3.6 Performance Assessment of Generic and Domain-Specific Skills in Higher Education Economics

Nagel, M.-T., Zlatkin-Troitschanskaia, O., Schmidt, S., and Beck, K.

Abstract

Following criticisms by employers about academic graduates' lack of 21st century skills, students need to develop skills such as professional knowledge, critical thinking and problem solving. Accordingly, there is a demand for suitable assessments of these skills. One approach is to develop a performance assessment using tasks adapted from real-world decision-making and judgment situations that students and graduates have to face in academic and professional domains. Such tasks employ real-life scenarios and require generic and domain-specific skills in different facets to handle a given problem adequately. In this paper, we present a newly developed performance assessment that aims to measure such skills among higher education economics students and graduates of economics and we report results from two validation studies.

Keywords

Generic skills, domain-specific knowledge, critical thinking, graduates of economics, performance assessment, validation

O. Zlatkin-Troitschanskaia et al. (Hrsg.), *Student Learning in German Higher Education* https://doi.org/10.1007/978-3-658-27886-1_14

Funding Details

This work was supported by the German Federal Ministry of Education and Research with the funding number 01PK15001A.

1 Introduction

In Germany, we are still lacking performance assessments that meet the methodological requirements for measuring university students' generic higher-order cognitive skills and that further meet the demands of the curriculum-instruction-assessment triad (Pellegrino et al. 2001) in higher education (Zlatkin-Troitschanskaia et al. 2018a). Initial attempts to adapt and validate existing performance tasks on critical thinking and problem solving from the U.S. for German contexts revealed significant limitations (for the adaptation and validation of CLA+ tasks in Germany, see Zlatkin-Troitschanskaia et al. 2018b). In particular, transferring the constructs underlying the tasks as well as developing the according scoring rubrics, which are necessary to rate the students' performance in critical thinking, have been challenging. It turned out that, for instance, cultural differences presented us with vast problems of interpretation and comparison.

Therefore, we developed an innovative performance assessment learning (PAL) task (Shavelson et al. 2019; Zlatkin-Troitschanskaia et al. 2019b) to measure these skills among higher education economics students and graduates in Germany. The PAL task consists of a realistic short-frame scenario, where test takers are confronted with the succinct description of a situation and a resulting problem. The scenario is complemented by a document library of additional background information that varies in relevance, reliability, credibility and validity. Test takers are asked to react to the presented problem by using this information and writing a well-founded recommendation for action (Section 3).

In this paper, we present first results of an assessment of university students of economics using the PAL task. To control for their domain-specific knowledge, we used the WiWiKom test (Zlatkin-Troitschanskaia et al. 2019a; see also Schlax et al. in this volume). Student general cognitive ability was assessed by the intelligence test IST-2000 R (Liepmann et al. 2007). In addition, the test takers' final school-leaving grades as well as their grades in attended study modules in higher education economics were assessed. This study design allows for measuring and investigating the relationships between different facets of domain-specific and generic skills.

2 Conceptual Background and Research Hypotheses

Critical thinking is receiving attention as a 21st century skill that is internationally considered indispensable for students of all disciplines who want to be successful not only in the national, but also in the global context after graduation (Allgood and Bayer 2016; Allgood et al. 2015). Due to global economic change and the increasing internationalization of markets, critical thinking skills are considered an ever more important requirement (e.g., Willingham 2007; McGoldrick and Garnett 2013) – particularly for business and economics professions but also for the public.

Critical thinking is described in literature as a complex, multi-dimensional construct that often includes the elements problem solving, communication ability, media literacy, and other 21st century skills. Thus, without further clarification, the concept seems to be quite vague (e.g., Lai and Viering 2012). For the context of this study, we develop a working definition below.

While the importance of such higher-order cognitive skills is commonly accepted, they are not yet systematically taught in higher education (e.g., Browne et al. 1995; Arum and Roksa 2010). A current nationwide analysis of 32 German higher education degree programs and module descriptions in economics showed that, although critical thinking is considered an objective of teaching and learning outcomes in the respective curricula, it is generally not implemented explicitly or actively on the instructional side (Zlatkin-Troitschanskaia et al. 2018b). According to the curriculum-instruction-assessment triad (Pellegrino et al. 2001), this is predominantly due to a lack of learning tools (i.e. appropriate methods to teach critical thinking) and corresponding testing instruments for conveying, assessing and fostering such skills in economics (Hoyt and McGoldrick 2012).

