EMPIRISCHE SONDERPÄDAGOGIK

ISSN 1869-4845 17. Jahrgang 2025 | Heft 2

PABST

Empirische Sonderpädagogik

17. Jahrgang · Heft 2 · 2025

Inhalt

Impressum

Herausgeber:
Prof. Dr. Matthias Grünke
Universität zu Köln
Humanwissenschaftliche Fakultät
Department Heilpädagogik und
Rehabilitation
Klosterstr. 79b
50931 Köln
matthias.gruenke@uni-koeln.de

Prof. Dr. Timo Lüke
Karl-Franzens-Universität Graz
Inklusive Bildung und Unterrichtsentwicklung
Merangasse 70/II
8010 Graz, Österreich
timo.lueke@uni-graz.at

Prof. Dr. Jürgen Wilbert
Universität Münster
Professur für Erziehungswissenschaft mit dem
Schwerpunkt Sonderpädagogik im Kontext
inklusiver Bildung
Georgskommende 26
48143 Münster
juergen.wilbert@uni-muenster.de

Verlag:
Pabst Science Publishers
Eichengrund 28
D-49525 Lengerich/Westfalen
Tel. +49(0) 5484-308
Fax +49(0) 5484-550
E-Mail: pabst@pabst-publishers.com
www.pabst-publishers.com
www.psychologie-aktuell.com

Erscheinungsweise:
4xjährlich

Preise:
(Jahresabonnement: 40,- €
Einzelheft: 15,- €
inkl. MwSt. und Versand)

Verlagsredaktion:
Erika Wiedenmann
wiedenmann@pabst-publishers.com

Herstellung:
Patrick Orths
orths@pabst-publishers.com

Administration:
Karin Moenninghoff
moenninghoff@pabst-publishers.com

Druck:
KM-Druck 2.0, D-64823 Groß-Umstadt

ISSN 1869-4845 (Printausgabe)
ISSN 1869-4934 (ebook)

Empirische Sonderpädagogik, 2025.17:124-141
DOI https://doi.org/10.2440/003-0044
ISSN 1869-4845 (Print) · ISSN 1869-4934 (ebook)

Scenario-Based Case Game to Promote Diagnostic Decision-Making in Inclusive Teacher Education

Judith Zellner & Markus Gebhardt

Universität München

Abstract

Diagnostic decision-making enables teachers to make accurate, data-informed instructional decisions. This study examined how teachers demonstrate diagnostic competences in simulations or how these vary by professional focus. Therefore *N* = 280 pre-service and in-service primary and special education teachers completed two digital status diagnostics cases using a scenario-based simulation game. Each case addressed either mathematics or reading difficulties and was randomly presented as case 1 or case 2. Participants selected diagnostic tests, interpreted data, and planned interventions. Click behavior (test selection, processing time) and open responses (interpretation and instructional planning quality) were analyzed. Results showed high diagnostic accuracy with in case 1 71.43% and in case 2 73.74 % of participants finding the appropriate test with the first click. Efficiency increased from case 1 to case 2, with the number of selected tests decreasing from *M* = 3.15 to *M* = 2.66 although confirmatory clicking persisted. No differences were found between reading and mathematics cases or professional groups—except for processing time: primary education participants responded significantly faster than special education participants (*p* < .001). A latent profile analysis identified three profiles: (1) uncertain-intuitive decision-makers with low diagnostic quality (*N* = 56), (2) reflective-deliberate decision-makers with high quality and long processing times (*N* = 60), and (3) efficient-intuitive decision-makers relying on heuristics, with high quality but low self-assessment (N = 144). The potential of scenario-based games in teacher education and research is discussed and underscores especially for profile 2 the need for adaptive, data-informed support for individual differences in diagnostic decision-making.

Keywords: Data-based decision-making, diagnostic competences, inclusive teacher education, E-learning, scenario-based learning game

Ein simulationsbasiertes Fallspiel zur Förderung diagnostischer Entscheidungen in der inklusiven Lehrkräftebildung

Zusammenfassung

Diagnostische Kompetenz befähigt Lehrkräfte zu präzisen, datengestützten Unterrichtsentscheidungen. Diese Studie untersuchte das diagnostische Entscheidungsverhalten von N = 280 Studierenden und Lehrkräften der Grundschul- und Sonderpädagogik in zwei szenariobasierte Onlinefällen zum statusdiagnostischen Prozess. Jeder Fall behandelte entweder Mathematik- oder Leseschwierigkeiten und wurde nach dem Zufallsprinzip als Fall 1 oder Fall 2 präsentiert. Dabei wurden das Klickverhalten (Testauswahl, Durchführungszeit) und offene Antworten (Interpretation, Fördermaßnahmen) analysiert. Die diagnostische Akkuratheit war von Beginn an hoch: bei Fall 1 fanden 71.43 % mit dem ersten Klick den passenden Test, bei Fall 2 73.74 %. Die diagnostische Effizienz stieg im zweiten Fall (M = 2.66 Tests vs. M = 3.15), obwohl weiterhin absicherndes Klickverhalten erkennbar blieb. Vorwissen und Praxiserfahrung beeinflussten die Bearbeitungsdauer, nicht jedoch die diagnostische Qualität des Klickverhaltens. Teilnehmende mit Grundschulstudium arbeiteten signifikant schneller als mit Studium der Sonderpädagogik ($p < .001$). Eine latente Profilanalyse identifizierte drei Entscheidungsprofile: (1) unsicher-intuitiv mit geringer Qualität (N = 56), (2) reflektiert-überlegt mit hoher Qualität und langer Bearbeitungszeit (N = 60), (3) effizient-intuitiv mit Heuristiken, hoher Qualität und niedriger Selbsteinschätzung (N = 144). Die Ergebnisse zeigen eine individuelle Entwicklung diagnostischer Kompetenz und betonen die Notwendigkeit adaptiver, datenbasierter Unterstützung insbesondere für Studierende und Lehrkräfte mit dem Profil 2. Das Potenzial szenariobasierter Spiele für die inklusive Lehrerkräfteausbildung und Forschung wird diskutiert.

Schlagwörter: Datenbasierte Entscheidungsfindung, diagnostische Kompetenzen, inklusive Lehrerbildung, E-Learning, simulationsbasiertes Lernspiel

Educational diagnostic decision-making

Educational diagnostic competence refers to the ability to recognize, interpret, and use relevant information to make sound instructional decisions (Gebhardt, 2023). Depending on the context, this competence is applied at two diagnostic approaches: Approach 1, individual-focused diagnostics, typically rooted in special education and school psychology, emphasize case-based decision-making (Klauer, 1978). In this context, diagnostic competence involves identifying a learner's specific needs, interpreting individualized data (e.g., behavioral observations, test results), and using this information to design individualized education programs—such as data-based decisions within response-to-intervention frameworks (Fuchs et al., 2012). Diagnostic decision-making in the classroom as approach 2 focusses on assessing the learning needs of many students in order to make instructional decisions that benefit the students of the class. Here, the emphasis is often on assessment accuracy—teachers' ability to evaluate student performance across multiple learners and adjust instruction accordingly (Herppich et al., 2017). In the context of inclusive schooling and the division of roles between special education professionals and general education teachers, these decisions are primarily attributed to special education professionals in the sense of diagnostic expertise (Kluge & Grosche, 2022). These approaches represent two complementary sides of the same coin, but can also contradict each other in the practice of inclusion, as each approach can lead to different decision-making and support in practice.

