Intelligence in Action – Effective Strategic Behaviors while Solving Complex Problems

Christin Lotz[a], Ronny Scherer[b], Samuel Greiff[c], & Jörn R. Sparfeldt[a]

[a] Saarland University, Department of Educational Science, Campus A5 4, D-66123 Saarbrücken, Germany

[b] Centre for Educational Measurement at the University of Oslo (CEMO), Faculty of Educational Sciences, Postbox 1161 Blindern, N-0318 Oslo, Norway

[c] University of Luxembourg, Education, Culture, Cognition, & Society unit, 4366 Esch-sur-Alzette, Luxembourg

Author Note

Correspondence concerning this article should be addressed to Christin Lotz, Department of Educational Science, Saarland University, Campus A5 4, D-66123 Saarbrücken, Germany. Email: c.lotz@mx.uni-saarland.de

We are particularly grateful to the contribution of Anna Auth, Elena Groh, and Kerstin Mayer, who participated in an excellent manner on the data collection. Furthermore, we express our gratitude to Sascha Ludwig who supported the analyses by extracting the relevant data from the log files. We are also grateful to the TBA group at DIPF (http://tba.dipf.de) for providing the authoring tool CBA Item Builder and technical support. Samuel Greiff is one of two authors of the commercially available COMPRO-test that is based on the multiple complex systems approach and that employs the same assessment principle as MicroDYN. For any research and educational purpose, a free version of MicroDYN is available.

Intelligence in Action – Effective Strategic Behaviors while Solving Complex Problems

Problem solving as well as reasoning are central aspects of human intelligence and intelligence tests are an excellent way of capturing these aspects (cf. Deary, 2012; Hunt, 2011). However, conventional intelligence tests often provide data only about the students' overall performance, but no process data about the students' behavior while solving the problems. The recently emerged construct of complex problem solving (CPS) may have the potential to complement conventional intelligence tests by supplying data about every single action made by a test taker over the course of the problem-solving process.

From a conceptual perspective, CPS reveals a high theoretical and empirical overlap with intelligence: The term 'problem solving' is an essential part of many intelligence definitions (e.g., Gottfredson, 1997). In addition, intelligence and CPS correlate substantially with each other (Stadler, Becker, Gödker, Leutner, & Greiff, 2015) and comparably high with external criteria such as educational success (Lotz, Sparfeldt, & Greiff, 2016). Furthermore, the computer-based assessment of CPS has the advantage of supplying process data, which could give direct and detailed information about how a problem solver interacts with the task (Csapó, Ainley, Bennett, Latour, & Law, 2012) and, moreover, how intelligence might facilitate successful exploration behavior. In past research, this potential was repeatedly praised but, nevertheless, only seldom implemented. Prior studies mainly utilized the potential of the CPS process measures by examining time-on-task (e.g., Scherer, Greiff, & Hautamäki, 2015) or by investigating rather specific strategies that were strongly dependent on the particular CPS test (e.g., Güss, Tuason, & Orduña, 2015; Strohschneider & Güss, 1999). Other studies that examined the usage of more generalized strategic behaviors primarily analyzed its relation to CPS performance either only for one single task (Greiff, Wüstenberg, & Avvisati, 2015) or averaged across all tasks (Greiff, Niepel, Scherer, & Martin, 2016; Wüstenberg, Stadler, Hautamäki, & Greiff, 2014). Thus, how problem solvers apply and adapt universal and domain-general strategic behaviors across a set of several CPS tasks and

how they might change their strategic behaviors across changing tasks is still requiring further examination.

Specifically, while exploring several independent CPS tasks as usually employed in the multiple complex system approach (Greiff, Wüstenberg, & Funke, 2012), students have to identify different effect types that pose different demands upon them as problem solvers (Hundertmark, Holt, Fischer, Said, & Fischer, 2015). Thus, different exploration strategies are optimal. On the one hand, the vary-one-thing-at-a-time strategy (VOTAT; Tschirgi, 1980) as the domain-general core strategy in scientific reasoning is suitable for discovering non-dynamic effects because it singles out effects of each problem element. On the other hand, the vary no-thing-at-a-time strategy (NOTAT, also known as non-interfering observation; Greiff et al., 2016) is optimal for exploring dynamic effects because the problem solver can observe how the dynamic system develops by itself. Successful problem solvers are characterized by applying VOTAT and NOTAT (Greiff et al., 2016; Kröner, Plass, & Leutner, 2005) and by increasing their proficient strategic behavior while working on a CPS test (Güss et al., 2015). Applying and flexibly adapting VOTAT and NOTAT when faced with different task types are reasonable strategic behaviors and, thus, should be related to the underlying levels of intelligence.

Therefore, this study examined in a first step how students applied and adapted the two domain-general strategic behaviors VOTAT and NOTAT across a set of CPS tasks with different demands by means of discontinuous latent growth curve modeling. Secondly, we investigated how intelligence is manifested in the effective application and adaption of these two optimal strategic behaviors.

1. Introduction

1.1. Complex Problem Solving

Out of the attempts to conceptualize CPS as a psychological construct, Buchner (in Frensch & Funke, 1995) provides one of the most specific definitions to describe complex problem solving. He defines CPS as:

> "The successful interaction with task environments that are dynamic (i.e., change as a function of user's intervention and/or as a function of time) and in which some, if not all, of the environment's regularities can only be revealed by successful exploration and integration of the information gained in that process" (p. 14).

Buchner's definition stresses the process-character of CPS as well as the role that strategies and interactions with the problem-solving environment play when engaging in complex problem solving. In fact, one of the defining characteristics of CPS is that action sequences are combined in a way that they work towards the resolution of a cognitively challenging and novel situation (Greiff, Wüstenberg, & Avvisati, 2015). To some extent, a perspective that focuses on the problem solving process is also found in definitions of intelligence (cf. Gottfredson, 1997), but the perspective that focuses on the problem solving process is more pronounced and has been considered a key to success in research on CPS.

CPS is often broken down into the two broad processes of knowledge acquisition and knowledge application (Wüstenberg, Greiff, & Funke, 2012). Consequently, in many CPS tests, final performance scores as well as process data on both processes are derived separately. Knowledge acquisition refers to the process of gathering knowledge about a non-transparent and potentially dynamically changing system and includes the exploration of the problem space through targeted interactions between problem solver and problem situation (Funke, 2001). Knowledge application refers to the process of applying the previously gathered knowledge to reach a given target state or, put simply, to solve the problem (Novick & Bassok, 2005). It is readily acknowledged that in real life both processes take place intermittently; for instance, when working towards the solution of a problem (i.e., knowledge

application), the problem solver might realize that there are some components of the problem structure that need further discovery and go back to re-explore the problem space (i.e., knowledge acquisition). However, when employing CPS tests, it might make sense to separate the two processes by first instructing participants to gather information and knowledge about a problem (i.e., knowledge acquisition) and to subsequently work towards a goal state (i.e., knowledge application; Wüstenberg et al., 2012).

Most current CPS assessments are based on multiple complex systems, so that participants work on several CPS tasks that contain the key features of CPS definitions. For instance, CPS tasks are all non-transparent at the outset and require a sequence of active interventions to successfully solve them. Due to their interactivity, the mode of administration for CPS tests is always computer-based. In the linear structural equation framework (Funke, 2001), the CPS tasks contain a set of input variables and a set of output variables that are related to each other; these relations are defined by linear structural equations. In addition, variables might change by themselves independent of other variables. These effects are often not transparent to the problem solver. Along the theoretical conceptualization of CPS, participants are asked to first gather information about the underlying problem structure for each task through exploration (i.e., knowledge acquisition) and then work toward a predefined solution of the problem (i.e., knowledge application).

When working on modern CPS tests, the underlying problem structure is varied across the tasks (e.g., MicroDYN, Greiff et al., 2012). Specifically, the number of input and output variables as well as the number and type of relations between them can be varied. In terms of the relations' type, direct and indirect effects can be distinguished (see Fig. 1). Direct effects, also referred to as non-dynamic effects, represent the relations between input and output variables. Indirect effects, also referred to as dynamic effects, represent the relations of output variables among themselves, which lead to dynamic changes of the problem situation irrespective of external interventions (Greiff, Fischer, Stadler, & Wüstenberg, 2015). Due to

their non-transparent nature, sequences of strategic behaviors are a prerequisite for effectively gathering information about the problem space and, subsequently, mastering CPS tasks. Thus, a theory-driven definition and analysis of the strategic behaviors during the exploration phase might help understand how problem solvers interact with the problem.

---Please insert Figure 1 about here---

Regarding the relation between CPS performance measures and intelligence, early studies on CPS showed only small connections between both constructs and viewed them as largely separate from one another. However, according to current research, CPS is mainly understood as a cognitive construct that is closely related to intelligence (Stadler et al., 2015); especially, if intelligence is broadly operationalized (Kretzschmar, Neubert, Wüstenberg, & Greiff, 2016; Kröner et al., 2005; Lotz et al., 2016). Beyond the major proportion of variance that CPS has in common with intelligence, CPS might exhibit some unique aspects. Next to the aspects of dynamics and interactivity on a conceptual level, there are also empirical findings that support the assumption of uniqueness. However, most studies claiming that CPS incrementally explained variance in educational contexts (e.g., Wüstenberg et al., 2012) or work settings (e.g., Danner, Hagemann, Schankin, Hager, & Funke, 2011) beyond intelligence relied on a rather narrow operationalization of intelligence (mostly only figural reasoning). Studies assessing broader measures of *g* and, thus, providing more reliable evidence for this claim, are still scarce. Nonetheless, their results indicated that a broader operationalization of intelligence led to higher correlations between intelligence and CPS as well as to lower increments of CPS beyond intelligence (Kretzschmar et al., 2016; Lotz et al., 2016). Besides the substantial relations between intelligence and CPS *performance* measures, the relations between broadly operationalized intelligence and CPS *process* measures such as strategic exploration behaviors have not yet been examined sufficiently. Thus, their inspection is a main research goal of the present study.

1.2. Strategic behaviors while solving complex problems

VOTAT. The optimal way to identify non-dynamic, direct effects during the knowledge acquisition phase of a CPS task is logical disconfirmation: Variables are systematically manipulated to produce conclusive tests (Inhelder & Piaget, 1958). This hypothesis testing strategy requires varying systematically one variable at a time while all other variables remain constant (VOTAT; Tschirgi, 1980; also known as "control of variables strategy" [CVS], Chen & Klahr, 1999). Thus, the problem solver can single out or isolate the one effective variable that is responsible for a particular direct effect.

Varying one thing at a time is regarded as a core strategy in scientific reasoning, which strongly resembles the knowledge acquisition phase in CPS tasks. The significance of VOTAT is founded in its domain-general applicability that was shown across many experimental studies. Different scientific domains were covered, such as determining the effects of different variables on how far a spring stretches (Chen & Klahr, 1999), finding out which factors influence the tilt of a balance apparatus (Zimmerman, Raghavan, & Sartoris, 2003), ascertaining the features that make TV-programs more popular (Kuhn, Garcia-Mila, Zohar, & Andersen, 1995), or even identifying baking ingredients that made a cake runny (Tschirgi, 1980). Furthermore, when switching from one experimentation problem to another between different domains, children and adults alike, maintained their original frequency of strategy use or even improved it (Kuhn et al., 1995).

In educational contexts, VOTAT's importance was empirically supported by its prediction of learning progress in science education over a three year period even after controlling for intelligence (Bryant, Nunes, Hillier, Gilroy, & Barros, 2015). Nevertheless, although VOTAT is fundamental for (school) science, most students have no generalized understanding of it because it does not routinely develop without practice (Schwichow, Croker, Zimmerman, Höffler, & Härtig, 2016; Zimmerman & Croker, 2013). This was shown, for example, by Schauble (1996) who examined fifth- or sixth-graders and adults (each N =

20) who worked on multi-trial scientific problems. An inspection of the graph (Schauble, 1996, p. 108) revealed that only about 25 % of the children's inferences and about 60 % of the adults' inferences were valid and based on applying VOTAT during the first of six sessions, whereas by the last session nearly 60 % of the children's and 85 % of the adults' inferences were valid. Moreover, Schauble concluded that neither adults nor children simply abandoned invalid strategies once they discovered valid ones, but used a mixture of former invalid and newer valid strategies in which the usage of the new optimal strategy was progressively increased. Another experiment conducted with a sample of $N = 36$ university students revealed comparable results (Vollmeyer, Burns, & Holyoak, 1996). Over the course of a four-task learning phase of a CPS test, increasingly more students recognized VOTAT as an optimal strategic behavior to obtain knowledge about the relations among variables within the problem space. This was indicated by an increase in the percentage of students using VOTAT (i.e., from 19 % of the students in the first task to up to 56 % in the fourth task) and a decrease of the percentage of students applying unsystematic strategic behavior such as changing all variables at the same time (i.e., from 67 % of the students in the first task to 22 % in the fourth task). Chen and Klahr (1999) identified a similar development for a sample of seven- to ten-year-old children ($N = 87$) who worked on scientific problems. Thus, among children, university students, and adults, a development or learning of VOTAT may occur within a sequence of problems, whereas adults seem to start with a higher percentage of VOTAT-use than children.

