Computerized process-oriented dynamic testing of children's ability to reason by analogy using log data

Veerbeek, J.; Vogelaar, B.

Special Issue: Process Data in Computer-Based Assessment: Opening the Black Box
Original Article

Computerized Process-Oriented Dynamic Testing of Children's Ability to Reason by Analogy Using Log Data

Jochanan Veerbeek[1] and Bart Vogelaar[2]

[1]Education and Child Studies, Leiden University, The Netherlands
[2]Developmental and Educational Psychology, Leiden University, The Netherlands

Abstract: The study investigated the value of process data obtained from a group-administered computerized dynamic test of analogical reasoning, consisting of a pretest-training-posttest design. We sought to evaluate the effects of training on processes and performance, and the relationships between process measures and performance on the dynamic test. Participants were N = 86 primary school children (M_{age} = 8.11 years, SD = 0.63). The test consisted of constructed-response geometrical analogy items, requiring several actions to construct an answer. Process data enabled scoring of the total time, the time taken for initial planning of the task, the time taken for checking the answer that was provided, and variation in solving time. Training led to improved performance compared to repeated practice, but this improvement was not reflected in task-solving processes. Almost all process measures were related to performance, but the effects of training or repeated practice on this relationship differed widely between measures. In conclusion, the findings seemed to indicate that investigating process indicators within computerized dynamic testing of analogical reasoning ability provided information about children's learning processes, but that not all processes were affected in the same way by training.

Keywords: dynamic testing, log data, process assessment, learning potential, time measures

Tests for children's cognitive or academic abilities usually consist of a short moment of instruction, followed by a single test session in which children independently solve items. Such tests, also known as static tests, are suited to measure abilities and knowledge, which for a large part have already been acquired as a result of prior learning, but do not necessarily measure children's potential for learning (Resing et al., 2020). As such, they can lead to underestimating the cognitive abilities of certain groups of children. A different form of testing, dynamic testing, was developed to address these drawbacks, specifically by measuring children's ability to learn from instruction. To this end, a training procedure is included in the test, enabling measurement of children's prior knowledge and skills and the extent to which they could profit from help or instruction. Studies show that dynamic tests provide insight into intra- and inter-individual learning paths, in terms of the extent to which children show improvement, their task approach, their ability to generalize what they have learned, and the instructions they need in order to learn, all considered measures of potential for learning (Resing et al., 2020). More importantly, they also provided insight into the notion that these measures are highly domain-specific and can be different across domains. However, to date, this approach has not seen widespread implementation in practice, likely because of the time, labor, and costs involved in administering them (Resing et al., 2020). One solution that has been mentioned is using computerized dynamic tests. Computerized testing procedures do not only present opportunities to provide children with individualized, yet standardized instruction (Resing et al., 2020), but also enable recording fine-grained process data from children's interactions with the computerized dynamic test (Resing & Elliott, 2011; Veerbeek et al., 2019) while also enabling more time-efficient group-based testing. The current study sought to investigate the processes involved in solving analogical reasoning tasks in a dynamic testing format and sought to investigate the changes in these processes as a result of training.

Computerized Dynamic Testing

Dynamic tests exist in a variety of forms, but all of them have in common that the testing procedure includes help,

in the form of feedback, mediation, or hints (Resing et al., 2020). Often, training is provided by means of graduated prompts techniques (Campione & Brown, 1987; Resing & Elliott, 2011). As part of this training procedure, children work on tasks independently. Then, help is provided through a series of hierarchically structured prompts, starting with general, metacognitive hints which if these do not enable the child to correctly solve the task become more task-specific, cognitive prompts. If these do not lead to the correct answer, the child receives step-by-step modeling of the process of solving the task correctly. In graduated prompts, detailed hints are given only if a child cannot solve the item using more general help. Children are always provided with the minimal amount of information they need to solve the task, which minimizes working memory load and should prevent expertise reversal effects in children that are more experienced in solving the tasks of interest (Kalyuga et al., 2003).

Because of the high level of standardization, dynamic tests using graduated prompts training are well suited for computerization. In previous studies, computerized graduated prompts were found an effective approach which led to learning gains in a variety of tasks and domains (e.g., Poehner & Lantolf, 2013; Veerbeek et al., 2019; Vogelaar et al., 2021; Wu et al., 2017). In addition, computerized dynamic tests allow for analyzing process data, providing additional information on solving processes on inductive reasoning tasks (Veerbeek et al., 2019).

