Original Article

Measuring Cognitive Ability in Children and Adolescents

Development and Validation of a New Test Battery for Working Memory Capacity

Benjamin Goecke[1,4], Luc Zimny[1], Johanna Hartung[2], Patrick Lösche[3], Jessika Golle[4], and Oliver Wilhelm[1]

[1]Institute of Psychology and Education, Ulm University, Germany
[2]Department of Psychology, University of Bonn, Germany
[3]Leibniz Research Institute for International Educational Research (DIPF), Centre for Research on Human Development and Education, Frankfurt am Main, Germany
[4]Hector Research Institute of Education Sciences and Psychology, University of Tübingen, Germany

Abstract: Working memory capacity (WMC) measures for children are unsatisfying in terms of lack of operational continuity with measures for adolescents and adults. Thus, we developed and validated a multivariate WMC test battery that uses WMC paradigms that can be applied from 1st grade onwards. In Study 1, we developed child-contextualized WMC tests and investigated their psychometric properties (including gender differences) in N = 343 1st graders. In Study 2, we juxtaposed child-contextualized and decontextualized instantiations of our tasks in N = 379 5th–10th graders. Child-contextualized tests were essentially equivalent to structurally identical, decontextualized WMC tests, and the battery correlated strongly with a multivariate measure of fluid intelligence. Across studies, we found support for good psychometric properties of the tests. The battery bridges the gap between child-specific and decontextualized WMC tests, is applicable for all ability levels, can be adapted easily in terms of difficulty and length.

Keywords: working memory capacity, intelligence, children, adolescents, measurement

Cognitive abilities predict children's future achievements, including school readiness, success in school, and academic achievement (Duncan et al., 2007; Lan et al., 2011; Peng & Kievit, 2020; Spinath et al., 2006). Adults' cognitive abilities are associated with further important life outcomes, such as life expectancy (Gottfredson & Deary, 2004; Sánchez-Izquierdo et al., 2023). However, measuring cognitive abilities across the lifespan is difficult because the specific tests and their parametrization commonly differ across age groups. Applying different tests across age (groups) somewhat limits the analysis of change in cognitive functioning in longitudinal studies and valid age comparisons in cross-sectional studies, as they are confounded by potential task specific differences. Hence, having measures of cognitive abilities available that can be applied to both children and older individuals would be beneficial. In addition to that, educational requirements, like reading or numerical requirements, could distort valid comparisons between tests conducted in children and older individuals, like adolescents. Consequently, the availability of maximal cognitive performance tests that are not confounded by educational requirements and that are applicable from childhood to adolescence would be beneficial.

Tests of working memory capacity (WMC; Wilhelm et al., 2013) are ideally suited for this purpose because WMC is the best explanatory variable of fluid intelligence (gf, e.g., Kyllonen & Christal, 1990; Oberauer et al., 2005; Wilhelm, 2005), which in turn is the best single predictor of general cognitive functioning (Carroll, 1993). Arguably, visualized WMC paradigms require minimal numerical or verbal prior knowledge. This makes them easy to instruct and contextualize for children. Moreover, as item difficulty in WMC tests hinges almost exclusively upon cognitive load and the relevant item attributes are amenable to experimental manipulation, WMC tests can easily be adjusted in difficulty (e.g., Redick et al., 2012; Unsworth et al., 2005). Across two studies, we describe the development and validation of three WMC tests based on the

binding hypothesis of working memory (Oberauer, 2005, 2019; Wilhelm et al., 2013) suited for children and adolescents. We provide evidence for their applicability in the 1st and 5th–10th grades by establishing measurement models for different test parametrizations. Moreover, we provide evidence for the equivalence of child-contextualized and conventional (i.e., abstract) tests, their correlation with measures of further cognitive abilities (vocabulary, phonological awareness, mathematical abilities, non-verbal intelligence, and gf), self-reported school grades, and school achievement.

