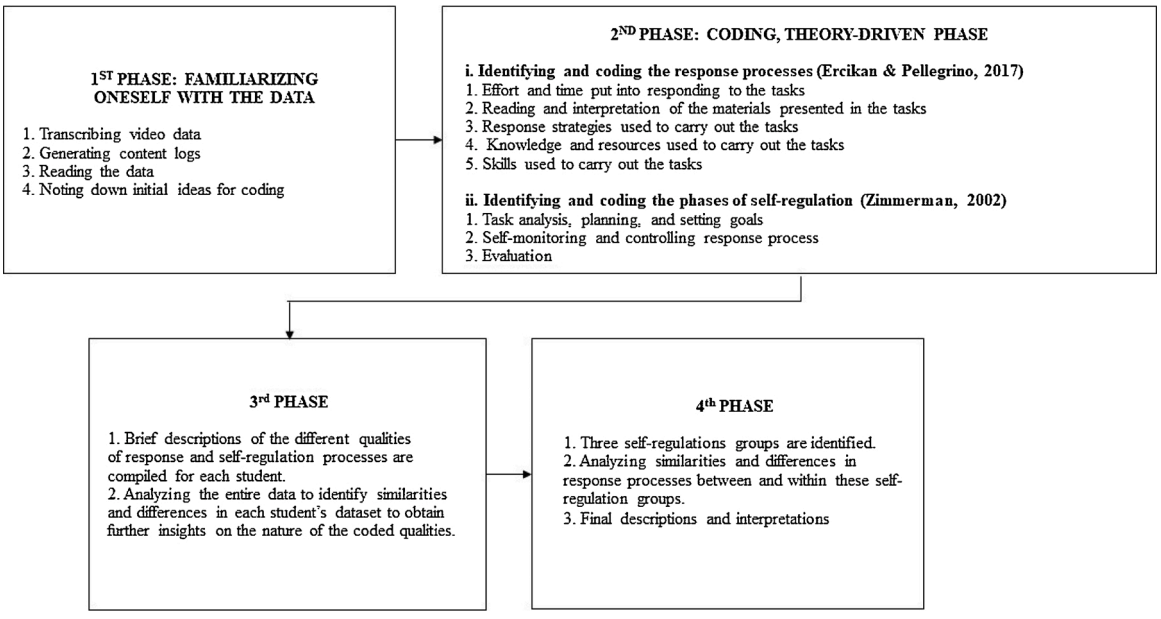

Fig. 1. The Four Main Phases of the Analytical Process.

more question-oriented and outcome-oriented than for the PT. At the start of each set of questions, the students ($n = 18$) often glanced at the first question and the source document(s). Thereafter, they compared the items (i.e., options) for each question with the source document(s), and in relation to each other, selecting their response on that basis. They continued in this way, following the set order. All the students were aware that there was just one correct item from the given alternatives. Thus, they sought to identify it and eliminate the wrong items.

4.1.4. Knowledge and resources

Differences between the students emerged in how they used knowledge and resources in performing the tasks. In the PT, some students ($n = 8$) referred intensively to all the documents in the Document Library, whereas others ($n = 7$) selected the most relevant documents and knowledge on the basis of their analysis. The latter used the information provided by the podcast transcript, the research memo, and the infographic in their answer, excluding the blog text and the newspaper article. Some students also compared the knowledge presented in the documents with their prior knowledge. A few students ($n = 5$) provided minimal analysis of the content of the documents or did not analyze the documents at all. These students reproduced some details from the documents, or else their response was based on their own opinion. Eight students processed all the information thoroughly and wrote comprehensive written responses. In contrast, the remaining students were likely to omit – often knowingly – essential information (e.g., contradictions presented in the documents or aspects students viewed as self-evident), and this led to a fragmented written response. In the SRQs, the students' use of resources was typically more question-oriented. Most students ($n = 18$) first glanced at the source document(s) for the set of questions. Thereafter, they identified the relevant sections from the document(s), and utilized the information provided therein to select their response. Two students did not read or interpret the source documents and selected the answer solely by guessing or on the basis of their prior knowledge.

4.1.5. Skills and task completion

Students adopted a variety of skills in constructing the response and in completing the tasks in relation to the instructions and goals. Some students ($n = 5$) browsed through the documents, many ($n = 15$) read them in detail, some made notes for constructing the response, and some revisited the documents while they responded. Our analysis indicated that the tasks triggered different skills. In contrast to the SRQs, the PT evoked more holistic problem-solving processes. Thus, to deal with the PT thoroughly, students needed to integrate several skills. For example, they had to read several documents, to combine, evaluate, elaborate, and synthesize the information from these documents, and to apply the information in order to reach a conclusion and formulate explanations regarding the questions presented in the task.

In the SRQs, each question required students to start the problem-solving process again. As noted above, at the start of each set of questions, most of the students ($n = 18$) first glanced at the source document(s). After that, they identified the sections related to each question (plus related items) and analyzed the sections in more detail. They then compared the items with each other and with the document (s) and selected their response accordingly. Thus, the SRQs allowed students to select their response by a process of elimination, without encouraging evaluation, elaboration, or synthesis of the information in the documents. Students sometimes arrived at the solution by a process of statistical reasoning or logical deduction (including identification of fallacies), or by identifying flawed assumptions within in the items – or simply by guessing.

