European Journal of Psychological Assessment

Extending the Assessment of Complex Problem Solving to Finite State Automata: Embracing Heterogeneity

--Manuscript Draft--

Manuscript Number:	EJPA-D-13-00126R1
Full Title:	Extending the Assessment of Complex Problem Solving to Finite State Automata: Embracing Heterogeneity
Short Title:	CPS Assessment with Finite State Automata
Article Type:	Original Article
Keywords:	Complex Problem Solving; Finite State Automata; MicroFIN; MicroDYN; Multitrait-multimethod
Corresponding Author:	Jonas C. Müller University of Luxembourg Luxembourg, LUXEMBOURG
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	University of Luxembourg
Corresponding Author's Secondary Institution:	
First Author:	Jonas C. Müller
First Author Secondary Information:	
Order of Authors:	Jonas C. Müller
	André Kretzschmar
	Sascha Wüstenberg
	Samuel Greiff
Order of Authors Secondary Information:	
Abstract:	Recent advancements in the assessment of Complex Problem Solving (CPS) build on the use of homogenous tasks that enable the reliable estimation of CPS skills. The range of problems featured in established instruments such as MicroDYN is consequently limited to a specific subset of homogeneous complex problems. This restriction is problematic when looking at domain-specific examples of complex problems, which feature characteristics absent from current assessment instruments (e.g., threshold states). We propose to utilize the formal framework of Finite State Automata (FSA) to extend the range of problems included in CPS assessment. An approach based on FSA, called MicroFIN, is presented, translated into specific tasks, and empirically investigated. We conducted an empirical study (N = 576), (1) inspecting the psychometric features of MicroFIN, (2) relating it to MicroDYN, and (3) investigating the relations to a measure of reasoning (i.e., CogAT). MicroFIN (1) exhibited adequate measurement characteristics and multitrait-multimethod models indicated (2) the convergence of latent dimensions measured with MicroDYN. Relations to reasoning (3) were moderate and comparable to the ones previously found for MicroDYN. Empirical results and corresponding explanations are discussed. More importantly, MicroFIN highlights the feasibility of expanding CPS assessment to a larger spectrum of complex problems.
Additional Information:	
Question	**Response**
Has the manuscript or any component of it already been published or is currently	No

Extending the Assessment of Complex Problem Solving to Finite State Automata: Embracing Heterogeneity

Due to their relevancy for dealing with the challenges of our times, an individuals' skills in coping with complex problems are generally considered as part of the so called 21st century skills (cf. Griffin, McGaw, & Care, 2012). Scientifically, these skills in coping with complex problems are investigated under the label of Complex Problem Solving (CPS; e.g., Buchner, 1995; Sternberg & Frensch, 1991), sometimes also called Dynamic Problem Solving (e.g., Greiff, Wüstenberg, & Funke, 2012), Dynamic Decision Making (e.g., Brehmer, 1992), or Interactive Problem Solving (e.g., OECD, 2013; see Greiff et al., 2013, for a discussion of different names). In the following, we will use the term CPS to refer to the construct. Recently, CPS has been added to a range of high-profile studies as representatives of domain-general and transversal skills, for example to the Programme for International Student Assessment in its 2012 cycle (PISA; OECD, 2013), the arguably most important large-scale assessment world wide.

One of the defining characteristics of the complex problems employed within CPS assessment in large-scale efforts and in general are their changes in reaction to user interaction and/or passage of time. Additionally, they demand active interventions and feature a multitude of interrelated factors (see Buchner, 1995). Dealing successfully with such complex problems has been shown to be related to, but separable from other (cognitive) constructs, such as reasoning ability and working memory capacity, both, conceptually and empirically (e.g., Schweizer, Wüstenberg, & Greiff, 2013; Sonnleitner, Keller, Martin, & Brunner, 2013; Wüstenberg, Greiff, & Funke, 2012; see also Bühner, Kröner, & Ziegler, 2008; Wittmann & Süß, 1999).

On the side of the problem solver, the dynamic, interactive, and complex problems result in characteristic requirements with regard to processes of knowledge acquisition and knowledge application (for details see Fischer, Greiff, & Funke, 2012; Osman, 2010). An individual’s skills to deal with such problems are called CPS skills and within this study we aim at broadening the view on their assessment.

Assessment of Complex Problem Solving

Looking for a viable assessment of CPS, essential elements are the utilization of formal frameworks and the application of several multiple complex systems within these formal frameworks (cf. Greiff et al., 2012). Formal frameworks allow for the analysis of the underlying structure of problems instead of surface features and hence, the systematic examination and comparison of assessment instruments (Funke, 2001; Greiff et al., 2012). Multiple complex systems built on a formal framework and employed within one assessment instrument are a viable and straightforward way to meet the psychometric requirement of stochastically independent indicators to estimate the target constructs (cf. Greiff et al., 2012; Wüstenberg et al., 2012).

Currently, CPS assessment instruments combining formal frameworks and multiple complex systems are mainly based on one specific formal framework: Linear Structural Equations (LSE)[1]. In LSE, problems are formally described as a set of linear equations of quantitative variables. That is, the problems are formalized as several input and output variables that are connected by linear relations between each other (e.g., the amount of different fertilizers changing the growth of flowers). Within the assessment, participants have to explore several of these relations by manipulating the input variables, observing the resulting changes in the output variables and, from this, deriving the causal relations between input and output variables (knowledge acquisition; cf. L. R. Novick & Bassok, 2005). Subsequently, participants have to use their knowledge to reach target values in the output variables (knowledge application).

[1] Please note that researchers focusing on basic human processes and a modeling of complex problems following real-world examples are successfully utilizing a much broader range of complex problems (see for example the work by Coty Gonzalez and colleagues; Gonzalez, Vanyukov, & Martin, 2005). For the sake of brevity, we focus on instruments directed towards skill assessment here.

Historical instances of instruments targeting CPS with the help of LSE featured a singular complex problem (e.g., MultiFlux; Bühner et al., 2008) and consequently suffered from psychometric problems due to one-item-testing (Greiff et al., 2012). To overcome this problem, CPS assessment nowadays employs multiple problems in succession and is consequently able to reliably estimate a persons CPS skills. Examples of instruments combining LSE as formal framework with multiple complex systems are *Genetics Lab* (Sonnleitner et al., 2012), *MicroDYN* (e.g., Greiff et al., 2012), and the assessment of CPS in PISA 2012, that is partially based on MicroDYN (OECD, 2013).

The utilization of the framework of LSE in combination with multiple complex systems has resulted in reliable instruments targeting core features of CPS with a focus on psychometrics requirements (e.g., Greiff & Wüstenberg, 2014). But when comparing the range of problems employed in assessment with the original breadth of the construct, instruments relying on LSE and, thus, on quantitative relations between a set of variables, are necessarily restricted as they have to follow a predefined, rather narrow pattern of relations between elements (cf. Funke, 2010). Consequently, the problems presented to participants are rather homogeneous with regard to the kinds of relations that have to be explored.

Complex problems found in specific domains, on the other hand, for example within the domain of chemistry, include features that are absent from current CPS assessment based on LSE. Examples are relations between input and output variables, which feature strong interactions between input variables as found in Le Chatelier's principle of chemical equilibria (e.g., Scherer & Tiemann, 2012) and qualitative changes in the problem after reaching a threshold (e.g., water freezing to ice). These and other features not present in current LSE-based CPS assessment are also part of historical notions of CPS (e.g., Dörner, Kreuzig, Reither, & Stäudel, 1983; Funke, 2010) and real-world examples of complex problems (e.g., Ackerman, 1992).