In our study, we follow a holistic understanding, assuming critical thinking not to be the sum of other individual skills, but an integration of various sub-capabilities (Shavelson et al. 2018a; Zlatkin-Troitschanskaia et al. 2019b). Based on the current state of research on critical thinking (e.g., Halpern 2014; Liu et al. 2014; Paul and Elder 2006; Facione 2000), we developed a systematical synthesis in which we differentiate between the following key dimensions and their central subdimensions to operationalize *critical thinking* and thereby form a basis for the assessment of higher education students (Shavelson et al. 2019; Zlatkin-Troitschanskaia et al. 2019b):

1. Evaluating and using information and sources in terms of relevance to the argument, reliability, validity, credibility of sources;

2. Recognizing, evaluating and using arguments and their components (such as claims, support, beliefs, assumptions or proven facts) with regard to evidentiality, objectivity, validity, consistency;
3. Developing sound and valid arguments based on information provided in the task, integrating additional information into coherent arguments, and structuring arguments consistently. This includes avoiding logical inconsistencies, identifying omissions and weaknesses, and evading decision-making errors or biases (e.g., due to "fast thinking" in contexts that call for "slow thinking"; Kahneman 2011; West et al. 2008);
4. Recognizing main and ancillary effects and evaluating consequences of decision-making and associated actions;
5. Appropriately communicating the most suitable course of action based on the given task prompt, i.e., making an evaluative judgment, explaining a decision, recommending a course of action by suggesting a solution to a problem.

Taking into account these facets and the various existing descriptions of critical thinking, the following working definition was developed for the study:

> Students with advanced critical thinking skills question existing assumptions and opinions, and recognize and evaluate the relevance and reliability of provided information. Based on this judgement, logical or causal interrelationships are identified, consequences are considered and conclusions are drawn. Consequently, own arguments and opinions are formed, which in turn are reflected upon and corrected if necessary. Finally, the arguments and conclusions are communicated (written or verbally) in an understandable and convincing way.

The PAL task is designed to measure the critical thinking skills of higher education students or graduates according to the aforementioned working definition and can be applied to different domains including economics, as the context of the task relates to socio-economic topics. The PAL task builds on prior research on generic higher-order cognitive skills and, in particular, on the concept of critical thinking, which – in short – is described as the ability to analyze problems, evaluate claims and information, draw inferences, and weigh decisions with regard to their consequences. The task represents the next generation of performance assessments due to its innovative approach to performance assessment (Shavelson et al. 2019; Zlatkin-Troitschanskaia et al. 2019b).

This newly developed PAL task was comprehensively validated in two subsequent studies in Germany (Shavelson et al. 2019; Zlatkin-Troitschanskaia et al. 2019b) in accordance with the internationally established Standards for Educational and Psychological Testing by AERA et al. (2014). In this paper, we focus on the valida-

tion criterion 'relation to other variables', i.e., convergent and discriminant validity, to examine the relationships between different facets of domain-specific and generic skills: critical thinking, domain-specific knowledge and general cognitive ability. Using the method of comparing known groups (Hattie and Cooksey 1984) we also investigate the specificity and sensitivity of the newly developed PAL task for the *domain of economics*.

The research literature remains inconclusive in terms of the extent to which critical thinking is a domain-specific or domain-independent higher-order ability. It is often hypothesized that critical thinking itself is a generic skill, which can, however, only be conveyed and learned in concrete domain-specific contexts (e.g., Fives and Dinsmore 2017). The ability to perceive and process a domain-specific problem such as the scenario presented in the PAL task as well as perform the according solution requires a substantial level of expertise in this domain, since a profound understanding of the subject area is necessary for the handling of such a complex task (Alexander et al. 2016; Pellegrino and Hilton 2012). Accordingly, *graduate students (master), who should have greater domain-specific expertise, can be expected to perform at a higher level in terms of their critical thinking abilities in the domain of economics than undergraduate students (bachelor) (Hypothesis 1).*

The PAL task focuses on critical thinking in a socio-economic context requiring not only domain-specific knowledge and an understanding of basic economic principles but also domain-independent higher-order skills, in terms of discriminant validity (e.g., Messick 1989). Therefore, we expect *a positive but relatively weak correlation between the critical thinking performance of students as measured by the PAL task and their performance in the domain-specific economic knowledge WiWiKom test (Hypothesis 2).*

As PAL also measures other higher-order cognitive abilities besides domain-specific knowledge, which are needed to complete this holistic task, it can be assumed *that a good performance in the PAL task is also slightly positively correlated with general cognitive abilities (Hypothesis 3)* (which were assessed through intelligence sub-tests from the task groups "Choosing figures" and "Matrices" from the German intelligence test IST-2000 R as well as final school-leaving grades).

