Evaluating the Benefits of Dynamic Testing of Arithmetic Skills for Developing Individualized Education Plans

Taina Gabriel,[1] Claudia Mähler,[2] Jürgen Wilbert,[3] and Moritz Börnert-Ringleb[1]
[1]Institute of Special Education, Leibniz University Hannover, [2]Institute for Psychology, University of Hildesheim, [3]Institute of Education, University of Münster

Abstract

Dynamic testing (DT) is a testing approach that allows tailoring instructions to students' needs. Previous research on DT has been mostly limited to studying its predictive validity in general domains. This study aimed to provide insights into the benefits of DT for planning individualized educational support. To that end, we developed a dynamic test of arithmetic skills for third graders with low math achievement. Math teachers were assigned to three experimental conditions in which they administered DT, a standard test of arithmetic skills, or no test and were asked to write individualized education plans (IEPs) for their students afterwards. A total of 99 IEPs were analyzed to discern patterns and variations across the conditions. Findings showed only a limited benefit of DT.

Keywords: Dynamic testing, individualized education plan, arithmetic skills, diagnostics, special educational needs

Introduction

Implementation of inclusive educational programs involves an emphasis on individualized learning and teaching (Lindner & Schwab, 2020). As a result, teachers are faced with the challenge of having to assess the initial learning status of participating students (Tiekstra et al., 2016). This process requires not only assessment competence (e.g., Herppich et al., 2018) but also tools and approaches that support teachers' decisions and provide valid information on the individual needs of learners. However, to date, there is no gold standard for diagnostic practice in inclusive schools, so bridging the gap between assessment and practice remains a central challenge (Bosma & Resing, 2008; Pameijer, 2006; Tiekstra et al., 2016).

Test-based approaches are repeatedly criticized for being of little use in individualized education planning as they provide little information on what is needed to accommodate students' potential (e.g., Resing, Elliott, et al., 2012; Tzuriel, 2000a; Tzuriel & Universin, 2001). Furthermore, concerns have been raised regarding the validity of outcomes obtained through static diagnostic tests (SDTs) for minority groups (Hessels, 1997; Tiekstra et al., 2009).

Dynamic Testing: A Promising Alternative?

To overcome these limitations of static diagnostic tests, dynamic testing (DT) is frequently proposed as an alternative as it allows more insight into problem-solving processes and enables teachers to derive recommendations for educational practices (Bosma & Resing, 2008, 2012; Haywood & Lidz, 2007; Tiekstra et al., 2016). For example, in contrast to static tests, DT includes a learning or training phase, where the examiner is asked to provide support to the student by teaching or demonstrating strategies for how to solve the tasks included in the test (Tiekstra et al., 2016). The student's response to this support is then regarded as crucial diagnostic information (Resing, Elliott, et al., 2012).

Gabriel, T., Mähler, C., Wilbert, J., & Börnert-Ringleb, M. (2026). Evaluating the Benefits of Dynamic Testing of Arithmetic Skills for Developing Individualized Education Plans. *International Journal for Research in Learning Disabilities, 9*(1), XX. https://doi.org/10.28987/ijrld.9.1.XX

Two facets of DT seem to be particularly relevant for designing individualized support. First, by providing guided instruction in the testing process, DT aims to identify children's learning potential (Sternberg & Grigorenko, 2002). Second, the insights gained during the instruction phase can serve as relevant information about the conditions under which children's learning potential can best be accommodated (Bosma et al., 2017; Bosma & Resing, 2008, 2012; Tiekstra et al., 2016).

A common variant of providing instruction during DT is the *graduated prompt approach* (e.g., Campione & Brown, 1978; Resing et al., 2009), whereby students receive a prestructured sequence of prompts (i.e., hints) during the test administration (Bosma et al., 2017; Resing, Stevenson, et al., 2012; Veerbeek et al., 2017). As soon as the child fails to solve a task, an initial prompt is given. If the child cannot provide the correct answer, the next prompt is given. This interaction is repeated until the task is solved or the final prompt (often modelling of the task) is presented. The predetermined sequence of prompts reflects a theoretical model of the problem-solving processes involved in solving the task. Additionally, emotional or motivational support (e.g., encouraging or reassuring the child) can be provided; however, research on DT shows that this is rarely done (Tiekstra et al., 2016).

Research on the use of DT in educational fields is limited. Nevertheless, the existing evidence on the validity of DT is positive, especially for predicting the achievement of students with disabilities (Caffrey et al., 2008). Even less research has addressed the potential of DT to help with individualized education planning. Bosma and colleagues (2012) examined teachers' preference for information regarding educational planning and compared information based on DT (e.g., learning processes) with other standard diagnostic information, such as a child's diagnosis or their achievement compared to that of peers. The authors highlighted teachers' positive appraisal of DT information, especially among more experienced teachers.

In another study, Bosma and Resing (2012) conducted DT of the analogical reasoning of 36 students with intellectual disabilities and compared teachers' evaluations of DT-based reports to static assessment results. Teachers appraised DT outcomes as more valuable because they seemed to be more useful in practice. Similar results were described by Deutsch and Reynolds (2000). In their study, participating educational psychologists stated that DT is a way to provide information on practical steps as well as potential learning barriers.

To date, no study has examined the extent to which the benefits of using DT are reflected in teachers' development of individualized education plans (IEPs). Instead, previous research has focused on DT in domain-general skills, such as analogical reasoning or intelligence (e.g., Resing, Stevenson, et al., 2012; Sternberg et al., 2002), which are distal to the subject-specific instruction in schools. Domain-specific DT might be more appropriate when it comes to developing subject-specific recommendations and has been found to improve "the efficacy of external validity in assessment-based decisions" (Kaniel, 2010, p. 104), especially for students at risk for learning disabilities (e.g., Dixon et al., 2023).

Applications of domain-specific DT are still rare, and existing examples have primarily addressed reading (Dixon et al., 2023; Dörfler et al., 2009) and math skills (Bosma et al., 2017; Fuchs et al., 2008; Kaskens et al., 2021, 2023; Tzuriel, 2000b). Findings show the advantages of DT compared to static testing for predicting learning development and success (Caffrey et al., 2008; Cho et al., 2014; D. Fuchs et al., 2011). For example, implementation of DT of algebraic learning in a two-stage screening for math problem-solving difficulty reduced the identification of false positives (L. S. Fuchs et al., 2011). Fuchs et al. (2008) noted that DT "might be used productively within an RTI framework to help identify students who will ultimately, 10–30 weeks later, prove unresponsive to secondary prevention" and argued that these "chronically unresponsive students are considered to have a learning disability" (p. 847). Despite the existing evidence on predictive validity, insight into the potential usefulness of domain-specific DT for educational planning in students at risk for learning disabilities is lacking.

