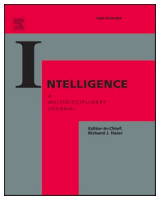
Intelligence 100 (2023) 101773

Contents lists available at ScienceDirect

Intelligence

journal homepage: www.elsevier.com/locate/intell



The role of learning in complex problem solving using MicroDYN

W. Herrmann [a], J.F. Beckmann [b], A. Kretzschmar [a,c,*]

[a] University of Zurich, Switzerland
[b] Durham University, UK
[c] University of Tübingen, Germany

ARTICLE INFO

Keywords:
Learning ability
ADAFI
MicroDYN
Multilevel
System characteristics

ABSTRACT

It is still an open question which cognitive and non-cognitive personality traits are useful for describing and explaining behaviour and performance in complex problems. During complex problem solving (CPS), problem solvers have to interact with the task in a way in which learning ability might be beneficial for successful task completion. By investigating the relationship between learning ability and CPS, while accounting for interactions between complex system characteristics and person characteristics, this paper aims to understand the role of learning processes in CPS more closely. In a sample of $N = 241$ participants, we performed a preregistered analysis to investigate the relationship between knowledge acquisition performance in a CPS test (MicroDYN) and learning test performance (ADAFI) with a multilevel modeling approach across 10 CPS systems with various characteristics. In line with our expectations, we replicated previous findings on a relationship between learning test and MicroDYN performance and found this relationship to be more pronounced in systems with (vs. without) autonomous changes. Further system and person characteristics also showed effects as expected, with better performance in systems with lower complexity, with more experience with the task, and with more strategic exploration behaviour. Our results provide further evidence for the notion that learning is an important component for the successful completion of CPS tasks.

1. Introduction

In the last four decades, different attempts to define complex problem solving (CPS) have been made (for an overview, see, e.g. Dörner & Funke, 2017; Frensch & Funke, 1995b). One of the most recent definitions states that CPS "is a collection of self-regulated psychological processes and activities necessary in dynamic environments to achieve ill-defined goals that cannot be reached by routine actions […] The problem-solving process combines cognitive, emotional, and motivational aspects, particularly in high-stakes situations" (Dörner & Funke, 2017, p. 6). Previous research has sought to describe and explain behaviour and performance in complex problems by examining the association with intelligence (e.g., Kretzschmar, Hacatrjana, & Rascevska, 2017; Sonnleitner, Keller, Martin, & Brunner, 2013; Stadler, Becker, Gödker, Leutner, & Greiff, 2015; Süß, 1996), working memory (e.g., Kretzschmar & Nebe, 2021; Schweizer, Wüstenberg, & Greiff, 2013; Wittmann & Hattrup, 2004), knowledge (e.g., Süß, 1996; Süß & Kretzschmar, 2018), and personality traits (e.g., Greiff & Neubert, 2014; Rudolph, Greiff, Strobel, & Preckel, 2018). However, the empirical associations in most studies were moderate at best, raising the question of what other personality traits and abilities might be potentially relevant for solving complex problems (e.g., Beckmann, 2019; Kretzschmar, Neubert, Wüstenberg, & Greiff, 2016).

In the early days of CPS research (e.g., Beckmann, 1994; Beckmann & Guthke, 1995), links between CPS and learning ability (Guthke, 1982) were studied. Conceptual similarities were identified in task demands, that is the challenge to acquire knowledge about the causal structure of a computerised system in CPS research, and on the receptiveness to feedback and thinking prompts as they are offered in so-called learning tests. In line with these conceptual considerations, these early studies provided empirical evidence for the link between performance in knowledge acquisition in CPS and learning test performance. However, it has been argued that in these early studies the CPS performance indices are afflicted by reliability issues (e.g., Herde, Wüstenberg, & Greiff, 2016; Süß, 1996), the studies were conducted with relatively small sample sizes according to today's standards, and they did not sufficiently account for potentially moderating variables when investigating CPS-learning ability relationships. Although more recent research

* Corresponding author at: University of Zurich, Department of Psychology, Individual Differences and Assessment, Binzmuehlestrasse 14/7, CH-8050 Zurich, Switzerland.
E-mail address: kretzsch.andre@gmail.com (A. Kretzschmar).

https://doi.org/10.1016/j.intell.2023.101773
Received 3 January 2023; Received in revised form 21 June 2023; Accepted 25 June 2023
Available online 5 July 2023

repeatedly discussed learning ability as relevant for CPS performance from a theoretical perspective (e.g., Beckmann & Guthke, 1999; Kretzschmar et al., 2016; Süß, 1996; Wüstenberg, Greiff, & Funke, 2012), it still remains, firstly, unclear whether previous empirical findings on the link between CPS and learning ability translate to currently widely-used instruments which were designed to ameliorate potential reliability issues (see Greiff, Fischer, Stadler, & Wüstenberg, 2015), and secondly, whether previously found associations are specific to or moderated by certain system characteristics, or can be generalised across a wider range of CPS tests. The current study therefore aims to investigate the relationship between learning ability and performance in a currently widely used CPS assessment approach and, thus, provides further insights into the nature of CPS as an ability construct.

1.1. Conceptual overlap between learning tests and complex problem solving tests

Learning tests represent a form of dynamic testing (Guthke, Beckmann, & Wiedl, 2003; Guthke & Wiedl, 1996; see also Lidz & Elliott, 2000). Dynamic Testing is defined "as a methodological approach to psychometric assessment that uses systematic variations of task characteristics and / or situational characteristics in the presentation of test items with the intention to evoke intra-individual variability in test performance. Interindividual differences in intraindividual variation are seen as more adequately reflecting the dynamics in the organisation of human behaviour." (Beckmann, 2014, p. 310). In learning tests – which may utilise typical reasoning items across domains (e.g., number series, verbal analogies, and figure series) – test takers are provided with item-by-item feedback and a graduated system of error-specific thinking prompts after an incorrect answer until the item is solved correctly (see Fig. 1). Subsequently presented items of comparable complexity then provide an opportunity for the test taker to demonstrate how they were able to benefit from the feedback and learning stimulation they were exposed to at preceding items. The diagnostic focus in such tests is on the receptiveness of the individual test taker to learning stimulations rather that the number of correct responses as is the case in conventional tests of cognitive abilities.

CPS tools represent another form of dynamic testing, especially in the case of minimal/multiple complex system approaches (Funke & Greiff, 2017; Greiff et al., 2015) such as MicroDYN (Greiff, Wüstenberg, & Funke, 2012) as they use several microsystems with systematic variations of task characteristics. More specifically, the ability to benefit from feedback and the receptiveness to learning stimulations might not only be relevant in learning tests, but also for CPS. In other words, both learning tests and CPS tools have the potential to not only capture inter-individual, but also intra-individual variability due to their dynamic nature.

Typical CPS tools confront the problem solver with two distinct tasks. In general terms, the first task requires the acquisition of knowledge about how the system variables are interconnected (similar to relational integration, see, e.g., Hannon & Daneman, 2014; Oberauer, Süß, Wilhelm, & Wittmann, 2008). To address the task of knowledge acquisition the problem solver is given the opportunity to manipulate the values of the input variables and, based on the observed changes in the output variables, draw inferences about the causal structure of the system. An example of a typical MicroDYN system during the knowledge acquisition phase is provided in Fig. 2. It is expected that during this goal-free (i.e., no target is given for the output variables) exploration of the system the problem solver acquires *effect* knowledge (Beckmann, 1994; Beckmann & Goode, 2017). This is knowledge about whether and how each of the input variables affect any of the output variables, including the effects that output variables might have on themselves (i.e., autonomous changes). For a successful acquisition of effect knowledge it is essential to create informative transitions from one system state to another that allow to conclusively attribute observed changes in the output variables to specific changes made in the input variables (Beckmann & Goode, 2014). As explained above, especially this process seems to require abilities which are also relevant in learning tests, which is in line with previous findings demonstrating that knowledge acquisition performance and learning test performance are associated (e.g., $r = 0.42$ in Beckmann, 1994; $r = 0.69$ in Freitag, 1993). However, the relatively small sample sizes in these studies in conjunction with reliability concerns (see Greiff et al., 2015) nurture the sense of necessity for a replication of these findings. The second task in CPS builds upon the first by asking problem solvers to manipulate the values of the input variables, hence utilising their acquired causal knowledge about the system, in order to reach or maintain given target values in the output variables (i.e., system control). It has been demonstrated that the performance in the control phase is causally determined by the success in the preceding knowledge acquisition phase (e.g., Goode & Beckmann, 2010), which is reflected in labelling it as the knowledge *application* phase. Whilst exploration interventions are aimed at acquiring *effect* knowledge, the control interventions are informed by *dependency* knowledge, that is knowledge about which input variable(s) a specific output variable depends on (Beckmann, 1994; Beckmann & Goode, 2017). The necessary transformation processes from effect to dependency knowledge are not immune to error ("lost in translation"), meaning that control performance can only be considered as mediated, or confounded. This issue imposes challenges to utilizing control performance scores for the investigation of links between CPS and learning ability, which is why Beckmann and Goode (2017) discuss the knowledge application performance as reflecting learning ability rather indirectly at best. The main challenges are: (1) inter-individual differences in the effectiveness of this translation process, (2) the fact that systems can be controlled to an acceptable standard based on intervention-by-intervention optimisation, which does not require any knowledge,[1] and (3) controlling the system successfully is possible even without any *acquired* knowledge in the exploration phase because the correct causal model is provided at the beginning of the knowledge application phase for each system in MicroDYN (Greiff et al., 2015). Consequently, interpreting control performance as an indicator for the application of acquired (i.e., learned) knowledge demands caution. We therefore focus on performance shown in the knowledge acquisition phase as the conceptual counterpart of learning test scores in the present study.

While these two CPS subtasks do not entail direct behavioural feedback or learning stimulations as in learning tests, they provide the problem solver with direct feedback on their actions and allow them to approach the desired outcome (acquiring knowledge about a system, and then applying it) in a stepwise manner. This might be one of the reasons why previous research has demonstrated CPS performance, especially in the knowledge acquisition task, to be substantially associated with learning test performance (Beckmann, 1994; Freitag, 1993). Relevant research has also highlighted that characteristics of the CPS systems used, such as their semantic embedment in terms of cover story or variable labels, moderate this relationship (Beckmann & Goode, 2014). From a conceptual point of view, one might thus argue that success in acquiring system knowledge in CPS is positively (co-)determined by the ability to learn, with a potential moderation of system characteristics.

1.2. System and person characteristics affect performance in CPS systems

The shared use of the label "CPS" across different studies tends to

[1] The simple, knowledge-free heuristic builds on step-by-step monitoring of system states after each control intervention (Beckmann & Guthke, 1995, p. 195; Beckmann & Goode, 2017). Guessing-based interventions that bring the system closer to the target state will be repeated, those that increase the distance between actual and target state will be reversed. Such intervention-by-intervention optimization, which can result in acceptable control performance, does not require knowledge.