To further examine the domain specificity of PAL, we also applied the method of comparing known groups by comparing the results of students with a major in economics with those of students without a major in economics. *We expect the students in economics to do at least slightly better in the PAL task than students enrolled in other subjects (Hypothesis 4).*

In addition, the influence of prior domain-specific knowledge, which might be due to previous education with an economic focus (e.g., completed commercial vocational training) was also controlled for.

3 Methods

3.1 Sample

For Germany, two samples have been surveyed within two subsequent validation studies that took place in the winter semester of 2017/2018 and the summer semester of 2018. Overall, 55 students from a German university participated – 25 undergraduates (bachelor's degree) and 30 graduates (master's degree). For their participation in the voluntary validation studies students could choose between an incentive of €20 or credits for a study module.

We contacted all 44 master's students enrolled in economics education at this university inviting them to participate in the study. Thus, we have a 68 % participation rate of all economics education master's students. Taking into account the sociodemographic characteristics, this sub-sample can be considered representative for this study domain (Tables 1a and 1b).

The procedure for recruiting bachelor's students was the same as for master's students, although in this case we managed to encourage slightly less than half of all students enrolled at the university in this course of study to participate. Based on the descriptive characteristics of the 25 participating bachelor's students and a descriptive comparison of characteristics with the results of the Germany-wide representative WiWiKom study with bachelor's students, this sub-sample can also be considered representative (Zlatkin-Troitschanskaia et al. 2019a; see also Schlax et al. in this volume). Since about half of the test takers were in the last year of their bachelor's studies and most of the graduate students were in the last year of a master's degree, this study not only provides preliminary data on advanced undergraduates' critical thinking skills but also indicates the level of critical thinking in graduate students towards the end of their studies.

Tables 1a and 1b show the descriptive statistics of the sample.

Table 1a Sample description

Attributes	***N* = 55**
Gender	
Female	38 (69.1 %)
Male	17 (30.0 %)
Degree	
Bachelor	25 (45.3 %)
Master	30 (54.5 %)
Degree course	
Economic studies	46 (83.7 %)
Other	9 (16.3 %)
Completed vocational training	
Yes	23 (41.8 %)
No	30 (54.5 %)
Not specified	2 (3.6 %)
Completed internship	
Yes	42 (76.4 %)
Not specified	9 (16.4 %)

Table 1b Sample description (continued)

	N = 55		
Attributes	N	Mean	Std. Dev.
Age	54	24	3.44
Semester (Bachelor)	25	4.36	1.47
Semester (Master)	30	2.03	1.22
University entry qualification grade*	54	2.21	0.60

Note. *Grades vary from 1.0 (best) to 5.0 (worst).

3.2 Test Instruments and Administration

The PAL task, called "Wind Turbine", has been constructed from authentic alternative energy source cases with meaningful consequences for a myriad of actors depending on the decisions and actions taken (Shavelson et al. 2019). More precisely, it focuses on the decision of a small-town council regarding whether or not to acquire and set up wind turbines on communal land near the town. Test-takers are presented with a document library and are asked to evaluate available information (e.g., a newspaper article, web documents, wind turbine schematics, stake-

holder interests as well as selected technical, economic, territorial law, settlement and wildlife data) that varies in terms of reliability, validity, and risk of bias or judgmental error. The task prompt requires participants to write an argumentative statement and recommend a course of action whether or not to set up the wind turbines and which further measures to take (for examples, see Shavelson et al. 2019; Zlatkin-Troitschanskaia et al. 2019b). Test-takers are asked to use only the information provided and told that there is no right or wrong answer but that answers can vary in terms of their justifiability. They are also informed of the main scoring criteria. In addition to judging the trustworthiness and relevance of the different library documents, test-takers need to develop arguments for or against wind turbine construction. This process requires them to assess the value of each document while taking into account possible bias or motives for hidden agendas, such as personal profit and consequences for the community or individual residents. Task difficulty is fine-tuned by the nature of the information presented (reliability, validity, bias/error), the number of information sources, and the various points to consider by the test taker (e.g., stakeholder interests, consequences of the decision).

A rating scheme for scoring the written responses for the wind turbine PAL task has been developed based on the previously described definition of critical thinking (for details, see Zlatkin-Troitschanskaia et al. 2019b). It divides the students' response texts into four main dimensions with 4–9 performance criteria each (23 categories in total), whereby the facets 1–3 were grouped into two dimensions for reasons of formulating adequate behavior anchors. Then, by assessing individual PAL responses, 6 scoring anchors were formulated; on this basis, the participants' task performance can be assessed on a scale of 1–6 (for rubric 1 as an example, see Shavelson et al. 2019).