Use of Dynamic Testing in Math

Starting school, children vary greatly in their mathematical abilities (Bodovski & Farkas, 2007), and arithmetic difficulties are a common and early phenomenon in schools. Specifically, Moll et al. (2014) and Morsanyi et al. (2018) reported that around 13% of primary-school children in their studies showed below-average math performance. Further, approximately 6% of students met the criteria for specific learning disorders in mathematics. Statistically, therefore, most teachers will encounter students in their classrooms who persistently struggle in math and are consequently faced with the challenge of having to provide adequate support for those students (Scherer et al., 2017). This is es-

pecially important as arithmetic skills are related to variables of short- and long-term development, and without appropriate support, students' differences in mathematical abilities will persist or even increase (Aunola et al., 2004; Navarro et al., 2012). Thus, recognizing and countering emerging difficulties early is important and requires appropriate tools. This is not only essential to foster successful learning, but also to reduce the risk of these difficulties developing into long-term underachievement or even learning disabilities (Lange & Thompson, 2006). Here, DT might be particularly supportive.

Most school-based assessments of arithmetic skills consist of static tests that primarily measure arithmetic operations or numerical precursor skills. However, arithmetic skills are also based on higher-order cognitive aspects such as working memory and executive processes, as well as specific mathematical knowledge (Dowker, 2008; Haberstroh & Schulte-Körne, 2022; Kaufmann et al., 2013; Maehler & Schuchardt, 2011). Additionally, mathematical difficulties are often accompanied by emotional and motivational responses that interfere with mathematical problem-solving and engagement, further hindering mathematical learning processes (Dowker et al., 2016; Schukajlow et al., 2023). Taking these issues into account seems to be crucial for providing adequate support.

DT might offer a way to integrate these aspects into a single diagnostic approach. So far, solely Kaskens et al. (2021) have examined the potential usefulness of math-related DT for identifying educational needs. Indeed, they found that a variation of DT (dynamic math interviews) facilitated 19 fourth-grade teachers' identification and understanding of the educational needs of students with low math achievement. However, the study did not examine whether DT improved the quality of educational planning. Additionally, Kaskens et al. (2021) did not have a control condition. Thus, it remains unclear whether the effects described were specific to the dynamic math interviews.

Research Questions

As mentioned, DT is repeatedly discussed as an alternative to SDTs as it promises to provide more information that support instructional decision-making (Bosma & Resing, 2008, 2012; Haywood & Lidz, 2007). Despite these promises, research on its advantages is limited, especially concerning benefits for educational planning of domain-specific DT in math-related areas and for students at risk of developing mathematical learning disabilities. DT might improve educational planning and result in distinct and more differentiated IEPs as it enables the integration of higher-order cognitive as well as emotional-motivational aspects that contribute to problem-solving. This advantage should be particularly apparent compared to IEPs that are based on SDTs. Therefore, this study explored the following research questions:

> RQ1: Does the application of DT lead to more differentiated IEPs than the application of a static diagnostic test (SDT) or teachers' individual diagnostic routines?

As DT allows for interactions between child and test administrator, it supports the identification of multiple processes, which might help teachers to gain a more nuanced image of children's learning. Therefore, it was expected that IEPs from teachers who used a dynamic test would contain a greater variety of processes than IEPs from teachers who used their own standard diagnostic routines. Furthermore, we expected that IEPs that were based on DT would contain more processes compared to IEPs that were based on a SDT.

One possible advantage of DT over the use of a SDT is that (higher-order) cognitive as well as emotional and motivational prompts can be included in the test, aiming at additional relevant processes in context of arithmetic (e.g., working memory or emotion regulation). Hence, test administrators could get a deeper understanding of these additional processes when applying DT and use these insights for educational planning. Therefore, we additionally aimed to examine differences between IEPs with regard to higher-order cognitive as well as motivational and emotional processes and answer the second research question:

> RQ2: Does the application of DT lead to more frequent descriptions of higher-order cognitive as well as motivational and emotional processes in IEPs than the application of a SDT or teachers' standard diagnostic routines?

We expected to find more motivational, emotional as well as higher-order cognitive aspects in IEPs written by teachers who used DT compared to SDTs and teachers' standard diagnostic routines.

Kaskens et al. (2021) found that teachers were able to identify students' need for support by using dynamic math interviews. However, as their study did not include a control condition, it remains unclear whether the participating teachers would be able to derive similar information by using other diagnostic approaches such as their own standard diagnostic routines or SDTs. To gain insight into this

issue, we explored not only whether different diagnostic approaches differed in the quantity and quality of processes mentioned, but also whether the IEPs differed in the level of identified need for support of students. Thereby, we aimed to answer the third research question:

> RQ3: Does the application of DT lead to more frequent descriptions of students' need for support in IEPs than the application of SDT or teachers' standard diagnostic routines?

Method

To address the three research questions, we conducted the following experiment. Teachers were randomly assigned to two experimental conditions, where they administered a dynamic test (DT group; DT-G) or a static diagnostic test (SDT group; SDT-G) of arithmetic skills with a sample of students with low math achievement. In addition, in a control condition (CG), teachers did not administer a predetermined test but followed their standard diagnostic routines. Afterwards, all teachers wrote IEPs for all of their students.

Participants

Thirty-seven 3rd-grade math teachers participated (DT-G: n = 18, SDT-G: n = 10, CG: n = 9) in the study. Initially, 50 teachers agreed to take part, but during data collection, 13 (26%) dropped out, leaving 37 to complete the study. The majority (78.3%) of the remaining teachers were between 35 and 54 years old. Their work experience varied between less than 5 years and more than 25 years (*Mdn* = 10 to 14 years). Most of them (73%) had earned a degree as primary school teacher and 56.8% were trained to be a math teacher.

Teachers were asked to develop IEPs for three students with low math achievement in their classes. Some teachers submitted fewer than three plans, resulting in 99 IEPs focused on students (M_{age} = 8.91 years, SD = 0.63; 64.2% girls) with low math achievement (T value: M = 35.89; SD = 8.09). The selection process of the children is described below.

Procedure

The study was approved by local school authorities as well as the ethical board of the Faculty of Educational and Social Sciences of the University of Hildesheim. The data collection took place at the end of 2022 and in the first half of 2023.

In the first step, third-grade math teachers were recruited via extensive outreach to schools across three regions in Germany. Next, all teachers who agreed to participate were asked to identify up to five students with low math achievement in their classes. After written informed consent was obtained from their parents, these students were screened using a curriculum-based math achievement test (German Math Test for Grade Two; DEMAT 2+; Krajewski et al., 2020). In each classroom, the three students with the lowest math achievement scores were included in the study.