Dynamic tests often utilize analogical reasoning tasks. Analogical reasoning, considered a subform of inductive reasoning, involves extracting novel information that could be relevant for more than one context and applying it to other contexts (Goswami, 2012). It is believed to be one of the central processes that allow for cognitive development, underlying a range of cognitive skills and processes such as fluid intelligence, and, implicitly, for everyday learning, and is thought to be closely related to important skills such as problem-solving, and transfer (Goswami, 2012; Richland & Simms, 2015).

Uncovering Process Indicators

Process-oriented dynamic testing is aimed at discovering individual differences and changes in task-solving processes that result from a training procedure (Resing & Elliott, 2011). Data on processes and strategies involved in task solving could provide information on the help a child needs to improve performance on a task (Greiff et al., 2015). Although aptitude-treatment interactions have not been reliably found in prior research, personalized education appears to more reliably lead to improved learning outcomes when a dynamic approach is used, taking into account changes in learning processes over time and as a result of instruction (Tetzlaff et al., 2021). Information such as the processing time or the steps taken during task solving can be used to construct process indicators, which are said to reflect behavior or information processing (Goldhammer et al., 2017). Prior research has successfully uncovered process indicators in a variety of domains, and shown that process indicators indeed reflected individual differences in ability and could provide information beyond test (product) scores (e.g., Naumann & Goldhammer, 2017; Stadler et al., 2020; Veerbeek et al., 2017, 2019).

Time measures are often used as process indicators. The allocation of time is a basic process that is relevant to performance on most tests and can be indicative of cognitive and metacognitive strategy use (Naumann, 2019). However, interpreting a time-on-task effect is complex, as the interpretation depends on an interaction between the complexity of the task and the solver's individual level of ability (Goldhammer et al., 2014, 2017; Naumann, 2019). For tasks requiring more controlled processing, more time spent on the task corresponds to better performance, while in tasks requiring automatic processing, the opposite seems true (Goldhammer et al., 2014, 2017).

Study Aims

In the current paper, it was investigated whether obtaining process indicators from digital log file data in a computerized process-oriented dynamic test of analogical reasoning ability was feasible, focusing specifically on the effect of training on children's test performance, and processes while problem-solving. Moreover, it was investigated how these process indicators were related to task success on individual items.

First, we aimed to evaluate the effects of the computerized graduated prompts training on both the performance and solving behavior of children on the computerized dynamic test of analogical reasoning. It was expected that children who received training would show more progression from pretest to posttest in terms of analogical reasoning (1) task-solving accuracy (number of items answered correctly), and (2) accuracy on transformations (transformations answered correctly), but this change was not expected to be reflected in (3) the solving processes as reflected by the process indicators obtained from the log data, as found in prior research (Veerbeek et al., 2019).

Next, it was investigated to what extent process indicators were related to task success (accuracy) and other process indicators. It was expected that time-based process indicators would be related to task-solving accuracy, but that the pattern of relations would change differentially as

a result of training, as compared to repeated practice (Resing & Elliott, 2011).

Methods

Participants

The participants for the current research consisted of $N = 86$ children, 52 boys and 34 girls (age $M = 8.11$ years, $SD = 0.63$), attending the 2nd or 3rd grade of primary school. All children were believed to be successful learners, as they all attended provisions for talented students. The participants were recruited from 12 schools, all located in the western part of the Netherlands. The study was approved by the ethical board of the Psychology Institute and informed consent was obtained for all children prior to testing.

Design and Procedure

The study consisted of a randomized blocking pretest-posttest experimental design. Based on their scores on an intelligence screener of the Intelligence and Developmental Scale – 2 (IDS-2; Ruiter et al., 2018) children were blocked and randomly assigned to either a group that received computerized training, or the group that only received repeated practice. The study consisted of a total of five sessions, and all testing took place at the children's school. The first session consisted of individualized administration of the IDS-2. The second session consisted of the dynamic test's pretest. During the third and fourth sessions, the children in the training group received computerized graduated prompts training, and children in the control group practiced independently with the same items. The fifth and final session consisted of the posttest and a reversal task. Each of these sessions lasted 30–45 min. Sessions took place once or twice a week. For these four sessions, groups of children worked independently on a computer with a student examiner present at all times to provide help when technical difficulties arose.

Materials

Intelligence and Developmental Scale – 2 (IDS-2)

To enable the blocking of children based on their general cognitive abilities, the IDS-2 intelligence screener was administered. This IQ screener consists of two subtests, Categories and Matrix Reasoning, and provides an indication of the child's crystallized intelligence (categories) and fluid intelligence (matrix reasoning). The screener takes about 10 min to administer (Ruiter et al., 2018). The non-verbal subtest Matrix Reasoning consists of 35 multiple-choice visual-spatial analogy items of the type A:B::C:?. The verbal subtest Category Naming consists of 34 multiple-choice items referring to children's verbal reasoning and prior knowledge of categories. The subtest Matrix Reasoning has a test-retest reliability of .86 and Categories of .93 (Grob & Hagmann-von Arx, 2018).

