Brückner and Zlatkin-Troitschanskaia
Empirical Research in Vocational Education and Training (2024) 16:2
https://doi.org/10.1186/s40461-024-00156-2

Empirical Research in Vocational Education and Training

RESEARCH Open Access

The role of confidence in the gaze bias effect among economics trainee teachers — results from a digital assessment of economic content knowledge

Sebastian Brückner[1]* and Olga Zlatkin-Troitschanskaia[1]

*Correspondence:
Sebastian Brückner
brueckner@uni-mainz.de
[1] Faculty of Law, Management, and Economics, Chair of Business and Economics Education, Johannes Gutenberg University Mainz, Jakob-Welder-Weg 9, D-55099 Mainz, Germany

Abstract
In the present study, we recorded the eye movements of 20 criterion-based selected trainee teachers in economics while they responded to 25 single choice (SC) items in an economic content knowledge (CK) test and rated their confidence for each response in a digital assessment. By using a multilevel modeling approach with crossed random effects, we confirmed prior findings from eye-tracking research on SC tests, which showed longer dwell time on the correct response options (attractor) and shorter dwell time on the distractors are positively linked to correct options. Furthermore, we identified an additional effect on dwell time on the attractor in a moderator model with participants who highly rated their confidence for correct response options. Thus, we identified a specific role of students' confidence in their CK on the gaze bias effect. We interpret these results in terms of students' actual understanding of test contents from assessments of their professional knowledge and draw implications for further research and teacher education.

Keywords Eye tracking, Content knowledge (CK), Student confidence, Gaze bias effect, Multilevel model, Teacher education

Introduction and research questions

Computer-based digital teaching-learning and assessment tools are increasingly being used in teacher education for vocational training to assess and promote professional teacher knowledge and its central facets such as content knowledge (CK) and pedagogical content knowledge (PCK) (Nilsson and Karlsson 2019; Zlatkin-Troitschanskaia et al. 2019a). Competent teachers should not only have sufficient professional knowledge (CK, PCK, PK or pedagogical knowledge), but also should be able to assess and apply it appropriately in a self-reflective manner (Schön 1987). The teachers' self-reflection on their role in the classroom, their personality, their situational teaching behavior as well as on their own learning processes and thus their professional knowledge has a high impact on the learning processes of the students (Korthagen 2004; Rodríguez et al. 2014). The development of the ability to self-reflect is not only relevant to correcting one's own misconceptions but also to establishing a strategic learning process (Shulman 1986). Self-reflection is shown to be particularly important in relation to a teacher's professional knowledge and its application (Feucht et al. 2017; Schön 1987).

In connection with teachers' learning and self-reflection abilities, the more specific question arises to what extent they are able to monitor and correctly assess the learning processes and their professional knowledge. While (teacher) students recognize the importance of professional knowledge, they often do not feel confident in their knowledge or, in contrast, overestimate it. One explanation in the literature refers to deficits in students' ability to evaluate their knowledge and recognize deficiencies (Kruger and Dunning 1999; Eva et al. 2004). In SC tests, which are widely established to measure professional knowledge in teacher studies, confidence in one's knowledge is not usually assessed. Therefore, teacher student participants may answer correctly for instance by simply guessing (Walstad et al. 2018). When responding to a content knowledge SC task and deciding one response option, this (type of) reasoning can be based on prior knowledge or subjective "good feelings", e.g., described by Kahneman (2011) as "System 2," i.e., a "subjective feeling of confidence". For instance, in a study that analyzed economics students while responding to economics knowledge test tasks, it was identified that naïve students tend to respond to these tasks, albeit with self-reported high uncertainty, with a bipolar superficial approach that reflects their good feelings more adequately than the elaborateness of their knowledge (Leiser and Aroch 2009). Some further studies explored to what extent students have confidence in their content knowledge or, in contrast, to what extent they respond to test questions based on a "strategically selected option" (e.g., guessing, Sanders et al. 2016).

Few studies consider not only the correctness of a response but also the students' confidence in their answers when investigating knowledge development (Khan et al. 2001; Gardner-Medwin 1995; Cordova et al. 2014). This research illustrates that the awareness of confidence in relation to professional knowledge has a major influence on the development of knowledge (Stankov 2013). Research on knowledge assessment and confidence testing (e.g., Bruno 1993; Davies 2002) indicates that confidence in the correctness of one's response to a task in a knowledge test can be considered as an appropriate indicator of the extent to which a student's response is based on knowledge vs. (strategic) guessing (Kolbitsch et al. 2008).

To confront teachers with the relation between their confidence and knowledge, confidence ratings have also been used in teacher education research (e.g., Dassa and Nichols 2019; Kim and Klassen 2018). The frequently observed tension between teachers' actual knowledge and their confidence rating has been researched intensively (e.g., Podgoršek and Lipovec 2017; Brückner and Zlatkin-Troitschanskaia 2018) and raises the question whether teachers are self-aware about the discrepancy in their perception of knowledge and their demonstrated understanding. Although this topic is relevant to a variety of domains in teacher education, initial eye tracking studies between the domains of physics and economics have shown that trainee teachers for vocational education in economics are less reliable in correctly assessing their content knowledge than trainee teachers in physics (Klein et al. 2019). Thus, a need has been identified to more comprehensively investigate the accuracy of self-assessments among teachers of vocational education in economics.

The CK of (trainee) teachers is usually assessed using (SC) items that are especially helpful in the context of the increasing digitalization of teaching-learning processes and are easy to implement and whose data are more readily fed into and scored in a Learning Management System (Parkes and Zimmaro 2016). SC tests are regularly used in courses as a part of audience or classroom response systems, as they are able to provide learners with immediate feedback on their actual performance and learning progress (Greving et al. 2020).[1] There are a number of tests available for assessing economic knowledge in vocational education which can be used in a valid way, especially for trainee teachers in economics (Walstad et al. 2007, 2013; Zlatkin-Troitschanskaia et al. 2019b).

To investigate actual comprehension and performance using computer-based digital CK and PCK tests, which are currently gaining increasing popularity in teacher education, it is necessary to analyze the processing of digital learning or testing materials by (trainee) teachers using eye-tracking in addition to the analysis of learning performance based on the CK/PCK test scores. An analysis of the response processes occurring while answering test items is important to gain insight into students' cognitive processing (Ercikan and Pellegrino 2017; Zumbo and Hubley 2017). To find out how attention is spatiotemporally directed to different item areas, eye-tracking studies have been carried out, which facilitate dedicated analyses of respondents' gaze behavior (Holmqvist et al. 2011). Recent literature reviews highlight the increasing importance of eye-tracking in in empirical research in (vocational) education (Mayer et al. 2023).

One particular focus of the current study is the interaction between information content processing and self-reflective abilities, which are measured as confidence in response correctness. Such self-reflective abilities are considered in research to be necessary components of economic teachers' professional economic knowledge in vocational education but have hardly been studied to date. Only the analysis of students' processing of content can reveal important information about the extent to which differences exist between trainee teachers that possess knowledge and rate their confidence in this knowledge differently. Therefore, in this study, the intraindividual spatiotemporal processing of trainee teachers of economics in their bachelor's program is investigated based on their response processes when answering a professional CK test in economics.

Based on the preliminary work of Brückner et al. (2020); Zlatkin-Troitschanskaia et al. (2019b), this paper presents an eye-tracking study examining how economics trainee teachers perform on an economics knowledge test administered digitally and focuses on two research questions (RQ). With the first RQ, we seek to confirm findings on SC items in teacher education using the gaze bias effect (for a definition, see Sect. 2.1) identified in prior research (Lindner et al. 2014).

RQ1 To what extent can the economics trainee teachers' dwell time on whole SC items and individual response-relevant or response-irrelevant parts of the items of a CK test in economics predict the correct or incorrect response of these SC items?

In addition, the confidence with which trainee teachers respond to the item is considered an "essential skill for efficient study and work practice" (Gardner-Medwin 1995, p. 81). Eye-tracking studies conducted by Brückner et al. (2020) and Klein et al. (2020) revealed that confidence also affects gaze behavior. This leads to the second RQ:

[1] The use of SC items in online assessments in domains with large student populations, like medicine or economics, has become more widespread (Calhoun and Mateer 2012).

RQ2 To what extent is item-related confidence related to dwell times on single elements of a correctly responded item?

