Exploring internal structure of a performance-based critical thinking assessment for new students in higher education

Kleemola, Katri

2022-05-11

Kleemola , K , Hyytinen , H & Toom , A 2022 , ' Exploring internal structure of a performance-based critical thinking assessment for new students in higher education ' , Assessment & Evaluation in Higher Education , vol. 47 , no. 4 , pp. 556-569 . https://doi.org/10.1080/02602938.2021

http://hdl.handle.net/10138/352614
https://doi.org/10.1080/02602938.2021.1946482

Exploring internal structure of a performance-based critical thinking assessment for new students in higher education

Katri Kleemola[a]*, Heidi Hyytinen[a] and Auli Toom[a]

[a] *University of Helsinki, Centre for University Teaching and Learning*

* Katri Kleemola, P.O. Box 9, 00014 University of Helsinki, Finland, (e-mail: katri.kleemola@helsinki.fi)

Acknowledgements:

The authors would like to thank the Council for Aid to Education (CAE) for providing the CLA+ International, the participating institutions for administering the assessments and recruiting participants and the Kappas project team for their efforts.

Exploring internal structure of a performance-based critical thinking assessment for new students in higher education

Critical thinking is a combination of complex cognitive skills that are used for purposeful thinking. It is important for the successful acquisition of disciplinary skills in higher education and thus, it is a valuable competency for a new student. The complex nature of critical thinking leads to challenges for its assessment even in performance assessments such as CLA+ International (Collegiate Learning Assessment). The aim with this study is to examine internal associations of a critical thinking assessment for new students in higher education. The sample consisted of 1469 first-year students in 18 higher education institutions in Finland. An open-ended performance task and multiple-choice tasks were used to assess six measures of critical thinking, namely analysis and problem solving, writing effectiveness, writing mechanics, scientific and quantitative reasoning, critical reading and critiquing an argument. Exploratory and confirmatory factor analyses were used to analyse the data. A latent component indicated by the measures derived from the performance task was identified. The measures derived from the multiple-choice tasks did not form a factorial structure. Multiple-choice questions are problematic in critical thinking assessment as they focus on individual skills instead of holistic use of skills. Implications for assessment development and higher education are discussed.

Keywords: critical thinking, performance assessment, internal structure, higher education

Introduction

Critical thinking is an important objective of higher education; it is crucial for employment and success in the work force (Arum and Roksa 2011; Tuononen, Parpala, and Lindblom-Ylänne 2019; Halpern 2014; Shavelson 2010). However, it is especially important when studying in successful learning and acquiring domain-specific knowledge and skills (Halpern 2014). It is a valuable competency for a new student. Critical thinking is a combination of complex cognitive skills such as problem solving,

analysis and evaluation, but it also requires affective disposition to utilize these skills (Hyytinen and Toom 2019; Halpern 2014; Facione 1990).

While assessment of students' critical thinking is important for developing teaching and learning in higher education, it is not without challenges. Critical thinking is often investigated by using self-report methods, which only enable access to students' perceptions of their critical thinking skills, but performance assessments can allow more direct access to students' thinking processes (e.g. Hyytinen et al. 2015; Zlatkin-Troitschanskaia, Shavelson, and Kuhn 2015). The central idea of a performance assessment is to face participants with a so-called criterion situation, a simulation of a real-world scenario in which critical thinking is required (Shavelson et al. 2019; Shavelson 2010; McClelland 1973). Considering the complex nature of critical thinking, a performance assessment for critical thinking should be challenging enough to trigger complex and holistic use of relevant skills (e.g. Shavelson et al. 2019). In addition, to tap critical thinking skills, students' prior knowledge of the contents should not influence the assessment (Klein et al. 2007). Creating a cross-disciplinary assessment for critical thinking is a balancing act between tasks that are challenging enough and generic enough in their content (cf. Hetmanek et al. 2018). Typically, performance assessments use open-ended performance tasks and/or multiple-choice tasks. While studies that focus on the equivalence and differences between the types of task are scarce, recent studies have pointed out that different types of task trigger different thinking processes and hence have an effect on the interpretation of findings (Hyytinen et al. 2020, 2015).

Understanding what the assessment measures is fundamental, yet studies evaluating evidence of validity are rare (cf. American Educational Research Association, American Psychological Association, and National Council on

Measurement in Education 2014). Care should be taken to ensure that the assessment is in line with the intended theoretical background (e.g. Shavelson, Zlatkin-Troitschanskaia, and Marino 2018; Messick 1995) and that it works accordingly in a range of contexts (e.g. Solano-Flores and Chía 2017). Qualitative methods such as cognitive labs (e.g. Hyytinen et al. 2020, 2014; Leighton 2019; Messick 1995), and quantitative methods such as factor analyses (e.g. Zlatkin-Troitschanskaia et al. 2019; Davey et al. 2015) are needed in order to understand what an assessment is telling us about the students' critical thinking.

The present study focuses on the internal structure of a performance-based critical thinking assessment for new students that were recently admitted in higher education in Finland. This assessment, namely CLA+ International (Collegiate Learning Assessment), uses two task types, open-ended performance tasks, and multiple-choice tasks. Investigation will increase understanding of the meaning and challenges of the present critical thinking assessment, but it will also give insights into the quality of new students' critical thinking skills. Further, the findings provide new initiatives for the development of critical thinking assessments in general and the significance of the task type and contents.

What is critical thinking?

Critical thinking is purposeful, self-regulatory thinking that integrates skills, knowledge and disposition to act (Halpern 2014). An ideal critical thinker is able to assess the reliability and relevance of information, to recognize biases and to reach a conclusion (Shavelson et al. 2019). Critical thinking can be viewed as a complex process, a holistic approach to a task (Zlatkin-Troitschanskaia et al. 2019), activating a variety of cognitive skills such as analytical reasoning, problem solving and argumentation (Halpern 2014; Facione 1990). Conceptually, communicative skills are often left outside the definition

of critical thinking. However, manifesting one's thoughts is necessary for successful critical thinking. Argumentation has a focus on displaying thoughts and it is most often included as an important or even central part of critical thinking (Kuhn 2019; Andrews 2015; Halpern 2014; Ennis 1991). From this perspective, it can be considered that even other communicative skills such as writing mechanics are functionally intertwined with critical thinking (Hetmanek et al. 2018; Halpern 2014; Messick 1995).