4.2. Self-regulation evoked by the tasks

Analysis revealed differences in students' self-regulation while completing the tasks. Based on these differences, students were classified into three groups, namely (1) *demonstrating versatile self-regulation,* (2) *demonstrating moderate self-regulation,* and (3) *lacking in self-regulation*. Depending on the demands of the tasks, the groups showed variation in the quality of the task interpretations, in how they planned and set goals, in how they self-monitored progress in the tasks, in their thinking, strategies, and behavior in the situation, and in their evaluation of their performance. Table 4 summarizes this variation in the self-regulation phases.

In the *versatile self-regulation group*, the students (8 out of 20 students) thoroughly interpreted the goals of the tasks before attempting to answer. Their interpretation of the tasks demonstrated a coherent understanding of the overall purpose and key ideas of the tasks. They deliberately planned, monitored and evaluated their thoughts and behaviors according to the demands of the tasks throughout completion of the tasks, especially in the PT. They asked themselves questions concerning the problems and the documents, and what was needed to solve them. They reviewed their work and their answers all the time during completion of the tasks. They also kept working, even on difficult questions and materials. The different phases of self-regulation were present throughout the task. The phases were tightly intertwined, with the students going back and forth smoothly between the phases of self-regulation. One student described how he self-monitored the progress in the PT as follows:

> First of all, I looked at them all [documents] and analyzed them a little, then I came back to each document in more detail, and then I started analyzing and identifying the most important things and then, like, compared these with each other and tried to find different perspectives. Based on that I started to consider recommendations in one direction and another. (ID5, interview)

In the *moderate self-regulation group*, the students ($n = 10$) aimed to understand the content of the tasks before they attempted to answer the questions. Most of them interpreted the key information and identified the most relevant knowledge in relation to the tasks or questions. However, most of the students ($n = 8$) in this group had difficulties in monitoring their performance while completing the tasks. One student described his situation in the following way:

> I forgot what I was doing, and so I needed to move back and forth [between the instructions and my answer]. This is so typical of me. I easily just start answering with what is nice to write, not what I'm asked to do. (ID7, interview)

Many of the students in this group felt frustrated. Six students expressed frustration while responding to the SRQs, and two students in relation to the PT. Furthermore, the students in this moderate group

Table 4
Summary of Variation Identified in Students' Self-Regulation.

Group	Self-regulation
1. Versatile self-regulation ($n = 8$)	• demonstrated task-analysis • planned and set goals • self-monitored • evaluated performance according to the demands of the tasks, especially in the PT • The different phases of self-regulation were present throughout the task; the phases were tightly intertwined.
2. Moderate self-regulation ($n = 10$)	• demonstrated task-analysis • might plan and set goals • might try self-monitoring, but struggle with it • did not evaluate performance • Self-regulation of learning was relatively rigid; the students did not move fluently between the phases of the regulation.
3. Lacking in self-regulation ($n = 2$)	• did not demonstrate task analysis • did not plan and set goals • did not monitor • did not evaluate performance • The various self-regulation phases were not present.

typically did not evaluate their performance or their answers. Self-regulation of learning was inflexible.

In the *lacking in self-regulation group*, the students ($n = 2$) tried to complete the tasks by relying on their intuition or previous experience, rather than planning, monitoring, or evaluating their cognitions or emotions in relation to goals. They struggled with completion of the tasks, and they felt very frustrated. These students emphasized that they did not read or interpret the documents. Nor did these students plan their course of action to solve the problems presented in the tasks:

> I don't have the strength to read these instructions [scans through the instruction and the documents of the constructed-response task]. Now, I don't understand what I must do. "Write your answer." Why? This sounds difficult somehow. You just have to write this kind of stuff in the classroom, and I just never have the energy to do that. I'm sorry but do not have the strength to do this, this is taking me so much time now. (ID14, think-aloud)
>
> One doesn't really have the energy to read this [the document for the first set of SRQs]. So, just like that, I select this and that. This item that sounds the most difficult must surely be the correct one. (ID14, think-aloud)
>
> I didn't plan or read, I just picked up the main things [from the documents for both tasks]. (ID19, interview)

4.3. Associations with self-regulation and response processes

As mentioned above, there was variation in the response processes used by students. The students who demonstrated *versatile self-regulation* skills were able to adapt their thinking, skills, strategies, and approaches to the task goals. In the PT, these students thoroughly analyzed and synthesized the knowledge presented in the documents in relation to the problem-solving situation, and to their own prior knowledge. However, they were aware that they were asked to carry out the tasks using only the documents provided. The students in this group created their own understanding of the problem-solving situation and generated their answer on the basis of a thorough analysis of the documents. They identified the key ideas presented in the source documents. They evaluated the reliability of the knowledge, and refuted contradictory evidence (i.e., the students noted the contradictions in the documents and pointed these out).

In the SRQs, the students in the versatile self-regulation group thoroughly analyzed the document(s) at the start of the task, and throughout completion of the responses. They usually selected their answer by eliminating the wrong items in relation to the source document(s). However, they also selected their response by comparing the given items with each other, and also with the document(s), utilizing reasoning, applying logical deduction, or identifying fallacies or flawed assumptions. They concentrated fully and kept working even on difficult items, using the time allowed. These students made a guess only if they did not arrive at a solution, for example if they perceived two of the given alternatives as correct. In both tasks, they evaluated their work to see if their answers made sense in the context of the task, and they put their best efforts into completing both tasks. They worked persistently until the time ended.