After rater training and a random designation of the individual PAL responses to two of a total of four raters, the responses were assessed using the developed rating scheme resulting in every task response of each participant being independently assessed by two raters.[1] To allow for a comparison of average performance results between the dimensions in spite of their different number of sub-categories and thus of attainable assessment points, the score of every dimension was divided by the number of its performance criteria. As a result, the calculated mean scores of every dimension can vary on a scale from 1 (requirements not fulfilled) to 6 (requirements fulfilled).

To assess knowledge and understanding of economics, we employed the *WiWiKom* test, which was validated in the representative, Germany-wide

1 There are sufficient inter-rater reliabilities between .7 and .85 (for further details, see Shavelson et al. 2019).

WiWiKom study with over 9,000 economics students according to the Standards for Educational and Psychological Testing (Zlatkin-Troitschanskaia et al. in 2019a; see also Schlax et al. in this volume). The WiWiKom test combines 15 items from the adapted and validated German version of the standardized *Test of Economic Literacy, Fourth Edition (TEL IV)* (Walstad, Rebeck, and Butters 2013) and 10 items from the adapted and validated German version of the *Test of Understanding in College Economics, Fourth Edition* (TUCE IV, Walstad et al. 2007; for validation and adaptation in Germany, see Zlatkin-Troitschanskaia et al. 2014). The items of TEL IV operationalize basic principles of economics (such as the supply-demand model), complemented by five TUCE IV items from the microeconomics part and five TUCE IV items from the macroeconomics part. Every question offered 4 possible answers in multiple-choice format, with only one correct option.

The figural-spatial intelligence as indicator of general (fluid) intelligence of the students was measured using the two task groups "Choosing figures" and "Matrices" from the German intelligence test IST-2000 R (Liepmann et al. 2007). In total, this test includes 12 task groups, of which the two aforementioned were considered most suitable as good indicators of general intelligence (Liepmann et al. 2007). Each scale consists of 20 tasks. The participants have 7 minutes to complete the tasks in "Choosing figures" and 10 minutes for "Matrices". In the former task group, students have to work out which of five given figures can be created by piecing together ten fragments. In the latter, students are given figure matrices built according to a certain rule and have to decide which of the five figure options would complete the matrix according to the rule.

The PAL task as well as the WiWiKom test were computer- and online-based. In addition to the students' written answers, some data on the response processes (required time, information used for problem solving) were also collected and included in the analyses. The intelligence test was administered on-site via paper-pencil questionnaires under controlled conditions to ensure that the task was carried out properly.

Further socio-demographic information expected to affect test performance was collected as well. Prior studies have shown that task readability impacts the test performance of migrant students (e.g., Happ et al. 2019) and was therefore systematically controlled for in our study. Furthermore, several studies have demonstrated the suitability of the higher education entrance qualification as a reliable indicator for generic cognitive skills (e.g., Kobrin et al. 2008). Other indicators of relevant expertise in the context of solving the PAL task, such as completed commercial or vocational training, were surveyed as well as they might influence task performance and should therefore be considered.

The total test time for the PAL task is 60 minutes plus about 60 minutes for the other tests and questionnaires. The limited time puts the participants under pressure, requiring them to limit their focus by choosing the most relevant documents from the library and narrowing down their arguments. This leads to the active decision to discard certain information without considering it as otherwise the task would not be completed in time. The test-takers completed the task on computers provided by the project coordinators. In addition to the students' written answers, further data was collected on the response processes (e.g., behavior while working on the tests)[2], which was part of different analyses.

4 Results

To test *Hypothesis 1* (H1), expecting that graduate students (master) perform at a higher level than undergraduate students (bachelor), we conducted various analyses based on the written and scored responses to the PAL task. The lengths of students' responses vary significantly: the longest response comprises 1365 words and the shortest 68 words, with a mean word count of 495 ($SD = 218.7$). A significant difference can also be seen in the length of the answers between the bachelor's and master's students, with the master's students writing longer answers on average: The t-test yields $p = 0.0185$, mean bachelor = 421.08 ($SD = 200.50$), and mean master = 561 ($SD = 216.31$).

The analysis of the test performance shows that the length of the responses (number of words in the written statement) significantly correlates with the assessed test result measurements of the respective texts (Pearson correlation $r = 0.6$, $p = .00$). The master's students accordingly perform better on average than the bachelor's students (Table 3).

On average, the participants scored 83 out of a maximum of 138 points, with a minimum of 32 points and a maximum of 117 points. The score distribution has a skewness of -1.5 and kurtosis of 7.2, indicating test scores slightly skewed to the left. Although none of the participants earned a perfect overall score, some earned a full six points on some of the individual dimensions. Bachelor's students' average score was 81 points, with a minimum of 32 points and a maximum of 107 points, whereas the master's students achieved 85 points on average with a minimum of

2 For instance, the observation of the test-takers captured a wide range of approaches, with one extreme being participants investigating all the provided links to external information, whilst students on the other extreme only took into account the information provided on the task sheet itself.