Critical thinking is not activated without affective components such as flexibility and disposition to reconsider, to be persistent and to reflect one's own thinking (Hyytinen, Toom, and Shavelson 2019; Halpern 2014). Disposition to utilize complex cognitive skills is associated with the actual performance in critical thinking (Heijltjes et al. 2014). Further, critical thinking cannot be separated from the knowledge. The domain-general and domain-specific nature of critical thinking has been discussed extensively over the years (e.g. Fischer et al. 2018). Some insist that critical thinking is firmly dependent on disciplinary knowledge and practices, whereas some perceive that critical thinking skills are transferable and that teaching them is entirely feasible (see Shavelson 2018). It has been shown that critical thinking requires at least some propositional knowledge on the topic and that expert knowledge facilitates thinking (Hyytinen and Toom 2019; Hyytinen, Toom, and Shavelson 2019; Hyytinen et al. 2014). It has been suggested that the domain-specificity of thinking and reasoning grow as expertise grows, in other words, the more advanced and demanding the topic, the more domain-specific skills are required (Hetmanek et al. 2018). However, according to Shavelson (2018), critical thinking does not require expertise, it is necessary already in the introductory phases of undertaking higher education.

Critical thinking is associated with study success and how new students in higher education adjust to it (van der Zanden et al. 2019; O'Hare and McGuinness

2015; Arum and Roksa 2011; Badcock, Pattison, and Harris 2010). Deficiencies in critical thinking in transition to higher education may turn into long-term challenges, as critical thinking does not always develop during higher education (Evens, Verburgh, and Elen 2013; Arum and Roksa 2011). A diagnostic assessment of new students' critical thinking helps in understanding their challenges and in supporting them in a constructive way.

Assessing critical thinking with performance-based assessments

Several studies in critical thinking have been based on self-report methods (see Zlatkin-Troitschanskaia, Shavelson, and Kuhn 2015). They are possibly problematic because they focus on participants' perceptions of their critical thinking skills or their critical thinking dispositions instead of their actual performance in critical thinking (Hyytinen et al. 2015; Zlatkin-Troitschanskaia, Shavelson, and Kuhn 2015). Students' perceptions may differ from their actual performance, either exaggerating or underrating their skills (Zlatkin-Troitschanskaia, Shavelson, and Kuhn 2015; Hyytinen et al. 2014; Bowman 2010). Therefore, measurements that seek to tap into processes of thinking are needed (Hyytinen and Toom 2019; Shavelson 2010).

Performance assessments are complex tasks that evoke authentic behaviour covering aspects of critical thinking through so-called criterion situations that resemble the real world (Davey et al. 2015; Shavelson 2010; Messick 1994; McClelland 1973). Thus, performance assessments enable a better understanding of students' thinking compared to self-report methods (Shavelson et al. 2019; Hyytinen and Toom 2019; Hyytinen et al. 2015;). However, critical thinking is a complex process and to reflect it, a performance assessment should require activation of a variety of skills in a holistic manner (Zlatkin-Troitschanskaia et al. 2019). In addition, cross-disciplinary performance assessments of critical thinking should not require prior knowledge of the

topic to be valid across study fields (Shavelson 2010; Klein et al. 2007). In developing a performance assessment, it is important to investigate its validity through diverse evidence, to understand what the assessment measures (e.g. American Educational Research Association, American Psychological Association, and National Council on Measurement in Education 2014).

A performance-based assessment may be an open-ended performance task or a multiple-choice task (e.g. Hyytinen et al. 2015; Messick 1994). While multiple-choice tasks are more widely used in critical thinking assessment due to their ease of use and inexpensiveness, their challenge is that guessing or eliminating incorrect options is always possible without profound understanding of the actual topic (Hyytinen et al. 2020, 2015; Shavelson 2010). Thus, the responses in a multiple-choice task tell little about thinking processes behind them (Messick 1994). Multiple-choice tasks often focus on a single skill, overlooking the complex nature of critical thinking which performance tasks often support (Shavelson 2010). However, while performance tasks address the complexity of critical thinking, students may not present their best skills in the test situation due to a lack of motivation, effort, or interest in the specific performance task especially if it is a low-stakes situation (e.g. Hyytinen et al. 2020; Arum and Roksa 2011). Additionally, performance tasks are often considered to be expensive and possibly unreliable due to the human aspect in interpreting and scoring responses (Davey et al. 2015). Nevertheless, it is possible to address the interpretative challenge through a rigorous scorer training and extensive scoring rubrics (Borowiec and Castle 2019; Shavelson, Baxter, and Gao 1993).

CLA+ International

CLA+ International by the Council for Aid to Education (CAE) is a test designed to assess critical thinking and written communication skills in higher education in a

domain-general manner (Shavelson 2010; Klein et al. 2007). It has been used earlier in Italy, Great Britain and Chile. CLA+ International is closely associated with CLA+ that is used in USA. The CLA+ tasks that are adaptable for international context are used in CLA+ International. Each country translates and adapts the assessment for their context in association with CAE. The assessment is typically used in low-stakes situations to help students and institutions develop their activities.

