

Re-evaluating the psychometric properties of MicroFIN: A multidimensional measurement of complex problem solving or a unidimensional reasoning test?

André Kretzschmar¹, Liena Hacatrijana² & Malgozata Rascevska²

Abstract

The present study investigated the psychometric properties of an extended version of the MicroFIN test as the latest development in complex problem solving (CPS) assessments within the multiple complex systems (MCS) approach. Specifically, we examined factorial validity, reliability, and the relation to reasoning using data from 362 Latvian high school students (mean age 16.82 years). Results indicated that the commonly applied 2-dimensional measurement model for MCS-based CPS tests did not fit the data better than a 1-dimensional measurement model. Furthermore, the extended version of MicroFIN showed satisfactory reliability. With regard to the relation to reasoning, we found latent correlations ranging from .38 to .78 depending on the operationalization of reasoning. Explanations for the findings (e.g., the impact of the Brunswik symmetry principle) and implications for MicroFIN, the assessment of CPS in general, and future CPS research are discussed.

Keywords: complex problem solving, reasoning, factorial validity, reliability, Brunswik symmetry

¹ *Correspondence concerning this article should be addressed to:* André Kretzschmar, PhD, University of Tübingen, Hector Research Institute of Education Sciences and Psychology, Europastraße 6, 72072 Tübingen, Germany; email: kretzsch.andre@gmail.com

² University of Latvia

Complex problem solving (CPS) skills are considered to be highly important in today's rapidly changing and increasingly complex world, (e.g., Funke, 1999; Kretzschmar & Süß, 2015; Neubert, Mainert, Kretzschmar, & Greiff, 2015; OECD, 2014) and hence, research about the assessment of CPS skills has been increasing in recent years (e.g., Greiff, Wüstenberg, & Funke, 2012; Kröner, Plass, & Leutner, 2005; Sonnleitner et al., 2012). MicroFIN (Neubert, Kretzschmar, Wüstenberg, & Greiff, 2015) is the latest development within the framework of multiple complex systems (MCS; Greiff et al., 2012), which was developed to overcome the psychometrical limitations of previous CPS tests and assessment frameworks (for an overview, see e.g., Greiff, Fischer, Stadler, & Wüstenberg, 2014).

Although the development of the MCS approach can be considered an important milestone in the CPS research field, the most prominent MCS-based tests MicroDYN (Greiff et al., 2012) and Genetics Lab (Sonnleitner et al., 2012) have been criticized as they only cover very selected characteristics of complex problems (e.g., Funke, 2010, 2014; Funke, Fischer, & Holt, 2017; Kretzschmar, 2017; Scherer, 2015; Schoppek & Fischer, 2015), particularly compared to more comprehensive CPS tests (a.k.a. microworlds) such as Tailorshop (Putz-Osterloh, 1981), FSYS (Wagener, 2001), or LEARN! (Grossler, Maier, & Milling, 2000). For example, MicroDYN and Genetics Lab tasks as applied in previous studies rely on only one specific strategy (vary-one-thing-at-a-time, VOTAT; see e.g., Chen & Klahr, 1999) or a very close adaptation of it (see Beckmann and Goode, 2014, who summarized the slightly different strategies as a vary-one-or-none-at-a-time-strategy, VONAT). Although the VOTAT (or VONAT) strategy is important in many contexts (see e.g., Wüstenberg, Stadler, Hautamäki, & Greiff, 2014), it is obvious that one strategy is not sufficient to solve the variety of problems that can arise in complex problem solving research or even in daily life (e.g., Funke, 2014; Funke et al., 2017). Therefore, the psychometrically advantageous homogeneity of current versions of MicroDYN and Genetics Lab can be considered a threat to a content-valid operationalization of CPS (see, e.g., Neubert, Kretzschmar, et al., 2015; Scherer, 2015; Schoppek & Fischer, 2015).

MicroFIN was developed to address the limitations of established MCS-based tests. Specifically, the development of MicroFIN was partially guided by the rationale to create tasks which are not solvable by solely applying the VOTAT strategy; instead, different problem solving strategies are required in each task (Kretzschmar, 2015; Neubert, Kretzschmar, et al., 2015). MicroFIN, therefore, has the potential to narrow the gap between highly reliable but homogeneous MCS-based assessment tools and the psychometrically less convincing but ecologically valid microworlds that have been applied in CPS research in recent decades (Kretzschmar, 2017). However, whether and to what extent MicroFIN fulfills this expectation still needs further investigation, although first evidence supports the view of MicroFIN as a heterogeneous CPS test (see Müller, Kretzschmar, & Greiff, 2013).

The present study aims to answer a more fundamental research question about MicroFIN: its psychometric quality. Apart from in the initial study by Neubert et al. (2015), the psychometric properties of MicroFIN have not yet been addressed comprehensively. Due to the heterogeneity of the included tasks, the test is expected to face challenges in terms

of reliability but more convincing findings in terms of construct validity (see attenuation paradox; Loevinger, 1954). As will be outlined below, empirical findings reveal an unclear pattern and the need for further research. Therefore, the purpose of the present study is to examine the psychometric properties of a further developed version of MicroFIN with regard to three important issues in CPS research: (1) factorial validity, (2) reliability, and (3) relation to reasoning (fluid intelligence; see McGrew, 2009).

Psychometric Properties of State-of-the-Art CPS Assessment Tools

Factorial validity: Structure of MCS-based tests

From a theoretical point of view, several sub-processes of CPS such as knowledge acquisition and knowledge application have been identified (e.g., Dörner, 1986; Fischer, Greiff, & Funke, 2012). Consequently, one of the main goals when developing CPS assessment tools is to represent these CPS sub-processes. Recent developments in CPS tools within the MCS framework have made remarkable progress with regard to the dimensionality of CPS tests. For example, Genetics Lab (Sonnleitner et al., 2012) provides performance scores on three different CPS dimensions: exploration behavior, knowledge acquisition, and knowledge application. Although these scores are highly correlated, as theoretically expected, a multi-dimensional measurement model with three distinguishable dimensions has been consistently validated empirically (e.g., Sonnleitner, Brunner, Keller, & Martin, 2014; Sonnleitner, Keller, Martin, & Brunner, 2013). MicroDYN (Greiff et al., 2012), the second MCS-based test, also started with a 3-dimensional measurement model (Greiff et al., 2012), but further research demonstrated that an empirical distinction between exploration behavior and knowledge acquisition was untenable (e.g., Wüstenberg, Greiff, & Funke, 2012). Consequently, a 2-dimensional measurement model based only on knowledge acquisition and knowledge application, omitting the performance score on exploration behavior, has been applied in most relevant studies (e.g., Greiff et al., 2013; Greiff, Kretzschmar, Müller, Spinath, & Martin, 2014; Kretzschmar, Neubert, Wüstenberg, & Greiff, 2016; Lotz, Sparfeldt, & Greiff, 2016). MicroFIN (Neubert, Kretzschmar, et al., 2015), the most recently developed CPS test, was also created to represent the sub-processes of CPS. Hence, a 2-dimensional measurement model similar to MicroDYN was empirically confirmed in the initial study (Neubert, Kretzschmar, et al., 2015). However, the only other study investigating the dimensionality of MicroFIN as of yet (Kretzschmar et al., 2016) did not provide evidence for a multidimensional measurement model. Instead, the authors argued for an empirically supported 1-dimensional model in which an aggregated score combining knowledge acquisition and knowledge application was used for each task. The resulting task scores were then used as indicators for a latent MicroFIN factor.

It should be noted that Kretzschmar et al.'s (2016) measurement model for MicroFIN also avoids a pitfall of the commonly applied measurement models for CPS tests within

the MCS framework.³ In these models, separate performance scores for knowledge acquisition and knowledge application (and for exploration behavior in the case of Genetics Lab) are calculated for each task (see e.g., Greiff et al., 2012; Neubert, Kretzschmar, et al., 2015). These performance scores are then used as independent indicators of the latent factors, even though they are in fact not independent. Acquired knowledge about a CPS task depends on exploration behavior, and performance regarding the application of knowledge depends on the acquired knowledge within a given task. Therefore, the three indicators for each task might share a correlated uniqueness (Brown, 2015; cf. local stochastic dependence in item response theory). As correlated uniqueness can have a substantial impact on the factor structure (see e.g., Brown, 2003; Marsh, 1996), it should be considered in the measurement model. Interestingly, previous studies investigating the measurement models of MCS-based CPS tests have ignored this issue.⁴ One might argue that presenting the correct causal structure immediately before each knowledge application item (partially) solves this problem (Greiff, Fischer, et al., 2014), meaning that there is no need to consider correlated uniqueness in the measurement models. However, to our knowledge, this assumption has not yet been empirically investigated. In fact, it is possible that (some) participants might nevertheless try to achieve the goals in the knowledge application phase on the basis of incorrect knowledge acquired during exploration. Kretzschmar et al.'s (2016) measurement model avoids the issue of correlated uniqueness as the interdependent performance indicators for each CPS task are aggregated into a task score, which then is used as an indicator for a latent factor. Therefore, this measurement model can be considered a parsimonious alternative (in the case of unidimensionality), avoiding the problem of correlated uniqueness completely.