Demands on rational and intuitive decisions in the classroom

Especially in the classroom, teachers usually have to react immediately and make quick decisions (approach 2). In contrast, approach 1 used for systematic diagnostic processes or team-based decisions are more reflective and time-intensive and therefore occur less frequently in everyday practice. Gigerenzer and Gaissmaier (2011) describe decision-making competence as the ability to distinguish between intuitive and reflective decisions and apply both contextually. Intuitive decisions draw on experience and emotion, enabling quick, often unconscious judgments (Gigerenzer & Goldstein, 1996). Such decisions require little cognitive effort, but in complex school settings, solely relying on a "gut feeling" is not sufficient. Instead, intuitive teacher decisions are based on professional experience and secured decision patterns. Reflective decisions, on the other hand, are typically slower and demand more cognitive resources. They involve the systematic evaluation of diagnostic data and are often used in complex situations where experience alone is insufficient. However, more information does not always improve decisions; it often strains resources (Gigerenzer & Gaissmaier, 2011). To meet the demands of everyday school, both approaches of decision-making are necessary—particularly reliable heuristics. Heuristics, trained action patterns, support intuition and enable fast, accurate decisions (Gigerenzer & Goldstein, 2011). In clinical reasoning, heuristics are effectively developed through case-based learning (Norman, 2005). In educational diagnostics, a secure heuristic emerges when teachers test purposefully without the need to fully verify every aspect of their decision (Reimer et al., 2007).

However, reflective, heuristic decisions are demanding and therefore tend to be avoided by teachers without sufficient decision-making competences (Stecker et al., 2005). Efficient decision-making means conducting few tests - accuracy means selecting the right test to gather the information needed (Heitzmann et al., 2019). Translating diagnostic data into appropriate interpretations and support measures necessitates subject-specific pedagogical knowledge and awareness of the multidimensional nature of learning difficulties (Brandt, 2022). Selecting interventions further depends on their practical feasibility and scientific basis to ensure sustainable, targeted support (Kuhl et al., 2021). How demanding such secure decisions are for teachers depends both on training and on certain key characteristics (Südkamp et al., 2017).

Teacher characteristics influencing decision-making

Key teacher characteristics affecting diagnostic decision-making competences in classrooms include experience, subject knowledge, attitudes, and intelligence (Südkamp et al., 2017). Experience is often assumed to improve diagnostic accuracy, as it offers more practice in making accurate and efficient decisions. Yet, findings vary. McElvany et al. (2009) found only a weak correlation; other studies showed none (Wild & Rost, 1995; Praetorius et al., 2011). Brunner et al. (2011) argued that subject-specific teaching skills, not experience, drive accuracy. Lorenz (2011) studied primary teachers' estimates of students' math and language skills. Accuracy varied little overall, with differences tied more to teachers' diagnostic attitudes. Negative attitudes and biases impair diagnostic decision-making (Klug et al., 2016; Glock & Kleen, 2023). Teachers with positive attitudes and higher self-efficacy invest more time in diagnostic decisions, reflecting on results and using them for planning (Ohle et al., 2015). Although research remains inconclusive regarding which teacher characteristics most strongly affect diagnostic decision-making in the classroom an in what manner, there is growing agreement that perceived self-efficacy influence actual diagnostic decision-making behavior (Südkamp et al., 2017). Based on this assumption, the present study measures self-efficacy using a self-assessment questionnaire. While this instrument does not capture underlying attitudes or cognitive ability directly, it reflects teachers' perceived diagnostic competence and their confidence in making educational decisions. Building on this understanding, the development of diagnostic decision-making competences requires targeted knowledge, repeated practice, and positive efficacy experiences—which need to be systematically embedded into teacher education programs.

Framework for fostering diagnostic decision-making competences

Approaches to fostering decision-making competence assume it develops through combining theoretical knowledge with guided practical reflection (Südkamp et al., 2017). Heitzmann et al. (2019) proposed a framework defining diagnostic competence as a disposition comprising knowledge, quality, and activities. Diagnostic knowledge includes conceptual understanding (concepts and their relations) and strategic knowledge (paths and heuristics for diagnosis). Diagnostic quality refers to the accuracy and efficiency of decisions. Diagnostic activities cover actions in the process, such as gathering information and selecting tools. Context moderates diagnostic performance, leading authors to later emphasize authenticity as a key factor (Chernikova et al., 2020). Interactive, scenario- or simulation-based learning enables strategic-diagnostic knowledge to be initiated and applied diagnostically. This study is therefore the first to apply the scenario-based approach to diagnostic decision-making competence in special education and elementary school students.

Fostering decision-making competences by scenario-based learning

Scenario-based learning, also called case-based or problem-oriented learning, prepares pre-service teachers for classroom practice (Caukin et al., 2016). It is based on the theories of situated learning (Lave & Wenger, 1991) and situated cognition (Brown et al., 1989), which hold that learning is most effective in realistic contexts. Working through case scenarios with interactive decision paths prompts pre-service teachers to reflect on diagnostic processes and develop professional decision-making competence (Zellner et al., 2024a). Research shows that scenario-based learning increases self-efficacy in training across fields like medicine, policing (McLean, 2016), and teacher education (Caukin et al., 2016).

Providing scenario-based learning in an online environment offers distinct advantages. First, digital scenarios reduce the need for real-time instructor presence and conserve time and spatial resources (Prilop et al., 2020). Online scenario-based learning scales easily to large groups, making it attractive for teacher training and research as user data offer insights into teachers' diagnostic competence. Second, digital platforms create a safe space to test diagnostic decisions. Pre-service teachers can simulate conflicts and challenges from future practice without exposing real students—or themselves—to ethical risks (Badiee & Kaufman, 2015). Third, online learning environments are cost-effective and flexible, using widely available technologies. It can be integrated into teacher education programs or serve as standalone professional development (Badiee & Kaufman, 2015).

Research questions

Pedagogical diagnostics requires developing robust heuristics to enable accurate and efficient decision-making (Heitzmann et al., 2019). Interactive learning formats that simulate diagnostic processes in realistic contexts can strengthen this competence, fostering data-driven decision confidence already in teacher education.

This study examines an online, scenario-based game that simulates diagnostic decisions through interactive processes. It offers practice opportunities for both pre-service and in-service teachers while generating research data on decision behavior. Comparing primary and special education teachers and students, the study addresses two research questions:

- RQ1: Can educators with different theoretical and practical backgrounds complete scenario-based tasks and find the appropriate test, data-driven interpretation and data-based instructional support decision?
- RQ2: Do primary and special education teachers and students differ in their decision behavior?
- RQ3: Which typical decision patterns indicate data-based training needs for teacher education?

Materials and Methods

Sample

The instrument was tested with 280 educators (242 female, 36 male, 2 non-binary; *M* age = 28.48, *SD* = 10.68) from University of Munich, specifically from the Chairs of Special Education and Primary Education, representing diverse theoretical and practical educational focus. Participants were divided into four groups based on their educational focus—special or primary education—and their professional status as students or in-service teachers. The group of 90 special education students (SES) was, on average, in their 6th semester (*M* = 6.43, *SD* = 2.1). Their curriculum includes diagnostic theory and case-based learning in semesters four and five, providing a solid theoretical foundation. In contrast, the 79 primary education students (PES) were, on average, in their 6th semester as well (*M* = 5.64, *SD* = 2.18), though diagnostic content is not systematically integrated into their study program (Brandt, 2022). Instead, diagnostic knowledge is typically conveyed implicitly within subject-specific didactics. The 63 in-service special education teachers (SET) had an average of *M* = 12 years of teaching experience (*SD* = 11.7) at the time of data collection, while the 48 in-service primary education teachers (PET) reported an average of *M* = 8.5 years of teaching experience (*SD* = 9.84). This difference in years of teaching experience between SET and PET was not statistically significant, t(104.81) = 1.69, p = .094, with a small effect size (*d* = 0.32, 95% CI: [-0.07, 0.71]).