As a consequence, LSE-based assessment instruments are currently unable to assess differences and commonalities with regard to a range of problem characteristics and corresponding requirements on the side of the problem solver. To give an example: Vary-one-thing-at-a-time (VOTAT; Tschirgi, 1980) can be considered an sufficient and optimal strategy in the LSE-based tasks mentioned above (Vollmeyer, Burns, & Holyoak, 1996; see Rollett, 2008, for a comprehensive discussion of strategies in LSE-based tasks), but not in the case of problems featuring threshold or equilibrium states (cf. McElhaney & Linn, 2011). Hence, we have to draw conclusions on general exploration behavior in complex problem situations based on one specific strategy (i.e., VOTAT) when using LSE-based tasks, whereas broader sets of strategies and an adaptive use of them are necessary to deal with complex problems in general (see Levy & Wilensky, 2011, for an example of adaptive exploration strategies).

To counter the limitation in the breadth of problems included in instruments and facilitate a greater heterogeneity in tasks, we propose to expand CPS assessment to include tasks based on another formal framework introduced by Buchner and Funke (1993), Finite State Automata (FSA).

The Formal Framework of Finite State Automata

In FSA, problems are formalized as predefined states with transitions between these states triggered by events such as user interactions or passing of time. In contrast to LSE, the type of relation formalized as transition does not have to follow a specific pattern (i.e., quantitative relations). Consequently, FSA-based CPS assessment can easily include a variety of features that overcome the current homogeneity of LSE-based tasks, for example by including complex problems that require different exploration strategies. That is, the use of FSA allows for the inclusion of heterogeneous causal relations in CPS assessment, thereby leading to a broader assessment of CPS skills.

At the same time, FSA-based CPS assessment can preserve advantages of established instruments based on LSE such as multiple tasks comparable by their causal structure and a general layout resembling that of established instruments (e.g., relating input variables to output variables). Additionally, FSA-based tasks were already utilized in laboratory studies targeting CPS, providing elaborations on the FSA framework itself and ways of mathematical and abstract representation (e.g., Buchner & Funke, 1993; Funke, 2001). Outside the realm of CPS research, FSAs are commonly used to formalize the workings of a broad range of appliances, for example for the purpose of programming vending machines, combination locks, and turnstiles (e.g., Anderson, 2006; Rich, 2008).

Empirical applications combining FSA and multiple complex systems, have not surfaced, yet (for a notable attempt, see the PISA framework for Problem Solving in 2012; OECD, 2013). To empirically investigate the utility of the combination of FSA with the approach of multiple complex systems, we therefore developed a set of tasks that represent different options exceeding the limitations of LSE-based instruments. The resulting approach is called MicroFIN, a label coined by Greiff and Funke (2009).

Like other assessment vehicles and along the theoretical understanding of CPS (e.g., MultiFlux, Bühner et al., 2008; MicroDYN, Greiff et al., 2012; Genetics Lab, Sonnleitner et al., 2012), MicroFIN tasks are structured into two separate phases: Tasks begin with the (1) knowledge acquisition phase, where participants can freely explore the simulation, followed by an assessment of the acquired knowledge. In the following (2) knowledge application phase, participants are then asked to reach given target states, assessing their capabilities in applying the acquired knowledge.

An illustration of a MicroFIN task, called ‘Fish-o-maton’, is depicted in Figure 1, showing the layout of the task as presented to participants. The corresponding state-transition diagram, an abstract formal representation of the underlying finite state automaton can be found in Figure A1 the Appendix.

The ‘Fish-o-maton’ includes three ordinal input variables, each with four input values and an output variable visualized as an aquarium (nominal variable with ordinal elements).

The output variable has five possible output values: Empty, soiled, few fish, moderate number of fish, and many fish. Whereas the general structure of the task is similar to LSE-based items, with input variables related to output variables, the specific relation between input and output variables is unfeasible to implement within LSE: Transitions are triggered when all input variables are moved into equivalent states and out of them. To give an example, if a participant brings all input variables to a low amount of input, few fish are displayed in the aquarium. If she moves the first input variable out of this equilibrium by increasing its value, the output variable changes to a display of a soiled aquarium, that is, a qualitative change in the output variable not feasible within LSE-based instruments.

Consequently, an exploration behavior following the application of VOTAT, which would be successful in the case of LSE-based tasks, that is, varying the inputs in isolation from zero to three, would lead to the omission of the central states of the 'Fish-o-maton'. Hence, the task is an example of expanding the range of necessary exploration behavior, while maintaining general features of established CPS assessment instruments.

Table 1 gives an overview of the features of all five utilized MicroFIN tasks. As explicated above for the 'Fish-o-maton', the table includes information on the specific kind of relations between input and output variables realized in each task, differentiating the tasks from each other, as well as from established instruments based on LSE such as MicroDYN. Furthermore, a detailed description of all utilized MicroFIN tasks is given in the Appendix.

Focus of the empirical study

The inquiries of the empirical study are directed towards the exploration of the empirical features of MicroFIN. Specifically, attention is directed towards three main directions.

(1) Securing psychometric properties and a measurement model for MicroFIN

A solid psychometric foundation is essential for assessment. Therefore, proficient levels of reliability and adequate item difficulty are expected when assessing the dimensions of CPS (Hypothesis 1a). With regard to these dimensions of CPS, a well-proven differentiation in CPS research is anticipated (e.g., Funke, 2001) with separable empirical dimensions for (1) knowledge acquisition and (2) knowledge application (Hypothesis 1b).

(2) Relating MicroFIN to an established instrument of CPS assessment

The latent dimensions of CPS as assessed via MicroFIN are expected to capture the same underlying constructs as established instruments based on LSE, pointing to convergent validity (i.e. knowledge acquisition and knowledge application). Therefore, a latent multitrait-multimethod model is expected to hold with instruments based on FSA and LSE loading on the same latent variables. Again, separate dimensions for knowledge acquisition and knowledge application as indicated by both instruments are expected. Additionally, method factors for both dimensions in MicroFIN are anticipated, emphasizing the specific requirements resulting from the broader range of problems implemented in MicroFIN (Hypothesis 2).

(3) Exploring the relations to reasoning

An important concept to check the instrument against in terms of discriminant validity is reasoning ability (cf. Sonnleitner et al., 2013; Wüstenberg et al., 2012). Extending the findings of Wüstenberg et al. (2012) for an LSE-based instrument, a positive relation but also separability is expected between reasoning and the dimensions of CPS as measured by the multitrait-multimethod model introduced in Hypothesis 2 (Hypothesis 3).

Materials and Methods

To empirically test these hypotheses, we conducted a study in the educational context, relating MicroFIN tasks to MicroDYN, an established instrument targeting CPS and based on LSE (Greiff et al., 2012). To inquire the relations to reasoning we employed the nonverbal scale of the Cognitive Abilities Test (CogAT; German adaptation by Heller & Perleth, 2000). All instruments were fully computer-based.

Participants

576 German high school students (262 males) attending grades 8 to 12 participated in the study (age between 13 and 18, *M* = 14.95, *SD* = 1.30). Participants attended one of three school tracks within the same school, together covering all educational tracks in German high schools. Participation was voluntary and we received informed consent from parents. With regard to incentives, participants received monetary compensation on a per-class basis that was given to the class inventory. Data of two participants had to be excluded from analyses due to technical problems during assessment.

