Cognitive engagement is a core dimension of good teaching. In this study, we conceptualize it with the ICAP framework, which distinguishes four levels of engagement identifiable through students' activities. We apply the framework to teaching English as a foreign language (TEFL) using a simulation in which pre- and in-service teachers learn to diagnose engagement levels when planning and implementing lessons. The study aimed to validate the simulation and replicate Roeben et al. (2025) to test the framework's cross-context applicability. Data from N = 118 pre-service teachers at a Bavarian university support the validity of the simulation. Applying signal detection theory further indicates that while ICAP is a useful analytical tool, it requires adaptation for subject-specific contexts like TEFL. To address these limits, we recommend incorporating the dual structure of foreign language learning and extending the framework with typical examples.

Schlagworte: Cognitive engagement; Diagnostic skills; Simulations; TEFL
Zitiervorschlag: *Roeben, Meral; Lütge, Christiane; Schultz-Pernice, Florian; Vejvoda, Johanna; Sailer, Michael; Fischer, Frank & Heitzmann, Nicole (2025). Diagnosing Cognitive Engagement in TEFL: Validating and Analyzing a Simulation. Zeitschrift für Fremdsprachenforschung, 36(2), 27-48. Bielefeld: wbv Publikation. https://doi.org/10.3278/ZFF2502W002*

E-Journal Einzelbeitrag
von: Meral Roeben, Christiane Lütge, Florian Schultz-Pernice, Johanna Vejvoda, Michael Sailer, Frank Fischer, Nicole Heitzmann

Diagnosing Cognitive Engagement in TEFL

Validating and Analyzing a Simulation

aus: Zeitschrift für Fremdsprachenforschung 2/2025 (ZFF2502W)
Erscheinungsjahr: 2025
Seiten: 27–48
DOI: 10.3278/ZFF2502W002

Diagnosing Cognitive Engagement in TEFL

Validating and Analyzing a Simulation

Meral Roeben[1], Christiane Lütge[2], Florian Schultz-Pernice[3], Johanna Vejvoda[4], Michael Sailer[5], Frank Fischer[6] & Nicole Heitzmann[7]

Abstract

Kognitive Aktivierung zählt zu den drei Basisdimensionen guten Unterrichts. In dieser Studie konzeptualisieren wir sie mit dem ICAP-Modell, das vier Stufen unterscheidet, die durch Beobachtung von Schüler:innen-Aktivitäten bestimmt werden können. Wir wenden das Modell auf den Englischunterricht an und nutzen eine Simulation, in der Lehramtsstudierende und Lehrkräfte im Dienst lernen, die Stufen kognitiver Aktivierung beim Planen und Realisieren von Unterricht zu diagnostizieren. Ziel war die Validierung der Simulation sowie die Replikation der fachübergreifenden Studie von Roeben et al. (2025), um die kontextübergreifende Anwendbarkeit des Modells zu prüfen. Mit N = 118 Lehramtsstudierenden einer bayerischen Universität fanden wir Belege für die Validität. Zudem zeigt die Signalentdeckungstheorie, dass sich die Übertragung in den englischdidaktischen Kontext unter Einschränkungen als erfolgversprechend erweist. Wir empfehlen daher, die Doppelstruktur des Fremdsprachenlernens in das Modell aufzunehmen und es um typische Beispiele zu erweitern.

1 Meral Roeben, Research associate, LMU Munich, Department of Psychology, Chair of Education and Educational Psychology, Leopoldstraße 13, 80802 Munich, Germany, and LMU Munich, Department of English and American Studies, Chair of Teaching English as a Foreign Language, Schellingstraße 3, 80799 Munich, Germany, E-Mail: Meral.Roeben@lmu.de

2 Christiane Lütge, Professor, LMU Munich, Department of English and American Studies, Chair of Teaching English as a Foreign Language, Schellingstraße 3, 80799 Munich, Germany, and LMU Munich, DigiLLab, Leopoldstraße 13, 80802 Munich, Germany, E-Mail: Christiane.Luetge@lmu.de

3 Florian Schultz-Pernice, Academic director, LMU Munich, Department of Psychology, Chair of Education and Educational Psychology, Leopoldstraße 13, 80802 Munich, Germany, and LMU Munich, DigiLLab, Leopoldstraße 13, 80802 Munich, Germany, E-Mail: Florian.Schultz-Pernice@psy.lmu.de

4 Johanna Vejvoda, Research associate, LMU Munich, Department of Psychology, Chair of Education and Educational Psychology, Leopoldstraße 13, 80802 Munich, Germany, and LMU Munich, DigiLLab, Leopoldstraße 13, 80802 Munich, Germany, E-Mail: Johanna.Vejvoda@psy.lmu.de

5 Michael Sailer, Professor, University of Augsburg, Faculty of Philosophy and Social Sciences, Department of Learning Analytics and Educational Data Mining, Alter Postweg 101, 86135 Augsburg, Germany, E-Mail: Michael.Sailer@uni-a.de

6 Frank Fischer, Professor, LMU Munich, Department of Psychology, Chair of Education and Educational Psychology, Leopoldstraße 13, 80802 Munich, Germany, and LMU Munich, DigiLLab, Leopoldstraße 13, 80802 Munich, Germany, E-Mail: Frank.Fischer@psy.lmu.de

7 Nicole Heitzmann, Assistant Professor, LMU Munich, Department of Psychology, Chair of Education and Educational Psychology, Leopoldstraße 13, 80802 Munich, Germany, and LMU Munich, DigiLLab, Leopoldstraße 13, 80802 Munich, Germany, E-Mail: Nicole.Heitzmann@psy.lmu.de

1 Introduction

One of the three basic dimensions of good teaching is cognitive engagement (Praetorius & Gräsel, 2021). This study investigates diagnosing cognitive engagement in teaching English as a foreign language (TEFL). Diagnostic skills are part of teachers' professional knowledge (Kramer et al., 2021). In modern TEFL classrooms, diagnosing is often conducted in the context of technology-related teaching. This process is complex and presupposes respective diagnostic skills. Simulations have shown to effectively support acquiring complex skills by reducing the complexity of a diagnostic situation and focusing only on aspects relevant for diagnosing (Heitzmann et al., 2019). In the present study, we aim at validating a simulation that supports acquiring diagnostic skills regarding cognitive engagement and instructional TEFL quality in technology-related TEFL lessons. Additionally, we aim at empirically investigating the concept of cognitive engagement within the context of foreign language teaching. More specifically, we investigate challenges that occur during diagnosing cognitive engagement in the TEFL context by replicating parts of a study conducted in a cross-domain context (Roeben et al., 2025). We hope to derive theoretical and practical implications for TEFL by comparing the patterns found in both studies.

1.1 Diagnostic Skills

Teachers make a variety of complex decisions which should be informed by aspects like the students' knowledge, motivation, or emotions (Kramer et al., 2021; Urhahne & Wijnia, 2021). Diagnosing such phenomena involves identifying a problem and finding a solution to it (Heitzmann et al., 2019). Based on these diagnoses, teachers can design content and tasks that fit students' needs, thus support their learning process (Urhahne & Wijnia, 2021). The respective diagnostic skills consist of conceptual knowledge (e. g., knowing the characteristics of a framework) and action-oriented knowledge (e. g., knowing how to put a framework to practice; Kopp et al., 2008). Today, teaching increasingly involves the use of technology. Technology within this article refers to any computer-based technology that supports teaching and learning. When making diagnostic decisions, teachers need to assess the potential of technology in enhancing students' learning. Technology, however, can also be used to inform the diagnostic processes, as technology allows insights into students' current activities from which phenomena central to the learning process, such as cognitive engagement, can be inferred.