In summary, it seems that the dimensionality of MCS-based CPS tests is not as clear as expected. Whereas most studies featuring Genetics Lab or MicroDYN have provided cumulative evidence for at least a 2-dimensional measurement model, the state of research about the factorial validity of MicroFIN is uncertain. Furthermore, previous studies did not consider the dependency of the indicators in a multidimensional measurement model and its impact on the dimensionality of CPS tests. The first research issue for the present study was thus to examine the dimensionality of an extended version of MicroFIN. We investigated five different measurement models: a 2-dimensional model without/with correlated errors, a 1-dimensional model without/with correlated errors, and a 1-dimensional model using aggregated task scores (see Figure 2). Based on the development rationale of MicroFIN (Neubert, Kretzschmar, et al., 2015), we hypothesized that a 2-dimensional model representing the CPS sub-processes knowledge acquisition and knowledge application would fit the data better than a 1-dimensional model (Hypothesis 1). Furthermore, we expected significant correlations between indicators for the same task (i.e., correlated uniqueness).

³ We thank an anonymous reviewer, who suggested elaborating on this issue in the present article.

⁴ However, see Sonnleitner et al. (2013) for a correlated uniqueness model based on a different rationale.

Reliability

High reliability is another key feature of MCS-based assessment tools (Greiff et al., 2012). More precisely, an internal consistency of up to $\alpha = .95$ has been reported for MicroDYN and Genetics Lab (e.g., Greiff et al., 2012; Sonnleitner et al., 2012). With regard to MicroFIN, a somewhat lower internal consistency has been found in previous studies ($\omega_h \approx .78$; Neubert, Kretzschmar, et al., 2015). The relatively small number of tasks in combination with their heterogeneity were considered to be the main reasons for this finding. Therefore, this study's second research aim was to inspect the internal consistency of an extended MicroFIN version with partially different tasks and more items per task compared to the initial study (see Neubert, Kretzschmar, et al., 2015). As reliability increases with test length, we expected the extended version of MicroFIN to show at least a similar internal consistency to the initial study (Hypothesis 2).

Furthermore, little is known about the reliability of MCS-based tests apart from internal consistency. In fact, to our knowledge no study has yet investigated the test-retest reliability of a MCS-based CPS assessment tool. We therefore exploratively examined the correlation between MicroFIN performances measured at two measurement times in a pilot study.

Relation between CPS and reasoning

The empirical relation between a potential CPS construct and established intelligence constructs (e.g., reasoning) is one of the most examined research questions in CPS research (see Stadler, Becker, Gödker, Leutner, & Greiff, 2015). Due to the close relation between CPS and intelligence, a relatively low correlation, indicating two distinct constructs, can be seen as the *raison d'être* of CPS research (Kretzschmar et al., 2016). Previous studies have found substantial but nevertheless significantly different from 1.0 correlations between CPS and intelligence. Specifically, a correlation of $M(g) = .59$ (corrected for reliability: $.72$) between CPS measured via the MCS approach and intelligence was reported in Stadler et al.'s (2015) meta-analysis. Unfortunately, previous studies featuring MicroFIN (Kretzschmar et al., 2016; Neubert, Kretzschmar, et al., 2015) did not directly examine the relation between CPS as measured by MicroFIN and intelligence. Instead, only a combined performance score on MicroFIN and MicroDYN was used. As MicroFIN aims to cater to more heterogeneous demands than other CPS tools in the MCS-based approach (e.g., Neubert, Kretzschmar, et al., 2015), it seems worthwhile to examine the relation between CPS measured by MicroFIN and intelligence separately.

In this sense, previous research has provided evidence that reasoning shows the highest relation to CPS compared to other intelligence constructs (e.g., mental speed or memory; Kretzschmar et al., 2016). However, not every operationalization of reasoning performs equally well when examining the correlation between the constructs. Specifically, the importance of a construct valid operationalization has been recently re-emphasized for CPS research in particular (Kretzschmar et al., 2016; Lotz et al., 2016). This means that a

reasoning operationalization based on only one task format and content type (e.g., figural matrices tasks such as Raven's Matrices test) usually underestimates the correlation between CPS and intelligence compared to a broad operationalization (i.e., based on different task formats and types of content; see Kretzschmar et al., 2016). However, the Brunswik symmetry principle (Wittmann, 1988) teaches us that a too broad operationalization of reasoning might also reduce the correlation between reasoning and CPS (e.g., Wittmann & Hatrup, 2004). For example, a very broad operationalization of reasoning based on figural, verbal, and numerical task content might be not the best operationalization of reasoning when its relation to a narrow operationalization of CPS based only on numerical stimuli is to be tested.

As the Brunswik symmetry principle has been barely acknowledged in previous studies investigating the empirical relation between CPS and reasoning (or intelligence in general), we examined the impact of different operationalizations of reasoning (in terms of task content) on the relation between reasoning and CPS. In doing so, we aimed to get a less biased view of this controversial, often-discussed relation. As the MicroFIN tasks primarily contain figural and verbal stimuli, with numerical stimuli present only to a marginal extent, we expected to find the highest correlation between CPS measured by MicroFIN and reasoning based on figural and verbal tasks, in comparison to other operationalizations of reasoning (e.g., numerical and figural tasks; Hypothesis 3a). However, in accordance with Stadler et al.'s (2015) findings, we hypothesized a strong but significantly lower than 1.0 latent correlation between CPS and reasoning independent of the specific operationalization (Hypothesis 3b).

Method

Participants

The study was part of a larger assessment in Latvian high schools ($n = 363$). Not every participant worked on every single test and, thus, a different proportion of missing data occurred for each test (see Table 1, and section below about data analysis). We excluded one participant due to invalid responses in every test. Of the remaining $n = 362$ high school students ($M_{\text{age}} = 16.82$, $SD_{\text{age}} = 1.03$), 54% were female, 33% were male, and 13% did not provide information about gender. All students were invited to participate in the study voluntarily and to take the test as part of computer class. The pilot study for test-retest reliability was conducted on $n = 39$ high school students.

Instruments

Complex problem solving.

MicroFIN is comprehensively described in Kretzschmar (2015) and Neubert, Kretzschmar, et al. (2015). Therefore, we only summarize the main concept and focus on the differences to previously applied versions. MicroFIN consists of several small complex tasks following the multiple complex systems (MCS) approach (Greiff et al., 2012). In

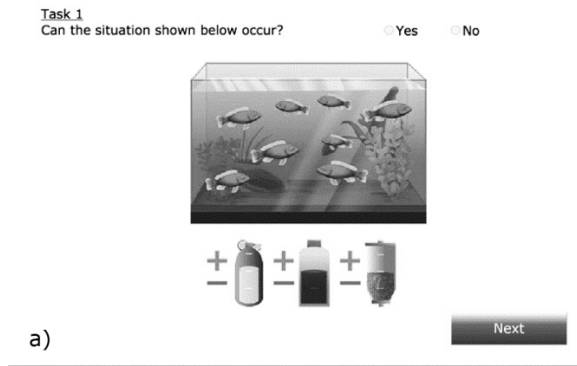
order to assess the two core processes of CPS (i.e., knowledge acquisition and knowledge application), the following procedure was applied in each of the seven tasks. First, participants had to interactively explore an unknown task to acquire as much knowledge as possible about the complex problem (exploration phase; 300 seconds). Second, participants had to answer several questions regarding their acquired knowledge (knowledge acquisition phase; no time limit). Third, participants were asked to reach certain goals (knowledge application phase; 60 seconds per item).

We used a Latvian version of MicroFIN with one warm-up task and six tasks. Specifically, we used the tasks Fish-o-maton (see Figure 1), Plan-o-maton, Concert-o-maton, Plant-o-maton (for details, see Neubert, Kretzschmar, et al., 2015), Green-o-maton (for details, see Kretzschmar et al., 2016), and Wash-o-maton (see Kretzschmar, 2015).