Measures

Complex problem solving

MicroFIN. The features of the set of five MicroFIN tasks employed in this study, as well as a detailed description of the task 'Fish-o-maton' has already been given (see also Table 1 and the Appendix)[2]. Instructions on MicroFIN included an additional trial task that was excluded from analysis. Generally, all MicroFIN tasks feature separate phases for the exploration of the problem's relations and the assessment of (1) knowledge acquisition and (2) knowledge application, each dimension targeted with the help of specific items. Overall, each MicroFIN task takes approximately 5 minutes to complete.

More specifically, after participants freely explore the task, items in the knowledge acquisition phase target the gathered knowledge. Within each item, an outcome state and a transition is given, and the initial state of the task needs to be selected by the participant (constructed response items, cf. Buchner, 1995). In the example of the 'Fish-o-maton', such items would present a specific manipulation of an input variable (e.g., raising the level of the first input variable from two to three) and the resulting state of the task (e.g., a soiled aquarium, with the input of the remaining two input variables being two). Participants are

[2] The MicroFIN tasks utilized in the study can be obtained from the first author upon request for academic purposes.

then asked to construct the initial state of the ‘Fish-o-maton’ prior to the manipulation from given elements (e.g., all inputs being on the level of two and an aquarium displaying a medium amount of fish). The item type was successfully introduced to CPS research in laboratory applications of FSA-based tasks, making them a natural choice to assess participants’ knowledge (e.g., Buchner & Funke, 1993). Two constructed response items are included per task.

In the knowledge application phase, items are featuring the task in a specific state and participants are asked to manipulate the input variables (i.e., trigger transitions) to reach a goal state that is presented visually and verbally at the beginning of each item (e.g., a specific amount of fish). Two of these items are included per task.

Scoring of knowledge acquisition: Participants receive credit for correctly constructing the initial state of an item and no credit if they fail to do so. A sum score over the two items ranging from 0 to 2 per task is utilized as a manifest indicator in latent analyses to reduce the number of estimated parameters.

Scoring of knowledge application: Participants receive credit for reaching the target state. No credit is given, when participants fail to reach the target state. Again, a sum score for both knowledge application items is used as manifest indicator with a range of 0 to 2 for each MicroFIN task.

MicroDYN. The MicroDYN approach assesses CPS based on the formal framework of LSE and has been covered extensively elsewhere (e.g., Greiff & Funke, 2009; Greiff et al., 2012; Greiff & Wüstenberg, 2014; Schweizer et al., 2013; Wüstenberg et al., 2012). Tasks are defined by up to three quantitative input variables relating to one or several output variables in a linear quantitative way. The connections between these variables have to be discovered and, subsequently, used to reach target values. The distinction between (1) knowledge acquisition and (2) knowledge application implemented in MicroFIN can also be found in MicroDYN.

Scoring of both phases of CPS assessment follows the recommendation of Wüstenberg et al. (2012), with credit given for correct models and no credit for wrong models in knowledge acquisition. In knowledge application, credit is given for reaching the target values and no credit otherwise. There were eight MicroDYN tasks employed in the study, each taking an average of 5 minutes to complete, plus an instructional task that was excluded from analysis. The underlying linear equations of the eight MicroDYN tasks can be found in the Appendix.

Reasoning

For the reasoning assessment, participants completed a computer-adapted version of the nonverbal scale of the Cognitive-Abilities Test (CogAT, Heller & Perleth, 2000). There, participants are asked to identify a figure, completing a 3 x 3 matrix of figures related by combination rules (similar to Ravens Advanced Progressive Matrices, Raven, Raven, & Court, 1998). Credit is given for correct solutions and no credit for wrong answers. Two items (item 9 and 14) of the 25 items included in the CogAT have been shown to be insolvable due to ambiguous solutions (Segerer, Marx, & Marx, 2012) and were excluded from analysis.

Statistical Analysis

Statistical analyses were conducted in MPlus Version 7 (L. K. Muthén & B. O. Muthén, 2012) utilizing descriptive analyses (Hypothesis 1a), confirmatory factor analysis (Hypothesis 1b), and structural equation modeling for the estimation of latent factors and their relations (Hypotheses 2 and 3). The empirical evaluation of constructs and instruments was done within multitrait-multimethod models (cf. Eid, Lischetzke, Nussbeck, & Trierweiler, 2003) allowing for the estimation of method specific effects for instruments targeting the same construct. As our indicators are ordered categorical variables, we used weighted least squares with means and variances adjusted estimation (WLSMV; B. O. Muthén, du Toit, & Spisic, 1997). Global goodness-of-fit was evaluated by χ^2-tests, Root Mean Square Error of Approximation (RMSEA), Comparative Fit Index (CFI), and Tucker Lewis Index (TLI).

According to Hu and Bentler (1999), a χ^2 to *df* ratio < 2 and RMSEA values ≤ .06 indicate a good global fit, as do values ≥ .95 for CFI and TLI. For the comparison of models, we utilized a specific procedure for WLSMV estimation integrated in MPlus to compute χ^2-difference values (L. K. Muthén & B. O. Muthén, 2012, p. 451).

Results

Hypothesis 1a: Psychometric properties of MicroFIN

Descriptive analyses for MicroFIN were the basis for analyzing item difficulties and reliability for both phases of CPS as targeted in Hypothesis 1a. In knowledge acquisition, the average success rate of participants (i.e., item difficulty) ranged from p = .18 to .49 (M = .34, SD = .14) across the five tasks. In knowledge application item difficulty ranged from p = .63 to .77 (M = .70 SD = .06), that is, items targeting knowledge application were easier compared to items targeting knowledge acquisition. Reliability for both dimensions was acceptable with McDonalds's omega (Zinbarg, Revelle, Yovel, & Li, 2005) of ω = .79 and ω = .78 for knowledge acquisition and knowledge application, especially when taking the number of only five MicroFIN tasks into account. The results supported Hypothesis 1a, confirming adequate item difficulty and reliability for MicroFIN.

Hypothesis 1b: A Measurement model for MicroFIN

As laid out under Hypothesis 1b, we expected a 2-dimensional measurement model differentiating between knowledge acquisition and knowledge application for MicroFIN. Results indicated a good model fit for this model (χ^2 (34) = 52.684, p = .021, RMSEA = 0.031, CFI = 0.989, TLI = 0.986) with a latent correlation between dimensions pointing to strongly related, but nonetheless separable dimensions of CPS as measured by MicroFIN (r = .81, p < .001).

An alternative measurement model conflating the two dimensions resulted in significantly worse model fit (χ^2 (35) = 90.118, p < .001, RMSEA = 0.053, CFI = 0.968, TLI

= 0.958, χ^2_Δ (1) = 24.398, p < .001). Therefore, Hypothesis 1b, which assumed a measurement model with separate dimensions for knowledge acquisition and knowledge application for MicroFIN, was supported.

Descriptives and measurement models for the remaining instruments

MicroDYN. For the eight MicroDYN tasks item difficulty for knowledge acquisition was in the range of p = .08 and .68 (M = .42, SD = .26) and p = .05 to .50 (M = .33, SD = .14) for knowledge application. Item difficulties were comparable to the ones reported in other applications of MicroDYN (e.g., Schweizer et al., 2013). Reliability was excellent with ω = .93 (knowledge acquisition) and ω = .88 (knowledge application).