In the *moderate self-regulation* group, most of the students demonstrated a thorough understanding of the content of the documents in both tasks. Typically, they first identified the most relevant knowledge pertaining to the situation or question. In the PT, seven students in this group analyzed the documents provided for the task. Based on their analysis, they excluded the unreliable knowledge, and generated their answer using only the most relevant knowledge presented in the documents. Three students, for their part, first roughly selected the documents according to their relevance to the situation, excluding more than half of the documents. Thereafter, they based their answer on their analysis of the remaining documents. Half of students in this group refuted the contradictory information.

Two students in this group indicated that at first, they did not know how to approach the PT, and thus tried to find more instructions on how to formulate the answer. These two students found the PT more challenging than the SRQs:

> How long should the report actually be? Is it instructed somewhere? Hmm, no. Lots of text here [in the documents]. What is here? Statistics, I could write some kind of report based on these, or give some recommendations, but I think I should read these other documents too. [break in thinking aloud, reads instructions]. Recommendations, arguments supporting recommendations, and address alternative recommendations. Okay [browses documents]. Okay, so there are these three or four documents, which include something a little bit different and also the same issues, and then statistics, but the references in these documents are strange. So, it means I have to figure out how to start constructing this fact-based report. I'll browse the instructions once again to see if there are some tips on the response… [revisits the instructions]. Okay, now I have challenges in structuring the report in a rational way and in what I write here… [starts writing the response; at the same time browses and reads the documents]. I find it difficult to find the facts. I cannot write these kinds of things in the report, because there are no facts about it. Okay, how do I start the report then? [break, thinks] Yeah, so, now I'll try to include the essential information in this report, but there is so much vague information. (ID16, think-aloud)

In the SRQs, the students with moderate self-regulation skills first read the question items and then browsed or read the source document (s). They then identified and interpreted the relevant section from the document (dealing with the issues pertaining to the question). Some expressed a need for a search function, which would have helped them to quickly locate relevant words from the document. However, the search function was blocked, so they had to concentrate on the whole document. At the start of the task, students put most effort into completing the SRQs. However, some ($n = 6$) lost interest towards the end of the task. The students in this group mainly selected their answer by eliminating the items in relation to each other and to the document (s). However, they also selected the answer by reasoning, identifying fallacies and flawed assumptions, and making a guess.

Students who *lacked self-regulation skills* did not adapt their skills, strategies, or approaches according to the task requirements. In the PT, they generated answers using prior knowledge or directly copying a single fact from one document, without processing. The students did not demonstrate understanding of the documents. In the SRQs, they selected the first answer that came to their mind, or simply guessed. They did not make a serious effort to analyze, evaluate, or interpret any of the documents; hence, they misrepresented the fragmented pieces of knowledge in both tasks. They did not use all the time allocated to complete the tasks. The characteristics of the response processes among the groups are illustrated in Table 5.

5. Discussion and conclusions

5.1. Findings in the light of previous literature

We aimed to provide new qualitative information on the characteristics of response processes and self-regulation that would go beyond traditional psychometric analyses of student responses, and hence, to present new insights into the validity of the CLA + International. The results indicated that response and self-regulatory processes are both based on students' reciprocal relationship with the task. Due to the features of the two task types, the tasks triggered a range of response processes among the students (Hyytinen et al., 2015; Ercikan & Pellegrino, 2017; Messick, 1994; Shavelson, 2010). In the PT, the students holistically elaborated two wide-ranging questions, giving possible alternative answers, plus justifications. In the SRQs, by contrast, each

Table 5
Characteristics of the Response and Self-regulatory Processes by which the Students Approached the Tasks.

Group	Response processes	Performance task	Selected-response task
1. Versatile self-regulation (n = 8)	Effort	Maintained effort throughout the task.	Maintained effort throughout the task.
	Reading and interpretation, knowledge and resources	Read and interpreted all the documents several times. Demonstrated a thorough understanding of the documents, identified the key ideas presented in the documents, refuted contradictory knowledge, evaluated the quality of the knowledge. Used prior knowledge in the analysis.	Read the document(s) throughout completion the task. Interpreted and analyzed the question and items and compared them with the document(s) and each other.
	Response strategies, skills	Based on the comprehensive analysis and understanding of the documents, planned and wrote an answer. Gave a reasoned explanation of the problem. Checked and evaluated the answer according to the demands of the tasks several times and modified it, so that it followed the instructions and included all necessary aspects.	Solved the question item by item. Selected the answer by eliminating the items in relation to each other and the document(s); also sometimes by reasoning, logical deduction, or identifying fallacies. If did not arrive at a solution, made a guess. Moved to the next question and started the process again. Persisted with the questions in the set order; might revisit the previous question and items pertaining to it.
2. Moderate self-regulation (n = 10)	Effort	Mainly maintained effort throughout the task.	Put effort into completing the task. Might lose concentration towards the end of the task.
	Reading and interpretation, knowledge and resources	Read or glanced at the documents. Aimed to understand the content of the documents. Based on the analysis, identified and selected the most relevant knowledge and excluded unreliable documents. Might evaluate the quality of information and refute contradictory information.	Read or glanced at the question and the document(s). Identified and interpreted the most relevant sections related to the question.
	Response strategies, skills	Might plan an answer. Might checked the instructions, if necessary, before started to write the answer. Wrote an answer by using the most relevant knowledge and facts. Did not evaluate or check the answer.	Solved the question item by item. Mainly selected the answer by eliminating the items in relation to each other and the document(s). Might also select the answer by reasoning, identifying fallacies, and making a guess. Moved to the next question and started the process again. Completed the questions in the set order.
3. Lacking in self-regulation (n = 2)	Effort	Did not make a serious effort to complete the task.	Did not make a serious effort to complete the task.
	Reading and interpretation, knowledge and resources	Did not read all the documents; glanced through some of them, omitted most of the source documents, and misrepresented knowledge; used prior knowledge	Glanced through the first document, read questions and items in the set order.
	Response strategies, skills	Generated a very short answer by using prior knowledge or copied a single fact from one document without processing.	Selected the solutions to the questions by randomly guessing or by using prior knowledge. Did not complete the task.