In the context of CPS, the application of VOTAT leads to better CPS performance as it is regarded as an effective strategic behavior for exploring the problem space (Greiff et al., 2015; Greiff et al., 2016; Kröner et al., 2005). For example, the study by Kröner et al. (2005), which examined a sample of $N = 101$ high school students from grades 9 to 12, revealed that students who explored the problem space of the CPS test *MultiFlux* (Kröner, 2001) more effectively (i.e., used VOTAT) were more successful in knowledge acquisition ($r = .47$) and

in knowledge application (r = .40). Accordingly, Greiff et al. (2016) investigated a sample of N = 1,476 high school 9th graders and showed that the VOTAT application mean score, averaged across nine MicroDYN tasks during the exploration phase, predicted knowledge acquisition (β = .55) as well as knowledge application (β = .56).

To conclude, VOTAT is a domain-general core strategy for scientific reasoning, and students' use of this strategy seems to progressively increase with practice across a set of problem tasks. Moreover, applying the principle of isolated variation to complex problems is an effective way to explore non-dynamic, direct effects within the problem space.

NOTAT. Besides VOTAT's efficiency for exploring direct effects, there are scenarios where other kinds of effects have to be identified. For example, "eigendynamics" such as growth or decay effects cause changes in the variables independent of the problem solvers manipulations and lead to autonomous changes of the situation without any intervention on the part of the problem solver (i.e., indirect effects). To detect such effects, VOTAT is not an optimal strategic behavior and, thus, the application of another strategic behavior such as systematically constraining all variables to simultaneously remain at a zero level is more appropriate (Funke, 2001). By doing so, the problem solver could actively observe how the system is changing itself without any interference (also referred to as non-interfering observations, Greiff et al., 2016). In accordance with the VOTAT abbreviation (vary-one-thing-at-a-time) non-interfering observations could also be understood in the sense of varying no-thing-at-a-time and, thus, be abbreviated as NOTAT.

In CPS contexts, most problem solvers struggle with resisting the temptation to manipulate the variables immediately and typically act too quickly. In contrast, proficient problem solvers who monitor the autonomously developing system tend to be more successful in solving dynamic CPS tasks (Dörner, 1980; Dörner & Schaub, 1994). Accordingly, a recent empirical study showed NOTAT's relevance for solving complex problems (Greiff et al., 2016): Although VOTAT and NOTAT correlated substantially (r = .35), NOTAT (mean score

averaged across nine MicroDYN tasks) showed unique effects on CPS performance even after controlling for VOTAT (knowledge acquisition: $\beta = .11$; knowledge application: $\beta = .08$). It should be noted that the authors did not differentiate between non-dynamic tasks (that contained only direct effects, for which NOTAT was not an effective strategic behavior) and dynamic tasks. Thus, considering the different task types might reveal a higher importance of NOTAT for dynamic tasks and, thereby, a deeper and more conclusive understanding of effective strategic behaviors while solving CPS tasks that involve dynamic effects.

1.3. The role of intelligence in applying and adapting strategic CPS behaviors

As mentioned above, a positive and substantial relation between students' intelligence and overall CPS performance has been established (Lotz et al., 2016; Stadler et al., 2015). Whereas this relation provides a rather *general* perspective of the role played by intelligence in CPS that primarily focuses on the overall CPS performance, the question arises to what extent intelligence and *specific* CPS behaviors are linked, as manifested in the application and adaptation of effective strategies. A study focusing on different facets of CPS, using the *MultiFlux* CPS test, provided an initial answer: Rule identification, understood as the ability to employ VOTAT as an effective strategic behavior to explore the relations among variables within a complex system, was positively related to reasoning ($r = .41$; Kröner et al., 2005). More specifically, those who scored higher on the reasoning test applied VOTAT more frequently. Furthermore, a large-scale study with $N = 3{,}191$ Finnish students in grades 6 and 9 showed a strong association between the use of VOTAT and reasoning ($r = .64$; Wüstenberg et al., 2014). Considering their finding, the authors concluded that "high fluid intelligence would then help students figure out the correct behavior in CPS tasks even if they did not have any prior knowledge about VOTAT" (p. 132). These findings suggest that there is a link between intelligence and the application of VOTAT.

Moreover, the positive relation between intelligence and VOTAT has been replicated in a number of studies in the domain of scientific reasoning (van der Graaf, Segers, & Verhoeven, 2015: $.42 \leq r \leq .47$; Künsting, Kempf, & Wirth, 2013: $r = .30$; Veenman, Bavelaar, De Wolf, & Van Haaren, 2014: $r = .17$; see Table 1 for an overview of studies concerning the VOTAT-intelligence-relations). Veenman, Wilhelm, and Beishuizen (2004) interpreted this link from a theoretical perspective on intelligence and argued that the ability to select and apply an effective problem solving behavior such as VOTAT is "an integral part of the intellectual toolbox" and should, therefore, be considered "a manifestation of intellectual ability" (p. 91).

---Please insert Table 1 about here---

As an extension of this argument, not only the selection and application of effective strategic behaviors such as VOTAT, but also the adaptation of strategic behaviors when confronted with changing task demands and, therefore, the ability to learn strategic behaviors may be considered an aspect of intelligence. In fact, in one of his early works on the conceptualization of intelligence, Thorndike (1922) proposed that "estimates of it [intelligence] are, or at least should be, estimates of the ability to learn. To be able to learn harder things, or to be able to learn the same things more quickly, would then be the single basis of evaluation" (pp. 17-18). Guthke and Stein (1996) adopted this proposal and conducted a study in which $N = 40$ participants worked on a CPS test, a figural reasoning test and a learning test; the latter comprised adaptive analogy items. The authors showed that the performances on learning and intelligence tests were highly related with correlations up to $r = .83$. The ability to learn was, in turn, positively related to the performance on CPS tests, as indicated by knowledge acquisition ($r = .36$) and application ($r = .50$). LePine, Colquitt, and Erez (2000) confirmed that intelligence and the ability to learn were significantly correlated; moreover, those with higher levels of intelligence were able to adapt their problem-solving strategy to changing situations more effectively. Deák (2003) concluded from existing

research findings that responding to changing task demands or instructions and, therefore, adapting effective problem solving strategies defines "flexible cognition" as a part of intelligence. Further empirical evidence for this claim showed that adapting and incorporating new strategies were positively related to fluid intelligence (Benedek, Jauk, Sommer, Arendasy, & Neubauer, 2014). Taken together, these findings suggest that the ability to adapt an effective problem solving behavior such as VOTAT may be positively related to intelligence. In other words, recognizing that a strategic behavior is effective to solve a given problem and, therefore, applying and adapting it indicates intelligence. Thus, an examination of this relation in the context of CPS by analyzing computer-generated log files might be very fruitful.

1.4. Log file analyses in CPS research

Administering computer-based CPS tests has the advantage of not only obtaining final outcome scores, but one can also discover the problem solvers steps towards the specific outcome. Revealing insights in applied strategies and tactics or committed errors might help understanding the problem-solving process. Concerning previous research on problem-solving process data, measures such as time-on-task were analyzed and seem to have a positive (e.g., Goldhammer et al., 2014; Scherer et al., 2015) or a reverse-U-shaped relation with performance (Greiff et al., 2016). However, in the area of examining strategic CPS behavior besides time-on-task, only a few studies have been conducted that reveal insights in students' actions while working on different CPS tests (Greiff et al., 2015; Greiff et al., 2016; Güss et al., 2015; Strohschneider & Güss, 1999; Wüstenberg et al., 2014).

For example, Strohschneider and Güss (1999) analyzed the strategic behavior of German and Indian (n = 34 each) university students while working on the CPS test *Moro* (providing developmental aid to a small African semi-nomadic tribe; Dörner, Stäudel, & Strohschneider, 1986). In this case, the authors focused mainly on comparing rather program-

specific features (e.g., number of alarm messages, number of questions posed to the system, or collision of two decisions) between the two cultures. More recently, Güss et al. (2015) administrated to $N = 130$ university students the CPS test *WINFIRE* (protecting a city from approaching fires; Gerdes, Dörner, & Pfeiffer, 1993) and, once again, examined program-specific indicators of strategies, tactics, and errors (e.g., number of helicopters sent to a fire that just started, use of patrol command, or number of trucks sent to a burned field). Nevertheless, the authors concluded that the participants' proficiency increased while working on the task (increase of good planning strategies and tactics: $.06 < \eta^2 < .14$; decrease of errors: $.04 < \eta^2 < .14$) and that the flexible adaption and shifting of strategies are important factors for successful performance. In sum, both studies analyzed rather program-specific behaviors; yet, an analysis of more universal strategic behavior (e.g., VOTAT) within larger samples might yield more general conclusions.

Accordingly, the mentioned afore study by Wüstenberg et al. (2014) analyzed the relations between the VOTAT behavior and CPS performance (assessed by MicroDYN) of $N = 3,191$ high school students. In this study, VOTAT was scored dichotomously and credit was given if students applied VOTAT for each input variable at least once. The results showed that the VOTAT application (averaged across all tasks) was mainly predicted by reasoning ability ($\beta = .56$). Moreover, VOTAT predicted the overall CPS performance ($\beta = .88$) and partially mediated the relation between the predictor reasoning ability on the criteria CPS performance (correlation between fluid intelligence and CPS performance: $r = .69$; mediation via CPS strategy: $\beta = .49$) with an additional direct effect of reasoning on CPS performance ($\beta = .11$). The authors concluded that reasoning mostly influenced CPS performance via its effect on applying VOTAT. In another study, using the MicroDYN approach, Greiff et al. (2015) analyzed the process data of one CPS task of $N = 16,219$ high school students who participated in the 2012 cycle of the Programme for International Student Assessment (PISA; OECD, 2014). The relations between the (again) dichotomous VOTAT application in one

single task and (a) CPS performance in this specific task (r = .67) as well as (b) the overall CPS performance in all tasks (r = .61) were strong. Taken together, there is only a small number of studies that analyzed strategic CPS behavior by investigating domain-general strategies such as VOTAT. These few studies, however, provided first indications that using VOTAT is an important exploration behavior.

1.5. The present investigation

The above-mentioned studies analyzed the strategic VOTAT or NOTAT behaviors by either focusing on only a single CPS task or averaging the strategy indicators across a set of tasks, not distinguishing between tasks with direct or dynamic effects. Moreover, the VOTAT or NOTAT indicators were mostly scored dichotomously per task by giving credit whenever the particular strategy was applied at least once for all input variables (cf. Greiff et al., 2015; Greiff et al., 2016; Wüstenberg et al., 2014). In contrast, this study (a) used a continuous scoring of the VOTAT and NOTAT indicators in the manner of relative frequencies per task (cf. Kröner et al., 2005), (b) analyzed every task within a set of CPS tasks, and (c) distinguished between tasks with direct and indirect effects. This might establish good prerequisites to display the temporal course of the strategic behaviors across the task set and facilitates the understanding of students' flexibility in adapting effective strategies when confronted with changes in the task type.

This study used the MicroDYN approach to examine students' strategic behaviors. The first five from a total of nine tasks comprised only direct effects, which could be explored most effectively by applying VOTAT. At the start of the remaining tasks (after task 5), students received another instruction, introducing dynamic effects. For the remaining tasks, students were confronted with the possible occurrences of dynamic effects in addition to direct effects. These tasks could be explored most effectively by applying NOTAT in addition to VOTAT. We expected the change in the task type to affect the application frequencies of

both strategic behaviors and, therefore, analyzed each strategy course separately across the CPS task set. Latent growth curve models were specified with discontinuous growth curves (see e.g., Diallo & Morin, 2015), representing the phases before the change in task type (task 1 – 5) and after (task 6 – 9). It is a main goal of this study to further understand the role of intelligence in the application and adaption of VOTAT and NOTAT while exploring unknown systems with different task demands. Thus, we examined the relations of intelligence with the relative application frequency and the adaption gradients of both strategic behaviors.

Hypothesis (1) examined the course of the relative VOTAT frequencies across the nine CPS tasks and its relations with intelligence. Regarding the VOTAT course, we expected a progressive increase of its relative frequency across the first five tasks with only direct effects (Hypothesis 1a). From the sixth task on, dynamic effects were introduced and could occur in addition to the direct effects. Thus, the application of NOTAT along with VOTAT seems to be optimal. Because relative frequencies were being used, a higher proportion of NOTAT might go along with a lower proportion of VOTAT. Thus, we expected a significant drop in the relative VOTAT frequency from the fifth to the sixth task (Hypothesis 1b). Across the last four tasks (tasks 6 – 9), we expected an increase in the relative VOTAT frequency (Hypothesis 1c) because the application of VOTAT was still important for exploring the direct effects across these tasks.