Replicating previous findings (e.g., Schweizer et al., 2013; Wüstenberg et al., 2012), a 2-dimensional model with separate dimensions for knowledge acquisition and knowledge application resulted in the best fitting measurement model for MicroDYN (χ^2 (103) = 225.982, p < .001, RMSEA = 0.047, CFI = 0.979, TLI = 0.976; latent correlation of dimensions r = .82, p < .001).

CogAT. For the CogAT, item difficulty ranged from p = .41 to .81 (M = .67, SD = .12) and reliability can be considered good with ω = .95. Due to the large number of 23 items, we used item-to-construct parceling according to Little, Cunningham, Shahar, and Widaman (2002) to reduce the number of estimated parameters. Three parcels were constructed with comparable average loading of the items on the parcels between M_λ = .64 and .66 and mean item difficulty ranging from M_p = 0.67 to .70. An essentially tau-equivalent measurement model (e.g., M. R. Novick, 1966) assuming equal loadings for parcels, but allowing for differences in intercepts fitted the data well (χ^2 (2) = 3.908, p = .142, RMSEA = 0.042, CFI = 0.998, TLI = 0.997).

Hypotheses 2 and 3: Structural models

CPS. Hypothesis 2 was directed towards the relation of MicroFIN and MicroDYN and the question, whether both instruments empirically target the same construct. Figure 2 shows the multitrait-multimethod model utilized to test the hypothesis, a correlated trait–correlated method minus one model (CT-C(M-1); Eid et al., 2003), including the same latent dimensions measured by both instruments and method factors for the number of methods minus one (i.e., one method served as reference method). This way, both instruments were assumed to target the same dimensions of CPS, namely knowledge acquisition and knowledge application, while specific aspects of MicroFIN resulting from the broader range of included problem features being explicitly modelled via method factors for both dimensions of CPS. MicroDYN as the established instrument served as the reference method. The resulting model fitted the data well (χ^2 (285) = 445.247, $p < .001$, RMSEA = 0.031, CFI = 0.980, TLI = 0.978).

Similar to the results for the separate measurement models of MicroFIN and MicroDYN, the latent indicators for knowledge acquisition and knowledge application correlated significantly in the CT-C(*M*-1) model ($r = .82$, $p < .001$, see Figure 2). Both method factors for MicroFIN were significantly correlated, pointing towards a somewhat generalized method effect ($r = .54$, $p < .001$, see Figure 2). Loadings of individual MicroFIN and MicroDYN items on the latent factors of CPS and on the method factors for the MicroFIN items can be found in the Appendix. An alternative model (not presented) with MicroFIN as the reference method yielded comparable results with regard to correlations and loadings. In summary, Hypothesis 2, which assumed the assessment of the same constructs by both instruments, was supported, with method specific factors differentiating their assessment through MicroFIN from MicroDYN.

Reasoning. Targeting Hypothesis 3, we included a separate latent factor for reasoning in the multitrait-multimethod model described above (see also Figure 2). The resulting model fitted the data well (χ^2 (361) = 540.359, $p < .001$, RMSEA = 0.029, CFI = 0.978, TLI = 0.976) and featured significant correlations between knowledge acquisition and reasoning ($r = .63$, $p < .001$) and knowledge application and reasoning ($r = .60$, $p < .001$). The method factors of MicroFIN were also significantly correlated with reasoning (knowledge acquisition, $r = .34$, $p < .001$, knowledge application, $r = .28$, $p < .001$).

The correlations of the CPS dimensions with reasoning did not indicate identity of the two concepts: For knowledge acquisition the 99% confidence interval of the correlation was [.55, .71], for knowledge application [.50, .71], both of them far and significantly different from values near one. Again, utilizing MicroFIN as the reference method led to similar results (not presented). In summary, the results supported Hypothesis 3, pointing to a moderate relation between the dimensions of CPS and reasoning but not identity and an influence of reasoning on the method specific factors.

Discussion

The extension of CPS assessment to include multiple tasks formalized as finite state automata led to a broader range of problem features to be included in our assessment, thereby overcoming the homogeneity of current instruments building on LSE-tasks (see Table 1 and the detailed description of tasks in the Appendix).

(1) Securing psychometric properties and a measurement model for MicroFIN. The reliability of MicroFIN naturally decreased with the range of included features not shared between tasks and the rather low number of tasks. However, adequate reliability could still be achieved for both dimensions of CPS. Together with the reasonable item difficulties, these findings highlight the empirical viability of including a broader range of problem features in CPS assessment via MicroFIN (Hypothesis 1a supported). On the other hand, the results also underline the benefits in terms of reliability when utilizing homogeneous tasks in assessment as in MicroDYN (e.g., Greiff & Wüstenberg, 2014). With regard to dimensionality, empirical results indicated the separability of knowledge acquisition and knowledge application in

MicroFIN, thereby supporting our expectations of an assessment allowing for the differentiation of both theoretically derived dimensions of CPS (Hypothesis 1b supported).

(2) Relating MicroFIN to an established instrument of CPS assessment. Combining MicroFIN with MicroDYN showed the expected proximity of both instruments when analyzed with the help of a multitrait-multimethod model. The approaches indicated the same latent traits, with method effects differentiating both instruments and their specific features (Hypothesis 2 supported). While MicroFIN generally targeted the same skills as MicroDYN, we also saw specific problem features realized within MicroFIN leading to differently accentuated requirements (e.g., a different set of necessary exploration strategies as in the case of the 'Fish-o-maton'). This finding is represented in the substantial loadings on the method factors for MicroFIN. Looking at the heterogeneity of the underlying construct and the general goal of this study, the expansion of heterogeneity in assessment instruments to new problem relations with high real-world importance, method effects like these come as no surprise. On the contrary, they show the need for a broader assessment of CPS that also includes heterogeneous sets of tasks not covered within LSE-based instruments alone. We would expect these additional elements included in MicroFIN and represented in the method factors to be generally useful, for example by providing additional value when predicting real-life indicators of successful problem solving.

(3) Exploring the relations to reasoning. The correlations between the dimensions of CPS as resulting from the multitrait-multimethod model and the CogAT as an indicator for reasoning ability were significant and in the range reported for MicroDYN by Wüstenberg et al. (2012). Both method factors of MicroFIN were also significantly correlated to reasoning, even though the relation was weaker than the relation between reasoning and the dimensions of CPS. That is, the additional requirements introduced by the more heterogeneous assessment of CPS were also associated with reasoning, as could be expected, but not exclusively due to a higher influence of reasoning in MicroFIN. In summary, both dimensions

of CPS and the method factors of MicroFIN showed substantial relations to reasoning, but not identity (Hypothesis 3 supported).