In the knowledge acquisition phase, two different task formats were used. In the first task format – called an identification task (Kretzschmar, 2015) and not investigated in previous studies featuring MicroFIN – four different states of the complex problem were presented. Participants had to decide whether each presented picture represents a valid state of the problem or not (see Figure 1a). This task format is based on Buchner and Funke's (1993) verification tasks and primarily aims to assess knowledge about the range of variables (e.g., number of fish in Fish-o-maton) and possible states (e.g., different colors of the water in Fish-o-maton). Each item was scored dichotomously and further summarized according to the sequential testing approach (Kubinger, 2009). That means full credit was given if all questions were answered correctly, partial credit if half of the questions were answered correctly, and no credit otherwise. In the second task format – named an initial state construction task (Kretzschmar, 2015; Neubert, Kretzschmar, et al., 2015) – a final state and an intervention were presented. Participants were asked to create a valid initial state for the complex problem with the help of predefined elements (see Figure 1b). This task format is based on Buchner and Funke's (1993) retrognostic tasks and primarily aims to assess rule knowledge (i.e., how interventions work). For each task, two items were presented, with the number of elements which had to be used to create the initial state ranging from 1 (as in Figure 1b) to 8. Each item was scored dichotomously and the mean average was calculated. Finally, the mean scores of both knowledge acquisitions task formats were averaged in order to calculate a total performance score for knowledge acquisition for each MicroFIN task. It should be noted that only a selection of knowledge about a problem was assessed in the knowledge acquisition phase. For example, the Fish-o-maton as presented in Figure 1 has 64 possible states (but only five different states with regard to the fish; see Figure A1 in Neubert, Kretzschmar, et al., 2015). Covering all possible states in the knowledge assessment would result in a very time-consuming assessment (not even considering the assessment of the underlying rules/intervention possibilities). Therefore, the applied knowledge items primarily focused on knowledge that was important for the specific control items in the knowledge application phase of each MicroFIN task.

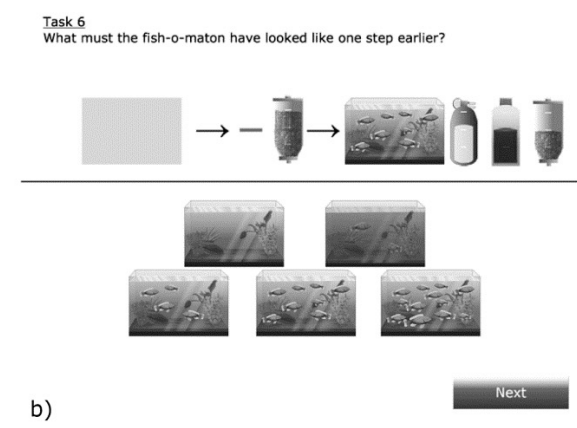
In the knowledge application phase, a specific state of the complex problem was presented and participants were asked to manipulate the complex problem in order to reach a given goal state in as few steps as possible (see Figure 1c). Two items per task were

Task 1
Can the situation shown below occur? Yes No



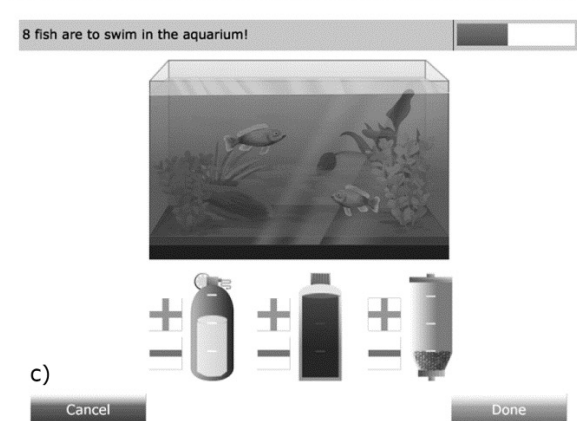
a)

Task 6
What must the fish-o-maton have looked like one step earlier?



b)

8 fish are to swim in the aquarium!



c)

Figure 1:

Screenshots of the knowledge acquisition and knowledge application phases of the MicroFIN task Fish-o-maton: (a) Identification task (knowledge acquisition phase); (b) Initial state construction task (knowledge acquisition phase); (c) Control task (knowledge application phase)

used. The items were scored dichotomously (i.e., goal achieved or not) and averaged to create a total performance score for knowledge application for each MicroFIN task.

In summary, we used a more comprehensive MicroFIN version than in previous studies (i.e., Kretzschmar et al., 2016; Neubert, Kretzschmar, et al., 2015). Specifically, we extended the assessment of acquired knowledge by way of a further task format which focuses more on knowledge about system states rather than knowledge about interventions. However, the knowledge assessed by each of the different task formats was not disjunct (i.e., you also need some knowledge about system states in the initial state construction task). Furthermore, we used a partially different compilation of tasks, with one additional task (i.e., Wash-o-maton) compared to previous studies.

Reasoning.

Reasoning was assessed with three tests, each representing a different type of task content. The assessment of verbal reasoning (VR; 20 items) consisted of originally created verbal analogies tasks in a paper-and-pencil format. The test showed a good reliability of $\omega_h = .81$. Numerical reasoning (NR; 16 items) was assessed with originally created paper-and-pencil tasks where participants had to complete a number sequence or matrix with a missing field. The reliability in this sample was $\omega_h = .80$. Figural reasoning (FR; 20 items) was assessed with a shortened version of Raven's Standard Progressive Matrices (SPM; Raven, 1938). The computerized version with 20 items was developed on the basis of previous psychometric investigations of SPM (Georgiev, 2008). Reliability was $\omega_h = .82$.

Procedure

Data for this study were collected during a period of several months from October 2014 to April 2015. All assessments were done in group settings. At the first session (about 40 min), each participant completed MicroFIN individually in the classroom using a personal computer. At the second session (about 40 min), participants conducted the reasoning tests individually in a classroom setting. To assess the test-retest validity of MicroFIN, a second assessment after four to five months was conducted in a manner similar to the first session.

Data analysis

We used the R software (version 3.3.3; R Core Team, 2016) with the packages lavaan (version 05-20; Rosseel, 2012) and psych (version 1.6.12; Revelle, 2016). The data for the following analyses is publicly available via the Open Science Framework (OSF) and can be accessed at <https://osf.io/wp3z4>.

We examined the measurement models for CPS and reasoning by computing confirmatory factor analyses (CFA). As the MicroFIN items were ordinal, we used weighted least squares means and variance adjusted (WLSMV) estimation for the measurement models

based on single items (i.e., Models M1 to M4 in Figure 2). For all other models (i.e., Model M5 in Figure 2, measurement models for reasoning, and the final models for Hypotheses 3a and 3b), robust maximum likelihood (MLR) estimation was used. The measurement models for reasoning were based on three parcels calculated according to the item-to-construct principle (Little, Cunningham, Shahar, & Widaman, 2002) for each reasoning test (not presented in detail). The measurement models were identified by fixing the variance of the latent factors to 1.00. In the case of higher-order measurement models for reasoning, in which only two indicators of the second-order factor were available (for an example, see Figure 3), we constrained two factor loadings of one first-order factor to be equal to avoid empirically underidentified estimations (Brown, 2015). All other model parameters were estimated freely. All reported coefficients for CFA were based on completely standardized solutions.

Model fit was evaluated on the basis of standard fit indices and commonly accepted cutoff values (see Schermelleh-Engel, Moosbrugger, & Müller, 2003). In order to compare different measurement models (Hypothesis 1), we used the Satorra-Bentler scaled χ^2 difference test (Satorra & Bentler, 2010) and differences between TLI values ($\Delta\text{TLI} > .01$; Gignac, 2007). Reliability (Hypothesis 2) in terms of internal consistency was investigated with McDonald's omega (ω_n ; Zinbarg, Revelle, Yovel, & Li, 2005). Test-retest reliability was examined on the basis of the Pearson correlation between the sum scores of the two CPS measurements. The empirical relation between CPS and reasoning was investigated on the basis of bootstrapped coefficients (500 draws; Hypotheses 3a and 3b).

Some participants only took part in the first or in the second session because of organizational issues. Therefore, missing data occurred for max. 43 % of the participants in our sample (see Table 1). Little's (1988) test indicated support for the assumption of missing completely at random (MCAR) ($\chi^2 = 154.433$, $df = 140$, $p = .191$). To adjust for missing data, we used pairwise deletion for WLSMV estimation and the full information maximum likelihood (FIML) procedure for MLR estimation. The minimum sample size for the MicroFIN measurement models (Hypothesis 1) was $n = 253$, while the remaining models (Hypotheses 3a and 3b; see Table 2) had a minimum sample size of $n = 330$. Tests of significance ($\alpha = .05$) were two-tailed.