Based on prior eye-tracking research (Brückner et al. 2020; Lindner et al. 2014; Klein et al. 2020), we propose working hypotheses and explain the research design, including the CK test used and the sample of students. After presenting the results from multilevel models that take the interactions between students and items into account, we discuss the limitations of the study and implications for future research in education of teachers in economics for vocational education.

Theoretical background

Eye-tracking research and gaze bias effect in standardized educational assessments

Eye tracking has been used in cognitive and educational research for many years (Holmqvist et al. 2011; Mayer et al. 2023). It is increasingly applied in the analysis of well-structured learning environments and standardized educational assessments in various disciplines (Han et al. 2017; Klein et al. 2019; Lindner et al. 2014; Saß et al. 2017; Tsai et al. 2012). Here, the focus has often been placed on the validation of the construct like graph comprehension and knowledge by elaborately investigating gaze behavior during task processing (Zumbo and Hubley 2017).

Eye-tracking research with respect to the knowledge of students assumes that there are associations between visual perception, interpretation, and understanding between learners that possess more or less knowledge within that domain. The cognitive theory of visual expertise (Gegenfurtner et al. 2023) is an example of a possible underlying theory. Classical approaches like the "immediacy assumption" establish the link between cognitive activity, the order of its processing, and the sequence of visual perception, i.e., cognitions that occur during an action, e.g., solving an economics task (Just and Carpenter 1980; Holmqvist and Andersson 2017). The "eye-mind assumption" associates the moment of visual perception with the moment of attention and information processing (Holmqvist et al. 2011), however does not adequately reflect the complex relationships between knowledge and visual processing.

For instance, novice learners who have a lower level of domain knowledge will foveally perceive, understand, and mentally process typical challenges of the domain differently than learners with a higher level of knowledge (Larkin et al. 1980). In addition, learners with a higher level of knowledge are considered to be more efficient at selecting relevant and ignoring irrelevant information than novice learners (Haider and Frensch 1999). They perceive information from the environment by foveal and parafoveal vision and keep them in a visual register for a short time, they bundle several pieces of visually perceived information into so-called image chunks, which, in addition to the classical assumptions, enable holistic mental representations that are kept as retrieval cues in working memory (Gegenfurtner et al. 2023) and thus allow faster information processing than perceiving information individually and sequentially. In this way, advanced learners are better able to connect their mental capacity and knowledge from long-term memory with their representations of e.g., economics concepts presented in the tasks, like economics principles and rules, and to attach meaning to them and to process them in a resource-efficient way connected with the suitable domain knowledge (Gegenfurtner et al. 2023).

Through this association of perception, cognitive processing, and memory, it can be assumed that both the relative frequency and duration of perception in certain relevant and irrelevant areas of interest (AoI) can serve as an indicator for naive or advanced learners, also among economics trainee teachers.[2] With respect to SC item responding, various cognitions are shown to play a role in determining the selection of a particular response option from multiple response options, e.g., in the initial reception of information and its interpretation to the prediction of a preference for a specific response option and its final selection and evaluation (Parkes and Zimmaro 2016). The correct options (attractor) and the incorrect ones (distractors) represent the central response-relevant and response-irrelevant features of the item, respectively, and indicate the intensity with which the students deal with certain item content.

A phenomenon often observed in the investigation of gaze behavior during SC tests is the so-termed 'gaze bias effect' or 'gaze cascade effect,' which plays a major role in visual decision-making (e.g., in marketing research or face recognition, Shimojo et al. 2003; Glaholt and Reingold 2009; Saito et al. 2017). Lindner et al. (2014, p. 738) describe the gaze bias effect as a positive correlation between the preference for an object and the duration with which this object is viewed. For example, when buying a car, the car that is purchased is more likely to be viewed and analyzed by the buyer for a longer period of time than cars that are ultimately not purchased. They transferred this effect to SC tests for the first time and showed that a gaze bias effect can also be detected in decision-making between given response options. When people are asked to choose one of several response options, they usually spend more time, i.e., have fixations of longer durations, looking at the response option they will ultimately choose than at the other options, e.g., students responding correctly to the item should focus on the attractor longer (Gegenfurtner et al. 2023; Lindner et al. 2014).

In further studies from physics education research, the eye-tracking studies using a kinematic graph comprehension test (Klein et al. 2020), incorrect responses were also associated with longer dwell times on attractive distractors and lower dwell times on attractors and vice versa for correct responses. In Tsai et al. (2012), students processing a meteorological task spent more time on the response options they chose. Moreover, incorrect respondents had more difficulties understanding the question and extracting relevant information.

These partly different findings might be due to different tests and analysis foci, e.g., analyzing eye movements in terms of dwell time to describe task-response behavior. Based on prior research on the gaze bias effect, we intend to replicate the findings from Lindner et al. (2014) and Klein et al. (2020) in a first step, showing that economics trainee teachers correctly responding to the item can be expected to have a longer dwell time on the attractor than those with incorrect responses.

Since an economics knowledge test has not yet been subjected to an eye-tracking analysis, the abovementioned findings are used as the theoretical foundation for the research hypotheses. Based on the gaze bias effect for SC tests (Lindner et al. 2014), we suggest:

[2] Eye tracking complements the introspective method of thinking aloud since unconscious cognitions can be tapped into by analyzing students' gaze. This method is non-reactive and creates no additional mental load that may influence test behavior compared to the think-aloud method in which verbalized thoughts can confound cognitions (Neuert and Lenzner 2019).

H1 The longer the average duration of fixation on the attractor, the higher the probability of a correct response.

H2 The shorter the average duration of fixation on the distractors, the higher the probability of a correct response.

Effect of economics trainee teachers confidence on gaze behavior and test scores

The relationship of knowledge to confidence, as an indicator of accuracy of teacher reflective abilities, is of great importance to teaching competence (Dassa and Nichols 2019; Podgoršek and Lipovec 2017). Confidence in one's own expertise in a knowledge test is critical in achieving learning success and applying acquired knowledge in learning environments (Gardner-Medwin 1995). As part of teacher competence, confidence influences the learner's actions and provides an insight regarding the likelihood with which a learner's task response might be correct (Stankov and Lee 2008).

The relationship between knowledge and confidence is intensively studied, e.g., in the heuristics-bias approach to explain why individuals overestimate or underestimate their performance and the ways in which this disparity manifests itself in practice (Stankov and Lee 2008). Confidence ratings have already been used in several educational assessments in various disciplines, e.g., to obtain an indication of whether guessing or competent learning behavior is used in responding to an item via the discrepancy between confidence and test score (Brückner and Zlatkin-Troitschanskaia 2018). Studies generally assume that higher knowledge is also associated with higher confidence (Gardner-Medwin 1995). In studies of graph comprehension with bachelor students and trainee teachers in economics and physics, trainee teachers in economics were found to estimate their knowledge of graphs in the economics domain less accurately than physics students in their own domain. While there was a positive correlation, there was still a domain difference that necessitates more specific investigation of the correlations in the domain of economics teaching in vocational education (Brückner et al. 2020; Klein et al. 2019). Therefore, we expect:

H3 The share of correct responses should be higher for responses with high confidence than for responses with medium or low confidence.

Connections with confidence were also explored using eye tracking (Brückner et al. 2020; Klein et al. 2020). These studies have demonstrated that, in general, higher test scores on a graph test in economics were correlated with higher confidence, indicating that high, medium, or low confidence can be reflected in gaze patterns. Building on prior studies on the gaze bias effect (Lindner et al. 2014) and the assumed positive correlation between confidence and test scores, we assume:

H4 The higher the confidence of economics trainee teachers, the higher the probability of a correct response due to the extended (shortened) dwell time on the attractor (distractors).

Design and sample

The descriptions in this chapter take into account the twelve reporting standards for eye tracking studies as recommended in Dunn et al. (2023), which we complemented by several aspects.