The Relation of WMC and gf

The most important first-order factor of general intelligence is gf (Carroll, 1993), which has even been equated with general intelligence for psychometric reasons (e.g., Gustafsson, 1984; Kan et al., 2011; Undheim & Gustafsson, 1987). Gf denotes the ability to reason and solve problems in unprecedented situations (e.g., Cattell, 1971). From a psychometric perspective, tests tapping crystallized intelligence (gc) should be free of confounding gf requirements and vice versa (Carroll, 1993). In this notion, gf stands in contrast to test performances requiring gc. The best measures of gf are reasoning tests (Wilhelm, 2005). It has been argued that performance in such tests relies on maintaining, mentally manipulating, and storing units of information (Wilhelm & Schroeders, 2019). Consequently, failure to build, maintain, and update such representations would cause erroneous responses to reasoning problems (Oberauer, Süß, et al., 2008). In turn, these operations are likely key to successfully performing several other cognitive tasks. Hence, good measures of gf should require building, maintaining, and updating relations between chunks of information presented in a test (Oberauer, Süß, et al., 2008).

Working Memory (WM) is the cognitive system postulated to underly this type of processing (i.e., build, maintain, and update chunks of information). It is considered a cognitive system that facilitates non-automatized cognitive processes by actively retaining information while simultaneously allowing the processing of new information (Conway et al., 2008). Competing theories of WM agree that it is a system that can only store and process a limited amount of information, which is reflected in the term *working memory capacity* (c.f., Benchmark 12.1 in Oberauer et al., 2018). However, there are competing explanatory accounts that consider WMC as an individual differences construct, which can inform test development. Individual differences in WMC have been considered (a) a function of different levels of executive attention (e.g., Engle, 2002; Miyake & Friedman, 2012), (b) an interplay of primary and secondary memory resources in dual-store models of memory (e.g., Unsworth & Engle, 2007), or (c) the ability to temporarily build, maintain, and update arbitrary bindings (Oberauer, 2005, 2019; Wilhelm et al., 2013). In this last perspective, the limited capacity of WM results from an interference of bindings (Oberauer, Süβ, et al., 2008; Wilhelm et al., 2013). Although there is a body of evidence for each of these accounts, from our view, the evidence supporting the binding hypothesis is the most convincing (e.g., Goecke et al., 2021; Oberauer, 2019; Oberauer, Süβ, et al., 2008; Wilhelm et al., 2013), because the basic cognitive operations formulated in the binding hypothesis can be found to be the common source of variance across several task paradigms. This makes it a strong theoretical basis for developing WMC tests. Hence, the here presented test battery is based on the binding hypothesis.

Importantly, WMC is theorized to account for the performance in gf tests independent of competing explanatory accounts of WMC, such that test takers with higher capacity usually outperform test takers with lower capacity (Kyllonen & Christal, 1990; Oberauer et al., 2005; Wilhelm et al., 2013). This relationship is well-documented empirically (Kane et al., 2005; Oberauer et al., 2005), and in fact, WMC has long been discussed as the explanatory construct of gf (Johnson-Laird, 1980; Kyllonen & Christal, 1990). Given the critical role of WMC for gf and cognition in general, it makes a lot of sense to use WMC as the construct of choice for studying cognitive abilities at a young age. Critically, measures of WMC can help overcome persistent problems of gf tests for children, which we outline in the next paragraph. In addition to that, WMC is a stronger predictor of academic success at the start of formal education than general intelligence, including subtests of gf (Alloway & Alloway, 2010).

Issues With gf Tests for Children and How WMC Can Solve Them

First, gf tests often require some level of formal education, which is a key contributor to crystallized intelligence (gc). For example, arithmetic reasoning tests require knowledge of basic mathematics, such as counting and understanding the number system (Zhang et al., 2017). Similarly, verbal analogies, require a substantial vocabulary besides reasoning ability itself. Even non-verbal figural matrices partly involve basic mathematical operations like counting, addition, or subtraction (Lösche et al., 2015). To the extent that such knowledge contributes to performance in gf tasks, these tests are at odds with the idea of gf tests as measures of decontextualized reasoning ability (e.g., Wilhelm & Schroeders, 2019). Although these issues may