1.2 Types of Knowledge Required in Different Phases of Teaching

Teachers diagnose cross-domain and subject-specific phenomena in technology-related teaching situations. The professional knowledge required to do so, is conceptualized by the TPACK model (Koehler et al., 2013). The TPACK model describes seven domains of knowledge: Content knowledge (CK) refers to knowledge about a subject's content, while pedagogical knowledge (PK) includes effective teaching methods; integrated, these types of knowledge form pedagogical content knowledge (PCK; Koehler et al., 2013). Technological knowledge (TK) refers to knowledge about how to use and

implement technology and the integration of CK, PK, and TK results in technological pedagogical and content knowledge (TPACK; Koehler et al., 2013). It is the teacher's task to combine the different domains and base the teaching on this combined knowledge (Koehler et al., 2013).

To provide students with beneficial learning opportunities, teachers apply TPACK in the different phases of teaching (Ertmer & Ottenbreit-Leftwich, 2010). In the present paper, we focus on diagnosing in the planning phase (i. e., planning lessons) and in the implementation phase (i. e., giving lessons in the classroom). We assume that the respective diagnostic decisions differ: In the planning phase, learning goals and tasks are diagnosed regarding aspects like cognitive engagement. Diagnosing in the implementation phase includes making decisions within the dynamic system of a classroom. In the present study, during the planning phase, participants diagnose phenomena related to PCK (see 1.4; Müller-Hartmann, 2017) and PK (see 1.3; Chi & Wylie, 2014), in the implementation phase a phenomenon related to PK (see 1.3; Chi & Wylie, 2014) is diagnosed. Results regarding the respective diagnostic performance (i. e., accurately diagnosing the different phenomena) in combination with conceptual knowledge inform about the validity of the simulation designed in the course of this study (see RQ1 in chapter 2).

1.3 Cognitive Engagement

One component of PK is being able to diagnose cognitive engagement – one of the core dimensions of good teaching – which is featured in prominent instructional quality models, such as the "Syntheseframework" and the MAIN-TEACH model (Praetorius & Gräsel, 2021). Cognitive engagement alludes to students' depth of cognitive involvement with instructional material (Praetorius & Gräsel, 2021). One promising way of conceptualizing this is the ICAP framework which distinguishes four observable levels of cognitive engagement – interactive, constructive, active, and passive – and assumes a hierarchy in terms of cognitive processing depth (Chi & Wylie, 2014). In contrast, the construct of cognitive activation (Klieme & Rakoczy, 2008) refers to internal cognitive stimulation through instructional design, which is not directly observable. This distinction is particularly relevant in TEFL, where communication plays a central role in learning and presupposes deep involvement with the language (Wilden, 2021). However, within the TEFL context the conceptualization of cognitive engagement varies greatly (Guttke, 2023). We acknowledge that the ICAP framework, while empirically established and offering diagnostically accessible levels of cognitive engagement, does not fully capture these deeper and subject-specific facets of cognitive activation. However, given the difficulty of conceptualizing cognitive engagement in cross-disciplinary contexts, our study adopts ICAP as a workable framework for empirical investigation, while making its assumptions and limitations transparent. By investigating the ICAP framework in the TEFL context, we aim to provide a nuanced understanding of cognitive engagement in TEFL teaching practice.

The ICAP framework assumes that each level of cognitive engagement describes a certain observable student activity (Chi et al., 2018): The passive level of cognitive en-

gagement refers to activities in which students do not get visibly active (e. g., reading, watching). However, the passive level does not imply the absence of learning – aligning with the constructivist notion that learning is never fully passive. The passive level merely refers to the absence of any other activity than attending to information presented by the teacher. In the ICAP framework, this level of activity is associated with low level cognitive process of storing information (Chi & Wylie, 2014). The active level is characterized by students being physically active but not generating new knowledge beyond the instructional material (e. g., copying what the teacher has written at the blackboard). The constructive level includes generating new knowledge, beyond the presented learning material (e. g., creating a mind-map based on the content of a video). The interactive level is reached when knowledge is co-generated (i. e., due to the exchange of ideas between two people, new knowledge emerges). For instance, a mind-map could be created jointly. The interactive level is only reached if the new knowledge is a product of the exchange. Thus, if students merely generate knowledge on their own and share it but do not build on the knowledge of the other students, they have not reached the interactive level.

Chi and Wylie (2014) note that although the ICAP framework assumes that the levels can be determined based on observation, distinguishing the active, constructive, and interactive level additionally requires inferring from student products. Consequently, we assume that these levels are more challenging to differentiate than to distinguish between passive and the other levels. (Roeben et al., 2025). This aligns with Chi et al. (2018) reporting that teachers struggled with differentiating active and constructive levels and with designing truly interactive learning activities.

However, there is little research on applying the ICAP framework in foreign language teaching. The present study examines how the context influences the difficulty of diagnosing levels of cognitive engagement, focusing on both the planning and implementation phase of technology-related TEFL lessons.

1.4 Instructional Quality of Tasks in Teaching English as a Foreign Language

One aim of this study is to validate the simulation we designed which is called Digivate-E. In order to do so, we assessed diagnostic skills regarding the instructional quality of tasks in TEFL lesson plans.

In contrast to cross-domain teaching criteria, evidence for subject-specific – especially TEFL- specific – criteria is limited (Praetorius & Gräsel, 2021; Wilden, 2021). In the present study, we conceptualize criteria for instructional quality of tasks in TEFL lesson plans with the core principles of modern TEFL proposed by Müller-Hartmann (2017) who bases his principles – action-orientation, interculturality, learner-centeredness, task-orientation, meaningful content, and self-regulated and cooperative learning – on the Common European Framework of Reference (Council of Europe, 2001). As participants were assumed to vary in their prior knowledge, the relevant information on these aspects of TEFL was integrated as supplementary material within the simulation and could be accessed any time.

1.5 Simulation-based Learning

Diagnosing cognitive engagement and instructional quality of technology-related TEFL lessons, both when planning and implementing them, requires diagnostic skills. As these are complex skills, they may be trained as early as in the first phase of teacher education. Simulations facilitate the acquisition of complex skills and offer the opportunity of imitating and approximating teaching practice to a certain degree while adapting the complexity to learners' current skill level (Grossman et al., 2009). Typical characteristics of simulations include the reduction in complexity of the presented situation and the opportunity for learners to interact with the simulation (Heitzmann et al., 2019). In order to ascertain that a simulation enables learners to acquire the intended knowledge and skills, it needs to be validated. Common criteria for validation are that individuals who dispose of pertinent conceptual knowledge perform better than those who do not (Kane, 2006). An additional indicator is intrinsic cognitive load – assuming that learners with lower pertinent knowledge will experience higher intrinsic cognitive load as the tasks are more difficult for them (Klepsch et al., 2017).

2 The Present Study[8]

The first aim of this study is to validate the simulation Digivate-E to ensure that it is effective for learning and to provide a valid basis for further research questions. We hope to gain insights into the predictiveness and correlations of conceptual PK, PCK, CK, TEFL-specific professional knowledge (i. e., PCK and CK; Kirchhoff, 2017) and the respective action-oriented PK and PCK (i. e., diagnostic skills in the respective domains). Moreover, by replicating parts of the study by Roeben et al., 2025, our second aim is to provide systematic research on the difficulty of diagnosing the different levels of cognitive engagement in a field different from the originally investigated one. These insights could point us towards necessary scaffolding or cues which are needed in the simulations to make it adaptive and thus more effective for learners with different learning prerequisites (Plass & Pawar, 2020).