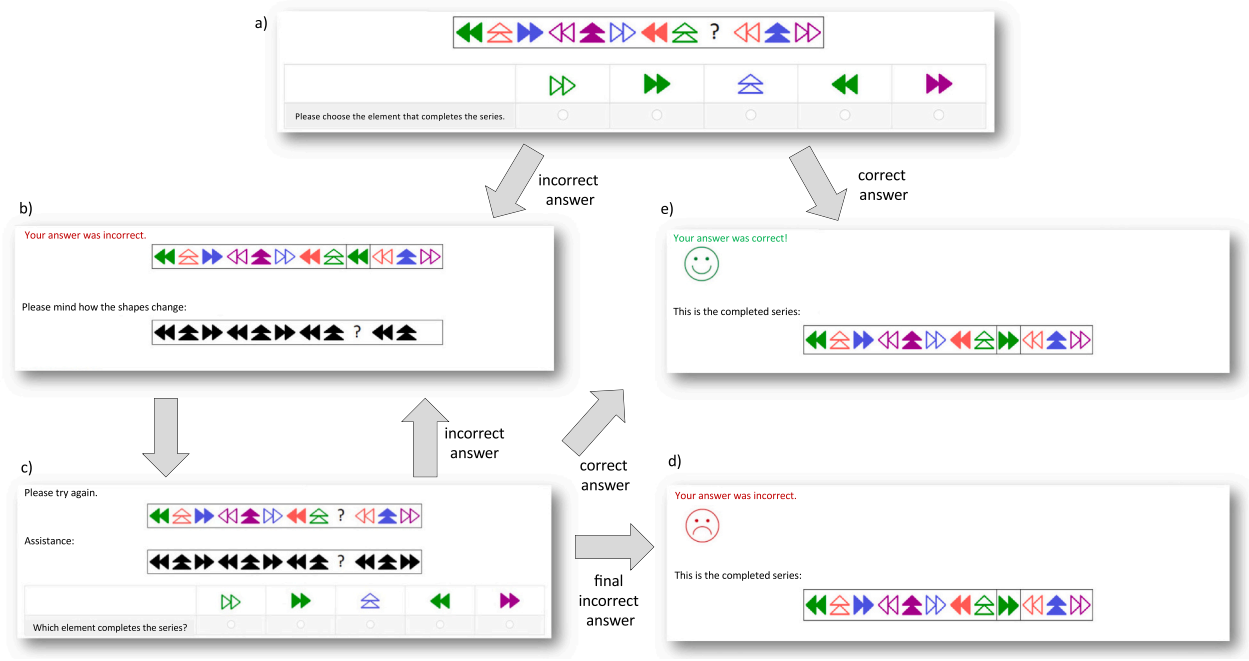


Fig. 1. *Flowchart for an example item of the learning test "Adaptiver Figurenfolgenlerntest" (ADAFI).*
Note. Figure a) is the starting point for each task. If an incorrect answer is selected, feedback and error-specific thinking prompts are given (Figure b), and the task is presented again (Figure c). Selecting the same type of error will result in the task being incorrectly solved and the next task will be presented (Figure d). If a correct answer is selected, feedback is given and the next task will be presented (Figure e). From Kretzschmar, A. (2023). *Adaptiver Figurenfolgenlerntest (ADAFI)* (Version 1). figshare. https://doi.org/10.6084/m9.figshare.23617476. CC BY 4.0

mask the fact that systems employed in this kind of research often vary wildly in terms of their characteristics (Beckmann, 2019). These include, but are by no means limited to the way information is presented, the ways in which problem solvers can interact with the system, the time frames, and the semanticity attached to the systems. The consideration of the role of system characteristics is generally important as they are contributors to complexity – and ultimately the levels of difficulty problem solvers experience, which is reflected in their performance scores. Apart from system characteristics, performance can further be affected by person-related variables, such as how much experience problem solvers have in interacting with CPS systems throughout the test (i.e., in the 10th system a problem solver has interacted with nine prior systems) and how this experience shapes their current behaviour, or how systematically they explore a system. Awareness of these effects is of particular relevance in the context of an exploration of the role of learning ability in MicroDYN – as is pursued in this study. In the following sections, we will provide an overview on a selection of characteristics of complex problems, how problem solvers might interact with them, and explain for each how we expect them to relate to learning ability.

1.2.1. Complexity

Complexity in CPS is conceptualised as a combination of system characteristics and the task (Beckmann & Goode, 2017). It is determined by the number of information cues that need to be processed in parallel when performing an instructed task. In case of system exploration, where the task is to acquire effect knowledge, system complexity can, for instance, be indicated by the number of effects associated with each of the input variables (e.g., Beckmann & Goode, 2017; Fischer, Greiff, & Funke, 2012). In contrast, difficulty is the reflection of a problem solver's ability to deal with the complexity imposed by the task and its system characteristics, as reflected in performance scores. Accordingly, higher levels of complexity tend to result in lower knowledge acquisition performance (Beckmann & Goode, 2017; Kluge, 2008). The complexity of a system might further have an impact on the association between learning test performance and knowledge acquisition performance, as systems with more complex demands likely provide a more disadvantageous situation for individuals with lower learning ability.

Autonomous changes, that means changes in the states of output variables that are *independent* from input variables, tend to contribute to the complexity of a system in a specific way, especially in the context of knowledge acquisition (e.g., Frensch & Funke, 1995a; Funke, 2001; Greiff, Krkovic, & Nagy, 2014). Stadler, Niepel, and Greiff (2019) have shown that individual performance differences in systems with versus without autonomous changes can be modelled as distinct portions of variances in factor analyses. In terms of the potentially moderating effect of complexity on the link between learning test and knowledge acquisition performance, autonomous changes might therefore demand special considerations. In other words, it is to be expected that – rather than in a mere additive sense – complexity indicators such as the relative number of effects and autonomous changes have an interactive effect on CPS performance.

1.2.2. Experience with the test as learning?

General conceptualisations of learning refer to it as a process resulting in enduring changes in behaviour (or the capability to behave) that is facilitated by practice and other forms of experience (e.g., Schunk, 2020). Individual differences in this process are assumed to be reflected in learning ability. Previous studies that explored the link between learning and CPS performance were conducted by using one CPS system as stimulus material and thus investigated changes in performance over time *within* a system (e.g., Beckmann, 1994; Beckmann & Guthke, 1995). For a MicroDYN approach, that means where problem solvers are confronted with a multitude of different (micro-)systems, learning processes (and therefore also individual differences in learning ability) should similarly be reflected in changes over time, which in this case means changes in performance *between* systems. From this follows that, depending on the relative position of a particular system within the MicroDYN test, performance might be influenced by the accumulated experiences made whilst dealing with the preceding systems. However,

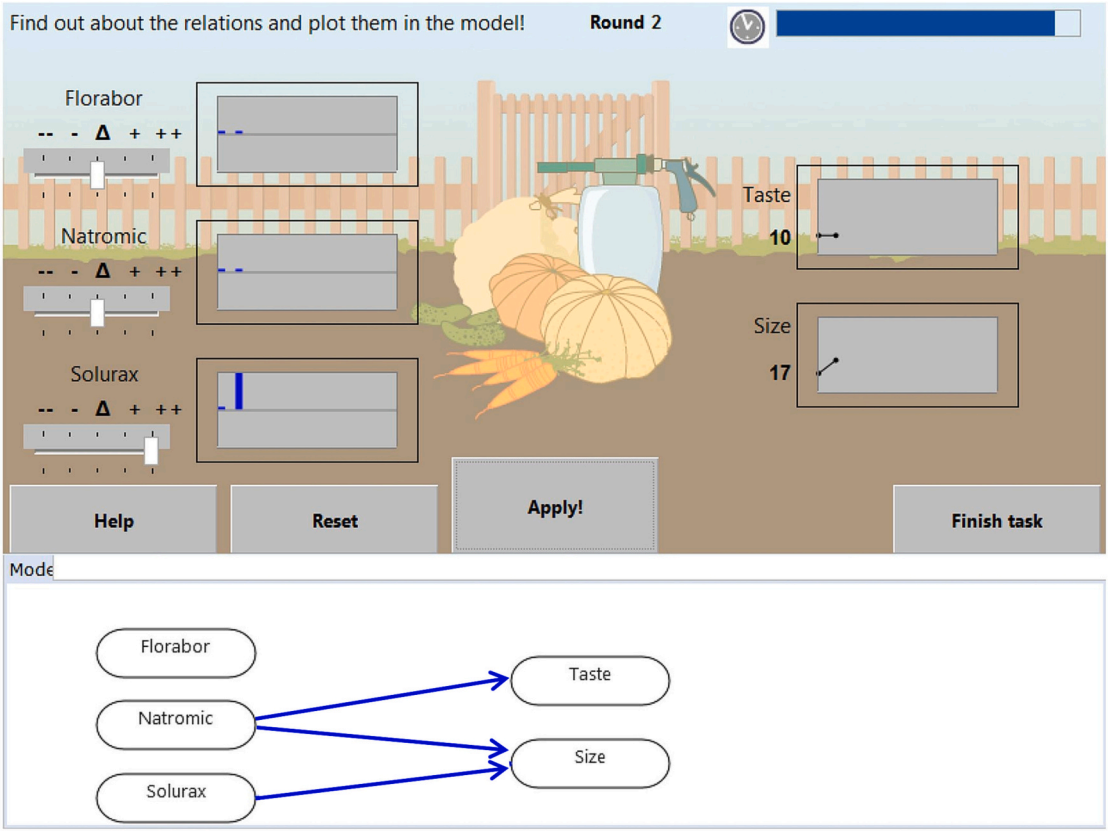


Fig. 2. *Screenshot of the MicroDYN System "Planting Pumpkins" in the Knowledge Acquisition Phase Note.* In this example system "Planting Pumpinks", problem solvers are asked to find out how the input variables Florabor, Natromic, and Solurax (left side) affect the output variables Taste and Size (right side) by manipulating them and observing the effects, and to enter the effects to their understanding as blue arrows into the model depicted underneath. Figure adapted from "A longitudinal study of higher-order thinking skills: working memory and fluid reasoning in childhood enhance complex problem solving in adolescence" by Greiff et al., 2015, *Frontiers in Psychology*, 6, 1060. CC BY 4.0. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

it yet remains open whether we can reasonably assume these learning processes in a single system in comparison to a MicroDYN approach to be the same, so to expect similar changes in performance over time. The answer to this question constitutes a key element for the generalisability of previous findings on the link between learning and CPS to a MicroDYN approach. When investigating knowledge acquisition performance in relation to learning throughout MicroDYN systems, the relative position of the system in the course of the test therefore also needs to be considered.

