Original Article

Assessing Complex Problem-Solving Skills in Under 20 Minutes

Florian Krieger[1], Matthias Stadler[2], Markus Bühner[2], Frank Fischer[2], and Samuel Greiff[1]

[1]Cognitive Science & Assessment, University of Luxembourg, Esch-sur-Alzette, Luxembourg
[2]Psychology and Education, Ludwig-Maximilians-University of Munich, Germany

Abstract. *Rationale:* Assessing complex problem-solving skills (CPS) is of great interest to many researchers. However, existing assessments require long testing times making them difficult to include in many studies and experiments. Here, we propose a specific composition of microworlds based on the MicroDYN approach, which allows for valid estimation of CPS in a substantially reduced amount of time (<20 min). *Methods:* Based on the reanalysis of a sample of $N = 232$ university students who worked on 11 microworlds of increasing difficulty, we conducted multiple confirmatory factor analyses to test all possible combinations of microworlds, which were theoretically justified in advance. *Results/Discussion:* We demonstrate one best fitting set with five microworlds, which shows excellent factorial validity and relates to both conventional measures of intelligence and to school grades. We hope that this will allow other researchers to include CPS into their study designs even when testing time is limited.

Keywords: complex problem-solving, MicroDYN, short scale, intelligence, validity

Being able to solve complex problems is an essential skill to succeed in the 21st century (Autor et al., 2003). Measuring complex problem-solving skills (CPS) has, therefore, been of great interest to psychologists for several decades (Schoppek et al., 2019). A recent culmination of this interest has been the assessment of CPS in the Program for International Student Assessment 2012, the arguably most well-known large-scale assessment worldwide (e.g., OECD, 2017). Following a widely adopted definition by Buchner (according to Frensch & Funke, 2014, p. 14), CPS is, throughout this paper, understood as "(...) the successful interaction with task environments that are dynamic (i.e., change as a function of the user's interventions and/or as a function of time) and in which some, if not all, of the environment's regularities can only be revealed by successful exploration and integration of the information gained in that process."[1]

Solving complex problems is closely associated with cognitive constructs, such as reasoning (Stadler et al., 2015; Wüstenberg et al., 2012) or working memory capacity (WMC; Meißner et al., 2016; Schweizer et al., 2013; Zech et al., 2017). In addition, it has been demonstrated to be significant in predicting job-related or educational success (Mainert et al., 2019; Sonnleitner et al., 2013).

Measuring CPS usually requires problem-solvers to interact with dynamic computer-simulated microworlds (Brehmer & Dörner, 1993). While there is a vast array of microworlds that were developed for this purpose (Stadler et al., 2015), most current research uses microworlds based on the MicroDYN approach (Greiff, Fischer et al., 2015). In this approach, problem-solvers need to interact with several microworlds that require determining the relation between multiple input and output variables. Studies usually use between 8 and 11 of these microworlds to assess CPS (Greiff, Fischer et al., 2015). This results in rather long testing times of more than 40 min, which may prevent many studies from including measures of CPS in their designs. In this paper, we suggest a set of microworlds that covers the whole theoretical range of microworlds based on the MicroDYN approach, provides empirically valid estimates of CPS scores, and reduces testing time to less than 20 min.

[1] Note that there are competing definitions of CPS (e.g., Dörner & Funke, 2017).

Psychological Test Adaptation and Development
https://doi.org/10.1027/2698-1866/a000009

What Is the Construct Being Measured?

First approaches to measuring CPS relied on intricate microworlds, including thousands of variables that were related to each other in ways that emulated real-world systems such as factories, cities, or even whole countries (Frensch & Funke, 2014). Due to the tremendous time and effort required by this approach, problem-solvers worked only on single microworlds, which made it challenging to estimate independent indicators of CPS. Moreover, performance in these microworlds was not related to performance in other microworlds beyond a shared measure of theoretically related constructs such as performance in more conventional measures of intelligence. Finally, the close emulation of real-world problems made it difficult to differentiate domain-general CPS from domain knowledge (Süß, 1996, 1999).

The multiple complex systems (MCS; Greiff, Fischer et al., 2015; see also Bühner et al., 2006; Zech et al., 2017) approach was suggested as a solution to those problems. In this approach, CPS is assessed as the performance in multiple smaller microworlds, rather than the performance in one single large microworld. This is made feasible by reducing both the number of variables and the complexity of their interrelations. The MicroDYN approach, as the most common operationalization of the MCS approach, describes the relation between variables by linear equations (Funke, 1993; Wüstenberg et al., 2012). Figure 1 graphically illustrates the underlying structure of three prototypical microworlds based on the MicroDYN approach. In the most complex example, three input variables (that can be manipulated by the problem-solver) are related to three output variables. Output variable *Z* affects itself, a so-called Eigendynamic (Funke, 1993), which may describe growth processes (such as interest rates) or decay processes (such as resource depletion) that are independent of the other variables. The relations are arbitrary and do not represent any real-world relations to minimize the impact of domain knowledge on the performance in these microworlds. However, to make the microworlds more appealing, they are usually semantically embedded in familiar scenarios. For instance, problem-solvers may need to take over the role of handball coaches investigating the impact of three types of training (generically labeled Training A–C) to three types of outcomes (e.g., exhaustion, speed, or power of the throw).

In line with the definition of CPS (see above), the problem-solving process is separated into two phases (Novick & Bassok, 2005). In the *knowledge acquisition phase*, problem-solvers determine the relations between input and output variables by systematically manipulating the input variables while registering the resulting changes in the output variables. The knowledge gained is then used to reach specific target values in the output variables (*knowledge application phase*). While the two phases are theoretically separate (Newell & Simon, 1972), they are empirically highly correlated and often aggregated into one CPS factor (Stadler et al., 2019; Wüstenberg et al., 2012).

The (a) number of variables and the (b) number and type of relations and (c) the number of Eigendynamics have been shown to fully describe the “microworlds” difficulty in both the knowledge acquisition phase and the knowledge application phase with Eigendynamics affecting the difficulty most strongly (Stadler et al., 2016). Correspondingly, most CPS measures based on the MicroDYN approach employ rather simple (few variables/relations) and complex (more variables/relations) microworlds without Eigendynamics, as well as microworlds with Eigendynamics to cover a broad range of difficulties (see Figure 1).

What Are the Intended Uses?

The intention of the current research is to provide a set of microworlds that allow for the estimation of a valid CPS score with low testing time. Since CPS assessment is traditionally time-consuming, we aim to communicate a short version, which allows an inclusion of CPS assessment in multiple research contexts (e.g., educational psychology, differential psychology, cognitive psychology, and artificial intelligence research).

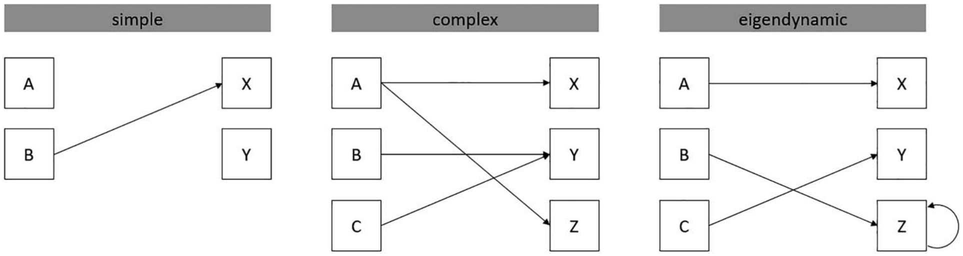

Figure 1. Example of structural equations underlying MicroDYN items: simple items, complex items, and items with eigendynamics.