Test and areas of interest (AoIs)

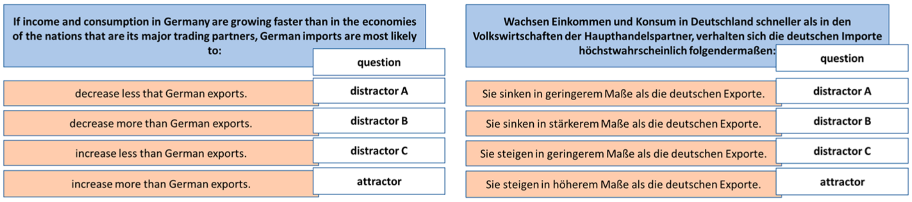
In this study, we used the economics knowledge test, which comprises 25 SC items (for details, see Zlatkin-Troitschanskaia et al. 2019b). Each item consists of one question, one attractor and three distractors (for an example, see Fig. 1). The test covers basic economic content that is generally required in economics teacher education worldwide (Holtsch et al. 2019). Each correct response is coded as 1 and incorrect responses are coded as 0. A maximum of 25 points can be achieved by each participant.

In this eye-tracking study, the five components of the items (four distractors and one attractor) were defined as AoIs for the analyses. Gaze data were collected specifically for these areas, which were spatially defined with a high degree of separation and without overlaps (Fig. 1). They reach beyond the text area, as deviations in the measured gazes were taken into account due to the precision values. Moreover, marginal areas were defined for each AoI that were at least 1° of the visual angle. A distance of 2° was defined between AoIs to avoid any confounding in the data (Holmqvist et al. 2011). An additional 'global' indicator was also created that showed students' overall processing of a task at the millisecond level.

After processing the test and selecting a response option for each item, economics trainee teachers were given a six-point Likert scale to assess their confidence in their response (1 = not confident, ..., 6 = very confident). This scale was aggregated into three categories of 0 = low confidence, 1 = medium confidence, 2 = high confidence (Gardner-Medwin 1995; Klein et al. 2020) to increase the robustness of statistical analyses in the cross-random effects model.

Apparatus

The items were implemented using the software Unipark[3]. The assessment was then implemented in the web-stimulus element of the eye-tracking software Tobii Pro Lab[3] with version 1.152.30002 (x64) and presented to the test participant on a desktop computer with a 22-inch monitor with a resolution of 1920×1080 pixels and a refresh rate of 60 Hz. The total system latency was 11 ms. Below the monitor, an infrared-based stationary eye-tracker Tobii Pro X3-120[4] using a pupil-corneal reflection method with a sampling frequency of 120 Hz was mounted, which allowed the trainee teachers to move their heads freely without a chinrest or similar objects and to assess both eye positions accurately. The laboratory in which the study took place was darkened and indirectly lit to prevent interference from other infrared sources. Precision was 0.24° of visual angle.



If income and consumption in Germany are growing faster than in the economies of the nations that are its major trading partners, German imports are most likely to:

decrease less that German exports.

decrease more than German exports.

increase less than German exports.

increase more than German exports.

Wachsen Einkommen und Konsum in Deutschland schneller als in den Volkswirtschaften der Haupthandelspartner, verhalten sich die deutschen Importe höchstwahrscheinlich folgendermaßen:

Sie sinken in geringerem Maße als die deutschen Exporte.

Sie sinken in stärkerem Maße als die deutschen Exporte.

Sie steigen in geringerem Maße als die deutschen Exporte.

Sie steigen in höherem Maße als die deutschen Exporte.

Fig. 1 Sample item from the WiWiKom test with the five labeled AoIs (colored rectangles) (translated version on the left, taken from Walstad et al. 2007)

[3] https://ww3.unipark.de/www/front.php.

[4] Tobii Pro Lab User Manual at https://www.tobiipro.com/siteassets/tobii-pro/user-manuals/Tobii-Pro-Lab-User-Manual/?v=1.152.1.

To calculate dwell times, the fixation duration metric was specified. Based on the manufacturer's specifications, fixations were measured on a millisecond basis using the identification by velocity threshold (I-VT) filter with a threshold of 30°/s of visual angle and a minimum fixation duration of 60 ms. Each fixation was always preceded and followed by a saccade, i.e., a metric to measure the eye movement between two fixations. Saccade classification was defined as when the acceleration of the eyes exceeded 8500° s−2 and velocity exceeded 30°s−1.

Procedure

The participants' distance to the eye-tracker was between 60 and 70 cm. A timed screen-based eye-tracking calibration was conducted with 9 black bullet points on a white background, 4 of which served as validation indicators to ensure the accurate and precise measurement of the trainee teachers' gazes. Measurement deviations of less than 0.8° of the visual angle were tolerated. The items were presented to the test participants in randomized order. By pressing a button, a response option (A, B, C, or D) could be selected. During the processing, the test administrator monitored the assessment on a second screen and gave correction instructions whenever the precision of the measurement was no longer sufficient or the participants were outside the measurement range.

The participants were allowed to work on the items freely, there was no time limit. The average processing time for the 25 items was 15.13 (Minimum=7.58 min; Maximum=25.42 min) minutes in total, which results in an average processing time of 36.32 s (Median=32.74 s) per item.

To ensure high test motivation, each participant received course-specific credit points for participating in this assessment, which trainee teachers need to accumulate a certain amount of to complete their course of studies.

Sample

For this study, we used the data of 20 economics trainee teachers from the degree course of economics education for vocational training from one German university. There were originally 22 participants in the study (15 female, 5 male, 2 missing), but the data from two of them were lost because the recording of gaze data was incomplete due to technical problems. The sampling approach used here was similar to that of the nation-wide representative main study (for details, see Zlatkin-Troitschanskaia et al. 2019b), i.e., the sample group was selected intentionally based on different descriptive criteria.

The average grade of the university entrance qualification in the sample was 2.25 (*SD*=0.500). The average age was 23.1 years (*SD*=4.518). The participants had, on average, completed 3.5 semesters (*SD*=3.488). Half of the participants had completed a commercial vocational training before starting their university studies.

Analysis

Since each participant completed 25 items, a total of *N*=500 response processes were available, which were clustered within students and items. Due to the nested data structure, variance splitting was necessary to account for the relationships between dwell times and test score, and confidence and dwell times. Therefore, multilevel models with crossed random effects (Rabe-Hesketh and Skrondal 2012; Snijders and Bosker 2012) were used, which are recommended for analyses of response process data (Strobel et al. 2018). Especially in unbalanced designs, random effects models have proven to be efficient and allow for the inclusion of dwell times and confidence as predictors of test scores (Rabe-Hesketh and Skrondal 2012).

Since the dependent variable item response is binary (correct vs. incorrect), a logit link function was used as a generalized linear mixed effects model. Moreover, it is assumed that confidence as a moderating variable can affect the dwell time on certain AoIs during the response process. Therefore, interaction terms, which accounts for the interaction between confidence and dwell time, were integrated into the multilevel model in addition to the main effects (moderator model).

To improve the interpretability of the results, some modifications were made: The dwell time was presented in seconds instead of the measured milliseconds. To compare the dwell time on the attractor with those on all three distractors combined, an additional variable was calculated showing the average dwell time on the three distractors.

Since a comparison between dwell times on attractor and distractor in terms of initiated cognitive processes is only possible if the stimuli are comparable, the average number of words of the two types of response options was also compared (Lindner et al. 2014). The average word count per attractor (M(SD) = 7.76(5.53) did not differ significantly from the average word count per distractor (7.31(3.63) (t = 0.697, d_{pooled}=0.013, p = .492). Each response option was phrased in approx. seven words (Fig. 1).

There was neither an item- nor sample-specific accumulation in the occurrence of the confidence and correct or incorrect responses. Therefore, to analyze the differences in the dwell times, responses in which the high confidence and low scores occurred were extracted (N = 41) and compared with responses that were also incorrect but had a medium or low confidence estimate or a correct response with high confidence. There were no values for 13 responses.

Results

Across all items, there were 302 correct responses and 198 incorrect responses. On average, 60% of the responses were correct. To obtain indications regarding the generalizability of the findings, the distribution of item difficulties calculated in the main study based on the performance of approximately 5,000 students (Zlatkin-Troitschanskaia et al. 2019b) were compared to the distribution of item difficulties in this eye-tracking study using a two-sample Kolmogorov-Smirnov test. The results (Z = 0.990, p = .281) indicated that neither distribution significantly differs from the other.