Regarding the VOTAT-intelligence-relations (Hypothesis 1d) as one of the key aspects of this study, we expected substantial and positive correlations between intelligence and the application levels (relative frequency) as well as the adaption gradients (changes in the relative frequency) of VOTAT for both task types because higher levels and steeper increases in the relative VOTAT frequencies indicate optimal and effective strategic behaviors for solving both task types.

Hypothesis (2) examined the course of the relative NOTAT frequencies and its relations with intelligence. Regarding the NOTAT course, its application across the first five non-dynamic tasks was not constructive because no dynamic effects occurred. Therefore, the relative NOTAT frequency should not have increased (i.e., remained constant on a low level or even decreased; Hypothesis 2a). After the change in the task type, the application of NOTAT was appropriate because dynamic effects might be apparent and could effectively be detected by using this strategic behavior. Thus, we expected a significant rise in the relative NOTAT frequency from the fifth to the sixth task (Hypothesis 2b). From the sixth to the ninth task, we expected an increase of the relative NOTAT frequencies because students should have experienced the benefits of NOTAT to detect the dynamic effects and apply it more often across these tasks (Hypothesis 2c).

Regarding the NOTAT-intelligence-relations (Hypothesis 2d) as another key aspect of this study, we assumed a substantial negative correlation between the level of the relative NOTAT frequency and intelligence across the non-dynamic tasks (tasks 1 – 5) because for these tasks, the application of NOTAT was not effective. Regarding the NOTAT adaption gradient across the non-dynamic tasks, we had no clear expectations. Across the dynamic tasks (tasks 6 – 9), we expected positive correlations between intelligence and the level as well as the gradient of the relative NOTAT frequencies because a higher level and an increasing gradient were effective strategic behaviors to detect the dynamic effects.

2. Method

2.1. Sample and procedure

The sample consisted of $N = 495$[1] German high school students ($n = 264$ females, $n = 228$ males, $n = 3$ without gender specification; age $M = 16.40$, $SD = 0.94$ years). The students

[1] Please note that the intelligence and CPS performance data of this sample have been used in a previous study (Lotz et al., 2016). However, analyzing the CPS process data and distinguishing between VOTAT and NOTAT

attended two academic-tracked school types from either 10th grade (Gymnasium, graduation after 12th grade; 12 classes out of 5 schools) or 11th grade (Gesamtschule, graduation after 13th grade; 14 classes out of 6 schools). The participation rate was 87%; the parents of 11% of the high school students did not allow their children to participate; 2% of the students were absent due to reasons unrelated to the study (e.g., illness). Students' and their parents' informed consent was obtained prior to testing. Students were not graded or rewarded in any way. Trained experimenters administered the assessments within three consecutive regular school lessons of 45 minutes each. During one lesson, students worked on intelligence subtests. Given the limited availability of computers for the computer-based CPS assessment, classes were randomly split into two halves at the beginning of the remaining two lessons: In one lesson, the first half worked on the CPS assessment, while the second half executed tasks irrelevant to the study. In the remaining lesson, the second half of the students completed the CPS assessment while the first half worked on other tasks.

2.2. Variables

Intelligence. Intelligence was assessed by the in Germany well-known and widely used Berlin Intelligence Structure test – Form 4 (BIS-4; Jäger, Süß, & Beauducel, 1997) that is based on the Berlin Model of Intelligence Structure (BIS; cf. Süβ, Oberauer, Wittmann, Wilhelm, & Schulze, 2002). The BIS-model postulates besides a general factor at the apex several ability components that belong to the two facets *contents* (verbal, numerical, and figural) and *operations* (processing capacity/reasoning, memory, speed, and creativity). On the basis of prior studies (Brunner & Süß, 2005; Valerius & Sparfeldt, 2014), we selected 10 subtests that covered a wide range of the BIS' content-operation-combinations (see Table 3). For example, the subtest 'verbal analogies' required the identification of analogous word pairs (content: verbal; operation: reasoning capacity), the subtest 'number sequences' the completion of

---

is entirely unique to this study. The different sample size ($N = 495$ instead of $N = 496$) resulted from one student who did not work on the CPS tasks (MicroDYN), but on all other variables analyzed in Lotz et al. (2016).

numbers in a series (content: numerical; operation: reasoning capacity), and the subtest 'city map' the memorization and recall of buildings in a city map (content: figural; operation: memory).

Complex Problem Solving. CPS was assessed by the entirely computer-based micro-world program MicroDYN[2] (Greiff et al., 2012) that consisted of nine fully independent CPS tasks. The time that participants worked on each task was limited to about 5 minutes (knowledge acquisition: 180 seconds; knowledge application: 90 seconds). Different cover stories such as feeding a cat, training a handball team, or providing medical aid were implemented for each task. Variables were labeled without deep semantic meaning or entirely fictitious to minimize the influence of prior knowledge (e.g., in the task "medical aid" different pharmaceuticals were labeled fictitiously as "Sarol", "Rexol", and "Menol"; see Fig. 1). Participants were required to explore the unknown system with a set of related input and output variables in order to first develop a mental model about the problem space (i.e. knowledge acquisition). Subsequently, students had to control the system by reaching given target values (i.e. knowledge application; see Greiff et al., 2012 for a more detailed description of MicroDYN). In this study, we focused on the knowledge acquisition phase because our interest was on how students were using different exploration strategies to develop a mental model about the problem space (e.g., Klahr & Dunbar, 1988; Lee, Jonassen, & Teo, 2011). In fact, the development of a mental model that represents students' knowledge about the relations among variables in the problem space is critical to problem solving success, because it forms the basis for subsequent steps of applying this knowledge to generate a solution of the problem (Goode & Beckmann, 2010; Sonnleitner, Keller, Martin, & Brunner, 2013).

During the instructional part of MicroDYN, students watched a video to familiarize themselves with the user interface and they were instructed to "find out about the relations" between input and output variables. During the main part of MicroDYN, students were

[2] A task example of MicroDYN can be found at: https://www.youtube.com/watch?v=sSKVFF6XE6g

confronted with non-dynamic and dynamic effects. In this study, the first five tasks contained only non-dynamic effects. After the fifth task, dynamic effects were introduced by an additional instruction. Therefore, students were aware of the potential occurrence of dynamic effects. In this study, three of the four remaining tasks after this change in task type had dynamic effects next to the non-dynamic effects (task 7: only non-dynamic effects). However, the students did not know in which of the remaining tasks dynamic effects would occur. During the main part of MicroDYN, students could revisit the general instruction again at any time by clicking the "help" button. It should be noted that neither hints about optimal strategies nor correct solutions were given thereby.

During the knowledge acquisition phase of each task, students freely explored the unknown task systems, which were based on linear equations that related a maximum of three input variables to a maximum of three output variables (non-dynamic effects). For example, in the task "medical aid" (see Fig. 1) students had to explore how different pharmaceuticals (input variables: "Sarol", "Rexol", and "Menol"; left part of Fig. 1) influenced some characteristics of human health (output variables: "Headache", "Diastolic blood pressure", and "Antibodies"; right part of Fig. 1). Regarding the dynamic effects, students had to consider that the output variables might change without manipulating the input variables. For example, "Headache" could increase over time by itself. To explore the underlying system, students manipulated the input variables by moving sliders from "0" to the right ("+" or "++") or to the left ("–" or "– –"), and clicked the "apply" button.

Whenever students clicked the "apply" button during the knowledge acquisition phase, their actions were documented in computer-generated log files. To score VOTAT, each exploration step was evaluated (i.e., the manipulation or action taken by the student before clicking "apply"). Credit was given (coded as 1), if only one input variable was varied and the other input variables were held at zero (cf. Kröner et al., 2005); otherwise, no credit was given (coded as 0). For each MicroDYN task, the relative VOTAT frequency was computed by

summing-up the number of explorations in which students used VOTAT and dividing it by the students' total number of exploration steps in this task. Thus, the relative VOTAT frequency ranged between 0 (i.e., the student never used VOTAT during the exploration of a particular task) and 1 (i.e., the student always used VOTAT during the exploration of a particular task). To score NOTAT, each exploration step was evaluated and credit was given (coded as 1), if all input variables were held zero when the student clicked "apply"; otherwise, no credit was given (coded as 0). In accordance to the relative VOTAT-frequency, the computed relative NOTAT-frequency could range between 0 (i.e., the student never used NOTAT) and 1 (i.e., the student always used NOTAT) in each task.

2.3. Analyses

All analyses were conducted with the statistical package M*plus* 7.11 (Muthén & Muthén, 1998-2013). We controlled for potential effects of students' clustering in classrooms on standard errors and χ^2 statistics by using the "type = complex" option (Maximum Likelihood Robust [MLR] estimator). Only few missing values occurred (maximum per variable: 4.6 %) and were handled by the full information maximum likelihood (FIML) algorithm (Enders & Bandalos, 2001). To describe the model fit, χ^2 with *df* were complemented by CFI, TLI, and RMSEA (Little, 2013; Schermelleh-Engel, Moosbrugger, & Müller, 2003). CFI and TLI values greater than .95 (.90) indicated a good (acceptable) fit. RMSEA values below .05 (.08) signified a close (acceptable) fit to the data.

The hypotheses referring to the relative VOTAT frequencies (Hypotheses 1a-d) and the relative NOTAT frequencies (Hypotheses 2a-d) were examined by latent growth curve models (LGCM; Bollen & Curran, 2006). LGCMs are well-suited to examine the change of a variable over time in a structural equation framework with an intercept factor (initial or reference level) and a slope factor (growth trajectory). Typically, these models estimate smoothed trajectories and, thus, tend to fail when sharp changes are present in the data. In the

current study, the change of the task type from non-dynamic to dynamic and the additional instruction after the fifth task could be regarded as an event that could have caused such a sharp change or a discontinuity in the growth curve, thus affecting the intercept as well as the slope factor. Therefore, a discontinuous LGCM-design (Hancock, Harring, & Lawrence, 2013; also often referred to as 'piecewise LGCM', Diallo & Morin, 2015), including two correlated intercept and slope factors that represent the levels and the growth trajectories before and after the change in the task type seemed especially well-suited to analyze the courses of VOTAT and NOTAT.

Concerning the course of the relative VOTAT frequencies (Hypotheses 1a-c), our model consisted of two LGCM parts that corresponded to the two types of MicroDYN tasks before the change in task type (tasks 1 – 5; only non-dynamic effects) and after (tasks 6 – 9; possibility of dynamic effects, occurring in addition to non-dynamic effects). Regarding the LGCM part before the change in task type, unstandardized factor loadings of the manifest VOTAT indicators of tasks 1 – 5 on the intercept factor before the change in task type (Int1-5) were fixed to 1. To model the latent slope factor before the change in task type (Slope1-5), we assumed a linear growth from the first to the fifth task. Factor loadings had numerical values of $\lambda_1 = -4$ for the first task, $\lambda_2 = -3$ for the second task up to $\lambda_5 = 0$ for the fifth task. Thus, the mean score of Int1-5 represented the relative VOTAT frequency of the fifth task (directly before the change in task type). Concerning the LGCM part after the change in task type, the intercept factor (Int6-9) was modeled by fixing the unstandardized loadings of the manifest VOTAT-indicators of the tasks 6 – 9 to 1. To model the latent slope factor after the change in task type (Slope6-9), we also assumed a linear growth from the sixth task onwards to the ninth. Factor loadings had numerical values of $\lambda_6 = 0$ for the sixth task up to $\lambda_9 = 3$ for the ninth task. Therefore, the mean score of Int6-9 represented the relative VOTAT frequency of the sixth task (directly after the change in task type). In addition, correlations among the slope and intercept factors were allowed.

However, before investigating the strategic behavior and its relation to intelligence, it is crucial to explore whether the hypothesized discontinuous (vs. continuous) growth curve and the assumed linear (vs. quadratic and vs. no-change) slopes of the above proposed LGCM adequately represented the data structure. The assumption of discontinuous LGCMs with a transition point between the fifth and the sixth CPS task was theoretically driven by the change in task type and the additional instruction after the fifth task. To statistically investigate whether the influence of this event on the growth curve was adequately represented in the model, we computed a continuous LGCM with only one intercept factor and one linear slope factor, both loading on all nine VOTAT indicators. We compared this continuous LGCM to the discontinuous LGCM described above by conducting a Satorra-Bentler corrected χ^2-difference test. If the discontinuous LGCM fitted significantly better than the continuous LGCM, it was the preferable model. Furthermore, we performed additional tests to assure the assumed linear growths of the discontinuous LGCM. For both slopes (before and after the change in task type), intercept-only models (assuming no change over time; i.e., slope = 0) and quadratic slope specifications were modeled. If they revealed significantly worse fit indices compared to the linear slope models (Satorra-Bentler corrected χ^2-difference tests), they could be rejected and, thus, removed from further analyses.