be less relevant in adolescent or adult samples in which required minimal levels of formal education can be readily assumed, it is impossible to rule out serious test disadvantages for children with a socio-economically and educationally challenged background (Bradley & Corwyn, 2002; Lee & Burkam, 2002), as language proficiency in children often strongly correlates with the educational levels of the parents (Dixon et al., 2012). Relatively language-free WMC tests can be developed as a possible remedy, for example, in the visuo-spatial stimulus domain. Although language-related assessments of intelligence were shown to be more predictive of school grades than language-free assessments (Roth et al., 2015), further potential benefits of relatively language-free assessments include better acceptance rates among individuals or parents with limited language skills, for example due to cultural reasons (non-native speakers), thereby improving inclusivity and reducing biases in cognitive evaluations. Whether or not better acceptance holds for the new WMC measures is an empirical question.

Second, other well-known problems of gf tests relate to their inherent task characteristics. For example, gf tests are prone to individual differences in speed-accuracy trade-offs (e.g., Goldhammer, 2015; Goldhammer & Klein Entink, 2011; Phillips & Rabbitt, 1995). Test takers show differences in the speed at which they complete test items, which in turn co-determines their accuracy level. In gf tests, such effects usually increase with stricter time-limits. Differential levels of speed-accuracy trade-offs threaten the validity of gf tests because they confound what shall be measured in the first place: the pure ability to reason. Well-designed WMC tests might mitigate these issues, as presentation times for stimuli are usually fixed for all subjects, and taking longer to respond to a certain trial does not bear costs in terms of lost time for other trials. Additionally, maximum performance tests require a veridical response standard relative to which test behavior is compared (Cronbach, 1949). While this is rarely an issue in adult gf tests, some measures of reasoning ability for children fail to meet this requirement (i.e., more than one response can be inferred from the premises given).

This relates to the third point, that developing gf tests has been described to be more art than science (Kyllonen & Christal, 1990). That is, it is hard to predict item difficulty – and hence test difficulty – based on item characteristics, such as the number of premises or the number of rules. Although some progress in this regard has been made (e.g., Arendasy, 2005; Becker et al., 2015; Blum et al., 2016; Loe & Rust, 2019; Primi, 2001), the effects of item attributes on item difficulty of various gf tests are still hardly understood (e.g., Hartung et al., 2022). In contrast, the effects of item attributes on item difficulty (i.e., trial difficulty) in WMC tests can be explained much better based on the previous literature (Oberauer et al., 2018). Item difficulty in WMC test can be understood as the requirements on working memory in terms of working memory load (Larson et al., 1988), which, in turn, could be understood as the number of bindings that have to be processed (Goecke et al., 2021). Crucially, this powerful simplification makes automatic item generation for WMC tests much easier. In sum, WMC tests are much more straightforward to develop than gf tests and because automatic item generation is easier to achieve, human errors in test development are easier to prevent.

Lastly, longitudinal studies of cognitive abilities from childhood to adulthood are often limited by applying different measurement instruments at different ages (e.g., Elliott & Shepherd, 2006). Although different gf tests arguably measure the same underlying ability, investigating trends in cognitive development across ages would be easier with tests that are linked to each other by overlapping (linking-) items that adhere to the same construction principles. Only a few gf measures are available for both children and adults that meet these criteria to some extent (e.g., figural matrices). However, to our knowledge, no multivariate gf test battery (i.e., including more than one test) exists that can be used from early childhood (i.e., preschool) to adolescence.

Extant Research of WMC in Children

A large body of research on WMC in children is available. Numerous paradigms such as simple (e.g., Ahmed et al., 2022; Panesi et al., 2022; Reynolds et al., 2022) and complex span (Gustafsson & Wolff, 2015; Peng & Fuchs, 2017), updating (e.g., Blume et al., 2022; Galeano-Keiner et al., 2022; Neubauer et al., 2019), binding (e.g., Cheng & Kibbe, 2022; Gray et al., 2017), corsi block (e.g., Cirino et al., 2018; Demetriou et al., 2014; Houwen et al., 2019), *n*-Back tests (e.g., Cabbage et al., 2017; Cirino et al., 2018; Watrin et al., 2022), and others (e.g., Cowan et al., 2011; Renner et al., 2021) are applied to measure WMC in children. Although some of these studies used WMC tests that were relatively abstract (e.g., using colored rectangles), many other studies applied child-contextualized stimuli (e.g., colored animals) by embedding the actual tests into a cover story (e.g., helping the farmer remember the animals). To the best of our knowledge, no study has tested the construct equivalence between abstract and child-contextualized stimuli so far. Additionally, some of these tests are already organized in multivariate test batteries (e.g., The Comprehensive Assessment Battery for Children – Working Memory; CABC-WM, Cabbage et al., 2017), but there is currently no test battery available that allows for efficient group test-settings starting at a young