Naturally, there are also limitations to take into account when considering the results of our study. Bearing in mind the enormous possibilities of the FSA framework, the sample of MicroFIN tasks was not representative of either restrictions or focus of FSA-based CPS assessment in general. We combined the framework of FSA with the approach of multiple complex systems, highlighting some first steps of how to exceed the margins of current LSE-based CPS assessments (see Table 1) and establishing the empirical feasibility of such an endeavor, but an exhaustive analysis of different problem features was beyond the scope of this study. Furthermore, an analysis of exploration strategies, necessary in response to the broader range of problem features implemented in MicroFIN, was not included here (see for example Rollett, 2008, for a comprehensive account on strategies in LSE-based tasks). The analyses of participants' process data certainly offers ample opportunity to dive into differences in terms of underlying process requirements in response to differently structured complex problems of various sizes (e.g., featuring a variety of possible states and problem features). And the availability of heterogeneous problem features combined in one approach of assessment might highlight interindividual differences in the adaptability of strategies across differently structured problems (see McElhaney & Linn, 2011). Future studies will also have to show whether the additional requirements introduced by MicroFIN allow for a better prediction of external outcomes of successfully handling complex problems outside of assessment contexts and how MicroFIN compares to the broader range of complex problems utilized in research (i.e., also in relation to instruments not focusing on assessment, e.g., Gonzalez, Vanyukov, & Martin, 2005). Finally, the question remains to be answered whether the semantic embedding used in MicroFIN represents a problem by triggering the formation of hypotheses that are not systematically tested during exploration (cf. Beckmann & Goode, 2013). Effects of semantic embedding as identified by Beckmann and Goode (2013) for LSE-based tasks and potential ways to minimize them remain to be researched for MicroFIN and other instruments building on FSA.

Looking at the assessment of CPS, MicroFIN represents the expansion of instruments to a formal framework that was previously limited to applications in laboratory settings, namely finite state automata. The framework is opening up the possibility of formalizing a broad range of problem features previously excluded from CPS assessment, thereby paving the way towards more heterogeneity of complex problems included in large-scale assessments, such as PISA or national school monitoring efforts. And by combining the formal framework of finite state automata with the use of multiple complex systems, the MicroFIN approach is facilitating a CPS assessment overcoming the current homogeneity of tasks building on LSE, while maintaining their advantage of a psychometrically sound foundation. In this paper we established the empirical applicability of the MicroFIN approach in a context of assessment and our findings give hope for a CPS assessment reaching out to a more valid reflection of the complex problems we encounter in the real world.

References

Ackerman, P. L. (1992). Predicting individual differences in complex skill acquisition: dynamics of ability determinants. *Journal of Applied Psychology*, *77*, 598–614.

Anderson, J. A. (2006). *Automata theory with modern applications*. Cambridge, MA: Cambridge University Press.

Beckmann, J. F., & Goode, N. (2013). The benefit of being naïve and knowing it: the unfavourable impact of perceived context familiarity on learning in complex problem solving tasks. *Instructional Science*, 1–20.

Brehmer, B. (1992). Dynamic decision making: Human control of complex systems. *Acta Psychologica*, *81*, 211–241.

Buchner, A. (1995). Basic topics and approaches to the study of complex problem solving. In P. A. Frensch & J. Funke (Eds.), *Complex problem solving: The European perspective* (pp. 27–63). Hillsdale, NJ: Erlbaum.

Buchner, A., & Funke, J. (1993). Finite-state automata: Dynamic task environments in problem-solving research. *The Quarterly Journal of Experimental Psychology*, *46*, 83–118.

Bühner, M., Kröner, S., & Ziegler, M. (2008). Working memory, visual–spatial-intelligence and their relationship to problem-solving. *Intelligence*, *36*, 672–680.

Dörner, D., Kreuzig, H. W., Reither, F., & Stäudel, T. (Eds.). (1983). *Lohausen. Vom Umgang mit Unbestimmtheit und Komplexität. [Lohhausen. On dealing with uncertainty and complexity]*. Switzerland, Bern: Huber.

Eid, M., Lischetzke, T., Nussbeck, F. W., & Trierweiler, L. I. (2003). Separating trait effects from trait-specific method effects in multitrait-multimethod models: A multiple-indicator CT-C(M-1) model. *Psychological Methods*, *8*, 38–60.

Fischer, A., Greiff, S., & Funke, J. (2012). The process of solving complex problems. *Journal of Problem Solving*, *4*, 19–42.

Funke, J. (2001). Dynamic systems as tools for analysing human judgement. *Thinking & Reasoning*, *7*, 69–89.

Funke, J. (2010). Complex problem solving: a case for complex cognition? *Cognitive Processing*, *11*, 133–142.

Gonzalez, C., Vanyukov, P., & Martin, M. K. (2005). The use of microworlds to study dynamic decision making. *Computers in Human Behavior*, *21*, 273–286.

Greiff, S., & Funke, J. (2009). Measuring complex problem solving: the MicroDYN approach. In F. Scheuermann & J. Björnsson (Eds.), *The transition to computer-based assessment: New approaches to skills assessment and implications for large-scale testing* (pp. 157–163). Luxembourg, Luxembourg: Office for Official Publications of the European Communities.

Greiff, S., & Wüstenberg, S. (2014). Assessment with microworlds using MicroDYN: Measurement invariance and latent mean comparisons: psychometric properties across several student samples and blue-collar workers. *European Journal of Psychological Assessment*, *Advance online publication*. doi:10.1027/1015-5759/a000194

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

Greiff, S., Wüstenberg, S., Molnár, G., Fischer, A., Funke, J., & Csapó, B. (2013). Complex problem solving in educational contexts—Something beyond g: Concept, assessment, measurement invariance, and construct Validity. *Journal of Educational Psychology*, *105*, 364–379.

Griffin, P., McGaw, B., & Care, E. (Eds.). (2012). *Assessment and teaching of 21st century skills*. New York, NY: Springer.

Heller, K. A., & Perleth, C. (2000). *Kognitiver Fähigkeitstest für 4. bis 12. Klassen, Revision. [Cognitive Abilities Test (CogAT; Thorndike, L. & Hagen, E., 1954-1986) - German adapted version]*. Göttingen, Germany: Beltz-Test.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*, 1–55.

Levy, S. T., & Wilensky, U. (2011). Mining students' inquiry actions for understanding of complex systems. *Computers & Education*, *56*, 556–573.

Little, T. D., Cunningham, W. A., Shahar, G., & Widaman, K. F. (2002). To parcel or not to parcel: Exploring the question, weighing the merits. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*, 151–173.

McElhaney, K. W., & Linn, M. C. (2011). Investigations of a complex, realistic task: Intentional, unsystematic, and exhaustive experimenters. *Journal of Research in Science Teaching*, *48*, 745–770.

Muthén, B. O., du Toit, S. H. C., & Spisic, D. (1997). *Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes*. Unpublished technical report. Retrieved from http://pages.gseis.ucla.edu/faculty/muthen/articles/Article_075.pdf

Muthén, L. K., & B. O. Muthén. (2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

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

Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology*, *3*, 1–18.

OECD. (2013). *PISA 2012 assessment and analytical framework*. Paris, France: Organisation for Economic Co-operation and Development.

Osman, M. (2010). Controlling uncertainty: A review of human behavior in complex dynamic environments. *Psychological Bulletin*, *136*, 65–86.

Raven, J., Raven, J. C., & Court, J. H. (1998). *Manual for Raven's progressive matrices and vocabulary scales*. Oxford, England: Oxford Psychologists Press.

Rich, E. (2008). *Automata, Computability and Complexity: Theory and Applications*. Upper Saddle River, NJ: Prentice Hall.

Rollett, W. (2008). *Strategieeinsatz, erzeugte Information und Informationsnutzung bei der Exploration und Steuerung komplexer dynamischer Systeme [Use of strategy, generated information and use of information when exploring and controlling complex dynamic systems].* Berlin, Germany: Lit.