In the second step, teachers were randomly assigned to one of three conditions, as described below. All teachers completed a two-hour training on basics and concepts of arithmetic development. Afterwards, teachers in the dynamic testing condition (DT-G) were trained to apply a newly developed dynamic test of arithmetic skills (DTR-ASM; DYNAMIK Project Team, 2022). Teachers in the static diagnostic test condition (SDT-G) were trained in applying the Heidelberger Calculation Test (HRT 1-4; Haffner et al., 2005), a SDT for arithmetic skills. All trainings were conducted online and synchronous by trained researchers. Finally, the teachers in the control condition (CG) did not receive a second training but were asked to apply their usual standard diagnostic routines.

After training, teachers were asked to apply the trained diagnostic approach with their identified students and to write an IEP for each of them using a prestructured document.

Instruments

Dynamic Test of Arithmetic Skills

The DTR-ASM (DYNAMIK Project Team, 2022) comprises three subscales (addition, subtraction, and multiplication), each with six tasks. The addition and subtraction subscales further distinguish between different levels of task difficulty (in-/excluding carrying tens as well as varying number range up to 20 or 100). The structure of the test follows the graduated-prompt approach (Campione & Brown, 1987) and was predetermined in a standardized protocol (see Figure 1). This protocol comprises higher-order cognitive (i.e., working memory and metacognition) as well as arithmetic (calculation and counting strategies) and supportive prompts addressing emotions and motivation. The prompts were derived from a task analysis and a review of the literature. In total, the students could receive a maximum of eight (meta-) cognitive prompts per task, which became increasingly explicit. Additionally, teachers could give an unlimited amount of emotional and motivation-

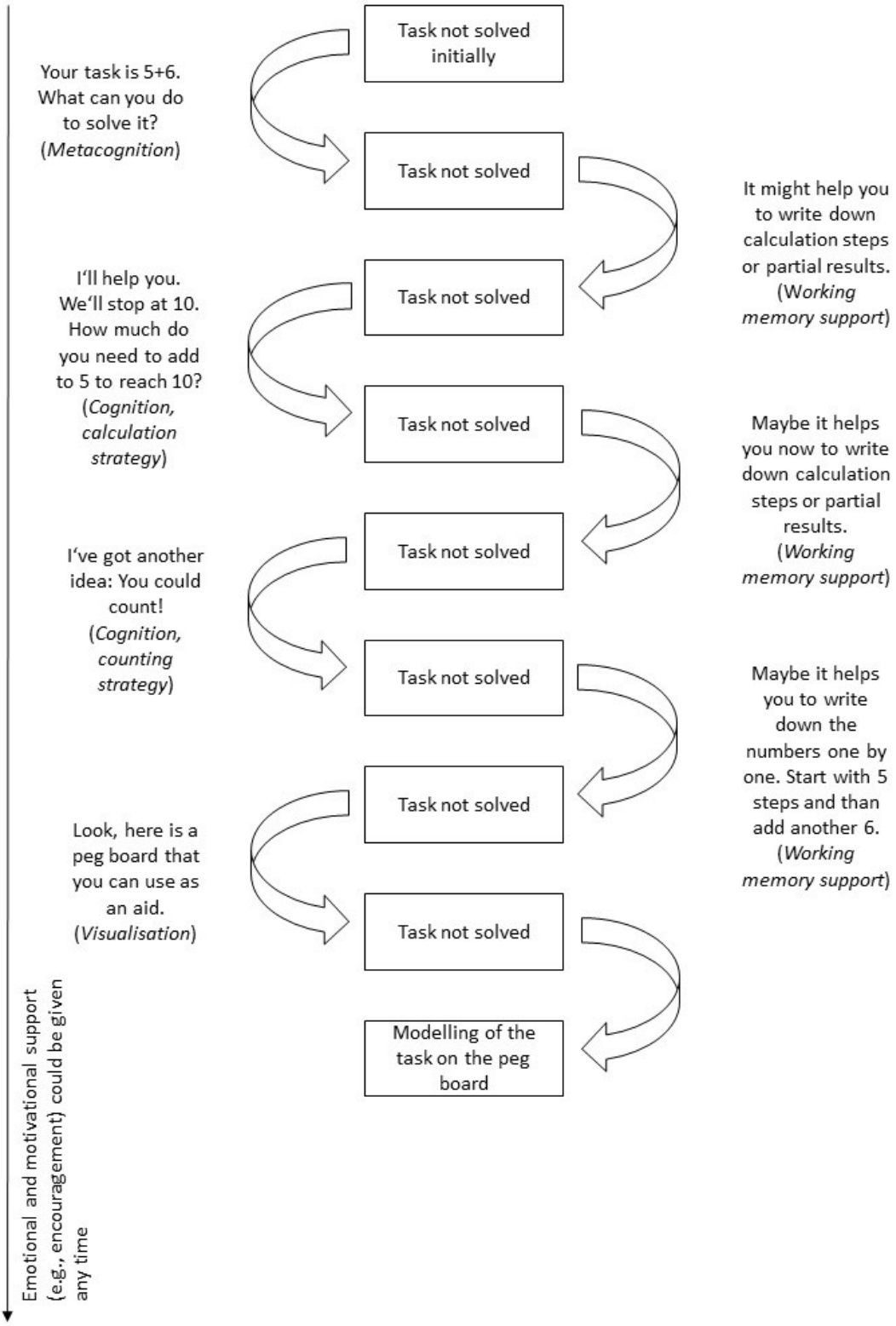

Figure 1
Nature and Structure of Prompts Included in the Dynamic Test

al support (e.g., "It's okay if you don't solve the task straight away." "Take a deep breath and try again.").

HRT 1-4

Teachers in the SDT condition administered the arithmetic skills scales of the Heidelberger Calculation Test (HRT 1-4; Haffner et al., 2005). Adequate reliability (r_{tt} = .69–.89) and validity of the HRT 1-4 was confirmed by the test authors.

Individualized Education Plans

Teachers were asked to fill in a prestructured IEP that consisted of three sections, as follows. They had to describe the child's *developmental status* (e.g., nature and extent of mathematical competencies) before formulating *learning goals* for the current school year and planning the *remediation*. A prestructured table was used for remediation planning. Here teachers had to specify which competencies they wanted to address, what the goal of the specific intervention was, how they wanted to reach it, and why they chose the aforementioned method. The table allowed the inclusion of up to three competencies. In this study, we focused on the information provided in the columns on addressed competencies and goals as they reflect the underlying processes teachers wanted to address. An English translation of the IEP template may be found in the supplement (https://doi.org/10.17605/OSF.IO/YGDC3).

Data Preparation and Coding. The IEPs were coded following the steps of qualitative content analysis (Mayring, 2015). That is, a coding scheme was developed inductively by the researchers, to involve a nuanced classification, distinguishing between arithmetic skills, higher-order cognitive processes, and motivational-emotional variables (see Supplement for a detailed description of all categories; https://doi.org/10.17605/OSF.IO/YGDC3). To support the coding process, a description of the categories and respective text examples from the IEPs were provided as training material. Coders were allowed to include additional subcategories during the development of the coding manual.