age (e.g., 1st grade). For example, the CABC-WM requires individual test sessions that last approximately 77 min (not including breaks).

The present studies report on developing and validating a multivariate WMC test battery that can be administered from a very young age to adolescence in more efficient group settings. Having a test battery available that can be administered in group settings would be beneficial for multivariate research purposes and swift screenings of cognitive abilities in educational settings (e.g., classroom settings).

Aims of the Present Studies

Across two studies, we developed a test battery of WMC suitable from childhood (i.e., 1st grade) to adolescence (i.e., 10th grade) and investigated its psychometric properties.

In Study 1, we were interested in the following aspects:

- First, based on the binding hypothesis of working memory, we predicted that trial difficulty is a function of load levels. The more information is needed to be held in working memory the more difficult a trial.
- Second, we aimed at finding a suitable measurement model by means of confirmatory factor analysis. Competing models for the test battery will be tested against each other. In the same vein, we examined the reliability (factor saturation) of our test battery.
- Third, we were interested in measurement invariance across gender and thus tested for the absence of gender differences.
- Fourth, we explored the nomological net of our test battery by testing associations between the WMC test battery and precursory cognitive skills (vocabulary, mathematical abilities, phonological awareness, and figural reasoning) relevant to early educational development in children.

In Study 2, we extended the validation of the WMC battery to an older age range, specifically adolescent students from 5th to 10th grade. In this study, we examined the following aspects:

- First, as in Study 1, trial difficulty should increase with increasing load levels.
- Second, we were interested in whether or not the child-contextualized stimuli used in the test battery in Study 1 could be exchanged with more conventional stimuli without altering what the test battery measures, thus assessing the equivalence of both task instantiations in a latent variable context.
- Third, to establish the convergent validity of the WMC battery, we examined its relationship with an established test of gf for adolescents. In line with previous literature, we expected to find a very high correlation between WMC and gf.
- Lastly, we explored associations between the WMC test battery and school-related achievement variables, such as self-reported grades and standardized test scores.

The study was not preregistered. We provide all newly developed tests and materials for interested researchers to use in their studies (https://github.com/luc-w/wmc-ulkabe). All data and code necessary to reproduce our results can be retrieved at https://osf.io/evfsd/ (Goecke et al., 2024).

Methods and Materials: Study 1

Sample and Procedure

Subjects were 343 children from 14 primary schools in southwestern Germany, of which 323 (94.2%) were included in the analyses after data cleaning (see SM Table 1 in the supplementary materials for more details). At the time of testing, the children attended the end of 1st grade and were, on average, 7.41 years old (SD = 0.53), and 59.4% of the sample were girls. Informed consent from the parents was obtained prior to the study. The anonymity of participants was guaranteed, and the study was conducted in accordance with the Declaration of Helsinki and was approved by a local ethics committee. Participation was voluntary and could be canceled at any time.

Testing took place during school hours in groups of no more than six children. Trained experimenters supervised the testing sessions. After an initial welcome, all tests were presented on 10.4-in. (2,000 × 1,200 pixels) Android tablets. Apart from the initial welcoming, subjects received all test instructions via headphones (i.e., audio instructions). The administrators ensured that all children started the individual tests at the same time, but the children could work on the tasks self-paced. All tests were programmed with Inquisit 6 (*Inquisit 6*, Version 6.5.2, 2022) and administered via the Inquisit app. Each session lasted 60 min, including short, flexible breaks between tests, and including preparation and wrap-up time. The single tests took approximately 15 min, including instruction times.