Consequently, we pose the following research questions:

RQ1: To what extent can evidence for the validity of the simulation be found? To address this question, we investigate in detail:

RQ1.1: To what extent can we reliably assess action-oriented pedagogical knowledge and pedagogical content knowledge within the simulation Digivate-E?

8 The first author's contribution by Meral Roeben was funded by a grant from the Hanns Seidel Foundation. This research was further funded by the European Union (Next Generation EU)—and by the German Federal Ministry of Education and Research under the grant number 01JA23S01E.

RQ1.2: To what extent is conceptual knowledge (i. e., TEFL-specific professional, pedagogical knowledge) predictive for action-oriented knowledge (i. e., pedagogical content, pedagogical knowledge) as assessed within the simulation?

RQ1.3: To what extent is the intrinsic cognitive load predictive for action-oriented pedagogical knowledge within the simulation?

H1.1: We hypothesize that action-oriented pedagogical and pedagogical content knowledge show acceptable reliability within the simulation.

H1.2a: We hypothesize that conceptual TEFL-specific professional knowledge is predictive for action-oriented pedagogical content knowledge and action-oriented pedagogical knowledge.

H1.2b: We hypothesize that conceptual pedagogical knowledge is predictive for action-oriented pedagogical knowledge.

H1.3: We hypothesize that a lower reported intrinsic cognitive load is predictive for higher action-oriented pedagogical knowledge within the simulation.

RQ2: To what extent does the difficulty of diagnosing levels of cognitive engagement depend on differences in the levels of cognitive engagement (inferring vs. no inferring) within the phases of teaching (i. e., planning phase, implementation phase)?

H2a: We hypothesize that levels of cognitive engagement with no need of inferring (passive) are easier to distinguish than levels that require inferring (active, constructive, interactive).

H2b: We hypothesize that the difficulty of diagnosing the levels of cognitive engagement is different in the planning and the implementation phase.

3 Methodology

3.1 Sample and Design

The present study is an observational study with a correlational design. Between June 2023 and December 2023 we collected data sets of N = 162 pre-service teachers (all school types), studying English at a Bavarian university (degree: state exam), of which we could use N = 118 complete data sets (no missing data). Of the participants, 71 % were female, 27 % male, and 2 % diverse. On average, they had taught 37 lessons (M = 37.20; SD = 86.89). Of the three TEFL modules, participants have to take during their studies, about half of the study participants had completed the first module and half the second one (M = 1.54, SD = .71). The study participants were recruited through advertisement in seminars and on university websites. The study was conducted as a laboratory study; the participants could earn 36€ for participating. The laptops the participants conducted the study on, were provided and set up by the researchers.

3.2 Learning Environment and Participants' Tasks

For the present study, we adapted the previous simulation Digivate (Roeben et al., 2025) and developed a simulation in the context of TEFL – Digivate-E. Digivate-E can be accessed via a website and is built as a point-and-click-adventure using comic-style visuals, audios, videos, and text documents. Study participants took on the role of teacher trainees conducting their teacher training at a secondary school in a class in their third year of learning English as a second language. The class is currently reading the graded Klett teamreader *The Magic Mirror* by Josh Lacey (2019).

In their role of teacher trainees, the study participants are greeted by the seminar teacher who introduces them to their first set of tasks (planning phase; see Fig. 1). This first task consists of looking at existing lesson plans on the sequence of *The Magic Mirror*. The study participants are told to diagnose the potential level of cognitive engagement of the learning goals and of the tasks within lesson plans. They are also asked to determine the quality of the tasks from a TEFL perspective. After a total of twelve lesson plans are diagnosed, the seminar teacher introduces the participants to the second phase, the implementation phase (see Fig. 2). The seminar teacher asks the participants to accompany her to a classroom and observe the students working on the tasks. While observing them, the study participants determine the students' current level of cognitive engagement. The student activity is represented by screen-videos of the students' tablets or phones.

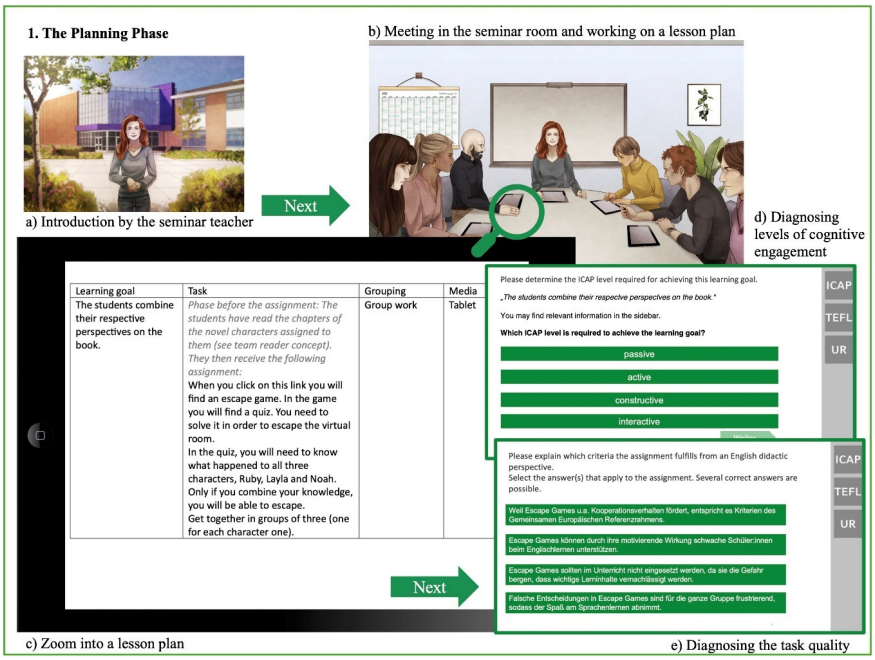

Figure 1: Overview of the Structure of the Planning Phase. (Source: "Overview of the Structure of the Planning Phase", created by Meral Roeben, pictures drawn by Nina Ploch, licensed as CC BY SA 4.0)