What Is the Intended Target Population?

The proposed subset is not intended for individual assessment as the resulting scores will likely have lower reliability than scores based on the full set of microworlds (Sijtsma & Emons, 2011). However, we believe that a short measure of CPS will be valuable for researchers interested in using the construct in experiments or studies. Due to the sets' composition, we expect the difficulties to be rather high making the short measure most suitable for adult samples of higher educational backgrounds.

Development of the Short Scale

To develop the short scale without losing the sound psychometric properties of MicroDYN (see Kemper et al., 2019; Ziegler et al., 2014), we suggest reducing the assessment to five microworlds with a specific composition. To maintain a wide range of difficulties, the set of microworlds should include both a simple and a complex microworld without Eigendynamics as well as three microworlds with Eigendynamics. The composition was based on two considerations. First, as suggested by Stadler et al. (2016), complexity (i.e., number of variables and number of relations among them) and the existence of Eigendynamics are the most important features in defining the difficulty of CPS tasks based on the MicroDYN approach (see also Greiff, Fischer et al., 2015). Therefore, we included both tasks with high and low complexity as well as with and without Eigendynamics. The other consideration alludes to the separation of connectivity and dynamics in CPS tasks, as suggested by Stadler et al. (2019). To estimate these two dimensions as latent factors, both require at least three indicators. As all tasks based on the MicroDYN approach include connectivity, the minimal number of tasks that complies with both of these considerations consists of two tasks of high and low complexity, respectively, without Eigendynamics and three tasks with Eigendynamics.

Furthermore, we propose to reduce the assessment to the knowledge acquisition phase only. Previous studies showed that the knowledge application phase was either redundant to the knowledge acquisition phase (Stadler et al., 2019; Wüstenberg et al., 2012) or did not contribute much in predictive validity when explaining external criteria (Lotz et al., 2016).

To find the optimal set of established microworlds, we reanalyze an existing data set that includes performance data on 11 microworlds. The three simple microworlds, four complex microworlds, and four microworlds including Eigendynamics allow comparing a total of 48 subsets that fulfill the criteria described above. We estimate factorial validity and inspecting item characteristics for all these subsets. Subsequently, we estimate the construct validity of the best-fitting subset by comparing the resulting "scores" in relation with proximal variables such as the performance on conventional measures of intelligence (Kretzschmar et al., 2016; Lotz et al., 2016) and WMC (Schweizer et al., 2013). Finally, we estimate the subset's criterion validity by comparing their relation to school grades, an established criterion of CPS (Kretzschmar et al., 2016; Stadler et al., 2019), to the relation found for the full set.

Methods

Participants

Analyses were based on data of a larger screening study, in which CPS was assessed besides other cognitive and noncognitive variables. Two hundred and thirty-four students of social sciences (mainly psychology and educational sciences) from universities in Munich (204 females and 30 males) participated in this study. We excluded two participants due to missing data for CPS. Remaining participants ($N = 232$) were between 18 and 52 years old ($M = 22.25$, $SD = 3.86$). The study was part of their course requirements, and the data of the study served for practicing statistical analyses in the following course. After assessing cognitive ability, participants were assigned to conditions of an experiment investigating impacts of instructional design on learning outcomes (for a full description, see Bichler et al., 2020). Participants received either €55 (~$60) or a certificate of participation in an empirical study after completing the study.

To estimate the optimal sample size, we referred to results of Monte Carlo simulations recommending a sample size of >200–500 (for an overview, see Kyriazos, 2018) for confirmatory factor analysis (CFA). Thus, we used more than 200 participants for our analyses.

Measures

Complex Problem-Solving

We measured CPS with the standard form of Complex Problem Solving Test (COMPRO) (Greiff & Wüstenberg, 2015). This test contains 11 microworlds (for structural equations, please see Appendix A), each differentiating between the knowledge acquisition and the knowledge application phase. Each microworld first describes a scenario embedded in a context belonging to the business sector. Afterward, the participant interacts with a system including both input and output variables connected through a structural equation system. The participant is instructed to manipulate the input variables by using slide controls and observe the effect on the dependent variables. The task of the participant is to identify the associations among input and output variables by systematically working with the input variables and draw the associations in a diagram. Participants have 3 minutes for this first part of the task. The performance in the first part of the microworld is an indicator of knowledge acquisition in complex situations (Greiff & Wüstenberg, 2015).

After a short transition, the participant works on the second part of the task, in which they apply the derived knowledge about the system of the first phase to reach specific, predefined goals. As described, we only focus on the first phase in the current study.

Note that COMPRO is one instance of the MicroDYN approach and thus relies on the same structures as other tests based on MicroDYN, which are widely employed in the literature. In detail, COMPRO consists of 11 items whose structure is based on the MicroDYN approach used in other studies (e.g., Greiff et al., 2012; see also Greiff & Wüstenberg, 2015). The specialty about COMPRO is that all cover stories of its items are from a business context. However, since the tasks from the MicroDYN approach are designed to not require any specific prior knowledge of the respective context, tasks are designed in that manner that their cover stories are fictitious and do not interfere with the solving process (see also Greiff et al., 2012).

Intelligence

Fluid and crystallized intelligence were measured using the computerized and adaptive intelligence structure battery (INSBAT; Hornke et al., 2004). Both facets of intelligence were operationalized with three subtests each. For each subtest, the number of tasks depended on the performance of the participants (adaptive testing). The results of all subtests for each facet of intelligence were transformed into raw scores for fluid and crystallized intelligence. All subtests are Rasch homogeneous and reliability of each subtest was $\alpha = .70$.

Fluid intelligence (gF) was measured by (a) numerical inductive reasoning, (b) figural inductive reasoning, and (c) verbal deductive reasoning. The task of participants in the numerical inductive reasoning subtest is to identify the rule underlying a series of numbers and complete the series. In the figural inductive reasoning subtest, participants see a 3 × 3 matrix with one empty field. Afterward, they must identify a rule for the symbols in the remaining eight fields. Then, the participants must choose the correct symbol out of eight possible ones to complete the matrix. In the deductive reasoning subtest, participants receive two given statements and have 45 s to draw a conclusion from these statements based on five possible answers.

Crystallized intelligence (gC) was assessed with the subtests (a) general knowledge, (b) word meaning, and (c) verbal fluency. In the general knowledge subtests, participants receive gap texts containing definitions for terms. Participants must complete these texts so that the definition is right by choosing among several clauses. In the word meaning subtest, participants initially receive a single term. Then, they must choose among four possible terms, the one which has the most similar meaning to the initially presented term. In the verbal fluency subtest, participants must arrange a series of letters to form a sensible noun. For this purpose, participants must click at the letters in the right sequence.

Working Memory Capacity

We used the shortened versions of the automated complex span tasks operation, reading, and symmetry span (Oswald et al., 2015) within the software E-Prime (version 2.0.10.356) as measures to assess WMC. The three complex span tasks consist of two phases and differ only concerning the stimulus material. In the first phase, participants have to initially indicate whether a simple mathematical equation is right or wrong (operation span task), a sentence makes sense or not (reading span task), or a pattern in a 8 × 8 matrix is symmetric according to the vertical axis (symmetry span task). Afterward, participants have to memorize a letter (operation and reading span task) or the position of a red square in a 4 × 4 matrix (symmetry span task). After a variable sequence alternating between processing and storing information, participants have to recall the letters (operation span and reading span task) the pattern of red squares (symmetry span task) in the right order (second phase). The dependent variable for all complex span tasks was the proportion of correctly recalled elements (see Conway et al., 2005, for different scoring methods). The reliability of WMC based on these three measures was $\omega = .65$.