The computer-based assessment consists of two main sections (see Zahner 2013; Zahner and Ciolfi 2018). The first section is a performance task in which students are faced with an authentic situation that requires them to use various critical thinking skills in a holistic manner (Hyytinen et al. 2015; Shavelson 2010). They write a report using a document library which includes both relevant and unreliable documents, namely a blog text, a podcast transcript, a research memorandum, a newspaper article and an infographic of statistics. Students can take up to 60 minutes to complete the performance task. An additional 30 minutes are allocated for 25 multiple-choice questions in three subsections. Students can go back and forth across subsections within the time limitation. Each of the subsections includes with five to ten questions that are based on one or two documents. CLA+ International measures analysis and problem solving, writing effectiveness, writing mechanics, scientific and quantitative reasoning, critical reading and evaluation, and ability to critique an argument (see Zahner 2013). Theoretically, the measures of CLA+ International should be indicators of a single latent variable of critical thinking (cf. Hyytinen and Toom 2019; Shavelson et al. 2019; Halpern 2014; Shavelson 2010). Further, the CLA+ International provides students not only with scores for individual measures, but also a total score for the entire assessment that is used for determining the competency level of the student, implying that the skills that are assessed are indicators of a macro construct. However, it has been questioned

whether CLA+ International actually is able to capture critical thinking (e.g. Aloisi and Callaghan 2018). Thus, there is a growing need for extensive investigations on different sources of validity regarding the assessment, such as cognitive processes, internal structure and correlational associations (see American Educational Research Association, American Psychological Association, and National Council on Measurement in Education 2014). Furthermore, as CLA+ International is intended for cross-cultural use such as benchmarking between different countries, investigations of validity need to be conducted in different cultural contexts (see Solano-Flores and Chía 2017).

While internal structure of CLA+ International has not been studied earlier, some earlier research raises questions about it. In research on the CLA+ International and its predecessor CLA, it has been found that students' performance in open-ended tasks and multiple-choice tasks differs due to students' processing strategies (Hyytinen et al. 2020, 2015; Beckman et al. 2019; Shavelson 2010). In examining correlations between different sections of the CLA+ International, an Italian study found that students' performance was not consistent across different sections of the CLA+ International (Zahner and Ciolfi 2018). This would indicate that the internal structure might be incoherent. The Italian study indicated that the correlations between sections are substantially higher in the American context. It is not known why such differences might exist between Italian and American contexts. It might be merely a methodological issue due to the use of different task types as indicated above. There may be also contextual aspects that influence critical thinking, or issues in the cultural adaptation of the test (e.g. Solano-Flores and Chía 2017; Hambleton, Merenda, and Spielberger 2005). As CLA+ International was implemented in the Finnish context for the first time, earlier findings warrant a thorough investigation of the internal structure of the

assessment in order to understand the meaning of the assessment and pinpoint development needs. Furthermore, Finnish higher education includes two types of institutions with slightly different emphases in their learning goals and education: research-intensive universities and universities of applied sciences. Therefore, it is important to address possible contextual differences in the internal structure.

Aim of the study

Our aim in the present study was to explore internal structure of a performance-based critical thinking assessment CLA+ International. We wanted to find out if the assessment reflects the construct of critical thinking when it is used for new students in Finnish higher education. We did so by exploring the internal structure of the assessment. We investigated the associations between the assessment sections and whether the associations and the structure of the assessment are equivalent in two separate contexts, namely different types of Finnish higher education institutions. Our research questions are:

> RQ1. Which latent components of critical thinking can be identified in the CLA+ International data for new students in Finnish higher education?
> RQ2. How are the components of critical thinking associated with each other in two types of Finnish higher education institutions?

As all the measures of the assessment are known components of critical thinking (e.g. Halpern 2014), we expect these measures to be strongly correlated. However, earlier research indicates that thinking processes between different task types such as open-ended performance tasks and multiple-choice tasks may differ (Hyytinen et al. 2020, 2015; Zahner and Ciolfi 2018). Thus, we assume that instead of one latent component

indicating a single construct of critical thinking, two latent components may emerge, setting apart two types of assessment task. Further, earlier findings on CLA+ International indicate a possibility of differences in the internal structure of the assessment across contexts (Zahner and Ciolfi 2018). We assume that such differences might emerge within the Finnish data, too, as the Finnish higher education system has a binary divide between two types of institution.

Materials and method

Context of the study

In Finland, higher education consists of two sectors with distinctly different objectives. Research-intensive universities base their teaching on research whereas universities of applied sciences focus on practical, work-related learning. Students who apply for entry to research-intensive universities have on average stronger prior school achievement compared to students who apply for universities of applied sciences (Heiskala, Erola, and Kilpi-Jakonen 2020). Bachelor's and master's degrees can be completed at both types of institution. However, in research-intensive universities, most students complete a research-intensive master's degree while in universities of applied sciences the bachelor's degree is more common target. Student admissions to bachelor's programs are competitive and until recently, they had been based on discipline-specific exams (e.g. Kleemola and Hyytinen 2019).

Participants and data collection

Data were collected in a national project that was funded by the Ministry of Education and Culture in accordance with the ethical principles of research with human participants by the Finnish National Board on Research Integrity (2019).

The study was conducted in 11 (out of 13) research-intensive universities and seven (out of 22) universities of applied sciences. The participating institutions were responsible for inviting students and administering the assessment. The instructions and tools were provided with the project. Additionally, the project was responsible for cultural adaptation and translation of the assessment. The process adhered to International Translation Committee (2018) guidelines including double translation, reconciliation, review and cognitive labs. Sampling was conducted by the project's statistician with the aim of achieving national representativeness of study fields. Students were invited to participate either via e-mail or as part of a course. Participation was voluntary and taken up by 25% of those invited. At some institutions, students received minor incentives like cinema tickets.

Participating students were first-year students in programmes in the humanities, social sciences, economics, law, natural sciences, engineering, forestry, medicine, nursing and services. A total of 1538 students participated in the assessment during their first term. Sixty-nine students did not complete the whole assessment and they were excluded, leaving a final data set that comprised 1469 students. Of these students, 785 were from research-intensive universities and 684 from universities of applied sciences. The mean age of the students was 23.13 (SD 5.38, Min 18, Max 64) and 52% reported being female, 45% male and 3% other.

Measures

CLA+ International was used to assess students' critical thinking (see description of the assessment in Theoretical background). Six measures were used. An open-ended performance task was used to measure *Analysis and problem-solving (APS), Writing Effectiveness (WE) and Writing mechanics (WM).* A multiple-choice questionnaire was used to measure *Scientific and quantitative reasoning (SQR), Critical reading and*

evaluation (CRE) and *Critiquing an argument (CA)*. For the APS, students showed their skills in utilizing, analysing and evaluating the provided information and reaching a conclusion. For the WE, students needed to elaborate and to provide arguments that were well constructed and logical. For the WM, students were evaluated in their skills in producing well-structured and grammatically correct text. For the SQR, students needed to make inferences and hypotheses, to identify connected and conflicting data, to evaluate conclusions and to make decisions. For the CRE, students evaluated reliability of information, made inferences and identified strengths and weaknesses of claims. For the CA, students needed to identify flaws in logic and biased assumptions.