For this analysis, we specifically focus on task processing at the individual level rather than on item comparisons. To investigate to what extent the dwell time differs depending on the selected response option, i.e., responding to the item either correctly or incorrectly, an exploratory repeated-measures analysis of variance was conducted (response option = within-factor; score = between-factor). The mean dwell time on the attractor (MW = 3.87) was significantly higher than the mean dwell time of the distractors, with a small effect size (MW = 3.41) (F(1, 519) = 9.064, p < .01; η^2_p=.017). This finding, however, is score-independent, since in each response process the students obviously paid more attention to the attractors than to the distractors, and it does not take into account compensatory effects. Although the overall dwell time on the response options is longer for incorrect responses (MW = 3.417) than for correct responses (MW = 3.87), it is not evident how the overall dwell time is distributed between the individual response options, and whether this distribution differs for correct and incorrect responses (Fig. 2). The results of the non-parametric Friedman test for dependent samples confirm this finding (χ^2(df)==8.23(1), $p<.01$). The average dwell time spent on the AoI *question* was about 11.39 s for both correct and incorrect responses. A response-specific examination of dwell times revealed that the mean dwell time for correct responses (MW=12.74, SD=12.063) differed significantly (MW=10.46, SD=10.35) from that for incorrect responses (t=2.310, $p<.05$, d_{pooled}=0.206) for the AoI question. The results of the non-parametric U-test for dependent samples confirm this finding (Z=-2.626, $p<.05$).

Considering the nested data structure, we first computed a variance component model, i.e., baseline model without covariates (Model 1 in Table 1). To assess the significance of the dwell times on each individual AoI for correct responses, we controlled for the effects of the dwell times on the other AoIs. In the multilevel model with crossed random effects, the log odds had different values (Model 2 in Table 1). When controlling for the dwell times on the attractor and distractors, no significant correlation was found between the dwell time on the AoI *question* and the test score (estimate=-0.009, z=-0.87, p=.382). However, the time spent on the distractor showed a highly significant negative correlation with the response (estimate=-0.356, z=-6.09, $p<.001$). Thus, if the dwell time on any distractor increases by one unit, the probability of a correct response decreases by 30%. Conversely, a longer dwell time on the attractor increases the probability of a correct response by 16% (estimate=0.151, z=3.93, $p<.001$). When comparing the two predictors, dwell time on the distractors proves to be more indicative of a correct response (Table 1). Therefore, H1 and H2 can be confirmed based on model 2, which shows that, as a predictor, the AoI question is no longer significant; this was also suggested by the ANOVA.

Relevant to H3, Table 2 illustrates that the proportion of correct responses is larger the higher the confidence rating is which corresponds to the general assumption. In line with previous research (Brückner and Zlatkin-Troitschanskaia 2018; Klein et al. 2020), the likelihood of a correct response is linked to the participant's confidence in their response ω (Cohen 1988) (χ^2(df)=43.5874(2), $p<.001$, ω=0.299) (Table 2). H3 can thus be confirmed.

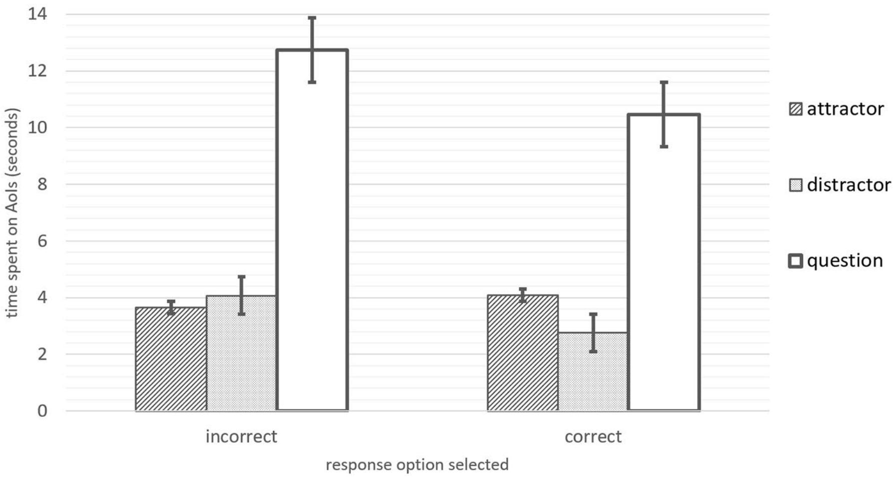


Fig. 2 Average dwell time (in seconds) of incorrect and correct respondents on the AoI attractor (left), mean of the three distractor AoIs (middle), and the AoI question (right). The error bars represent 1 standard error of the mean (SEM)

Table 1 Random intercept model with a binary logistic regression function and fixed effects on score

	Fixed effects on score							
Variable	Model 1 (VC)		Model 2		Model 3		Model 4	
	Coef.	*SE*	*Coef.*	*SE*	*Coef.*	*SE*	*Coef.*	*SE*
Intercept	0.471*	0.183	1.147***	0.262	0.426	0.330	0.831	0.435
Question (s)	-	-	−0.009	0.010	−0.006	0.010	−0.004	0.011
Distractor (s)	-	-	−0.356***	0.059	−0.296***	0.060	−0.414**	0.120
Attractor (s)	-	-	0.151***	0.038	0.157***	0.039	0.149*	0.072
Confidence[1]								
1	-	-	-	-	0.086	0.272	−0.389	0.489
2	-	-	-	-	1.221***	0.290	0.231	0.497
Conf. × Att. [1]								
1	-	-	-	-	-	-	−0.053	0.085
2	-	-	-	-	-	-	0.453**	0.166
Conf. × Dist. [1]								
1	-	-	-	-	-	-	0.202	0.137
2	-	-	-	-	-	-	−0.109	0.180
	Random effects and fit indices							
LL	-326.974		-300.675		-280.823		-272.753	
Particip. (var)	0.306		0.379		0.342		0.335	
Item (var)	0.212		0.021		0.040		0.076	
AIC/BIC	659.948/672.592		613.350/638.626		577.645/611.134		569.508/619.742	

Note: VC=Variance Component Model, SE=standard error, var=variance, LL=log likelihood, AIC=Akaike information criterion; BIC=Bayesian information criterion, s=seconds, Confidence with; *p<.05, **p<.01, ***p<.001,[1]lowest confidence rating as reference group

Table 2 Cross table score × confidence

Score	Confidence n (%)			
	low	***Medium***	***high***	***Total***
incorrect	66 (53)	87 (50)	41 (22)	194 (40)
Correct	58 (47)	86 (50)	149 (78)	293 (60)
Total	124 (100)	173 (100)	190 (100)	487 (100)

χ^2 (2)=43.587***, ω=0.299

To test H4, first, a random intercept model that only includes confidence as a fixed effect was calculated. Taking into account the nested and unbalanced data structure and compared with responses made with low confidence, the likelihood that an item was responded to correctly was four times higher when students' confidence ratings were high (odds ratio=4.312, z=5.26, p<.001). There was no significant effect when students were medium confident (odds ratio=1.097, z=0.36, p=.772). In Model 3, confidence was included as a covariate in addition to the dwell times on the AoIs (Model 3: odds ratio=3.391, z=4.21, p<.001), indicating that the assessment of confidence is a significant predictor for the likelihood of responding to the items correctly. Taking into account the dwell times on AoIs, it can be seen — in addition to Table 2 — that in the group comparison, the group with high confidence in particular shows a large correlation with a correct response (H3).