1.2.3. Systematicity of system exploration

Previous research has established that success in acquiring knowledge about the causal structure of a CPS system depends on how systematically it is explored (e.g., Beckmann & Goode, 2014; Kröner, Plass, & Leutner, 2005; Wüstenberg, Greiff, Molnár, & Funke, 2014). Research on the role of systematic exploration of CPS systems discusses various strategies, with the so-called VOTAT strategy ("Vary-One-Thing-At-a-Time", i.e., single interventions for every input variable; Tschirgi, 1980; see also, e.g., Inhelder & Piaget, 1958; Shayer, 2008; Zimmerman, 2007) being arguably the most prominent one (for an overview, see, e.g., Wüstenberg, Stadler, Hautamäki, & Greiff, 2014). In the context of a MicroDYN approach to CPS, the frequency of VOTAT was found to increase over the course of the eight systems of the test (Wüstenberg et al., 2012). However, the VOTAT strategy is only functional if it is preceded by a zero intervention, in which all input variables are kept at zero in order to identify potential autonomous changes in the to be explored system. Therefore, the "desired" strategy would begin with a zero intervention, followed by every input variable varied at a time, which has been labelled the strict VONAT strategy ("Vary-One-or-None-At-a-Time", Beckmann, Birney, & Goode, 2017, p. 4; Beckmann, 2019; Beckmann & Goode, 2014, p. 279, 2017, p. 9).[2]

Systematic exploration behaviour is necessary (but not sufficient) for successful performance (Beckmann et al., 2017), so when considering that systematicity in system exploration behaviour increases over time, it seems plausible that performance improvements across MicroDYN systems might be – at least partly – enabled through some form of strategy learning. In other words, systematicity in exploration behaviour may act as a moderator of the relationship between the amount of experience with CPS systems over time and knowledge acquisition performance. It further seems plausible that systematic exploration behaviour might moderate the effects of complexity on knowledge acquisition performance, as detecting effects without systematic exploration might be possible in systems of very low complexity or systems in which the chosen variable labels provide "strong hints" regarding which input might be linked to which output. For systems with higher levels of complexity and/or neutral variable labels, systematic exploration is indispensable for successful knowledge acquisition.

[2] Previous studies have often considered only whether a zero intervention or a VOTAT strategy was applied, ignoring the order of the steps. However, this can lead to "inconclusive experiments" (de Jong & van Joolingen, 1998) in which ambiguous information is produced (e.g., if one starts with a single variable intervention, one would not be able to decide whether the observed changes are due to the intervention or to autonomous changes). Following the strict VONAT approach, each step produces unambiguous information.

1.3. The present study

In the present study we investigate whether and how individual differences in CPS performance are associated with learning test performance. Our analysis is guided by conceptually informed considerations of various characteristics of the CPS systems in question. We analysed MicroDYN (Greiff et al., 2012) performance, focussing on the knowledge acquisition phase, as a measure of CPS ability based on a sample of $N = 241$ participants. Their performance in the "Adaptive Figurenfolgenlerntest" (ADAFI, Beckmann & Guthke, 1999; Guthke & Beckmann, 2000; Guthke, Beckmann, & Stein, 1995) served as the operationalization of learning ability. Although the relationship between CPS and learning ability has been studied previously (e.g., Beckmann, 1994; Beckmann & Guthke, 1995; Guthke & Beckmann, 2003; Guthke, Beckmann, & Stein, 1995), we conceive the current investigation as rather explorative, because we are, to our current knowledge, the first to investigate the role of learning ability in the context of a MicroDYN approach to CPS assessment. Despite the explorative nature of this study, we have pre-registered our expectations and analysis strategy prior to accessing the already existing dataset (https://doi.org/10.17605/OSF.IO/E5KZJ).

In summary, the overarching aim of this study is to explore the role of learning in MicroDYN performance. We address this aim by investigating the association between learning test performance and MicroDYN performance, as well as whether and how it is impacted by complexity-related system characteristics and systematicity in exploration behaviour. We further explore additional interactions between system and person characteristics for a closer understanding of how the numerous variables included might affect performance in interaction with each other. An overview of all relevant study variables is provided in Fig. 3.

1.3.1. Expectations and hypotheses

In our prepending analyses to replicate the role of system characteristics and systematicity on MicroDYN performance, we expected higher levels of system complexity (as indicated by the relative number of effects) and the presence (vs. absence) of autonomous changes to be associated with a decrease in MicroDYN performance. We further expected the position of a system in the test (as a proxy for experience/time with the task) to explain performance variance. We however had no directional expectation for this effect, as it remains open whether, for example, positive effects of learning or negative effects of fatigue might dominate, which could further confound the effects of complexity in different directions. Further, we expected that higher levels of systematicity in exploration behaviour (i.e., strategy use) are associated with better MicroDYN performance. Establishing the existence of expected effects of system complexity, autonomous changes, system position, and strategy use on knowledge acquisition performance could be seen as conceptual replications of previous findings on the role of system characteristics and strategy use in CPS.

Our main analysis focused on the association between learning test and MicroDYN performance. We expected learning test performance to be generally positively related to MicroDYN performance while simultaneously accounting for the effects of system characteristics and strategy use on MicroDYN performance as mentioned in our prepending analyses. As we assume that the relevancy of learning ability increases with the complexity of the CPS system, we expect this to be reflected by a positive learning test performance by complexity interaction on MicroDYN performance. In focussing on the presumed special role of autonomous changes to the complexity of a system, we expect a similar result pattern in terms of a positive learning test performance by autonomous change interaction on MicroDYN performance.

Lastly, we aimed to explore further interactions between system and person-related characteristics (i.e., complexity, autonomous changes, system position, strategy use, learning test performance), and MicroDYN performance. Based on the assumption that, for successful knowledge acquisition, strategy use is even more important in later systems with higher levels of complexity, we deemed it likely that strategy use could interact with a system's position in the test, as well as its complexity. Furthermore, complementing the assumption that complexity and autonomous changes would lead to decreased performance, we deemed it likely that both variables in combination would lead to even more performance decrements than their additive effects alone, as indicated by a complexity-autonomous changes interaction.

2. Method

2.1. Participants

The study is part of a larger research project aiming at the construct validity of cognitive ability tests (see https://doi.org/10.17605/OSF.IO/2YM3X). For the present study, we only considered participants who provided data on MicroDYN and the learning ability test, resulting in a sample of $N = 241$ participants (76% female) with a mean age of 23.22 years ($SD = 3.57$) of which nearly all were university students.

2.2. Measures

2.2.1. MicroDYN

We used the MicroDYN (Greiff et al., 2012) approach to assess CPS. MicroDYN is based on the formal framework of linear structural equations (Funke, 1985). Previous research has demonstrated the reliability (e.g., Wüstenberg et al., 2012) and validity (e.g., Greiff et al., 2013; Neubert, Kretzschmar, Wüstenberg, & Greiff, 2015) of the MicroDYN approach. In a typical MicroDYN system, participants were first asked to explore an unknown system in order to acquire knowledge about causal relations between input and output variables (and sometimes, effects output variables have on themselves, i.e., autonomous changes) and record their proposed causal relations into a causal diagram (knowledge acquisition phase; see Fig. 2). In a second phase, participants were then asked to apply their knowledge by manipulating the system to reach a given goal state (knowledge application phase). For the present study, we used a version of MicroDYN in which ten systems were presented. Participants had a maximum of five minutes per system to tackle both tasks. In the present analysis we only focus on the task of knowledge acquisition. As recent research provided evidence for the importance of autonomous changes (e.g., Stadler et al., 2019), every second system of the ten included had autonomous changes (see Supplement section 1.2 for formal task descriptions). The ten systems differed regarding the number of input and output variables and the number of relationships between them, which serves as manipulation of complexity across these systems. We used the relative number of effects as a proxy for system complexity (Beckmann & Goode, 2017). It is defined as the maximum ratio between the number of actual effects and the number of possible effects any of the input variables has on any of the output variables in the given system. For simplicity, we will use "complexity score" for this variable that serves as proxy for system complexity in relation to the knowledge acquisition task. For example, a system as depicted in the lower section of Fig. 2 has a complexity score of 1, because the input variable with the largest relative number of effects (*Natromic*) has two out of two possible effects on the output variables. We further derived an alternative indicator for complexity: Instead of focusing on the input variable with the largest relative number of effects, we aggregated the relative number of effects across all input variables in the given system ("complexity sum score"). Aggregating the relative number of effects across all variables in a model as depicted in Fig. 2 results in a complexity sum score of 1.5 because the relative number of effects for the first input variable (*Florabor*; zero out of two possible effects, therefore 0), for the second input variable (*Natromic*; two out of two, therefore 1), and for the third input variable (*Solurax*; one out of two, therefore 0.5) are summed up. Complexity scores for both operationalisations for the ten systems used in this study are presented in the Supplement section 1.2.

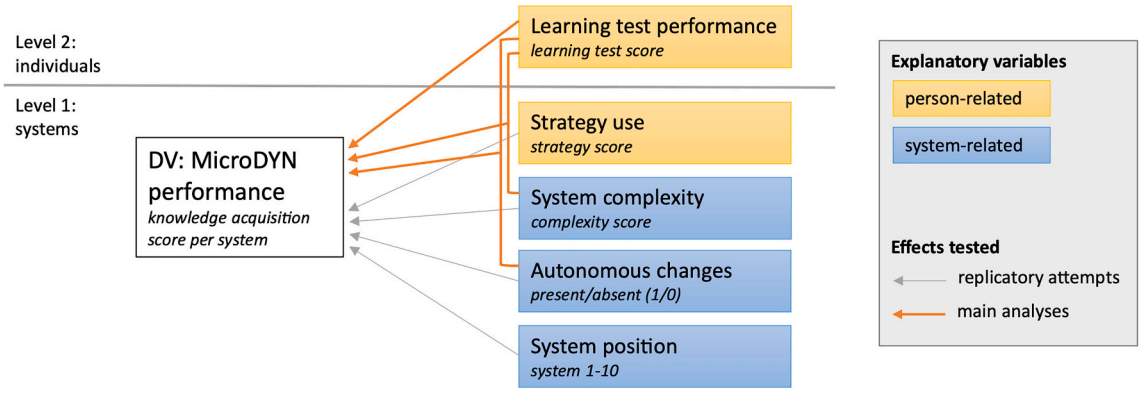


Fig. 3. *Overview on the Data Structure and Variables Included in the Current Analysis.*
Note. The presentation of 10 MicroDYN systems per individual leads to a nested data structure with explanatory variables varying either on level 1 (systems) or level 2 (individuals), and varying either systematically (marked in blue) or depending on individual behaviour (marked in yellow). Grey arrows represent effects tested in our prepending analysis (replicatory attempts), orange arrows represent the main and interaction effects tested in our main analyses. For a better overview, our planned exploratory analyses are not represented in this Figure.