56.5 points and a maximum of 117 points. This shows a higher level of critical thinking ability in graduate students and thus, at the descriptive level, the results are in line with our assumption (H1). Observing, however, the quartile level with respect to both groups, most graduate students place within the third quartile and only few in the fourth quartile of attainable points. Table 2 shows the distribution of students' PAL scores subdivided in quartiles. In spite of the obvious differences between performance scores of both groups, a t-test does not yield significant results, with $p = 0.15$, mean bachelor = 81.12 ($SD = 16.84$), and mean master = 87.6 ($SD = 16.04$).

Table 2 Distribution of students' PAL scores in quartiles

Quartile	Bachelor		Master	
	N	%	*N*	%
1 (0 – 34.5 points)	1	4 %	-	-
2 (35 – 69 points)	3	12 %	5	17.24 %
3 (69.5 – 103.5 points)	20	80 %	19	76 %
4 (104 – 138 points)	1	4 %	5	17.24 %

Further analyses of the four subdimensions reveal in which subskills students, on average, have better or worse performance. Table 3 shows a comparison of the average PAL subscale scores of bachelor's vs. master's students. In all subscales, master's students perform better than bachelor's students. However, both groups score lowest in subscale 3, "Recognizing and evaluating consequences of decision-making and actions", with average scores below 3 points, which constitutes less than half of the points that could be reached and again underlines the difference to the other sub-scales. According to the working definition, this is a vital facet of critical thinking, as in comparison to the other facets it requires students to apply their highest cognitive and meta-cognitive abilities.

Bachelor's students achieved their best results in subscale 1, "Recognizing and evaluating the relevance, reliability, and validity of given information", in which both groups scored approximately 4 points; therefore, compared to the other categories, all of the students' abilities rank high for this subscale. The same holds true for most students with regard to subscale 4, "Writing effectiveness and mechanics", in which the master's students achieved their best results (bachelor: 3.8 points, master: 4.2 points). The main difference between the two samples might be traced back to training in the context of bachelor's theses. The bachelor's thesis

is the first opportunity in the course of study for which a longer, more systematic scientific text must be written, which could significantly enhance students' writing skills. For subscale 2, "Evaluating and making a decision", in comparison to the other subscales, both groups' mean results are rather average, with only a small difference between bachelor's (3.5) and master's (3.8) students.

Table 3 Group performances for the PAL subscales

	Group Average Performance in PAL	
Subscale	Bachelor	Master
1: Recognizing and evaluating the relevance, reliability, and validity of given information	4.03 (SD = .70)	4.15 (SD = .62)
2: Evaluating and making a decision	3.47 (SD = .80)	3.81 (SD = .75)
3: Recognizing and evaluating consequences of decision-making and actions	2.69 (SD = .84)	2.87 (SD = .79)
4: Writing effectiveness and mechanics	3.84 (SD = .92)	4.20 (SD = .95)

Overall, the findings from the four subdimensions indicate a significant difference between undergraduate and graduate students only for „Writing effectiveness and mechanics". With regard to the three abovementioned subdimensions 1, 2, and 3, students only show a high level of underlying abilities, indicated by high performance scores, in one subdimension, "Recognizing and evaluating the relevance, reliability, and validity of given information". In the two other subdimensions, performance levels are remarkably low in comparison, particularly for graduate students. Based on these findings, H1 cannot be rejected, as – even if only marginally in some cases – overall the graduate students performed better than the undergraduates.

As described in *Hypothesis 2* (H2), due to the economics-related context of the PAL task, it can be inferred that successfully solving the PAL task correlates with a high level of domain-specific knowledge and economic expertise. The 25 items of the WiWiKom test were evaluated as either incorrect (0 points) or correct (1 point) and the individual total scores as well as the mean scores were calculated for each of the 36 participants who completed the test[3]. On average, the participants scored 16.7 out of 25 possible points (*SD* = 4.17), with a minimum of 7 points and a maximum of 23 points. The distributions have a skewness of -.65 and kurtosis of 2.69 which implies a skewed to the left distribution of test scores. Although none

3 Not all students completed all three tests: The results of 8 students are missing for the intelligence test and of 19 students for the WiWiKom test.

of the participants achieved a maximum score, an examination of the answers at item level indicates that the participants were able to complete the test in the given time. In comparison, a nationwide German sample of 7,571 bachelor's students of economics in the WiWiKom study achieved an average score of 13.3 points (*SD* = 4.39; Zlatkin-Troitschanskaia et al. 2019a). The students in our test sample distinguish themselves from this latter sample by a comparatively high level of economics knowledge.