Grade Point Average

We used participants' self-reports on their grade point average (GPA). GPA represented the German 6-point grading scale, on which 1 represents the best and 6 the worst grade. For the sake of consistency with the previous literature, we transformed GPA in the sense that higher values represent a higher GPA in the current article (i.e., multiplied GPA by −1).

Statistical Analyses

Item Selection

As described above, the aim was to select five microworlds from the CPS long measure by considering both theoretical assumptions on item composition and psychometric properties of the items. To this end, we applied *three steps*, including a combination of a conceptual approach for item selection and a multi-indicator approach for evaluating item-scale fit, which goes beyond a mere inspection of fit values in factor analyses (see Figure 2).

In Step 1, we predefined a *theoretical criterion* that one item should be included with a simple structure, one with a complex structure, and three items including eigendynamics. The applied long measure assesses CPS with 11 items, including three items with simple structure (set SIM), four items with a complex structure (set COM), and four items with a complex structure with Eigendynamics (set ED). To select five items according to the theoretical criterion, one item out of set SIM, one item out of set COM, and three items out of set ED have to be drawn, resulting in $\binom{3}{1} \times \binom{4}{1} \times \binom{4}{3} = 48$ models.

In Step 2, we conducted 48 CFA (one for each model) with a one-factor solution and inspected the respective model fit. Specifically, we focused on the χ^2 statistic, the respective *p* value, CFI, SRMR, and RMSEA. We adhered to the following conventions to assess the global fit of the model besides a nonsignificant χ^2 statistic: RMSEA < .06, SRMR < .09, and CFI > .96 (see combination rules by Hu & Bentler, 1999).

Since there are reasonable considerations that the mere focus on fit indices can lead to a misleading model acceptance when factor loadings are low or false model rejection when factor loadings are high (e.g., Greiff & Heene, 2017; Heene et al., 2011), we applied further criteria in Step 3. To avoid a misleading model acceptance due to low factor loadings, we only included models in this step, which have a minimum factor loading of $\lambda = .50$ (i.e., *medium* factor loading; e.g., see Heene et al., 2011) of one of the five microworlds. This is backed-up by the criterion to only include models with an internal consistency of McDonald's ω >.70. In addition, only models were included, for which the reliability does not become substantially better when one of the five microworlds would have been dropped to allow that only microworlds are included, which contribute to an at least acceptable reliability. For this, we defined that the difference between ω for the scale with all five microworlds and ω for the remaining four microworlds when one item was dropped should be around 0 or positive (i.e., diff-$\omega \geq -.01$).

Out of the 48 models, we selected the models with the best properties regarding the described criteria. Validity analyses were conducted using the Pearson-moment correlation. The chosen approach as a combination of top-down definition of task characteristics and a data-driven task selection is increasingly becoming best

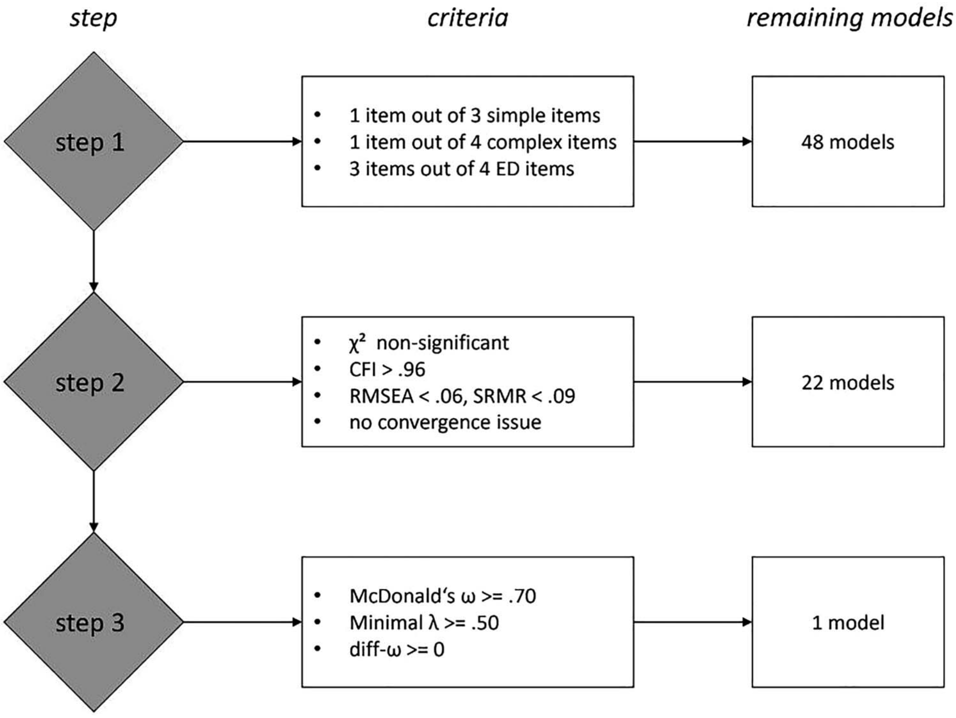

Figure 2. Applied three-step scheme for short-measure construction.

practice in the development of short scales from existing longer ones (e.g., Schroeders et al. 2016). In this case, where there is a relatively small number (k = 48) of permutations (i.e., sets of tasks with the predefined characteristics), we can estimate all models in a reasonable time. More complex scenarios use more sophisticated optimization algorithms to limit the number of models to be estimated (e.g., Leite et al., 2008).

Analysis

Analyses were performed using both *Python* (create combinations, CFA syntaxes, and plots) and *R* (statistical analyses). Specifically, for statistical analyses, the *R* packages *lavaan* (Rosseel, 2012) and *jmv* (The jamovi project, 2020) were used. For CFA, WLSMV was used as an estimator since microworld scores were dichotomous. To compare the magnitude of dependent correlations, we used analyses for paired correlation using *paired.r* from *psych*-package in *R*.

All analyses, the used data set, task characteristics of all microworlds, and output of all 48 model comparisons can be retrieved at the open science framework repository (OSF; https://osf.io/5grmj).

Results

Three Steps for Item Selection

After defining the models based on theoretical criteria (Step 1), 48 CFA models were conducted in Step 2. Four models (Models 2, 14, 18, and 30) did not converge, potentially due to large collinearity between items, and were not further considered. The remaining models revealed 44 values for χ^2 statistic, respective p values, CFI, SRMR, and RMSEA (see also Appendix B). The minimum for χ^2 statistic was 1.769; the maximum 50.66, p values ranged from < .001 to .88; for CFI, the minimum was .95 and the maximum was 1.00; for RMSEA, the minimum was .00 and the maximum was .20; and for SRMR, the minimum was .03 and the maximum was .15. Based on the defined criteria for model fit, 22 models were further excluded in Step 2, resulting in 22 remaining models.

In Step 3, item analysis was performed for all remaining models, resulting in 22 values for each indicator for item analysis of the respective short measure (internal consistency McDonald's ω, diff-ω, and factor loadings). Values for McDonald's ω ranged from .69 to .78, values for diff-ω ranged from −.04 to 0, and values for minimal factor loading of one of the microworlds for each model ranged from .35 to .64. Applying the defined criteria for Step 3, Model 43 remained.

Figure 3 displays the distributions of fit indices for all 44 models. Models are categorized as acceptable models according to Step 2 and nonacceptable models according to Step 2. The winning Model 43 is displayed at first position on the x-axis.