Knowledge acquisition performance was scored based on the notion of a "relationship detection task" for which a two-high-threshold model (Snodgrass & Corwin, 1988) was employed (for details, see Beckmann, 1994 and Beckmann et al., 2017; for a similar approach in CPS research, see, e.g., Beckmann & Goode, 2014; Beckmann & Guthke, 1995; Goode & Beckmann, 2016)). We used the sensitivity index P_r which represents the difference between the rate of correctly identified relationships (hit rate) and the rate with which relationships were identified that do in fact not exist in the given system (false alarm rate). This index reflects the accuracy with which existing and nonexisting relationships, including autonomous changes, have been identified by the problem solver. The two high-threshold approach also enables the derivation of a bias index B_r which is based on the false alarm rate relative to the sensitivity of correctly identifying the relationship structure. This bias index reflects a problem solver's tendency to either "see" or "not to see" relationships when guessing.

As discussed, success in acquiring knowledge about the causal structure of a system depends on the level of systematicity of the exploration behaviour. We thus created a "strategy score" to reflect the degree of systematicity with which a problem solver has interacted with a system. The score reflects whether problem solvers perform two components of a systematic exploration, ideally completely and in the optimal order. That is at least one zero intervention followed by single interventions applied to each of the input variables. The derived strategy score represents the sum of points, with each given for (1) at least one single intervention, (2) at least one zero intervention, (3) at least one single intervention *after* a zero intervention, and (4) single interventions for all input variables *after* a zero intervention. This results in a score ranging from 0 (no component of systematicity present) to 4 (representing the execution of the ideal exploration strategy).

2.2.2. *Learning test*

We administered the "Adaptive Figurenfolgenlerntest" (ADAFI) from the "Adaptive Computergestützte Intelligenzlerntestbatterie" (ACIL; Beckmann & Guthke, 1999). The ADAFI is a computerised test of learning potential that follows the principles of Dynamic Testing[3] in the fashion of a so-called sandwich paradigm (i.e., test and training components are integrated within the same testing session; Sternberg & Grigorenko, 2002). Test takers are asked to respond to series completion items that contain abstract geometric figures as stimuli. The item pool of the ADAFI comprises a total of 32 items that are subdivided into three complexity levels. Roughly speaking, complexity levels are determined by the number of dimensions that are to be considered for determining the rule that governs the series of figures in an item. For instance, whilst the rule for items in complexity level I is based on the change of colour or shape of the individual elements across the figure series, the complexity for items in level II is based on the combination of colour and shape. Items in complexity level III represent series of figures that vary systematically in terms of colour, shape, and gestalt.

In learning tests such as the ADAFI there are two categories of items for each complexity level, which are test items and training items. After correct responses to test items in one complexity level, the test taker progresses to test items of the next complexity level. In case of an incorrect response to a test item, the algorithm presents the test taker with a training item of similar complexity. Incorrect responses to those items will be followed up with error specific hints, which are intended to provide the test taker with opportunities to learn how to tackle the reasoning challenge posed by items of increasing complexity. Additional thinking prompts and retries are provided until items at this complexity level can be solved successfully. As a result of this doubly adaptive procedure (i.e., a combination of failure-adaptive feedback and thinking prompts, and "classical" adaptive testing), test takers vary in the amount of training items needed and the amount of thinking prompts used. The combination of both represents the number of steps through the item pool, which is the operationalisation of learning test performance (for more details see Guthke & Beckmann, 2000). The fewer steps needed to work through the item pool the better the test performance. The learning test score used in the analyses presented here is the reverse of the number of steps in relation to the optimal and pessimal number of steps, which is akin to the POMP score approach (Cohen, Cohen, Aiken, & West, 1999). The learning test score ranges from 0 to 1, with higher scores reflecting better performance, suggesting higher learning ability.

2.3. *Procedure*

Participants engaged in daily online assessments over the course of one week (see https://doi.org/10.17605/OSF.IO/2YM3X). Learning ability was assessed on the first day, whereas MicroDYN was assessed on the fifth day. All assessments tools were presented on a computer. Participants received an invitation via e-mail every day and were able to attend to the respective assessment independently on their own computer. Participants received 9.50€ per hour as compensation for their participation in the study. Ethics approval for the study was granted from the ethics committee for psychological research of the University of Tübingen.

[3] For an overview and a critical discussion of the concept of Dynamic Testing and learning tests, see, for example, Beckmann (2006, 2014), Elliott, Resing, and Beckmann (2018), and Lidz and Elliott (2000).

2.4. Statistical analysis

To accommodate the nested structure of the data, capturing both intra- and interindividual variability (i.e., 10 MicroDYN systems nested within $N = 241$ individuals, see Fig. 3), we have adopted a multilevel model approach in our data analyses. We built up the model in a stepwise fixed manner with system- and person-related, partly time-variant characteristics (i.e., system position, complexity score, autonomous changes, strategy score, learning test score), and their interactions as variables regressed in a fixed order on MicroDYN performance operationalised as sensitivity index P_r (for the order of included effects per step, see Supplement). We decided on a stepwise approach in order to replicate previously reported, isolated effects of variables (e.g., the effect of complexity without simultaneously accounting for system position) in our data before investigating our main research question. This approach further allowed us to investigate whether these effects are not only replicable, but also remained stable when further variables are accounted for. We decided against removing any non-significant lower-level effects, as this would prevent the investigation of potential higher-order interactions.

For each step in building up the model, we evaluated whether the added main or interaction effect met two criteria. First, we consider an improved model fit as a decrease in the Bayesian information criterion (BIC; Schwarz, 1978) of at least 6 (see Raftery, 1995). Secondly, we consider a meaningful extension of a model to result in a higher proportion of explained variance in the criterion. Typically, this is indicated in an increase in the overall R^2. As discussed in Nakagawa and Schielzeth (2013), R^2 estimated for multilevel models can also decrease when a new predictor is added to a model. Hence, we focus on an increase in *conditional* R^2 (which also accounts for random effects) in each multilevel model to determine whether the addition of the respective predictor to the model makes a meaningful contribution (as estimated by using the r2 function from the sjstats package in R; Lüdecke, 2018).

As model fit and variance explained do not reveal the direction of effects, we also evaluated model coefficients and their 95% confidence intervals (CIs). For a better contextualisation and interpretation of the potential effects identified in our analyses we report confidence intervals as plausible values in the population (Cumming, 2014), but refrain from explicit hypothesis testing on the basis of *p*-values. We estimated the multilevel models with full (instead of restricted) maximum-likelihood estimation to be able to compare fixed effects with the BIC (cf. Hox, 2010). All metric variables were z-standardised to minimise the likelihood of model convergence issues.

We preregistered several additional and sensitivity analyses which serve the purpose to discern the robustness of the result patterns obtained in the main analyses. In these analyses we used the complexity sum score and the bias index (B_r) as described above. Results of these analyses are in line with results presented below (complexity sum score), or indicate that the operationalisation was not suitable for the measurement context (bias index B_r). Results of these analyses are reported in the Supplement.

2.5. Transparency, openness, and reproducibility

We registered our expectations and planned analyses in relation to this already existing dataset on 14th September 2022 prior to accessing the data for the present study. The analysis was performed as planned. Open data, a reproducible analysis script as R markdown, a codebook, and the online supplement material are permanently available under https://doi.org/10.17605/OSF.IO/E5KZJ.

3. Results

3.1. Descriptive analysis

Descriptive statistics are provided in Table 1. Mean MicroDYN performance indices (sensitivity index P_r per system) showed the to be expected pattern of higher performance scores in systems without autonomous changes (odd system position numbers) in contrast to systems with autonomous changes (even system position numbers). This is further reflected descriptively in more left-skewed distributions of performance scores for systems without autonomous changes, in which 80.52% of participants achieved a perfect score of 1.0, compared to only 44.05% of participants in systems with autonomous changes. Table 1 also presents descriptive statistics for the learning test scores (based on the number of steps, transformed to a reversed POMP score) across the three complexity levels and the total score. As to be expected, average performance, skewness, and kurtosis tended to decrease over the course of the test, suggesting that, although performance scores are relatively high, their degree of variability seems to reduce the risk of a ceiling effect. Data further indicate that in complexity level I, 82.16% of participants achieved optimal performance as indicated by requiring only the minimal number of steps, while it was 42.32% in complexity level II, and 16.18% in complexity level III. Overall, only 12.86% of participants achieved the maximum scores in all three complexity levels.

Table 1
Descriptive Statistics for MicroDYN and the Learning Test.

Measure		Performance			
		M	*SD*	*Skewness*	*Kurtosis*
MicroDYN (P_r)	1	0.87	0.36	−3.07	11.31
	2	0.48	0.47	−0.24	1.73
	3	0.93	0.21	−3.63	18.28
	4	0.67	0.36	−0.61	2.20
	5	0.89	0.26	−2.80	10.88
	6	0.69	0.36	−0.75	2.51
	7	0.92	0.2	−3.10	12.42
	8	0.75	0.28	−0.94	3.02
	9	0.91	0.22	−2.90	11.23
	10	0.56	0.44	−0.37	1.51
	Total	0.77	0.25	−1.18	5.16
Learning Test Score	I	0.97	0.08	−1.90	5.05
	II	0.93	0.08	−0.95	3.44
	III	0.86	0.10	−0.55	2.89
	Total	0.90	0.08	−0.75	3.51

Note. $N = 241$. P_r = Sensitivity index. MicroDYN consists of 10 systems, resulting in 10 performance scores per participant, which are presented in the table. Items in the learning test (ADAFI) are subdivided into three complexity levels, which are presented in the table (in addition to the total performance that was used for the analysis). Performance in the learning test is operationalised as a reversed POMP score (Cohen et al., 1999) based on the number of steps, resulting in values ranging from 0 to 1, with higher values indicating better performance.

The average strategy score and the relative frequencies of its four components, reflecting the systematicity of exploration behaviour, are depicted over the course of the ten systems in Fig. 4. Nearly all participants applied single interventions at some stage, while the relative frequencies of the other strategy criteria were considerably lower. The total strategy score had a mean of $M = 2.4$ points ($SD = 0.73$).