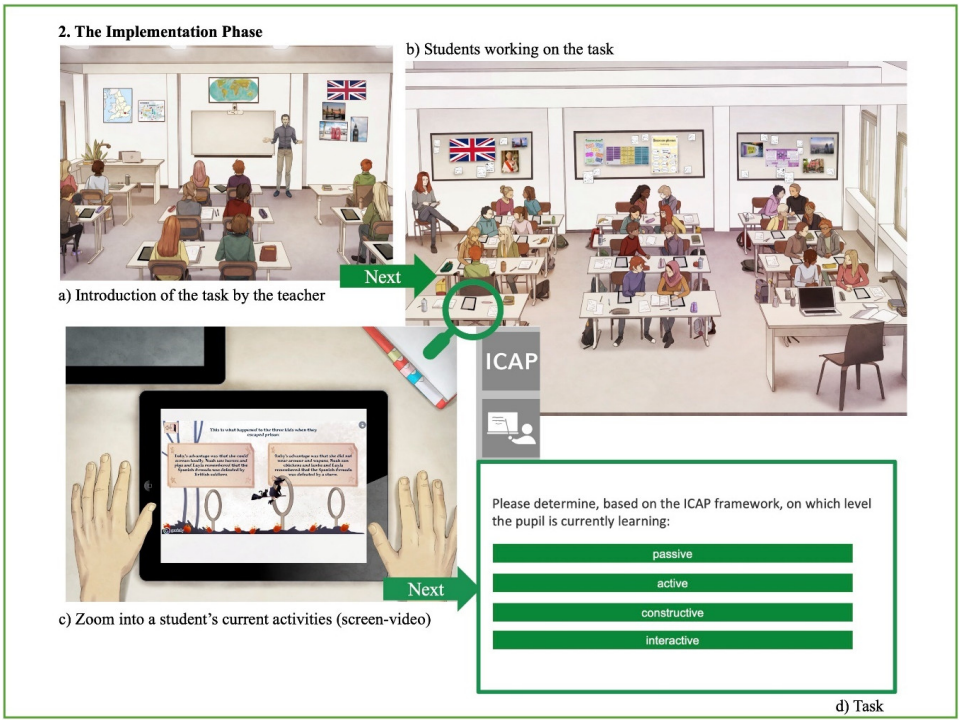

Figure 2: Overview of the Structure of the Implementation Phase (Source: "Overview of the Structure of the Implementation Phase", created by Meral Roeben, pictures drawn by Nina Ploch, licensed as CC BY SA 4.0)

In the following, we speak of cases. One case in the planning phase consists of analyzing one lesson plan (i. e., evaluating the potential level of cognitive engagement of either the learning goal or the task and justifying the TEFL quality of the task), and one case in the implementing phase consists of diagnosing the current level of cognitive engagement of one student. The cases were presented in the same sequence to all participants.

Before starting on their first case, the study participants completed a comprehensive test on conceptual PCK and CK. During their first case of the planning phase, the study participants had access to a video explaining the ICAP framework. Participants had 25 minutes to complete their first case (i. e., test case) in the planning phase. Afterwards, the participants were tested on conceptual knowledge on the ICAP framework. After that, the participants worked on eleven more cases for which they were granted six minutes each to solve them. Subsequently, in the implementation phase, the participants again first worked on a test case (ten minutes) and on eleven further cases (six minutes per case). In both phases, the participants were asked to report their intrinsic cognitive load after they had determined students' level of cognitive engagement (i. e., one validation claim; see 1.5).

We assumed participants' prior knowledge to differ as the sample included preservice teachers from different semesters and studying different school types. There-

fore, during the whole study, we offered additional information on the ICAP framework (slides from the video that was available during the first case) and on TEFL (see 1.4) in a sidebar. In the implementation phase, they also found information on the respective student they were diagnosing.

3.3 Measures

Intrinsic cognitive load was measured on a seven-point Likert scale, conceptual knowledge with open ended and closed questions, action-oriented knowledge as the accuracy of diagnostic decisions within the simulation, levels of cognitive engagement as levels of the ICAP framework, and the difficulty of diagnosing these levels by sensitivity and specificity.

3.3.1 Intrinsic Cognitive Load

The intrinsic cognitive load was measured in both phases. In the planning phase, it was measured after the level of cognitive engagement was determined for the learning goal or task and before the TEFL quality of the task was diagnosed. It was measured on a seven-point Likert scale ("do not agree at all" to "fully agree") for the two questions: "When assessing this lesson plan/student activity, I had to work on many things in my head at the same time." and "The assessment of this lesson plan/student activity was very complex." (Klepsch et al., 2017).

3.3.2 Conceptual Knowledge (TEFL-specific professional knowledge, PCK, CK, PK)

Conceptual knowledge was assessed by different tests. Conceptual PCK and conceptual CK knowledge was assessed by the FALKO-E test (Kirchhoff, 2017). FALKO-E consists of two parts. One part is a test on PCK (Cronbach's $\alpha = .50$) and one is a test on CK (Cronbach's $\alpha = .55$), both for the TEFL context. When combing the two parts of the FALKO-E test, it assesses the conceptual TEFL-specific professional knowledge (Cronbach's $\alpha = .68$; Kirchhoff, 2017). The PCK part of FALKO-E consist of 12 items, with two closed and ten open questions. For the CK part of the FALKO-E test, we assessed nine items, with four closed and five open questions. Conceptual knowledge on PK was assessed by eight single-choice items on the ICAP framework (Cronbach's $\alpha = .68$).

3.3.3 Action-oriented Knowledge (PCK, PK)

To test PCK and PK in action, we assessed action-oriented PK and PCK within the simulation. We assessed the action-oriented PCK and PK as the participants' performance, which was measured as the participants' diagnostic accuracy. For action-oriented PCK, we measured the participants' performance regarding diagnosing the TEFL-related instructional quality of the task during the planning phase of the simulation ("Please explain which criteria the task fulfills from an English didactic perspective. Select the answer(s) that apply/applies to the task. Several correct answers are possible."). The participants could choose one or more answers from a total of four answer options per question. For each accurately selected and accurately not selected option, they gained a

point. Thus, per question, they could achieve four points and a total of 20 points in the planning phase.

To evaluate the action-oriented PK, we assessed the performance regarding accurately diagnosing the levels of cognitive engagement during the planning phase ("Please assess which ICAP level is required to achieve the following learning objective/task. You may find helpful information in the sidebar.") and during the implementation phase ("Based on the ICAP framework, please assess the ICAP level the student is actually engaged on – regardless of which ICAP level the task was intended to stimulate!"). The participants could choose one of the four levels of cognitive engagement and gained one point for selecting the accurate level. They could reach eleven points in the planning phase as well as in the implementation phase, thus a maximum of 22 points for diagnosing the levels of cognitive engagement, as we did not take the first case of each phase into account for the performance measure.

3.3.4 Levels of Cognitive Engagement

Levels of cognitive engagement are conceptualized as the levels of the ICAP framework (i. e., passive, active, constructive, interactive). The accurate levels we assigned to each learning goal, task, and student activity were carefully validated in the complex process of an expert evaluation and workshop with researchers and practitioners. For diagnosing levels of cognitive engagement, we divided the levels. Active, constructive, and interactive levels require inferring from the student products to accurately determine them as merely observing students is not sufficient for these. In contrast, the passive level does not require inferring from the student product but can be determined by observing students because, in contrast to all other levels, students do not get physically active.

3.3.5 Difficulty of Diagnosing Levels of Cognitive Engagement

We assessed the difficulty of diagnosing the levels of cognitive engagement by creating confusion matrices which in the present study compare the accurate level (predicted) of cognitive engagement to the selected level (actual) of cognitive engagement. This way, we can find out which levels of cognitive engagement tend to get confused with each other. Based on confusion matrices, sensitivity and specificity can be calculated. Sensitivity is the likelihood that the accurate level is selected. If it is easy to diagnose a level, its sensitivity is high. Specificity is the likelihood that a level is not selected when it is inaccurate. If a level is easy to diagnose, its specificity is high.

3.3.6 Statistical Analyses

For RQ1, we conducted simple linear regressions and Pearson correlations to assess the associations between conceptual knowledge and action-oriented knowledge within the simulation. Additionally, we conducted simple linear regressions to check whether a reported lower intrinsic cognitive load is predictive for higher performance within the simulation (i. e., higher action-oriented PK).