Design and Instruments

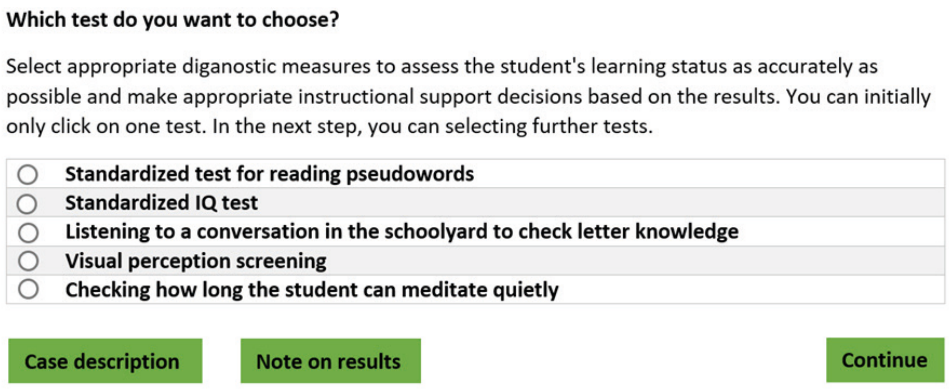
DaKI. All participants completed the *DaKI* self-assessment questionnaire on diagnostic competence in inclusive schools (Jungjohann & Gebhardt, 2023) before the task. The *DaKI* is a psychometric tool designed to measure teachers' diagnostic competence in inclusive education settings with 28 items on a five-point Likert scale ("strongly disagree" to "strongly agree"). It assesses four dimensions: (1) instructional decision-making (e.g. „I know ways to promote written language or mathematical skills."), (2) educational assessment (e.g. „I know several informal and standardized school performance tests."), (3) identification of special educational needs (e.g. „I know the necessary steps to determine the need for special educational support."), and (4) progress monitoring (e.g. „I can diagnose and evaluate learning progression."). Scale evaluation with *N* = 252 participants, including 152 special education students and 100 in-service teachers, showed high internal consistency across all dimensions (Cronbach's α = .82–.93). For our sample Cronbach's Alpha was across all dimensions at α = .87–.93.

Scenario-based game (SBG). The study employed a 2x2 experimental design with randomized case order. Each participant worked through four cases, with participants assigned case 1 and 2 as well as case 3 and 4 in randomized order. To account for increasing task complexity (Ebenbeck et al., 2022), the first two cases focused on status diagnostics (= case 1 and case 2), followed by two cases on learning progress diagnostics (= case 3 and 4), whereby only the data of the first two cases on status diagnostics are analyzed in this article. The interactive click-based task was administered as an online survey via SoSciSurvey (Leiner, 2024) on participants' personal digital devices.

The SBG simulates diagnostic processes leading to pedagogical decisions. The focus is not on evaluating the SBG itself but on assessing the decisions and click behavior of participants when processing these digital interactive cases. Structured as a decision tree with branching points, it offers a more interactive practice environment than traditional case vignettes with static solution keys (Zellner et al., 2024a). Administering the SBG via SoSciSurvey allows the collection of log data, including click paths, response patterns, and processing times, enabling detailed tracking of participants' practice behavior. The SBG includes two status diagnostics cases (case 1 and case 2) —one in mathematics and one in reading —with random assignment of participants to the subjects. The detailed case description, data and code can be viewed here: http://bit.ly/4idAnan

Each SBG case begins with a task and case description, providing comparable baseline information for both mathematics and reading. Case descriptions are designed so that the primary learning difficulty becomes clear, while potential comorbid issues can be ruled out. Both cases (mathematics and reading) are designed to be as unambiguous as possible, so that even beginners without in-depth didactic knowledge can work through them, ensuring that the focus remains on diagnostic decision-making competence. Accordingly, questions are posed regarding the selection of an appropriate test (Figure 1), including choices between standardized and informal procedures. In this first active step, participants select a diagnostic instrument to clarify the described difficulty. Five options are provided, differing in their relevance to the case and their degree of standardization. One standardized test fits the case perfectly and provides sufficient diagnostic information. The other tests merely indicate that the issue does not lie within their respective domains. In line with efficient classroom-based diagnostics, resource-intensive procedures such as intelligence testing could be excluded based on the case description, as they were neither appropriate nor necessary for the presenting problem.

Next, participants interpret all available case information—both the description and test results—in an open-text field. The goal is a comprehensive, data-driven interpretation that identifies the primary difficulty, confirms it based on evidence, and labels it with a precise diagnostic term. Following this, participants formulate data-based instructional support decisions in a second open-text field. These should be case-specific, reference the test results, state a clear support goal, and suggest con-



Which test do you want to choose?

Select appropriate diganostic measures to assess the student's learning status as accurately as possible and make appropriate instructional support decisions based on the results. You can initially only click on one test. In the next step, you can selecting further tests.

○ **Standardized test for reading pseudowords**

○ **Standardized IQ test**

○ **Listening to a conversation in the schoolyard to check letter knowledge**

○ **Visual perception screening**

○ **Checking how long the student can meditate quietly**

Figure 1
Test selection of SBG

crete measures, including materials or evidence-based interventions.

During the task, participants can revisit the case description and test results (e.g., percentile ranks) at any time. However, no back button is provided to prevent changes to prior selections and ensure traceability of decision paths. Between cases, brief prompts support diagnostic accuracy and efficiency, such as excluding areas before selecting a test (“Consider where the difficulty likely lies”), aiming for test accuracy (“Identify the most fitting test”), and recognizing sufficiency (“One appropriate test can yield adequate information”).

Analysis

All data analysis was performed in *R* and *RStudio* with the dplyr package for descriptive analysis (Wickham et al., 2023). Analyses focused on the two randomized status diagnostics cases (case 1 and case 2) within the SBG, describing the decisions and click behavior in each case and not whether there was an intervention effect.

Participants were expected to demonstrate high diagnostic quality, defined by accurate, efficient, and data-based decision-making (Heitzmann et al., 2019). To operationalize this, each correct test selection and each appropriate omission of irrelevant tests was rated with one point, with a maximum of five points. The variables used to assess diagnostic quality included the accuracy of test selection, the quality of interpretation and the quality of instructional support planning. The processing time was recorded for each case and evaluated as an indicator of decision efficiency.

Open-text responses for data-driven interpretation and instructional support planning were coded by two independent raters based on a theory-driven rubric (Heitzmann et al., 2019; Brandt, 2022), with scores ranging from 0 (“inadequate”) to 5 (“fully adequate”). Points were awarded for domain-specific precision, consideration of multidimensionality, data literacy, and application-oriented planning, ensuring individual, evidence-based support (Brandt, 2022; Kuhl et al., 2021). Interrater reliability was assessed using Cohen’s Kappa for ordinal data, indicating good initial agreement (Döring & Bortz, 2016): Item 1: $\kappa = 0.62$, 95% CI [0.55, 0.69]; Item 2: $\kappa = 0.52$, 95% CI [0.44, 0.60]; Item 3: $\kappa = 0.72$, 95% CI [0.65, 0.79]; Item 4: $\kappa = 0.66$, 95% CI [0.59, 0.73]. Discrepant cases were discussed, leading to final, weighted Kappa values indicating very high agreement: Item 1: $\kappa = 0.86$, 95% CI [0.81, 0.90]; Item 2: $\kappa = 0.85$, 95% CI [0.79, 0.90]; Item 3: $\kappa = 0.92$, 95% CI [0.88, 0.97]; Item 4: $\kappa = 0.93$, 95% CI [0.89, 0.96].