If the assumed discontinuous and linear growth curves fitted the data best, the LGCM could be augmented by an intelligence factor, and all further analyses could be based on this augmented model. Specifically, for modeling the intelligence factor, manifest scores of the 10 intelligence subtests loaded on the corresponding content-facet specific factors of verbal (V_{BIS}), numerical (N_{BIS}), and figural (F_{BIS}) intelligence; and these three first-order content factors indicated the second-order *g*-factor (g_{BIS}) to specify the common higher order structure of intelligence. The unstandardized loadings of the first indicator of a corresponding latent factor were fixed to 1. In accordance with the theoretical BIS-framework (see Jäger et al., 1997), we also considered the operation facets within the BIS-model. Specifically, we

specified residual correlations among those manifest subtest indicators that belong to the same operation facet of the BIS-model [i.e., correlations among the residuals of (a) the six subtests that tapped the reasoning capacity operation facet, (b) the two subtests that tapped the speed operation facet, and (c) the two subtests that tapped the memory operation facet; see Table 3]. The g_{BIS}-factor was correlated with the two LGCM slope factors and the two LGCM intercept factors (before and after the change in task type).

Responding to Hypotheses (1a) and (1c), we inspected the (unstandardized) means and variances of the slope factors. Regarding Hypothesis (1b), the drop in the relative VOTAT frequency after the change in task type was tested by conducting a Satorra-Bentler corrected χ^2-difference test. This test compared the above-specified model to a model with equality constrained means of the intercept factors before and after the change in task type. Moreover, we computed the effect size *h* as a measure of the difference between two proportions (Cohen, 1988). To investigate Hypothesis (1d), we inspected the correlations between the intelligence factor (g_{BIS}), the two slope factors, and the two intercept factors.

Regarding NOTAT (Hypotheses 2a-d), we established an analogous LGCM for NOTAT by replacing the VOTAT indicators of the LGCM of Hypotheses (1a-d) with the NOTAT indicators (relative NOTAT frequencies). Similar to the VOTAT-LGCM, we performed analyses of the structure of the NOTAT growth curves to assure that the proposed discontinuous (vs. continuous) and linear (vs. quadratic and vs. no-change) growth curves fitted the data best. Accordingly, we augmented the hypothesized NOTAT model by the above described intelligence factor and conducted all further analyses with this augmented model. Responding to Hypotheses (2a) and (2c), we inspected the means and variances of the slope factors, as in Hypotheses (1a) and (1c). We then tested the hypothesized rise in the relative NOTAT frequency after the change in task type (Hypothesis 2b) by conducting a Satorra-Bentler corrected χ^2-difference test, comparing the intercept means, and computed the

effect size *h*. To test Hypothesis (2d), we inspected the correlations of the intelligence factor (g_{BIS}) with the two slope and the two intercept factors, as in Hypothesis (1d).

3. Results

Descriptive statistics. Means and standard deviations of the relative VOTAT and NOTAT frequencies across the nine tasks are shown in Table 2; their course is displayed in Fig. 2.[3] Means and standard deviations of the intelligence subtests were always well within the possible range, indicating an absence of bottom or ceiling effects (see Table 3). The manifest correlations of the intelligence subtests and the relative VOTAT and NOTAT frequencies are presented in Table 4 (reported task wise as well as averaged across the two task types: task 1-5 and task 6-9). Regarding VOTAT, correlations with the intelligence subtests were for the most part substantially positive. Regarding NOTAT, correlations with the intelligence subtests were mostly negative across the non-dynamic tasks, but mostly positive across the dynamic tasks as NOTAT became an effective behavior. As expected, VOTAT correlated negatively with NOTAT. It should be noted that these correlations were based on relative frequencies (i.e., a higher proportion of the one might coincide with a lower proportion of the other).

---Please insert Tables 2, 3, and 4 about here---

---Please insert Figure 2 about here---

Results for VOTAT (Hypothesis 1). The assumption of a discontinuous LGCM with a transition point between the fifth and the sixth CPS task was empirically supported by a significantly worse fit of the continuous LGCM (Table 5, Model 1; cont. LGCM without

[3] One anonymous reviewer suggested that high *SD*s might indicate not only high variability, but also different performance patterns. Therefore, we ran supplemental hierarchical cluster analyses that revealed a three-cluster solution for VOTAT (two clusters with increasing VOTAT rates either starting on a low level or on a medium level and a third cluster with a low level VOTAT rate throughout the task set) and a two cluster solution for NOTAT (one cluster was characterized by low NOTAT rates throughout all tasks, the other cluster represented the overall NOTAT course as reported in this manuscript). However, these cluster analytical results should be interpreted with caution (too small sample sizes of the subgroups); consistent with our expectations as well as the results presented below, intelligence correlated substantially with the grouping variable (VOTAT: $r = .35$; NOTAT: $r = .15$).

intelligence) compared to the discontinuous LGCM (Table 5, Model 2; disc. LGCM without intelligence; $p < .05$). Comparing this discontinuous LGCM with entirely linear slopes (Table 5, Model 2) to the intercept-only models with either no slope before the change in task type (Table 5, Model 3a, disc. LGCM without intelligence, no slope before) or no slope after the change in task type (Table 5, Model 3b, disc. LGCM without intelligence, no slope after) revealed significantly worse fit indices ($p < .05$). Furthermore, comparing the discontinuous LGCM with entirely linear slopes (Table 5, Model 2) to models with quadratic slope specifications either before the change in task type (Table 5, Model 3c, disc. LGCM without intelligence, quad. slope before) or after (Table 5, Model 3d, disc. LGCM without intelligence, quad. slope after) revealed both Heywood cases (i.e., correlations > |1|) and were, therefore, rejected from further analyses. Thus, for VOTAT the theoretically proposed discontinuous LGCM with entirely linear slopes was selected for further analyses and augmented by intelligence. This augmented VOTAT-LGCM fitted at least acceptably to the data (Table 5, Model 4, disc. LGCM with intelligence; Fig. 3; standardized loadings of the intelligence subtests are shown in Table 3).

---Please insert Table 5 about here---

---Please insert Figure 3 about here---

Hypothesis (1a) presumed an increase of the relative VOTAT frequency from the first to the fifth task (see Table 2 for descriptive statistics). The (unstandardized) mean and variance of the slope factor before the change in task type were significant (Slope1-5; $M = .05$, $p < .05$; $Var = .01$, $p < .05$). This slope factor mean corresponded to a standardized coefficient (interpreted as the regression trajectory) of $\beta = .66$ and indicated a substantial increase of the relative VOTT frequency across the first five tasks. Hypothesis (1b) assumed a drop of the relative VOTAT frequency from task 5 to task 6 when the task type changed from non-dynamic to dynamic tasks. The means of the two intercept factors Int1-5 ($M = .71$, $p < .05$; $Var = .15$, $p < .05$) and Int6-9 ($M = .63$, $p < .05$; $Var = .11$, $p < .05$) differed significantly

($p < .05$; Table 5, Model 5, disc. LGCM with intelligence difftest intercept means; $h = -.15$, below the cutoff for small effects suggested by Cohen, 1988) and, thereby, supported the hypothesized drop. Hypothesis (1c) presumed an increase of the relative VOTAT frequency from the sixth to the ninth task. The slope factor coefficients after the change in task type revealed only a significant mean but a non-significant variance (Slope6-9; $M = .03$, $p < .05$; $Var < .01$, $p = .23$); this slope factor mean corresponded to a standardized coefficient of $\beta = .78$, indicating the assumed rise. Overall, the students showed increasing VOTAT frequencies across both task types with a drop right after the change in task type.

As a main goal of this study, Hypothesis (1d) examined the relations of intelligence and the course of the relative VOTAT frequencies (as described by the two intercept and slope factors)[4]. Intelligence significantly correlated with both intercept factors ($r_{Int1\text{-}5} = .48$, $p < .05$; $r_{Int6\text{-}9} = .40$, $p < .05$) and with the slope factor before the change in task type ($r_{Slope1\text{-}5} = .21$, $p < .05$). Thus, more intelligent students showed higher relative frequency levels of VOTAT across both task types and a steeper frequency increase across the non-dynamic tasks. However, intelligence did not correlate substantially ($r_{Slope6\text{-}9} = .01$, $p = .95$) with the slope factor after the change in task type.

Results for NOTAT (Hypothesis 2). A comparison of the continuous LGCM (Table 5, Model 6, cont. LGCM without intelligence) with the discontinuous LGCM (Table 5, Model 7, disc. LGCM without intelligence) empirically supported the assumption of a discontinuous growth curve with a breaking point between the fifth and sixth CPS task for NOTAT, as well ($p < .05$). Moreover, the discontinuous model with entirely linear slope specifications (Table 5, Model 7) revealed also for NOTAT significantly better fit indices compared to the intercept-only models with either no slope before the change in task type (Table 5, Model 8a, disc. LGCM without intelligence, no slope before; $p < .05$) or after (Table 5, Model 8b, disc.

[4] Regarding the relations between intelligence and the CPS performance measures, a latent higher-order g_{BIS} factor correlated substantially with the latent CPS performance factors knowledge acquisition ($r = .65$) and knowledge application ($r = .70$; see Lotz et al., 2016 for details).

LGCM without intelligence, no slope after; $p < .05$). Comparably to the VOTAT models, the NOTAT models with quadratic slope specifications either before the change in task type (Table 5, Model 8c, disc. LGCM without intelligence, quad. slope before) or after (Table 5, Model 8d, disc. LGCM without intelligence, quad. slope after) revealed Heywood cases (i.e., correlations > |1|) and were excluded from further analyses. Thus, the theoretically proposed discontinuous NOTAT-LGCM with entirely linear slopes was selected for further analyses and augmented by intelligence. This augmented NOTAT model fitted the data well (Table 5, Model 9, disc. LGCM with intelligence; Fig. 4; see Table 3 for standardized loadings of the intelligence subtests).

---Please insert Figure 4 about here---

Hypothesis (2a) presumed no increase of the relative NOTAT frequency from the first to the fifth task (see Table 2 for descriptive statistics). The (unstandardized) mean of the slope factor before the change in task type revealed to be significant but not the variance (Slope1-5; $M = -.01$, $p < .05$; $Var < .01$, $p = .15$). This negative slope factor mean corresponded to a standardized coefficient of $\beta = -.26$ and indicated a substantial decrease of the relative NOTAT frequency across the first five tasks. Hypothesis (2b) assumed a rise of the relative NOTAT frequency from task 5 to task 6 as the task type changed from non-dynamic to (potentially) dynamic tasks. The means of the two intercept factors Int1-5 ($M = .04$, $p < .05$; $Var = .01$, $p < .05$) and Int6-9 ($M = .10$, $p < .05$; $Var = .02$, $p < .05$) differed significantly ($p < .05$; Table 5, Model 10, disc. LGCM with intelligence difftest intercept means; $h = .24$, small effect size) and, thereby, supporting the hypothesized rise. In contrast to Hypothesis (2a), Hypothesis (2c) presumed an increase of the relative NOTAT frequency from the sixth to the ninth task. The slope factor coefficients after the change in task type revealed a significant mean but a non-significant variance (Slope6-9; $M = .01$, $p < .05$; $Var < .01$, $p = .34$). The positive slope factor mean corresponded to a standardized coefficient of $\beta = .30$, indicating the assumed increase. Overall, students showed a slight decrease in their NOTAT frequencies

across the first five non-dynamic tasks, a substantial rise that corresponded to the change in task type, and a small NOTAT increase across the remaining tasks with potentially dynamic effects.

Addressing another central aim of this study, Hypothesis (2d) examined the relations of intelligence and the course of the relative NOTAT frequencies (as described by the two intercept and slope factors). Intelligence correlated substantially with both intercept factors: negatively with Int1-5 ($r_{Int1\text{-}5} = -.16$, $p < .05$) and positively with Int6-9 ($r_{Int6\text{-}9} = .27$, $p < .05$). Thus, students who scored higher in the BIS applied a lower proportion of NOTAT during the first five tasks, but after the change in task type, they applied higher NOTAT proportions. Furthermore, intelligence correlated non-substantially with the slope factor before the change in task type ($r_{Slope1\text{-}5} = -.14$, $p = .15$) as well as after the change in task type ($r_{Slope6\text{-}9} = -.05$, $p = .77$). Thus, there was no significant relation between intelligence and the NOTAT gradients across both task types.

4. Discussion

The present study focused on the courses of two domain-general and effective strategic behaviors VOTAT and NOTAT across a set of nine CPS tasks with different demands (detecting direct and dynamic effects). Furthermore, our study investigated the extent to which intelligence was related to the application and adaption of the strategic behaviors. Log file-based analyses revealed that students showed higher application rates when a strategic behavior was effective but lower application rates when a behavior was not effective. In addition, students progressively adapted their strategic behaviors across the task set when they were confronted with a change of the task type. Specifically, they showed increasing application gradients when a behavior was effective but a decreasing application gradient when a behavior was not effective. Regarding the relations of the strategic process measures and intelligence, more intelligent students applied higher levels of VOTAT and NOTAT when

the behaviors were effective. Moreover, they also showed steeper gradients when flexibly adapting their strategic behaviors across tasks 1–5. This clearly indicated that intelligence manifested itself in the use of effective strategic behaviors.