Scherer, R., & Tiemann, R. (2012). Factors of problem-solving competency in a virtual chemistry environment: The role of metacognitive knowledge about strategies. *Computers & Education*, *59*, 1199–1214.

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.

Segerer, R., Marx, A., & Marx, P. (2012). Unlösbare Items im KFT 4-12+R [Unsolvable items in the CFT 4-12+R]. *Diagnostica*, *58*, 45–50.

Sonnleitner, P., Brunner, M., Greiff, S., Funke, J., Keller, U., Martin, R., … Latour, T. (2012). The Genetics Lab: Acceptance and psychometric characteristics of a computer-based microworld assessing complex problem solving. *Psychological Test and Assessment Modeling*, *54*, 54–72.

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

Sternberg, R. J., & Frensch, P. A. (1991). *Complex problem solving: Principles and mechanisms*. Hillsdale, NJ: Erlbaum.

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

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

Wittmann, W., & Süß, H.-M. (1999). Investigating the paths between working memory, intelligence, knowledge, and complex problem-solving performances via Brunswik symmetry. In P. L. Ackerman, P. C. Kyllonen, & R. D. Roberts (Eds.), *Learning and individual differences: Process, trait, and content determinants* (pp. 77–108). Washington, DC: American Psychological Association.

Wüstenberg, S., Greiff, S., & Funke, J. (2012). Complex problem solving — More than reasoning? *Intelligence*, *40*, 1–14.

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α, Revelle's β, and Mcdonald's ωH: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika*, *70*, 123–133.

Figure 1

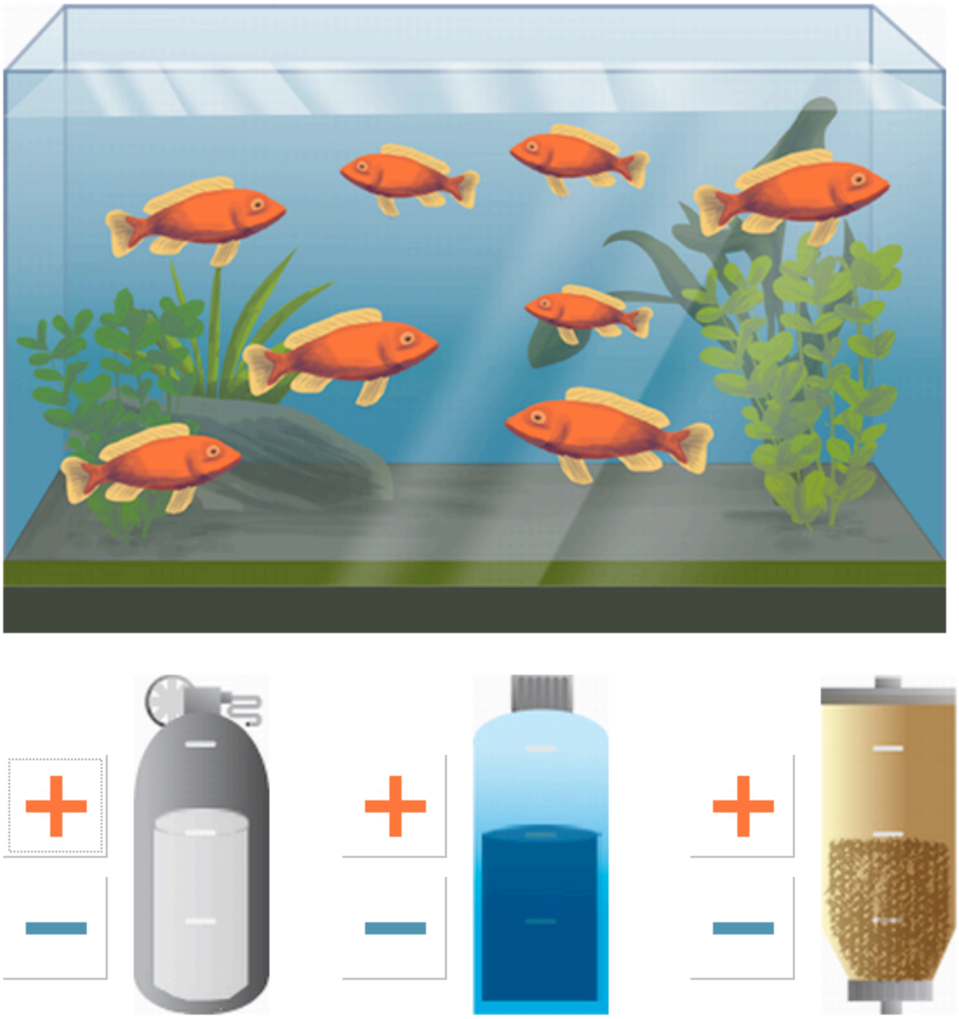

Figure 1: Screenshot of the MicroFIN task Fish-o-mat. The three containers at the bottom represent the input variables, the aquarium on top the output variable. Input variables' values can be varied by pressing the plus and minus signs next to them. Each of the input variables features four input values (empty, low, medium, high).