3.2. Prepending analysis

We started the analysis with a null model (see Table 2: Null Model), which only included a random intercept for participant but no other effects. An analysis of the variability indicated that 69% of the variability in the data occurred within participants, and 31% between participants. As a first step, we added system position and the complexity score separately into the null model to investigate whether they might contain redundant information (as they both increase over the course of the test, but should contribute to the difficulty of a system in distinct ways). Model fit decreased and conditional R^2 remained the same size when system position was included ($BIC = 6315$, $R^2_{\text{conditional}} = 0.313$), indicating that system position alone did not add meaningful information to the model. When adding complexity alone into the null model,

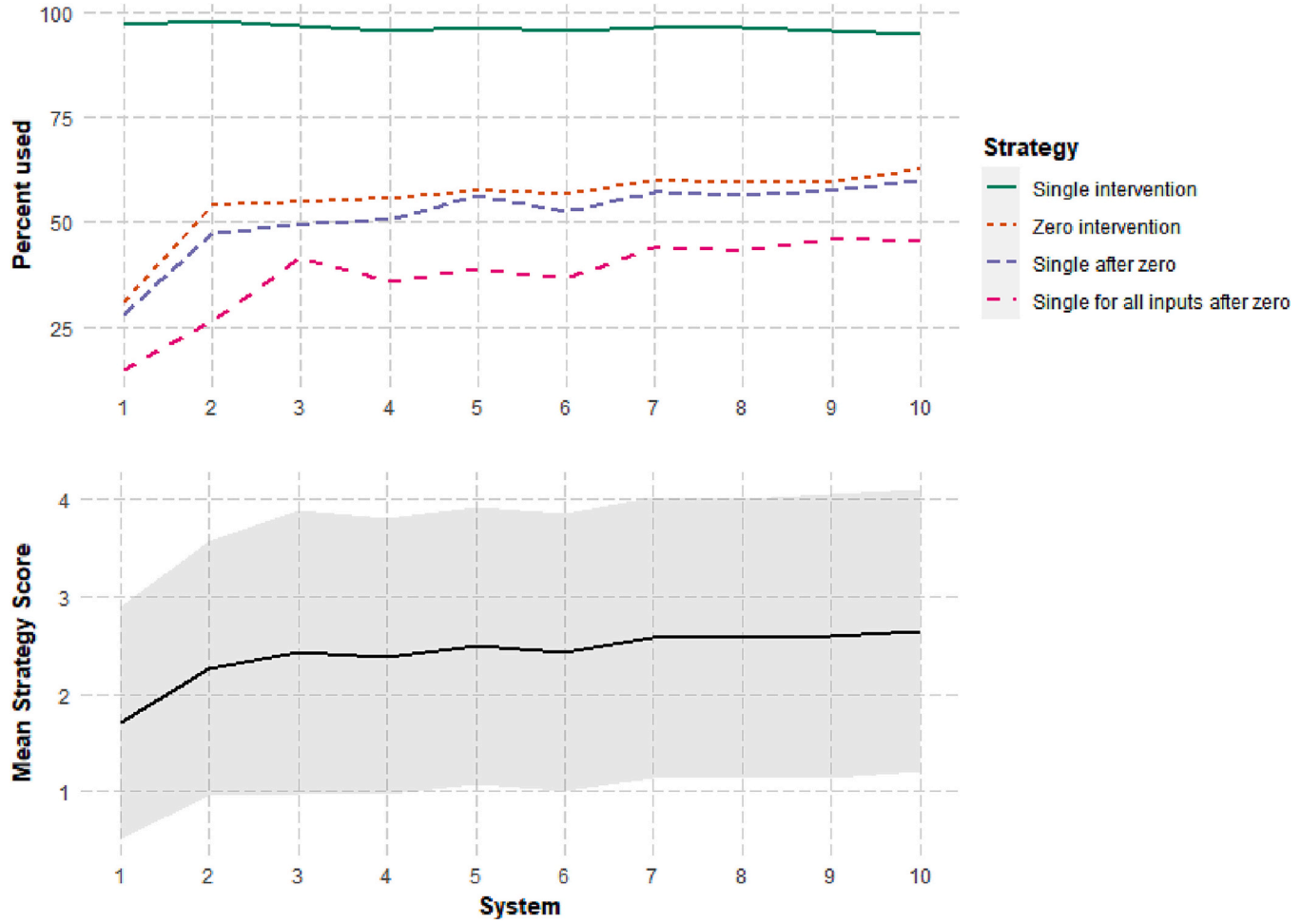


Fig. 4. *Relative Frequencies of the Four Criteria of the Strategy Score and the Mean Strategy Score Across the Ten Systems Note*. Scores for each strategy were coded dichotomously (0 or 1) per participant, relative frequencies therefore indicate the percentage of participants who applied the respective strategy. The strategy score reflects the sum of these four criteria. The shaded area reflects ±1 *SD*.

model fit and explained variance both increased ($BIC = 6159$, $R^2_{\text{conditional}} = 0.361$). Subsequently both system position and complexity were simultaneously added into the null model. Model fit and explained variance further increased ($BIC = 6117$, $R^2_{\text{conditional}} = 0.376$), indicating that, in combination, both system position and complexity contribute systematic information to the model and are therefore not redundant.

We then added autonomous changes (present vs. absent) as a further indicator of complexity to the previous model. Model fit and explained variance showed a pronounced increase ($BIC = 5738$, $R^2_{\text{conditional}} = 0.482$), indicating that autonomous changes also contribute meaningfully to the explanation of MicroDYN performance. As a last step of our prepending analysis model (see Table 2: Prepending Analysis), we added the strategy score to the previous model. Model fit increased, while explained variance decreased ($BIC = 5591$, $R^2_{\text{conditional}} = 0.425$), indicating that the inclusion of the strategy score into the model might improve fit with the data but does not reach our preregistered criterion of an increase in explained variance in order to interpret the effect as meaningful. All coefficients in this model showed confidence intervals not including zero.

In line with expectations, we found negative effects for both complexity and autonomous changes, indicating lower performance with higher complexity and/or in the presence of autonomous changes. Also as expected, the effect of the strategy score was clearly positive, indicating better performance with a higher strategy score, although it needs to be noted that, strictly speaking, the effect did not reach our predefined criterion of an increase in explained variance. We further found a positive effect of system position, indicating that more experience over the course of the test tended to result in better performance, which supports the notion that system position and complexity contribute different information to the model.

3.3. Main analysis

To address the main question posed in this paper, which was whether performance in MicroDYN is linked to learning, we included the learning test score and its interactions with complexity and autonomous changes in the model including all variables mentioned in the previous section, again in a stepwise process. Firstly, the learning test score was added to the model, which resulted in an improvement in model fit and an increase in explained variance ($BIC = 5560$, $R^2_{\text{conditional}} = 0.443$). Model fit and explained variance also increased when the interaction of the learning test score with complexity was added ($BIC = 5552$, $R^2_{\text{conditional}} = 0.447$), as well as when the interaction of the learning test score with autonomous changes was added ($BIC = 5542$, $R^2_{\text{conditional}} = 0.452$). In summary, all three effects added into the model seem to contribute systematic information in terms of explaining variance in knowledge acquisition performance in MicroDYN.

When including all three effects in relation to our main research question (in addition to the effects investigated in the prepending analysis), the reported effects modelled in the prepending analysis remained largely unchanged (see Table 2: Main Analysis). More specifically, the inclusion of the learning test score revealed the expected

Table 2
Selected Steps from the Stepwise Multilevel Model Explaining MicroDYN Performance.

	Null Model			Prepending Analysis (Replication)			Main Analysis			Exploratory Analysis		
Predictors	β	*95% CI*	*SE*	β	*95% CI*	*SE*	β	*95% CI*	*SE*	β	*95% CI*	*SE*
Intercept	−0.01	[−0.08, 0.07]	0.04	0.47	[0.40, 0.54]	0.04	0.47	[0.40, 0.54]	0.03	0.49	[0.41, 0.58]	0.04
System position				**0.10**	[0.05, 0.14]	0.02	**0.10**	[0.06, 0.14]	0.02	**0.12**	[0.07, 0.16]	0.02
Complexity score				**−0.13**	[−0.18, −0.08]	0.03	**−0.13**	[−0.18, −0.08]	0.03	**−0.17**	[−0.27, −0.08]	0.05
Autonomous changes				**−0.95**	[−1.04, −0.86]	0.05	**−0.95**	[−1.04, −0.86]	0.05	**−0.98**	[−1.08, −0.88]	0.05
Strategy score				**0.32**	[0.28, 0.36]	0.02	**0.29**	[0.25, 0.33]	0.02	**0.31**	[0.27, 0.35]	0.02
Learning test score							**0.10**	[0.03, 0.16]	0.03	**0.12**	[0.05, 0.18]	0.03
Learning test score * complexity score							−0.02	[−0.05, 0.02]	0.02	0.02	[−0.01, 0.06]	0.02
Learning test score * autonomous changes							**0.15**	[0.08, 0.22]	0.04	**0.10**	[0.03, 0.17]	0.04
System position * strategy score										**0.15**	[0.12, 0.19]	0.02
Complexity score * strategy score										**−0.22**	[−0.26, −0.19]	0.02
Complexity score * autonomous changes										0.04	[−0.04, 0.13]	0.04
Random Effects												
σ^2	0.69			0.53			0.52			0.48		
τ_{00}	0.31			0.14			0.12			0.11		
ICC	0.31			0.21			0.19			0.19		
Conditional R^2	0.31			0.43			0.45			0.49		
BIC	6307			5591			5542			5387		

Note. $N = 241$, with 10 observations per participant. System position = number ranging from 1 to 10; Complexity score = largest relative number of effects among the input variables; Autonomous changes = absent [intercept] vs. present; Strategy score = sum score of four strategy criteria ranging from 0 to 4; Learning test score = reversed POMP score of the number of steps ranging from 0 to 1; σ^2 = residual variance, or within-participant variance; τ_{00} = random intercept variance, or between-participant variance; ICC = Intraclass correlation coefficient. Effects interpreted as meaningful in bold. The table represents selected steps of our analysis as the basis to interpret findings with regard to our expectations. Results of all model steps are displayed in the Supplement.

Fig. 5. Interaction between Learning Test Score and Autonomous Changes on the Sensitivity Index.
Note. The figure presents regression lines (based on model step 8, see Supplement section 5) for the learning test score at the two levels of autonomous changes (present or absent) with 95% confidence intervals. The learning test score is indicated by the reversed POMP score of the number of steps needed in the learning test.

positive effect on model fit and explained variance overall. The same is true for the inclusion of the respective interaction terms with the complexity score and autonomous changes. The central findings of the analysis are that knowledge acquisition performance was generally better in systems without autonomous changes (see Fig. 5), that knowledge acquisition was positively associated with learning test scores, and that this association was more pronounced in systems with autonomous changes. Including the interaction between the learning test score and complexity score did not add further to the explanation of MicroDYN performance once the interaction between the learning test score and autonomous changes was included.

3.4. Exploratory and additional analyses

In our preregistered exploratory analysis, we investigated three further two-way interactions (see Table 2: Exploratory Analysis), for which we found improved model fit and increased explained variance ($BIC = 5387$, $R^2_{\text{conditional}} = 0.492$). Firstly, we found an interaction between system position and the strategy score, indicating a positive association between the strategy score and P_r, suggesting higher relevancy of systematicity for systems presented later in process (see Figure in Supplement section 6.4). Secondly, we found a negative interaction between the complexity score and the strategy score, suggesting that – in conjunction with the positive association between the strategy score and P_r – the employment of a highly systematic exploration strategy seems less positively associated with performance outcomes in more complex systems. This finding stands in contrast with our pre-registered expectations as we would have expected a positive effect. We found no interaction effect between the complexity score and autonomous changes. Generally, the combination of these various main and interaction effects of system and person characteristics explains nearly half of the total variance in MicroDYN performance.

Full results of all model steps reported in the prepending, main and exploratory analysis are displayed in the sections 3–5 of the Supplement. The full model, including all higher-order interactions of all predictors, revealed considerably larger confidence intervals than all other models (probably due to the much larger number of degrees of freedom in that model), with no higher-order interactions with confidence intervals not including zero (see Supplement section 5).