4.1. VOTAT's course and relation to intelligence

Our hypotheses concerning the course of VOTAT (Hypotheses 1a-c) were confirmed: First, students gradually increased their relative VOTAT frequency across the first five non-dynamic tasks; subsequently, the VOTAT frequency dropped slightly after the change in task type (from task 5 to task 6), and finally, the relative VOTAT use progressively increased again across the last four tasks with potential dynamic effects. Reviewing the VOTAT rate of the first CPS task in more detail (see Table 2), students applied VOTAT in about 50% of their exploration steps. In previous studies, the VOTAT rate exhibited by students of comparable age who worked on a scientific discovery learning computer program either resembled the students of our sample (group without metacognitive support and nonspecific learning goals: 54%; Künsting et al., 2013) or was lower (group with nonspecific problem solving goals: 29%; group with nonspecific learning goals: 24%; Künsting, Wirth, & Paas, 2011). Thus, a VOTAT rate from about half of all exploration steps could probably be regarded as a rather usual to high application frequency for the very first task when confronted with an unknown problem space. Probably, tenth or eleventh graders might have already been taught VOTAT (implicitly or explicitly) in science education, and their knowledge about how to apply VOTAT and its domain-general utility enabled them to show these moderate to high VOTAT rates. However, one should keep in mind that MicroDYN, as used in our study and other computer programs that were used, for example, in the studies of Künsting and colleagues are not directly comparable. Because this study was the first to operationalize the VOTAT application by relative frequencies during the exploration phase of MircoDYN, no direct comparisons to former studies are possible. To examine how much VOTAT use can be

expected from students of different age groups while solving complex problems remains, therefore, an open question and might be examined in future research.

Second, regarding the VOTAT course across the non-dynamic tasks, the relative frequency gradually increased from 47% in the first task to 71% in the fifth task. Students probably recognized VOTAT as an effective strategy to detect the direct effects between input and output variables, and students' emerging generalized understanding of this strategic behavior revealed its manifestation in the increased application frequency. Furthermore, the gradual increase of VOTAT was in line with previous research, which found that ineffective strategies were not simply abandoned but that students changed their behavior rather slowly but progressively (Schauble, 1996; Vollmeyer et al., 1996). When the task type changed (after task 5) and students were faced with potential dynamic effects in addition to direct effects, the application of NOTAT in addition to VOTAT became an optimal strategic behavior. Because of analyzing relative frequencies, an increase in the application of one strategy, as for example NOTAT, might cause a decrease in the application of another strategy such as VOTAT. Thus, the relative VOTAT frequency dropped after the change in task type, as expected. This connection was also corroborated by the small negative correlation between VOTAT and NOTAT (see Table 4). Although the drop of the VOTAT application after the change in task type was statistically significant, its effect size was very small. This might indicate that students quickly understood that applying VOTAT was still an effective behavior.

Third, the increase of the relative VOTAT frequency across the last four tasks gives further evidence for students' deeper understanding about the utility of this effective strategic behavior. One may argue that the increasing trend might reach a plateau (around 70%) which might represent a reasonable upper bound for one single strategic behavior, leaving some space for trying out some other strategic behaviors which might be effective if an unexpected change in task type occurs. One could imagine tasks in which other relations between input

and output variables appear that might not have been introduced to the students. For example, in the case of interaction effects between the input variables, systematically holding one variable constant and setting its value different from zero and, at the same time, manipulating the other variables (so-called "HOTAT": hold-one-thing-at-a-time; Tschirgi, 1980) would be an effective strategic behavior. Given that many real-life contexts provide students with novel and uncertain situations that might be subjected to unexpected and randomly occurring changes, testing a set of different strategic behaviors such as VOTAT, NOTAT, and HOTAT instead of rigidly adhering to only one single strategy represents an effective strategic behavior as well.

The relations between the VOTAT process measures and intelligence (Hypothesis 1d) were in three out of four cases substantially positive and, thus, indicated the expected direction. Medium correlation coefficients between intelligence and the levels of the relative VOTAT frequencies (intercept before the change in task type: $r = .48$; intercept after the change in task type $r = .40$) were higher than in prior studies (Veenman et al., 2014: $r = .17$) or comparably high (e.g., Kröner et al, 2005: $r = .41$; van der Graaf et al., 2015: $.42 \leq r \leq .47$). Thus, more intelligent students typically applied VOTAT more often than less intelligent students. This consistent result provides further evidence for the link between intelligence and the application of effective strategic problem-solving behavior, although the correlations in this study as well as in the literature do not seem to be large in absolute terms. Thus, it is important to search for additional predictors that account for further variance in strategic behavior. For example, strategy knowledge (metacognitive knowledge about strategies, their applicability, and usefulness) was shown to significantly predict the application of VOTAT after controlling for shared variance with intelligence (Wüstenberg et al., 2014). Because strategic knowledge does not routinely develop without practice, individual differences in strategic knowledge might originate from differences in students' science education or other (formal or informal) learning experiences. Probably, more intelligent students can refer better

to their prior knowledge about VOTAT and, thus, produce conclusive variable manipulations. Furthermore, one might expect for personality variables like conscientiousness that more conscientious students conduct variable manipulations more systematically and, therefore, apply VOTAT more consistently. However, to examine this hypothesis was beyond the scope of the current study.

Concerning the correlation between intelligence and the VOTAT gradient (slope) across the non-dynamic tasks, the coefficient was small, but statistically substantial (r = .21). Thus, it can be concluded that more intelligent students, once they discovered VOTAT as an effective strategic behavior, were able to progress towards using VOTAT more and more frequently over the course of the structurally similar non-dynamic tasks. In line with our expectations and previous literature, higher intelligence was associated with a steeper gradient (Guthke & Stein, 1996; Wüstenberg, et al., 2014). In contrast, the slope-intelligence-correlation across the tasks with potential dynamic effects was near zero. Methodologically, this result could be explained by the very small and non-significant variance of the corresponding VOTAT slope factor that indicates negligible interindividual differences in the VOTAT increase. One might further argue that on the one hand more intelligent students may have reached a plateau of their VOTAT performance, because they started with a generally higher VOTAT level in the non-dynamic tasks. On the other hand, more intelligent students might have used a higher variety of different strategic behaviors (as for example HOTAT) instead of using nonsystematic strategic behavior; they might have been expecting yet another change in task type that has not been introduced to them. However, to clarify whether more intelligent students used other strategic behaviors next to VOTAT and NOTAT remains a topic for future research. Nevertheless, when faced with non-transparent situations in environments that become more and more complex, an application of a variety of systematic strategic behaviors would lead to more success in real-life situations, as more intelligent people have been known to do (cf. Sternberg, 1997).

4.2. NOTAT's course and relation to intelligence

The hypotheses concerning the NOTAT course (Hypotheses 2a-c) were confirmed as well: Across the non-dynamic tasks, students showed a slightly decreasing frequency; subsequently, the NOTAT rate rose after the change in task type and increased across the last four tasks. More specifically, students started applying NOTAT in only 7% of their exploration steps of the first task (Table 2). This comparatively low frequency is not surprising because students were mainly instructed to identify the relations between the input and output variables. Thus, keeping all input variables at zero could be regarded as a rather unsystematic behavior. Across the following non-dynamic tasks, the NOTAT rate decreased slightly, but statistically significant, probably because students realized that this behavior was ineffective for identifying direct effects. However, directly after the change in task type, the relative NOTAT frequency rose to 10% and increased up to 13% across the remaining tasks with potential dynamic effects. Nevertheless, the rise after the change in task type had a small effect size and its following increase was rather minor. Probably students understood that applying rather moderately increasing rates of NOTAT in addition to VOTAT was an optimal and especially efficient strategic behavior to detect the dynamic effects in addition to the direct effects. Nevertheless, because this was the first study to examine the relative NOTAT frequency in addition to VOTAT, it will be the task of future studies to replicate our results.

Regarding the correlations between the NOTAT process measures and intelligence (Hypothesis 2d), our expectations, especially with respect to the slopes, were only partly confirmed. Nevertheless, the correlations between intelligence and the levels of NOTAT (intercepts) were in accordance with our hypothesis. The negative correlation before the change in task type (r = -.16) indicated that the more intelligent students showed lower levels of this ineffective strategic behavior. The positive correlation after the change in task type (r = .27) indicated the more effective strategy use of more intelligent students. This differential correlational pattern before and after the change in task type illustrated that intelligence was

systematically related to the adaption of strategies when task demands were changing. Furthermore, these results provided evidence that higher intelligence enabled students to better identify the effectiveness of a particular strategic behavior.

The slightly negative correlation between intelligence and the NOTAT gradient (slope) before the change in task type indicated that more intelligent students showed a less steep decrease in their NOTAT frequency. Since NOTAT could be regarded as a rather ineffective behavior across these tasks, because no dynamic effects occurred, one could have expected a positive correlation with intelligence that indicated that more intelligent students showed a steeper decrease of this suboptimal behavior. To further clarify this result, we conducted an additional analysis by changing the reference point of the NOTAT intercept factor from task 5 to task 1 (model is covariance equivalent to Model 9, Table 5). This change in the model specifications revealed that more intelligent students showed lower NOTAT frequencies already in the first task. One should keep in mind that the relative NOTAT frequency in the first task reached only 7%. Therefore, the more intelligent students had numerically almost no potential to show a steeper decrease of their NOTAT frequencies compared to the less intelligent students that showed a higher NOTAT level in the first task.

4.3. Limitations, Outlook, and Conclusion

One limitation of this study is that both VOTAT and NOTAT were measured by only one indicator per CPS task. Consequently, no adjustment for measurement error, for instance with second-order latent growth curve models, could be implemented. Therefore, it was unclear whether these indicators were reliable or whether the observed VOTAT or NOTAT behaviors might have appeared randomly. However, acceptable to good model fit indices for both the VOTAT and the NOTAT models indicated that our data was properly represented by the hypothesized discontinuous models. To provide further evidence that no deficiencies in the reliabilities of the VOTAT or NOTAT indicators led to our results, we computed the manifest

inter-task-correlations between two adjacent tasks separately for VOTAT and NOTAT, and separately for the two task types. Regarding the relative VOTAT frequency, the correlation coefficients were $.51 \leq r \leq .82$ with a median of $Mdn(r) = .73$ across the tasks 1–5 and $.75 \leq r \leq .88$ with a median of $Mdn(r) = .86$ across the tasks 6–9. Regarding the relative NOTAT frequency, the coefficients were smaller, but still substantial: $.33 \leq r \leq .51$ with a median of $Mdn(r) = .42$ across the tasks 1–5 and $.51 \leq r \leq .60$ with a median of $Mdn(r) = .58$ across the tasks 6–9. Numerical values of this magnitude indicated that students did not just randomly show the strategic behaviors, but that students who had higher frequencies in one task also had higher frequencies in the subsequent task. Furthermore, the VOTAT and NOTAT indicators mainly revealed the expected courses and the expected substantial convergent correlations with intelligence. Thus, these results were reasonable markers for the reliability and the validity of the indicators of students' strategic behavior, as used in this study.

Second, the choice of intelligence indicators, that is 10 subtests of the BIS, might be criticized. However, Jensen and Weng (1994) argued that such a selection of subtests fulfilled the criteria of a "good *g*" (also cf. Lotz et al., 2016). Most prior studies only assessed measures of (figural) reasoning as indicators of intelligence to examine the relation between intelligence and CPS performance measures. In contrast, our study overcame this frequent limitation and prevented possible impairments of the interpretation that may have resulted from too narrow intelligence operationalizations.

Next to the relatively broad operationalization of intelligence, this study incorporated some other strengths such as its theory-driven definition of domain-general strategic behaviors and their extraction from the log files (instead of strategies that strongly depended on the particular CPS test; cf. Güss et al., 2015). This allowed for more general conclusions about VOTAT's and NOTAT's importance for exploring the problem space of CPS tests, especially since we distinguished between the two task types. Furthermore, this study extended prior research by analyzing the VOTAT and NOTAT application of each

administered CPS task instead of either examining the strategic behaviors of only one single task (Greiff et al., 2015) or averaging across all tasks (Greiff et al., 2016; Wüstenberg et al., 2014). Obtaining strategy indicators for each task from the log files allowed drawing fine-grained conclusions about the course of the examined strategic behaviors and how students adapt those behaviors when confronted with a change in task type. Another important extension to most prior CPS research was analyzing VOTAT and NOTAT by relative frequencies (cf. Kröner et al., 2005). Earlier CPS studies mostly relied on a dichotomous scoring and gave credit if the problem solver applied the strategic behavior at least once during a task. A continuous scoring, as used in this study, gave a much more detailed picture about how (relatively) often students applied a particular strategic behavior and, thus, made the interpretation of the results even more trustworthy.