In the performance task (APS, WE, WM), participants provided a written response to a complex real-life issue utilizing five documents that varied in their source and reliability. Students were asked to analyse, give recommendations and identify counterarguments. The task that was used in the present study is confidential, but a similar task was presented in Shavelson (2010). Performance task responses were qualitatively analysed by a group of seven trained scorers, based on a scoring rubric (the rubric is published in Aloisi and Callaghan 2018). Each of the skills was scored on a six-point scale with an additional option for responses that could not be scored, for instance empty response fields. Scorers took part in a two-day training to ensure that the scoring rubric was interpreted consistently (Borowiec and Castle 2019; Shavelson, Baxter and Gao 1993). Two independent scorers evaluated each response. Agreement between the two scorers was acceptable (Analysis and Problem Solving r=.73, Writing Effectiveness r=.72, Writing Mechanics r=.71). Agreement in the scoring was further controlled by monitoring the difference of the sum of given scores in APS, WE and WM. If the sum had more than a two-point difference between the scorers, a third scorer evaluated the response. If the third scorers' scores were within the two-point limit with

either of the first two scorers, the discrepant scores were dropped. This was done in 95 cases within the data of 1469 students. The third scorer enabled agreement in each of these cases and therefore, no fourth scorer was needed. Any discrepancies between the scorers within the two-point limit was interpreted in favour of the student. The scoring system assigned responses randomly to scorers. All authors participated in the scoring process.

In the multiple-choice questionnaire, ten questions targeted SQR, ten questions targeted CRE and five questions targeted CA. The questions in each section were based on one or more documents. Each question had four options. Scores for each section were based on the number of correct answers. Cronbach alfa's that were calculated for the multiple-choice sections (SQR α=.47, CRE α=.65, CA α=.57) indicated that the reliability of these sections was not desirable.

Scores were transformed into comparable Z scores, setting the mean of each measure at zero and standard deviation at one. This levelled off any differences in difficulty of the sections. Descriptive values of scores and correlations between the measures are presented in Table 1. While the performance task sections are highly correlated with each other, the multiple-choice sections have low correlations between each other and performance task section.

Data analyses

First, we used exploratory factor analysis to identify any latent components. Exploratory method allows for any unexpected dimension or measurement error (see Davey et al. 2015), and is suitable for exploring an instrument in a new context. For extraction, maximum likelihood was used, and for rotation, the oblique direct oblimin (see Goretzko, Pham, and Bühner 2019; Costello and Osborne 2005). To determine the most appropriate model, a parallel analysis and MAP test were conducted in addition to

examination of communalities and factor loadings (Goretzko, Pham, and Bühner 2019; O'Connor 2000; Fabrigar, Wegener, MacCallum, and Strahan 1999).

Acknowledging the findings of the exploratory factor analysis and the theory behind the assessment, we further tested a measurement model using confirmatory factor analysis to understand associations between the components of critical thinking. The goodness-of-fit of the models was tested with the Root Mean Square Error of Approximation (RMSEA), Comparative Fit Index (CFI), Tucker-Lewis Index (TLI) and Standardized Root Mean Square Residual (SRMR) (Hu and Bentler 1998). Cut off points for good fit were used (RMSEA<.08, CFI>.95, TLI>.95 and SRMR<.08) (e.g. Hu and Bentler 1999). In order to make sure that the model was equivalent in two different higher education contexts, namely groups of students from research-intensive universities (RIU students) and universities of applied sciences (UAS students), we tested measurement invariance with configural, metric and scalar models moving stepwise towards more constrained models (Vandenberg and Lance 2000). Differences between models were investigated with χ^2 difference tests and, as χ^2 is sensitive to large sample sizes, CFI difference tests (Meade, Johnson, and Braddy 2008).

Exploratory factor analysis was conducted with SPSS version 25. Confirmatory factor analysis was conducted with Mplus version 8.4 (Muthén and Muthén 1998-2017).

Results

Components of critical thinking

Exploratory factor analysis was used to determine the number of factors that could be identified in the data. First, a parallel analysis and a MAP test were conducted (O'Connor 2000). While the parallel analysis indicated a 2-factor solution, the MAP test indicated a 1-factor solution. Both alternatives were examined in detail. In both

solutions, communalities (Tables 2 and 3) for APS, WE and WM were moderate to high, but communalities for SQR, CRE and CA were very low, well below the desired .40 (Costello and Osborne 2005). This was not surprising, given that the correlations between these measures were low and reliability of these measures was not good. APS, WE and WM loaded strongly on a factor in both solutions (Tables 2 and 3). In the 1-factor solution (Table 2), the loadings for SQR, CRE and CE were very low. In the 2-factor solution (Table 3), these measures loaded more strongly on the second factor. However, loadings were rather low even though they passed the desired .32 mark (see Tabachnick and Fidell 2014). Examination of both solutions indicated that while APS, WE and WM load on a factor, SQR, CRE and CE do no form clear factorial structure.

Associations between critical thinking components in different higher education contexts

Considering the findings of the exploratory factor analysis, no latent component is indicated by SQR, CRE and CA measures while the APS, WE and WM measures indicate a latent component. To understand more about the associations between these components, a model including all measures of the assessment was suggested and tested. This model included a latent variable that is indicated by APS, WE and WM with correlations with observed variables SQR, CRE and CA (see Figure 1).