To implement the moderator model (Model 4 in Table 1), the significant main effects of the log odds of the average distractor and the attractor were each extended by an interaction effect with confidence. No significant interaction effect was found for the distractor, but a significant interaction effect beyond the significant mean effect was evident for the attractor when students were highly confident (odds ratio=1.573, z=2.73,

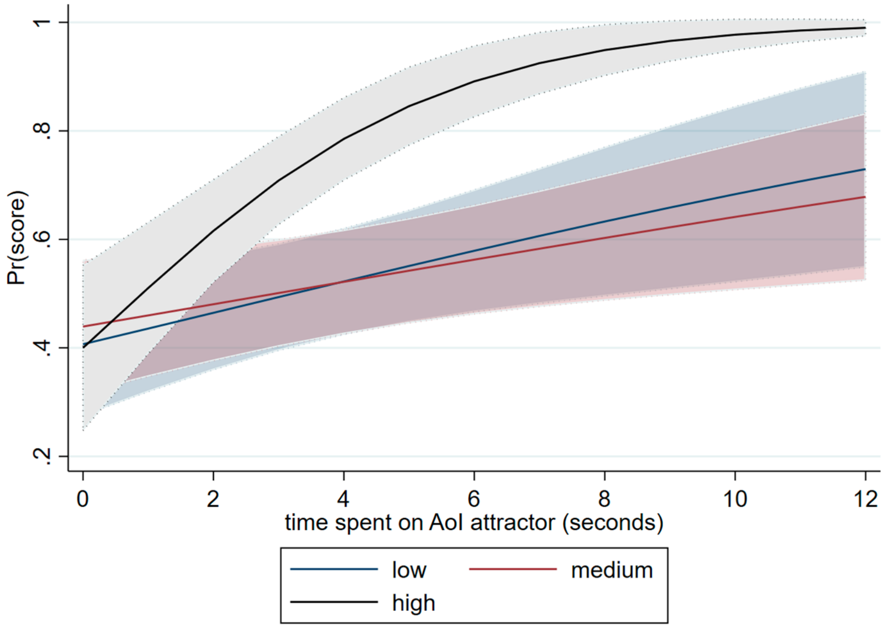


Fig. 3 Interaction between dwell time on attractor and students with low, medium, and high confidence and its predictive power regarding the average test score for all participants (with 95% confidence interval, dashed lines) (seconds <= 12)

Table 3 t-tests with the total dwell time

Group	Total dwell time				
	HCFS M (SD)	*non-HCFS M (SD)*	*T*	Cohen's *d*	U-test *z*
HCFS vs. LCFS	30.229 (18.263)	40.172 (22.124)	2.411*	0.479	2.358*
HCFS vs. MCFS	30.229 (18.263)	26.860 (15.722)	-1.172	−0.207	-1.159
HCFS vs. HCCS	30.229 (18.263)	45.100 (22.527)	3.692***	0.699	3.799***

Note. LCFS=low confidence and incorrect response; MCFS=medium confidence and incorrect response; HCCS=high confidence and correct response

*$p<0.05$. **$p<0.01$. ***$p<0.001$

$p<.01$). Thus, for participants with a high confidence, the probability of a correct response increases significantly with a dwell time on the attractor longer than 2.5 s (see Fig. 3). Hypothesis 4 can therefore only be partially confirmed, since a significant correlation between high confidence and longer dwell time on the attractor was found but not between a change in dwell time on the distractors and correct response.

These findings were also confirmed when dwell times on attractors and distractors for cases with correct solutions at high confidence are compared to those with correct solutions at low confidence. Part of the group with high confidence achieved incorrect responses(in 41 cases) (Table 2). To compare the dwell times, t-tests with independent samples were calculated. The different total dwell times, depending on the response process (Table 3) indicate that, for a comparison of the dwell times per AoI, the relative dwell times have to be used.

The analyses (Table 4) illustrate that students with high confidence and a correct solution were more able to identify the correct solution shown by a higher preference for the attractor. Since students with a high degree of confidence and a correct solution are more likely to have the domain knowledge required to answer the task, the longer focus

Table 4 t-tests with the dwell time on attractor

Group	Attractor				
	HCFS M (SD)	***non-HCFS M (SD)***	***T***	**Cohen's *d***	**U-test *z***
HCFS vs. LCFS	0.079 (0.065)	0.107 (0.084)	1.838	0.365	1.589
HCFS vs. MCFS	0.079 (0.065)	0.101 (0.097)	1.321	0.250	0.659
HCFS vs. HCCS	0.079 (0.065)	0.136 (0.096)	3.615***	0.638	3.447***

Note. LCFS=low confidence and incorrect response; MCFS=medium confidence and incorrect response; HCCS=high confidence and correct response

*$p<0.05$. **$p<0.01$. ***$p<0.001$

Table 5 *t*-tests with the dwell time on distractor

Group	Distractor				
	HCFS M (SD)	***non-HCFS M (SD)***	***T***	**Cohen's *d***	**U-test *z***
HCFS vs. LCFS	0.082 (0.053)	0.108 (0.042)	2.855**	0.568	2.781**
HCFS vs. MCFS	0.082 (0.053)	0.105 (0.050)	2.347*	0.445	2.336*
HCFS vs. HCCS	0.082 (0.053)	0.078 (0.045)	-0.492	−0.087	-0.297

Note. LCFS=low confidence and incorrect response; MCFS=medium confidence and incorrect response; HCCS=high confidence and correct response

*$p<0.05$. **$p<0.01$. ***$p<0.001$

on the attractor also reflects their preference for this answer option. They were able to identify the correct answer from a variety of incorrect answer options.

Conversely, the relative dwell times on the distractors show that, in cases with high confidence and a correct solution, the dwell time on the distractors is similar to that with high confidence and an incorrect solution (Table 5). However, students with greater uncertainty and incorrect solutions spend longer on the preferred incorrect solutions. This suggests that the relative dwell times may also reflect different task solving strategies, as solutions with greater uncertainty are worse at distinguishing between the correct and incorrect answer options, while cases with high confidence and correct solutions use the attractor purposefully.

Discussion

Given the tension between self-reflective skills and knowledge of prospective economics teachers for vocational education, this study examined the extent to which a change in the length of time spent on different AoIs influenced the test score depending on confidence.

Regarding RQ1: The assumption that correct responses are associated with shorter total dwell times was confirmed at the response process level (person × item), with a small effect size. The findings related to the comparative analysis of dwell times on distractors and attractors between trainee teachers who responded to the items correctly or incorrectly are consistent with prior research (Klein et al. 2020; Lindner et al. 2014). This indicates that participants who responded to the items correctly tend to dwell longer on attractors (H1) than on distractors and vice versa (H2). The reversed effect found for the AoI 'question' may be due to the fact that the eye-tracking metric 'dwell time‘ can refer to fixations on the question or the response options and can therefore serve as an indicator for different cognitive functions/processes (gaze bias effect) (Lindner et al. 2014). With regard to the response options, economics trainee teachers tend to spend more time focusing on certain areas of an item or a particular response option if they are inclined to choose that response option (Thomas et al. 2019).

Using multilevel models with crossed random effects for each AoI, as expected, there was a positive correlation between a correct response and a longer dwell time on the attractors (H1) and a negative correlation between an incorrect response and the dwell time on the distractors. A shorter dwell time on the AoI question indicates a tendency to respond to the item correctly, however, this correlation was not significant. Since the economics knowledge test items focus primarily on the activation of (mental) schemes and less on the activation of complex mental activities like in problem-solving processes, one explanation might be that the performance of participants who possess the required knowledge can faster infer the meaning of the questions. This is in contrast to the findings of Klein et al. (2020) but replicates the findings of Lindner et al. (2014), who also found that students who responded to the item correctly tend to have shorter total dwell times than students who did not. This indicates, for economics knowledge test items, a longer dwell time on the question is associated with comprehension difficulties or more elaborate information processing by test participants (Tsai et al. 2012). Once all AoIs had been integrated into one model, the dwell time on the question did not appear to have any significant negative correlation with the response. However, the dwell times on the individual response options were highly significant, confirming previous findings on SC tests from other domains (Lindner et al. 2014).

Regarding RQ2: Confirming Klein et al. (2020) and another study that assumed a positive relation between economics knowledge and self-reported confidence (e.g., Leiser and Aroch 2009), economics trainee teachers' confidence was positively correlated with the overall test score (H3). Thus, economics trainee teachers with higher economics knowledge tend to be able to self-reflect adequately. A high or low level of confidence was also reflected in the dwell times on the individual AoIs, which in turn were predictive of whether the item was responded to correctly. Longer dwell time at low confidence can be explained by actions characterized by higher doubt and hesitant deliberation (Stankov and Lee 2008). A high level of confidence was linked to faster response processes in the economics knowledge test, as the individual aspects of the item content were more quickly evaluated by the economics trainee teachers in terms of their relevance. However, the interaction model (Model 4) no longer shows a general confidence effect, indicating that the dwell times on the distractors and the attractor essentially determine the probability of responding to the item correctly. The moderator effect becomes evident in the interaction between dwell time on the attractor and high confidence. When confident responses were accompanied by a longer dwell time on the attractor, the probability of a correct response increased (H4).

Since the distractors address typical misconceptions of economics, they may also provide more in-depth insights into low confident economics trainee teachers' misunderstandings. For example, the sample item (Fig. 1) describes that the income of the population in Germany is increasing overall, which apparently also leads to a general increase in consumption. If the economics trainee teachers chose one of the first two response options (distractors), it can be assumed that they do not understand the significance of a general increase in income and its effect on consumption.