To assess overall diagnostic quality, test selection accuracy and total scores were analyzed using the Wilcoxon signed-rank test for paired samples, due to non-normal distribution. In addition, Spearman’s correlation analysis (visualized with the corrplot package; Wei & Simko, 2024) was used to examine associations between test selection, interpretation quality, and planning quality.

To examine group differences, diagnostic behavior across the four subgroups (pre-service and in-service teachers from primary and special education) was analyzed for all key indicators: The primary outcome variables were the accuracy and efficacy of test selection (0-5 points), interpretation (0-5 points) and the instructional support decision (0-5 points). Additionally, processing time (in seconds) was analyzed as a behavioral indicator of confidence in diagnostic decision making and click behavior in digital learning environments. Processing time was examined using the Kruskal-Wallis test, followed by Dunn’s post hoc tests with Bonferroni correction for multiple comparisons. Group differences in test scores and open-response evaluations were tested using ANOVA for continuous variables and chi-square tests for categorical outcomes, with post hoc comparisons where appropriate.

To identify diagnostic decision patterns, a Latent Profile Analysis (LPA) was conducted using the mclust package (Scrucca et al., 2023). Variable selection was guided by theoretical decision-making models (Gigerenzer & Gaissmaier, 2011) and Variable Importance Analysis from Random Forest in Machine Learining by using tidymodels package in R (Kuhn & Wickham, 2020), identifying the strongest predictors of overall diagnostic performance. The Variables Importance Analysis in Random Forest models estimates how important each variable is to the model's ability to make accurate predictions. The Mean Decrease in Accuracy method assesses the importance of a variable by measuring how much the predictive accuracy of the model decreases when the values of that variable are randomly interchanged. A significant decrease in accuracy indicates that the variable is important to the model. Profile-specific differences were analyzed using ANOVA for continuous measures and chi-square tests for categorical variables.

Results

DaKI

Self-assessments of diagnostic competence across all four groups revealed moderate to below-average ratings on all *DaKI* dimensions. All dimensions combined, SET reported the highest competence levels (*M* = 3.01, *SD* = 1.00), followed by PET (*M* = 2.73, *SD* = 0.80). SES (*M* = 2.69, *SD* = 1.18) and PES (*M* = 2.64, *SD* = 0.68) rated their competence lowest. Group differences were not statistically significant (*p* > .05), diverging from prior DaKI evaluations, where in-service teachers rated their competence significantly higher than students (Jungjohann & Gebhardt, 2023).

Diagnostic Quality Across Cases

Test Selection Accuracy. Analysis of participants' click behavior revealed high initial accuracy in selecting diagnostic tests. In case 1, 71.43% of participants chose the correct test on their first click. Accuracy remained similarly high in case 2, with 73.74% selecting the correct test on the first attempt. As diagnostic performance did not differ significantly (*p* > .05) between subject areas (mathematics vs. reading) or between students and teachers, data from case 1 and case 2 were analyzed based solely on sequence.

Despite the frequent initial selection of the correct test, participants showed a tendency toward confirmatory clicking. Out of five available options, participants selected, on average, *M* = 3.15 tests (*SD* = 1.09) in case 1, which decreased to *M* = 2.66 tests (*SD* = 1.17) in case 2. Overall, participants selected fewer tests in case 2, and the correct test was identified more frequently (39.21%). Nonetheless, additional information was often retrieved across all groups, suggesting persistent uncertainty. After the appropriate standardized test, "meditation" was the second most frequently selected option in both cases (20.2% in case 1; 19.53% in case 2), followed closely by "informal observation" (17.60% in case 1; 16.45% in case 2).

Combining accuracy and efficiency scores showed an improvement in diagnostic quality from case 1 to case 2 (Table 1). In case 1, 45% of participants scored four points, while 13.21% achieved the maximum of five points. In case 2, the proportion scoring five points increased to 37.14%, and 33.21% scored four points. A stable subgroup of 16–19% demonstrated consistently low diagnostic quality. A Wilcoxon signed-rank test indicated a significant increase in test-selection scores in case 2 (*p* < .05). On the individual level, 128 participants (45.71%) improved, 101 (36.07%) remained unchanged, and 51 (18.21%) performed worse from case 1 to case 2.

Interpretation and Intervention Planning. Scores for interpretation and intervention showed a concentration around the midpoint, with few extreme values (*M* interpretation case 1 = 2.41, *SD* = 1.27; *M* interpretation case 2 = 2.32, *SD* = 1.14; *M* interventions case 1 = 2.30, *SD* = 1.23; *M* interventions case 2 = 2.48, *SD* = 1.03). Although diagnostic quality in test selection improved significantly from case 1 to case 2 (Wilcoxon, $p < .001$), open response quality for interpretation ($p = .22$) and interventions ($p = .08$) showed no significant change. Spearman´s correlation analysis revealed moderate positive associations between interpretation and intervention scores in case 1 ($r = .56$). These correlations remained stable across cases ($r = .52$ for interpretation; $r = .48$ for interventions), suggesting consistency in participants' diagnostic reasoning over both cases.

Differences between primary and special education teachers and students

Group comparisons revealed no significant differences in overall test selection scores ($p > .05$). Descriptive data showed that in case 1, PES and PET selected more tests (*M* = 3.45; *M* = 3.56) than SES and SET (*M* = 2.97; *M* = 2.89). In case 2, PES again made the most selections (*M* = 3.13), while SES and SET showed the fewest clicks (*M* = 2.48). As shown in Figure 2, in open-response quality, also no significant differences between groups emerged across either case ($p > .05$).

Processing time decreased from 9.50 minutes (*M* = 570.52 seconds, *SD* = 633.43) in case 1 to 5.50 minutes (*M* = 332.78 seconds, *SD* = 327.56) in case 2. A Kruskal-Wallis test revealed significant group differences ($p < .001$). PES and PET consistently completed both cases faster than SES and SET. While SES and SET required more time in case 1, they narrowed the gap in case 2, suggesting a possible learning effect.

Profiles of Diagnostic Decision-Making

Although participants grouped by educational focus showed largely similar task behavior, further differences emerged in performance patterns. To capture these individual variations, a person-centered LPA was conducted.

Table 1

Descriptive statistics (M and SD) by group for each outcome variable.

Variable	SES	SET	PES	PET
DaKI	*M* = 2.69, *SD* = 1.18	*M* = 3.01, *SD* = 1	*M* = 2.64, *SD* = 0.68	*M* = 2.73, *SD* = 0.8
Case 1 – Test selection	*M* = 3.53, *SD* = 1.14	*M* = 3.38, *SD* = 1.24	*M* = 3.52, *SD* = 1.12	*M* = 3.1, *SD* = 1.36
Case 2 – Test selection	*M* = 3.9, *SD* = 1.34	*M* = 3.43, *SD* = 1.74	*M* = 3.78, *SD* = 1.34	*M* = 3.62, *SD* = 1.57
Case 1 – Interpretation	*M* = 2.58, *SD* = 1.26	*M* = 2.84, *SD* = 1.21	*M* = 1.95, *SD* = 1.06	*M* = 2.29, *SD* = 1.44
Case 2 – Interpretation	*M* = 2.68, *SD* = 1.16	*M* = 2.62, *SD* = 1.16	*M* = 1.94, *SD* = 0.79	*M* = 1.9, *SD* = 1.26
Case 1 – Intervention	*M* = 2.54, *SD* = 1.29	*M* = 2.72, *SD* = 1.33	*M* = 1.94, *SD* = 0.91	*M* = 2.06, *SD* = 1.33
Case 2 – Intervention	*M* = 2.74, *SD* = 0.95	*M* = 2.65, *SD* = 1.19	*M* = 2.23, *SD* = 0.86	*M* = 2.19, *SD* = 1.1
Case 1 – Processing Time (s)	*M* = 1014.85, *SD* = 3001.63	*M* = 2056.38, *SD* = 9836.92	*M* = 347.95, *SD* = 127.36	*M* = 350.79, *SD* = 279.5
Case 2 – Processing Time (s)	*M* = 428.2, *SD* = 391.35	*M* = 381.21, *SD* = 401.03	*M* = 232.64, *SD* = 136.24	*M* = 685.49, *SD* = 3017.38

Note. Values represent means and standard deviations per group. Time measured in seconds.