Evaluation of the Winning Model

Table 1 displays the structural properties of Model 43. A one-factorial solution for Model 43 is plausible with $\chi^2(5)$ = 8.372, p = .14, CFI = 1.00, RMSEA = .054,

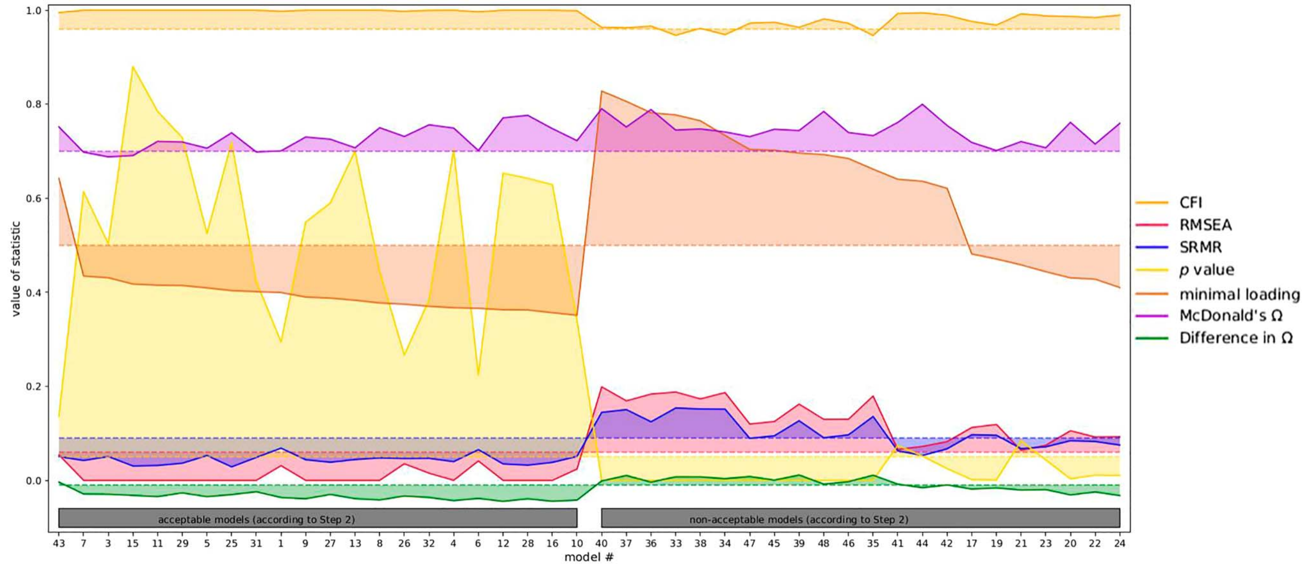

Figure 3. Properties of short measure (y) plotted against model number (x). The winning model is displayed at first position at the x-axis. Values are sorted by model fit (see the criteria above).

Table 1. Item characteristics and item analyses of Model 43

Item	Input	Output	Relations	ED	*M*	*SD*	λ	ω if dropped
Task 3 (SIM)	2	2	2	0	0.80	0.40	.64	.756
Task 9 (COM)	3	3	4	0	0.74	0.44	.82	.717
Task 7 (ED)	3	2	1	1	0.22	0.42	.72	.737
Task 10 (ED)	3	3	3	1	0.42	0.50	.94	.653
Task 11 (ED)	3	3	4	1	0.43	0.50	.87	.673

Note. COM = complex, ED = eigendynamic, SIM = simple, λ = factor loading.

SRMR = .05. The short measure has a mean difficulty of 0.52 (*SD* = 0.32), an internal consistency of ω = .75, a diff-ω of −.004, and a minimum factor loading of one of the items of .64. The results of item analyses for Model 43 are displayed in Table 1. The correlation between the sum score of long measure and the sum score of short measure based on Model 43 was *r* = .93, 95% CI [.91, .95], indicating strong similarity between both short and long measures.

To further determine internal validity, we also replicated the proposed two-factor model by Stadler et al. (2019), in which a second, orthogonal factor for Eigendynamics (with loadings constrained to equality) has been proposed. Our model received an excellent fit ($\chi^2[4]$ = 2.89, *p* = .58, CFI = 1.00, RMSEA = .00, SRMR = .03), replicating the model with the short version. In addition, we tested whether the use of the vary-one-thing-at-a-time strategy (VOTAT) is a strong predictor of performance since several studies on CPS have demonstrated its positive impact on solving CPS tasks (Greiff, Wüstenberg & Avvisati, 2015; Lotz et al., 2017; Wu & Molnár, 2021; Wüstenberg et al., 2014). We could replicate that successful VOTAT usage is a strong predictor of CPS performance for both long (*r* = .76, 95% CI [.70, .81]) and short version (*r* = .60, 95% CI [.51, .68]). Although the correlation of VOTAT was smaller for the short version than for the long version (*t* = 10.99, *p* < .001), it is still a strong relation replicating results from previous fields. For instance, Kröner et al. (2005) report an effect size of *r* = .47 of VOTAT usage (labeled "rule identification") and overall performance, or Greiff et al. (2016) report an effect size of β = .55.

To compare the two forms' construct validity, we first correlated the sum scores of the long measure with WMC, *gF*, and *gC*. The results revealed that the long measure is moderately related to WMC (*r* = .21, 95% CI [.09, .33]), which is comparable to literature reporting correlations between CPS and WMC (Greiff, Wüstenberg, Goetz et al., 2015; Meißner et al., 2016; Zech et al., 2017) or other measures of intelligence with similar WMC estimates on a manifest level (Krieger et al., 2019; Unsworth & Engle, 2005). Also, the long measure was substantially related to both *gF* (*r* = .41, 95% CI [.30, .51]) and *gC* (*r* = .38, 95% CI [.26, .48]), which is also in line with previous literature (e.g., Stadler et al., 2015). More importantly, the short measure based on Model 43 is related with the respective construct to a similar extent (WMC, *r* = .23, 95% CI [.10, .34]; *gF*, *r* = .43, 95% CI [.32, .53]; *gC* = 0.40, 95% CI [.28, .50]), indicating no decrease in validity from long to short measure. Regarding predictive validity, both long measure and short measure showed a moderate positive relation with GPA (for the long measure, *r* = .27, 95% CI [.15, .38]; for the short measure, *r* = .30, 95% CI [.18, .41]). All comparisons and statistical verifications are displayed in Table 2 besides descriptives of the respective measures.

Table 2. Descriptives and correlations of measures and short form of winning Model 43

	Descriptives		Intercorrelations						Difference long–short version	
	M	*SD*	Long	Short	WMC	*gF*	*gC*	GPA	|*t*|	*p*
Long	0.65	0.26								
Short	0.52	0.32	0.93							
WMC	0.70	0.14	0.21	0.23					0.46	.56
gF	0.21	0.89	0.41	0.43	0.37				0.75	.45
gC	−0.13	0.56	0.38	0.40	0.28	0.48			1.03	.30
GPA	−1.88	0.62	0.27	0.30	0.06	0.14	0.08		1.36	.18
VOTAT	8.91	3.35	0.76	0.60	0.09	0.29	0.25	0.13	10.99	<.001

Note. Mean score was used for both long and short versions; for WMC mean proportion correct of measures, OSpan, SSpan, and RSpan were used; for GPA, values were transformed and high values represent high GPA; statistical verification for differences of correlations for long or short version with respective measure is displayed in "difference long-short version." GPA = grade point average; WMC = working memory capacity.