question required its own process. This tended to lead to a situation in which students' response processes were interrupted after each question and started again with each new question. Both tasks mostly maintained students' interest throughout, especially among those students who demonstrated versatile self-regulation. The tasks, especially the PT, offered learning situations to students, and invited them to use various critical thinking skills (Kane et al., 2005; Shavelson et al., 2018).

In the PT, students typically maintained their effort throughout the task, and combined a variety of skills when completing the task. In the SRQs, students aimed to define and select the correct response. The SRQs were often answered via lower-level processing, such as selection by eliminating the obviously wrong items (cf. Braun et al., 2020). Some students simply tried to guess the correct alternative when completing the SRQs (cf. Aloisi & Callaghan, 2018). However, in this study, some students occasionally selected the answer by applying logical deduction, reasoning, or identifying fallacies and flawed assumptions. Some students seemed to lose concentration while working on the SRQs. Most of the students were unable to complete the SRQs in the time given. Moreover, in the SRQs, the use of documents was more restricted and fragmented than was the case with the PT, which entailed a more comprehensive use of the documents provided. In addition, the SRQs entailed outcome- and question-oriented strategies, while the response strategies in the PT were more holistic in nature and varied more among the students. This study supports the evidence from earlier observations, which have suggested that SRQs do not capture the complex nature of critical thinking (Braun et al., 2020; Kane et al., 2005; Shavelson, 2010).

The tasks did not distinguish students merely in terms of the response processes but also in respect of their self-regulation activities, which involved task analysis, strategy formation and monitoring, evaluation, and reflection on their performance in solving the tasks. In line with previous studies (Hoyle & Dent, 2018; Winne, 2018), the current study found that to complete the tasks – which were indeed challenging – self-regulation was crucial. The holistic and open-ended PT in particular involved setting up aims and goals, applying and evaluating skills, strategies, the use of knowledge resources, and the effort needed to achieve the goals in question. In so doing, the students had to monitor their progress and make the necessary adjustments (Lau, 2015). By contrast, the SRQs did not provide students with opportunities for comprehensive regulation, since the problem-solving process was more fragmented and question-oriented than in the PT. The differences between the task types in terms of response and self-regulation processes may explain why students perform differently in tasks that are intended to trigger the same construct (Hyytinen et al., 2015; Cook et al., 2014; Ercikan & Pellegrino, 2017; Kane et al., 2005; Messick, 1994; Perie, 2020). The differences would appear to threaten the coherence of the assessment and should considered in developing the assessment.

Interestingly, the students who participated in this study showed relatively good self-regulation skills. The students with versatile self-regulation skills completed the tasks thoroughly and with determination, monitoring their thinking and performance in order to adapt their thinking processes to the demands of the tasks and their prior knowledge. All the phases of self-regulation were present, especially in the PT, and the phases intertwined with one another. In this respect, our findings are in line with those obtained in earlier studies that highlight intertwinement of the phases self-regulation (Saariaho et al., 2016; Pintrich, 2004; Usher & Schunk, 2018). However, students with moderate self-regulation skills did indeed demonstrate task-analysis; nevertheless, they did not evaluate their performance in relation to the demands of the task. They also faced challenges in monitoring their performance, and they expressed frustration in either the PT or the SRQs. These students seemed to be more fact-oriented than the students in the versatile group, who more often demonstrated very good comprehension of the documents. Additionally, the students with low self-regulation skills struggled both with the tasks and their own progress (cf. Beckman et al., 2021;

Zimmerman & Campillo, 2003). Their answers were based on their own opinion, and they did not make a serious effort to complete the tasks. The results reported here shed light on the entanglement of response and self-regulatory processes among the students. Variety in self-regulation was associated with the increased complexity of the response processes (Hoyle & Dent, 2018). Similarly, previous research has shown that there is a relationship between the quality of students' self-regulation, task completion, effort and engagement (Beckman et al., 2021), suggesting that students who are able to analyze and articulate task goals in open-ended task and based on that make their own decision about how they approach task are more engaged to complete the task.

5.2. Practical implications

In line with previous literature (Hyytinen et al., 2015; Braun et al., 2020; Shavelson, 2010), this study found that PT activates a variety of cognitive skills such as analytical reasoning, problem solving and argumentation in a holistic manner. However, the results also indicated some challenges related to the use of PT. In the PT, many students knowingly omitted important points from their written responses, including contradictions they observed in the documents, or aspects they viewed as self-evident. In other words, these students' analyses and thinking processes were more versatile than their written answers. These limitations must be considered when students' performances are evaluated solely on the basis of their responses or test scores.