To assess interrater agreement, 20% of the IEPs were coded by three raters until a reasonable consensus was reached. Interrater agreement was substantial for the reliability coding (Fleiss K = .648). Finally, a third of the remaining IEPs were randomly assigned to each rater to complete the coding process. Coders were given only the text of the IEP so that they were blind to any other aspect such as school district, condition, or children's gender. MAXQDA 2022 (VERBI Software, 2021) was used for the coding process and data extraction.

Analyses

In order to answer the first research question (RQ1) and to analyze whether diagnostic condition (DT-G; ST-G; CG) led to differing levels of differentiation in the IEPs, binomial regressions were calculated. Here, the total number of different categories assigned during coding was included as the dependent variable (DV). We chose to use binomial regressions as the DV can be regarded as the proportion of mentioned categories out of a predefined (maximum) number and hence follows a binomial distribution. As the main interest of the study was the effects of DT, the results of the DT-G were included as the intercept and compared to the other two groups.

Similar analyses were applied to answer the second and third research questions. For RQ2, the number of higher-order cognitive processes as well as motivational and emotional processes were included as DV in two separate regressions. Concerning RQ3, the need for support was included as DV. Data analysis was conducted using *R* version 4.4 (R Core Team, 2024) with the psych (Revelle, 2025) and companion (Mangiafico, 2025) packages.

Results

On a descriptive level, differences in the total frequencies of processes can be described. An overview of all descriptive statistics may be found in Tables 1 and 2.

Total Number of Categories

Binomial regressions were used to investigate whether the educational plans differed across conditions with regard to the total number of categories mentioned (see Table 3) and to address RQ1. The number of different categories (relative to the maximum number of categories) was included as DV and the condition as predictor. No differences in the total number of processes mentioned were found between the conditions, regardless of whether the educational plans were considered as a whole or if the various sections of the plans were investigated individually.

Arithmetic Skills

To gain more detailed information, we further analyzed the number of categories mentioned within each dimension. Here, arithmetic skills were a crucial area of interest as the IEPs were written for students with low math achievement. Differences between DT-G and SDT-G became clear in the description of the developmental status (B = -0.40, p =

Table 1
Descriptive Statistics of the Number of Categories in the IEPs

Section of the IEP	Average Number of Categories (*SD*) DT-G	SDT-G	CG
Total	12.53 (5.24)	13.15 (3.87)	13.10 (6.06)
Description of Developmental Status	6.07 (2.88)	5.59 (2.08)	7.10 (3.62)
Description of Learning Goals	3.19 (1.52)	3.96 (1.56)	3.17 (1.73)
Description of Remediation	3.28 (1.74)	3.59 (1.85)	2.83 (1.79)

Table 2
Descriptive Statistics of the Categories

Category	Average Number of Processes (*SD*) DT-G	SDT-G	CG
		Description of Developmental Status	
Arithmetic Skills	3.42 (1.28)	2.63 (1.04)	4.00 (1.77)
Emotional/Motivational Processes	0.63 (1.07)	1.30 (1.20)	1.10 (1.32)
Cognitive Processes	1.23 (1.11)	1.37 (1.01)	1.59 (1.27)
Need for Support	0.79 (0.94)	0.30 (0.47)	0.41 (0.63)
		Description of Learning Goals	
Arithmetic Skills	2.21 (1.08)	2.56 (1.12)	2.10 (1.42)
Emotional/Motivational Processes	0.07 (0.26)	0.30 (0. 61)	0.31 (0.66)
Cognitive Processes	0.74 (0.79)	1.04 (1.02)	0.48 (0.69)
Need for Support	0.16 (0.43)	0.07 (0.27)	0.28 (0.53)
		Description of Remediation	
Arithmetic Skills	2.51 (0.98)	2.37 (1.31)	2.10 (1.54)
Emotional/Motivational Processes	0.16 (0.57)	0.19 (0.56)	0.10 (0.41)
Cognitive Processes	0.49 (0.74)	0.93 (1.04)	0.52 (0.78)
Need for Support	0.12 (0.45)	0.11 (0.32)	0.10 (0.31)

.025). According to the regression results, teachers in DT-G mentioned, on average, 3.42 different arithmetic skills whereas teachers in SDT-G described only 2.63. No differences between conditions were found regarding the learning goals or the planned remediation (see Table 4).

Motivational and Emotional Variables

To answer RQ2, motivational and emotional variables (e.g., description of motivation or emotions) as well as higher-order cognitive variables (e.g., descriptions of memory and attention) were analyzed. For the description of motivational and emotional categories (see Table 5), binomial regressions revealed differences between groups in the description of developmental status; DT-G vs. SDT-G: $B = 0.79$, $p = .003$; DT-G vs. CG: $B = 0.61$, $p = .025$. Teachers in DT-G (predicted mean = 0.63) mentioned significantly fewer categories than teachers from the other groups (predicted mean SDT-G = 1.29; CG = 1.10). The same pattern was found for the definition of learning goals; DT-G vs. SDT-G: $B = 1.47$, $p = .031$; DT-G vs. CG: $B = 1.51$, $p = .024$; with a predicted mean of 0.07 for DT-G, 0.30 for SDT-G and 0.31 for CG. No differences were found for the description of the remediation.

Table 3
Results of the Binomial Regression for the Number of Categories in the IEPs

	B	*SE*	*z*	*p*
		Total Number of Categories		
Intercept[1]	-1.859	0.046	-40.152	<.001***
SDT-G	0.055	0.074	0.752	.452
CG	0.051	0.072	0.713	.476
Nagelkerke's R^2 = .008				
		Description of Developmental Status		
Intercept[1]	-1.413	0.069	-20.468	<.001***
SDT-G	-0.101	0.113	-0.890	.374
CG	0.200	0.105	1.898	.058
Nagelkerke's R^2 = .067				
		Description of Learning Goals		
Intercept[1]	-2.167	0.090	-24.023	<.001***
SDT-G	0.247	0.137	1.796	.073
CG	-0.005	0.142	-0.034	.973
Nagelkerke's R^2 = .039				
		Description of Remediation		
Intercept[1]	-2.135	0.089	-23.969	<.001***
SDT-G	0.103	0.140	0.734	.463
CG	-0.164	0.146	-1.124	.261
Nagelkerke's R^2 = .030				

Note. [1]Refers to DT-G.
***$p < .001$.

Cognitive Variables

Binomial regressions were used to determine differences in the number of higher-order cognitive variables (see Table 6). No differences between groups were found for the description of developmental status or the learning goals. However, in the description of the remediation, teachers from DT-G mentioned significantly fewer processes than teachers in SDT-G; $B = 0.709$, $p = .023$. Specifically, DT-G mentioned, on average, 0.49 processes whereas SDT-G had a predicted mean of 0.93 processes.