Results

Factorial validity

Descriptive statistics for the MicroFIN tasks and their intercorrelations are reported in Table 1. To examine Hypothesis 1, we first applied the 2-dimensional model with one latent factor each for knowledge acquisition and knowledge application (Model M1 in Figure 2), in accordance with Neubert et al. (2015). The measurement model showed a good fit (Model M1, Table 2). The correlation between the two latent factors was .99 (95% CI [.93, 1.06]), meaning that the two latent factors were empirically indistinguishable. A 1-dimensional model (i.e., all items of knowledge acquisition and knowledge

Table 1:
Manifest Correlations and Descriptive Statistics of MicroFIN Tasks and Reasoning Tests

Measure	MicroFIN						Reasoning								
	1)	2)	3)	4)	5)	6)	7)	8)	9)	10)	11)	12)	13)	14)	15)
MicroFIN															
1) MicroFIN: Task 1	-														
2) MicroFIN: Task 2	.51	-													
3) MicroFIN: Task 3	.48	.55	-												
4) MicroFIN: Task 4	.43	.51	.52	-											
5) MicroFIN: Task 5	.42	.47	.43	.38	-										
6) MicroFIN: Task 6	.38	.33	.37	.30	.34	-									
Reasoning															
7) Figural: Parcel 1	.29	.37	.34	.24	.29	.39	-								
8) Figural: Parcel 2	.39	.44	.49	.31	.37	.46	.62	-							
9) Figural: Parcel 3	.29	.42	.41	.29	.30	.42	.64	.64	-						
10) Numerical: Parcel 1	.21	.15	.25	.23	.21	.14	.35	.32	.27	-					
11) Numerical: Parcel 2	.19	.23	.20	.23	.12	.16	.34	.35	.33	.69	-				
12) Numerical: Parcel 3	.19	.13	.18	.21	.11	.23	.34	.29	.24	.56	.68	-			
13) Verbal: Parcel 1	.25	.28	.42	.25	.30	.22	.31	.36	.36	.30	.35	.28	-		
14) Verbal: Parcel 2	.21	.26	.30	.19	.28	.20	.24	.33	.31	.26	.28	.18	.64	-	
15) Verbal: Parcel 3	.23	.21	.34	.32	.25	.18	.25	.39	.33	.35	.34	.31	.61	.56	-
Mean	0.45	0.56	0.53	0.54	0.38	0.16	0.52	0.41	0.54	0.63	0.57	0.53	0.60	0.63	0.66
Standard deviation	0.24	0.27	0.28	0.24	0.22	0.18	0.23	0.28	0.24	0.28	0.23	0.28	0.22	0.24	0.24
<i>n</i>	253	253	252	252	250	247	239	239	239	228	228	228	208	208	208

Note. Based on Pearson's correlation coefficients; nonsignificant correlations (pairwise deletion) are written in italics.

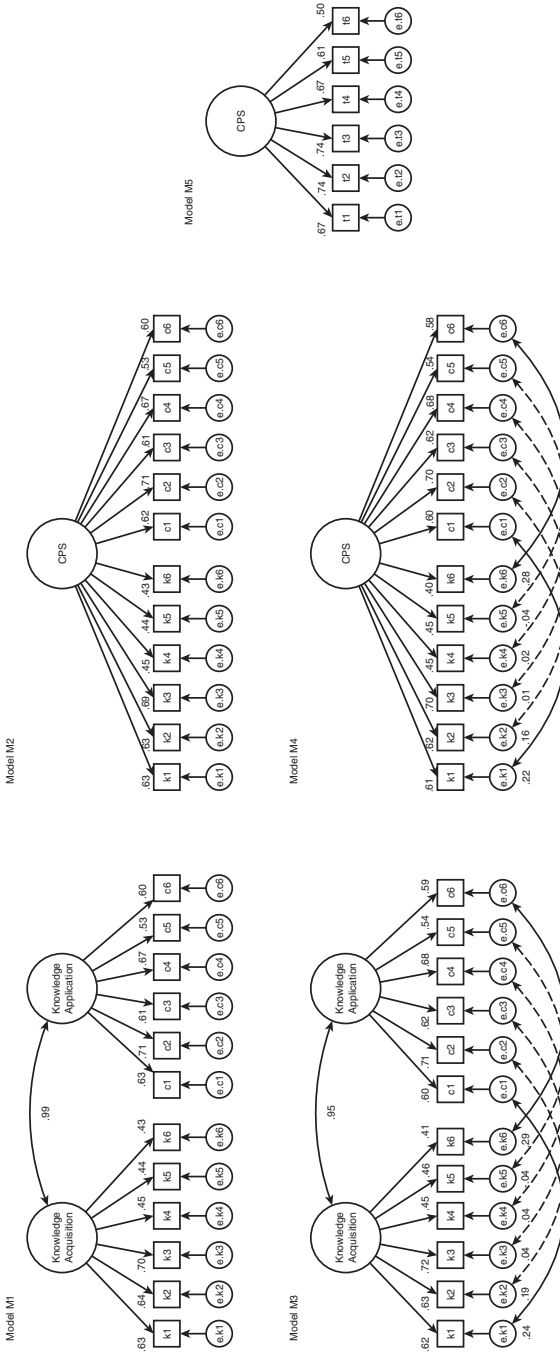


Figure 2:

Different measurement models for MicroFIN. Model M1 = 2-dimensional model without correlated errors; Model M2 = 1-dimensional model without correlated errors; Model M3 = 2-dimensional model with correlated errors; Model M4 = 1-dimensional model with correlated errors; Model M5: 1-dimensional model based on aggregated task scores; CPS = complex problem solving; k_i = knowledge acquisition for MicroFIN task; c_i = knowledge application for MicroFIN task; t_i = MicroFIN task; dashed lines = non-significant paths ($p > .05$)

application load on one latent factor; Model M2 in Figure 2) also showed a good fit (Model M2, Table 2). The difference test between the 2-dimensional model and the 1-dimensional model showed no significant difference: $\Delta\chi^2 = 0.041$, $df = 1$, $p = .839$; $\Delta TLI = .001$.

In order to investigate whether and to what extent the indicators shared a correlated uniqueness, we applied the 2-dimensional model as above and added paths for each pair of indicator errors belonging to the same MicroFIN task (Model M3 in Figure 2). The measurement model showed a very good fit (Model M3, Table 2). Two pairs of indicator errors showed a small significant correlation (i.e., Task 1 and Task 6), one pair showed a small but non-significant correlation (i.e., Task 2), and the other three pairs showed no substantial correlation (i.e., Tasks 3 to 5). The correlation between the two latent factors was .95 (95% CI [.88, 1.02]), again indicating unidimensionality. Consequently, a 1-dimensional model with correlated indicator errors (Model M4 in Figure 2) showed a very good fit (Model M4, Table 2). The difference test between these two models showed no significant difference: $\Delta\chi^2 = 1.737$, $df = 1$, $p = .188$; $\Delta TLI = .001$. We also compared the 1-dimensional model without correlated indicator errors (Model M2) to the 1-dimensional model with correlated indicator errors (Model M4) and found a significantly better fit for the latter: $\Delta\chi^2 = 13.918$, $df = 6$, $p = .030$; $\Delta TLI = .000$. The result was clearer when we omitted the non-significant error correlations (i.e., for Tasks 2 to 5): $\Delta\chi^2 = 12.079$, $df = 2$, $p = .002$; $\Delta TLI = .014$.

Finally, a 1-dimensional measurement model based on aggregated task scores in line with Kretzschmar et al. (2016) was investigated (Model M5 in Figure 2). The model showed a very good model fit, better than any of the other models (Model M5, Table 2).

In summary, a 2-dimensional model did not fit the data better than a 1-dimensional model, meaning that Hypothesis 1 was rejected. Furthermore, we found substantial correlations between some of the indicator errors. Therefore, correlated uniqueness plays a role in the commonly applied measurement models of MCS-based CPS tests. The 1-dimensional measurement model based on aggregated task scores was superior in terms of model fit and parsimony, and additionally avoids the issue of correlated uniqueness. Thus, it was used in all further analyses.

Reliability

The internal consistency of MicroFIN based on the 1-dimensional model (Model M5 in Figure 2) was $\omega_h = .82$. Thus, it was similar to the internal consistency of $\omega_h \approx .78$ reported by Neubert et al. (2015). Hypothesis 2 was therefore not rejected.

In order to examine test-retest reliability, we used a total sum score based on the 1-dimensional model ($\omega_{h_retest} = .90$). The correlation between the two measurement time points of MicroFIN was $r = .67$, $p = .000$.