This study has educational significance in identifying the dynamic interplay between response processes and self-regulation in critical thinking assessment situations. It is known from prior research that not all higher education students are automatically competent in critical thinking (Arum & Roksa, 2011; Badcock et al., 2010; Evens et al., 2013); thus, critical thinking skills need to be practiced, and the acquisition of the skills needs to be supported in a variety of ways throughout higher education (Hyytinen et al., 2019). Critical thinking involves self-regulation skills, as the data here indicate. The teaching of critical thinking skills should thus be expanded so that it explicitly enhances self-regulation in the use of these skills (Lau, 2015). The development of regulation is supported when teachers provide students with concrete advice and support in how to set goals, plan, monitor and reflect on their learning (Beckman et al., 2021; Räisänen et al., 2016, 2020; Winne, 2018). Students who have challenges in self-regulation would benefit from the kinds of instructions, tasks, and feedback that would help them to monitor and reflect on their learning processes.

The results of this study indicate that special attention needs to be paid to the characteristics of tasks that truly enhance the learning of critical thinking and self-regulation (cf. Beckman et al., 2021). Teachers could enhance both critical thinking and self-regulation by giving students complex open-ended tasks that require goal directed and self-regulatory act of thinking to assess the reliability and relevance of information, to recognize biases and to find explanations for the task. The complexity of the open-ended task triggers students' self-regulation: they need to analyze the information, consider alternative strategies in addressing the assessment questions, and reflect on their success and performance throughout the task (Beckman et al., 2021; Braun et al., 2020). Once students have opportunities to become aware of their critical thinking and self-regulation skills, they will have more options for enhancing such skills. In an open-ended task, students are engaged and challenged in terms of their own learning (Kane et al., 2005; Shavelson, 2010; Zlatkin-Troitschanskaia et al., 2015).

5.3. Limitations and methodological reflections

A limitation of this study was the small number of participants, with the risk of bias in the data. It is possible that the students who participated in our study were more competent than those who did not volunteer for the study. However, for the purposes of this study the low sample size is not necessarily a problem. The aim was to examine the variety in students' response and self-regulatory processes in the two types of critical thinking assessment included in CLA + International, rather than to generalize from a small sample to a target population. Notwithstanding the limited sample, the study sheds light on the qualitative variation in the constructs in question, and their interconnections.

Another limitation of this study relates to the data collection situations. On the one hand, the data collection situations – together with the tasks and instructions – were interventions; they could thus have prompted students to apply their critical thinking skills and be more self-regulated, putting more effort into the tasks than they would have otherwise (Beckman et al., 2021). On the other hand, the data collection situation, in conjunction with the think-aloud method, was highly challenging for students, with the cognitive and self-regulatory effort increasing in accordance with the complexity of the situation (Hoyle & Dent, 2018; Winne, 2018). It should also be noted that in the data collection situations, the students first completed the demanding PT and thereafter answered the difficult SRQs. The effort of completing the PT could well have been found by students to be exhausting. This, in turn, could have adversely affected the students' response and self-regulatory processes in the SRQs.

We were also aware that the students' ability to articulate their thoughts might itself influence the quality of the think-aloud data. To avoid bias in the data collection, the students were informed what was meant by thinking aloud, and they were given some initial training in thinking aloud. It has been reported that even a short exercise task helps participants to understand the think-aloud procedure (Leighton, 2017b, 2019). Note also that the neutral nature of the think-aloud protocol ensured that the probing questions were not asked until the follow-up interview, i.e., after the students had completed the task.

In future, it would be useful to examine students' response and self-regulatory processes in critical thinking assessment situations with a larger dataset, and in different contexts. Other unexplored aspects of students' critical thinking and self-regulation might be identified. Studies along these lines would also facilitate the continuing development of performance-based assessments of complex constructs.

Acknowledgements

This research was funded by FinnishMinistry of Education and Culture. The authors would like to thank the Council for Aid to Education for providing the CLA+International. The funder and the CAE had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

Aloisi, C., & Callaghan, A. (2018). Threats to the validity of the Collegiate Learning Assessment (CLA+) as a measure of critical thinking skills and implications for Learning Gain. *Higher Education Pedagogies, 3*(1), 57–82. https://doi.org/10.1080/23752696.2018.1449128

American Educational Research Association (Ed.). (2014). *Standards for educational & psychological testing*. American Educational Research Association.

Arum, R., & Roksa, J. (2011). *Academically adrift: Limited learning on college campuses*. University of Chicago Press.

Badcock, P. B. T., Pattison, P. E., & Harris, K.-L. (2010). Developing generic skills through university study: A study of arts, science and engineering in Australia. *Higher Education, 60*(4), 441–458. https://doi.org/10.1007/s10734-010-9308-8

Beck, K. (2020). Ensuring content validity of psychological and educational tests – the role of experts. *Frontline Learning Research, 8*(6), 1–37. https://doi.org/10.14786/flr.v8i6.517

Beckman, K., Apps, T., Bennett, S., Dalgarno, B., Kennedy, G., & Lockyer, L. (2021). Self-regulation in open-ended online assignment tasks: The importance of initial task interpretation and goal setting. *Studies in Higher Education, 46*(4), 821–835. https://doi.org/10.1080/03075079.2019.1654450

Braun, H. I., Shavelson, R. J., Zlatkin-Troitschanskaia, O., & Borowiec, K. (2020). Performance assessment of critical Thinking: Conceptualization, design, and implementation. *Frontiers in Education, 5*, 156. https://doi.org/10.3389/feduc.2020.00156