Configural, metric and scalar invariance were examined in two higher education contexts (RIU and UAS) in order to ensure equivalence. Differences in fit indices are presented in Table 4. The unconstrained, configural model had acceptable model fit (RMSEA=.07, CFI=.97, TLI=.96, SRMR=.05) indicating that both groups have same number of factors. In the metric model, factor loadings and covariances were constrained across groups. The metric model had a good fit, (RMSEA= .06, CFI=.97, TLI=.96, SRMR=.05), and both $\Delta\chi 2$ and ΔCFI tests indicated metric invariance. Thus,

factor loadings were equal across groups. Finally, the scalar model was investigated. In addition to the metric model, factor intercepts were constrained. The model fit was still acceptable (RMSEA=.07, CFI=.96, TLI=.95, SRMR=.05), but the significance of the $\Delta\chi^2$ and the difference greater than .002 in the ΔCFI test indicated scalar non-invariance. Next, partial scalar invariance was tested freeing factor intercepts one by one. Freeing the intercept of WM provided a good model fit (RMSEA=.06, CFI=.97, TLI=.96, SRMR=.05) and good results in both $\Delta\chi 2$ and ΔCFI tests. Therefore, partial scalar invariance was supported. The freed intercept of WM was lower in UAS students than RIU students. This finding indicated that when a RIU student and a UAS student had equal scores in APS and WE, the RIU student had higher score in WM.

The findings of partial scalar invariance should be considered when interpreting the findings, but invariance can be considered acceptable when most of the items are invariant (e.g. Vandenberg and Lance 2000). Full configural and metric invariance meant that the suggested model represented both student groups, in other words the factorial structure and loadings were equal across the groups. Therefore, the associations in the model were investigated without groupings (see Figure 2). APS, WE and WM loaded strongly on the latent component as indicated by the exploratory factor analysis. The measure WE had the strongest loading. Conversely, correlations between the latent component and measures SQR, CRE and CA were significant yet very low. This, combined with the fact that SQR, CRE and CA did not form a factorial structure, indicate that students have varying skillsets.

Discussion

The findings from the present study demonstrate that the measures of CLA+ International do not indicate a single latent component of critical thinking when administered to the new students in Finnish higher education. Based on earlier studies

(e.g. Hyytinen et al. 2020, 2015) we expected that students' performance might be different in the performance task section and the multiple-choice section due to different processes they trigger. However, we were surprised to find that subsections of the multiple-choice section (SQR, CRE and CA) did not indicate a latent component at all, but were only weakly associated with each other. Nevertheless, this finding is in line with Italian findings in their CLA+ International administration (Zahner and Ciolfi 2018). While the multiple-choice subsections turned out to be problematic, the performance task section with measures APS, WE and WM functioned as expected, indicating a latent component. The strongest indicator of the component was WE (writing effectiveness) that focused on argumentation and helped students' make their ideas manifest (see Kuhn 2019; Andrews 2015). While the identified latent component includes a variety of skills that are considered essential in critical thinking, such as analysis, evaluation and problem solving, it has a strong emphasis on the communicative aspect through WE and WM. It can be assumed that writing skills are essential in the competency that the latent component represents (see also Aloisi and Callaghan 2018). Therefore, we call the latent component "Critical thinking and writing".

The internal structure of the assessment was tested in both sectors of the Finnish higher education system, namely research-intensive universities and universities of applied sciences. Configural and metric invariance was detected, but only partial scalar invariance was supported. In testing the scalar model, it was found that RIU students had relatively stronger skills in writing mechanics compared to UAS students. This could be due to differences in student population in the two types of institution (Heiskala, Erola, and Kilpi-Jakonen 2020). Conversely, other internal associations did not differ across the groups of new students. However, as SQR, CRE and CA are

weakly associated not only with each other but also with the latent component of Critical thinking and writing, it seems that students have somewhat uneven skillsets when they enter higher education, meaning that their strengths and weaknesses vary.

Some of the incoherence in the internal structure of the assessment can be explained with the difference in the task type. Open-ended performance tasks and multiple-choice tasks trigger different cognitive processes (Hyytinen et al. 2020). However, questions arise concerning the lack of unity within the multiple-choice section. Task types cannot entirely explain this finding. An inherent difference in the task types is that the performance task requires holistic use of different skills whereas the multiple-choice sections are designed to measure one skill at a time (e.g. Shavelson 2010). It may be that in holistic use of skills, the differences in skillsets level off while the same differences emerge more clearly when the task focuses on an individual skill. Additionally, it is possible that the scoring process adds to the coherence of the performance task section. The significance of the task content also needs to be acknowledged. Even if tasks are designed to require no expertise (e.g. Shavelson 2018; Klein et al. 2007), the prior knowledge on the topic probably facilitates a student's performance (see Hyytinen and Toom 2019; Hetmanek et al. 2018).

It is noteworthy that some research instruments, particularly self-report surveys, function unexpectedly within new students, who are in the midst of transition to higher education (e.g. Bowman 2010). However, as a performance-based assessment was used in the present study, the findings were expected to reflect the skills of this group. To find out if associations between critical thinking skills change during higher education studies, a longitudinal research design is needed. Furthermore, disciplinary differences should be focused when studying development of critical thinking skills.

To understand antecedents of students' performance in the assessment, future research should focus on the interaction between the skills, task content and students' background such as prior academic achievement. Writing skills are necessarily emphasised in a written task, and in order to develop assessments that tap thinking processes, more understanding is needed on association of writing and critical thinking. Further, cross-cultural investigations are needed in order to find out if internal structure of the assessment is different across cultural contexts (see also Zahner and Ciolfi 2018) to develop the present assessment.

Findings give guidelines for developing critical thinking assessments. The target construct should be defined carefully (Shavelson et al. 2019) and all sections of the assessment should be aligned with characteristics of the target construct (cf. Shavelson 2010; McClelland 1973). Appropriate sources of validity should always be investigated when developing a new assessment or implementing one in a new context (Solano-Flores and Chía 2017; American Educational Research Association, American Psychological Association, and National Council on Measurement in Education 2014). The multiple-choice tasks appear problematic, as they do not require holistic thinking inherent to critical thinking in the same way as performance tasks (e.g. Shavelson 2010). To capture complex and holistic critical thinking, tasks should trigger holistic thinking processes. Open-ended performance tasks are best suited for this purpose even though their emphasis on written communication needs to be considered. If multiple-choice tasks are used, intertwining them with open-ended tasks could be beneficial, e.g. making students argue about their responses in the multiple-choice task. Issues of the task type that are indicated in earlier research (e.g. Hyytinen et al. 2020, 2015; Shavelson 2010) should be acknowledged, and the equivalence between assessment

sections should be ensured particularly if different task types are used (see also Aloisi and Callaghan 2018).