To address RQ2, we made use of the signal detection theory which measures skills by considering the accurate and wrong answers (Wixted, 2020). This way, errors can be

understood better, and interventions or scaffolds can be used to prevent these errors from reoccurring (Wixted, 2020). So-called confusion matrices display how often a level of cognitive engagement was selected, for instance, if passive is the accurate level, how often was passive, active, constructive, or interactive selected. Based on this, the likelihood that an answer is accurately selected or accurately not selected can be calculated. In the present study we created confusion matrices for both phases of teaching (see Table 2): for each level of cognitive engagement in the planning phase (see Table 3 and 4), and for each level of cognitive engagement in the implementation phase (see Table 5). For the planning phase, we conducted a further differentiation by creating separate confusion matrices for learning goals (see Table 3) and tasks (see Table 4). One confusion matrix depicts the accurate level (predicted) on the x-axis and the selected level (actual) on the y-axis. Thus, the diagonal of the matrix displays the accurately selected accurate levels. These are called True Positives. Within one column, for example the active column, the incorrect levels (passive, constructive, interactive) are called False Negatives. To calculate the sensitivity, the True Positive value is divided by the sum of True Positive and False Negative values. To calculate the specificity, all occasions on which an incorrect level was accurately not selected (True Negative) are divided by the sum of the True Negative value and the False Positive value. False Positives describe all instances in which the level concerned was selected although a different level would have been the accurate choice.

Additionally, to explore whether certain levels of cognitive engagement were systematically confused more often than others, we conducted Chi-square tests, comparing the observed frequencies of confusion to the expected values under the assumption of independence. To assess whether there are significant differences between sensitivities and specificities for the different levels, we calculated confidence intervals using the Wilson method. There are significant differences between them if the confidence intervals of the sensitivity or specificity do not show any overlap.

4 Results

Regarding RQ1, validating the simulation, we summarized the descriptive results in Table 1. It shows that conceptual PCK is higher than conceptual CK, but conceptual PK exceeds all conceptual knowledge types. This is also true for the performance within the simulation as action-oriented PK is higher than action-oriented PCK.

Table 1: Standardized Descriptive Results of the Conceptual and Action-oriented Knowledge

Variables	*N*	*M*	*SD*	*Min*	*Max*
Conceptual PCK	118	.43	.14	.13	.75
Conceptual CK	118	.33	.14	.06	.67
TEFL-specific professional knowledge	118	.38	.12	.15	.71

(Continuing Table 1)

Variables	*N*	*M*	*SD*	*Min*	*Max*
Conceptual PK	118	.70	.25	.13	1.00
Action-oriented PCK	118	.68	.13	.33	.94
Action-oriented PK	118	.74	.11	.40	.95

Action-oriented PK showed an acceptable internal consistency (Cronbach's $\alpha = .66$) while it was low for action-oriented PCK (Cronbach's $\alpha = .43$; RQ1.1). Other validation claims are that a low intrinsic cognitive load is predictive for high performance (i. e., high level of action-oriented PK) and that conceptual knowledge is predictive for the performance within the simulation (i. e., higher action-oriented knowledge; RQ1.3). The intrinsic cognitive load was assessed after participants had diagnosed the level of cognitive engagement. Linear regressions show that the intrinsic cognitive load explains 8 % of the variance in action-oriented PK ($R2 = .08$) with a standardized regression coefficient for intrinsic cognitive load of $\beta = -.30$ ($p < .005$; H1.3). With regard to RQ1.2, we found that the conceptual PK explains 44 % of the variance of the action-oriented PK ($R2 = .44$) with a standardized regression coefficient for conceptual PK of $\beta = .67$ ($p < .001$; H1.2b). Moreover, conceptual TEFL-specific professional knowledge explains 8 % of variance in action-oriented PK ($R2 = .08$) with a standardized regression coefficient for conceptual TEFL-specific professional knowledge of $\beta = .29$ ($p < .001$; H1.2a). Regarding the action-oriented PCK, the conceptual TEFL-specific professional knowledge accounts for 11 % in variance ($R2 = .11$; $\beta = .34$, $p < .001$; H1.2a). The results of all linear regressions support our validity claim as they show moderate to strong, statistically significant associations of concept knowledge with action-oriented knowledge within the simulation. The residuals were approximately normally distributed, supporting the use of linear regression.

For the linear regressions, we combined conceptual PCK and conceptual PK in the construct of conceptual TEFL-specific professional knowledge. However, we also looked at these knowledge domains separately, conducting unidirectional Pearson correlations. Conceptual CK correlates higher with action-oriented PCK ($r = .39$, $p < .01$) than conceptual PCK correlates with action-oriented PCK ($r = .20$, $p < .05$). Another finding was that conceptual CK shows a significant correlation with action-oriented PK ($r = .36$, $p < .01$) while this is not the case for conceptual PCK ($r = .15$, n. s.). There is also no significant correlation between conceptual PK and action-oriented PCK ($r = .12$, n. s.). The assumption of normally distributed residuals was met.

To address RQ2, identifying the difficulty in diagnosing different levels of cognitive engagement in the planning and implementation phase of teaching, we created confusion matrices and calculated sensitivities and specificities. We created one confusion matrix including both the planning and implementation phase (see Table 2), two confusion matrices for the planning phase, namely for the learning goal (see Table 3) and the task (see Table 4), and one for implementation phase (see Table 5). We summed up the sensitivities and specificities overall (i. e., both phases), for learning goals

(i. e., part one of the planning phase), for tasks (i. e., part one of the planning phase), and for the student activities (i. e., the implementation phase) in Table 6.

While we did not include the test cases (i. e., first case of the planning and implementation phase) into the calculations regarding RQ1, we included them in the results of RQ2 as only the first case of the planning phase includes the passive level (i. e., the accurate level for the learning goal of the first case). Yet, due to this being the first case in the simulation, the respective results need to be treated with caution.

Overall (i. e., the planning and implementation phase; see Table 2) the True Positives (i. e., the occasions when the accurate level was correctly selected by participants) were the highest for all levels but for the passive level. Passive was most often diagnosed as active by participants. This confusion was confirmed as significant by a Chi-square test ($\chi 2$ (1) = 41.40, $p < .001$). Active and constructive were frequently confused with each other. Chi-square tests confirm a significant association between the active and constructive level ($\chi 2$ (1) = 286.96, $p < .001$). Interactive was sometimes confused with constructive, mostly however, it was accurately determined. This is reflected in the high sensitivity of interactive (see Table 6.). Overall, specificity was highest for passive and interactive (see Table 6). Confidence intervals for both sensitivity and specificity did not overlap, indicating that the sensitivities and specificities of all levels differed significantly from each other.

Table 2: Confusion Matrix for Diagnostic Accuracy in the Planning and Implementation Phase

		Predicted			
		P	**A**	**C**	**I**
Actual	P	39 (33 %)	55 (9 %)	15 (2 %)	4 (1 %)
	A	68 (58 %)	443 (75 %)	232 (28 %)	20 (3 %)
	C	5 (4 %)	87 (15 %)	439 (53 %)	60 (10 %)
	I	6 (5 %)	4 (1 %)	140 (17 %)	506 (86 %)

Note: The cells describe the overlap of predicted (i. e., accurate) levels and actual (i. e., selected) levels of cognitive engagement for all four levels (i. e., passive, active, constructive).

In the planning phase, we find similar patterns as we find overall (see Table 2, Table 3, Table 4) with frequent confusions of active and constructive and a high sensitivity for interactive.