In addition to findings from previous studies (Lindner et al. 2014; Klein et al. 2020), this study shows a correlation between confidence assessment dwell times on specific task parts and economics knowledge among trainee teachers in economics. At the same time, however, different effects can occur. Thus, it is necessary to capture the self-assessed confidence of trainee teachers in economics from a metacognitive perspective. It seems obvious that the dwell time on the different AoIs may be an expression of different task solving strategies that are used when confidence is high or low. At the same time, however, it seems necessary to diagnose the different facets of teachers' professional knowledge even more precisely to find out what the explanations might be for different levels of confidence in answering tasks. Since knowledge tasks also depict different topics and concepts, it is obvious that teacher knowledge also varies and that alternative task solving strategies are used in cases of self-assessed uncertainty because the correct answer is not directly recognized. This has been emphasized before, e.g., Leiser and Aroch (2009, p. 381) conclude from their study: "On the one hand, they declare on average not to understand the concepts very well. On the other, they are quite willing to judge how changes in one economic variable would affect another. Our interpretation is that what enables the economically untrained to answer is their superficial approach to the issues." Thus, comprehensive assessment of economics teacher knowledge also requires the measurement of teachers' self-reflection to find out about their strategies for answering the content knowledge tasks. However, to diagnose how this is reflected in performing specific, professional tasks in their teaching job, such as responding to CK items, more extensive analyses using authentic tasks and log data analysis beyond eye tracking are required).

Limitations and future research

While the presented results are mostly in line with previous studies, further areas of research emerge for a more in-depth analysis of the significance of self-reflection as part of the response process. CK represents only one facet of teaching competence. The extent to which the phenomenon of correct and incorrect solutions with different levels of confidence and its effect on dwell times might also be evident in other knowledge dimensions, e.g., PCK, has not yet been explored. Likewise, the relationship of this phenomenon and the associated eye movements to teachers' actual classroom performance is only vaguely suggestible. In particular, the effects of high and low confidence with high or low economic knowledge on classroom behavior, e.g., instruction or economics teachers' detailed attention to student errors, remain to be investigated. Further studies using the corresponding assessments are still required.

In the present study, the eye movements of economics trainee teachers were investigated for the first time in the context of self-assessed confidence and economics knowledge for vocational education. This is a domain-specific finding. Moreover, the question arises whether these findings can be generally assumed for other teaching domains. First interdisciplinary studies on the domain comparison of graph comprehension suggest that there might be domain-specific differences (Klein et al. 2019; Brückner et al. 2020). However, empirical evidence has yet to be provided.

In this study, only SC tests in a traditional task format with one correct and several incorrect options were applied, as they were also commonly used in other studies (Klein et al. 2020; Tsai et al. 2012; Han et al. 2017). However, comparisons are not always possible, as these studies refer to other disciplines and do not exclusively focus on teacher education research. Including other constructs e.g., PCK entails also including tasks with other format representations, e.g., graphical rather than textual, which might be the addition of representation on a whiteboard in a classroom to the tasks. Since the response process can be affected by the type of representation (textual vs. graphical) as well as the specified cognitive demands (simply recalling content from memory vs. problem analysis) of an SC item or by specific content and teacher knowledge demands, different expectations should be formulated for different types of SC items (Saß et al. 2017).

When responding to SC items, participants have to choose one of several response options. Here, too, comprehension plays a role, but the focus mainly lies on the 'attractiveness' of the response options, one of which must be selected by the economics trainee teachers. As studies from other disciplines indicate (Lindner et al. 2014; Klein et al. 2020), the time the trainee teachers spend looking at the response options, i.e., dwell time, tends to be indicative of which response option they prefer and will eventually choose. In further studies, the individual distractors should be taken into consideration in a more differentiated manner, e.g., by analyzing them based on their 'attractiveness' and by matching eye-movement data with the item difficulty and discrimination parameters (derived from more comprehensive field studies) or other classroom specific parameters. For instance, in assessing PCK, the distractors and attractors might include different economic student or teacher statements that need to be evaluated.

Another (general) methodological limitation lies in the definition of AoIs (Bojko 2013; Holmqvist et al. 2011), which include textual content. The size of the AoIs significantly determines the dwell times and fixation frequencies to be assessed and was standardized across all items for this study.

Moreover, the question arises whether similar findings would have been obtained with mobile eye-trackers and paper-based SC tests. For instance, due to the particular setup of the experimental situations with a participant-to-administrator ratio of 1:1 (which differs from field surveys), the survey situations were highly controlled in terms of time, place, and person, and the participants always made an effort to work intensively on the items, which is less common for low-stakes surveys (with large samples). In future studies, a variation of audience-response systems or clickers should be implemented to find out how the feedback affects the (visual) perception of items. In addition, mobile eye trackers are often used to analyze classroom events (Goldberg et al. 2021).

How difference in content knowledge and teachers self-reflection could therefore also be investigated in the context of specific actions in the classroom and, together with teacher educators, an objectified evaluation of the actions could be compared with the trainee teachers' self-reflections of these situations. For example, the controversially discussed Dunning-Kruger effect (Kruger and Dunning 1999), could also be a significant factor that needs to be investigated in more detail to obtain indications of different task strategies. The effect describes the phenomenon of deficits in one's (here: content) knowledge with a concurrent high self-assessment of this knowledge, and is therefore to be seen critically, especially with regard to the necessary self-reflection in teaching (Dassa and Nichols 2019). To date, it is largely unclear how this manifests itself in the visual perception and selection of SC response options and it is discussed whether it is just a statistical artefact or not (Gignac et al. 2020).

In cross-linked mixed-effects models, the effects of dwell times on scores have been investigated by taking into account the cross-classification of dwell times in relation to both items and participants simultaneously (Strobel et al. 2018). Further predictors can be used at different levels, and future studies should also analyze gaze behavior in relation to item difficulty. For this purpose, adjustments and estimates of random effects are necessary, which require a larger sample — a greater number of items and participants.

When expanding the sample, different levels of expertise should be systematically taken into account, e.g., advanced students and first-year students, to analyze developments over the course of the study (Brückner et al. 2020). Furthermore, the present study did not aim to analyze how dwell times changed during the response process; thus, no analysis of the chronological sequence of dwell times on AoIs in specific time intervals has been conducted so far. In particular, multilevel models with autoregressive covariance structures and crossed-random effects might provide some valuable insights into time-dependent analyses. However, Lindner et al. (2014) showed in a gaze-likelihood analysis (across the response process of SC items) that the fixation times of participants with higher and lower performance levels on different task intervals was overall comparable in terms of their attention distribution over time, which was not the matter of this study.

Conclusion

Competent economics teachers should not only have sufficient professional knowledge (CK, PCK, PK), but also assess and apply it appropriately in a self-reflective manner. Self-reflection such as self-confidence is shown to be particularly important in relation to teachers' professional knowledge and its application. In previous studies, economics teachers' self-reflective competencies were theoretically modelled and empirically assessed (Brückner and Zlatkin-Troitschanskaia 2018). These studies empirically identified significant correlations between these self-reflective competencies and teaching skills. They suggest that such self-reflective competencies, in addition to professional knowledge, are a necessary foundation for professional action in the classroom (Feucht et al. 2017; Schön 1987).

Our study is based on prior research, in which a diagnosis and analysis of the differences between students' confidence and their knowledge has already been used to explain differences in economics knowledge test performance based on isolated eye movements that provide insight into participants' analytic information processing. To date, little research has been conducted to analyze the relationship between confidence, knowledge and eye movements as it pertains to (prospective) teachers, and no study was available for the domain of economics for vocational education. Therefore, based on research from the other domains (physics, biology), this eye-tracking study contributes towards bridging this research gap. The findings indicate that trainee teachers who exhibit differences between confidence and knowledge also differ in their gaze behavior from students who correctly assess their CK in economics. The results of this study thus not only indicate deficits in self-reflective skills in line with previous studies on teachers' self-reflective competencies, but also point to the significant role these skills play in the acquisition and application of correct CK.