Computerized Dynamic Test of Analogical Reasoning

The current study utilized a computerized dynamic test of geometrical analogies (Vogelaar et al., 2021; see Figure 1 for a screenshot of the program). The constructed response items were presented as pictures and followed an A:B::C:? format. The items were constructed utilizing a maximum of six different geometrical shapes (circle, square, triangle, ellipse, hexagon, pentagon), which could be modified by adding figures, changing position, size, rotation, or halving. Children had to construct their own solutions by dragging shapes and dropping them into place. All the different shapes were provided in the screens, and additional buttons could be used to transform the shapes or to reset the solution window.

Children's answers were automatically scored by the computer program. For each figure that was placed, the program recorded the basic event information, consisting of the timestamp, place, size, and identity of each placement into log files. When an item was finished, a summary was written into the log file as well, consisting of information on the correctness of the item, and which transformations were successfully used.

Pretest and Posttest

The test contained a pretest and posttest, which were constructed with 15 items each, that increased in level of difficulty and were constructed to contain parallel items. The pretest started with a short, general instruction and two example items. Beyond this, no help was provided during the pre- and posttest.

Graduated Prompts Training

The graduated prompts training was provided fully by the laptop, through audio that was provided contingent upon the child's answers. It consisted of two sessions, both consisting of six items. If a child could not independently provide a correct answer, prompts were provided to help the child reach the correct answers. The prompts were hierarchically ordered, and started with the most general, metacognitive prompt, and ended with item-specific cognitive prompts. If these did not lead the child to provide a correct answer, modeling was provided to show the child the correct process for solving the task. The training was constructed to address the processes involved in solving analogical reasoning tasks, and was based on task analysis of analogy problem-solving.

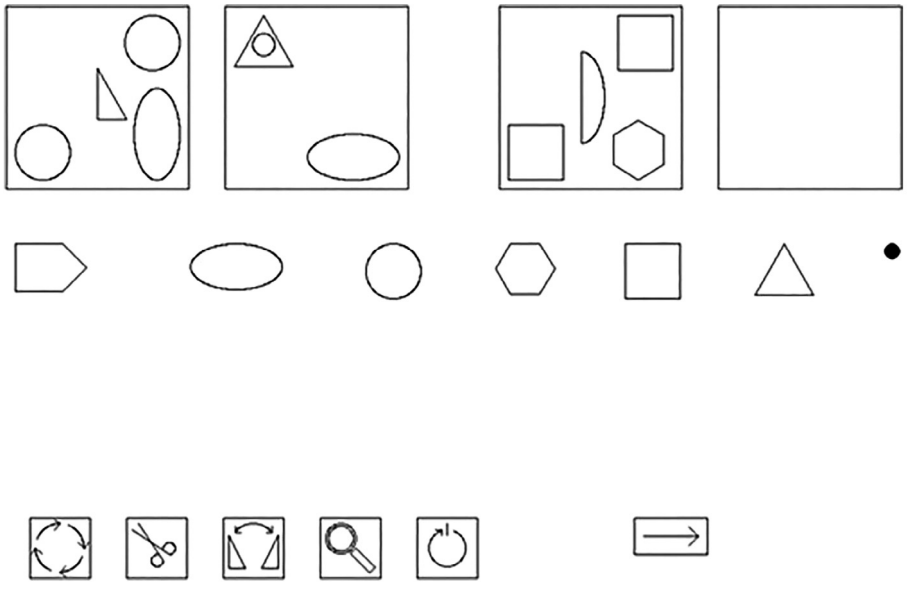

Figure 1. Screenshot of the computerized dynamic test of analogical reasoning, showing the shapes and the buttons to transform the shapes (left-to-right: rotate, cut, flip, size, reset/correct).

Scoring

Accuracy

Two measures for analogy-solving accuracy were used, the primary measure was the number of items correctly solved on the pretest and posttest. To allow for a finer resolution in solving accuracy, additionally, the number of elements solved correctly was calculated as well, counting each separate figure within an item.

Transformation Measures

Additionally, the program used computer-automated scoring of incorrectly applied transformations when constructing the analogy solution. Based on the identifiers of the unique shapes, it was counted how many instances there were of mistakes in (1) size, (2) place, (3) shape, (4) cut (half or whole figure), (5) flip (mirrored image, only applicable to some half shapes), (6) rotation, or (7) omissions. In addition, it was scored how many times the children used the option to "reset" the answer, to make a new start constructing the solution. These were scored as corrections.