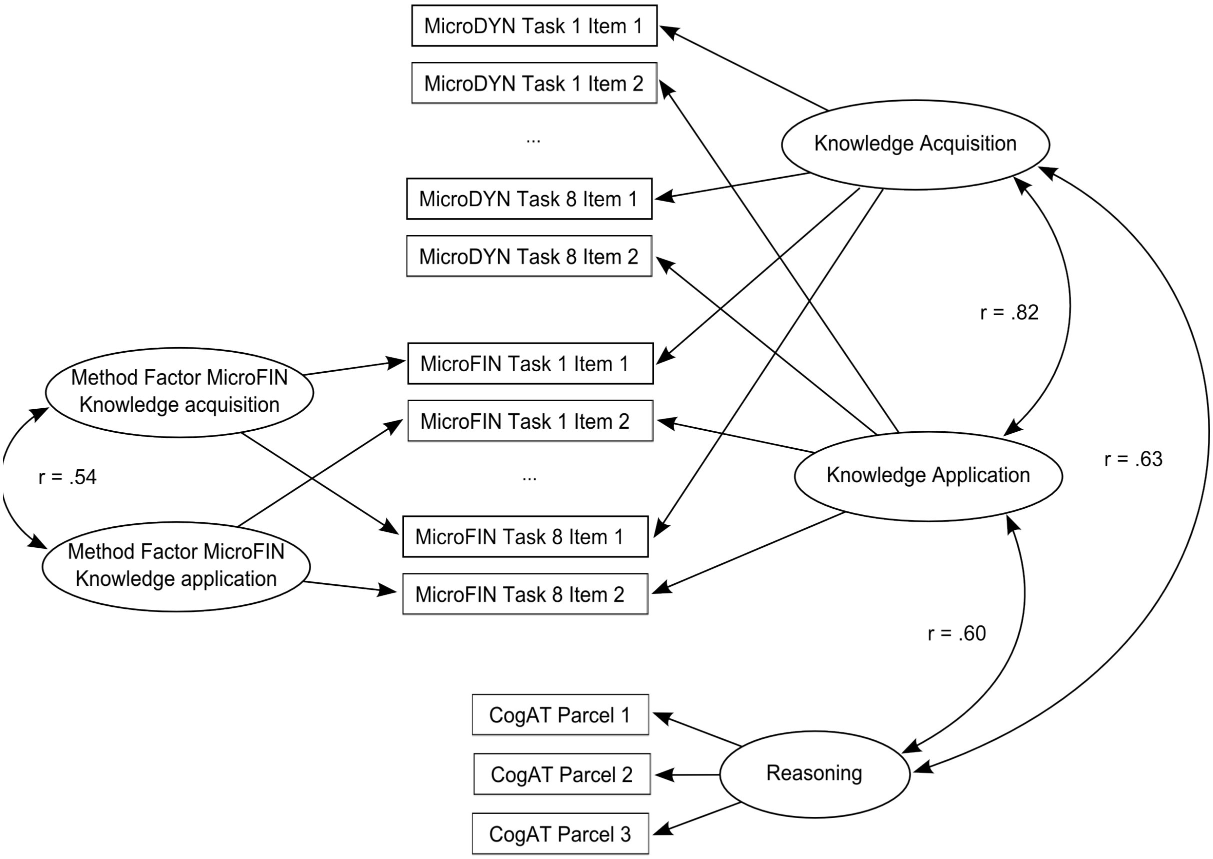

Figure 2: CT-C(M-1) multitrait-multimethod model. The model depicts the latent relations between the dimensions of CPS, knowledge acquisition and knowledge application, two method factors modeling specific aspects of MicroFIN for both dimensions, and a latent factor indicating participants' reasoning ability. Latent correlations for the method factors, error variances, and manifest indicators for some MicroFIN and MicroDYN tasks were omitted from the figure for better accessibility. The factor loadings of the model are available in the Appendix.

Table 1

Table 1

Features of the utilized MicroFIN tasks

Task	Nb Input Variables[a]	Nb of States[b]	Input Functioning[c]	Output Functioning[c]	Other Features[d]
Task 1: Concert-o-maton	2 x 2 x 3	12	Quantitative (ordinal) and qualitative (dichotomous)	Quantitative (ordinal)	Opposite subsystems
Task 2: Plan-o-maton	4 x 4 x 4 x 4	24	Qualitative (nominal)	Qualitative (dichotomous)	Relative positioning
Task 3: Plant-o-maton	4 x 4 x 4	64	Quantitative (ordinal)	Qualitative (nominal)	Threshold value
Task 4: Fish-o-maton	4 x 4 x 4	64	Quantitative (ordinal)	Quantitative (ordinal) and Qualitative (nominal)	Equivalence
Task 5: Flooz-o-maton	3 x 3 x 3	27	Quantitative (ordinal) and qualitative (nominal)	Quantitative (ordinal) and Qualitative (nominal)	Start button and Threshold values

Note: Table intended to give an overview of the features varied between MicroFIN tasks.

[a] Number of input variables and possible values per variable. Example: 2 x 3 means one variable with two possible values and another variable with three possible values

[b] Number of different states, the finite state automaton can be in (i.e., combinations of input and output variables)

[c] Input/Output variables are about qualitative (i.e., dichotomous or nominal) or quantitative (i.e., ordinal) variation, or combinations of both

[d] Other features of the task, representing unique characteristics: Opposite subsystems: The task features two subsystems with opposite reaction to input variation. Relative positioning: Relative position of subelements as relevant input for transitions between output states. Start button: Effects of input manipulation are only observable after pressing a button (comparable to MicroDYN). Threshold values: The tasks behavior changes if one variable is put above a certain level. Equivalence: The task features equilibrium states as relevant input. For further information please refer to the detailed description of tasks in the Appendix.

The commonalities between task features as shown in the table are considerably lower in FSA than for tasks based on LSE, where a comparable line in Table 1 for *all* items would read: Number of input variables ranging from two to four with up to 20 values per input variable (i.e., finer granulation in inputs). Quantitative inputs and outputs and a start button as other feature.

Appendix

Detailed description of the MicroFIN tasks

Task 1: Concert-o-maton:

The task 'Concert-o-maton' features combinations of the input variables 'music group' (two qualitatively different options) and 'stage' (two options). Both can be varied independently. Additionally, the factor 'admission fee' can be varied (three ordinal options). Based on the combination of 'music group' and 'stage', the influence of 'admission fee' on the number of visitors (outcome variable, three ordinal values) is varied. One option of both input variables is matched to another one, leading to variations in consequence to the height of the admission fee for a specific combination of 'music group' and 'stage'. The effect of admission fee is reversed for a second combination of 'music group' and 'stage', while other combinations result in no influence of admission fee on the output variable at all.

Task 2: Plan-o-maton

Four input values (depictions of different buildings) are presented in a 2 x 2 matrix format. The position of the values can be exchanged on a bilateral level (e.g., changing the position of the picture from top left to top right, and vice versa) by pressing a button. Output values (four variables, each with two options) are presented between the pictures, their value based on the combination of the pictures (two pairs of the buildings can be matched). The combination of pictures leads to output values independently of the place the pictures are shown, just based on the combination.

Task 3: Plant-o-maton

Three input variables (four ordinal values each) are related to one output variable (three ordinal values). One of the input variables features a threshold value, after which the direction of influence on the output variable is changed from positive to negative. The other two variables are related to the output variable in a continuous positive and negative way.

Task 4: Fish-o-maton

Three input variables (four ordinal values each) can be manipulated, leading to different values in one output variable (an aquarium, five values). Variations in the output variable occur, if the input variables are brought to equivalent values (e.g., all input variable on medium setting leading to the display of a medium amount of fish). See also the article for a detailed description.

Task 5: Flooz-o-maton

Combinations of three different inputs with three values each (ordinal) have to be explored to create a cocktail. Three types of cocktails with different features (e.g., the amount of sugar) can be created, while some input combinations lead to no cocktail at all (i.e., threshold values). The microworld features a start button.

Online Appendix 2

Table A1

Standardized trait and method factor loadings, consistency, and method specifity of MicroFIN and MicroDYN items

			CPS dimensions		Method factors MicroFIN			
Instrument	Task	Item	Knowledge acquisition	Knowledge application	Knowledge acquisition	Knowledge application	Consistency	Method specifity
MicroFIN	Task 1: Concert-o-maton	Item 1	.502		.572		.435	.565
		Item 2		.358		.557	.292	.708
	Task 2: Plan-o-maton	Item 1	.510		.351		.678	.321
		Item 2		.339		.502	.313	.687
	Task 3: Plant-o-maton	Item 1	.499		.458		.543	.457
		Item 2		.355		.606	.255	.745
	Task 4: Fish-o-maton	Item 1	.511		.442		.572	.428
		Item 2		.430		.582	.353	.647
	Task 5: Flooz-o-maton	Item 1	.332		.374		.441	.559
		Item 2		.308		.439	.330	.670
MicroDYN	Task 1: Cat	Item 1	.842				1	0
		Item 2		.623			1	0
	Task 2: Moped	Item 1	.777				1	0
		Item 2		.817			1	0
	Task 3: Game night	Item 1	.810				1	0
		Item 2		.898			1	0
	Task 4: Perfume	Item 1	.897				1	0
		Item 2		.910			1	0
	Task 5: Gardening	Item 1	.896				1	0
		Item 2		.804			1	0
	Task 6: Handball team	Item 1	.729				1	0
		Item 2		.525			1	0
	Task 7: Spaceship	Item 1	.590				1	0
		Item 2		.315			1	0
	Task 8: First aid	Item 1	.750				1	0
		Item 2		.570			1	0

Note. All loadings reported in the table are significant on a $p \leq .01$ level. The differences in loadings between MicroFIN and MicroDYN on the latent dimensions of CPS reverse when the reference method is changed from MicroDYN to MicroFIN.