Further research is needed to investigate this phenomenon in other teacher professional knowledge areas such as PCK and PK. To this end, we are currently conducting an analogue eye-tracking study using a validated PCK test among economics students for vocational education (Kuhn et al. 2016). Here, it is of particular interest whether differences between confidence, eye movements, and knowledge that became evident in this study using a CK test can also be found in economics trainee teachers while responding to a PCK test.

In terms of practical implications, it can be concluded that such self-reflective skills need to be more explicitly addressed in economics teacher education. This is especially true in the context of increased digital learning and the use of freely available online information in economics teacher education, to prevent the acquisition of erroneous knowledge and misconceptions.

Abbreviations

CK	Content Knowledge
PCK	Pedagogical Content Knowledge
SC	Single-Choice
AoI	Areas of Interest
I-VT	Identification by Velocity Threshold

Acknowledgements

We would like to thank all economics trainee teachers who participated in this study. Furthermore, we would also like to thank the reviewers for their differentiated feedback and their contribution to the improvement of this manuscript. Detailed comments, efforts, and expertise of the reviewers were outstanding and certainly contributed significantly to the quality of this manuscript.

Author contributions

Both authors contributed substantially to this work. Conceptualization, SB and OZT; methodology, SB; validation, SB and OZT; investigation, SB; resources, writing—original draft preparation, SB and OZT; project administration, OZT and SB; all authors have read and agreed to the published version of the manuscript.

Funding

This study was funded by the Center of School, Education and Higher Education Research at the Johannes Gutenberg University Mainz and the PLATO project funded by the Ministry for Science, Continued Education and Cultural Affairs Rhineland-Palatinate, Germany.

Data availability

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Declarations

Competing interests

The authors declare no conflict of interest.

Received: 15 May 2023 / Accepted: 10 January 2024

Published online: 28 January 2024

References

Bojko A (2013) Eye tracking the user experience - a practical guide to research. Rosenfeld Media, Brooklyn, New York

Brückner S, Zlatkin-Troitschanskaia O (2018) Threshold concepts for modeling and assessing higher education students' understanding and learning in Economics. In: Zlatkin-Troitschanskaia O, Toepper M, Pant HA, Lautenbach C, Kuhn C (eds) Assessment of learning outcomes in higher education: cross-national comparisons and perspectives, vol 40. Springer, Cham, pp 103–121. doi:https://doi.org/10.1007/978-3-319-74338-7_6

Brückner S, Zlatkin-Troitschanskaia O, Küchemann S, Klein P, Kuhn J (2020) Changes in students' understanding of and visual attention on Digitally Represented Graphs across Two Domains in higher education: a postreplication study. Front Psychol 11,2090:1–20. https://doi.org/10.3389/fpsyg.2020.02090

Bruno JE (1993) Using testing to provide feedback to support instruction: a reexamination of the role of assessment in educational organizations. In: Leclerq DA, Bruno JE (eds) Item Banking: interactive testing and Self-Assessment, vol 112. Springer, Berlin, Heidelberg, pp 190–209. doi: https://doi.org/10.1007/978-3-642-58033-8_16

Calhoun J, Mateer D (2012) Incorporating media and response systems in the economics classroom. In: Hoyt G, McGoldrick K (eds) International handbook on teaching and learning Economics. Edward Elgar Publishing, Cheltenham, pp 149–159. doi:https://doi.org/10.4337/9781781002452.00025

Cohen J (1988) Statistical power analysis for the behavioral sciences, 2nd edn. Routledge, New York. https://doi.org/10.4324/9780203771587

Cordova JR, Sinatra GM, Jones SH, Taasoobshirazi G, Lombardi D (2014) Confidence in prior knowledge, self-efficacy, interest and prior knowledge: influences on conceptual change. Contemp Educ Psychol 39(2):164–174. https://doi.org/10.1016/j.cedpsych.2014.03.006

Dassa L, Nichols B (2019) Self-efficacy or overconfidence? Comparing preservice teacher self-perceptions of their content knowledge and teaching abilities to the perceptions of their supervisors. New Educ 15(2):156–174. https://doi.org/10.1080/1547688X.2019.1578447

Davies P (2002) There's no confidence in multiple-choice testing. Paper presented at the 6th CAA Conference, University of Loughborough, July 2002

Dunn MJ, Alexander RG, Amiebenomo OM et al (2023) Minimal reporting guideline for research involving eye tracking (2023 edition). Behavior Research. https://doi.org/10.3758/s13428-023-02187-1

Ercikan K, Pellegrino JW (eds) (2017) NCME applications of educational measurement and assessment book series. Validation of score meaning for the next generation of assessments: the use of response processes. Routledge, New York

Eva KW, Cunnington JPW, Reiter HI, Keane DR, Norman GR (2004) How can I know what I don't know? Poor self assessment in a well-defined domain. Adv Health Sci Educ 9(3):211–224. https://doi.org/10.1023/B:AHSE.0000038209.65714.d4

Feucht FC, Lunn Brownlee J, Schraw G (2017) Moving beyond reflection: Reflexivity and epistemic cognition in teaching and teacher education. Educational Psychol 52(4):234–241. https://doi.org/10.1080/00461520.2017.1350180

Gardner-Medwin AR (1995) Confidence assessment in the teaching of basic science. ALT-J 3(1):80–85. https://doi.org/10.1080/0968776950030113

Gegenfurtner A, Gruber H, Holzberger D, Keskin Ö, Lehtinen E, Seidel T, Stürmer K, Säljö R (2023) Towards a cognitive theory of visual expertise: methods of Inquiry. In: Damşa C, Rajala A, Ritella G, Brouwer J (eds) Re-theorising Learning and Research methods in Learning Research. Routledge, London, pp 146–163. doi: https://doi.org/10.4324/9781003205838-10

Gignac GE, Zajenkowski M (2020) The Dunning-Kruger effect is (mostly) a statistical artefact: valid approaches to testing the hypothesis with individual differences data. Intelligence 80:101449

Glaholt MG, Reingold EM (2009) Stimulus exposure and gaze bias: a further test of the gaze cascade model. Atten Percept Psychophysics 71(3):445–450. https://doi.org/10.3758/APP.71.3.445

Goldberg P, Schwerter J, Seidel T, Müller K, Stürmer K (2021) How does learners' behavior attract preservice teachers' attention during teaching? Teaching and teacher education, 97,103213. https://doi.org/10.1016/j.tate.2020.103213

Greving S, Lenhard W, Richter T (2020) Adaptive retrieval practice with multiple-choice questions in the university classroom. J Comput Assist Learn 36(6):799–809. https://doi.org/10.1111/jcal.12445

Haider H, Frensch PA (1999) Eye movement during skill acquisition: more evidence for the information-reduction hypothesis. J Experimental Psychology: Learn Memory Cognition 25(1):172–190

Han J, Chen L, Fu Z, Fritchman J, Bao L (2017) Eye-tracking of visual attention in web-based assessment using the force concept inventory. Eur J Phys 38(4). https://doi.org/10.1088/1361-6404/aa6c49

Holmqvist K, Andersson R (2017) Eye-tracking: a comprehensive guide to methods, paradigms and measures. Lund Eye-Tracking Research Institute

Holmqvist K, Nyström M, Andersson R, Dewhurst R, Jarodzka H, van de Weijer J (2011) Eye tracking: a comprehensive guide to methods and measures, 1st edn. Oxford University

Holtsch D, Brückner S, Förster M, Zlatkin-Troitschanskaia O (2019) Gender gap in Swiss vocational education and training teachers' economics content knowledge and the role of teaching experience. Citizsh Social Econ Educ 18(3):218–237. https://doi.org/10.1177/2047173419893595

Just MA, Carpenter PA (1980) A theory of reading: from eye fixations to comprehension. Psychol Rev 87(4):329–354

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

Khan KS, Davies DA, Gupta JK (2001) Formative self-assessment using multiple true-false questions on the internet: feedback according to confidence about correct knowledge. Med Teach 23(2):158–163. https://doi.org/10.1080/01421590031075

Kim LE, Klassen RM (2018) Teachers' cognitive processing of complex school-based scenarios: differences across experience levels. Teach Teacher Educ 73:215–226. https://doi.org/10.1016/j.tate.2018.04.006

Klein P, Küchemann S, Brückner S, Zlatkin-Troitschanskaia O, Kuhn J (2019) Student understanding of graph slope and area under a curve: a replication study comparing first-year physics and economics students. Phys Rev Phys Educ Res 15(2):1–17. https://doi.org/10.1103/PhysRevPhysEducRes.15.020116