Need for Support

To answer RQ3, we investigated whether the groups differed in their description of a need for support (see Table 7). Here, teachers from DT-G described more need for support than the other two groups when describing developmental status; DT-G vs. SDT-G: $B = -1.184$, $p = .005$; DT-G vs. CG: $B = -0.805$, $p = .029$. The predicted mean of DT-G was 0.79 processes whereas the predicted means of SDT-G and CG were 0.30 and 0.41, respectively.

Discussion

This study evaluated the benefits of using a dynamic test of arithmetic skills for planning individualized educational support. To that end, we compared IEPs that were either based on a dynamic test, a static diagnostic test, or teachers' standard diagnostic routines. Considering the results described above, it did not become clear whether DT was superior to other diagnostic approaches when it comes to individualized education planning. That is, contrary to our expectations, IEPs from the three groups did not differ significantly in most of the investigated areas, implying that DT did not yield in advantages compared to the alternative approaches. DT-based IEPs

Table 4
Results of the Binomial Regression for Arithmetic Skills

	B	*SE*	*z*	*p*
		Description of Developmental Status		
Intercept[1]	-0.490	0.105	-4.680	<.001***
SDT-G	-0.395	0.176	-2.246	.0247*
CG	0.267	0.163	1.641	.101
Nagelkerke's R^2 = .124				
		Description of Learning Goals		
Intercept[1]	-1.123	0.118	-9.507	<.001***
SDT-G	0.198	0.185	1.070	.284
CG	-0.065	0.188	-0.343	.731
Nagelkerke's R^2 = .019				
		Description of Remediation		
Intercept[1]	-0.949	0.133	-8.375	<.001***
SDT-G	-0.079	0.185	-0.430	.667
CG	-0.238	0.185	-1.288	.198
Nagelkerke's R^2 = .018				

Note. [1]Refers to DT-G.
*$p < .05$. ***$p < .001$.

Table 5
Results of the Binomial Regression for Motivational and Emotional Processes

	B	*SE*	*z*	*p*
		Description of Developmental Status		
Intercept[1]	-2.90	0.198	-14.652	<.001***
SDT-G	0.785	0.267	2.945	.003**
CG	0.606	0.271	2.237	.025*
Nagelkerke's R^2 =.100				
		Description of Learning Goals		
Intercept[1]	-5.142	0.579	-8.880	<.001***
SDT-G	1.465	0.681	2.153	.031*
CG	1.513	0.670	2.257	.024*
Nagelkerke's R^2 = .109				
		Description of Remediation		
Intercept[1]	-4.287	0.381	-11.264	<.001***
SDT-G	0.131	0.590	0.222	.825
CG	-0.458	0.694	-0.661	.509
Nagelkerke's R^2 = .012				

Note. [1]Refers to DT-G.
*$p < .05$. **$p < .01$. ***$p < .001$.

Table 6
Results of the Binomial Regression for Higher-Order Cognitive Processes

	B	*SE*	*z*	*p*
		Description of Developmental Status		
Intercept[1]	-1.543	0.151	-10.197	<.001***
SDT-G	0.130	0.237	0.548	.584
CG	0.315	0.226	1.397	.162
Nagelkerke's R^2 = .020				
		Description of Learning Goals		
Intercept[1]	-2.129	0.187	-11.385	<.001***
SDT-G	0.380	0.277	1.370	.171
CG	-0.474	0.334	-1.417	.156
Nagelkerke's R^2 = .071				
		Description of Remediation		
Intercept[1]	-2.590	0.226	-11.449	<.001***
SDT-G	0.709	0.312	2.274	.023*
CG	0.062	0.351	0.176	.860
Nagelkerke's R^2 = .066				

Note. [1]Refers to DT-G.
*p < .05. ***p < .001.

Table 7
Results of the Binomial Regression for the Need for Support

	B	*SE*	*z*	*p*
		Description of Developmental Status		
Intercept[1]	-1.028	0.200	-5.142	<.001***
SDT-G	-1.184	0.423	-2.800	.005**
CG	-0.805	0.370	-2.178	.029*
Nagelkerke's R^2 = .119				
		Description of Learning Goals		
Intercept[1]	-2.858	0.389	-7.354	<.001***
SDT-G	-0.818	0.815	-1.004	.315
CG	0.568	0.537	1.057	.290
Nagelkerke's R^2 = .058				
		Description of Remediation		
Intercept[1]	-3.211	0.456	-7.039	<.001***
SDT-G	-0.047	0.744	-0.063	.949
CG	-0.121	0.744	-0.163	.870
Nagelkerke's R^2 = .001				

Note. [1]Refers to DT-G.
*p < .05. **p < .01. ***p < .001.

did not contain more categories than IEPs based on a SDT or teachers' diagnostic routines and cannot be described as more differentiated. Furthermore, based on the results, the answer to RQ2 – whether DT leads to a more nuanced description of (higher-order) cognitive as well as emotional and motivational categories in teachers' IEPs – was negative. In fact, IEPs by teachers that applied DT even seemed to contain slightly fewer descriptions of these processes.

Although DT is often discussed as being of use or even superior to standard static tests when it comes to tailoring instruction to students' needs (Bosma et al., 2017; Bosma & Resing, 2008), these advantages did not become clear in the present study. However, this is not surprising as Bosma and Resing (2008), who observed teachers' behavior in the classroom after receiving reports based on DT or a STD, did not find a clear picture regarding the effect of DT either. The sole significant effect involved task regulation, whilst teachers in both conditions improved after receiving diagnostic reports in all remaining areas. In contrast, Kaskens et al. (2023) described effects of dynamic math interviews on teaching behavior and found improvements in teaching aspects such as differentiation and adaption of lessons or teaching (math) learning strategies.

Tiekstra et al. (2016) discussed the usefulness of DT with regard to its consequential validity, arguing that different levels of consequential validity exist, ranging from consequences within the testing procedure (proximal consequential validity) to implications beyond the testing (distal consequential validity). The present study employed the graduated prompt approach, which adapts the testing procedure to students' responses, and thus had immediate consequences for their learning process. As a result, proximal consequential validity can be assumed. However, if teachers are to use DT results for individualized education planning, establishing distal consequential validity of DT is necessary.

According to Tiekstra et al. (2016), information about the level of mediation required can provide valid insights that help to derive implications for future teaching. In the dynamic test applied in the current study, teachers potentially gained insights into the level of mediation in different ways. They filled in a protocol sheet where they noted all the prompts they had to apply before the student was able to solve a task. Additionally, they rated how much and which kind of motivational and emotional support they gave. In addition to documenting the level of mediation, teachers needed to interpret the information. Before applying the dynamic test, teachers followed a training procedure in which they were taught which prompt relates to which underlying difficulties. However, they still had to apply this knowledge and combine the information in order to derive implications for their educational practice.