A next crucial step in research on strategic CPS behavior might be to examine its incremental validity for real-life criteria above and beyond intelligence. Next to the CPS-features, such as the tasks being dynamic and non-transparent, there is another major difference to conventional intelligence tests: the interaction with the task during the exploration phase. As an assumption, the specific type of interaction behavior assessed during the exploration phase which is unique to CPS tasks might substantially contribute to the prediction of real-world criteria by CPS beyond intelligence. At least, the study of Bryant et al. (2015) gave first evidence that the application of VOTAT could explain unique proportions of variance in science achievement beyond intelligence. Desirably, a future study that examines the increment of CPS process measures would assess CPS as well as intelligence computer-based to obviate possible effects of the mode of the test administration (computer-based vs. paper-and-pencil-based; cf. Sonnleitner et al., 2013). As another direction for future research, it might be fruitful to combine measures of CPS strategy use with measures of CPS response times. Because of the computer-based assessment of CPS, the exact time stamps of every single action of a student can be documented. For example, it

would be interesting to investigate whether the students with the better understandings of specific strategies are also more efficient in the sense of shorter times-on-task during the exploration phase.

To conclude, this study examined the course of the two strategic behaviors, VOTAT and NOTAT, across a set of CPS tasks with different task demands through log file analyses. The assessments of VOTAT and NOTAT were sufficiently reliable and valid, as well as useful process measures that gave new insights in how students used effective strategic behaviors. Additionally, it was possible to show how a certain development of VOTAT and NOTAT across a task set can be described by intercepts and slopes of discontinuous latent growth curve models. Over the course of a CPS task set, students showed higher levels and progressively increasing rates of a strategic behavior when it was effective, lower levels and decreasing rates when it was ineffective, besides a flexible adaption of the behaviors when task requirements changed. Furthermore, this study's main focus was on the relation between intelligence and the application and adaption of the effective strategic behaviors during the CPS exploration phase. Intelligence, as the (theoretically assumed) underlying construct, manifested itself in the observed specific behaviors. The corresponding substantial correlations with the process measures reflected that more intelligent students applied more effective strategic behaviors and were able to adapt them more rapidly. Although some conventional intelligence tests could be administered computer-based as well and, thus, provide log-files about the specific actions during the problem-solving process, a free exploration phase in which students can interact with the task is, nevertheless, unique to modern CPS tests. Understanding the mechanisms of how intelligence manifests itself in specific strategic actions during the exploration phase provided information above and beyond what conventional intelligence tests could offer and might be a key to understand the increment of CPS for real-life criteria. The results of this study gave first insights in how

intelligence acts during the exploration process by facilitating students' use of effective strategic behaviors.

References

Benedek, M., Jauk, E., Sommer, M., Arendasy, M., & Neubauer, A. C. (2014). Intelligence, creativity, and cognitive control: The common and differential involvement of executive functions in intelligence and creativity. *Intelligence, 46*, 73-83. doi:10.1016/j.intell.2014.05.007

Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation perspective.* New York: Wiley.

Brunner, M., & Süß, H.-M. (2005). Analyzing the reliability of multidimensional measures: An example from intelligence research. *Education and Psychological Measurement, 65,* 227–240. doi:10.1177/0013164404268669

Bryant, P., Nunes, T., Hillier, J., Gilroy, C., & Barros, R. (2015). The importance of being able to deal with variables in learning science. *International Journal of Science and Mathematics Education, 13,* 145–163. doi: 10.1007/s10763-013-9469-x

Chen, Z., & Klahr, D. (1999). All other things being equal: Acquisition and transfer of the control of variables strategy. *Child Development, 70*, 1098–1120. doi:10.1111/1467-8624.00081

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York, NY: Psychology Press.

Csapó, B., Ainley, J., Bennett, R. E., Latour, T., & Law, N. (2012). Technological issues for computer-based assessment. In P. Griffin, B. McGaw, & E. Care (Eds.), *Assessment and teaching of 21st century skills* (pp. 143–230). Dordrecht: Springer. doi: 10.1007/978-94-007-2324-5_4

Danner, D., Hagemann, D., Schankin, A., Hager, M., & Funke, J. (2011). Beyond IQ: A latent state-trait analysis of general intelligence, dynamic decision making, and implicit learning. *Intelligence, 39*, 323–334. doi: 10.1016/j.intell.2011.06.004

Deák, G. O. (2003). The development of cognitive flexibility and language abilities. *Advances in Child Development and Behavior, 31*, 271-327. doi: 10.1016/S0065-2407(03)31007-9

Deary, I. J., (2012). Intelligence. *Annual Review of Psychology, 63*, 453–482. doi: 10.1146/annurev-psych-120710-100353

Diallo, T. M. O., & Morin, A. J. S. (2015). Power of latent growth curve models to detect piecewise linear trajectories. *Structural Equation Modeling: A Multidisciplinary Journal, 22,* 449–460. doi: 10.1080/10705511.2014.935678

Dörner, D. (1980). On the difficulties people have in dealing with complexity. *Simulation & Gaming, 11,* 87–106. doi: 10.1177/104687818001100108

Dörner, D. & Schaub, H. (1994). Errors in planning and decision-making and the nature of human information processing. *Applied Psychology, 43*, 433–453. doi: 10.1111/j.1464-0597.1994.tb00839.x

Dörner, D., Stäudel, T., & Strohschneider, S. (1986). *Moro: Programmdokumentation [Moro: Program documentation]* (Memorandum No. 23). Bamberg, Germany: University of Bamberg, LS Psychologie II.

Enders, C. K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling, 8*, 430–457. doi: 10.1207/S15328007SEM0803_5

Frensch, P. A. & Funke, J. (1995). Definitions, traditions, and a general framework for understanding complex problem solving. In P. A. Frensch & J. Funke (Eds.), *Complex problem solving: The European perspective.* (pp. 3–25). Hillsdale, NJ: Erlbaum.

Funke, J. (2001). Dynamic systems as tools for analyzing human judgment. *Thinking and Reasoning, 7,* 69–89. doi: 10.1080/13546780042000046

Geiser, C. (2013). *Data analysis with Mplus.* New York, NY: Guilford Press.

Gerdes, J., Dörner, D., & Pfeiffer, E. (1993). *Interaktive Computersimulation „WINFIRE" [The interactive computer simulation "WINFIRE"].* Bamberg, Germany: University of Bamberg, LS Psychologie II.

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*, 608–626. doi: 10.1037/a0034716

Goode, N., & Beckmann, J. F. (2010). You need to know: There is a causal relationship between structural knowledge and control performance in complex problem solving tasks. *Intelligence, 38*, 345–352. doi:10.1016/j.intell.2010.01.001

Gottfredson, L. S. (1997). Mainstream science on intelligence: An editorial with 52 signatories, history, and bibliography. *Intelligence, 24,* 13–23. doi: 10.1016/S0160-2896(97)90011-8

van der Graaf, J., Segers, E., & Verhoeven, L. (2015). Scientific reasoning abilities in kindergarten: Dynamic assessment of the control of variables strategy. *Instructional Science, 43*, 381–400. doi: 10.1007/s11251-015-9344-y

Greiff, S., Fischer, A., Stadler, M., & Wüstenberg, S. (2015). Assessing complex problem solving skills with Multiple Complex Systems. *Thinking & Reasoning, 21*, 356–382. doi:10.1016/S0160-2896(97)90011-8

Greiff, S., Niepel, C., Scherer, R., & Martin, R. (2016). Understanding students‘performance in computer-based assessment of complex problem solving: An analysis of behavioral data from computer-generated log files. *Computers in Human Behavior, 61,* 36–46. doi:10.1016/j.chb.2016.02.095

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 & Education, 91*, 92–105. doi: 10.1016/j.compedu.2015.10.018

Greiff, S., Wüstenberg, S., & Funke, J. (2012). Dynamic problem solving: A new assessment perspective. *Applied Psychological Measurement, 36*, 189–213. doi: 10.1177/0146621612439620

Guthke, J., & Stein, H. (1996). Are learning tests the better version of intelligence tests? *European Journal of Psychological Assessment, 12*, 1–13. doi: 10.1177/0146621612439620

Güss, C. D., Tuason, M, M. T., & Orduña, L. V. (2015). Strategies, tactics, and errors in dynamic decision making in an Asian sample. *Journal of Dynamic Decision Making, 1*, 3. doi: 10.11588/jddm.2015.1.13131

Hancock, G. R., Harring, J. R., & Lawrence, F. R. (2006). Using latent growth models to evaluate longitudinal change. In G. R. Hancock & R. O. Mueller (Eds.), *Structural*

equation modeling: A second course (2nd ed., p. 309–341). Charlotte, NC: Age Publishing.

Hundertmark, J., Holt, D. V., Fischer, A., Said, N., & Fischer, H. (2015). System structure and cognitive ability as predictors of performance in dynamic system control tasks. *Journal of Dynamic Decision Making, 1,* 5. doi:10.11588/jddm.2015.1.26416

Hunt, E. (2011). *Human intelligence.* New York, NY: Cambridge University Press. doi:10.1017/CBO9780511781308.002

Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence.* New York, NY: Basic. doi: 10.1037/10034-000

Jäger, A. O., Süß, H.-M, & Beauducel, A. (1997). *Berliner Intelligenzstruktur-Test. BIS-Test, Form 4 [Berlin Intelligence-Structure Test. Version 4]*. Göttingen, Germany: Hogrefe.

Jensen, A. R., & Weng, L.-J. (1994). What is a good g? *Intelligence, 18,* 231–258. doi:10.1016/0160-2896(94)90029-9

Klahr, D., & Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive Science, 12,* 1–48. doi: 10.1207/s15516709cog1201_1

Kretzschmar, A., Neubert, J. C., Wüstenberg, S., & Greiff, S. (2016). Construct validity of complex problem solving: A comprehensive view on different facets of intelligence and school grades. *Intelligence, 54,* 55–69. doi:10.1016/j.intell.2015.11.004

Kröner, S. (2001). *Intelligenzdiagnistik per Computersimulation [Intelligence assessment via computer simulation].* Münster, D: Waxmann.

Kröner, S., Plass, J. L., & Leutner, D. (2005). Intelligence assessment with computer simulations. *Intelligence, 33,* 347–368. doi: 10.1016/j.intell.2005.03.002

Kuhn, D., Garcia-Mila, M., Zohar, A., & Andersen, C. (1995). Strategies of knowledge acquisition. *Society for Research in Child Development Monographs, 60*, Serial No. 245. doi: 10.2307/1166059

Künsting, J., Kempf, J., & Wirth, J. (2013). Enhancing scientific discovery learning through metacognitive support. *Contemporary Educational Psychology, 38*, 349–360. doi:10.1016/j.cedpsych.2013.07.001

Künsting, J., Wirth, J., & Paas, F. (2011). The goal specificity effects on strategy use and instructional efficiency during computer-based scientific discovery learning. *Computers & Education, 56*, 668–679. doi: 10.1016/j.compedu.2010.10.009

Lee, C. B., Jonassen, D., & Teo, T. (2011). The role of model building in problem solving and conceptual change. *Interactive Learning Environments, 19,* 247–265. doi: 10.1080/10494820902850158

LePine, J. A., Colquitt, J. A., & Erez, A. (2000). Adaptability to changing task contexts: Effects of general cognitive ability conscientiousness, and openness to experience. *Personnel Psychology, 53*, 563–569. doi:10.1111/j.1744-6570.2000.tb00214.x

Little, T. D. (2013). *Longitudinal structural equation modeling*. New York, NY: Guilford.

Lotz, C., Sparfeldt, J. R., & Greiff, S. (2016). Complex problem solving in educational contexts – Still something beyond a "good *g*"? *Intelligence*, 59, 127-138. doi:10.1016/j.intell.2016.09.001

Muthén, L. K., & Muthén, B. O. (1998-2013). *Mplus user's guide.* Los Angeles, CA: Muthén & Muthén.

Novick, L. R., & Bassok, M. (2005). Problem solving. In K. J. Holyoak & R. G. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning* (p. 321-349). Cambridge, NY: University Press.

OECD. (2014). *PISA 2012 results: What students know and can do – student performance in mathematics, reading and science (Volume I).* Paris, France: PISA OECD Publishing.

Satorra, A. (2000). Scaled and adjusted restricted tests in multi-sample analysis of moment structures. In Heijmans, R.D.H., Pollock, D.S.G. & Satorra, A. (eds.), *Innovations in multivariate statistical analysis. A Festschrift for Heinz Neudecker* (pp.233-247). London: Kluwer Academic Publishers.

Schauble, L. (1996). The development of scientific reasoning in knowledge-rich contexts, *Developmental Psychology, 32*, 102–119. doi: 10.1037/0012-1649.32.1.102

Scherer, R., Greiff, S., & Hautamäki, J. (2015). Exploring the relation between time on task and ability in complex problem solving. *Intelligence, 48,* 37–50. doi: 10.1016/j.intell.2014.10.003

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Test of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online, 8*, 23–74.