Measures

Working Memory Capacity

The newly developed WMC tests were presented in a fixed order, as presented in the following sections. The

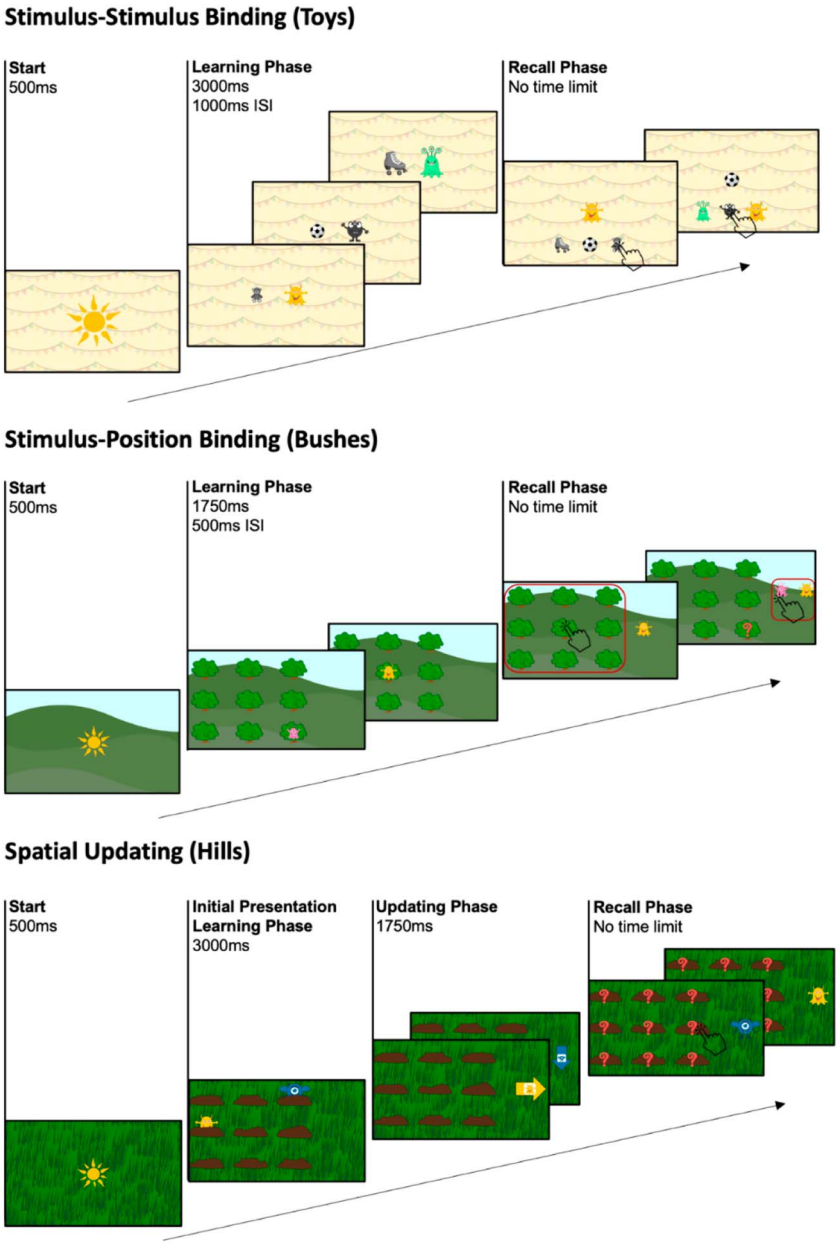

Figure 1. Examples of the three working memory capacity tests for children. For the stimulus-stimulus binding test, a Load 3 trial is presented. For the stimulus-position binding test, a Load 2 trial is presented. For the spatial updating test, a load 2 trial is presented that contains two consecutive updating steps (one update for each stimulus).

development of all tests was based on the rationale of the binding hypothesis of WMC (Oberauer, 2005; Wilhelm et al., 2013). An overview of the triallists for all tests can be found in the supplementary materials at https://osf.io/evfsd/. Figure 1 provides an overview of each test.