Brückner, S., & Pellegrino, J. W. (2016). Integrating the analysis of mental operations into multilevel models to validate an assessment of higher education students' competency in business and economics: Integrating the analysis of mental

operations. *Journal of Educational Measurement, 53*(3), 293–312. https://doi.org/10.1111/jedm.12113
Cook, D. A., Zendejas, B., Hamstra, S. J., Hatala, R., & Brydges, R. (2014). What counts as validity evidence? Examples and prevalence in a systematic review of simulation-based assessment. *Advances in Health Sciences Education, 19*(2), 233–250. https://doi.org/10.1007/s10459-013-9458-4
Davey, T., Ferrara, S., Holland, P. W., Shavelson, R. J., Webb, N. M., & Wise, L. L. (2015). *Psychometric considerations for the next generation of performance assessment*. Center for K-12 Assessment & Performance Management, Educational Testing Service. https://www.ets.org/Media/Research/pdf/psychometric_considerations_white_paper.pdf.
Denzin, N. K. (2012). Triangulation 2.0. *Journal of Mixed Methods Research, 6*(2), 80–88. https://doi.org/10.1177/1558689812437186
Ennis, R. H. (2015). Critical thinking: A streamlined conception. In M. Davies, & R. Barnett (Eds.), *The Palgrave handbook of critical thinking in higher education* (pp. 31–47). Palgrave Macmillan US. https://doi.org/10.1057/9781137378057.
Ercikan, K., & Oliveri, M. E. (2016). In search of validity evidence in support of the interpretation and use of assessments of complex constructs: Discussion of research on assessing 21st century skills. *Applied Measurement in Education, 29*(4), 310–318. https://doi.org/10.1080/08957347.2016.1209210
Ercikan, K., & Pellegrino, J. W. (2017). Validation of score meaning using examinee response processes for the next generation of assessments. In K. Ercikan, & J. W. Pellegrino (Eds.), *Validation of score meaning for the next generation of assessments: The use of response processes* (pp. 1–8). Routledge.
Ercikan, K., & Por, H. H. (2020). Comparability in multilingual and multicultural assessment contexts. In A. Berman, E. Haertel, & J. Pellegrino (Eds.), *Comparability of large-scale educational assessments: Issues and recommendations* (pp. 205–225). National Academy of Education. https://doi.org/10.31094/2020/1.
Ercikan, K., Seixas, P. C., Lyons-Thomas, J., & Gibson, L. (2015). Cognitive validity evidence for validating assessments of historical thinking. In K. Ercikan, & P. C. Seixas (Eds.), *New directions in assessing historical thinking* (pp. 206–220). Routledge/Taylor & Francis Group.
Evens, M., Verburgh, A., & Elen, J. (2013). Critical thinking in college freshmen: The impact of secondary and higher education. *International Journal of Higher Education, 2*(3), 139. https://doi.org/10.5430/ijhe.v2n3p139
Facione, P. (1990). Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction (The Delphi Report). *Educational Resources Information Center (ERIC)* (pp. 1–112).
Gu, L., Liu, O. L., Xu, J., Kardonova, E., Chirikov, I., Li, G., Hu, S., Yu, N., Ma, L., Guo, F., Su, Q., Shi, J., Shi, H., & Loyalka, P. (2018). Validating the use of translated and adapted HEIghten ® Quantitative Literacy Test in Russia. In O. Zlatkin-Troitschanskaia, M. Toepper, H. A. Pant, C. Lautenbach, & C. Kuhn (Eds.), *Assessment of learning outcomes in higher education: Cross-national comparisons and perspectives* (pp. 267–284). Springer International Publishing. https://doi.org/10.1007/978-3-319-74338-7.
Halpern, D. F. (2014). *Thought and knowledge: An introduction to critical thinking* (5th ed.). Psychology Press.
Hoyle, R. H., & Dent, A. L. (2018). Developmental trajectories of skills and abilities relevant for self-regulation of learning and performance. In D. H. Schunk, & J. A. Greene (Eds.), *Handbook of self-regulation of learning and performance* (2nd ed., pp. 49–63). Routledge, Taylor & Francis Group.
Hyytinen, H., Nissinen, K., Ursin, J., Toom, A., & Lindblom-Ylänne, S. (2015). Problematising the equivalence of the test results of performance-based critical thinking tests for undergraduate students. *Studies in Educational Evaluation, 44*, 1–8. https://doi.org/10.1016/j.stueduc.2014.11.001.
Hyytinen, H., & Toom, A. (2019). Developing a performance assessment task in the Finnish higher education context: conceptual and empirical insights. *British Journal of Educational Psychology, 89*(3), 551–563. https://doi.org/10.1111/bjep.12283
Hyytinen, H., Toom, A., & Shavelson, R. J. (2019). Enhancing scientific thinking through the development of critical thinking in higher education. In M. Murtonen, & K. Balloo (Eds.), *Redefining scientific thinking for higher education: Higher-order thinking, evidence-based reasoning and research skills* (pp. 59–78). Palgrave Macmillan.
International Test Commission. (2018). ITC guidelines for translating and adapting tests (2nd ed.). *International Journal of Testing, 18*(2), 101–134. https://doi.org/10.1080/15305058.2017.1398166
Kane, M., Crooks, T., & Cohen, A. (2005). Validating measures of performance. *Educational Measurement: Issues and Practice, 18*(2), 5–17. https://doi.org/10.1111/j.1745-3992.1999.tb00010.x
Karabenick, S. A., Woolley, M. E., Friedel, J. M., Ammon, B. V., Blazevski, J., Bonney, C. R., Groot, E. D., Gilbert, M. C., Musu, L., Kempler, T. M., & Kelly, K. L. (2007). Cognitive processing of self-Report items in educational research: Do they think what we mean? *Educational Psychologist, 42*(3), 139–151. https://doi.org/10.1080/00461520701416231
Kleemola, K., Hyytinen, H., & Toom, A.. https://doi.org/10.1080/02602938.2021.1946482.
Klein, S., Benjamin, R., Shavelson, R., & Bolus, R. (2007). The collegiate learning assessment: Facts and fantasies. *Evaluation Review, 31*(5), 415–439. https://doi.org/10.1177/0193841X07303318
Lau, J. Y. F. (2015). Metacognitive education: Going beyond critical thinking. In M. Davies, & R. Barnett (Eds.), *The Palgrave handbook of critical thinking in higher education* (pp. 373–390). Palgrave Macmillan US. https://doi.org/10.1057/9781137378057.