In the planning phase, we will first look at the confusion matrix for learning goals (see Table 3). The True Positives (i. e., predicted level equals actual level) for passive and constructive are not the highest values. Both were most frequently determined as active. As mentioned before, for the passive level, this result is based on one case which was also the test case. As for constructive, this result is derived from four different cases, two rather in the beginning of the simulation (case 3, case 4) and two in the end (case 8, case 9). For active and interactive, the levels of cognitive engagement were most of the time determined accurately. While interactive was hardly ever confused, active was frequently

mistakenly classified as constructive. Chi-square tests confirm the significant association of active and constructive ($\chi 2\ (1) = 7.45$, $p < .05$). Sensitivity for learning goals is highest for interactive followed by active (see Table 6) with an overlap of confidence intervals for passive (.25–.42) and constructive (.30–.42), indicating that there is no significant difference in sensitivity between them. Specificity is highest for passive, followed by interactive with no overlap of any confidence intervals. In comparison to tasks (i. e., second part of the planning phase) sensitivity for learning goals is lower (see Table 6).

Table 3: Confusion Matrix for Diagnostic Accuracy for Learning Goals (Planning Phase)

		Predicted			
		P	**A**	**C**	**I**
Actual	P	39 (33 %)	2 (2 %)	6 (3 %)	0 (0 %)
	A	68 (58 %)	86 (73 %)	124 (53 %)	4 (2 %)
	C	5 (4 %)	28 (24 %)	84 (36 %)	30 (13 %)
	I	6 (5 %)	2 (2 %)	22 (10 %)	202 (86 %)

Tasks (see Table 4) are the second part in the planning phase. Here, the True Positives are the highest values for all levels. Again, the active and constructive levels are frequently confused which is confirmed by a Chi-square test ($\chi 2\ (1) = 145.39$, $p < .001$). Interactive is rarely confused, reflected in its high sensitivity (see Table 6). For sensitivity, there is no overlap of confidence intervals for any levels within tasks. Specificity is high for all levels with confidence intervals overlapping for active (.86–.91), constructive (.86–.91), and interactive (.89–.94), indicating that there is no significant difference between these specificities.

Table 4: Confusion Matrix for Diagnostic Accuracy for Tasks (Planning Phase)

		Predicted			
		P	**A**	**C**	**I**
Actual	P	NA	8 (3 %)	3 (1 %)	1 (0 %)
	A	NA	185 (79 %)	45 (19 %)	7 (3 %)
	C	NA	41 (17 %)	152 (64 %)	12 (5 %)
	I	NA	1 (0 %)	36 (15 %)	216 (92 %)

We will now focus on the implementation phase (i. e., the levels of cognitive engagements participants diagnosed for the student activities). In contrast to the planning phase, the pattern of confusing levels of cognitive engagement differs from the overall (i. e., planning and implementation phase) pattern. For the implementation phase, it is not the active and constructive levels that mostly get confused but instead the construc-

tive and interactive levels (see Table 5). Chi-square tests confirm a significant association between constructive and interactive ($\chi 2\ (1) = 90.33$, $p < .001$).

Table 5: Confusion Matrix for Diagnostic Accuracy for Student Activities (Implementation Phase)

		Predicted			
		P	A	C	I
Actual	P	NA	45 (19 %)	6 (2 %)	3 (3 %)
	A	NA	172 (73 %)	63 (18 %)	9 (8 %)
	C	NA	18 (8 %)	203 (57 %)	18 (15 %)
	I	NA	1 (0 %)	82 (23 %)	88 (75 %)

For sensitivity and specificity of interactive, we also find a new pattern in the implementation phase as both sensitivity and specificity are lower than overall and in the planning phase (see Table 6). This hints towards a higher difficulty in diagnosing interactive in the implementation phase. Confidence intervals for sensitivity overlap for active (.67–.78) and interactive (.66–.82). For specificity there is an overlap of confidence intervals between active (.81–.88) and interactive (.83–.89).

Table 6: Sensitivity and Specificity Overall, for Learning Goals, Tasks, and Student Activities

	Overall		Planning Phase				Implementation Phase	
	Both Phases		Learning Goal		Task		Student Activity	
	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
P	33 %	96 %	33 %	99 %	NA	98 %	NA	92 %
A	75 %	79 %	73 %	68 %	79 %	89 %	73 %	85 %
C	53 %	88 %	36 %	87 %	64%	89 %	57 %	90 %
I	86 %	90 %	86 %	94 %	92 %	92 %	75 %	86 %

5 Discussion

5.1 Summary of the Results

Regarding RQ1, we found that overall, the validity claim tested in the present study is supported: Conceptual knowledge (TEFL-specific professional knowledge, PK) is predictive for the respective action-oriented knowledge as shown in the performance in the simulation (PCK, PK; H1.2a; H1.2b). Yet, while conceptual PK explains a substantial proportion in variance of action-oriented PK, conceptual TEFL-specific professional knowledge shows a small effect size in predicting action-oriented PCK. Additionally, conceptual TEFL-specific professional knowledge is predictive for action-oriented PK

but shows a small effect size (H1.2a). Moreover, a low intrinsic cognitive load is predictive for higher performance regarding action-oriented PK, however the respective effect size is small (H1.3). We also found that the correlation between conceptual CK and action-oriented PCK is higher than the one between conceptual PCK and action-oriented PCK. Due to the low internal consistency of action-oriented PCK, both correlations may in fact even be stronger (Stadler et al., 2021; H1.1). Additionally, conceptual CK significantly correlates with action-oriented PK, while this is not the case for conceptual PCK.

For RQ2, we replicated Roeben et al. (2025) to find out whether the findings from the previous study also apply in a TEFL context. Overall, we found similar results: Over both phases and for the planning phase, confusion matrices show that the active and constructive levels of cognitive engagement are confused most frequently (H2). For the implementation phase, constructive and interactive levels are mostly confused (H2b). Moreover, in the implementation phase, the interactive level shows a comparatively low sensitivity (H2a). Passive, the level that does not need to be inferred from student products when determining it, shows high specificities overall and for both phases (H2a). However, in contrast to Roeben et al. (2025), we found that for learning goals, passive and constructive was surprisingly often mistakenly determined as active (i. e., True Positives < number of selected active levels). This matches the low sensitivity found for passive.

5.2 Practical and Theoretical Implications

Regarding RQ1, correlations suggest that the curriculum may place greater emphasis on CK-related courses and topics than PCLK-related ones: There is a high correlation between conceptual CK and action-oriented PK as well as a significant correlation of conceptual CK and action-oriented PK. Consequently, it may cautiously be suggested that PCK is taught insufficiently at universities. However, our finding that the average conceptual CK is lower than the average conceptual PCK (see Table 1) challenges the assumption that the focus on PCK is insufficient. One explanation may be that the test on conceptual PCK is easier than the one on conceptual CK (Kirchhoff, 2017). PCK answers might be easier to transfer from general knowledge, while for CK a participant either has a particular piece of knowledge or cannot answer the question. From this, we may derive that a more reliable test on conceptual PCK knowledge is required. With a closer look at the questions on conceptual CK, we find that the questions on linguistics and literature are framed to be relevant to the school-context (Kirchhoff, 2017). This may indicate that teaching-related conceptual CK is more important for teaching than conceptual PCK.