Schwichow, M., Croker, S., Zimmerman, C., Höffler, T., & Härtig, H. (2016). Teaching the control-of-variables strategy: A meta-analysis. *Developmental Review, 39*, 37-63. doi: 10.1016/j.dr.2015.12.001

Sonnleitner, P., Keller, U., Martin, R., & Brunner, M. (2013). Students‘ complex problem-solving abilities: Their structure and relations to reasoning ability and educational success. *Intelligence, 41*, 289–305. doi: 10.1016/j.intell.2013.05.002

Stadler, M., Becker, N., Gödker, M., Leutner, D., & Greiff, S. (2015). Complex problem solving and intelligence: A meta-analysis. *Intelligence, 53*, 92–101. doi: 10.1016/j.intell.2015.09.005

Sternberg, R. J. (1997). The concept of intelligence and its role in lifelong learning and success. *American Psychologist, 52,* 1030–1037. doi: 10.1037/0003-066X.52.10.1030

Strohschneider, S. & Güss, D. (1999). The fate of the moros: A cross-cultural exploration of strategies in complex and dynamic decision making. *International Journal of Psychology, 34*, 235–252. doi: 10.1080/002075999399873

Süβ, H.-M., Oberauer, K., Wittmann, W. W., Wilhelm, O., & Schulze, R. (2002). Working-memory capacity explains reasoning ability – And a little bit more. *Intelligence, 30*, 261–288. doi: 10.1016/S0160-2896(01)00100-3

Thorndike, E. L. (1922). Practice effects in intelligence tests. *Journal of Experimental Psychology, 5*, 101-107. doi: 10.1037/h0074568

Tschirgi, J. E. (1980). Sensible reasoning: A hypothesis about hypotheses. *Child Development, 51*, 1–10. doi: 10.2307/1129583

Valerius, S. & Sparfeldt, J. (2014). Consistent g- as well as consistent verbal-, numerical- and figural-factors in nested factor models? Confirmatory factor analyses using three test batteries. *Intelligence, 44*, 120–133. doi: 10.1016/j.intell.2014.04.003

Veenman, M. V. J., Bavelaar, L., De Wolf, L., & Van Haaren, M. G. P. (2014). The on-line assessment of metacognitive skills in a computerized learning environment. *Learning and Individual Differences, 29*, 123–130. doi:10.1016/j.lindif.2013.01.003. doi: 10.1016/j.lindif.2013.01.003

Veenman, M. V. J., Wilhelm, P., & Beishuizen, J. J. (2004). The relation between intellectual and metacognitive skills from a developmental perspective. *Learning and Instruction, 14*, 89–109. doi:10.1016/j.learninstruc.2003.10.004. doi: 10.1016/j.learninstruc.2003.10.004

Vollmeyer, R., Burns, B. D., & Holyoak, K. J. (1996). The impact of goal specificity on strategy use and the acquisition of problem structure. *Cognitive Science, 20*, 75–100. doi:10.1207/s15516709cog2001_3. doi: 10.1207/s15516709cog2001_3

Wüstenberg, S., Greiff, S., & Funke, J. (2012). Complex problem solving – More than reasoning? *Intelligence, 40*, 1–14. doi: 10.1016/j.intell.2011.11.003

Wüstenberg, S., Stadler, M., Hautamäki, J., & Greiff, S. (2014). The role of strategy knowledge for the application of strategies in complex problem solving tasks. *Technology, Knowledge and Learning, 19*, 127–146. doi:10.1007/s10758-014-9222-8

Zimmerman, C. & Croker, S. (2013). Learning science through inquiry. In G. Feist & M. Gorman (Eds.), *Handbook of the psychology of science* (pp. 49–70). New York, NY: Springer.

Zimmerman, C., Raghavan, K., & Sartoris, M. L. (2003). The impact of the MARS curriculum on students' ability to coordinate theory and evidence. *International Journal of Science Education, 25,* 1247–1271. doi: 10.1080/095006902200003830

Table 1

Studies on the relation between intelligence and the use of VOTAT in the domains of CPS and scientific reasoning.

	Domain	*Test instrument*	*VOTAT operationalization*	*Intelligence measure*	*VOTAT-Intelligence - relation*	*Sample size*	*Participants grade levels and mean age*
Kröner et al. (2005)	CPS	MultiFlux	Relative frequency	Verbal, numerical, and figural reasoning (BIS-K)	$r = .41$	101	Grades: 9-12; *Age: M* = 15.0, *SD* = 1.0
Wüstenberg et al. (2014)	CPS	MicroDYN	Dichotomous per task (participant applied VOTAT at least once for every input variable)	Verbal and numerical reasoning (two subtests)	$r = .64$	3,191	Grades: 6 or 9; *Age: M* = 13.59, *SD* = 1.56
van der Graf et al. (2015)	Scientific reasoning	Ramp task	(a) absolute frequency of correctly designed experiments after two trials each; (b) absolute frequency of variables set correctly	Nonverbal reasoning (one subtest)	$r_a = .42$, $r_b = .47$	46	Age Range: 4;6 – 6;3, *M* = 5;3
Künsting et al. (2013)	Scientific reasoning	Computer-based learning environment	Relative frequency	Figural reasoning (one subtest)	$r = .30$	129	Grade: 9; *Age: M* = 14.33, *SD* = 0.69
Veenman et al. (2014)	Scientific reasoning	Otter task	Absolute frequency	Verbal, numerical, and figural reasoning (three subtests)	$r = .17$	52	Grade: 7; *Age: M* = 13;2

Table 2

Means and standard deviations of the relative VOTAT and NOTAT frequencies across the nine MicroDYN tasks, complemented by the corresponding Intraclass Correlation Coefficients (ICC-1) and design effects

	VOTAT				NOTAT			
Task name	*M*	*SD*	*ICC-1*	*Design effect*	*M*	*SD*	*ICC-1*	*Design effect*
1. Lemonade	.47	.32	.030	1.54	.07	.13	.026	1.47
2. Drawing	.58	.37	.060	2.08	.06	.16	.038	1.69
3. Cat	.63	.37	.071	2.28	.04	.14	.025	1.45
4. Moped	.67	.39	.067	2.21	.04	.12	.013	1.23
5. Game	.71	.39	.083	2.50	.05	.17	.031	1.56
6. Gardening	.60	.36	.050	1.90	.10	.19	.037	1.67
7. Handball	.69	.38	.073	2.32	.10	.19	.029	1.52
8. Spaceship	.71	.38	.055	1.99	.10	.20	.048	1.87
9. Medical Aid	.70	.37	.049	1.88	.13	.24	.028	1.51

Table 3

Name of tests, description, content facet, standardized loadings of the subtests on their corresponding latent content factor, operation facet, number of items, means, and standard deviations for the 10 BIS-4 subtests complemented by the corresponding Intraclass Correlation Coefficients (ICC-1), and design effects

Test (abbr.)	Description	Content facet	λ on content factor in Model 4/9	Operation facet	No. of Items	*M*	*SD*	*ICC-1*	*Design effect*
Figural analogies (AN)	Identification of figural analogies	F	.68*/.68*	R	8	3.32	1.59	.077	2.39
Crossing out letters (BD)	Crossing out letters in a series	F	.25*/.26*	S	130	54.22	13.41	.090	2.62
Charkow (CH)	Completion and generalization of figures in a series	F	.70*/.70*	R	6	2.26	1.54	.087	2.57
City map (OG)	Recall of buildings in a city map	F	.30*/.30*	M	27	14.71	4.48	.069	2.24
Number sequences (ZN)	Completion of numbers in a series	N	.68*/.68*	R	9	4.18	2.56	.138	3.49
X greater (XG)	Crossing out numbers x greater than the prior one	N	.61*/.62*	S	44	19.41	8.56	.150	3.75
Estimation (SC)	Estimation of complex arithmetic	N	.65*/.64*	R	7	3.54	1.77	.088	2.59
Story (ST)	Recall of text information	V	.26*/.27*	M	22	8.74	3.43	.064	2.15
Fact-opinion (TM)	Conclusion of fact or opinion of verbal statements	V	.54*/.53*	R	16	9.91	2.54	.078	2.41
Verbal analogies (WA)	Identification of analogous word pairs	V	.52*/.52*	R	8	2.35	1.52	.092	2.66

Note. Standardized loading of the latent first-order content factors on the second-order *g*-factor were identical for the VOTAT model (Table 5, Model 4; Fig 3) and the NOTAT model (Table 5, Model 9; Fig. 4): $\lambda_{verbal} = .62*$, $\lambda_{numerical} = .80*$, and $\lambda_{figural} = .90*$. F = figural; N = numerical; V = verbal; R = reasoning capacity; S = speed; M = memory.
* $p < .05$.

Table 4

Manifest correlations of the 10 Intelligence subtests with the relative VOTAT and NOTAT frequency (reported task wise as well as averaged across the two task types).

Tests	*Intelligence*									
	1	2	3	4	5	6	7	8	9	10
Intelligence										
1. ST	--									
2. TM	.15*	--								
3. WA	.07	.30*	--							
4. ZN	.14*	.15*	.16*	--						
5. XG	.21*	.21*	.23*	.43*	--					
6. SC	.14*	.19*	.20*	.41*	.41*	--				
7. AN	.09*	.15*	.20*	.39*	.23*	.34*	--			
8. BD	.13*	-.01	.07	.09*	.20*	.14*	.14*	--		
9. CH	.08	.20*	.21*	.35*	.22*	.34*	.49*	.16*	--	
10. OG	.28*	.00	-.01	.18*	.16*	.12*	.18*	.31*	.20*	--
CPS, relative VOTAT frequency										
11. Task 1	.04	.11*	.04	.10*	.12*	.13*	.17*	.03	.20*	.06
12. Task 2	.03	.18*	.09	.17*	.20*	.19*	.25*	.09*	.30*	.09*
13. Task 3	.07	.13*	.15*	.17*	.17*	.19*	.27*	.02	.29*	.10*
14. Task 4	.06	.15*	.15*	.19*	.16*	.18*	.28*	.06	.28*	.12*
15. Task 5	.01	.15*	.11*	.20*	.16*	.20*	.26*	.05	.31*	.08
16. Task 6	.01	.12*	.08	.16*	.07	.19*	.20*	.00	.26*	.08
17. Task 7	.04	.14*	.08	15*	.12*	.17*	.19*	.02	.25*	.08
18. Task 8	.02	.17*	.08	.17*	.14*	.16*	.22*	.05	.27*	.07
19. Task 9	.01	.11*	.07	.11*	.11*	.16*	.19*	.04	.27*	.06
20. Tasks 1-5	.05	.18*	.13*	.20*	.20*	.21*	.29*	.06	.34*	.12*
21. Tasks 6-9	.04	.15*	.09	.17*	.12*	.18*	.22*	.03	.29*	.10*
CPS, relative NOTAT frequency										
22. Task 1	.01	-.02	-.07	-.02	.00	-.02	-.08	-.03	-.10*	.07
23. Task 2	.09	.01	.02	-.05	-.05	-.03	-.05	-.06	-.08	-.07
24. Task 3	.04	-.02	-.06	-.08	-.06	-.13*	-.10*	-.08	-.13*	-.05
25. Task 4	-.01	-.04	-.09	-.04	-.03	.00	-.12*	-.12*	-.10*	-.03
26. Task 5	.01	.01	-.04	-.10*	-.04	-.07	-.16*	-.11*	-.14*	-.03
27. Task 6	.04	.16*	.10*	.13*	.13*	.08	.05	-.05	.10*	.01
28. Task 7	.07	.16*	.07	.15*	.17*	.11*	.07	-.03	.11*	.02
29. Task 8	.00	.16*	.10*	.10*	.06	.07	.02	-.05	.10*	-.03
30. Task 9	.08	.14*	.13*	.09*	.13*	.01	.02	-.05	.08	.02
31. Tasks 1-5	.04	-.02	-.07	-.09*	-.05	-.08	-.15*	-.12*	-.16*	-.03
32. Tasks 6-9	.05	.12*	.03	.10*	.12*	.07	.01	-.05	.03	.06

Note. Please refer to Table 3 for the names of the abbreviated intelligence subtests. Not displayed in the table: $r_{VOTAT1\text{-}5\text{–}NOTAT1\text{-}5}$ = -.35*; $r_{VOTAT6\text{-}9\text{–}NOTAT6\text{-}9}$ = -.20*;
* $p < .05$.

Table 5

Fit indices of the latent growth curve models of VOTAT and NOTAT.