Discussion

This paper aimed to propose a set of microworlds based on the microworlds that allow for the estimation of a valid CPS score with short testing time. In line with other measures of cognitive ability, CPS scores have consistently proved themselves valid predictors of both academic (Stadler et al., 2018; Stadler et al., 2019) and professional success (Mainert et al., 2019). Moreover, CPS scores provide a valid approximation of general intelligence (Lotz et al., 2016). A short yet valid measure of CPS may, thus, enrich future studies in various fields that would, otherwise, not have had the opportunity to include measures of CPS. To find a reliable and valid set of microworlds to be used in a short measure, we defined several sets of five microworlds composed of one simple microworld, one complex microworld, and three microworlds with Eigendynamics that showed excellent indicators of both factorial validity and item analyses. We propose choosing one set that clearly showed the best properties using multi-indicators for model fit and item analysis. CPS scores based on this set showed high correlations with CPS scores based on 11 microworlds. Furthermore, there were no substantial differences in the relation between performance in a conventional measure of intelligence and CPS scores based on the five microworld subset as compared to the full set of 11 microworlds. Finally, CPS scores based on the five-microworlds subset are meaningfully related to school grades, an established criterion of CPS, as strongly as CPS scores based on the full set.

Limitations

Some limitations need to be considered in interpreting these results. Most importantly, we reanalyzed an existing data set including data on 11 microworlds, rather than gathering a new validation sample with participants only working on five microworlds. This allowed us to compare the relations found directly between the two sets of microworlds but can only be considered the first step in the scale's validation as the presence of other tasks may affect performance in the individual microworlds. There is a slight increase in problem-solving performance with increased familiarity with the microworlds (Lotz et al., 2017) that could alter the tasks' difficulties as observed in the long version. In other words, the reduced opportunities to familiarize with the tasks might increase item difficulties for complex microworlds or microworlds with Eigendynamics in the shorter version. Moreover, instructions for COMPRO, as with most established implementations of MicroDYN tasks, are split into two parts. The first part of the instruction, which is provided at the onset of the assessment, provides a general introduction without introducing Eigendynamics. Eigendynamics are then introduced separately in the second part of the introduction, right before the first microworld with Eigendynamics. This separation seems unnecessary for a short measure with only five microworlds, and both instructions should be provided at the onset of the assessment. Taken together, this strongly suggests a follow-up study assessing data based on only the items of the short version proposed here. That means including only the knowledge acquisition phase and providing a combined instruction. However, it is promising that a study using a set of six microworlds (Kretzschmar et al., 2016) found results that were similar to ours, leading to the assumption that the direct application of the subset of five microworlds will reveal similar properties. Therefore, we hope that this paper will instigate researchers to use the scale in their research but urge them to keep this limitation in mind.

In addition, limiting our assessment of CPS on only the knowledge acquisition phase, it ignores the knowledge application phase, which is a theoretically separate aspect of CPS. Tasks based on the MicroDYN approach attempt to measure the two processes independently in two phases by providing the full solution to the knowledge acquisition phase (i.e., the relation between input and output variables) in the knowledge application phase. Empirically, however, very high correlations (>.85) are reported between knowledge acquisition and knowledge application (e.g., Gnaldi et al., 2020; Wüstenberg et al., 2012). Other publications even regressed the two factors onto a single higher-order factor of CPS (e.g., Greiff & Fischer, 2013; Kretzschmar et al., 2014) with knowledge acquisition consistently showing the stronger loadings. Thus, while our short measure does not capture the construct of CPS in its entirety, the measure represents a good estimate of CPS.

Finally, we would advise against interpreting individual CPS scores based on the proposed subset of microworlds. The microworlds' internal consistency was acceptable but still indicated more measurement error than is desirable for individual assessment (Greiff et al., 2012). This is despite our rather homogenous sample that may have slightly increased internal consistencies. In addition, gender was not balanced in our sample, and thus, the homogeneous sample during construction needs to be considered depending on the purpose of the use of the short form. However, we can assume that composition of sample has no effect on the reported measurement models in this article as it has been shown that CPS shows strong measurement invariance across gender (Wüstenberg et al., 2012).

Implications and Practical Use

Our results support the use of CPS scores based on the proposed composition of microworlds in research contexts such as cognitive psychology, the learning sciences, or even clinical psychology. While this study used microworlds in the COMPRO design, the results should be generalizable to all realizations of the MicroDYN approach such as the conventional design (e.g., Wüstenberg et al., 2012) or the Genetics Lab (Hazotte et al., 2011; Sonnleitner et al., 2012). For a study relating performance in different realizations of tasks based on the MicroDYN approach, see Greiff, Stadler et al. (2015).

In conclusion, we hope that the possibility to assess CPS in a short testing time will instigate more researchers to include measures of CPS into their studies and experiments. This is necessary to increase our understanding of the construct and help people succeed in the 21st century. The items of the proposed short version of this article were developed with the CBA Item Builder (Rölke, 2012). A software for displaying items and collecting data can be found on OSF (https://osf.io/5grmj).

References

Autor, D. H., Levy, F., & Murnane, R. J. (2003). The skill content of recent technological change: An empirical exploration. *The Quarterly Journal of Economics*, *118*(4), 1279–1333. https://doi.org/10.1162/003355303322552801

Bichler, S., Schwaighofer, M., Stadler, M., Bühner, M., Greiff, S., & Fischer, F. (2020). How working memory capacity and shifting matter for learning with worked examples-A replication study. *Journal of Educational Psychology*, *112*(7), 1320–1337. https://doi.org/10.1037/edu0000433

Brehmer, B., & Dörner, D. (1993). Experiments with computer-simulated microworlds: Escaping both the narrow straits of the laboratory and the deep blue sea of the field study. *Computers in Human Behavior*, *9*(2–3), 171–184. https://doi.org/10.1016/0747-5632(93)90005-D

Bühner, M., Krumm, S., Ziegler, M., & Pluecken, T. (2006). Cognitive abilities and their interplay. *Journal of Individual Differences*, *27*(2), 57–72. https://doi.org/10.1027/1614-0001.27.2.57

Conway, A. R. A., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review*, *12*(5), 769–786. https://doi.org/10.3758/bf03196772

Dörner, D., & Funke, J. (2017). Complex problem solving: What it is and what it is not. *Frontiers in Psychology, 8,*1153. https://doi.org/https://doi.org/10.3389/fpsyg.2017.01153.

Frensch, P. A., & Funke, J. (2014). *Complex problem solving: The European perspective*. Psychology Press.

Funke, J. (1993). Chapter 14 microworlds based on linear equation systems: A new approach to complex problem solving and experimental results. In G. Strube & K. F. Wender (Eds.), *The cognitive psychology of knowledge* (*Vol. 101*, pp. 313–330). https://doi.org/10.1016/S0166-4115(08)62663-1

Gnaldi, M., Bacci, S., Kunze, T., & Greiff, S. (2020). Students' complex problem solving profiles. *Psychometrika*, *85*(2), 469–501. https://doi.org/10.1007/s11336-020-09709-2

Greiff, S., & Fischer, A. (2013). Der Nutzen einer komplexen Problemlösekompetenz: Theoretische Überlegungen und empirische Befunde [The benefits of complex problem-solving skills: Theoretical considerations and empirical findings]. *Zeitschrift* Für *Pädagogische Psychologie*, *27*(1–2), 27–39. https://doi.org/10.1024/1010-0652/a000086

Greiff, S., Fischer, A., Stadler, M., & Wüstenberg, S. (2015). Assessing complex problem-solving skills with multiple complex systems. *Thinking & Reasoning*, *21*(3), 356–382. https://doi.org/10.1080/13546783.2014.989263

Greiff, S., & Heene, M. (2017). Why psychological assessment needs to start worrying about model fit. *European Journal of Psychological Assessment*, *33*(5), 313–317. https://doi.org/10.1027/1015-5759/a000450

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

Greiff, S., Wüstenberg, S., & Funke, J. (2012). Dynamic problem solving: A new assessment perspective. *Applied Psychological Measurement*, *36*(3), 189–213. https://doi.org/10.1177/0146621612439620

Greiff, S., Wüstenberg, S., Goetz, T., Vainikainen, M.-P., Hautamäki, J., & Bornstein, M. H. (2015). A longitudinal study of higher-order thinking skills: Working memory and fluid reasoning in childhood enhance complex problem solving in adolescence. *Frontiers in Psychology*, *6*, 1060. https://doi.org/10.3389/fpsyg.2015.01060

Greiff, S., & Wüstenberg, S. (2015). *Komplexer Problemlösetest COMPRO* [Complex problem-solving test COMPRO]. Schuhfried.