Furthermore, the predictiveness of conceptual TEFL-specific professional knowledge for action-oriented PK indicates that in a TEFL context, cross-domain conceptual PK is not sufficient but additional TEFL-specific professional knowledge is required (Seidel & Shavelson, 2007) to diagnose cognitive engagement. Hence, a practical implication for the first phase of teacher education may be to integrate courses on PCK and PK more closely.

With RQ2, we replicated Roeben et al. (2025) and found similar results, indicating that the ICAP framework can be applied to the TEFL context. In contrast to the active level, the constructive level presupposes a knowledge generation process. Hence, this confusion (i. e., active with constructive) may indicate that participants struggled with grasping the concept of how or whether knowledge is generated (Chi et al., 2018). To address the challenges of identifying knowledge generation, scaffolds such as guiding questions that support pre-service teachers towards understanding this process could be implemented in an updated version of Digivate-E. Additionally, feedback on learners' solution may help reflect on the reasoning process. The confusion of the constructive and interactive level as well as the relatively low sensitivity of interactive may be due to an inaccurate overgeneralization. When observing an activity in which more than one student is involved, participants may jump to the conclusion that the students are engaged on an interactive level. However, it is possible that students have a conversation to which everyone contributes; nevertheless, not all or no student at all may be engaged on an interactive level because knowledge is generated on their own and not co-generated. Manipulating small aspects of the exchange between students that highlight whether the previous information is integrated in an answer, may help making the difference between a truly interactive interaction and one in which students remain on a constructive level more salient (Plass & Pawar, 2020). These adaptive and supporting elements may be based on metrics, such as the numbers of correctly or incorrectly solved cases, as well as on the time required to answer, or on detecting certain confusion patterns (i. e., confusing active and constructive or constructive and interactive). Both in this study and in Roeben et al. (2025) learning goals showed lower sensitivity in comparison to tasks. The levels within the ICAP framework are conceptualized in terms of student activities; hence, applying them to learning goals requires extensive inferential processes. Learning goals do not focus on a certain activity but describe which skill is supposed to be acquired with an activity. Inferring the level of cognitive engagement that can potentially be achieved by the activity that a learning goal aims at is more complex than inferring the levels for tasks or student activities. This complexity can be addressed by adding a meta-perspective to the framework, such as a step-by-step instruction to the diagnostic process. Moreover, Wekerle et al. (2024) suggest redesigning the levels, which are currently rather fixed categories, into fine-grained dimensions. Hence, the strict categories could be dissolved to a certain extent, allowing for more differentiation within the levels. These precise sub-dimensions could support understanding the activity better and facilitate diagnosing the respective learning goals as they may become more feasible. Both approaches may make the ICAP framework more comprehensive in terms of including not only the implementation but also the planning phase of teaching.

The present study showed new findings regarding diagnosing the levels from stated learning goals. For the passive and the constructive level, the majority of participants diagnosed the active level (i. e., active outnumbering the passive and constructive level). The confusion of the active and constructive level fits the overall confusion pattern. Yet, the confusion of passive and active level is novel and may hint towards certain

restrictions in applying the ICAP framework to the TEFL context. In contrast to subjects taught in students' mother tongue, in foreign language teaching the medium (i. e., the language) which is used to convey content is still being learned (König et al., 2016). Thus, language has two functions: It is both the content and the means of communication (Wilden, 2021). The levels of cognitive engagement within the ICAP framework describe observable student activities. Only when taking students' products into account may we be able to make more sophisticated statements about the students' covert cognitive processes. Yet, even then, the descriptions of the levels are merely focused on generating knowledge in terms of content. The medium is not considered, leading to the conclusion that the ICAP framework neglects mediality. This is especially problematic when it comes to novice foreign language learners. For them, coordinating those two functions may be challenging while later on, when the rules of language are more automized, expert students have the capacity to focus mostly on the language as means of communication. Digivate-E, however, is set in the classroom of year seven when students are still learning the rules of a language. With that in mind, when determining the level of cognitive engagement of learning goals in the TEFL context, participants may analyze these learning goals in great depth and realize that there are two sides to a learning goal. For instance, the learning goal may consist in being able to read a certain text. This presupposes both, decoding skills (i. e., knowing the words) and linguistic comprehension (García & Cain, 2014). Participants may have struggled with considering such a complex goal as merely passive in the ICAP classification. In contrast, in their description of the ICAP framework, Chi and Wylie (2014) suggest reading as one typical example for a passive student activity. As just argued, when teaching a foreign language, reading is not an ideal example to describe a passive student activity, especially for novice students. Thus, one implication to fit the ICAP framework to the TEFL context may include considering the aspect of mediality, for instance by adding typical TEFL-specific activities to each level of the framework. At this point it may also be considered taking students' proficiency into account as reading an English text is more challenging for novice students and it is questionable whether novice learners are able to engage on a truly interactive level due to a lack of language skills. This dual perspective enables a more accurate interpretation of learners' engagement, especially when analyzing ambiguous activities such as reading, where cognitive demands may be underestimated by ICAP alone. To improve diagnostic accuracy, TEFL-specific indicators (e. g., task types, language level) should be considered in interpreting levels of cognitive engagement. This subject-specific refinement supports a more valid application of the ICAP framework in language teacher education.

5.3 Limitations

Although we validated the simulation and replicated results from a previous study, this study has limitations. Most importantly, due to its low internal consistency, interpreting results regarding action-oriented PCK should be treated with caution. The low consistency may stem from action-oriented PCK being a latent construct including various knowledge facets assessed within the simulation (Stadler et al., 2021). Additionally, the

criteria used to assess action-oriented PCK pose a limitation as empirically tested criteria are lacking. Reliable tests for (action-oriented) PCK are urgently needed.

The sample is another limitation, as the study was conducted with pre-service teachers from one Bavarian university. It would be interesting to test whether similar results emerge with pre-service teachers from other universities or with in-service teachers.

A further limitation concerns the number of passive cases. Assumptions toward the passive level are based on one case (i. e., the test case). The high confusion with the active level may thus be due to participants' unfamiliarity with the ICAP framework and simulation when working on the first case. Future studies may control for case sequence and include several cases for each level. Additionally, the findings for the passive level may imply that the ICAP framework in its current form is not able to capture the full complexity of cognitive engagement in the TEFL context. Thus, we suggest using an updated version that accounts for subject-specific particularities and better reflects the assumed depth structures of cognitive activation.

Finally, as Digivate-E was designed for validation, it does not allow participants to manipulate the simulation and thus lacks one characteristic of simulations (Heitzmann et al., 2019). A next step would be to test whether similar results occur in a version with more freedom to select or skip cases and choose additional scaffolding if needed.

5.4 Conclusion

Overall, the validity claims are supported, suggesting that Digivate-E is valid for the TEFL context and can be developed into an adaptive learning environment (Plass & Pawar, 2020). The study underscores the importance of subject-specific knowledge for cross-domain diagnostic skills (i. e., diagnosing cognitive engagement) and of integrating subject-specific and cross-domain pedagogical knowledge. Applying the ICAP framework – originally from cognitive psychology – to TEFL was largely successful: In both contexts, levels of cognitive engagement vary in diagnostic difficulty across teaching phases, and confusion patterns are similar. Still, revised versions of ICAP may need to address causes and remedies of diagnostic challenges, especially in TEFL, where adaptations are required. For instance, the framework should be extended by subject-specific activities, as it currently only lists cross-domain ones. The transfer to TEFL also enabled a more holistic analysis of language teaching, exposing the complexity of learning goals and their dual function: language as both content and medium (Wilden, 2021). While our data only allow us to infer this dual structure, it highlights that ICAP neglects mediality (i. e., language) more generally. This duality may also apply to other subjects, but further research is needed. Finally, the study provided methodological insights: By using signal detection theory (confusion matrices, sensitivity, specificity), we showed how it can be applied to evaluate teaching decisions.