Using a Spearman test, we found no significant correlation between the PAL scores and the WiWiKom test scores (r_s = .11, p = 51.2, n = 36). The correlation also remains insignificant if calculated only for economics students (r_s = -.04, p = 0.85, n = 28).

For further analyses, the relationship between the PAL task and economic knowledge was examined by means of cross-classified tables which investigated the interrelations of different performance levels in both tests. For this purpose, using the median of the PAL scores and of the WiWiKom test, the sample was divided up into equally large groups of high and low performers. Table 4 shows the cross-classification of overall performance in PAL and in the WiWiKom test. For the cross-classified tables of the subscales, the groups were clustered according to the score of the respective subscale in the PAL task.

Table 4 Cross-classification of overall performance in PAL and the WiWiKom test

Performance in PAL	Performance in WiWiKom test		
	Low (=0)	High (=1)	Total
Low	7 46.67 %	8 53.33 %	15 100.00 %
High	8 38.10 %	13 61.90 %	21 100.00 %
Total	15 41.67 %	21 58.33 %	36 100.00 %

The cross-classification shows that 47 % of students who performed low on the PAL task are also low-performers on the WiWiKom test. However, the majority (53 %) of low performers comprises students who performed well in the WiWiKom test. Yet again, it becomes apparent that despite high levels of domain-specific knowledge, overall, economics students did not perform particularly well in the PAL task. This further supports the hypothesis (H2) that many of these students were not able to apply their domain-specific knowledge to solve the economics problem.

Conversely, 38 % of the students with high performance in the PAL task show low performance in the WiWiKom test. Over one third of students achieved good results in the PAL task despite low levels of domain-specific knowledge, which they evidently were able to compensate with higher general cognitive abilities (probably by reasoning and adequate use of the given information). A majority (62 %) performed well in both tests. The chi-squared test is not significant for the cross-classified table ($\chi^2(36) = 0.26$, $p = .607$), which leads to the overall conclusion that high economics knowledge does not necessarily indicate a high performance in the PAL task.

Similar findings were derived from the cross-classified table for the subdimension "Recognizing and evaluating the relevance, reliability, and validity of given information" ($\chi^2(36) = 0.39$, $p = .53$), in which, on average, participants achieved the highest scores. The findings for the subdimension "Evaluating and making a decision" ($\chi^2(36) = 1.03$, $p = .31$) showed that two thirds of high performers in the WiWiKom test performed poorly in the PAL task. Based on these results, *H2*, assuming a positive but relatively weak correlation between the critical thinking performance of students and their level of economic knowledge, must be rejected.

As the PAL task is constructed to assess higher cognitive skills, a slight correlation with the general cognitive abilities measured in the intelligence test was expected (H3).

The results of the 20 intelligence test items in each of the two employed scales were calculated (0 = wrong answer, 1 = right answer) and participants could achieve a maximum of 40 points in total on the test. On average, the students achieved a score of 16.9 ($SD = 5.33$) with a minimum of 7 and a maximum of 29 points. Because of skewness to the left of the PAL results, a Spearman correlation was conducted. As expected, the results show a significant weak correlation between the PAL scores and the intelligence test scores ($r_s = .35$, $p = .02$, $n = 46$).

With regard to the students' school-leaving grades, there were no related differences in performance in the PAL task; students with a final grade of 2.0–2.9 performed better than students with either a final grade of 1.0 – 1.9 or 3.0 and higher. Table 5 shows the means of PAL scores subdivided by different groups. The Spearman correlation of these two performance measures indicates no significant correlation ($r_s = -.03$, $p = .82$, $n = 47$). Based on these findings, H3 cannot be rejected.

To test for the group differences assumed in *Hypothesis* 4 (H4) (Table 5), we conducted a t-test which shows no significant differences between students of economics and students with other majors with regard to average test scores ($p = 0.69$). Thus, contrary to our expectations (H4), studying economics does not appear to

provide domain-specific knowledge that enhances the ability to solve these tasks.[4] Based on these results, *H4* must be rejected.