Varying skillsets of new students in higher education should be acknowledged in the teaching of introductory courses. Because challenges at the beginning of the studies may grow into more severe problems (e.g. Arum and Roksa 2011), it is important to address deficiencies in skillsets. Students with challenges in their critical thinking should be identified and supported from the very beginning to avoid further problems and student attrition.

References

Aloisi, C. and A. Callaghan. 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.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. 2014. *Standards for Educational and Psychological Testing*. Washington (D.C.): American Educational Research Association.

Andrews, R. 2015. "Critical Thinking and/Or Argumentation in Higher Education." In *The Palgrave Handbook of Critical Thinking in Higher Education*, edited by Martin Davies and Donald Barnett, 49-62. New York: Palgrave Macmillan.

Arum, R. and J. Roksa. 2011. *Academically Adrift: Limited Learning on College Campuses*. Chicago: University of Chicago Press.

Badcock, P., P. Pattison, and K.L. Harris. 2010. "Developing Generic Skills through University Study: A Study of Arts, Science and Engineering in Australia." *Higher Education* 60 (4): 441-458.

Beckman, K., T. Apps, S. Bennett, B. Dalgarno, G. Kennedy, and L. Lockyer. 2019. "Self-Regulation in Open-Ended Online Assignment Tasks: The Importance of Initial Task Interpretation and Goal Setting." *Studies in Higher Education*: 1-15.

Borowiec, K. and C. Castle. 2019. "Using Rater Cognition to Improve Generalizability of an Assessment of Scientific Argumentation." *Practical Assessment, Research, and Evaluation* 24 (1): 8.

Bowman, N. A. 2010. "Can 1st-Year College Students Accurately Report their Learning and Development?" *American Educational Research Journal* 47 (2): 466-496.

Costello, A. and J. Osborne. 2005. "Best Practices in Exploratory Factor Analysis: Four Recommendations for Getting the most from Your Analysis." *Practical Assessment, Research, and Evaluation* 10 (1): 7.

Davey, T., S. Ferrara, R. Shavelson, P. Holland, N. Webb, and L. Wise. 2015. "Psychometric Considerations for the Next Generation of Performance Assessment." *Washington, DC: Center for K-12 Assessment & Performance Management, Educational Testing Service*.

Ennis, R. 1991. "Critical Thinking: A Streamlined Conception." *Teaching Philosophy* 14 (1): 5-24.

Evens, M., A. Verburgh, and J. Elen. 2013. "Critical Thinking in College Freshmen: The Impact of Secondary and Higher Education." *International Journal of Higher Education* 2 (3): 139-151.

Fabrigar, L., D. Wegener, R. MacCallum, and E. Strahan. 1999. "Evaluating the use of Exploratory Factor Analysis in Psychological Research." *Psychological Methods* 4 (3): 272.

Facione, P. 1990. *Critical Thinking: A Statement of Expert Consensus for Purposes of Educational Assessment and Instruction.* Newark: The American Philosophical Association.

Finnish National Board on Research Integrity. 2019. *The Ethical Principles of Research with Human Participants and Ethical Review in the Human Sciences in Finland*. Helsinki: Finnish National Board on Research Integrity TENK.

Fischer, F., C. Chinn, K. Engelmann, and J. Osborne, eds. 2018. *Scientific Reasoning and Argumentation: The Roles of Domain-Specific and Domain-General Knowledge*. New York: Routledge.

Goretzko, D., T. Pham, and M. Bühner. 2019. "Exploratory Factor Analysis: Current use, Methodological Developments and Recommendations for Good Practice." *Current Psychology*: 1-12.

Halpern, D. 2014. *Thought and Knowledge: An Introduction to Critical Thinking*. 5th ed. New York, London: Psychology Press.

Hambleton, R., P. Merenda, and C. Spielberger. 2005. *Adapting Educational and Psychological Tests for Cross-Cultural Assessment*. Mahwah, N.J.: Psychology Press.

Heijltjes, A., T. van Gog, J. Leppink, and F. Paas. 2014. "Improving Critical Thinking: Effects of Dispositions and Instructions On economics Students' Reasoning Skills." *Learning and Instruction* 29: 31-42.

Heiskala, L., J. Erola, and E. Kilpi-Jakonen. 2020. "Compensatory and Multiplicative Advantages: Social Origin, School Performance, and Stratified Higher Education Enrolment in Finland." *European Sociological Review*: 1–15.

Hetmanek, A., K. Engelmann, A. Opitz, and F. Fischer. 2018. "Beyond Intelligence and Domain Knowledge." In *Scientific Reasoning and Argumentation - the Roles of Domain-Specific and Domain-General Knowledge*, edited by F. Fischer, C. Chinn, K. Engelmann and J. Osborne, 205-226. New York: Routledge.

Hu, L. and P. Bentler. 1999. "Cutoff Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria Versus New Alternatives." *Structural Equation Modeling: A Multidisciplinary Journal* 6 (1): 1-55.

Hu, L. and P. Bentler. 1998. "Fit Indices in Covariance Structure Modeling: Sensitivity to Underparameterized Model Misspecification." *Psychological Methods* 3 (4): 424.

Hyytinen, H., K. Holma, A. Toom, R. Shavelson, and S. Lindblom-Ylänne. 2014. "The Complex Relationship between Students' Critical Thinking and Epistemological Beliefs in the Context of Problem Solving." *Frontline Learning Research* 2 (5): 1-25.

Hyytinen, H., K. Nissinen, J. Ursin, A. Toom, and S. Lindblom-Ylänne. 2015. "Problematising the Equivalence of the Test Results of Performance-Based Critical Thinking Tests for Undergraduate Students." *Studies in Educational Evaluation* 44: 1-8.

Hyytinen, H. and A. Toom. 2019. "Developing a Performance Assessment Task in the Finnish Higher Education Context: Conceptual and Empirical Insights." *British Journal of Educational Psychology* (89): 551-563.