Hazotte, C., Mayer, H., Djaghloul, Y., Latour, T., Sonnleitner, P., Brunner, M., Keller, U., Francois, E., & Martin, R. (2011). The genetics lab: An innovative tool for assessment of intelligence by mean of complex problem solving. In A. Abd Manaf, S. Sahibuddin, R. Ahmad, S. Mohd Daud, & E. El-Qawasmeh (Eds.), *Informatics engineering and information science* (*Vol. 254*, pp. 296–310). Springer. https://doi.org/10.1007/978-3-642-25483-3_24s

Heene, M., Hilbert, S., Draxler, C., Ziegler, M., & Bühner, M. (2011). Masking misfit in confirmatory factor analysis by increasing unique variances: A cautionary note on the usefulness of cutoff values of fit indices. *Psychological Methods*, *16*(3), 319–336. https://doi.org/10.1037/a0024917

Hornke, L., Arendasy, M., Sommer, M., Häusler, J., Wagner-Menghin, M., Gittler, G., Bognar, B., & Wenzl, M. (2004). *Intelligenz-Struktur-Batterie (INSBAT): Eine Testbatterie zur Messung von Intelligenz* [Intelligence Structure Battery (INSBAT): A test battery for the measurement of intelligence]. Schuhfried.

Hu, L. t., & 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), 1–55. https://doi.org/10.1080/10705519909540118

Kemper, C. J., Trapp, S., Kathmann, N., Samuel, D. B., & Ziegler, M. (2019). Short versus long scales in clinical assessment: Exploring the trade-off between resources saved and psychometric quality lost using two measures of obsessive-compulsive symptoms. *Assessment*, *26*(5), 767–782. https://doi.org/10.1177/1073191118810057

Kretzschmar, A., Neubert, J. C., & Greiff, S. (2014). Komplexes Problemlösen, schulfachliche Kompetenzen und ihre Relation zu Schulnoten [Complex problem solving, school subject 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

Krieger, F., Zimmer, H. D., Greiff, S., Spinath, F. M., & Becker, N. (2019). Why are difficult figural matrices hard to solve? The role of selective encoding and working memory capacity. *Intelligence*, *72*, 35–48. https://doi.org/10.1016/j.intell.2018.11.007

Krieger, F., Stadler, M., Bühner, M., Fischer, F., & Greiff, S. (2020). *Cut to the chase: Assessing complex problem-solving skills in under 20 minutes* [Data set]. https://doi.org/https://osf.io/5grmj

Kröner, S., Plass, J., & Leutner, D. (2005). Intelligence assessment with computer simulations. *Intelligence*, *33*(4), 347–368. https://doi.org/10.1016/j.intell.2005.03.002

Kyriazos, T. A. (2018). Applied psychometrics: Sample size and sample power considerations in factor analysis (EFA, CFA) and SEM in general. *Psychology*, *9*(8), 2207–2230. https://doi.org/10.4236/psych.2018.98126

Leite, W. L., Huang, I.-C., & Marcoulides, G. A. (2008). Item selection for the development of short forms of scales using an ant colony optimization algorithm. *Multivariate Behavioral Research*, *43*(3), 411–431. https://doi.org/10.1080/00273170802285743

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

Lotz, C., Scherer, R., Greiff, S., & Sparfeldt, J. R. (2017). Intelligence in action –Effective strategic behaviors while solving complex problems. *Intelligence*, *64*, 98–112. https://doi.org/10.1016/j.intell.2017.08.002

Mainert, J., Niepel, C., Murphy, K. R., & Greiff, S. (2019). The incremental contribution of complex problem-solving skills to the prediction of job level, job complexity, and salary. *Journal of Business and Psychology*, *34*(6), 825–845. https://doi.org/10.1007/s10869-018-9561-x

Meißner, A., Greiff, S., Frischkorn, G. T., & Steinmayr, R. (2016). Predicting complex problem Solving and school grades with working memory and ability self-concept. *Learning and Individual Differences*, *49*, 323–331. https://doi.org/10.1016/j.lindif.2016.04.006

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Prentice-Hall Englewood Cliffs.

Novick, L. R., & Bassok, M. (2005). Problem solving. In K. J. Holyoak & R. G. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning* (*Vol. 137*). University Press.

OECD. (2017). *In the nature of problem solving: Using research to inspire 21st century learning*. B. Csapó & J. Funke Eds.). https://doi.org/10.1787/9789264273955-en

Oswald, F. L., McAbee, S. T., Redick, T. S., & Hambrick, D. Z. (2015). The development of a short domain-general measure of working memory capacity. *Behavior Research Methods*, *47*(4), 1343–1355. https://doi.org/10.3758/s13428-014-0543-2

Rölke, H. (2012). The ItemBuilder: A graphical authoring system for complex item development. In T. Bastiaens & G. Marks (Eds.), *E-Learn: World conference on e-learning in corporate, government, healthcare, and higher education* (pp. 344–353). AACE.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2). https://doi.org/10.18637/jss.v048.i02

Schoppek, W., Fischer, A., Funke, J., & Holt, D. (2019). On the future of complex problem solving: Seven questions, many answers? *Journal of Dynamic Decision Making*, *5*(2019). https://doi.org/10.11588/JDDM.2019.1.69294

Schroeders, U., Wilhelm, O., & Olaru, G. (2016). Meta-heuristics in short scale construction: Ant colony optimization and genetic algorithm. *PLoS One*, *11*(11), e0167110. https://doi.org/10.1371/journal.pone.0167110

Schweizer, F., Wüstenberg, S., & Greiff, S. (2013). Validity of the MicroDYN approach: Complex problem solving predicts school grades beyond working memory capacity. *Learning and Individual Differences*, *24*, 42–52. https://doi.org/10.1016/j.lindif.2012.12.011

Sijtsma, K., & Emons, W. H. M. (2011). Advice on total-score reliability issues in psychosomatic measurement. *Journal of Psychosomatic Research*, *70*(6), 565–572. https://doi.org/10.1016/j.jpsychores.2010.11.002

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 to assess complex problem solving. *Psychological Test and Assessment Modeling*, *54*(1), 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*(5), 289–305. https://doi.org/10.1016/j.intell.2013.05.002

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

Stadler, M., Niepel, C., & Greiff, S. (2016). Easily too difficult: Estimating item difficulty in computer simulated microworlds. *Computers in Human Behavior*, *65*, 100–106. https://doi.org/10.1016/j.chb.2016.08.025

Stadler, M., Becker, N., Schult, J., Niepel, C., Spinath, F. M., Sparfeldt, J. R., & Greiff, S. (2018). The logic of success: The relation between complex problem-solving skills and university achievement. *Higher Education*, *76*(1), 1–15. https://doi.org/10.1007/s10734-017-0189-y

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

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 action in computer-simulated problems]. Hogrefe.