Relation to reasoning

We analyzed seven different models in order to investigate the impact of different operationalizations of reasoning on the relation between reasoning and CPS. All reported correlations were statistically significant ($p < .01$). In the first three models, only one type of task content (i.e., figural, numerical, or verbal) with regard to reasoning was included. This means the corresponding g-factor (based on three parcels) only reflected figural reasoning, numerical reasoning, or verbal reasoning, respectively. The model fits (see Models R1–R3, Table 2) were very good. The correlations with CPS were .62 for figural reasoning (95% CI [.48, .76]), .38 for numerical reasoning (95% CI [.19, .54]), and .53 for verbal reasoning (95% CI [.35, .67]). In the next three models, combinations of two types of task content (i.e., figural-numerical, figural-verbal, numerical-verbal) were used. This means the g-factor (based on two latent first-order factors, which were each based on three parcels) reflected a combination of two types of task content. The model fits were very good (see Models R4–R6, Table 2). The correlations with CPS were .67 for figural-numerical reasoning (95% CI [.51, .85]), .78 for figural-verbal reasoning (95% CI [.61, .94]; see Figure 3), and .63 for numerical-verbal reasoning (95% CI [.47, .81]). In the last model, with very good model fit (see Model R7, Table 2), all three types of task content (i.e., figural, numerical, verbal) were included. The correlation between figural-numerical-verbal reasoning and CPS was .72 (95% CI [.58, .86]).

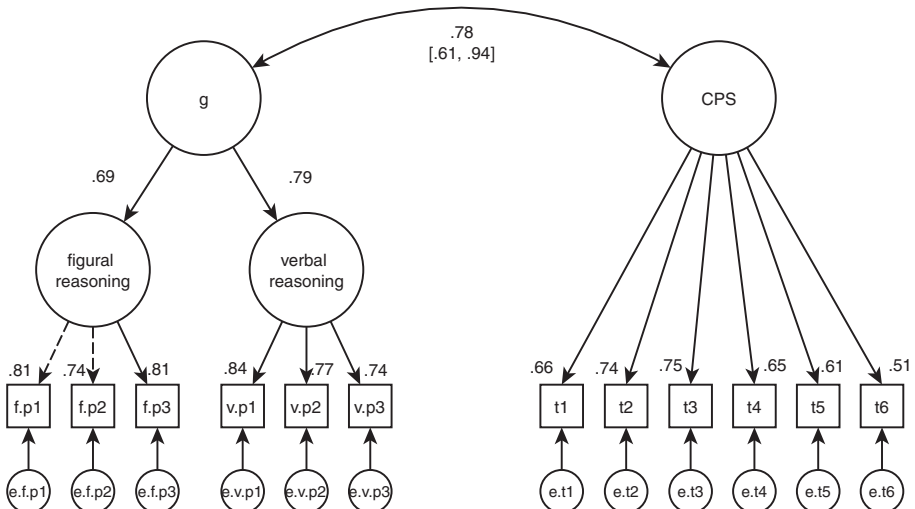


Figure 3:

Model R5 on the latent correlation between CPS measured by MicroFIN and figural-verbal reasoning; g = general reasoning factor; CPS = complex problem solving; p_i = parcels; t_i = MicroFIN task; dashed lines = unstandardized parameters were constrained to be equal; 95% confidence interval for the correlation between reasoning and CPS in brackets

Table 2: Goodness of Fit Indices for Different Measurement Models of MicroFIN and Operationalizations of Reasoning

Models	<i>n</i>	χ^2	<i>df</i>	<i>p</i>	CFI	TLI	RMSEA	[95%CI]	SRMR/ WRMR†
M1: MicroFIN 2-dimensional without correlated errors	253	70.664	53	.053	.986	.982	.036	[.000, .057]	.637 †
M2: MicroFIN 1-dimensional without correlated errors	253	71.037	54	.060	.986	.983	.035	[.000, .056]	.637 †
M3: MicroFIN 2-dimensional with correlated errors	253	50.290	47	.345	.997	.996	.017	[.000, .046]	.521 †
M4: MicroFIN 1-dimensional with correlated errors	253	52.233	48	.313	.997	.995	.019	[.000, .046]	.531 †
M5: MicroFIN 1-dimensional based on task scores	253	6.121	9	.728	1.000	1.012	.000	[.000, .005]	.018
R1: MicroFIN & figural reasoning	331	29.610	26	.284	.995	.993	.020	[.000, .049]	.044
R2: MicroFIN & numerical reasoning	330	22.498	26	.661	1.000	1.007	.000	[.000, .036]	.030
R3: MicroFIN & verbal reasoning	321	22.815	26	.643	1.000	1.007	.000	[.000, .037]	.036
R4: MicroFIN & figural-numerical reasoning	362	61.718	52	.168	.991	.988	.023	[.000, .042]	.052
R5: MicroFIN & figural-verbal reasoning	357	61.960	52	.162	.990	.987	.023	[.000, .043]	.058
R6: MicroFIN & numerical-verbal reasoning	332	48.286	52	.621	1.000	1.005	.000	[.000, .031]	.039
R7: MicroFIN & figural-numerical-verbal reasoning	362	92.397	86	.299	.995	.994	.014	[.000, .033]	.051

Note. *n* = sample size; *df* = degrees of freedom; CFI = Comparative Fit Index; TLI = Tucker Lewis Index; RMSEA = Root Mean Square Error of Approximation; SRMR = Standardized Root Mean Square Residual; WRMR = Weighted Root Mean Square Residual; χ^2 and *df* estimates are based on MLR and WLSMV, respectively.

In conclusion, our findings aligned with our expectations that figural-verbal reasoning would show a stronger association with CPS measured by MicroFIN than other reasoning operationalizations. Thus, Hypothesis 3a was not rejected. Furthermore, the correlation between reasoning and CPS was significantly different from $r = 1.00$ based on a 95% confidence interval. Therefore, Hypothesis 3b was not rejected either.

Discussion

The aim of the present study was to examine the psychometric properties of the MicroFIN test as the latest development in MCS-based CPS assessment tools. Specifically, we investigated the factorial validity and reliability as well as the correlation between MicroFIN and different operationalizations of reasoning.

Factorial Validity: Structure of MCS-Based Tests

Contrary to our expectations and the initial publication of MicroFIN (Neubert, Kretzschmar, et al., 2015), the commonly applied multidimensional model for MCS-based assessment tools did not fit the data better than a 1-dimensional model. This finding is in line with Kretzschmar et al.'s (2016) study, in which a 2-dimensional model for a somewhat different version of MicroFIN could not be confirmed either. As we used an extended version of MicroFIN (i.e., additional knowledge acquisition items for each task, a few different tasks compared to previous studies), we additionally checked whether a MicroFIN version more comparable to the version in Neubert et al. (2015; i.e., only the initial state construction task for knowledge acquisition) would alter the results, but it did not (i.e., the latent correlation between knowledge acquisition and knowledge application was .95 (95% CI [.88, 1.03])). Furthermore, additional analyses revealed that the assumption of independent indicators in the commonly applied measurement model for MCS-based tests was violated for MicroFIN. Taken together, as two of three studies could not confirm a 2-dimensional measurement model for MicroFIN, the alternative unidimensional model based on aggregated task scores seems to be the most appropriate measurement model for the current version of MicroFIN. Consequently, it has to be concluded that MicroFIN does not provide as fine-grained a measurement as MicroDYN (2 dimensions; Greiff et al., 2012) or Genetics Lab (3 dimensions; Sonnleitner et al., 2012), for which a variety of studies have consistently found evidence for multidimensional models. In the next section, we will discuss some possible explanations for the present finding, particularly in light of the differences between the various MCS-based assessment tools.

The first noticeable difference among the MicroDYN, Genetics Lab, and MicroFIN is the assessment of acquired knowledge. In MicroDYN and Genetics Lab, acquired knowledge is assessed on the basis of causal diagrams (e.g., Funke, 1985). Causal diagrams are a widely used format to assess structural knowledge about a problem, particularly for assessment tools based on the linear structural equation approach (Blech & Funke, 2006). However, a causal diagram is only one of many appropriate task formats,

each of which has specific features (for a summary of a few task formats, see Greiff, Fischer, et al., 2014). One specific characteristic of causal diagrams is that only a subset of knowledge (i.e., structural knowledge; “knowing that”) is assessed. This means that knowledge about how to control the system (e.g., which interventions should be applied in a specific situation; “knowing how”), and general problem-solving heuristics are neglected in assessments via causal diagrams. As these types of knowledge are also relevant for performance in the knowledge application phase (see e.g., Beckmann, 1994), it can be concluded that knowledge acquisition assessments via causal diagrams tap only part of the acquired knowledge that is necessary to solve a problem. Furthermore, it should be noted that even the assessment of structural knowledge is limited in MicroDYN and Genetics Lab. As Funke (1985) outlined, there are three different types of structural knowledge: knowledge about the existence of relations, knowledge about the directions of relations, and knowledge about the strength of relations. In MicroDYN, for example, only knowledge about the existence and directions of relations is assessed. Applying the Brunswik symmetry principle (Wittmann, 1988) to the measurement model of MicroDYN and Genetics Lab, it becomes clear that these two assessments exhibit a symmetry mismatch between the knowledge acquisition dimension (i.e., assessing performance rather narrowly with regard to one specific knowledge type) and the knowledge application dimension (i.e., assessing performance rather broadly based on different knowledge types). This asymmetry would prevent an empirical correlation of 1.00 between the two dimensions (i.e., unidimensionality), even in the case of a perfect true correlation (see Wittmann, 1988).