We conducted an unplanned (i.e., ad hoc), additional analysis to further explore the unexpected negative (instead of positive) interaction effect between the complexity score and the strategy score. This effect might have been caused by the fact that, in the MicroDYN test used in our study, systems with autonomous changes tended to have lower complexity scores than systems without autonomous changes (mean complexity was $M = 0.50$ with, versus $M = 0.67$ without autonomous changes). In other words, the complexity score and autonomous changes were confounded. We therefore added the interaction between the strategy score and autonomous changes into the model including all other previously mentioned effects to explore this interpretation further. Indeed, the interaction between the strategy score and autonomous changes improved model fit and increased explained variance ($BIC = 5281$, $R^2_{\text{conditional}} = 0.392$), and was highly pronounced ($\beta = 0.49$, 95% CI [0.40–0.58], $SE = 0.05$), while both the positive interaction effect of the strategy score with system position, as well as with the complexity score, disappeared completely (for full results see Supplement section 7.1).

4. Discussion

The main objective of this study was to explore the role of learning in MicroDYN performance. To that end we analysed an existing data set of MicroDYN performance obtained from $N = 241$ undergraduate students employing a hierarchical set of steps. The analyses targeted three aims. First, to establish whether previous findings related to effects of system characteristics and systematicity in exploration behaviour on CPS performance can be replicated in a MicroDYN context. Second, to investigate the associations between learning test performance and MicroDYN performance, as well as whether and how it is impacted by complexity-related system characteristics. And third, to explore further interactions between system and person characteristics. Each of these analysis steps are underpinned by a set of conceptually derived expectations, which we – together with an *ex ante* determined analysis strategy have pre-registered.

4.1. System and person characteristics affect performance

With regard to the first aim, results obtained from our analyses lend support to the generalisability of previous findings. As has been shown for other CPS tests that are not based on the minimal/multiple complex system approach (Funke & Greiff, 2017; Greiff et al., 2015), system characteristics of MicroDYN tests tend to influence how problem solvers interact with these systems. They also have an impact on problem solvers' level of success in tackling them. More specifically, the presence of autonomous changes as a system characteristic, and the employment of a systematic approach to system exploration as a person characteristic, showed the strongest associations with MicroDYN performance.

This finding corroborates evidence for the importance of a systematic exploration of MicroDYN systems in order to successfully acquire knowledge about their individual causal structure, especially when they feature some form of dynamisms as in the presence of autonomous changes. The comparison between system position and complexity (operationalised as the maximum of the number of relative effects across all input variables in a system) in terms of their predictive utility of knowledge acquisition performance can be interpreted as confirmation for (a) distinguishing between complexity and difficulty (e.g., Beckmann, 2019), (b) determining task demands *ex ante* (e.g., complexity metric) rather than relying on post hoc interpretations of difficulty estimates, and (c) our chosen approach to the operationalisation of system complexity.

4.2. Knowledge acquisition performance is associated with learning

With regard to our second aim, which constitutes the master theme for our analyses presented here, we found evidence for an association between learning ability (measured by using a learning test for abstract reasoning) and knowledge acquisition performance in MicroDYN. This association tends to be substantially stronger for systems that require the identification of autonomous changes. The absence of an interaction effect of learning test scores and complexity draws attention as it seems to suggest that learning in MicroDYN is unrelated to increases in complexity across the ten systems.

4.3. Role of systematicity in the context of different system characteristics

To further our understanding of the processes underpinning MicroDYN performance, our third aim was to explore effects of interactions between selected system and person characteristics on knowledge acquisition scores. Results seem to suggest that performance in systems presented later in the MicroDYN test seems to depend more strongly on a systematic approach to knowledge acquisition. Furthermore, the negative interaction between complexity and strategy, which seems to suggest that the employment of a systematic strategy is *less* important in system with higher complexity, appears to be counterintuitive and contradictory at first. But when considering that systems with the lowest complexity index (i.e., systems 4 and 6) are systems *with* autonomous changes and systems with the highest complexity index (i.e., systems 7 and 9) are systems *without* autonomous changes, this finding becomes more plausible. A high strategy score depends on the employment of a zero intervention that precedes single interventions. Zero interventions are essential to a successful identification of autonomous changes. A failure to employ zero interventions, which would result in a lower strategy score, in systems without autonomous changes tends to be

inconsequential, while it is detrimental for performance in systems with autonomous changes. Considering that systems without autonomous changes had the highest complexity scores, the negative interaction between complexity and strategy was thus most likely caused by the fact that complexity is confounded by autonomous changes.

When controlling for the confounding effect of autonomous changes,[4] both interactions of the strategy score with system position and complexity disappeared. Although we did not explicitly include this effect in our expectations, this finding also provides strong support for the notion that a systematic approach to knowledge acquisition is particularly important in systems with autonomous changes. In sum, this finding highlights the importance of autonomous changes in complex problem solving tasks; they are, after all, an essential differentiator between "conventional" problem solving and complex problem solving (see, e.g., Frensch & Funke, 1995a; Funke, 2001; Stadler et al., 2019).

Based on an integrative perspective, the combination of (a) the absence of an interaction between learning ability and complexity and (b) the trajectory of strategy use shown in Fig. 4, seems to suggest that learning in MicroDYN might be predominantly a matter of learning how to interact with the systems in terms of acquiring a functional strategy that allows the effective (and efficient, given the time constraints) identification of the causal structure of the respective systems. This learning tends to take place in the early stages in working through the MicroDYN test. As can be gleaned from analyses discussed earlier, this is of particular importance for systems that feature autonomous changes.

4.4. Learning as a dynamic process in complex problem solving

The central finding is that MicroDYN performance tends to be associated with learning test scores. This is reassuring and promising. As it was one of CPS' initial promises, MicroDYN has also the potential to reflect the *dynamics* of cognitive functioning and to go beyond capturing cognitive abilities, albeit in a psychometrically fine-tuned and controlled, yet static fashion. Learning is on the one hand part of our conceptual understanding of intelligence. On the other hand, learning – in its manifestation as intra-individual change – poses a psychometric threat (most notably in terms of a traditional notion of reliability). Unsurprisingly, this tension has attracted attention in intelligence research. For instance, Verguts and de Boeck (2002) identified processes of learning across items in Raven's Progressive Matrices through the tendency of test takers to employ rules identified in preceding items. Emphasising the role of item order in Raven's Progressive Matrices, Ren, Wang, Altmeyer, and Schweizer (2014) modelled learning processes as being distinct from performance (i.e., reasoning) processes. (Birney, Beckmann, Beckmann, & Double, 2017) provide an example for how individual differences in learning trajectories can be separated from conventionally operationalised performance indicators in a constrained, standardised test such as Raven's Progressive Matrices. They distinguish between psychometric complexity (ψ_C) and processes of psychometric learning (ψ_L). Whilst psychometric complexity (ψ_C) is conceptualised as a statistical moderation of the cognitive demand (i.e., complexity) of items on performance trajectories, processes of psychometric learning (ψ_L) is conceptualised as a statistical moderation of accumulated experience across items on performance trajectories. The perspective on learning as accumulation of experience also informed the study of the potential influence of so-called non-cognitive factors in complex decision making tasks (Birney, Beckmann, Beckmann, Double, & Whittingham, 2018).

4.5. Future directions and limitations

As is the case with any pre-registered exploration, our analysis has limitations. For instance, we adhered to our planned analysis rather strictly and only pursued "unforeseen" analytic steps to follow up on certain findings in one case. Due to the albeit planned, but explorative nature of this study, our attempts to project the findings and their interpretations into the bigger picture of CPS research therefore must remain tentative and even speculative to a certain degree.

MicroDYN's potential to evoke learning processes is in fact twofold. It includes learning within systems in form of knowledge acquisition processes for each individual item (i.e., system) in a MicroDYN test, and it includes learning across systems in form of acquiring the competency (e.g., employment of a suitable exploration strategy) to deal effectively with increasing levels of complexity. The study presented here is limited to the analysis of the effects of the latter process. As there is now tentative evidence for learning processes both within and across systems, future studies might want to explore whether these two perspectives represent distinct learning processes. One question of interest would be whether the learning processes studied *within* systems as in the context of the classic DYNAMIS approach (Funke, 1992, 2001), in which one system is to be explored and controlled over an extended number of trials, is qualitatively equivalent to learning processes observed *across* systems in a MicroDYN test. One way to address such questions would require an extension of the time available for system exploration and system control in a MicroDYN context to be able to study learning processes within systems. This, however, tends to contravene the rationale that has motivated the development of MicroDYN in the first place. An alternative approach could include to investigate associations between learning ability and performance in a DYNAMIS-type task and MicroDYN tests in the same study.

Contrasting performance trajectories across systems that are difficulty-ordered vs. complexity-ordered (Beckmann & Goode, 2017) are expected to also help to better understand learning in MicroDYN. It would also be interesting to see whether a complexity-ordered item pool will benefit the psychometric properties of a MicroDYN task as a properly structured (albeit speeded) power test that has the potential to capture individual differences in learning.

A better utilisation of MicroDYN's potential as a tool for measuring learning – be it in terms of a research instrument or an assessment device – requires an operationalisation of knowledge acquisition performance that goes beyond a pragmatically simple dichotomous scoring rubric. The sensitivity index, as used in our analyses, represents a conceptually informed approach to not only differentiate between problems solvers, but also to trace within-person trajectories of knowledge acquisition. Especially the latter is essential to capturing learning as intra-individual change processes. MicroDYN further offers the potential to investigate the specific effects of system characteristics more closely by systematically varying them across systems. For instance, in order to better understand the potentially distinct effects of complexity and system position, or more precisely, person-related characteristics that go along with system position, such as experience or fatigue, future studies might present systems with varying complexities in various orders instead of one fixed order.

As our analyses provide corroboratory evidence for the importance of autonomous changes in CPS tests, MicroDYN's potential to evoke learning processes is likely limited if systems are used that do not contain autonomous changes. The result pattern obtained in our analyses suggests that performance differs not just quantitively, but also qualitatively between systems with and without autonomous changes. This has implications for the aggregation of performance scores across systems in a MicroDYN test. If at all, systems without autonomous changes might serve as some form of "distractor tasks" within the item pool of a CPS test; performance scores reflective of learning should,

[4] As a reminder, with regard to system position, every second system in the MicroDYN test had autonomous changes. With regard to complexity, systems with autonomous changes (which require the employment of a systematic exploration strategy comprising a zero intervention) have an average complexity index of 0.50, whilst the average complexity index for systems without autonomous changes is 0.67. System position and complexity in this MicroDYN test are rather loosely aligned ($r = 0.48$).

however, be based solely on systems that feature autonomous changes.