Time-Based Measures

The primary log data used to create the process indicators were the timestamps for the pieces of the answers that children placed in the program. Difference scores for the timestamps were calculated by subtracting them from each other. The resulting time intervals were thought to be indicative of the time allocated to solve that specific piece of the task. Because there was not a set amount of elements that items consisted of, and children differed in the number of elements they used to build up their answers, the number of intervals available differed both between items and between participants. Based on the available time intervals and timestamps, several process indicators were created. All times were in milliseconds.

Completion Time

The first process indicator was item completion time, which indicated the time from the start of the item, up to the placement of the last element. As such, it represents the time it took the children to construct their answers.

Log Item Time

The log item time indicator consisted of the total time children spent on an item, up to the moment they clicked the button to move on to the next item. As such it incorporated the time between the placement of the last element and the child clicking the "next" button.

Encoding Time

To estimate the time children spent on encoding the information contained in the item, the time between the first presentation of the item and the placement of the first element for the construction of an answer was used.

Checking Time

At the end of the process of solving an item, the time was measured between children's placement of the last element to complete the item solution, and clicking the button to move to the next item. This time was thought to indicate the time children took to check their answers.

Encoding Proportion

To create a measure for the time taken for encoding, which would be less dependent on the overall speed of processing and completing items, the proportion of time taken for encoding was used. This indicator consisted of the

encoding time, divided by the completion time (Kossowska & Nęcka, 1994; Resing & Elliott, 2011; Veerbeek et al., 2017).

Checking Proportion

Similar to encoding proportion, checking proportion was thought to provide a measure of time taken for checking provided answers that would be less dependent on overall speed. This indicator consisted of checking time divided by log item time.

Time Variation

An additional process indicator used was the time variation in children's speed of providing a solution. The variation in intervals between each action was calculated by using all the interval times, calculating their variance, and in turn taking the square root.

Results

Effects of Training

Univariate analyses of variance (ANOVAs) were used to check for a priori differences in abilities between the two conditions (training/control). No significant effect for the condition was found on intelligence screener scores (p = .428) or items correct on the pretest (p = .425). To estimate the effects of training on analogy-solving accuracy, transformations, and process indicators, three mixed-model repeated measures Multivariate analyses of variance (MANOVAs) were used. The means and standard deviations were displayed in Table 1, and the statistics for the multivariate and univariate effects were displayed in Table 2.

First, the effect of training on children's accuracy in solving analogy items was investigated, using a mixed-model repeated measures MANOVA with session (pretest/posttest) as the within-subjects factor and condition (training/control) as the between-subjects factor. The total number of correctly solved items and the number of elements solved correctly were the dependent variables. Significant multivariate effects were found for session, condition, and session × condition (see Table 2). Significant univariate effects were found for both items correct and elements correct. A significant univariate interaction effect was found for session × condition for items correct and elements correct, which demonstrated a larger improvement in analogy-solving accuracy for the group that had received training than for the children who had received repeated practice.

Transformation measures were also investigated using mixed-model repeated measures MANOVA, with a session (pretest/posttest) as the within-subjects factor and condition (training/control) as the between-subjects factor. Multivariate tests showed a significant effect for session, for condition, and an effect bordering on significance (p = .050) for session × condition. Univariate effects of the session were significant for all measures except incorrect flips and omissions. The mean scores provided in Table 1 indicate that all types of incorrect transformations decreased from the pretest to the posttest. Univariate effects for the condition were only significant for incorrect sizes. The answers of the children in the control group contained more incorrect sizes than those of the trained children. Univariate interaction effects for session × condition were only significant for incorrect rotations. Trained children showed a stronger decrease in the number of incorrect rotations from the pretest to the posttest than the control group. No other significant interaction effects were found for transformation measures.

To investigate the effects of training on children's processes when solving analogy items, a third repeated-measures MANOVA was used, with a session (pretest/posttest) as the within-subjects factor and condition (training/control) as the between-subjects factor. Multivariate tests showed a significant effect for session, but not for condition or session × condition, indicating that in general the processes changed between the sessions, but did not change differentially as a consequence of training. Univariate effects of the session were found to be significant for completion time, log item time, check time and check proportion. Encoding time, encoding proportion and time variation did not yield significant univariate effects for session. The mean scores further demonstrated that completion time, log item time and check time all decreased from the pretest to the posttest, while check proportion increased.