Stimulus–Stimulus Binding (Toys)

Children had to memorize a sequence of pairs of monsters and toys (Figure 1). Monster-Toy pairs appeared in the middle of the screen for 3,000 ms with an interstimulus-interval (ISI) of 1,000 ms. In the Load 2 condition, a sequence of two pairs was presented. Analogously, in the Load 3 and the Load 4 conditions, sequences of three and four monsters-toy pairs were presented, respectively. At the end of each sequence, the children had to recall the bindings: either a monster was presented in the middle of the screen, and the children had to choose the matching toys from an array below, or a toy was presented in the middle of the screen, and the children had to indicate the matching monster from the array below. Each monster-toy pair was probed once.

The number of possible responses was equivalent to the load level of a sequence. The order of the recall trials was not identical to the previously presented sequences, which assured that only the memory for pairwise relations was tested, but not item memory (i.e., sequences) itself. The

stimulus material included images of eight monsters and eight toys. Hence, it was possible for the same monsters to be paired with different toys several times, but no monsters and no toys were shown more than once in each sequence. A fixation point (the picture of a sun) was presented for 500 ms between single trials to indicate the start of a new trial. The instruction included one Load 2 trial as an example with explanations. Before the test part started, three practice trials (2× Load 2 trials and 1× Load 3 trial) were presented with feedback about the correctness of the response. Each load condition comprised six trials, resulting in 54 responses per child.

Stimulus-Position Binding (Bushes)

Monsters were presented on different positions of a 3 × 3 grid (where the cells were depicted as bushes), and children were instructed to memorize their position but not the sequence. One trial consisted of the sequential presentation of a list of stimulus-position bindings (Figure 1). Immediately after the presentation of these bindings, a recall test of each stimulus-position binding followed in a pseudo-random order. For 50% of the recall trials, a position in the grid (i.e., a bush) was marked, and all monsters used in the current sequence were lined up to the right of the grid. Children were required to select the monster that previously had been paired with the presented bush. For the other 50% of the recall trials, children were presented with one of the monsters and had to choose the position in the grid where the monster was presented before. This procedure ensured that only the memory for pairwise relations was tested, but not item memory (i.e., sequences) itself. To assure the requirement of memorizing and remembering only temporary but not long-term memory bindings, again, only a small set of eight monsters was used repeatedly across trials. Hence, it was possible for the same monsters to be presented in different positions across trials. A fixation point (the picture of a sun) was presented for 500 ms between single trials to indicate the start of a new trial. The load levels of this test ranged from 1 to 4, and prior to 16 test trials, children had to work on four practice trials (2× Load 1 trials, 2× Load 2 trials). During the learning phase, the presentation time for each stimulus-position binding was held constant across trials and amounted to 1,750 ms with an ISI of 500 ms between single stimuli-position bindings.

Spatial Updating (Hills)

Monsters were initially presented in different positions on a 3 × 3 grid matrix (where the cells were depicted as hills). After the initial simultaneous presentation of the stimuli, the monsters disappeared, and the updating phase of the respective trial began (Figure 1). In this updating phase, arrows presented on the right-hand side of the screen indicated the movement of the monsters underneath the grid to other positions within the grid. The children's' task was to bear in mind the initial position of each presented monster and to update this position according to the presented arrows. Children were required to memorize the most recent position of each monster in the grid. Immediately after the presentation of the last updating step, a recall test of each stimulus-position binding followed in a pseudo-random order. Children were required to select the hill that, in the last updating step, had been paired with the monster presented to the right of the grid at the recall trial. Again, a set of eight monsters was used repeatedly across trials. Hence, it was possible for the same monsters to be presented in different positions across trials. The load levels of this test ranged from 1 to 2, with 2 and 3 updating steps for load one trials and from 4 to 14 updating steps within a load level. A fixation point (the picture of a sun) was presented for 2,000 ms between single trials to indicate the start of a new trial. There were 14 test trials, and children had to work on five practice trials beforehand (two trials with explanations during the instructions and three practice trials without explanations). During the learning phase, the presentation time for each stimulus-position binding was held constant across trials and amounted 3,000 ms for the initial presentation of the stimuli in the grid and 1,750 ms for each updating operation (i.e., arrows depicting where a target stimulus moved). The scoring of all WMC tests followed a partial credit unit (Conway et al., 2005). For all tests, a score of accuracy was computed as the proportion of correct responses across queries in a trial.