Our analyses and subsequent insights are also limited by a slight mismatch between the sample used in the study and the population for which the learning test was developed. The learning test was developed, standardised, and normed for a population aged 10 to 15 (for further details see Beckmann, 2001; Beckmann & Guthke, 1999). Although the learning test has been used exploratorily in samples of older test takers in the past, the estimates of learning ability that current analyses are based on might be limited in terms of sufficiently differentiating among highly functioning learners. An additional potentially limiting aspect to consider is the fact that the distribution of learning ability in the current, rather homogeneous sample of university students is expected to be naturally restricted. University students are likely to show above-average cognitive abilities and problem-solving skills (e.g., exploration strategies), which might limit generalizability. The result pattern obtained in our analyses therefore should be interpreted as a conservative, in other words lower bound estimate, of the strength of the association between learning test and MicroDYN performance. Future research interested in the elicitation of further insights into learning in MicroDYN might consider employing assessment tools validated for the measurement of learning ability for cognitively high functioning adults.

In addition, it should be noted that data are based on unsupervised online assessments. Previous studies on the comparability of online and offline assessments of cognitive abilities have provided evidence that online assessments provide data that is reliable, valid, and comparable to supervised offline assessments (for CPS, see, e.g., Schult, Stadler, Becker, Greiff, & Sparfeldt, 2017; for cognitive abilities in general, see, e.g., Mead & Drasgow, 1993; Steger, Schroeders, & Gnambs, 2020). However, due to their self-selection participants in online studies tend to be more dedicated (Wilhelm & McKnight, 2002). Future studies might want to aim at replicating the present study with more heterogeneous samples and ideally with different test settings.

To further our understanding of learning processes in the context of CPS tests, future studies may also want to explore the generalisability of current findings across different CPS paradigms. Initially, the link between learning and CPS performance has been studied using the classic DYNAMIS approach (e.g., Beckmann, 1994; Beckmann & Guthke, 1995; Funke, 1992). Learning processes have also been studied in the context of so-called microworlds (e.g., Süß & Kretzschmar, 2018; Wood, Beckmann, & Birney, 2009). Findings obtained in these contexts have partially informed the research reported here using the MicroDYN approach. Although the minimal/multiple complex system (MCS) approach (Funke & Greiff, 2017; Greiff et al., 2015) and in particular MicroDYN (Greiff et al., 2012) are currently one of the most prominent measurement approaches in the field of CPS (see, for example, the integration in international largescale assessments such as the Programme for International Student Assessment, PISA; OECD, 2013), it is not without its drawbacks (for a critical discussion, see, e.g. Süß & Kretzschmar, 2018). For example, Funke (2014) argued that the MCS approach does not capture the same complex cognitions compared to other, more complex CPS tests such as the Tailorshop (Süß, Kersting, & Oberauer, 1993; Süß, Oberauer, & Kersting, 1993). Moreover, some interpret CPS measures as a computerised fluid intelligence tasks (e.g., Kretzschmar et al., 2017; Süß, 1996), whereas others emphasize that the moderately sized associations between CPS measures and intelligence tests might suggest that different cognitive abilities are captured (e.g., Stadler et al., 2015). In summary, the construct validity of CPS measures seems still an open question (for an overview, see Kretzschmar et al., 2016) and, therefore, it would be interesting to investigate the role of learning in more complex operationalisations of CPS in future studies.

Lastly, future research needs to go beyond correlation-based descriptions of associations between outcome measures and should aim for employing research designs that (a) better reflect the *processes* that underpin problem solving and learning, and (b) warrant (directed) *causal* interpretations in the interplay between task characteristics or task demands and processes of learning to solve complex, dynamic problems. Such research needs to ideally build on (cognitive) process theories. Correlation-focussed research tends to produce evidence of predictive utility, which is of importance in various applied contexts. For conceptual progress in CPS research, however, it is important to acknowledge that predictive utility is not to be confused with (construct) validity.

4.6. Conclusion

Our analyses have revealed a number of promising findings and have prompted further questions. We were able to replicate and refine previous findings on the role of learning in CPS in the context of knowledge acquisition in MicroDYN. The presence of autonomous changes played a central role in our analyses, as (1) the relationship between learning and knowledge acquisition performance was more pronounced in systems with autonomous changes, and (2) the systematic exploration of the system was more strongly associated with knowledge acquisition performance in systems with autonomous changes. Of course, as is the case with any research, our study is limited in its conclusiveness, but taken together, our findings suggest that a MicroDYN approach might have the potential to capture processes of acquiring a functional strategy to identify the causal structure of a system (i.e., learning). Our findings suggest also that this potential has not been exhaustively utilised. It is our hope that our findings motivate and orientate future research into the relationship between CPS performance and learning or beyond.

Declaration of Competing Interest

None.

Data availability

Data and code can be found online: https://doi.org/10.17605/OSF.IO/E5KZJ.

Acknowledgements

This research project was supported by the Tübingen Postdoctoral Academy for Research on Education (PACE) at the Hector Research Institute of Education Sciences and Psychology, Tübingen; PACE is funded by the Baden-Württemberg Ministry of Science, Research, and the Arts. The preparation of this manuscript was also supported by a grant from the Research Talent Development Fund of the University of Zurich.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.17605/OSF.IO/E5KZJ.

References

Beckmann, J. F. (1994). *Lernen und komplexes Problemlösen: Ein Beitrag zur Konstruktvalidierung von Lerntests*. Holos.

Beckmann, J. F. (2001). *Zur Validierung des Konstrukts des intellektuellen Veränderungspotentials: Validation of the construct intellectual change potential. logos*.

Beckmann, J. F. (2006). Superiority: Always and everywhere? On some misconceptions in the validation of dynamic testing. *Educational and Child Psychology, 23*(3), 35–49.

Beckmann, J. F. (2014). The umbrella that is too wide and yet too small: Why dynamic testing has still not delivered on the promise that was never made. *Journal of Cognitive Education and Psychology, 13*(3), 308–323. https://doi.org/10.1891/1945-8959.13.3.308

Beckmann, J. F. (2019). *Heigh-ho: Cps and the seven questions – Some thoughts on contemporary complex problem solving research*. Advance online publication. https://doi.org/10.11588/JDDM.2019.1.69301 (Journal of Dynamic Decision Making, Vol 5 (2019)).

Beckmann, J. F., Birney, D. P., & Goode, N. (2017). Beyond psychometrics: The difference between difficult problem solving and complex problem solving. *Frontiers in Psychology, 8*, 1–13. https://doi.org/10.3389/fpsyg.2017.01739. Article 1739.

Beckmann, J. F., & Goode, N. (2014). The benefit of being naïve and knowing it: The unfavourable impact of perceived context familiarity on learning in complex