In our study, teachers did not receive further guidance on this and thus might have struggled to translate the DT information into educational implications. Accordingly, it is unclear whether the IEPs did not differ because DT failed to provide deeper insights or because teachers were unable to draw conclusions based on the DT results. Similarly, Tiekstra et al. (2016) highlighted that the existing research on DT does not fulfill on its promise of bridging the gap between assessment and instruction as distal consequential validity is not guaranteed.

In sum, the results of the current study indicate that teachers from DT-G more frequently emphasized a need for support when describing students' current learning status. Thus, DT might help teachers to identify students' educational needs, as previously stated by Kaskens et al. (2021). However, those researchers did not have any control conditions, so it remains unclear whether identification of need for support is specific to DT. The current study extends Kaskens et al.'s findings, as a more frequent identification of need for support was found in comparison to static testing and teachers' standard diagnostic routines.

Limitations

The findings of the present study need to be discussed in light of several limitations. First, the IEPs (Level 1) were nested in the teachers (Level 2). Each teacher developed several plans, which potentially influenced the quality and the content, as plans from the same person might be more similar. In addition to the experimental condition, teachers' characteristics might explain differences between IEPs. Unfortunately, the sample size on Level 1 (IEPs) as well as Level 2 (teachers) did not allow for adequate control of the nested data structure as simulation studies for multilevel data suggest that approximately 50 Level 2 units are needed to obtain less biased estimates (Maas & Hox, 2005; Moineddin et al., 2007; Paccagnella, 2011). Furthermore, Moineddin et al. (2007) noted that the sample sizes might have to be even larger if the prevalence of events is low. As some of the categories were scarcely mentioned in the IEPs in the current study, multilevel estimates might lead to biased results. Nevertheless, to gain exploratory insights into the potential stability of the findings under control of the nested data structure, we calcu-

lated multilevel binomial regressions. Although the main results are similar to the simpler models, the results support the hypothesis that the influence of teachers' characteristics on the content of the IEPs is high (see Supplement for detailed results: https://doi.org/10.17605/OSF.IO/YGDC3). This point is especially relevant, as we experienced a high dropout rate among teachers during the data collection. Thus, the remaining teachers may represent a sample of highly motivated teachers, which can reflect a potential bias. Moreover, the group sizes differed strongly due to the unbalanced dropout, meaning that IEPs from individual teachers had a different impact on the group outcome. Furthermore, the results only allow for interpretations concerning the processes that were mentioned in the educational plans.

Second, Kaskens et al. (2021) reported that only 6 out of the 19 teachers in their study conducted dynamic interviews that were considered to be completely adequate, despite receiving an extensive 16-hour training beforehand. This finding emphasizes that DT is more complex and requires appropriate training as it also is an unfamiliar source of information in educational contexts (Freeman & Miller, 2001). In the current study, teachers only received 90 minutes of specific training on DT before applying the DTR-ASM. Even though the training was designed to sufficiently prepare participants for conducting DT, the results of our study need to be interpreted in light of the possibility that the teachers might have needed additional training. This assumption is further stressed in a study by Deutsch and Reynolds (2000), in which educational psychologists stated that to be adequate, multiple days of training on DT are needed. However, we did not control for the quality of the implementation.

Third, the interpretation of findings between the CG and the DT-G is complex as we have no information on the diagnostic measures applied by teachers in the CG. In addition, teachers in all conditions had known their students for up to three years and were consequently able to draw upon a broad knowledge about them. As such, potential advantages of DT might mainly become clear in situations, where teachers are unable to draw on experiences and insights about individual students. (Formalized) Diagnostic actions often (e.g., in Germany) still fall under the purview of special education teachers, who identify special educational needs (Sansour & Bernhard, 2018) and design IEPs. In parts of Germany, special education teachers have to provide services to various schools in a district (e.g., support of classroom teachers, diagnostics) (Reiser et al., 2003). Here, time for diagnostic processes is limited, and teachers might benefit from diagnostic approaches that enable holistic insights in a short time so that the effects of DT might be stronger in such contexts. A conclusive evaluation of the usefulness of DT, however, needs to be based on the extent to which the IEPs adequately relate to students' needs and subsequently lead to effective instruction. Fourth and finally, the dynamic test of arithmetic skills was newly developed. Adequate insights in the reliability and validity of the procedure are sparse. At the same time, the SDT was selected among a plurality of alternative approaches and the application of different tests might have provided different insights.

References

Aunola, K., Leskinen, E., Lerkkanen, M.-K., & Nurmi, J.-E. (2004). Developmental Dynamics of Math Performance From Preschool to Grade 2. *Journal of Educational Psychology, 96*(4), 699–713. https://doi.org/10.1037/0022-0663.96.4.699

Bodovski, K., & Farkas, G. (2007). Mathematics Growth in Early Elementary School: The Roles of Beginning Knowledge, Student Engagement, and Instruction. *The Elementary School Journal, 108*(2), 115–130. https://doi.org/10.1086/525550

Bosma, T., Hessels, M. G. P., & Resing, W. C. M. (2012). Teachers' preferences for educational planning: Dynamic testing, teaching' experience and teachers' sense of efficacy. *Teaching and Teacher Education, 28*(4), 560–567. https://doi.org/10.1016/j.tate.2012.01.007

Bosma, T., & Resing, W. C. M. (2008). Bridging the Gap Between Diagnostic Assessment and Classroom Practice. *Journal of Cognitive Education and Psychology, 7*(2), 174–198. https://doi.org/10.1891/194589508787381854

Bosma, T., & Resing, W. C. M. (2012). Need for instruction: Dynamic testing in special education. *European Journal of Special Needs Education, 27*(1), 1–19. https://doi.org/10.1080/08856257.2011.613599

Bosma, T., Stevenson, C. E., & Resing, W. C. M. (2017). Differences in Need for Instruction: Dynamic Testing in Children with Arithmetic Difficulties. *Journal of Education and Training Studies, 5*(6), 132–145. https://doi.org/10.11114/jets.v5i6.2326

Caffrey, E., Fuchs, D., & Fuchs, L. S. (2008). The Predictive Validity of Dynamic Assessment: A Review. *The Journal of Special Education, 41*(4), 254–270. https://doi.org/10.1177/0022466907310366

Campione, J. C., & Brown, A. L. (1978). Toward a theory of intelligence: Contributions from research with retarded children. *Intelligence, 2*(3), 279–304. https://doi.org/10.1016/0160-2896(78)90020-X

Campione, J. C., & Brown, A. L. (1987). Linking Dynamic Assessment with School Achievement. In C. S. Lidz (Ed.), Dynamic Assessment: *An Interactional Approach to Evaluating Learning Potential* (pp. 82–109). The Guilford Press.