Süß, H.-M. (1999). Intelligenz und komplexes Problemlösen [Intelligence and complex problem solving]. *Psychologische Rundschau*, *50*(4), 220–228. https://doi.org/10.1026//0033-3042.50.4.220

The Jamovi Project. (2020). *Jamovi* (Version 1.2) [Computer software]. https://www.jamovi.org

Unsworth, N., & Engle, R. (2005). Working memory capacity and fluid abilities: Examining the correlation between operation span and raven. *Intelligence*, *33*(1), 67–81. https://doi.org/10.1016/j.intell.2004.08.003

Wu, H., & Molnár, G. (2021). Logfile analyses of successful and unsuccessful strategy use in complex problem-solving: A cross-national comparison study. *European Journal of Psychology of Education*. Advance online publication. https://doi.org/10.1007/s10212-020-00516-y

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. *Knowledge and Learning,19*,(1–2) 127–146. https://psycnet.apa.org/doi/10.1007/s10758-014-9222-8

Zech, A., Bühner, M., Kröner, S., Heene, M., & Hilbert, S. (2017). The impact of symmetry: Explaining contradictory results concerning working memory, reasoning, and complex problem solving. *Journal of Intelligence*, *5*(2), 22. https://doi.org/10.3390/jintelligence5020022

Ziegler, M., Kemper, C. J., & Kruyen, P. (2014). Short scales – Five misunderstandings and ways to overcome them. *Journal of Individual Differences*, *35*(4), 185–189. https://doi.org/10.1027/1614-0001/a000148

History
Received August 14, 2020
Revision received May 5, 2021
Accepted May 27, 2021
Published online August 17, 2021

Section: Differential Psychology

Open Data

All analyses, the used data set, task characteristics of all microworlds, output of all 48 model comparisons, and a software for displaying items, and collecting data can be retrieved from the open science framework repository at https://osf.io/5grmj (Krieger et al., 2020).

Conflict of Interest

Samuel Greiff is one of two authors of the commercially available COMPRO test that is based on the multiple complex systems approach and that employs the MicroDYN approach. However, for any research and educational purposes, a free version of MicroDYN tasks is available and he receives royalties for COMPRO.

Funding
The contributions by Matthias Stadler and Frank Fischer were supported by a grant from the Deutsche Forschungsgemeinschaft DFG (COSIMA; DFG-Forschungsgruppe 2385).

Authorship

Florian Krieger and Matthias Stadler have joint first authorship.

ORCID
Florian Krieger
https://orcid.org/0000-0001-9981-8432

Florian Krieger
Université de Luxembourg
11, porte des Sciences
4366 Esch-sur-Alzette
Luxembourg
florian.krieger@uni.lu

Appendix A. Task descriptions and linear structural equations for all items of long version

Task	Short?	Classification	Input	Output	Relations	Eigendynmic	Linear structural equations
1	No	Simple	2	1	0	No	$X_t + 1 = 1*X_t + 2*A_t + 2*B_t$
2	No	Simple	2	2	2	No	$X_t + 1 = 1*X_t + 2*A_t + 0*B_t$ $Y_t + 1 = 1*Y_t + 0*A_t + 2*B_t$
3	Yes	Simple	2	2	2	No	$X_t + 1 = 1*X_t + 0*A_t + 2*B_t$ $Y_t + 1 = 1*Y_t + 0*A_t + 2*B_t$
4	No	Complex	3	2	3	No	$X_t + 1 = 1*X_t + 2*A_t + 2*B_t + 0*C_t$ $Y_t + 1 = 1*Y_t + 0*A_t + 0*B_t + 0*C_t$
5	No	Complex	3	3	3	No	$X_t + 1 = 1*X_t + 0*A_t + 2*B_t + 0*C_t$ $Y_t + 1 = 1*Y_t + 2*A_t + 0*B_t + 0*C_t$ $Z_t + 1 = 1*Z_t + 0*A_t + 0*B_t + 2*C_t$
6	No	Complex	3	3	5	No	$X_t + 1 = 1*X_t + 2*A_t + 2*B_t + 0*C_t$ $Y_t + 1 = 1*Y_t + 0*A_t + 2*B_t + 2*C_t$ $Z_t + 1 = 1*Z_t + 0*A_t + 0*B_t + 2*C_t$
7	Yes	Eigendynamic	3	2	2	Yes	$X_t + 1 = 1*X_t + 0*A_t + 0*B_t + 0*C_t$ $Y_t + 1 = (1*Y_t + 0*A_t + 2*B_t + 0*C_t) + 3$
8	No	Eigendynamic	3	2	3	Yes	$X_t + 1 = 1*X_t + 2*A_t + 0*B_t + 0*C_t$ $Y_t + 1 = (1*Y_t + 0*A_t + 2*B_t + 0*C_t) + 3$
9	Yes	Complex	3	3	4	No	$X_t + 1 = 1*X_t + 2*A_t + 0*B_t + 0*C_t$ $Y_t + 1 = 1*Y_t + 0*A_t + 2*B_t + 2*C_t$ $Z_t + 1 = 1*Z_t + 0*A_t + 0*B_t + 2*C_t$
10	Yes	Eigendynamic	3	3	4	Yes	$X_t + 1 = 1*X_t + 0*A_t + 0*B_t + 0*C_t$ $Y_t + 1 = (1*Y_t + 2*A_t + 2*B_t + 0*C_t) + 3$ $Z_t + 1 = 1*Z_t + 0*A_t + 0*B_t + 2*C_t$
11	Yes	Eigendynamic	3	3	5	Yes	$X_t + 1 = (1*X_t + 2*A_t + 2*B_t + 0*C_t) + 3$ $Y_t + 1 = 1*Y_t + 2*A_t + 0*B_t + 0*C_t$ $Z_t + 1 = 1*Z_t + 0*A_t + 0*B_t + 2*C_t$

Note. Short? = Is this task used in the short version of model 43?