Relationships Between Process Indicators and Accuracy

To investigate the relationships between time-based process indicators and solving accuracy on the analogy items, Pearson's correlations were calculated on both the pretest and the posttest. For the posttest, the analyses were split by condition. Differences in correlations were compared using Fisher's r-to-z. All significant differences were mentioned in the text. The results were displayed in Table 3. The results indicated that completion time and log item time showed similar patterns of relationships with solving accuracy. Both were strongly and positively related to the accuracy, indicating that, for pretest and posttest, those who spent more time on the task solved more items correctly. Encoding time and time variation also showed comparable patterns of relationships, as both were moderately related

Table 1. The means (*M*s) and standard deviations (*SD*s) for the dependent variables of the MANOVAs

	Pretest		Posttest	
	Training *N* = 43	Control *N* = 40	Training *N* = 43	Control *N* = 40
	M (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)
Items correct	3.79 (3.55)	2.85 (2.97)	8.74 (4.15)	5.43 (4.56)
Elements correct	25.21 (14.90)	21.40 (14.34)	33.95 (13.69)	24.55 (15.72)
Corrections	6.36 (4.97)	6.23 (4.95)	3.17 (3.15)	2.22 (2.38)
Incorrect size	0.09 (0.09)	0.12 (0.12)	0.04 (0.08)	0.10 (0.13)
Incorrect place	0.28 (0.25)	0.32 (0.21)	0.14 (0.16)	0.23 (0.19)
Incorrect shape	0.99 (0.89)	1.21 (0.86)	0.53 (0.78)	0.94 (0.90)
Incorrect cut	0.04 (0.06)	0.05 (0.12)	0.01 (0.04)	0.01 (0.03)
Incorrect flip	0.01 (0.03)	0.02 (0.09)	0.00 (0.00)	0.00 (0.00)
Incorrect rotation	0.17 (0.14)	0.16 (0.16)	0.09 (0.09)	0.15 (0.12)
Omissions	0.00 (0.00)	0.02 (0.09)	0.00 (0.00)	0.00 (0.00)
Completion time	88,629 (34,607)	87,018 (31,327)	59,493 (23,772)	69,226 (82,443)
Log item time	93,091 (35,498)	92,821 (31,982)	63,072 (24,288)	72,909 (82,492)
Encoding time	35,034 (15,022)	36,807 (17,566)	21,297 (8,305)	33,840 (77,231)
Check time	4,462 (1,820)	5,803 (5,590)	3,579 (1,961)	3,684 (1,909)
Encoding proportion	0.42 (0.08)	0.44 (0.12)	0.43 (0.12)	0.46 (0.12)
Check proportion	0.07 (0.02)	0.08 (0.04)	0.09 (0.06)	0.09 (0.04)
Time variation	17,445 (8,898)	17,861 (8,648)	10,634 (5,107)	16,443 (34,540)

Table 2. Univariate effects for the RM MANOVAs investigating effects of training

	Session			Condition			Session × Condition		
	F(1, 81)	*p*	η^2_p	*F*(1, 81)	*p*	η^2_p	*F*(1, 81)	*p*	η^2_p
Accuracy	**50.81**	**.000**	**.56**	**4.75**	**.011**	**.11**	**4.61**	**.013**	**.10**
(multivariate)		**Wilks' λ = .44**			**Wilks' λ = .89**			**Wilks' λ = .90**	
Items correct	**93.30**	**.000**	**.54**	**8.03**	**.006**	**.09**	**9.31**	**.003**	**.10**
Elements correct	**25.25**	**.000**	**.24**	**4.86**	**.030**	**.057**	**5.59**	**.021**	**.07**
Transformations	**16.16**	**.000**	**.64**	**2.10**	**.046**	**.19**	**2.07**	**.050**	**.19**
(multivariate)		**Wilks' λ = .36**			**Wilks' λ = .81**			**Wilks' λ = .82**	
Corrections	**43.81**	**.000**	**.35**	0.58	.448	.01	0.56	.458	.01
Incorrect size	**8.46**	**.005**	**.10**	**4.51**	**.037**	**.05**	1.54	.218	.02
Incorrect place	**17.74**	**.000**	**.18**	3.32	.072	.04	0.86	.357	.01
Incorrect shape	**23.14**	**.000**	**.22**	3.20	.077	.04	1.54	.219	.02
Incorrect cut	**9.83**	**.002**	**.11**	0.22	.644	.00	0.23	.635	.00
Incorrect flip	2.27	.136	.03	0.22	.644	.00	0.22	.644	.00
Incorrect rotation	**7.58**	**.007**	**.09**	1.32	.253	.02	**4.74**	**.032**	**.06**
Omissions	1.05	.308	.01	1.05	.308	.01	1.05	.308	.01
Time-based	**12.85**	**.000**	**.51**	0.91	.495	.07	0.49	.816	.04
(multivariate)		**Wilks' λ = .49**							
Completion time	**13.04**	**.001**	**.14**						
Log item time	**14.68**	**.000**	**.16**						
Encoding time	1.98	.163	.02						
Check time	**9.31**	**.003**	**.10**						
Encoding proportion	1.03	.313	.01						
Check proportion	**6.26**	**.014**	**.07**						
Time variation	2.35	.129	.03						

Note. Significant effects were displayed in bold font.