Cho, E., Compton, D. L., Fuchs, D., Fuchs, L. S., & Bouton, B. (2014). Examining the Predictive Validity of a Dynamic Assessment of Decoding to Forecast Response to Tier 2 Intervention. *Journal of Learning Disabilities, 47*(5), 409–423. https://doi.org/10.1177/0022219412466703

Deutsch, R., & Reynolds, Y. (2000). The Use of Dynamic Assessment by Educational Psychologists in the UK. *Educational Psychology in Practice, 16*(3), 311–331. https://doi.org/10.1080/713666083

Dixon, C., Oxley, E., Nash, H., & Gellert, A. S. (2023). Does Dynamic Assessment Offer An Alternative Approach to Identifying Reading Disorder? A Systematic Review. *Journal of Learning Disabilities, 56*(6), 423–439. https://doi.org/10.1177/00222194221117510

Dörfler, T., Golke, S., & Artelt, C. (2009). Dynamic assessment and its potential for the assessment of reading competence. *Studies in Educational Evaluation, 35*(2–3), 77–82. https://doi.org/10.1016/j.stueduc.2009.10.005

Dowker, A. (2008). Individual differences in numerical abilities in preschoolers. *Developmental Science, 11*(5), 650–654. https://doi.org/10.1111/j.1467-7687.2008.00713.x

Dowker, A., Sarkar, A., & Looi, C. Y. (2016). Mathematics Anxiety: What Have We Learned in 60 Years? *Frontiers in Psychology, 7*. https://doi.org/10.3389/fpsyg.2016.00508

DYNAMIK Project Team (2022). *DTR-ASM – Dynamischer Test der Rechenfertigkeiten — Addition, Subtraktion, Multiplikation* [DTR- ASM – Dynamic Test of Arithmetic Skills – Addition, Subraction, Mulitplication]. [Unpublished test].

Freeman, L., & Miller, A. (2001). Norm-referenced, Criterion-referenced, and Dynamic Assessment: What exactly is the point? *Educational Psychology in Practice, 17*(1), 3–16. https://doi.org/10.1080/02667360120039942

Fuchs, D., Compton, D. L., Fuchs, L. S., Bouton, B., & Caffrey, E. (2011). The Construct and Predictive Validity of a Dynamic Assessment of Young Children Learning to Read: Implications for RTI Frameworks. *Journal of Learning Disabilities, 44*(4), 339–347. https://doi.org/10.1177/0022219411407864

Fuchs, L. S., Compton, D. L., Fuchs, D., Hollenbeck, K. N., Craddock, C. F., & Hamlett, C. L. (2008). Dynamic assessment of algebraic learning in predicting third graders' development of mathematical problem solving. *Journal of Educational Psychology, 100*(4), 829–850. https://doi.org/10.1037/a0012657

Fuchs, L. S., Compton, D. L., Fuchs, D., Hollenbeck, K. N., Hamlett, C. L., & Seethaler, P. M. (2011). Two-Stage Screening for Math Problem-Solving Difficulty Using Dynamic Assessment of Algebraic Learning. *Journal of Learning Disabilities, 44*(4), 372–380. https://doi.org/10.1177/0022219411407867

Haberstroh, S., & Schulte-Körne, G. (2022). The Cognitive Profile of Math Difficulties: A Meta-Analysis Based on Clinical Criteria. *Frontiers in Psychology, 13,* 842391. https://doi.org/10.3389/fpsyg.2022.842391

Haffner, J., Baro, K., Parzer, P., & Resch, F. (2005). *HRT 1-4 – Heidelberger Rechentest* [HRT 1-4 - Heidelberger Calculation Test]. Hogrefe.

Haywood, H. C., & Lidz, C. S. (2007). *Dynamic assessment in practice: Clinical and educational applications*. Cambridge University Press.

Herppich, S., Praetorius, A.-K., Förster, N., Glogger-Frey, I., Karst, K., Leutner, D., Behrmann, L., Böhmer, M., Ufer, S., Klug, J., Hetmanek, A., Ohle, A., Böhmer, I., Karing, C., Kaiser, J., & Südkamp, A. (2018). Teachers' assessment competence: Integrating knowledge-, process-, and product-oriented approaches into a competence-oriented conceptual model. *Teaching and Teacher Education, 76,* 181–193. https://doi.org/10.1016/j.tate.2017.12.001

Hessels, M. G. P. (1997). Low IQ but high learning potential: Why Zeyneb and Moussa do not belong in special education. *Educational and Child Psychology, 14*(4), 121–136.

Kaniel, S. (2010). Domain Specific vs Domain General: Implications for Dynamic Assessment. *Gifted Education International, 26*(1), 96–109. https://doi.org/10.1177/026142941002600112

Kaskens, J., Goei, S. L., Van Luit, J. E. H., Verhoeven, L., & Segers, E. (2021). Dynamic maths interviews to identify educational needs of students showing low math achievement. *European Journal of Special Needs Education, 37*(3), 432–446. https://doi.org/10.1080/08856257.2021.1889848

Kaskens, J., Segers, E., Goei, S. L., Van Luit, J. E. H., & Verhoeven, L. (2023). Dynamic Mathematics Interviews in Primary Education: The Relationship Between Teacher Professional Development and Mathematics Teaching. *Mathematics Teacher Education and Development, 25*(1), 61–80.

Kaufmann, L., Mazzocco, M. M., Dowker, A., Von Aster, M., Göbel, S. M., Grabner, R. H., Henik, A., Jordan, N. C., Karmiloff-Smith, A. D., Kucian, K., Rubinsten, O., Szucs, D., Shalev, R., & Nuerk, H.-C. (2013). Dyscalculia from a developmental and differential perspective. *Frontiers in Psychology, 4.* https://doi.org/10.3389/fpsyg.2013.00516

Krajewski, K., Dix, S., & Schneider, W. (2020). DEMAT 2+ *Deutscher Mathematiktest für zweite Klassen* [DEMAT 2+ German Math Test for Grade Two]. Hogrefe.

Lange, S. M., & Thompson, B. (2006). Early Identification and Interventions for Children at Risk for Learning Disabilities. *International Journal of Special Education, 21*(3), 108–119.