Table A2

Linear structural equations, system size, and type of effects for the MicroDYN tasks

Task	Linear Structural Equations	System size	Effects
Task 1: Cat	X_{t+1} =$1*X_t + 0*A_t + 2*B_t$ Y_{t+1} =$1*Y_t + 0*A_t + 2*B_t$	2 x 2 - System	only effects of inputs
Task 2: Moped	$X_{t+1} = 1*X_t + 2*A_t + 2*B_t + 0*C_t$ $Y_{t+1} = 1*Y_t + 0*A_t + 0*B_t + 2*C_t$	2 x 3 - System	only effects of inputs
Task 3: Game night	$X_{t+1} = 1*X_t + 0*A_t + 2*B_t + 0*C_t$ $Y_{t+1} = 1*Y_t + 2*A_t + 0*B_t + 0*C_t$ $Z_{t+1} = 1*Z_t + 0*A_t + 0*B_t + 2*C_t$	3 x 3 - System	only effects of inputs
Task 4: Perfume	$X_{t+1} = 1*X_t + 2*A_t + 0*B_t + 0*C_t$ $Y_{t+1} = 1*Y_t + 0*A_t + 2*B_t + 2*C_t$ $Z_{t+1} = 1*Z_t + 0*A_t + 0*B_t + 2*C_t$	3 x 3 - System	only effects of inputs
Task 5: Gardening	$X_{t+1} = 1*X_t + 2*A_t + 2*B_t + 0*C_t$ $Y_{t+1} = 1*Y_t + 0*A_t + 2*B_t + 0*C_t$ $Z_{t+1} = 1*Z_t + 0*A_t + 0*B_t + 2*C_t$	3 x 3 - System	only effects of inputs
Task 6: Handball team	$X_{t+1} = 1.33*X_t + 2*A_t + 0*B_t + 0*C_t$ $Y_{t+1} = 1*Y_t + 0*A_t + 0*B_t + 2*C_t$	2 x 3 - System	effects of inputs & outputs
Task 7: Spaceship	$X_{t+1} = 1*X_t + 0*A_t + 0*B_t + 0*C_t$ $Y_{t+1} = 1.33*Y_t + 2*A_t + 2*B_t + 0*C_t$ $Z_{t+1} = 1*Z_t + 0*A_t + 0*B_t + 2*C_t$	3 x 3 - System	effects of inputs & outputs
Task 8: First aid	$X_{t+1} = 1*X_t + 2*A_t + 0*B_t + 0*C_t$ $Y_{t+1} = 1*Y_t + 2*A_t + 0*B_t + 0*C_t$ $Z_{t+1} = 1.33*Z_t + 0*A_t + 0*B_t + 2*C_t$	3 x 3 - System	effects of inputs & outputs

Note: Features of the MicroDYN tasks.
Linear Structural Equations: Values of the output variables (X, Y, Z) at time t+1 depending on input and output (A, B, C) variables at time t.
System size: Number of input and output variables.
Effects: Only effects of input variables on output variables or effects of both, input and output variables on output variables (e.g., including Eigendynamics)

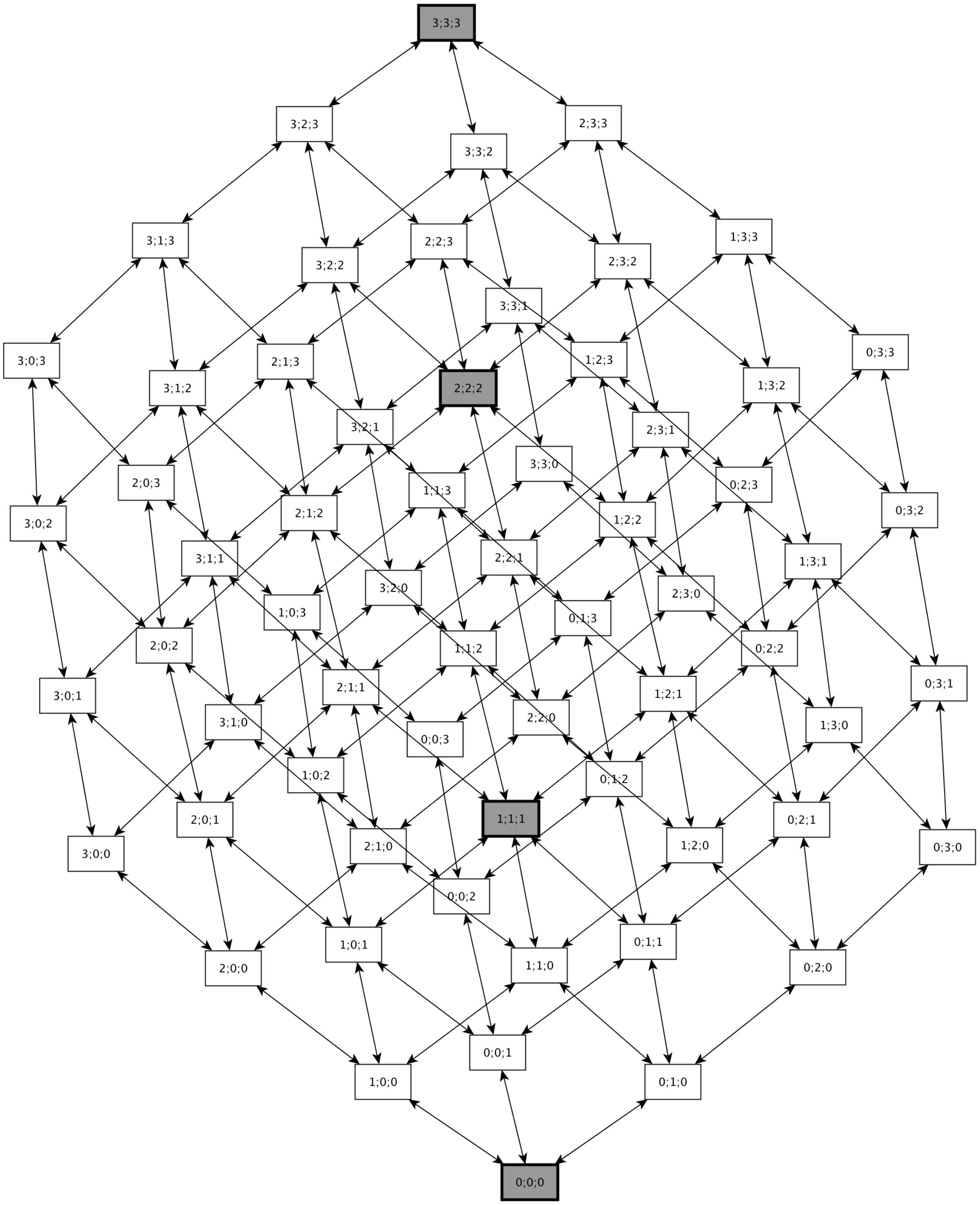

Figure A1: State-transition diagram of the MicroFIN task Fish-o-mat. Each rectangle represents one state of the task, with the numbers indicating the input levels for all three input variables (see Figure 1a). Arrows represent transitions between states due to changes in input variable values. Gray states in the state-transition diagram are indicating states with equilibrium in input values, and hence, fish in the aquarium.