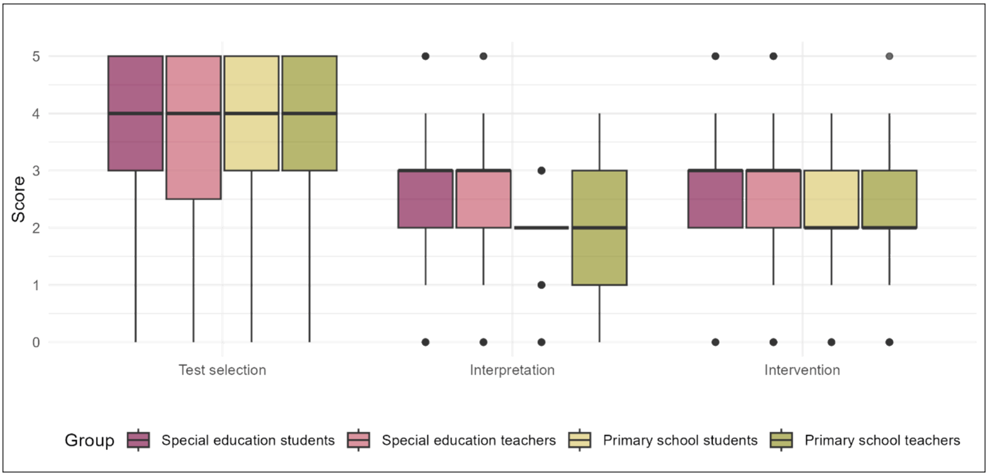


Figure 2
Diagnostic quality across subtasks and groups

The variable-importance analysis using a Machine Learning Random Forest model identified the click-based item test selection as the strongest predictor of overall diagnostic quality in the SBG task (Importance = 2.75). Open-response components (interpretation and intervention planning) had notably lower predictive value (Importance < 0.57). This decision aligns with theoretical reasoning: test selection represents the most standardized and objectively assessable subtask, making it the most valid indicator of judgment accuracy. DaKI scores and processing time were included as additional indicators of decision-making patterns (Gigerenzer & Gaissmaier, 2011).

To identify the optimal number of profiles, models with one to five latent profiles were estimated using the Mclust package. Table 1 summarizes model fit indices including log-likelihood, Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), entropy, average classification probabilities, and the proportion of the smallest class. The three-profile solution showed a favorable balance of model fit and class separation, with all classes containing more than 20% of participants. Although the BIC continued to decrease for four- and five-profile solutions, these models introduced very small classes (< 10%), raising concerns about stability and interpretability (Nylund et al., 2007). Therefore, the three-profile model was retained for further interpretation in line with theoretical considerations and pragmatic model evaluation criteria. Profiles comprised N = 56, N = 60, and N = 144 participants. The model assumed a diagonal covariance matrix with varying volume and shape.

Analysis of processing behavior identified three distinct decision-making profiles, differing in diagnostic quality, processing time, and self-assessed competence (Figure 3). Profile 1 (N = 56) reflected moderate self-assessment, average processing time, and low diagnostic quality. Profile 2 (N = 60) showed high self-assessment, high diagnostic quality, and comparatively long processing time. Notably, 76.67% of this group were SES. Profile 3 (N = 144) displayed moderate to low self-assessment, high diagnostic quality, and short processing time. This profile included two-thirds of all PET, 50% of SET, as well as a share of SES. As shown in Figure 3, Profile 1 shows low quality, and Profile 2 requires more time. Therefore, Profile 3 represents the desired decision-making pattern, combining high diagnostic quality with short processing time.

Table 2

Model fit indices and classification quality for latent profile solutions.

Profiles	Log Likelihood	AIC	BIC	Entropy	Mean Classification Probability	Smallest Class (%)
1	-1,087.38	2,192.76	-2,224.81			100.0
2	-967.57	1,969.13	-2,029.67	0.80	0.95	32.7
3	**-944.56**	**1,929.13**	**-2,000.34**	**0.73**	**0.88**	**21.5**
4	-930.37	1,914.74	-2,010.88	0.74	0.85	10.0
5	-917.93	1,895.86	-2,002.68	0.79	0.85	8.1

Note. Entropy and mean classification probability are only reported for models with G ≥ 2 profiles.

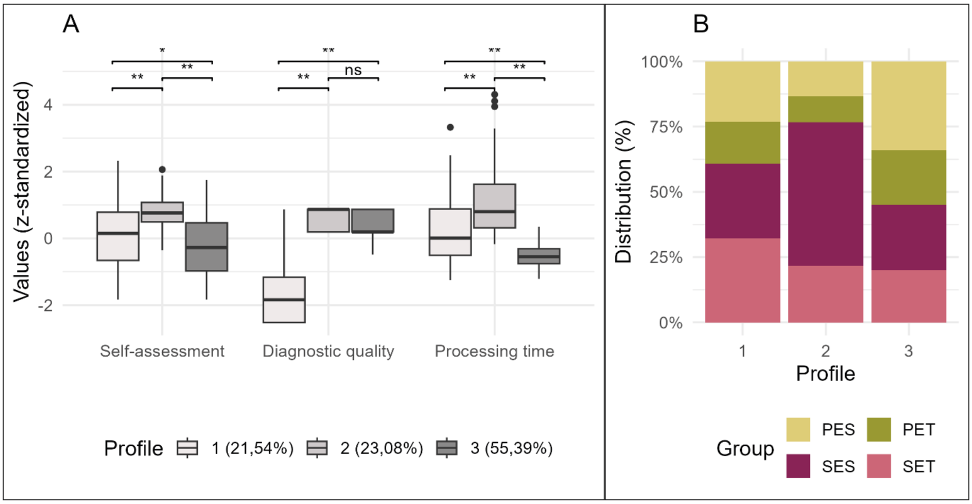


Figure 3

*(A) Differences in self-assessment, diagnostic quality, and processing time across profiles 1 to 3. Significant differences are indicated as follows: ns = not significant, *p < .05, ***p < .001. (B) Distribution of participant groups across profile 1 to 3.*

Age differed significantly between profiles, as shown by an ANOVA: $F(1, 220) = 4.78$, $p = .0298$. However, Bonferroni-corrected t-tests revealed that pairwise differences were marginal or non-significant ($p > .05$). Age was thus not retained as a key distinguishing feature. No significant differences were found regarding years of professional experience ($p > .05$). Chi-square tests revealed significant associations between profiles and sample groups, $\chi^2(6) = 24.65$, $p < .001$. Pairwise comparisons indicated differences between Profiles 2 and 3, and between Profiles 1 and 2. Descriptive analysis showed that two-thirds of PET were in Profile 3, with the remainder distributed across Profiles 1 and 2. PES were mainly in Profile 3 (70%), with 18.57% in Profile 1. SES were equally represented in Profiles 2 and 3, with only 18.82% in Profile 1. SET were predominantly in Profile 3 (50%), with the remainder evenly distributed across Profiles 1 and 2. The educational focus (special vs. primary education) was also significantly related to profile membership, $\chi^2(2) = 17.77$, $p < .001$. Pairwise comparisons showed this effect was driven by differences between Profiles 2 and 3. Profile 2 consisted predominantly of SES (76.67%), whereas Profiles 1 and 3 showed a more balanced mix of participants from special and primary education focus. Gender showed no association with profile membership.