References

Chi, M. T. H., Adams, J., Bogusch, E. B., Bruchok, C., Kang, S., Lancaster, M., Levy, R., Li, N., McEldoon, K. L., Stump, G. S., Wylie, R., Xu, D., & Yaghmourian, D. L. (2018). Translating the ICAP theory of cognitive engagement into practice. *Cognitive Science*, 42(6), 1777–1832. https://doi.org/10.1111/cogs.12626

Chi, M. T. H., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist*, 49(4), 219–243. https://doi.org/10.1080/00461520.2014.965823

Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge University Press. https://rm.coe.int/1680459f97

Ertmer, P. A., & Ottenbreit-Leftwich, A. T. (2010). Teacher technology change. *Journal of Research on Technology in Education*, 42(3), 255–284. https://doi.org/10.1080/15391523.2010.10782551

García, J. R., & Cain, K. (2014). Decoding and reading comprehension: A meta-analysis to identify which reader and assessment characteristics influence the strength of the relationship in English. *Review of Educational Research*, 84(1), 74–111. https://doi.org/10.3102/0034654313499616

Grossman, P., Compton, C., Igra, D., Ronfeldt, M., Shahan, E., & Williamson, P. W. (2009). Teaching practice: A cross-professional perspective. *Teachers College Record*, 111(9), 2055–2100. https://doi.org/10.1177/016146810911100905

Guttke, J. (2023). Kognitive Aktivierung im Fremdsprachenunterricht: Ein systematisches Review von Forschungsarbeiten aus dem deutschsprachigen Raum. *Zeitschrift für Fremdsprachenforschung*, 34(2), 145–176. https://doi.org/10.17185/duepublico/79249

Heitzmann, N., Seidel, T., Opitz, A., Hetmanek, A., Wecker, C., Fischer, M., Ufer, S., Schmidmaier, R., Neuhaus, B., Siebeck, M., Stürmer, K., Obersteiner, A., Reiss, K., Girwidz, R., & Fischer, F. (2019). Facilitating diagnostic competences in simulations in higher education: A framework and a research agenda. *Frontline Learning Research*, 7(4), 1–24. https://doi.org/10.14786/flr.v7i4.384

Müller-Hartmann, A. (2017). Didaktische Leitideen, Konzepte und Prinzipien im Englischunterricht der Gegenwart. In F. Haß (Ed.), *Fachdidaktik Englisch* (2nd ed., pp. 26–30). Ernst Klett Sprachen.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational Measurement* (pp. 17–64). Praeger.

Kirchhoff, P. (2017). FALKO-E: Fachspezifisches professionelles Wissen von Englischlehrkräften: Entwicklung und Validierung eines domänenspezifischen Testinstruments. In S. Krauss, A. Lindl, A. Schilcher, M. Fricke, A. Göhring, B. Hofmann, P. Kirchhoff, & R. H. Mulder (Eds.), *FALKO: Fachspezifische Lehrerkompetenzen: Konzeption von Professionswissenstests in den Fächern Deutsch, Englisch, Latein, Physik, Musik, Evangelische Religion und Pädagogik* (pp. 113–152). Waxmann.

Klepsch, M., Schmitz, F., & Seufert, T. (2017). Development and Validation of Two Instruments Measuring Intrinsic, Extraneous, and Germane Cognitive Load. *Frontiers in Psychology*, 8. https://doi.org/10.3389/fpsyg.2017.01997

Klieme, E., & Rakoczy, K. (2008). Empirische Unterrichtsforschung und Fachdidaktik. Outcome-orientierte Messung und Prozessqualität des Unterrichts. *Zeitschrift für Pädagogik*, 54(2), 222–237. https://doi.org/10.25656/01:4348

Koehler, M. J., Mishra, P., & Cain, W. (2013). What is technological pedagogical content knowledge (TPACK)? *The Journal of Education*, 193(3), 13–19. https://doi.org/10.1177/002205741319300303

König, J., Lammerding, S., & Nold, G. (2016). Teachers' professional knowledge for teaching English as a foreign language: Assessing the outcomes of teacher education. *Journal of Teacher Education*, 67(4), 320–337. https://doi.org/10.1177/0022487116644956

Kopp, V., Stark, R., & Fischer, M. R. (2008). Fostering diagnostic knowledge through computer-supported, case-based worked examples: Effects of erroneous examples and feedback. *Medical Education*, 42(8), 823–829. https://doi.org/10.1111/j.1365-2923.2008.03122.x

Kramer, M., Förtsch, C., Seidel, T., & Neuhaus, B. J. (2021). Comparing two constructs for describing and analyzing teachers' diagnostic processes. *Studies in Educational Evaluation*, 68, 100973. https://doi.org/10.1016/j.stueduc.2020.100973

Lacey, J. (2019). *The Magic Mirror*. Ernst Klett Sprachen.

Plass, J. L., & Pawar, S. (2020). Toward a taxonomy of adaptivity for learning. *Journal of Research on Technology in Education*, 52(3), 275–300. https://doi.org/10.1080/15391523.2020.1719943

Praetorius, A.-K., & Gräsel, C. (2021). Noch immer auf der Suche nach dem heiligen Gral: Wie generisch oder fachspezifisch sind Dimensionen der Unterrichtsqualität? *Unterrichtswissenschaft*, 49(2), 167–188. https://doi.org/10.1007/s42010-021-00119-6

Roeben, M., Vejvoda, J., Murböck, J., Fischer, F., Schultz-Pernice, F., Lohr, A., Stadler, M., Sailer, M., & Heitzmann, N. (2025). Simulations in Teacher Education: Learning to Diagnose Cognitive Engagement. *Education Sciences*, 15(3), 261. https://doi.org/10.3390/educsci15030261

Seidel, T., & Shavelson, R. J. (2007). Teaching effectiveness research in the past decade: The role of theory and research design in disentangling meta-analysis results. *Review of Educational Research*, 77(4), 454–499. https://doi.org/10.3102/0034654307310317

Stadler, M., Sailer, M., & Fischer, F. (2021). Knowledge as a formative construct: A good alpha is not always better. *New Ideas in Psychology*, 60, 100832. https://doi.org/10.1016/j.newideapsych.2020.100832

Urhahne, D., & Wijnia, L. (2021). A review on the accuracy of teacher judgments. *Educational Research Review*, 32, 100374. https://doi.org/10.1016/j.edurev.2020.100374

Wekerle, C., Daumiller, M., Janke, S., Dickhäuser, O., Dresel, M., & Kollar, I. (2024). Putting ICAP to the test: How technology-enhanced learning activities are related to cognitive and affective-motivational learning outcomes in higher education. *Scientific Reports*, 14(1), 16295. https://doi.org/10.1038/s41598-024-66069-y

Wilden, E. (2021). Fachspezifische Aspekte von Unterrichtsqualität im Schulfach Englisch. *Unterrichtswissenschaft*, 49(2), 211–219. https://doi.org/10.1007/s42010-021-00105-y

Wixted, J. T. (2020). The forgotten history of signal detection theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 46(2), 201–233. https://doi.org/10.1037/xlm0000732