Table 3. Correlations between items and elements total correct and process indicators based on time

			Posttest			
	Pretest (n = 85)		Trained (n = 44)		Control (n = 40)	
	Items correct	Elements correct	Items correct	Elements correct	Items correct	Elements correct
Completion time	**.53****	**.59****	**.35***	**.43****	**.36***	**.35***
Log item time	**.54****	**.58****	**.35***	**.42****	**.36***	**.35***
Encoding time	**.39****	**.40****	.16	.23	.28	.24
Check time	.18	.09	.04	.01	−.06	−.06
Encoding proportion	−.20	**−.27***	**−.47****	**−.55****	−.18	−.14
Check proportion	−.21	**−.38****	**−.45****	**−.54****	**−.46****	**−.55****
Time variation	**.38****	**.37****	.01	.02	.28	.26

Note. Significant correlations were displayed in bold font. **Correlation is significant at the .01 level (2-tailed); *Correlation is significant at the .05 level (2-tailed).

to solving accuracy on the pretest, but no longer were significantly related to accuracy on the posttest for either group. The moderate positive relationships indicated that spending more time on encoding was related to higher accuracy, and more variation in working tempo was related to higher solving accuracy.

The difference in correlation for time variation between the pretest and posttest for the trained children was significant on items correct (p = .041), but not for elements correct (p = .054). Check time was not related to solving accuracy. On the pretest, encoding proportion showed a weak negative relationship to elements correct only. On the posttest, it showed a moderate to a strong negative relationship with solving accuracy for the trained children only. This indicated that a smaller proportion of the time spent on encoding was related to higher solving accuracy. The difference between the trained children and the control group children on the posttest in relationships between encoding proportion and elements correct was significant (p = .035), but not for encoding proportion and items correct (p = .147). For checking proportion, on the pretest a moderate negative relationship was found with elements correct only. On the posttest, checking time showed a moderate to strong negative relationship with all measures of solving accuracy, for both groups, indicating that a smaller portion of the time spent checking the answer was related to higher solving accuracy.

Discussion

The current paper utilized a computerized dynamic testing format to investigate the processes involved in analogical reasoning and whether these processes change as a result of training.

In line with our expectations and a myriad of prior studies, children's analogy-solving accuracy improved from pretest to posttest and improved more for children that had received training than for children in the control group (Resing et al., 2020). However, this pattern of improvement was not found for all process measures. To be precise, only one transformation measure (incorrect rotation) showed more improvement for the trained than for the control group children. Similarly, the time-based process indicators in general showed to change from pretest to posttest. In line with prior studies, none of these measures showed differential effects as a result of training compared to the control group (Veerbeek et al., 2019).

This finding might indicates that if we look at smaller, separate components of performance, these do not necessarily show the same effects of training seen on the resulting product of performance. Instead, all these separate components seem to interact and accumulate into eventual performance gains where the whole is bigger than the sum of its parts. However, the training may have influenced other processes, such as children's schemas and their construction of mental representations (e.g., Kalyuga et al., 2003). These processes interact with children's use and availability of working memory and may have been influenced by training, but were beyond the scope of our process measures.

Relationship Between Process Indicators and Accuracy

In terms of the relationships between time-based process indicators and accuracy in solving analogy items, a dynamic picture emerged. In line with prior research involving complex tasks, more time spent on the task was related to better outcomes (Goldhammer et al., 2014, 2017; Naumann, 2019). Both encoding time and time variation were related to better outcomes, but only on the pretest. This was in line with previous findings, where processes that were not addressed by the training were no longer

related to performance on the posttest (Veerbeek et al., 2019). Check time was unrelated to performance on the analogy task.

The negative relationship between encoding proportion and solving accuracy for trained children indicated that a smaller portion of time spent on the initial stages of the task was related to better outcomes. Given the complexity of the analogy items used in the current research, encoding all relations might exceed children's working memory capacity, so part-by-part encoding might be the more promising approach for successfully solving the tasks (Kossowska & Nęcka, 1994).