Hyytinen, H., A. Toom, and R. Shavelson. 2019. "Enhancing Scientific Thinking through the Development of Critical Thinking in Higher Education." In *Redefining Scientific Thinking for Higher Education*, edited by M. Murtonen and K. Balloo, 59-78. Cham: Palgrave Macmillan.

Hyytinen, H., J. Ursin, K. Silvennoinen, K. Kleemola, and A. Toom. 2020. "The Dynamic Relationship between Cognitive Response Processes and Self-

Regulation of Cognition in Critical Thinking Assessments." *Manuscript Submitted for Publication*.

International Test Commission. 2018. “ITC Guidelines for Translating and Adapting Tests (Second Edition)”. *International Journal of Testing, 18*(2), 101-134.

Kleemola, K. and H. Hyytinen. 2019. "Exploring the Relationship between Law Students’ Prior Performance and Academic Achievement at University." *Education Sciences* 9 (3): 236.

Klein, S., R. Benjamin, R. Shavelson, and R. Bolus. 2007. "The Collegiate Learning Assessment: Facts and Fantasies." *Evaluation Review* 31 (5): 415-439.

Kuhn, D. 2019. "Critical Thinking as Discourse." *Human Development* 62 (3): 146-164.

Leighton, J. 2019. "The Risk–return Trade-off: Performance Assessments and Cognitive Validation of Inferences." *British Journal of Educational Psychology* 89 (3): 441-455.

McClelland, D. 1973. "Testing for Competence rather than for "Intelligence"." *American Psychologist* 28 (1): 1-14.

Meade, A., E. Johnson, and P. Braddy. 2008. "Power and Sensitivity of Alternative Fit Indices in Tests of Measurement Invariance." *Journal of Applied Psychology* 93 (3): 568.

Messick, S. 1994. *Alternative Modes of Assessment, Uniform Standards of Validity*. Princeton: Educational Testing Service.

———. 1995. "Validity of Psychological Assessment: Validation of Inferences from Persons' Responses and Performances as Scientific Inquiry into Score Meaning." *American Psychologist* 50 (9): 741.

Muthén, L. and B. Muthén. 2017. *Mplus User’s Guide. Eight Edition*. Los Angeles: Muthén & Muthén.

O’Connor, B.. 2000. "SPSS and SAS Programs for Determining the Number of Components using Parallel Analysis and Velicer’s MAP Test." *Behavior Research Methods, Instruments, & Computers* 32 (3): 396-402.

O’Hare, L. and C. McGuinness. 2015. "The Validity of Critical Thinking Tests for Predicting Degree Performance: A Longitudinal Study." *International Journal of Educational Research* 72: 162-172.

Shavelson, R. 2018. "Discussion of Papers and Reflections on "Exploring the Limits of Domain-Generality"." In *Scientific Reasoning and Argumentation: The Roles of*

Domain-Specific and Domain-General Knowledge, edited by F. Fischer, C. Chinn, K. Engelmann and J. Osborne, 112-118. New York: Routledge.

———. 2010. *Measuring College Learning Responsibly: Accountability in a New Era*. Stanford: Stanford University Press.

Shavelson, R., G. Baxter, and X. Gao. 1993. "Sampling Variability of Performance Assessments." *Journal of Educational Measurement* 30 (3): 215-232.

Shavelson, R., O. Zlatkin-Troitschanskaia, K. Beck, S. Schmidt, and J. Marino. 2019. "Assessment of University Students' Critical Thinking: Next Generation Performance Assessment." *International Journal of Testing*: 1-20.

Shavelson, R., O. Zlatkin-Troitschanskaia, and J. Marino. 2018. "International Performance Assessment of Learning in Higher Education (iPAL): Research and Development." In *Assessment of Learning Outcomes in Higher Education*, edited by O. Zlatkin-Troitschanskaia, M. Toepper, H. Pant, C. Lautenbach and C. Kuhn, 193-214. Cham: Springer.

Solano-Flores, G. and M. Chía. 2017. "Validation of Score Meaning in Multiple Language Versions of Tests." In *Validation of Score Meaning for the Next Generation of Assessments: The use of Response Processes*, edited by K. Ercikan and J. Pellegrino, 127-137. New York: Routledge.

Tabachnick, B. and L. Fidell. 2014. *Using Multivariate Statistics*. 6th ed. Harlow: Pearson.

Tuononen, T., A. Parpala, and S. Lindblom-Ylänne. 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*: 1-14.

van der Zanden, P., E. Denessen, A. Cillessen, and P. Meijer. 2019. "Patterns of Success: First-Year Student Success in Multiple Domains." *Studies in Higher Education* 44 (11): 2081-2095.

Vandenberg, R. and C. Lance. 2000. "A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research." *Organizational Research Methods* 3 (1): 4-70.

Zahner, D.. 2013. *Reliability and Validity CLA+*. New York: Council for Aid to Education.

Zahner, D. and A. Ciolfi. 2018. "International Comparison of a Performance-Based Assessment in Higher Education." In *Assessment of Learning Outcomes in*

Higher Education, edited by O. Zlatkin-Troitschanskaia, M. Toepper, H. Pant, C. Lautenbach and C.Kuhn, 215-244. Cham: Springer.

Zlatkin-Troitschanskaia, O., R. Shavelson, S. Schmidt, and K. Beck. 2019. "On the Complementarity of Holistic and Analytic Approaches to Performance Assessment Scoring." *British Journal of Educational Psychology* (89): 468-484.

Zlatkin-Troitschanskaia, O., R. Shavelson, and C. Kuhn. 2015. "The International State of Research on Measurement of Competency in Higher Education." *Studies in Higher Education* 40 (3): 393-411.

Table 1. Descriptive values and correlations between critical thinking measures in transformed Z scores (** significant at .01 level)

	APS	WE	WM	SQR	CRE	CA
Analysis and problem-solving (APS)	-					
Writing effectiveness (WE)	.771**	-				
Writing mechanics (WM)	.576**	.662**	-			
Scientific and quantitative reasoning (SQR)	.163**	.175**	.160**	-		
Critical reading and evaluation (CRE)	.201**	.225**	.227**	.138**	-	
Critiquing an argument (CA)	.192**	.192**	.166**	.143**	.157**	-
Mean	.00	.00	.00	.00	.00	.00
SD	1.00	1.00	1.00	1.00	1.00	1.00
Min	-3.95	-3.63	-4.09	-2.15	-2.54	-2.17
Max	2.66	2.44	2.15	2.57	2.02	1.55

Table 2. Factor loadings and communalities for a 1-factor solution.