Models	χ^2	*df*	CFI	TLI	RMSEA	Comp. model	T	Δ*df*	*p*
VOTAT Models									
1. Cont. LGCM without intelligence	419.702*	40	.881	.892	.138				
2. Disc. LGCM without intelligence	164.868*	31	.958	.951	.093	1	241.032	9	< .01
3. a. Disc. LGCM without intelligence, no slope before	603.117*	36	.822	.822	.178	2	428.060	5	< .01
3. b. Disc. LGCM without intelligence, no slope after	225.211*	36	.940	.940	.103	2	59.508	5	< .01
3. c. Disc. LGCM without intelligence, quad. slope before[a]	(105.883*)	(25)	(.975)	(.963)	(.081)				
3. d. Disc. LGCM without intelligence, quad. slope after[a]	(99.934*)	(25)	(.976)	.(966)	(.078)				
4. Disc. LGCM with intelligence	348.423*	132	.955	.942	.058				
5. Disc. LGCM with intelligence difftest intercept means	391.457*	133	.946	.931	.063	4	38.350	1	< .01
NOTAT Models									
6. Cont. LGCM without intelligence	102,415*	40	.822	.840	.056				
7. Disc. LGCM without intelligence	39.077*	31	.977	.973	.023	6	53.212	9	< .01
8. a. Disc. LGCM without intelligence, no slope before	59.912*	36	.932	.932	.037	7	16.427	5	< .01
8. b. Disc. LGCM without intelligence, no slope after	49.598*	36	.964	.964	.027	7	12.117	5	< .05
8. c. Disc. LGCM without intelligence, quad. slope before[a]	(32.643*)	(25)	(.978)	(.969)	(.025)				
8. d. Disc. LGCM without intelligence, quad. slope after[a]	(29.158*)	(25)	(.988)	(.983)	(.018)				
9. Disc. LGCM with intelligence	180.688*	132	.964	.953	.027				
10. Disc. LGCM with intelligence difftest intercept means	218.101*	133	.937	.919	.036	9	24.504	1	< .01

Note. Cont. = continuous; Disc. = discontinuous; LGCM = latent growth curve model; χ^2 = chi-square goodness-of-fit statistic; *df* = degrees of freedom; CFI = Comparative Fit Index; TLI = Tucker-Lewis Index; RMSEA = Root Mean Square Error of Approximation; Comp. model = comparison model; T = value of the chi-square difference test; T and Δ*df* were estimated by the Satorra-Bentler corrected chi-square difference test (see Satorra, 2000).

The models 4 and 9 correspond to the Figures 3 and 4, respectively.

[a] Model estimation revealed a Heywood case as indicated by a correlation > |1| between the linear and quadratic slope factors, pointing towards an over-specification of the models with quadratic slopes; models were rejected from further analyses (cf. Geiser, 2013, p. 182).

* $p < .05$.

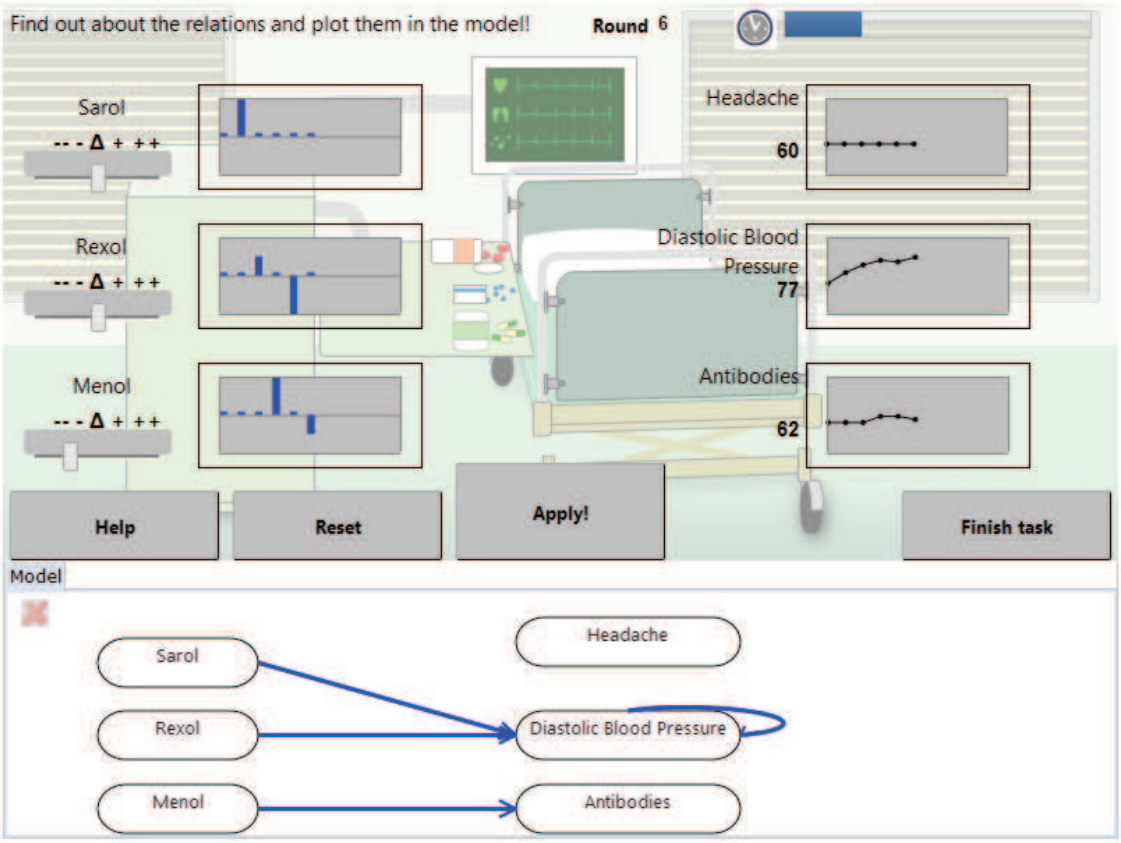

Figure 1. Screenshot of the MicroDYN task "Medical Aid" during the knowledge acquisition phase.

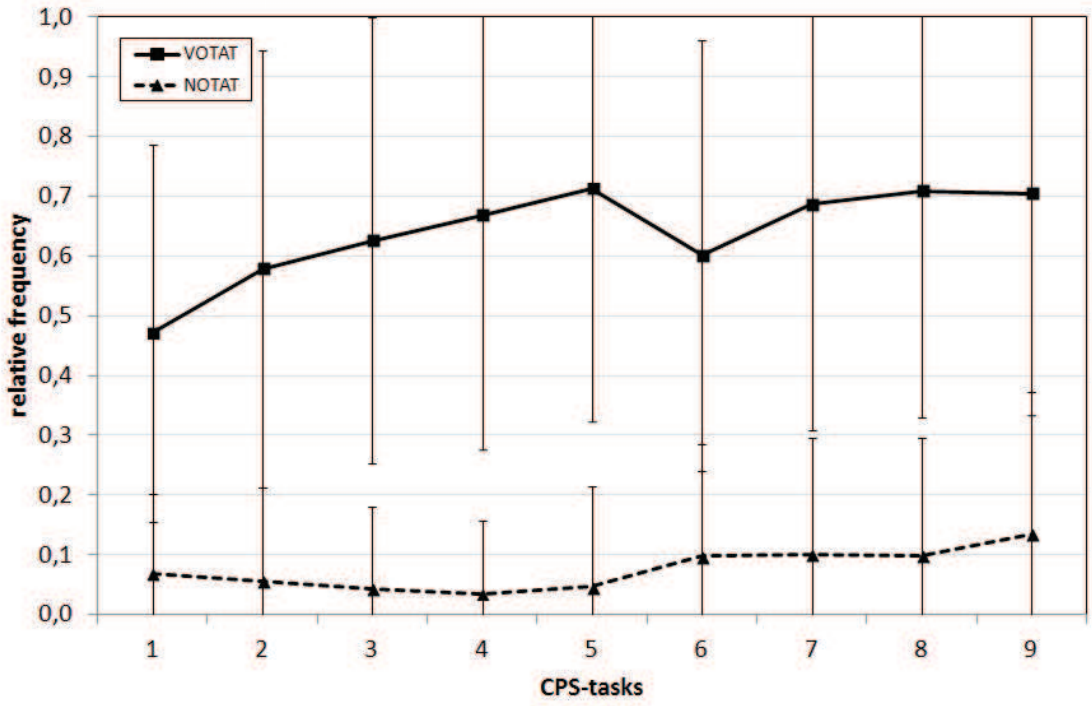

Figure 2. Courses of the relative VOTAT and NOTAT frequency; means and standard deviations across the nine CPS tasks. Change in task type after task 5.

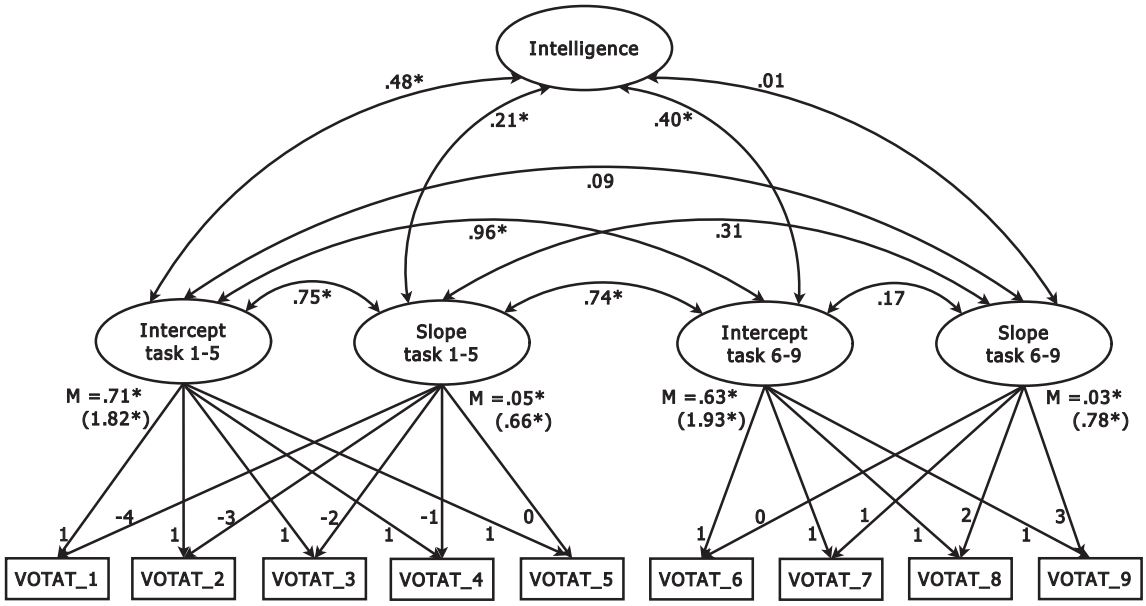

Figure 3. Discontinuous latent growth curve model for the relative VOTAT frequencies across the nine CPS tasks augmented by intelligence (see Table 5, Model 4).

Note. Non-standardized (standardized) means are depicted for the intercept and slope factors. Relations among latent variables are shown as correlation coefficients. The measurement model of intelligence is not illustrated.
Please note that small discrepancies compared to the descriptive statistics presented in Table 2 and Fig. 2 are due to the model-based estimates.
* $p < .05$.

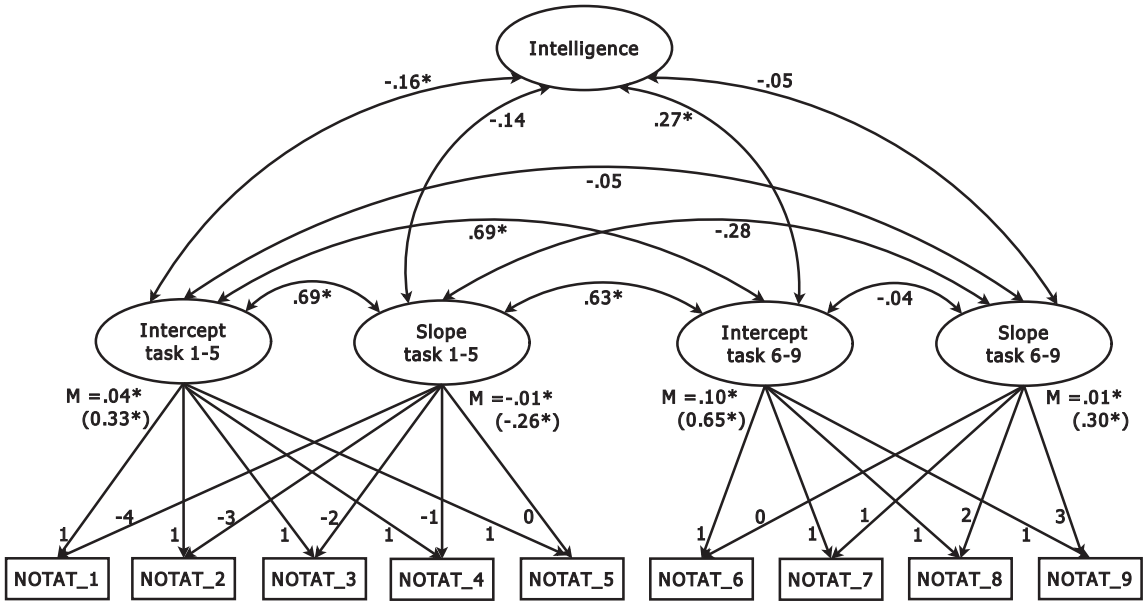

Figure 4. Discontinuous latent growth curve model for the relative NOTAT frequencies across the nine CPS tasks augmented by intelligence (see Table 5, Model 9).

Note. Non-standardized (standardized) means are depicted for the intercept and slope factors. Relations among latent variables are shown as correlation coefficients. The measurement model of intelligence is not illustrated.
Please note that small discrepancies compared to the descriptive statistics presented in Table 2 and Fig. 2 are due to the model-based estimates.
* $p < .05$.