Klein P, Lichtenberger A, Küchemann S, Becker S, Kekule M, Viiri J, Baadte C, Vaterlaus A, Kuhn J (2020) Visual attention while solving the test of understanding graphs in kinematics: an eye-tracking analysis. Eur J Phys 41(2):25701. https://doi.org/10.1088/1361-6404/ab5f51

Kolbitsch J, Ebner M, Nagler W, Scerbakov N Can confidence assessment enhance traditional multiple-choice testing. Paper presented at the ICL, Conference (2008) Carinthia Tech Institute, Villach, 24–28 September 2008

Korthagen FAJ (2004) In search of the essence of a good teacher: towards a more holistic approach in teacher education. Teach Teacher Educ 20(1):77–97. https://doi.org/10.1016/j.tate.2003.10.002

Kruger J, Dunning D (1999) Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. J Personal Soc Psychol 77(6):1121–1134. https://doi.org/10.1037/0022-3514.77.6.1121

Kuhn C, Alonzo AC, Zlatkin-Troitschanskaia O (2016) Evaluating the pedagogical content knowledge of pre- and in-service teachers of business and economics to ensure quality of classroom practice in vocational education and training. Empir Res Vocat Educ Train 8(1):1–18. https://doi.org/10.1186/s40461-016-0031-2

Larkin J, McDermott J, Simon DP, Simon HA (1980) Expert and novice performance in solving physics problems. Science 208(4450):1335–1342

Leiser D, Aroch R (2009) Lay understanding of macroeconomic causation: the good-begets-Good Heuristic. Appl Psychol 58(3):370–384. https://doi.org/10.1111/j.1464-0597.2009.00396.x

Lindner MA, Eitel A, Thoma GB, Dalehefte IM, Ihme JM, Köller O (2014) Tracking the decision-making process in multiple-choice Assessment: evidence from eye movements. Appl Cogn Psychol 28(5):738–752. https://doi.org/10.1002/acp.3060

Mayer WM, Rausch A, Seifried J (2023) Analysing domain-specific problem-solving processes within authentic computer-based learning and training environments by using eye-tracking: a scoping review. Empir Res Vocat Educ Train 15(2). https://doi.org/10.1186/s40461-023-00140-2

Neuert CE, Lenzer T (2019) Use of eye tracking in cognitive pretests. Leibniz Institute for the Social Sciences, Mannheim. https://doi.org/10.15465/gesis-sg_en_025

Nilsson P, Karlsson G (2019) Capturing student teachers' pedagogical content knowledge (PCK) using CoRes and digital technology. Int J Sci Educ 41(4):419–447. https://doi.org/10.1080/09500693.2018.1551642

Parkes J, Zimmaro D (2016) Learning and assessing with multiple-choice questions in college classrooms. Routledge, New York. https://doi.org/10.4324/9781315727769

Podgoršek M, Lipovec A (2017) Self-assessment ability of pre-service teachers. The New Educational Review 48(2):213–223. https://doi.org/10.15804/tner.2017.48.2.17

Rabe-Hesketh S, Skrondal A (2012) Continuous responses, 3rd edn. In: Rabe-Hesketh S, Skrondal A. (eds) Multilevel and longitudinal modeling using stata, vol 1. Stata Press, College Station

Rodríguez S, Regueiro B, Blas R, Valle A, Piñeiro I, Cerezo R (2014) Teacher self-efficacy and its relationship with students' affective and motivational variables in higher education. Eur J Educ Psychol 7(2):107–120. https://doi.org/10.1989/ejep.v7i2.183

Saito T, Nouchi R, Kinjo H, Kawashima R (2017) Gaze bias in preference judgments by younger and older adults. Front Aging Neurosci 9:285. https://doi.org/10.3389/fnagi.2017.00285

Sanders JI, Hangya B, Kepecs A (2016) Signatures of a statistical computation in the human sense of confidence. Neuron 90(3):499–506. https://doi.org/10.1037/t64341-000

Saß S, Schütte K, Lindner MA (2017) Test-takers' eye movements: effects of integration aids and types of graphical representations. Comput Educ 109:85–97. https://doi.org/10.1016/j.compedu.2017.02.007

Schön DA (1987) Educating the reflective practitioner. Jossey-Bass, London

Shimojo S, Simion C, Shimojo E, Scheier C (2003) Gaze bias both reflects and influences preference. Nat Neurosci 6(12):1317–1322. https://doi.org/10.1038/nn1150

Shulman LS (1986) Those who understand: knowledge growth in teaching. Educational Researcher 15(2):4–14. https://doi.org/10.3102/0013189X015002004

Snijders TAB, Bosker RJ (2012) Multilevel analysis: an introduction to basic and advanced multilevel modeling, 2nd edn. Sage, London

Stankov L (2013) Noncognitive predictors of intelligence and academic achievement: an important role of confidence. Pers Indiv Differ 55(7):727–732. https://doi.org/10.1177/0013164492052004025

Stankov L, Lee J (2008) Confidence and cognitive test performance. J Educ Psychol 100(4):961–976. https://doi.org/10.1037/a0012546

Strobel B, Lindner MA, Saß S, Köller O (2018) Task-irrelevant data impair processing of graph reading tasks: an eye tracking study. Learn Instruction 55:139–147. https://doi.org/10.1016/j.learninstruc.2017.10.003

Thomas AW, Molter F, Krajbich I, Heekeren HR, Mohr PNC (2019) Gaze bias differences capture individual choice behaviour. Nat Hum Behav 3(6):625–635. https://doi.org/10.1038/s41562-019-0584-8

Tsai MJ, Hou HT, Lai ML, Liu WY, Yang FY (2012) Visual attention for solving multiple-choice science problem: an eye-tracking analysis. Comput Educ 58(1):375–385. https://doi.org/10.1016/j.compedu.2011.07.012

Walstad WB, Watts M, Rebeck K (2007) Test of understanding in college economics: examiner's manual, 4th edn. National Council on Economic Education

Walstad WB, Rebeck K, Butters RB (2013) The test of economic literacy: development and results. J Econ Educ 44(3):298–309. https://doi.org/10.1080/00220485.2013.795462

Walstad WB, Schmidt S, Zlatkin-Troitschanskaia O, Happ R (2018) Pretest-posttest measurement of the economic knowledge of undergraduates – estimating guessing effects. Paper presented at the Annual AEA Conference on Teaching and Research in Economic Education, Philadelphia, PA. https://doi.org/10.4300/JGME-D-11-00324.1

Zlatkin-Troitschanskaia O, Kuhn C, Brückner S, Leighton JP (2019a) Evaluating a technology-based Assessment (TBA) to measure teachers' action-related and reflective skills. Int J Test 19(2):148–171. https://doi.org/10.1080/15305058.2019.1586377

Zlatkin-Troitschanskaia O, Jitomirski J, Happ R, Molerov D, Schlax J, Kühling-Thees C, Förster M, Brückner S (2019b) Validating a test for measuring knowledge and understanding of Economics among University students. Z Für Pädagogische Psychologie 33(2):119–133. https://doi.org/10.1024/1010-0652/a000239

Zumbo BD, Hubley AM (eds) (2017) Understanding and investigating response processes in validation research. Springer, Cham. https://doi.org/10.1007/978-3-319-56129-5

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Dr. Sebastian Brückner (SB) is a postdoctoral researcher at the chair of business and economic at the Johannes Gutenberg University (JGU) in Mainz, Germany. In 2015 he received his Ph.D. from the JGU. His research focuses on modeling and assessment of teaching and learning in digital and analog environments in the domain of business and economics using technological assessments (e.g., eye-tracking).

Prof. Dr. Olga Zlatkin-Troitschanskaia (OZT) has held the Chair of Business and Economics Education at Johannes Gutenberg University (JGU) Mainz since 2006. After her studies at Humboldt University (HU) Berlin she completed her doctorate (summa cum laude) in 2004 with two research awards, habilitated in 2006 and received several awards for outstanding research achievements in international competence research in higher education, including fellowships at the International Academy of Education (IAE), the National Academy of Science and Engineering (Acatech) and the Gutenberg Research College. He has been researching the competence development of students in higher education in large-scale longitudinal and in experimental studies.