In MicroFIN, by contrast, the knowledge acquisition tasks aim to tap a broader range of knowledge (i.e., “knowing that”; “knowing how”) using different task formats. For example, in the identification tasks (see Figure 1a), the problem solver has to analyze whether a certain state of the problem situation is possible or not. In order to do this, the problem solver mainly has to know about different states of variables and their relations with each other (“knowing that”). In the initial state construction task (see Figure 1b), knowledge about interventions and rules (“knowing how”) is primarily assessed. Although knowledge about the specific problem is not exhaustively assessed in MicroFIN either (i.e., only a selection of the most important knowledge with regard to the specific control items in each task is assessed; see above), it seems that MicroFIN achieves better symmetry matching between the knowledge acquisition dimension and knowledge application dimensions. Therefore, a perfect correlation between these two types of performance (i.e., a unidimensional measurement model) is more likely to be observable (Wittmann, 1988) than in MicroDYN or Genetics Lab.

Another relevant issue with regard to the different task format is the manner of knowledge representation. For example, it is open to debate whether problem solvers have a causal diagram of acquired knowledge in their mind or use a different internal knowledge representation (see Kluwe & Haider, 1990; Tergan, 1989). In the latter case, a mental transformation of some sort between the internal knowledge representation and the representation via causal diagrams in the CPS assessment has to be applied. As this transformation requires additional cognitive processes and introduces noise in the form of transformation errors (Süß, 1996), assessment via causal diagrams might lead to spe-

cific method variance. It is also worth noting that the systematic variance inherent in the knowledge acquisition phase is less (or not at all) relevant for performance in the knowledge application phase.

In MicroFIN, by contrast, different task formats (i.e., identification task; initial state construction task) presented in a similar manner as in the knowledge application phase are used. Thus, problem solvers are not required to transform their internal knowledge representation into an abstract knowledge representation (e.g., causal diagrams). In fact, in both task formats in the knowledge acquisition phase (see Figures 1a and 1b), the presentation of the problem applied is almost identical to the presentation in the exploration or knowledge application phase (see Figure 1c). As a consequence, one could argue that less specific method variance is unilaterally introduced by the task format in the knowledge acquisition phase of MicroFIN, meaning that an observable correlation of 1.00 is again more likely (see Wittmann, 1988).

However, both explanations are of a speculative nature and have to be addressed in future research. In fact, the differences between task formats outlined above highlight the need to scrutinize the influence of different task types on the factorial validity of MCS-based assessment tools. For example, several different task formats could be integrated into the knowledge acquisition phase of different MCS-based CPS tests. Although not suitable for all developed tasks, it is possible to use causal diagrams to assess acquired knowledge in MicroFIN. Furthermore, the easy-to-implement state identification tasks can be integrated into MicroDYN or Genetics Lab. If the explanation of the present findings outlined above is applicable, a 2-dimensional model will be empirically confirmed for MicroFIN (i.e., only with causal diagrams), but a 1-dimensional model will be confirmed for MicroDYN/Genetics Lab (e.g., with state identification tasks).

The second issue we want to discuss is the impact of correlated uniqueness. Previous research (e.g., Brown, 2003; Marsh, 1996) has demonstrated that multidimensionality can arise simply from neglecting the relation between indicator errors of latent variables. Therefore, examining correlated uniqueness is recommended, particularly when the indicators share common features (e.g., two items rely on the same task). For MCS-based CPS tests, the commonly applied multidimensional measurement model did not reflect the possible mutual dependency of the indicators (see, e.g., Greiff et al., 2012; Neubert, Kretzschmar, et al., 2015), although it is reasonable to expect such a correlation (i.e., the control performance in the knowledge application phase depends on acquired knowledge about the task). For MicroFIN, we showed that correlated uniqueness is an issue for some of the indicator pairs. It is unknown whether and to what extent correlated uniqueness has an impact on the multidimensionality of the other MCS-based CPS tests. Therefore, we re-analyzed the data from Kretzschmar et al. (2014) with regard to the measurement model of MicroDYN for illustrative purposes. We extended the commonly applied 2-dimensional measurement model using correlated error pairs (similar to Model M3 in Figure 2). The model showed a good model fit ($\chi^2(125) = 303.737$, $p = .000$, CFI = .989, TLI = .986, RMSEA = .031 (CI 95% [.027; .036]), WRMR = 1.212), very similar to the original 2-dimensional measurement model for MicroDYN. We found correlations between indicator pairs ranging between -.15 and .36 (4 out of 9 correlations were greater than .20). Thus, it seems that correlated uniqueness is also important for MicroDYN,

although it has heretofore been assumed that the indicators are independent of each other (see, e.g., Greiff, Fischer, et al., 2014). Therefore, future research should investigate whether the multidimensionality of MicroDYN and Genetics Lab might be influenced by the not yet considered correlated uniqueness of their indicators.

In summary, we can conclude that commonly applied multidimensional measurement models for MCS-based CPS tests are not suitable for MicroFIN. In fact, an alternative unidimensional model based on aggregated task scores, avoiding the issue of correlated uniqueness, seems to be more appropriate. At the first glance, this can be considered a shortcoming of the MicroFIN test compared to the other MCS-based tests. If a fine-grained assessment of different CPS sub-processes is needed, MicroFIN should probably not be used in its current version. However, we also discussed issues that have an impact on the multidimensionality of MCS-based tests in general, and thus call for closer scrutiny of previous findings supporting the assumption of multidimensionality for MicroDYN and Genetics Lab. In this regard, Gignac and Kretzschmar (2017) recently demonstrated that the approach to investigating factorial validity (i.e., examining whether the latent correlation is significantly smaller than 1.00) most commonly applied in CPS research (e.g., Neubert, Kretzschmar, et al., 2015; Sonnleitner et al., 2013; Wüstenberg et al., 2012) might be highly misleading when used as a criterion for multidimensionality. Obviously, further research is needed to investigate the factorial validity of MCS-based CPS tests.

Reliability

With regard to the second research issue about the reliability of MicroFIN, the present findings provide evidence that a further developed MicroFIN version based on a different measurement model (see above) showed a similar internal consistency to the initial MicroFIN version (Neubert, Kretzschmar, et al., 2015). As high reliability is one of the main features of the MCS approach (Greiff et al., 2012), it was important to replicate previous findings with partly different tasks. It can be concluded that the relatively more heterogeneous MicroFIN test is a suitable extension of MCS-based CPS assessment tools with regard to reliability. Future research should examine how far task heterogeneity can be extended (e.g., considering dynamic changes; Scherer, 2015) before a substantial decrease in reliability is observed.

Furthermore, the study provides the very first insights into the test-retest reliability of MicroFIN and MCS-based assessment tools in general. Although the MCS approach has drawn praise for its high reliability, surprisingly, no studies investigating reliability estimations apart from internal consistency with regard to the MCS approach have yet been published. At first glance, the results of the pilot study indicate that a satisfactory test-retest reliability is achievable with MicroFIN. However, it should be strongly emphasized that the present findings are embedded in an exploratory context. Correlations based on a sample size of $n = 39$ (with selection biases) should be approached with skepticism. Due to the obvious limitations of this pilot study, the present findings regarding test-retest reliability should be considered at most a stimulus for further research.

Relation between CPS and reasoning

Our findings in terms of the last research issue about the empirical relation between MicroFIN and reasoning fit in with a variety of previous findings providing evidence of a strong relation between CPS skills and intelligence (see Stadler et al., 2015). We found latent correlations between .38 and .78, depending on the operationalization of reasoning. The wide range of correlations underscores the importance of the Brunswik symmetry principle (Wittmann, 1988) when examining the relation between these two constructs (e.g., Kretzschmar et al., 2016).

Specifically, the lowest correlations were found for reasoning operationalizations based on a single type of task content. The highest correlation was found for an operationalization of reasoning based on two different types of task content that matched the demands of the CPS operationalization MicroFIN (i.e., figural-verbal). However, a combination of all three task formats (i.e., figural, verbal, numerical), following the recommendations of a “good g” operationalization (e.g., Jensen & Wang, 1994), reduced the correlation. This is an interesting finding with several implications.