Appendix B. Results of all 44 CFA, which showed convergence

		Fit indices (Step 2)						Further evaluation criteria (Step 3)			Descriptives for score		Intercorrelation between score and ...					
No.	Tasks	χ^2	*p*	*df*	CFI	RMSEA	SRMR	ω	Min λ	Diff ω	*M*	*SD*	WMC	GPA	*gF*	*gC*	VOTAT	Long version
1	1, 4, 7, 8, 10	6.122	.295	5	.998	.031	.068	.70	.40	−.04	0.54	0.29	0.21	0.29	0.39	0.36	0.53	0.90
3	1, 4, 7, 10, 11	4.327	.503	5	1.000	.000	.051	.69	.43	−.03	0.55	0.29	0.21	0.29	0.42	0.37	0.54	0.91
4	1, 4, 8, 10, 11	2.974	.704	5	1.000	.000	.040	.75	.37	−.04	0.58	0.31	0.19	0.26	0.40	0.34	0.54	0.90
5	1, 5, 7, 8, 10	4.173	.525	5	1.000	.000	.053	.71	.41	−.03	0.55	0.29	0.18	0.31	0.39	0.38	0.57	0.91
6	1, 5, 7, 8, 11	6.958	.224	5	.997	.041	.065	.70	.37	−.04	0.55	0.29	0.19	0.34	0.44	0.38	0.59	0.90
7	1, 5, 7, 10, 11	3.561	.614	5	1.000	.000	.042	.70	.43	−.03	0.56	0.29	0.18	0.31	0.42	0.39	0.58	0.91
8	1, 5, 8, 10, 11	4.769	.445	5	1.000	.000	.048	.75	.38	−.04	0.59	0.31	0.17	0.29	0.40	0.36	0.58	0.91
9	1, 9, 7, 8, 10	3.999	.550	5	1.000	.000	.044	.73	.39	−.04	0.53	0.30	0.21	0.28	0.42	0.40	0.49	0.88
10	1, 9, 7, 8, 11	5.665	.340	5	.999	.024	.051	.72	.35	−.04	0.53	0.30	0.22	0.30	0.46	0.39	0.51	0.88
11	1, 9, 7, 10, 11	2.442	.785	5	1.000	.000	.032	.72	.42	−.03	0.54	0.30	0.21	0.27	0.45	0.40	0.50	0.88
12	1, 9, 8, 10, 11	3.303	.653	5	1.000	.000	.035	.77	.36	−.04	0.57	0.33	0.20	0.25	0.42	0.37	0.50	0.89
13	1, 6, 7, 8, 10	2.994	.701	5	1.000	.000	.044	.71	.38	−.04	0.54	0.29	0.18	0.28	0.40	0.39	0.54	0.89
15	1, 6, 7, 10, 11	1.769	.880	5	1.000	.000	.031	.69	.42	−.03	0.55	0.29	0.18	0.28	0.43	0.40	0.55	0.90
16	1, 6, 8, 10, 11	3.462	.629	5	1.000	.000	.038	.75	.36	−.04	0.59	0.31	0.17	0.26	0.41	0.37	0.55	0.90
17	2, 4, 7, 8, 10	19.594	.001	5	.976	.112	.097	.72	.48	−.02	0.52	0.30	0.20	0.26	0.36	0.37	0.56	0.91
19	2, 4, 7, 10, 11	21.314	.001	5	.969	.119	.096	.70	.47	−.02	0.53	0.30	0.20	0.26	0.39	0.38	0.58	0.92
20	2, 4, 8, 10, 11	17.856	.003	5	.987	.106	.084	.76	.43	−.03	0.56	0.33	0.19	0.24	0.37	0.35	0.57	0.92
21	2, 5, 7, 8, 10	9.696	.084	5	.993	.064	.067	.72	.46	−.02	0.53	0.30	0.18	0.28	0.37	0.39	0.61	0.92
22	2, 5, 7, 8, 11	14.807	.011	5	.985	.092	.083	.72	.43	−.02	0.53	0.30	0.18	0.31	0.41	0.39	0.63	0.92
23	2, 5, 7, 10, 11	11.439	.043	5	.988	.075	.071	.71	.44	−.02	0.54	0.30	0.17	0.28	0.40	0.40	0.62	0.93
24	2, 5, 8, 10, 11	14.922	.011	5	.990	.093	.075	.76	.41	−.03	0.57	0.32	0.17	0.26	0.38	0.37	0.62	0.93
25	2, 9, 7, 8, 10	2.875	.719	5	1.000	.000	.029	.74	.40	−.03	0.51	0.31	0.21	0.25	0.39	0.41	0.53	0.90
26	2, 9, 7, 8, 11	6.431	.266	5	.998	.035	.046	.73	.37	−.03	0.51	0.31	0.22	0.27	0.43	0.40	0.55	0.90
27	2, 9, 7, 10, 11	3.721	.590	5	1.000	.000	.038	.73	.39	−.03	0.52	0.31	0.21	0.25	0.42	0.42	0.54	0.91
28	2, 9, 8, 10, 11	3.374	.643	5	1.000	.000	.032	.78	.36	−.04	0.55	0.34	0.20	0.23	0.40	0.38	0.54	0.91
29	2, 6, 7, 8, 10	2.812	.729	5	1.000	.000	.036	.72	.41	−.03	0.52	0.30	0.17	0.25	0.37	0.40	0.58	0.91
31	2, 6, 7, 10, 11	4.940	.423	5	1.000	.000	.049	.70	.40	−.02	0.53	0.30	0.17	0.25	0.41	0.41	0.59	0.93
32	2, 6, 8, 10, 11	5.265	.384	5	1.000	.015	.047	.76	.37	−.04	0.56	0.32	0.16	0.23	0.38	0.38	0.59	0.92
33	3, 4, 7, 8, 10	45.730	.000	5	.947	.188	.154	.75	.78	.01	0.53	0.31	0.22	0.31	0.37	0.35	0.62	0.93
34	3, 4, 7, 8, 11	45.244	.000	5	.948	.187	.151	.74	.73	.00	0.53	0.31	0.23	0.33	0.42	0.35	0.64	0.93

(Continued on next page)

Appendix B. (Continued)

		Fit indices (Step 2)						Further evaluation criteria (Step 3)			Descriptives for score		Intercorrelation between score and . . .					
No.	Tasks	χ^2	*p*	*df*	CFI	RMSEA	SRMR	ω	Min λ	Diff ω	*M*	*SD*	WMC	GPA	*gF*	*gC*	VOTAT	Long version
35	3, 4, 7, 10, 11	42.151	.000	5	.946	.179	.136	.73	.66	.01	0.54	0.31	0.22	0.31	0.40	0.36	0.64	0.94
36	3, 4, 8, 10, 11	43.928	.000	5	.966	.184	.125	.79	.78	.00	0.57	0.33	0.21	0.29	0.38	0.33	0.63	0.93
37	3, 5, 7, 8, 10	38.039	.000	5	.963	.169	.150	.75	.81	.01	0.53	0.31	0.19	0.34	0.37	0.37	0.67	0.94
38	3, 5, 7, 8, 11	39.686	.000	5	.962	.173	.152	.75	.76	.01	0.53	0.31	0.20	0.36	0.42	0.37	0.69	0.94
39	3, 5, 7, 10, 11	35.351	.000	5	.964	.162	.127	.74	.70	.01	0.54	0.31	0.19	0.33	0.40	0.38	0.68	0.94
40	3, 5, 8, 10, 11	50.663	.000	5	.964	.199	.145	.79	.83	.00	0.57	0.33	0.18	0.31	0.38	0.35	0.67	0.94
41	3, 9, 7, 8, 10	10.027	.074	5	.993	.066	.063	.76	.64	−.01	0.51	0.32	0.23	0.30	0.40	0.39	0.59	0.92
42	3, 9, 7, 8, 11	12.835	.025	5	.990	.082	.067	.75	.62	−.01	0.52	0.32	0.24	0.32	0.44	0.39	0.61	0.93
43	3, 9, 7, 10, 11	8.372	.137	5	.995	.054	.050	.75	.64	.00	0.52	0.32	0.23	0.30	0.43	0.40	0.60	0.93
44	3, 9, 8, 10, 11	10.994	.052	5	.995	.072	.053	.80	.64	−.02	0.56	0.35	0.21	0.28	0.41	0.37	0.60	0.92
45	3, 6, 7, 8, 10	23.111	.000	5	.975	.125	.095	.75	.70	.00	0.53	0.31	0.19	0.30	0.38	0.38	0.64	0.93
46	3, 6, 7, 8, 11	24.491	.000	5	.972	.130	.097	.74	.68	.00	0.53	0.31	0.20	0.33	0.43	0.38	0.66	0.93
47	3, 6, 7, 10, 11	21.578	.001	5	.972	.120	.089	.73	.70	.01	0.54	0.31	0.19	0.31	0.41	0.39	0.65	0.94
48	3, 6, 8, 10, 11	24.443	.000	5	.982	.130	.091	.78	.69	−.01	0.57	0.33	0.18	0.28	0.39	0.36	0.64	0.93