For checking proportion, weak negative relationships were found with accuracy on the pretest, which became stronger on the posttest. Children that took a bigger portion of time for the final stage of the task, between the construction of the answer and moving on to the next item, were less likely to provide a correct answer. This might indicate that children detected that their answers were not correct or doubted their correctness, and lingered to think if they could come up with a better answer. This would be in line with the findings of Eichmann and colleagues (2020), who found that correct responses were accompanied by a high proportion of goal-directed behavior, whilst incorrect responses were preceded by a high proportion of non-targeted exploration, possibly indicating confusion about distraction. The lack of a distinction between goal-directed and non-targeted behavior may explain why training, which included metacognitive prompts that could affect both encoding/exploration and checking, did not appear to have a discernable effect on these process measures or the relationship they showed with performance.

Implications

The findings of the current paper provide additional support for the interactive relationship between time on task, item complexity, and the solver's ability (Goldhammer et al., 2014, 2017; Naumann, 2019). More research will be necessary to address the complex data that are involved in exploring the relationships between task-solving processes, ability, task complexity, and the effects of training. To account for complex, interactive, non-linear relationships, as well as skewed data, future research should be directed at using bigger datasets with more advanced modeling and non-linear analysis methods to evaluate the robustness and generalizability of results. However, the current findings could indicate that not just a solver's ability, but also their familiarity with the task might play a role. In this context, using a dynamic test could provide compensation for children's (lack of) familiarity with the type of task. The current findings contribute to the support for using dynamic measures, specifically process-oriented dynamic tests that take into account children's learning processes, to personalize education (Tetzlaff et al., 2021).

With regard to the process measures, some appeared to be more valuable within the context of a dynamic test of analogical reasoning than others. Notably, the time taken for checking the item (the final stage of the task) was unrelated to performance and therefore does not seem to provide any value in terms of understanding children's performance on the tasks. With regard to task time, one could think that a more accurate indicator of the time spent on a task would outperform one that seemingly involved task-unrelated time. However, in the current study, the accurate time spent on the task and the less accurate measure until the child moved on to the next item did not seem to differ in any of the analyses.

Overall, the findings seem to indicate that investigating process indicators within computerized dynamic testing of inductive reasoning ability provide information about children's learning processes. Not all processes are affected in the same way by training. For the further development of dynamic tests, this implies that the connection between training and the task-solving processes could be kept closer in mind when designing training or determining which effects of training are to be expected on the process indicators. Future dynamic tests could combine elements from intelligent tutoring systems to adapt instruction to learner characteristics using student modeling (e.g., Tetzlaff et al., 2021) with existing formats such as graduated prompts to evaluate children's needs for instruction in a dynamic framework.

References

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). Guilford Press.

Eichmann, B., Greiff, S., Naumann, J., Brandhuber, L., & Goldhammer, F. (2020). Exploring behavioural patterns during complex problem-solving. *Journal of Computer Assisted Learning, 36*, 933–956. https://doi.org/10.1111/jcal.12451

Goldhammer, F., Naumann, J., Rölke, H., Stelter, A., & Tóth, K. (2017). Relating product data to process data from computer-based competency assessment. In D. Leutner, J. Fleischer, J. Grünkorn, & E. Klieme (Eds.), *Competence assessment in education* (pp. 407–425). Springer. https://doi.org/10.1007/978-3-319-50030-0

Goldhammer, F., Naumann, J., Stelter, A., Tóth, K., Rölke, H., & Klieme, E. (2014). The time on task effect in reading and problem solving is moderated by task difficulty and skill: Insights from a computer-based large-scale assessment. *Journal of Educational Psychology, 106*(3), 608–626. https://doi.org/10.1037/a0034716

Goswami, U. C. (2012). Analogical reasoning by young children. In N. M. Seel (Ed.), *Encyclopedia of the sciences of learning* (pp. 225–228). Springer. https://doi.org/10.1007/978-1-4419-1428-6_993

Greiff, S., Wüstenberg, S., & Avvisati, F. (2015). Computer-generated log-file analyses as a window into students' minds? A showcase study based on the PISA 2012 assessment of problem solving. *Computers and Education, 91*, 92–105. https://doi.org/10.1016/j.compedu.2015.10.018

Grob, A., & Hagmann-von Arx, P. (2018). *IDS-2, Intelligentie- en ontwikkelingsschalen voor kinderen en jongeren: Verantwoording en psychometrie* (Dutch revision, S. Ruiter) [IDS-2, Intelligence and Development Scales for Children and Adolescents: Justification and psychometrics]. Hogrefe Uitgevers B.V.