Leighton, J. P. (2019). The risk–return trade-off: Performance assessments and cognitive validation of inferences. *British Journal of Educational Psychology, 89*(3), 441–455. https://doi.org/10.1111/bjep.12271
Leighton, J. P. (2017a). Collecting and analyzing verbal response process data in the service of interpretive and validity arguments. In K. Ercikan, & J. W. Pellegrino (Eds.), *Validation of score meaning for the next generation of assessments: The use of response processes* (pp. 25–37). Routledge.
Leighton, J. P. (2017b). *Using think-aloud interviews and cognitive labs in educational research*. Oxford University Press.
McClelland, D. C. (1973). Testing for competence rather than for 'intelligence'. *American Psychologist, 28*(1), 1–14. https://doi.org/10.1037/h0034092
Messick, S. (1994). Alternative modes of assessment, uniform standards of validity. A Research Report. *ETS Research Report Series, 1994*(2). https://doi.org/10.1002/j.2333-8504.1994.tb01634.x. i–22.
Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741–749. https://doi.org/10.1037/0003-066X.50.9.741
Mislevy, R. J., Haertel, G., Cheng, B. H., Ructtinger, L., DeBarger, A., Murray, E., Rose, D., Gravel, J., Colker, A. M., Rutstein, D., & Vendlinski, T. (2013). A "conditional" sense of fairness in assessment. *Educational Research and Evaluation, 19*(2–3), 121–140. https://doi.org/10.1080/13803611.2013.767614
Nichols, P., & Huff, K. (2017). Assessments of complex thinking. In K. Ercikan, & J. W. Pellegrino (Eds.), *Validation of score meaning for the next generation of assessments: The use of response processes* (pp. 62–73). Routledge.
Oranje, A., Gorin, J., Jia, Y., & Kerr, D. (2017). Collecting, analyzing, and interpreting response time, eye-tracking and log data. In K. Ercikan, & J. W. Pellegrino (Eds.), *Validation of score meaning for the next generation of assessments: The use of response processes* (pp. 34–44). Routledge.
Pascarella, E. T., Blaich, C., Martin, G. L., & Hanson, J. M. (2011). How robust are the findings of academically adrift? *Change: The Magazine of Higher Learning, 43*(3), 20–24. https://doi.org/10.1080/00091383.2011.568898
Pekrun, R. (2020). Self-Report is indispensable to assess students' learning. *Frontline Learning Research, 8*(3), 185–193. https://doi.org/10.14786/flr.v8i3.637
Perie, M. (2020). Comparability across different assessment systems. In A. Berman, E. Haertel, & J. Pellegrino (Eds.), *Comparability of large-scale educational assessments: Issues and recommendations* (pp. 123–148). National Academy of Education. https://doi.org/10.31094/2020/1.
Pintrich, P. R. (2004). A conceptual framework for assessing motivation and self-regulated learning in college students. *Educational Psychology Review, 16*(4), 385–407. https://doi.org/10.1007/s10648-004-0006-x
Räisänen, M., Postareff, L., & Lindblom-Ylänne, S. (2016). University students' self- and co-regulation of learning and processes of understanding: A person-oriented approach. *Learning and Individual Differences, 47*, 281–288. https://doi.org/10.1016/j.lindif.2016.01.006
Räisänen, M., Postareff, L., Mattsson, M., & Lindblom-Ylänne, S. (2020). Study-related exhaustion: First-year students' use of self-regulation of learning and peer learning and perceived value of peer support. *Active Learning in Higher Education, 21*(3), 173–188. https://doi.org/10.1177/1469787418798517
Saariaho, E., Pyhältö, K., Toom, A., Pietarinen, J., & Soini, T. (2016). Student teachers' self- and co-regulation of learning during teacher education. *Learning: Research and Practice, 2*(1), 44–63. https://doi.org/10.1080/23735082.2015.1081395.
Saunders, B., Sim, J., Kingstone, T., Baker, S., Waterfield, J., Bartlam, B., Burroughs, H., & Jinks, C. (2018). Saturation in qualitative research: Exploring its conceptualization and operationalization. *Quality & Quantity, 52*(4), 1893–1907. https://doi.org/10.1007/s11135-017-0574-8
Schunk, D. H., & Greene, J. A. (2018). Historical, contemporary, and future perspectives on self-regulated learning and performance. In D. H. Schunk, & J. A. Greene (Eds.), *Handbook of self-regulation of learning and performance* (2nd ed., pp. 1–16). Routledge, Taylor & Francis Group.
Shavelson, R. J. (2010). *Measuring college learning responsibly: Accountability in a new era*. Stanford University Press.
Shavelson, R. J., Zlatkin-Troitschanskaia, O., Beck, K., Schmidt, S., & Marino, J. P. (2019). Assessment of university students' critical thinking: Next generation performance assessment. *International Journal of Testing, 19*(4), 337–362. https://doi.org/10.1080/15305058.2018.1543309
Shavelson, R. J., Zlatkin-Troitschanskaia, O., & Mariño, J. P. (2018). International performance assessment of learning in higher education (iPAL): Research and development. In O. Zlatkin-Troitschanskaia, M. Toepper, H. A. Pant, C. Lautenbach, & C. Kuhn (Eds.), *Assessment of learning outcomes in higher education: Cross-national comparisons and perspectives* (pp. 193–214). Springer International Publishing. https://doi.org/10.1007/978-3-319-74338-7.
Solano-Flores, G., & Chia, M. (2017). Validation of score meaning using examinee response processes for the next generation of assessments. In K. Ercikan, & J. W. Pellegrino (Eds.), *Validation of score meaning for the next generation of assessments: The use of response processes* (pp. 127–137). Routledge.
Timmermans, S., & Tavory, I. (2012). Theory construction in qualitative research: From grounded theory to abductive analysis. *Sociological Theory, 30*(3), 167–186. https://doi.org/10.1177/0735275112457914
Toering, T., Elferink-Gemser, M. T., Jonker, L., van Heuvelen, M. J. G., & Visscher, C. (2012). Measuring self-regulation in a learning context: Reliability and validity of the Self-Regulation of Learning Self-Report Scale (SRL-SRS). *International Journal of Sport and Exercise Psychology, 10*(1), 24–38. https://doi.org/10.1080/1612197X.2012.645132
Tremblay, K., Lalancette, D., & Roseveare, D. (2012). *AHELO, Assessment of higher education learning outcomes. Feasibility study report, vol. 1. Design and implementation [Study report]*. OECD. https://www.oecd.org/education/skills-beyond-school/AHELOFSReportVolume1.pdf.