Lindner, K.-T., & Schwab, S. (2020). Differentiation and individualisation in inclusive education: A systematic review and narrative synthesis. *International Journal of Inclusive Education, 129*(12), 2199–2219. https://doi.org/10.1080/13603116.2020.1813450

Maas, C. J. M., & Hox, J. J. (2005). Sufficient Sample Sizes for Multilevel Modeling. *Methodology, 1*(3), 86–92. https://doi.org/10.1027/1614-2241.1.3.86

Maehler, C., & Schuchardt, K. (2011). Working Memory in Children with Learning Disabilities: Rethinking the criterion of discrepancy. *International Journal of Disability, Development and Education, 58*(1), 5–17. https://doi.org/10.1080/1034912X.2011.547335

Mangiafico, S. (2025). *rcompanion: Functions to Support Extension Education Program Evaluation* [Computer software]. https://cran.r-project.org/package=rcompanion

Mayring, P. (2015). Qualitative Content Analysis: Theoretical Background and Procedures. In A. Bikner-Ahsbahs, C. Knipping, & N. Presmeg (Eds.), *Approaches to Qualitative Research in Mathematics Education* (pp. 365–380). Springer Netherlands. https://doi.org/10.1007/978-94-017-9181-6_13

Moineddin, R., Matheson, F. I., & Glazier, R. H. (2007). A simulation study of sample size for multilevel logistic regression models. *BMC Medical Research Methodology, 7*(1), 34. https://doi.org/10.1186/1471-2288-7-34

Moll, K., Kunze, S., Neuhoff, N., Bruder, J., & Schulte-Körne, G. (2014). Specific Learning Disorder: Prevalence and Gender Differences. *PLoS ONE, 9*(7), e103537. https://doi.org/10.1371/journal.pone.0103537

Morsanyi, K., Van Bers, B. M. C. W., McCormack, T., & McGourty, J. (2018). The prevalence of specific learning disorder in mathematics and comorbidity with other developmental disorders in primary school-age children. *British Journal of Psychology, 109*(4), 917–940. https://doi.org/10.1111/bjop.12322

Navarro, J. I., Aguilar, M., Marchena, E., Ruiz, G., Menacho, I., & Van Luit, J. E. H. (2012). Longitudinal study of low and high achievers in early mathematics. *British Journal of Educational Psychology, 82*(1), 28–41. https://doi.org/10.1111/j.2044-8279.2011.02043.x

Paccagnella, O. (2011). Sample Size and Accuracy of Estimates in Multilevel Models: New Simulation Results. *Methodology, 7*(3), 111–120. https://doi.org/10.1027/1614-2241/a000029

Pameijer, N. (2006). Towards needs-based assessment: Bridging the gap between assessment and practice. *Educational and Child Psychology, 23*(3), 12–24. https://doi.org/10.53841/bpsecp.2006.23.3.12

R Core Team (2024). *R: A language and environment for statistical computing* [Computer software]. https://www.R-project.org

Reiser, H., Willmann, M., Urban, M., & Sanders, N. (2003). Different models of social and emotional needs consultation and support in German schools. *European Journal of Special Needs Education, 18*(1), 37–51. https://doi.org/10.1080/0885625032000042302

Resing, W. C. M., Elliott, J. G., & Grigorenko, E. L. (2012). Dynamic Testing and Assessment. In N. M. Seel (Ed.), *Encyclopedia of the Sciences of Learning* (pp. 1055–1058). Springer US. https://doi.org/10.1007/978-1-4419-1428-6

Resing, W. C. M., Stevenson, C. E., & Bosma, T. (2012). Dynamic Testing: Measuring Inductive Reasoning in Children With Developmental Disabilities and Mild Cognitive Impairments. *Journal of Cognitive Education and Psychology, 11*(2), 159–178. https://doi.org/10.1891/1945-8959.11.2.159

Resing, W. C. M., Tunteler, E., de Jong, F. M., & Bosma, T. (2009). Dynamic testing in indigenous and ethnic minority children. Learn*ing and Individual Differences, 19*(4), 445–450. https://doi.org/10.1016/j.lindif.2009.03.006

Revelle, W. (2025). psych: *Procedures for Psychological, Psychometric, and Personality Research* [Computer software]. https://CRAN.R-project.org/package=psych

Sansour, T., & Bernhard, D. (2018). Special needs education and inclusion in Germany and Sweden. Alter, 12–3, 127–139. https://doi.org/10.1016/j.alter.2017.12.002

Scherer, P., Beswick, K., DeBlois, L., Healy, L., & Moser Opitz, E. (2017). Assistance of Students with Mathematical Learning Difficulties – How Can Research Support Practice? – A Summary. In G. Kaiser (Ed.), *Proceedings of the 13th International Congress on Mathematical Education: ICME-13* (pp. 249–259). Springer International Publishing. https://doi.org/10.1007/978-3-319-62597-3

Schukajlow, S., Rakoczy, K., & Pekrun, R. (2023). Emotions and motivation in mathematics education: Where we are today and where we need to go. *ZDM – Mathematics Education, 55*(2), 249–267. https://doi.org/10.1007/s11858-022-01463-2

Sternberg, R. J., & Grigorenko, E. L. (2002). *Dynamic Testing: The Nature and Measurement of Learning Potential*. Cambridge University Press.

Sternberg, R. J., Grigorenko, E. L., Ngorosho, D., Tantufuye, E., Mbise, A., Nokes, C., Jukes, M., & Bundy, D. A. (2002). Assessing intellectual potential in rural Tanzanian school children. *Intelligence, 30*(2), 141–162. https://doi.org/10.1016/S0160-2896(01)00091-5

Tiekstra, M., Hessels, M. G. P., & Minnaert, A. E. M. G. (2009). Learning capacity in adolescents with mild intellectual disabilities. *Psychological Reports, 105*(3), 804–814. https://doi.org/10.2466/PR0.105.3.804-814

Tiekstra, M., Minnaert, A., & Hessels, M. G. P. (2016). A review scrutinising the consequential validity of dynamic assessment. *Educational Psychology, 36*(1), 112–137. https://doi.org/10.1080/01443410.2014.915930

Tzuriel, D. (2000a). Dynamic Assessment of Young Children: Educational and Intervention Perspectives. *Educational Psychology Review, 12*(4), 385-435. https://doi.org/10.1023/A:1009032414088

Tzuriel, D. (2000b). The Seria-Think Instrument: Development of a Dynamic Test for Young Children. *School Psychology International, 21*(2), 177–194. https://doi.org/10.1177/0143034300212005

Tzuriel, D., & Universin, B. I. (2001). Dynamic Assessment is not Dynamic Testing. *Issues in Education, 7*(2), 237–249.

Veerbeek, J., Hessels, M. G. P., Vogelaar, S., & Resing, W. C. M. (2017). Pretest Versus No Pretest: An Investigation Into the Problem-Solving Processes in a Dynamic Testing Context. Jou*rnal of Cognitive Education and Psychology, 16*(3), 260–280. https://doi.org/10.1891/1945-8959.16.3.260

VERBI Software (2021). *MAXQDA – Software für qualitative Datenanalyse* (Version 2022) [MAXQDA – Software for Qualitative Data Analysis] [Computer software]. maxqda.com

Acknowledgements

This research was funded by the German Ministry of Education and Research (BMBF; Grant numbers 01NV2120A, 01NV2120B and 01NV2120C. DeepL and ChatGPT 4 were used to support language editing in English.

The authors report there are no competing interests to declare.

We would like to thank Alea Kreyes, Lydia Küttner, Fiona Ladisch and Linda Kuhr for their work during the development of the DTR-ASM as well as the data collection.