Table 5 Means of PAL-scores of different groups

	Group Average Performance in PAL
Variables	$N = 55$
Degree	
Bachelor	3.53 (SD = .73)
Master	3.68 (SD = .67)
Subject	
Economics studies	3.62 (SD = .68)
Other	3.52 (SD = .82)
University entry qualification grade	
1.0 – 1.9	3.62 (SD = .85)
2.0 – 2.9	3.66 (SD = .60)
3.0 – 3.9	3.55 (SD = .69)
Completed Vocational Training	
No	3.58 (SD = .79)
Yes (commercial)	3.69 (SD = .59)

5 Discussion and Conclusion

This study primarily presented findings on performance-oriented assessment of key facets of domain-specific and generic skills such as critical thinking and content knowledge among economics students at a German university. In this validation study, the approach of comparing known groups by assessing undergraduate economics students as well as a control group of master's students in economics degree programs has been applied. Additionally, for a domain-specific comparison, a control group of students with different major subjects was assessed in comparison to students having economics as their main study subject.[5] Beside the main finding that the construct of 'critical thinking' measured by the PAL task has turned out to be of discriminant validity, the results provide two important

4 The comparison between economics students (n=45, 12 bachelor's and 33 master's students) and students with other main subjects (n=8) shows that, on average, economics students achieve slightly better test results (economics: 3.6 points, others: 3.5 points).

5 The descriptive statistics of the sample are largely in line with the results of a Germany-wide survey of graduates of economics education at 20 universities, which further increases the external validity of the results of this study (see Kuhn et al. in this volume).

findings: (1) bachelor's students performed rather poorly in the PAL task on average and the level of critical thinking in master's students is much lower than one would expect of graduate students at the end of their 5-year academic education. (2) There were no significant differences between economics students and students with other main study subjects.

This study provides several pieces of evidence that many students in economics are not able to apply their domain-specific knowledge to solve realistic economics-related problems. This finding appears even more dramatic in light of performance in each of the four subdimensions. Overall test performance, which on average is mediocre, is primarily based on well-developed abilities in rather basic (1) information processing and (4) writing (Table 3). With respect to the two other subdimensions of critical thinking that were considered vital in the construct definition (i.e. (2) decision making and (3) dealing intellectually with consequences), most students displayed substantial deficits – even towards the end of their academic education, and also at the graduate level.

Overall, the findings indicate that although critical thinking skills are required both in curricula and as learning outcomes in economic higher education, these key aspects have been insufficiently or ineffectively nurtured so far, both in undergraduate and in graduate studies. Although students have a relatively solid level of economic knowledge and understanding, they clearly lack the ability to transfer this knowledge and to solve a concrete economics-related problem. Based on the results from the innovative performance assessment, it is urgently necessary to give careful consideration to why the highly ambitious teaching-and-learning objectives and the expected outcomes of higher education are possibly not reached at all.

Our results raise a number of questions, which require more in-depth analyses to gain differentiated insights as to how such higher cognitive skills can be effectively taught and enhanced during higher education in economics. Multiple-choice tests are, at least in Germany, usually used for examinations, while tests based on a case study are rare (e.g., Walstad 2001). Therefore, a testing effect cannot be ruled out; it is possible that, although students do gain domain-specific knowledge, they cannot demonstrate it in a PAL test due to the unfamiliarity with this type of test instrument. A controlled, experimental intervention study should be conducted with a posttest measurement design to measure such effects.

Additionally, this validation study was conducted using a small sample from a single university. Thus, the results presented here should be considered only as preliminary evidence for the level of critical thinking measured by this particular PAL task and they still leave room to expect better results based on an ample sample. Furthermore, in future studies, not only performance data, but also behavior

data and students' response processes while working on PAL tasks, such as log and gaze data, should be collected and integrated in analyses to assess which concrete cognitive and non-cognitive subskills are used to solve performance tasks.

References

Alexander, P. A., Singer, L. M., Jablansky, S., & Hattan, C. (2016). Relational reasoning in word and in figure. *Journal of Educational Psychology*, 108 (8), 1140–1152.

American Educational Research Association, American Psychological Association and National Council on Measurement in Education (AERA, APA and NCME). (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.

Allgood, S., & Bayer, A. (2016). Measuring College Learning in Economics. MPRA Paper No. 85104. Online: https://mpra.ub.uni-muenchen.de/85104/.

Allgood, S., Walstad, W. B., & Siegfried, J. J. (2015). Research on Teaching Economics to Undergraduates. *Journal of Economic Literature*, 53 (2), 285–325.

Arum, R., & Roksa, J. (2010). Academically adrift: Limited learning on college campuses. Chicago, IL: University of Chicago Press.

Browne, M. N., Hoag, J. H., & Boudreau, N. (1995). Critical Thinking in Graduate Economic Programs: A Study of Faculty Perceptions. *The Journal of Economic Education*, 26 (2), 177–181.

Facione, P. A. (2000). The disposition toward critical thinking: Its character, measurement, and relation to critical thinking skill. *Informal Logic*, 20 (1), 61–84.

Fives, H., & Dinsmore, D. L. (Eds.). (2017). The Model of Domain Learning: Understanding the Development of Expertise. New York, Abingdon: Routledge.