Tuononen, T., Parpala, A., & Lindblom-Ylänne, S. (2019). Graduates' evaluations of usefulness of university education, and early career success – a longitudinal study of the transition to working life. *Assessment & Evaluation in Higher Education, 44*(4), 581–595. https://doi.org/10.1080/02602938.2018.1524000

Usher, E. L., & Schunk, D. H. (2018). Social cognitive theoretical perspective of self-regulation. In D. H. Schunk, & J. A. Greene (Eds.), *Handbook of self-regulation of learning and performance* (2nd ed., pp. 17–35). Routledge, Taylor & Francis Group.

van Someren, M. W., Barnard, Y. F., & Sandberg, J. A. C. (1994). *The think aloud method: A practical guide to modelling cognitive processes*. Academic Press.

Vipunen. (2021). *Education statistics Finland. Statistics database*. https://vipunen.fi/en-gb/home.

Winne, P. H. (2018). Cognition and metacognition within self-regulated learning. In D. H. Schunk, & J. A. Greene (Eds.), *Handbook of self-regulation of learning and performance* (2nd ed., pp. 36–48). Routledge, Taylor & Francis Group.

Zahner, D., & Ciolfi, A. (2018). International comparison of a performance-based assessment in higher education. In O. Zlatkin-Troitschanskaia, M. Toepper, H. A. Pant, C. Lautenbach, & C. Kuhn (Eds.), *Assessment of learning outcomes in higher education: Cross-national comparisons and perspectives* (pp. 215–244). Springer International Publishing. https://doi.org/10.1007/978-3-319-74338-7.

Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory Into Practice, 41*(2), 64–70. https://doi.org/10.1207/s15430421tip4102_2

Zimmerman, B. J., & Campillo, M. (2003). Motivating self-regulated problem solvers. In J. E. Davidson, & R. J. Sternberg (Eds.), *The psychology of problem solving* (1st ed., pp. 233–262). Cambridge University Press. https://doi.org/10.1017/CBO9780511615771.009.

Zlatkin-Troitschanskaia, O., Shavelson, R. J., & Kuhn, C. (2015). The international state of research on measurement of competency in higher education. *Studies in Higher Education, 40*(3), 393–411. https://doi.org/10.1080/03075079.2015.1004241

Zlatkin-Troitschanskaia, O., Toepper, M., Molerov, D., Buske, R., Brückner, S., Pant, H. A., Hofmann, S., & Hansen-Schirra, S. (2018). Adapting and validating the collegiate learning assessment to measure generic academic skills of students in Germany: Implications for international assessment studies in higher education. In O. Zlatkin-Troitschanskaia, M. Toepper, H. A. Pant, C. Lautenbach, & C. Kuhn (Eds.), *Assessment of learning outcomes in higher education: Cross-national comparisons and perspectives* (pp. 245–266). Springer International Publishing. https://doi.org/10.1007/978-3-319-74338-7.

Zlatkin-Troitschanskaia, O., Shavelson, R. J., Schmidt, S., & Beck, K. (2019). On the complementarity of holistic and analytic approaches to performance assessment scoring. *British Journal of Educational Psychology, 89*(3), 468–484. https://doi.org/10.1111/bjep.12286