First, as most previous studies investigating the correlation between CPS and intelligence relied on a reasoning operationalization based on only one form of task content (e.g., Greiff et al., 2013; Neubert, Kretzschmar, et al., 2015; Wüstenberg et al., 2012), but recent CPS tests are based without exception on a combination of at least two types of task content (e.g., MicroDYN contains mainly numerical and figural stimuli), it can be concluded that these studies systematically underestimated the relation between CPS and intelligence. However, recent attempts to examine the relation between CPS and intelligence based on a very broad operationalization of intelligence (see Kretzschmar et al., 2016; Lotz et al., 2016) might have also underestimated the correlation. The key issue is to use a symmetrical operationalization of both constructs, leading to a fair examination of their relation (Wittmann, 1988). This might also explain why the correlation found in the present study is higher than the meta-analytically averaged correlation of .59 reported in Stadler et al. (2015). Although Stadler et al. (2015) controlled for the operationalization of intelligence (i.e., general intelligence vs. reasoning), they did not consider the symmetry match in terms of task content between intelligence and CPS operationalizations. This is even more important in the context of faceted models of intelligence (e.g., Berlin Intelligence Structure Model; Jäger, 1982; for a description in English, see Süß & Beauducel, 2015) in which the stimulus material (i.e., figural, numerical, verbal) defines specific (sub-)constructs of intelligence with specific cognitive demands (e.g., figural intelligence). According to these models and the findings of the present study, the question might not be whether CPS and reasoning are distinct constructs (Stadler et al., 2015) but whether CPS and specific sub-constructs of intelligence (e.g., numerical reasoning; figural-verbal reasoning) tap into different cognitive demands. Therefore, we highly recommend that future studies use a broad operationalization of reasoning (or intelligence in general) but also consider the Brunswik symmetry principle (Wittmann, 1988) in terms of the stimulus material for both CPS and reasoning in order to estimate a less biased relation.

Second, previous studies emphasized the independence of the CPS construct compared to established intelligence constructs, and reasoning in particular, based on a correlation significantly smaller than 1.00 (e.g., Greiff et al., 2013; Kretzschmar et al., 2016; Lotz et al., 2016; Sonnleitner et al., 2013; Wüstenberg et al., 2012). From a psychometric perspective, the present findings support this argumentation (but see Gignac & Kretzschmar, 2017). However, from a conceptual point of view, a correlation of .78 is usually interpreted as convergent validity with regard to different assessment tools for the same construct. For example, correlations between three MCS-based CPS tests ranging from .62 to .73 were interpreted as convergent validity (Greiff, Stadler, Sonnleitner, Wolff, & Martin, 2015). In this sense, there has been much speculation as to how CPS and intelligence might differ (see Kretzschmar et al., 2016). It is worth noting that the differentiation between CPS and intelligence constructs has been based mainly on the level of operationalization (e.g., interactive CPS tasks vs. static reasoning tasks) rather than the construct level (i.e., different cognitive processes). In fact, the theoretical overlap between these two constructs is substantial (for an overview, see e.g., Kretzschmar et al., 2016), which makes it difficult to identify an exclusive CPS cognition. Therefore, according to the strong correlation found in the present study, the implications of the mostly disregarded Brunswik symmetry principle for previous studies, and the rather vague theoretical foundation for an independent CPS construct, it might be wise to (re-) consider MicroFIN and CPS assessment tools in general as intelligence measurements instead of tests covering an independent construct (see e.g., Kröner et al., 2005). This conclusion was already drawn by Süß (1996, 1999) based on more comprehensive studies with more ecologically valid CPS tests (i.e., microworlds) and broad operationalizations of intelligence and knowledge. Of course, as long as the findings of the present study have not been replicated with a heterogeneous sample, different operationalizations of intelligence and CPS tests, and ideally using longitudinal study designs (see Süß, 1996) as well as multitrait-multimethod analyses (see Greiff et al., 2013), the question of whether CPS should be considered an independent ability construct cannot be conclusively answered.

Conclusion

We can conclude that MicroFIN, the latest development within the MCS framework, has comparable psychometric properties to more established MCS-based CPS tests such as MicroDYN or Genetics Lab. However, the present findings raise some important questions about the factorial validity of MicroFIN and CPS tests in general. In this sense, the present study extends the recent discussion about how to further improve CPS assessment tools (e.g., Funke, 2010, 2014; Greiff & Martin, 2014; Scherer, 2015; Schoppek & Fischer, 2015). Furthermore, there seems to be little evidence that MicroFIN covers a specific and independent cognitive construct. Therefore, it should be considered a modern operationalization of reasoning with several advantages, such as the possibility to investigate problem solving behavior via log file analyses (e.g., Greiff, Niepel, Scherer, & Martin, 2016; Müller et al., 2013). Ultimately, the present study emphasizes the importance of the Brunswik symmetry principle (Wittmann, 1988), which should be paid

much more attention in research about the validity of CPS assessment tools and psychological assessment in general.

Author note

This research was supported by the Postdoc Academy of the Hector Research Institute of Education Sciences and Psychology, Tübingen, funded by the Baden-Württemberg Ministry of Science, Education and the Arts. We are grateful to the TBA group at DIPF (<http://tba.dipf.de>) for providing the authoring tool CBA Item Builder and technical support.

References

- Beckmann, J. F. (1994). Lernen und komplexes Problemlösen: Ein Beitrag zur Konstruktvalidierung von Lerntests. [Learning and complex problem solving. A contribution to the validation of learning tests]. Bonn: Holos.
- 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>
- Blech, C., & Funke, J. (2006). Zur Reaktivität von Kausaldiagramm-Analysen beim komplexen Problemlösen [On the reactivity of causal diagrams in complex problem solving]. *Zeitschrift für Psychologie*, *214*(4), 185–195. <https://doi.org/10.1026/0044-3409.214.4.185>
- Brown, T. A. (2003). Confirmatory factor analysis of the Penn State Worry Questionnaire: Multiple factors or method effects? *Behaviour Research and Therapy*, *41*(12), 1411–1426. <https://doi.org/10.1111/1467-8624.00081>
- Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (Second Edition). New York : London: The Guilford Press.
- Buchner, A., & Funke, J. (1993). Finite-state automata: Dynamic task environments in problem-solving research. *The Quarterly Journal of Experimental Psychology Section A*, *46*(1), 83–118. <https://doi.org/10.1080/14640749308401068>
- Chen, Z., & Klahr, D. (1999). All other things being equal: Acquisition and transfer of the control of variables strategy. *Child Development*, *70*, 1098–1120.
- Dörner, D. (1986). Diagnostik der operativen Intelligenz [Assessment of operative intelligence]. *Diagnostica*, *32*, 290–208.
- Fischer, A., Greiff, S., & Funke, J. (2012). The process of solving complex problems. *The Journal of Problem Solving*, *4*(1), 19–41. <https://doi.org/10.7771/1932-6246.1118>
- 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. (Ed.). (1999). Themenheft “Komplexes Problemlösen” [Special Issue: Complex Problem Solving]. *Psychologische Rundschau*, *50*(4).
- Funke, J. (2010). Complex problem solving: A case for complex cognition? *Cognitive Processing*, *11*, 133–142. <https://doi.org/10.1007/s10339-009-0345-0>

- Funke, J. (2014). Analysis of minimal complex systems and complex problem solving require different forms of causal cognition. *Frontiers in Psychology*, 5. <https://doi.org/10.3389/fpsyg.2014.00739>
- Funke, J., Fischer, A., & Holt, V. D. (2017). When less is less: Solving multiple simple problems is not complex problem solving – a comment on Greiff et al. (2015). *Journal of Intelligence*, 5(1). <https://doi.org/10.3390/jintelligence5010005>
- Georgiev, N. (2008). Item analysis of C, D and E series from Raven's standard progressive matrices with item response theory two-parameter logistic model. *Europe's Journal of Psychology*, 4(3). <https://doi.org/10.5964/ejop.v4i3.431>
- Gignac, G. E. (2007). Multi-factor modeling in individual differences research: Some recommendations and suggestions. *Personality and Individual Differences*, 42(1), 37–48. <https://doi.org/10.1016/j.paid.2006.06.019>
- Gignac, G. E., & Kretzschmar, A. (2017). Evaluating dimensional distinctness with correlated-factor models: Limitations and suggestions. *Intelligence*, 62, 128–147. <https://doi.org/10.1016/j.intell.2017.04.001>
- Greiff, S., Fischer, A., Stadler, M., & Wüstenberg, S. (2014). 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., Kretzschmar, A., Müller, J. C., Spinath, B., & Martin, R. (2014). The computer-based assessment of complex problem solving and how it is influenced by students' information and communication technology literacy. *Journal of Educational Psychology*, 106(3), 666–680. <https://doi.org/10.1037/a0035426>
- Greiff, S., & Martin, R. (2014). What you see is what you (don't) get: A comment on Funke's (2014) opinion paper. *Frontiers in Psychology*, 5. <https://doi.org/10.3389/fpsyg.2014.01120>
- Greiff, S., Niepel, C., Scherer, R., & Martin, R. (2016). Understanding students' performance in a computer-based assessment of complex problem solving: An analysis of behavioral data from computer-generated log files. *Computers in Human Behavior*, 61, 36–46. <https://doi.org/10.1016/j.chb.2016.02.095>
- Greiff, S., Stadler, M., Sonnleitner, P., Wolff, C., & Martin, R. (2015). Sometimes less is more: Comparing the validity of complex problem solving measures. *Intelligence*, 50, 100–113. <https://doi.org/10.1016/j.intell.2015.02.007>
- 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>
- Grossler, A., Maier, F. H., & Milling, P. M. (2000). Enhancing learning capabilities by providing transparency in business simulators. *Simulation & Gaming*, 31(2), 257–278. <https://doi.org/10.1177/104687810003100209>
- Jäger, A. O. (1982). Mehrmodale Klassifikation von Intelligenzleistungen: Experimentell kontrollierte Weiterentwicklung eines deskriptiven Intelligenzstrukturmodells [Multimodal classifications of intelligence performance: Experimentally controlled development of a descriptive model of intelligence]. *Diagnostica*, 28, 195–225.
- Jensen, A. R., & Wang, L.-J. (1994). What is a good g? *Intelligence*, 18, 231–258. [https://doi.org/10.1016/0160-2896\(94\)90029-9](https://doi.org/10.1016/0160-2896(94)90029-9)