For self-assessed diagnostic competence (DaKI score), ANOVA revealed a significant effect of profile membership emerged, $F(2, 257) = 33.29$, $p < .001$. Tukey post hoc tests showed that Profile 2 scored significantly higher than Profile 1 ($p < .001$). Profile 3 reported significantly lower scores than both Profile 1 ($p = .041$) and Profile 2 ($p < .001$). Processing time in case 2 also differed significantly between profiles, $F(2, 257) = 100.6$, $p < .001$. Profile 2 worked significantly slower than Profile 1 ($p < .001$), while Profile 3 was significantly faster than both Profile 1 ($p < .001$) and Profile 2 ($p < .001$). Test selection accuracy in case 2 showed a significant profile effect, $F(2, 257) = 285.8$, $p < .001$. Profile 2 scored significantly higher than Profile 1 ($p < .001$), as did Profile 3 compared to Profile 1 ($p < .001$). The difference between Profile 2 and Profile 3 was not significant ($p = .083$).

Discussion

Effects of Prior Knowledge and Experience on Diagnostic Quality

To address RQ1 and RQ2, we examined whether participants with different theoretical and practical backgrounds (RQ1) differed in their diagnostic performance and whether group-specific differences in decision-making behavior emerged (RQ2). Although it was assumed that special needs teachers achieved slightly better scores than primary school teachers in practice and study due to additional seminars and experience in diagnostics, the results indicated similar diagnostic competences across the different groups, with no significant differences between teacher students and teachers in self-assessment, accuracy, efficiency (test selection), or open responses (interpretation and intervention planning) ($p > .05$). Despite self-rating their competence as moderate to low, over 70% of all participants—regardless of group—demonstrated accurate diagnostic judgment from the outset, selecting the appropriate standardized mathematics or reading test on their first click with no difference between mathematics or reading. The absence of differences between mathematics and reading reflects the SBG design, which assessed diagnostic rather than didactic knowledge. Surprisingly, prior diagnostic knowledge and experience had no effect, despite previous findings linking them to competence (McElvany et al., 2009). However, research is inconsistent (Praetorius et al., 2011), with Ohle et al. (2015) emphasizing self-efficacy and attitudes in shaping diagnostic confidence. Uncertainty about efficiency was present across all groups, as participants gathered extensive information before reaching a decision. Descriptive differences in the number of tests selected by participants in case 1 (SES/SET < $M = 3$; PES/PET > $M = 3$) were observed, but these did not reach statistical significance. In case 2, all groups showed slight improvements in efficiency, although confirmatory clicking persisted. Teachers in the classroom, however, must make fast and accurate decisions. Ideally, participants would have selected the appropriate test with a single click, thereby obtaining all the relevant information needed for the subsequent process. According to Gigerenzer and Gaissmaier (2011), this could be considered an accurate and efficient decision, based on secure heuristics, without needing to complehensively secure the decision. However, it remains debatable whether the case vignette sufficiently reflected a classroom situation that genuinely demands rapid, intuitive decision-making, as the scenario may have implicitly allowed for more deliberate and reflective processing.

SES and SET scored slightly higher in interpretation and intervention planning (SES/SET > M = 3; PES/PET < M = 3), but this pattern was stable across both cases and not statistically significant (p > .05). As with test selection, participants made partially accurate choices in open-ended responses, often with uncertainty and little data support, resulting in few full scores (5 points) and few zero scores (0). These findings challenge the assumption that diagnostic competence is a core strength of special education teacher training. This is particularly noteworthy given that, in inclusive educational settings, diagnostic expertise is primarily attributed to special education professionals (Kluge & Grosche, 2022). As the SBG task standardized situational and structural conditions for all participants, individual factors likely drove performance differences (Lorenz, 2011). Attitudes, particularly negative ones, can exert a strong influence (Glock & Kleen, 2023), potentially explaining why few participants achieved the expected high diagnostic quality. This performance pattern—moderate competence with rare extremes—emerged regardless of prior knowledge or teaching experience. It suggests that SBGs with their unambigous design and descriptions are broadly suitable for diverse user groups and can yield positive learning effects (Caukin et al., 2016). Nevertheless, it is important to critically emphasize that, especially in inclusive settings and within multiprofessional teams, special education teachers are still expected to provide in-depth diagnostic expertise. Therefore, the systematic professionalization of diagnostic competence in special education training remains a key priority.

Prior knowledge and professional experience significantly influenced processing speed (*p* < .001) but had no significant effect on diagnostic quality. Earlier research primarily examined the link between teaching experience and judgment accuracy, finding weak (McElvany et al., 2009) or no clear associations (Wild & Rost, 1995; Praetorius et al., 2011). In the present study, group differences emerged only in processing time, suggesting that prior diagnostic knowledge primarily affected task efficiency. Participants with less diagnostic training (PES and PET) consistently worked faster than those with formal diagnostic preparation (SES and SET). This may indicate that participants with less knowledge relied more heavily on intuitive decision-making, compensating for their lack of diagnostic expertise. SES and SET showed a notable reduction in processing time in case 2, suggesting a learning effect. They may have better understood the task requirements and more efficiently applied their diagnostic knowledge. These findings suggest that prior knowledge and experience shape the efficiency and approach to diagnostic decision-making rather than improving accuracy. However, the observed reduction in processing time for the second case must be interpreted with caution, as it remains unclear whether this reflects more efficient diagnostic reasoning or merely increased familiarity with the simulation format.

Tailered Support for Decision-Making-Profiles in Teacher Education

To address RQ3, we exploratively analyzed participants' processing behavior to identify typical decision-making profiles that reflect different levels of diagnostic competence and indicate data-based training needs. Processing behavior revealed three profiles representing distinct decision-making types, differing in the extent to which participants made data-based, reflective, or intuitive diagnostic decisions. Profile 1 reflected uncertain, inaccurate decision-makers who acted quickly and intuitively. These participants appeared to rely on faulty or underdeveloped heuristics to support their intuitive judgments (Gigerenzer & Goldstein, 2011). Profile 2 represented reflective, meticulous decision-makers, likely drawing on substantial prior knowledge. This type aligns with findings by Ohle et al. (2015), suggesting that teachers with higher self-efficacy invest more time in diagnostic tasks. Heuristics in this group were likely emerging but not yet fully developed, resulting in predominantly analytical, time-intensive decisions.

Profile 3 comprised efficient, accurate decision-makers with prior knowledge and professional experience. Their performance suggests reliance on established, functional heuristics (Reimer et al., 2007). However, their comparatively low self-assessment indicated underlying uncertainty, possibly underestimating their actual competence.

While these profiles are consistent with theoretical models of decision-making (Gigerenzer & Gaissmaier, 2011) and reflect different stages of diagnostic decision-making competence, it remains an open question whether they represent developmental stages. One might hypothesize a progression from intuition based on weak heuristics (Profile 1), to reflective analysis (Profile 2), and ultimately to efficient, expert-like decisions guided by robust heuristics (Profile 3). However, our data do not allow for causal or developmental conclusions. Future research—ideally using longitudinal designs and methods such as Latent Transition Analysis—is needed to examine whether and how individuals transition between these profiles over time. This knowledge can inform teacher education programs, enabling data-based, targeted support for developing diagnostic competence at different proficiency levels.