Halpern, D. F. (2014). *Thought and knowledge: An introduction to Critical Thinking*. New York: Psychology Press.

Happ, R., Nagel, M.-T., Zlatkin-Troitschanskaia, O., & Schmidt, S. (2019). How migration background affects master degree students' knowledge of business and economics. *Studies in Higher Education*. doi: 10.1080/03075079.2019.1640670.

Hattie, J., & Cooksey, R. W. (1984). Procedures for assessing the validities of tests using the" known-groups" method. *Applied Psychological Measurement, 8* (3), 295–305.

Hoyt G. M., & McGoldrick, K. (Eds.). (2012). *International Handbook on Teaching and Learning Economics*. Cheltenham, Northampton: Edward Elgar.

Kahneman, D. (2011). *Thinking, fast and slow*. New York, NY: Farrar, Straus and Giroux.

Kobrin, J. L., Patterson, B. F., Shaw, E. J., Mattern, K. D., & Barbuti, S. M. (2008). Validity of the SAT® for Predicting First-Year College Grade Point Average. *College Board Research Report* (5). New York: The College Board.

Lai, E. R., & Viering, M. (2012). Assessing 21st Century Skills: Integrating research findings. Paper presented at the annual meeting of the National Council on Measurement in Education. Vancouver, B.C., Canada.

Liepmann, D., Beauducel, A., Brocke, B., & Amthauer, R. (2007). *Intelligenz-Struktur-Test 2000 R (I-S-T 2000 R)*. Available on https://psychowissen.jimdo.com/psychologische-tests/intelligenztests/i-s-t-2000-r/

Liu, O. L., Frankel, L., & Roohr, K. C. (2014). Assessing Critical Thinking in Higher Education: Current State and Directions for Next-Generation Assessments. *ETS Research Report*, 14 (10).

McGoldrick, K., & Garnett, R. (2013). Big Think: A Model for Critical Inquiry in Economics Courses. *The Journal of Economic Education*, 44 (4), 389–398.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement (3rd ed., pp. 13–103)*. New York: Macmillan Publishing.

Paul, R. W., & Elder, L. (2006). Critical thinking: The nature of critical and creative thought. *Journal of Developmental Education*, 30 (2), 34–35.

Pellegrino, J. W., & Hilton, M. L. (2012). *Education for Life and Work: Developing Transferable Knowledge and Skills in the 21st Century*. Washington, DC: The National Academies Press.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing What Students Know: The Science and Design of Educational Assessment*. Washington, DC: National Academy Press.

Shavelson, R. J., Zlatkin-Troitschanskaia, O., Beck, K., Schmidt, S., & Marino, J. (2019). Assessment of University Students' Critical Thinking: Next Generation Performance Assessment. *International Journal of Testing*. doi: 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*. Wiesbaden: Springer.

Walstad, W. B. (2001). Improving Assessment in University Economics. *The Journal of Economic Education*, 32 (3), 281–294.

Walstad, W. B., Rebeck, K., & Butters, R. B. (2013). *Test of economic literacy: Examiner's manual*, 4th ed. New York: Council for Economic Education.

Walstad, W. B., Watts, M., & Rebeck, K. (2007). *Test of Understanding in College Economics: Examiner`s manual*, 4th ed. New York: National Council on Economic Education.

West, R. F., Toplak, M. E., & Stanovich, K. E. (2008). Heuristics and biases as measures of critical thinking: Associations with cognitive ability and thinking dispositions. *Journal of Educational Psychology*, 100 (4), 930–941.

Willingham, D. T. (2007). Critical thinking: Why is it so hard to teach? *American Educator*, 31 (2), 8–19.

Zlatkin-Troitschanskaia, O., Jitomirski, J., Happ, R., Molerov, D., Schlax, J., Kühling-Thees, C., Förster, M. & Brückner, S. (2019). Validating a Test for Measuring Knowledge and Understanding of Economics Among University Students. *Zeitschrift für Pädagogische Psychologie, 33*(2), 119–133.

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

Zlatkin-Troitschanskaia, O., Shavelson, R. J., & Pant, H. A. (2018a). Assessment of Learning Outcomes in Higher Education – International Comparisons and Perspectives. In C. Secolsky and B. Denison (Eds.). *Handbook on Measurement, Assessment and Evaluation in Higher Education (2nd ed.)*. New York: Routledge.

Zlatkin-Troitschanskaia, O., Toepper, M., Molerov, D., Buske, R., Brückner, S., Pant, H. A., Hofmann, S., & Hansen-Schirra, S. (2018b). 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)*. Cham: Springer.