problem solving tasks. *Instructional Science, 42*(2), 271–290. https://doi.org/10.1007/s11251-013-9280-7
Beckmann, J. F., & Goode, N. (2017). Missing the wood for the wrong trees: On the difficulty of defining the complexity of complex problem solving scenarios. *Journal of Intelligence, 5*(2). https://doi.org/10.3390/jintelligence5020015
Beckmann, J. F., & Guthke, J. (1995). Complex problem solving, intelligence, and learning ability. In P. A. Frensch, & J. Funke (Eds.), *Complex problem solving - the European perspective*. Lawrence Erlbaum Associates.
Beckmann, J. F., & Guthke, J. (1999). *Psychodiagnostik des schlußfolgernden Denkens: Handbuch zur Adaptiven Computergestützten Intelligenz-Lerntestbatterie für Schlußfolgerndes Denken (ACIL)*. Hogrefe.
Birney, D. P., Beckmann, J. F., Beckmann, N., & Double, K. S. (2017). Beyond the intellect: Complexity and learning trajectories in Raven's progressive matrices depend on self-regulatory processes and conative dispositions. *Intelligence, 61*, 63–77. https://doi.org/10.1016/j.intell.2017.01.005
Birney, D. P., Beckmann, J. F., Beckmann, N., Double, K. S., & Whittingham, K. (2018). Moderators of learning and performance trajectories in microworld simulations: Too soon to give up on intellect!? *Intelligence, 68*, 128–140. https://doi.org/10.1016/j.intell.2018.03.008
Cohen, P., Cohen, J., Aiken, L. S., & West, S. G. (1999). The problem of units and the circumstance for POMP. *Multivariate Behavioral Research, 34*(3), 315–346. https://doi.org/10.1207/S15327906MBR34032
Cumming, G. (2014). The new statistics: Why and how. *Psychological Science, 25*(1), 7–29. https://doi.org/10.1177/0956797613504966
Dörner, D., & Funke, J. (2017). Complex problem solving: What it is and what it is not. *Frontiers in Psychology, 8*. https://doi.org/10.3389/fpsyg.2017.01153. Article 1153.
Elliott, J. G., Resing, W. C. M., & Beckmann, J. F. (2018). Dynamic assessment: A case of unfulfilled potential? *Educational Review, 70*(1), 7–17. https://doi.org/10.1080/00131911.2018.1396806
Fischer, A., Greiff, S., & Funke, J. (2012). The process of solving complex problems. *The Journal of Problem Solving, 4*(1). https://doi.org/10.7771/1932-6246.1118
Freitag, C. (1993). *Validierung des Kurzzeitlerntests ADAFI anhand eines Vergleichs mit Problemlöseverhalten in computersimulierten dynamischen Systemen [Diplomarbeit (unveröff.), Universität Bonn]*.
Frensch, P. A., & Funke, J. (Eds.). (1995a). *Complex problem solving - the European perspective*. Lawrence Erlbaum Associates.
Frensch, P. A., & Funke, J. (1995b). Definitions, traditions, and a general framework for understanding complex problem solving. In P. A. Frensch, & J. Funke (Eds.), *Complex problem solving - the European perspective*. Lawrence Erlbaum Associates.
Funke, J. (1985). Steuerung dynamischer Systeme durch Aufbau und Anwendung subjektiver Kausalmodelle [control of dynamic systems by building up and using subjective causal models]. *Zeitschrift für Psychologie, 193*, 435–457.
Funke, J. (1992). Dealing with dynamic systems: Research strategy, diagnostic approach and experimental results. *German Journal of Psychology, 16*(1), 24–43.
Funke, J. (2001). Dynamic systems as tools for analysing human judgement. *Thinking & Reasoning, 7*(1), 69–89. https://doi.org/10.1080/13546780042000046
Funke, J. (2014). Analysis of minimal complex systems and complex problem solving require different forms of causal cognition. *Frontiers in Psychology, 5*, 739. https://doi.org/10.3389/fpsyg.2014.00739
Funke, J., & Greiff, S. (2017). Dynamic problem solving: Multiple-item testing based on minimally complex systems. In D. Leutner, J. Fleischer, J. Grünkorn, & E. Klieme (Eds.), *Competence assessment in education: An introduction* (pp. 427–443). https://doi.org/10.1007/978-3-319-50030-0_25
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*(3), 345–352. https://doi.org/10.1016/j.intell.2010.01.001
Goode, N., & Beckmann, J. F. (2016). With a little help …: On the role of guidance in the acquisition and utilisation of knowledge in the control of complex, dynamic systems. *Journal of Dynamic Decision Making, 2*(5), 1. https://doi.org/10.11588/jddm.2016.1.33346
Greiff, S., Fischer, A., Stadler, M., & Wüstenberg, S. (2015). Assessing complex problem-solving skills with multiple complex systems. *Thinking & Reasoning, 21*(3), 356–382. https://doi.org/10.1080/13546783.2014.989263
Greiff, S., Fischer, A., Wüstenberg, S., Sonnleitner, P., Brunner, M., & Martin, R. (2013). A multitrait–multimethod study of assessment instruments for complex problem solving. *Intelligence, 41*(5), 579–596. https://doi.org/10.1016/j.intell.2013.07.012
Greiff, S., Krkovic, K., & Nagy, G. (2014). The systematic variation of task characteristics facilitates the understanding of task difficulty: A cognitive diagnostic modeling approach to complex problem solving. *Psychological Test and Assessment Modeling, 56*(1), 83–103. https://pure.ipn.uni-kiel.de/portal/files/506377/05_Greif.pdf.
Greiff, S., & Neubert, J. C. (2014). On the relation of complex problem solving, personality, fluid intelligence, and academic achievement. *Learning and Individual Differences, 36*, 37–48. https://doi.org/10.1016/j.lindif.2014.08.003
Greiff, S., Wüstenberg, S., & Funke, J. (2012). Dynamic problem solving: A new assessment perspective. *Applied Psychological Measurement, 36*(3), 189–213. https://doi.org/10.1177/0146621612439620
Guthke, J. (1982). The learning test concept—An alternative to the traditional static intelligence test. *German Journal of Psychology, 6*(4), 306–324.
Guthke, J., & Beckmann, J. F. (2000). The learning test concept and its application in practice. In C. S. Lidz, & J. G. Elliott (Eds.), *Advances in cognition and educational practice: Vol. 6. Dynamic assessment: Prevailing models and applications* (1st ed., pp. 17–69). JAI Press.
Guthke, J., & Beckmann, J. F. (2003). Dynamic Assessment With Diagnostic ProgramsDynamic assessment with diagnostic programs. In R. J. Sternberg, J. Lautrey, & T. I. Lubart (Eds.), *Models of intelligence: International perspectives*. American Psychological Association.
Guthke, J., Beckmann, J. F., & Stein, H. (1995). Recent research evidence on the validity of learning tests. In J. S. Carlson (Ed.), *Vol. 3. Advances in cognition and educational practice* (pp. 117–143). JAI Press.
Guthke, J., Beckmann, J. F., & Wiedl, K. H. (2003). Dynamics in dynamic testing. *Psychologische Rundschau, 54*(4), 225–232. https://doi.org/10.1026//0033-3042.54.4.225
Guthke, J., & Wiedl, K. H. (1996). *Dynamisches Testen: Zur Psychodiagnostik der intraindividuellen Variabilität*. Hogrefe.
Hannon, B., & Daneman, M. (2014). Revisiting the construct of "relational integration" and its role in accounting for general intelligence: The importance of knowledge integration. *Intelligence, 47*, 175–187. https://doi.org/10.1016/j.intell.2014.09.010
Herde, C. N., Wüstenberg, S., & Greiff, S. (2016). Assessment of complex problem solving: What we know and what we don't know. *Applied Measurement in Education, 29*(4), 265–277. https://doi.org/10.1080/08957347.2016.1209208
Hox, J. J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed.). Routledge/Taylor & Francis Group. https://doi.org/10.4324/9780203852279
Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking: From childhood to adolescence*. Routledge & Kegan Paul.
de Jong, T., & van Joolingen, W. R. (1998). Scientific discovery learning with computer simulations of conceptual domains. *Review of Educational Research, 68*(2), 179–201. https://doi.org/10.3102/00346543068002179
Kluge, A. (2008). Performance assessments with microworlds and their difficulty. *Applied Psychological Measurement, 32*(2), 156–180. https://doi.org/10.1177/0146621607300015
Kretzschmar, A., Hacatrjana, L., & Rascevska, M. (2017). Re-evaluating the psychometric properties of MicroFIN: A multidimensional measurement of complex problem solving or a unidimensional reasoning test? *Psychological Test and Assessment Modeling, 59*(2), 157–182. https://doi.org/10.5167/UZH-185323
Kretzschmar, A., & Nebe, S. (2021). Working memory, fluid reasoning, and complex problem solving: Different results explained by the Brunswik symmetry. *Journal of Intelligence, 9*(1). https://doi.org/10.3390/jintelligence9010005
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. https://doi.org/10.1016/j.intell.2015.11.004
Kröner, S., Plass, J. L., & Leutner, D. (2005). Intelligence assessment with computer simulations. *Intelligence, 33*(4), 347–368. https://doi.org/10.1016/j.intell.2005.03.002
Advances in cognition and educational practice. In Lidz, C. S., & Elliott, J. G. (Eds.) (1. ed.), *Vol. 6. Dynamic assessment: Prevailing models and applications*, (2000). JAI Press.
Lüdecke, D. (2018). Sjstats: Statistical functions for regression models. R package version 0.18.1. https://CRAN.R-project.org/package=sjstats.
Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin, 114*(3), 449–458. https://doi.org/10.1037/0033-2909.114.3.449
Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution, 4*(2), 133–142. https://doi.org/10.1111/j.2041-210x.2012.00261.x
Neubert, J. C., Kretzschmar, A., Wüstenberg, S., & Greiff, S. (2015). Extending the assessment of complex problem solving to finite state automata. *European Journal of Psychological Assessment, 31*(3), 181–194. https://doi.org/10.1027/1015-5759/a000224
Oberauer, K. K., Süß, H.-M., Wilhelm, O. O., & Wittmann, W. W. (2008). Which working memory functions predict intelligence? *Intelligence, 36*(6), 641–652. https://doi.org/10.1016/j.intell.2008.01.007
OECD. (2013). *PISA 2012 assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy*. Paris, France: OECD Publishing.
Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology, 25*, 111–163. https://doi.org/10.2307/271063
Ren, X., Wang, T., Altmeyer, M., & Schweizer, K. (2014). A learning-based account of fluid intelligence from the perspective of the position effect. *Learning and Individual Differences, 31*, 30–35. https://doi.org/10.1016/j.lindif.2014.01.002
Rudolph, J., Greiff, S., Strobel, A., & Preckel, F. (2018). Understanding the link between need for cognition and complex problem solving. *Contemporary Educational Psychology, 55*, 53–62. https://doi.org/10.1016/j.cedpsych.2018.08.001
Schult, J., Stadler, M., Becker, N., Greiff, S., & Sparfeldt, J. R. (2017). Home alone: Complex problem solving performance benefits from individual online assessment. *Computers in Human Behavior, 68*, 513–519. https://doi.org/10.1016/j.chb.2016.11.054
Schunk, D. H. (2020). *Learning theories: An educational perspective* (8th ed.). Pearson.
Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*(2). https://doi.org/10.1214/aos/1176344136
Schweizer, F., Wüstenberg, S., & Greiff, S. (2013). Validity of the MicroDYN approach: Complex problem solving predicts school grades beyond working memory capacity. *Learning and Individual Differences, 24*, 42–52. https://doi.org/10.1016/j.lindif.2012.12.011
Shayer, M. (2008). Intelligence for education: As described by Piaget and measured by psychometrics. *The British Journal of Educational Psychology, 78*(Pt 1), 1–29. https://doi.org/10.1348/000709907X264907
Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General, 117*(1), 34–50. https://doi.org/10.1037/0096-3445.117.1.34
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*(5), 289–305. https://doi.org/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. https://doi.org/10.1016/j.intell.2015.09.005

Stadler, M., Niepel, C., & Greiff, S. (2019). Differentiating between static and complex problems: A theoretical framework and its empirical validation. *Intelligence, 72*, 1–12. https://doi.org/10.1016/j.intell.2018.11.003

Steger, D., Schroeders, U., & Gnambs, T. (2020). A meta-analysis of test scores in proctored and unproctored ability assessments. *European Journal of Psychological Assessment, 36*(1), 174–184. https://doi.org/10.1027/1015-5759/a000494

Sternberg, R. J., & Grigorenko, E. L. (Eds.). (2002). *Dynamic testing: The nature and measurement of learning potential*. Cambridge University Press.

Süß, H.-M. (1996). *Intelligenz, Wissen und Problemlösen: Kognitive Voraussetzungen für erfolgreiches Handeln bei computersimulierten Problemen [Intelligence, knowledge and problem solving: Cognitive prerequisites for successful behavior in computer-simulated problems]*. Hogrefe.

Süß, H.-M., Kersting, M., & Oberauer, K. (1993). Zur Vorhersage von Steuerungsleistungen an computersimulierten Systemen durch Wissen und Intelligenz [On the predictability of control performance on computersimulated systems by knowledge and intelligence]. *Z. Differ. Diagnostische Psychol., 14*, 189–203.

Süß, H.-M., & Kretzschmar, A. (2018). Impact of cognitive abilities and prior knowledge on complex problem solving performance - empirical results and a plea for ecologically valid microworlds. *Frontiers in Psychology, 9*, 626. https://doi.org/10.3389/fpsyg.2018.00626

Süß, H.-M., Oberauer, K., & Kersting, M. (1993). Intellektuelle Fähigkeiten und die Steuerung komplexer Systeme [Intelligence and control performance on computer-simulated systems]. *Spr. Kognition, 12*, 83–97.

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

Verguts, T., & de Boeck, P. (2002). The induction of solution rules in Raven's progressive matrices test. *European Journal of Cognitive Psychology, 14*(4), 521–547. https://doi.org/10.1080/09541440143000230

Wilhelm, O., & McKnight, P. E. (2002). Ability and achievement testing on the world wide web. In B. Batinic, U.-D. Reips, & M. Bosnjak (Eds.), *Online social sciences* (pp. 151–180). Hogrefe & Huber Publishers.

Wittmann, W. W., & Hattrup, K. (2004). The relationship between performance in dynamic systems and intelligence. *Systems Research and Behavioral Science, 21*(4), 393–409. https://doi.org/10.1002/sres.653

Wood, R. E., Beckmann, J. F., & Birney, D. P. (2009). Simulations, learning and real world capabilities. *Education + Training, 51*(5/6), 491–510. https://doi.org/10.1108/00400910910987273

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

Wüstenberg, S., Greiff, S., Molnár, G., & Funke, J. (2014). Cross-national gender differences in complex problem solving and their determinants. *Learning and Individual Differences, 29*, 18–29. https://doi.org/10.1016/j.lindif.2013.10.006

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*(1–2), 127–146. https://doi.org/10.1007/s10758-014-9222-8

Zimmerman, C. (2007). The development of scientific thinking skills in elementary and middle school. *Developmental Review, 27*(2), 172–223. https://doi.org/10.1016/j.dr.2006.12.001