The identified decision-making profiles from SBGs allow for adaptive data-based training in teacher education at university toward secure diagnostic heuristics and data-driven pedagogical decisions. Profile 3 represents the desirable pattern: They demonstrate efficient and accurate diagnostic decision-making, likely supported by well-established heuristics, and are expected to thrive in school practice as well. Current diagnostic training at university fits their needs, while targeted, interactive practice (Zellner et al., 2024b) can further strengthen their procedural confidence. Profile 1 reflects a known risk profile in schools, which is also reported by research: uncertain, low-performing teachers who rely on faulty heuristics and show limited diagnostic understanding. Prior studies on teacher characteristics influencing diagnostic decisions (Klug et al., 2016; Glock & Kleen, 2023) suggest similar patterns—superficial decisions, often unchecked due to a lack of external review mechanisms in school practice. They require foundational theoretical knowledge, structured case examples (Praetorius & Südkamp, 2017), and strong instructional guidance to establish basic diagnostic routines (Südkamp et al., 2017). While profile 1 is a well known risk profile, profile 2 is known from studies on professionalization, but has not yet been discussed in the field of diagnostics. This group shows high diagnostic accuracy but requires extensive processing time—indicating reflective but insecure learners. Despite their potential, they struggle to apply knowledge fluently and expend considerable cognitive resources. Without targeted support, they risk burnout in fast-paced school environments. Yet they actively seek feedback, engage in reflection, and aim to improve. They benefit from high instructional support, guided practice, and gradual release in increasingly complex settings (Chernikova et al., 2020). However, at university they rarely get it to the extent they require. These risk profiles should be addressed in university and in-school training to provide pre-service and in-service teachers with individualized, competence-oriented support (Gebhardt, 2023).

Implications for Integrating SBG in Research and Teacher Education

Building on the findings from RQ1 to RQ3, implications for the design of SBGs and their integration into future teacher education and diagnostic research will be discussed. Findings indicate that targeted practice, combined with theoretical study, can support improvements in diagnostic efficiency and accuracy. Although we did not explicitly measure improvements in a pre-post follow-up design, the observed processing behaviors-such as reduced decision time and more targeted test selection in the second case-suggest that repeated exposure to the SBG may help participants internalize diagnostic routines. These changes can be interpreted as signs of developing diagnostic competence. Participant profiles further suggest that a gradual increase in case complexity and interaction demands supports the transfer of theoretical knowledge into school practice. Future SBG design should incorporate progressively complex cases and incrementally expand interactive elements. This scaffolding approach may better prepare educators for the nuanced demands of real-world diagnostics. A promising development would be the inclusion of ambiguous, less clear-cut cases, requiring more sophisticated diagnostic reasoning. Additionally, integrating open-ended conversational simulations with virtual avatars—offering less structured, dynamic dialogue paths—could strengthen diagnostic flexibility and decision-making under realistic, uncertain conditions (Zellner et al., 2024b).

SBG is an innovative educational approach that incorporates strategic problem-solving knowledge in a resource-efficient manner and promotes diagnostic decision-making through structured, data-driven exercises. SBGs can be seen as a valuable building block for simulated learning in a secure and structured environment, allowing learners to practice without real-world consequences. However, given the novelty of this format, more in-depth analysis using a control group design is needed to more rigorously evaluate its effectiveness. The case-based learning approach exists for a long time, but has not gained widespread acceptance as it makes high cognitive demands. However, advances in digital technology offer new opportunities for tailoring interactive learning environments to the needs of students and teachers. This approach bridges research and differentiated instruction by using click data to assess individual processing behavior, facilitating data-driven exercises that develop students' diagnostic decision-making competences. This is particularly relevant for inclusive education, as there is still an urgent need for reliable, data-supported diagnostic decisions.

References

Badiee, F., & Kaufman, D. (2015). Design Evaluation of a Simulation for Teacher Education. *Sage Open*, *5*(2). https://doi.org/10.1177/2158244015592454

Brandt, J. (2022). *Diagnose und Förderung erlernen. Untersuchung zu Akzeptanz und Kompetenzen in einer universitären Großveranstaltung*. Springer.

Brown, J. S., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. Educational Researcher, 18(1), 32–42. https://doi.org/10.3102/0013189X018001032

Brunner, M., Anders, Y., Hachfeld, A., & Krauss, S. (2011). 10 Diagnostische Fähigkeiten von Mathematiklehrkräften. *Professionelle Kompetenz von Lehrkräften: Ergebnisse des Forschungsprogramms COACTIV*, 215.

Caukin, N., Dillard, H., & Goodin, T. (2016). A Problem-Based Learning Approach to Teacher Training: Findings after Program Redesign. *SRATE Journal*, 25(2), 26–32.

Chernikova, O., Heitzmann, N., Stadler, M., Holzberger, D., Seidel, T., & Fischer, F. (2020). Simulation-based learning in higher education: A meta-analysis. Review of Educational Research, 90(4), 499–541. https://doi.org/10.3102/0034654320933544

Döring, N., & Bortz, J. (2016). *Forschungsmethoden und Evaluation* (Vol. 5). Springer.

Ebenbeck, N., Jungjohann, J., & Gebhardt, M. (2022). *Testbeschreibung des Lesescreenings LES-IN für dritte inklusive Klassen. Version 1*. Universität Regensburg. https://epub.uni-regensburg.de/53204/

Fuchs, D., Fuchs, L. S., & Compton, D. L. (2012). Smart RTI: A next-generation approach to multilevel prevention. *Exceptional children*, *78*(3), 263–279. https://doi.org/10.1177/001440291207800301

Gebhardt, M. (2023). *Pädagogische Diagnostik. Leistung, Kompetenz und Entwicklung messen, bewerten und interpretieren für individuelle Förderung.* (Version 0.3). Ludwig-Maximilians-Universität München.

Gigerenzer, G., & Goldstein, D. G. (2011). The recognition heuristic: A decade of research. *Judgment and Decision Making*, *6*(1), 100–121.

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. Annual Review of Psychology, 62, 451–482. https://doi.org/10.1146/annurev-psych-120709-145346

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. Psychological Review, 103(4), 650–669. https://doi.org/10.1037/0033-295X.103.4.650

Glock, S., & Kleen, H. (2023). The role of preservice teachers´ implicit attitudes and causal attributions: a deeper look into students´ ethnicity. *Current Psychology*, *42*(10), 8125–8135. https://doi.org/10.1007/s12144-021-01992-4

Heitzmann, N., Seidel, T., Opitz, A., Hetmanek, A., Wecker, C., Fischer, M., ... & Fischer, F. (2019). Facilitating diagnostic competences in simulations: A conceptual framework and a research agenda for medical and teacher education. *Frontline Learning Research*, *7*(4), 1–24. https://doi.org/10.14786/flr.v7i4.492

Herppich, S., Praetorius, A. K., Hetmanek, A., Glogger-Frey, I., Ufer, S., Leutner, D., … Südkamp, A. (2017). Diagnostische Kompetenz von Lehrkräften: Theoretische und methodische Weiterentwicklungen. In A. Südkamp & A. K. Praetorius (Hrsg.), Diagnostische Kompetenz von Lehrkräften (S. 75–93). Waxmann.

Jungjohann, J., & Gebhardt, M. (2023). *Questionnaire on teachers' diagnostic competence related to classroom-based assessment in inclusive schools (DaCI)-English Translation of the Version 0.2*. Universität Regensburg.