Kalyuga, S., Ayres, P., Chandler, P., & Sweller, J. (2003). The expertise reversal effects. *Educational Psychologist, 38*(1), 23–31. https://doi.org/10.1207/S15326985EP3801_4

Kossowska, M., & N□cka, E. (1994). Do it your own way: Cognitive strategies, intelligence, and personality. *Personality and Individual Differences, 16*(1), 33–46. https://doi.org/10.1016/0191-8869(94)90108-2

Naumann, J. (2019). The skilled, the knowledgeable, and the motivated: Investigating the strategic allocation of time on task in a computer-based assessment. *Frontiers in Psychology, 10*, Article 1429. https://doi.org/10.3389/fpsyg.2019.01429

Naumann, J., & Goldhammer, F. (2017). Time-on-task effects in digital reading are non-linear and moderated by persons' skills and tasks' demands. *Learning and Individual Differences, 53*, 1–16. https://doi.org/10.1016/j.lindif.2016.10.002

Poehner, M. E., & Lantolf, J. P. (2013). Bringing the ZPD into the equation: Capturing L2 development during Computerized Dynamic Assessment (C-DA). *Language Teaching Research, 17*(3), 323–342. https://doi.org/10.1177/1362168813482935

Resing, W. C. M., & Elliott, J. G. (2011). Dynamic testing with tangible electronics: Measuring children's change in strategy use with a series completion task. *The British Journal of Educational Psychology, 81*(Pt 4), 579–605. https://doi.org/10.1348/2044-8279.002006

Resing, W. C. M., Elliott, J. G., & Vogelaar, B. (2020). Assessing potential for learning in school children. *Oxford Research Encyclopedia of Education, June*, 1–19. https://doi.org/10.1093/acrefore/9780190264093.013.943

Richland, L. E., & Simms, N. (2015). Analogy, higher order thinking, and education. *Wiley Interdisciplinary Reviews: Cognitive Science, 6*(2), 177–192. https://doi.org/10.1002/wcs.1336

Ruiter, S., Timmerman, M., & Visser, L. (2018). *IDS-2: Intelligentie- en ontwikkelingsschalen voor kinderen en jongeren. Afnamehandleiding* [IDS-2, Intelligence and Development Scales for Children and Adolescents: Administration manual]. Hogrefe Uitgevers.

Stadler, M., Hofer, S., & Greiff, S. (2020). First among equals: Log data indicates ability differences despite equal scores. *Computers in Human Behavior, 111*, Article 106442. https://doi.org/10.1016/j.chb.2020.106442

Tetzlaff, L., Schmiedek, F., & Brod, G. (2021). Developing personalized education: A dynamic framework. *Educational Psychology Review, 33*, 863–882. https://doi.org/10.1007/s10648-020-09570-w

Veerbeek, J., Verhaegh, J., Elliott, J. G., & Resing, W. C. M. (2017). Process-oriented measurement using electronic tangibles. *Journal of Education and Learning, 6*(2), 155–170. https://doi.org/10.5539/jel.v6n2p155

Veerbeek, J., Vogelaar, B., Verhaegh, J., & Resing, W. C. M. (2019). Process assessment in dynamic testing using electronic tangibles. *Journal of Computer Assisted Learning, 35*(1), 127–142. https://doi.org/10.1111/jcal.12318

Vogelaar, B., Veerbeek, J., Splinter, S. E., & Resing, W. C. M. (2021). Computerized dynamic testing of children's potential for reasoning by analogy: The role of executive functioning. *Journal of Computer Assisted Learning, 37*, 632–644. https://doi.org/10.1111/jcal.12512

Wu, H. M., Kuo, B. C., & Wang, S. C. (2017). Computerized dynamic adaptive tests with immediately individualized feedback for primary school mathematics learning. *Educational Technology and Society, 20*(1), 61–72.

History
Received January 21, 2022
Revision received July 6, 2022
Accepted October 3, 2022
Published online March 21, 2023
EJPA Section / Category Educational Psychology

Publication Ethics
The study was approved by the ethical board of the Psychology Institute and informed consent was obtained for all children prior to testing.

Open Science
Open Data: For further inquiries about the data and materials, please contact the lead author. Data and materials can be made available upon reasoned request.
Preregistration of Studies and Analysis Plan: This study has no preregistration.

ORCID
Jochanan Veerbeek
https://orcid.org/0000-0001-9280-9066

Jochanan Veerbeek
Education and Child Studies
Leiden University
Wassenaarseweg 52
2333 AK Leiden
The Netherlands
j.veerbeek@fsw.leidenuniv.nl