- Kluwe, R. H., & Haider, H. (1990). Modelle zur internen Repräsentation komplexer technischer Systeme [Models of internal representations of complex technical systems]. *Sprache & Kognition*, 9(4), 173–192.
- Kretzschmar, A. (2015). *Konstruktivität des komplexen Problemlösens unter besonderer Berücksichtigung moderner diagnostischer Ansätze [Construct validity of complex problem solving with particular focus on modern assessment approaches]* (Doctoral dissertation). University of Luxembourg, Luxembourg. Retrieved from <http://hdl.handle.net/10993/22584>
- Kretzschmar, A. (2017). Sometimes less is not enough: A commentary on Greiff et al. (2015). *Journal of Intelligence*, 5(4). <https://doi.org/10.3390/jintelligence5010004>
- Kretzschmar, A., Neubert, J. C., & Greiff, S. (2014). Komplexes Problemlösen, schulfachliche Kompetenzen und ihre Relation zu Schulnoten [Complex problem solving, school competencies and their relation to school grades]. *Zeitschrift für Pädagogische Psychologie*, 28(4), 205–215. <https://doi.org/10.1024/1010-0652/a000137>
- 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>
- Kretzschmar, A., & Süß, H.-M. (2015). A study on the training of complex problem solving competence. *Journal of Dynamic Decision Making*, 1, 4. <https://doi.org/10.11588/jddm.2015.1.15455>
- 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>
- Kubinger, K. D. (2009). Psychologische Computerdiagnostik [Computer-based psychological assessment]. *Zeitschrift für Psychiatrie, Psychologie und Psychotherapie*, 57(1), 23–32. <https://doi.org/10.1024/1661-4747.57.1.23>
- Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, 83(404), 1198–1202. <https://doi.org/10.1080/01621459.1988.10478722>
- 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*, 9(2), 151–173. https://doi.org/10.1207/S15328007SEM0902_1
- Loevinger, J. (1954). The attenuation paradox in test theory. *Psychological Bulletin*, 51(5), 493–504. <https://doi.org/10.1037/h0058543>
- Lotz, C., Sparfeldt, J. R., & Greiff, S. (2016). Complex problem solving in educational contexts – Still something beyond a “good g”? *Intelligence*. <https://doi.org/10.1016/j.intell.2016.09.001>
- Marsh, H. W. (1996). Positive and negative global self-esteem: A substantively meaningful distinction or artifacts? *Journal of Personality and Social Psychology*, 70(4), 810–819. <https://doi.org/10.1037/0022-3514.70.4.810>
- McGrew, K. S. (2009). CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. *Intelligence*, 37(1), 1–10. <https://doi.org/10.1016/j.intell.2008.08.004>
- Müller, J. C., Kretzschmar, A., & Greiff, S. (2013). Exploring exploration: Inquiries into exploration behavior in complex problem solving assessment. In S. K. D’Mello, R. A. Calvo, & A. Olney (Eds.), *Proceedings of the 6th International Conference on Educational Data Mining* (pp. 336–337).

- Neubert, J. C., Kretzschmar, A., Wüstenberg, S., & Greiff, S. (2015). Extending the assessment of complex problem solving to finite state automata: Embracing heterogeneity. *European Journal of Psychological Assessment, 31*(3), 181–194. <https://doi.org/10.1027/1015-5759/a000224>
- Neubert, J. C., Mainert, J., Kretzschmar, A., & Greiff, S. (2015). The assessment of 21st century skills in industrial and organizational psychology: Complex and collaborative problem solving. *Industrial and Organizational Psychology, 8*(02), 238–268. <https://doi.org/10.1017/iop.2015.14>
- OECD. (2014). *PISA 2012 results: Creative problem solving: Students' skills in tackling real-life problems (Volume 1)*. Paris: OECD Publishing.
- Putz-Osterloh, W. (1981). Über die Beziehung zwischen Testintelligenz und Problemlöseerfolg [On the relationship between test intelligence and success in problem solving]. *Zeitschrift für Psychologie, 189*(1), 79–100.
- R Core Team. (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <https://www.R-project.org/>
- Raven, J. C. (1938). *Progressive matrices: A perceptual test of intelligence*. London: HK Lewis.
- Revelle, W. (2016). *psych: Procedures for psychological, psychometric, and personality research*. Evanston, Illinois: Northwestern University.
- Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software, 48*(2), 1–36. <https://doi.org/10.18637/jss.v048.i02>
- Satorra, A., & Bentler, P. M. (2010). Ensuring positiveness of the scaled difference chi-square test statistic. *Psychometrika, 75*(2), 243–248. <https://doi.org/10.1007/s11336-009-9135-y>
- Scherer, R. (2015). Is it time for a new measurement approach? A closer look at the assessment of cognitive adaptability in complex problem solving. *Frontiers in Psychology, 6*. <https://doi.org/10.3389/fpsyg.2015.01664>
- Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online, 8*(2), 23–74.
- Schoppek, W., & Fischer, A. (2015). Complex problem solving – single ability or complex phenomenon? *Frontiers in Psychology, 6*(1669). <https://doi.org/10.3389/fpsyg.2015.01669>
- Sonnleitner, P., Brunner, M., Greiff, S., Funke, J., Keller, U., Martin, R., Hazotte, C., Mayer, H., & 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*(1), 54–72. <https://doi.org/10.1037/e578442014-045>
- Sonnleitner, P., Brunner, M., Keller, U., & Martin, R. (2014). Differential relations between facets of complex problem solving and students' immigration background. *Journal of Educational Psychology, 106*(3), 681–695. <https://doi.org/10.1037/a0035506>
- 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>

- 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]*. Göttingen: Hogrefe.
- Süß, H.-M. (1999). Intelligenz und komplexes Problemlösen: Perspektiven für eine Kooperation zwischen differentiell-psychometrischer und kognitionspsychologischer Forschung [Intelligence and complex problem solving: Perspectives for a cooperation between differential-psychometric and cognition-psychological research]. *Psychologische Rundschau*, 50(4), 220–228. <https://doi.org/10.1026//0033-3042.50.4.220>
- Süß, H.-M., & Beauducel, A. (2015). Modeling the construct validity of the Berlin Intelligence Structure Model. *Estudos de Psicologia (Campinas)*, 32(1), 13–25. <https://doi.org/10.1590/0103-166X2015000100002>
- Tergan, S.-O. (1989). Psychologische Grundlagen der Erfassung individueller Wissensrepräsentationen. Teil I: Grundlagen der Wissensmodellierung [Psychological foundations of assessing individual representations of knowledge. Part 1: Foundations of knowledge modeling]. *Sprache & Kognition*, 8(3), 152–165.
- Wagener, D. (2001). *Psychologische Diagnostik mit komplexen Szenarios - Taxonomie, Entwicklung, Evaluation [Psychological assessment with complex scenarios - taxonomy, development, evaluation]*. Lengerich: Pabst Science Publishers.
- Wittmann, W. W. (1988). Multivariate reliability theory: Principles of symmetry and successful validation strategies. In J. R. Nesselroade & R. B. Cattell (Eds.), *Handbook of multivariate experimental psychology* (2nd ed., pp. 505–560). New York: New York.
- 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>
- 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., 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*. <https://doi.org/10.1007/s10758-014-9222-8>
- Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α , Revelle's β , and McDonald's ω H: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika*, 70(1), 123–133. <https://doi.org/10.1007/s11336-003-0974-7>