Klauer, K. J. (1985). Framework for a theory of teaching. Teaching and Teacher Education, 1(1), 5–17. https://doi.org/10.1016/0742-051X(85)90003-2

Kluge, J. & Grosche, M. (2022). Aufgaben sonderpädagogischer Lehrkräfte in inklusiven Schulen - Ein Vergleich der Aufgabenwahrnehmung und Wünsche von Regelschul- und sonderpädagogischen Lehrkräften. In S. Fränkel, M. Grünke, T. Hennemann, D. C. Hövel, C. Melzer, & K. Ziemen (Hrsg.), *Teilhabe in allen Lebensbereichen? Ein Blick zurück und nach vorn* (S. 122–129). Verlag Julius Klinkhardt.

Kuhl, U., Sobotta, S., Legascreen Consortium, & Skeide, M. A. (2021). Mathematical learning deficits originate in early childhood from atypical development of a frontoparietal brain network. *PLoS Biology*, *19*(9), e3001407. https://doi.org/10.1371/journal.pbio.3001407

Kuhn, M., & Wickham, H. (2020). Tidymodels: A collection of packages for modeling and machine learning using tidyverse principles [R package]. https://www.tidymodels.org

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge University Press.

Leiner, D. J. (2024). *SoSci Survey* (Version 3.5.02) [Computer software]. https://www.soscisurvey.de

Lorenz, C. (2011). *Diagnostische Kompetenz von Grundschullehrkräften: Strukturelle Aspekte und Bedingungen*. University of Bamberg Press. https://doi.org/10.20378/irb-3956

McElvany, N., Schroeder, S., Hachfeld, A., Baumert, J., Richter, T., Schnotz, W., ... & Ullrich, M. (2009). Diagnostische Fähigkeiten von Lehrkräften: bei der Einschätzung von Schülerleistungen und Aufgabenschwierigkeiten bei Lernmedien mit instruktionalen Bildern. *Zeitschrift für pädagogische Psychologie*, *23*(34), 223–235. https://doi.org/10.1024/1010-0652.23.34.223

McLean, P. (2016). *Culture in networks*. John Wiley & Sons.

Norman, G. (2005). Research in clinical reasoning: past history and current trends. *Medical education, 39*(4), 418–427. https://doi.org/10.1111/j.1365-2929.2005.02127.x

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling, 14*, 535–569. https://doi.org/10.1080/10705510701575396

Ohle, A., Boone, W. J., & Fischer, H. E. (2015). Investigating the impact of teachers` physics ck on students outcomes. *International Journal of Science and Mathematics Education*, 13, 1211-1233. https://doi.org/10.1007/s10763-014-9547-0

Praetorius, A.-K., Karst, K., Dickhäuser, O., & Lipowsky, F. (2011). Wie gut schätzen Lehrer die Fähigkeitsselbstkonzepte ihrer Schüler ein? Zur diagnostischen Kompetenz von Lehrkräften. *Psychologie in Erziehung und Unterricht*, *58*(2), 81–91. https://doi.org/10.2378/peu2010.art30d

Prilop, C. N., Weber, K. E., & Kleinknecht, M. (2020). Effects of digital video-based feedback environments on pre-service teachers` feedback competence. *Computers in Human Behavior*, 102, 120–131. https://doi.org/10.1016/j.chb.2019.08.011

Reimer, T., Hoffrage, U., & Katsikopoulos, K. (2007). Entscheidungsheuristiken in Gruppen. *NeuroPsychoEconomics*, *2*(1), 7–29.

Scrucca, L., Fraley, C., Murphy, T., Raftery, A. (2023). *Model-Based Clustering, Classification, and Density Estimation Using mclust in R*. Chapman and Hall/CRC. https://doi.org/10.1201/9781003277965

Stecker, P. M., Fuchs, L. S., & Fuchs, D. (2005). Using curriculum based measurement to improve student achievement: Review of research. *Psychology in the Schools*, *42*(8), 795–819. https://doi.org/10.1002/pits.20113

Südkamp, A., & Praetorius, A. K. (Eds.). (2017*)*. *Diagnostische Kompetenz von Lehrkräften: theoretische und methodische Weiterentwicklungen*. Waxmann Verlag.

Südkamp, A., Kaiser, J., & Möller, J. (2012). Accuracy of teachers` judgments of students` academic achievement: a meta-analysis. *Journal of Educational Psychology*, *104*(3), 743–762. https://doi.org/10.1037/a0027627

Südkamp, A., Kaiser, J., & Möller, J. (2017). Ein heuristisches Modell der Akkuratheit diagnostischer Urteile von Lehrkräften. In A. Südkamp & A.-K. Praetorius (Hrsg.), *Diagnostische Kompetenz von Lehrkräften* (S. 33–38). Waxmann.

Tversky, A., & Kahneman, D. (1986). The framing of decisions and the evaluation of prospects. In Studies in Logic and the Foundations of Mathematics (Vol. 114, pp. 503–520). *Elsevier.*

Wei, T., & Simko, V. (2024). corrplot: Visualization of a correlation matrix (Version 0.95) [R package]. https://github.com/taiyun/corrplot

Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2023). dplyr: A grammar of data manipulation (Version 1.1.4) [R package]. https://dplyr.tidyverse.org

Wild, K. P., & Rost, D. H. (1995). Klassengröße und Genauigkeit von Schülerbeurteilungen. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, *27*(1), 78–90.

Zellner, J., Ebenbeck, N., & Gebhardt, M. (2024a). Entwicklung digitaler Simulationsspiele mit integrierten Entscheidungsbäumen zur Förderung der diagnostischen Entscheidungskompetenzen in der sonderpädagogischen Lehrkräfteausbildung. *QfI-Qualifizierung für Inklusion*, *6*(2). https://doi.org/10.25656/01:33042

Zellner, J., Koch, J., & Gebhardt, M. (2024b) Schulung der diagnostischen Kompetenz in diskursiven Gesprächen mit einem GPT-Avatar. [Tagungspräsentation] *Herbsttagung der Arbeitsgruppe Empirische Sonderpädagogische Forschung 2024 an der Ludwig-Maximilians-Universität München*, 11.10.2024. http://dx.doi.org/10.13140/RG.2.2.36015.14240

Autorinnen- und Autorenhinweis

Judith Zellner
https://orcid.org/0009-0006-5113-9872

Markus Gebhardt
https://orcid.org/0000-0002-9122-0556

Korrespondenzadresse

Judith Zellner
Universität München
Department für Pädagogik und
Rehabilitation
Lehrstuhl für Sonderpädagogik
Leopoldstraße 13, D-80802 München
judith.zellner@edu.lmu.de

Erstmals eingereicht: 03.03.2025
Überarbeitung eingereicht: 16.05.2025
Angenommen: 26.05.2025

Offene Daten	Datenfiles, Kodierungshinweise, Link zum Fragebogen und der zur Analyse verwendete R Code ist verfügbar unter: http://bit.ly/4idAnan
Offener Code	Datenfiles, Kodierungshinweise, Link zum Fragebogen und der zur Analyse verwendete R Code ist verfügbar unter: http://bit.ly/4idAnan
Offene Materialien	Datenfiles, Kodierungshinweise, Link zum Fragebogen und der zur Analyse verwendete R Code ist verfügbar unter: http://bit.ly/4idAnan
Präregistrierung	NA
Votum Ethikkommission	NA
Finanzielle und weitere sachliche Unterstützung	Das Projekt wurde aus Eigenmitteln des Lehrstuhls finanziert.
Autorenschaft	JZ und MG haben die Studie geplant. JZ hat die Daten erhoben, analysiert und das Manuskript geschrieben, MG hat die Supervision des Beitrags übernommen
KI und generative Modelle	Der R Code wurde mithilfe von ChatGPT gekürzt und optimiert. Zur Übersetzung der deutschen Zusammenfassung des Manuskripts ins Englische wurde DeepL genutzt.