	Factor 1	**Communalities**
Analysis and problem-solving (APS)	.83	.68
Writing effectiveness (WE)	.93	.87
Writing mechanics (WM)	.71	.50

Scientific and quantitative reasoning (SQR)	.20	.04
Critical reading and evaluation (CRE)	.26	.07
Critiquing an argument (CA)	.22	.05

Table 3. Factor loadings and communalities for a 2-factor solution.

	Factor 1	**Factor 2**	**Communalities**
Analysis and problem-solving (APS)	.82	-.00	.67
Writing effectiveness (WE)	.99	-.07	.90
Writing mechanics (WM)	.65	.10	.50
Scientific and quantitative reasoning (SQR)	-.01	.35	.12
Critical reading and evaluation (CRE)	.02	.41	.18
Critiquing an argument (CA)	-.00	.38	.15

Table 4. Differences in fit indices between configural, metric, scalar and partial scalar models.

	χ^2	**df**	**CFI**	**$\Delta\chi^2$ p-value**	**ΔCFI**
Configural model	85.16	20	.970	-	-
Metric model	89.12	23	.970	-	-
vs. configural	-	-	-	.266	.000
Scalar model	109.30	25	.961	-	-
vs. metric	-	-	-	<.001	.009
Partial scalar model	89.46	24	.970	-	-
vs. metric	-	-	-	.560	.000

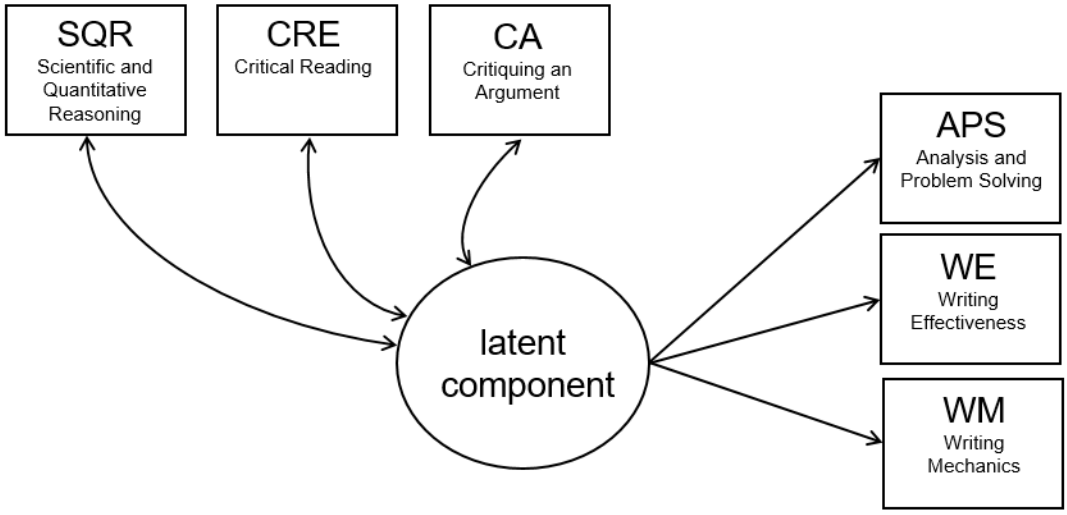

Figure 1. Suggested measurement model.

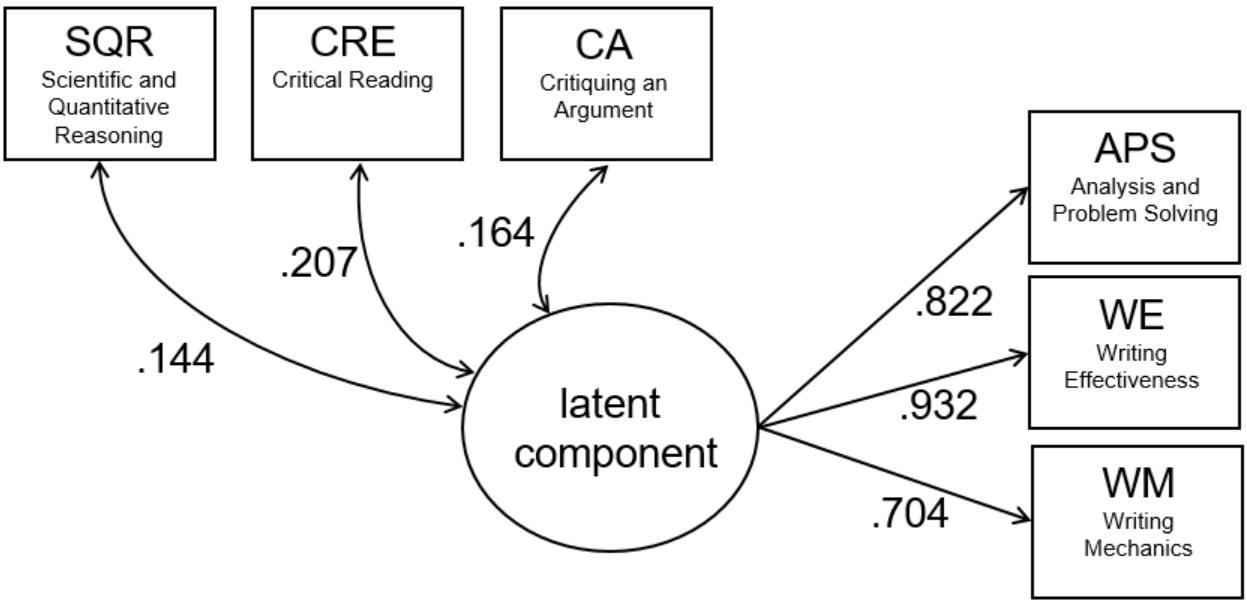

Figure 2. Factor loadings and correlations