Performance Indicators in Primary Schools

As proxy measures of intelligence, we applied the subtests of vocabulary (e.g., identifying objects embedded within a picture), phonological awareness (e.g., repeating or rhyming words), and mathematical abilities (e.g., number reading, picture-based addition and subtraction, arithmetic tasks with dots, or formal arithmetic tasks) of a German adaption of the Performance Indicators in Primary School (c.f., Bäuerlein et al., 2021; Tymms et al., 2014). Additionally, we applied a single figural matrix reasoning task from another test battery (BUEGA-II; Esser et al., 2021) used to identify developmental disorders in primary-school-aged children. This task contains 38 items with increasing difficulty but was aborted after three consecutive or four total errors. Children were asked to identify the missing picture in a matrix of pictures and select the correct picture out of a set of at least five response alternatives. We built four parcels for all subtests that were subsequently used as indicators in a latent variable analysis approach. Please note that, as participation was voluntary, only N = 211 children participated in this test, which was applied in a separate test session.

Statistical Analyses

We provide all materials (stimuli, programs), data, code, and analysis outputs on a repository of the open science framework at https://osf.io/evfsd/.

All analyses were conducted with R (R Core Team, 2024). General data handling and data visualization were performed with packages from the *tidyverse* (Wickham et al., 2019). Descriptive statistics were computed with the *psych* package (Revelle, 2024), and latent variable models were estimated with the *lavaan* package (Rosseel, 2012). Information regarding the versions of the software used can be found in the online repository. We used full information maximum likelihood estimation under the assumption of missing completely at random to combine missing data and parameter estimation in a single step (Enders, 2010; Schafer & Graham, 2002). We computed four parcels with similar means (i.e., proportion correct scores) for each WMC test as indicators for the latent variable models (Little et al., 2002). Models were estimated using robust maximum likelihood (MLR) estimation. We accounted for the nested data structure (children nested in classes) using a design-based estimation approach to estimate cluster-robust standard errors for all models (DiStefano & Zhang, 2022).

Regarding model fit, we pursued a twofold strategy. First, the classical following fit statistics were considered to indicate good model fit: CFI (comparative fit index) ≥ .95, RMSEA (root-mean-square error of approximation) ≤ .06, and SRMR (standardized root-mean-square residual) ≤ .08 (Hu & Bentler, 1999). For acceptable model fit, the following thresholds were used: CFI ≥ .90, RMSEA ≤ .08, and SRMR ≤ .10 (Bentler, 1990; Browne & Cudeck, 1992). Second, we computed dynamic fit indices (McNeish & Wolf, 2023; Wolf & McNeish, 2023), as concerns regarding fixed cutoffs for evaluating model fit are growing (e.g., Groskurth et al., 2023). Specifically, we evaluated the empirical model fits (CFI and RMSEA) of our data against the suggested dynamic cutoffs for mediocre, fair, and close model fit, whereby "fair fit on both fit indices [CFI and RMSEA] would be [the] minimally acceptable criteria to confidently declare acceptable fit" (McNeish & Wolf, 2023, pp. 37–38) and close model fit would indicate excellent fit of a model. Throughout the results, we transparently report both traditional and dynamic fit indices but do not reject models based on single fit indices. Please note that detailed numerical values for each model are available online. We computed McDonald's ω as an estimate of factor saturation (McDonald, 1999). Factor saturation indicates how much variance is accounted for by a latent variable in all underlying indicators (Brunner et al., 2012).

Gender differences were examined in a multigroup confirmatory factor analysis (MGCFA) with reference group approach (boys as reference group; Schroeders & Gnambs, 2020). Measurement invariance conditions were tested in the following order (c.f., Vandenberg & Lance, 2000): configural, metric, scalar, and strict measurement invariance. Scalar invariance can be understood as a precondition for comparing latent group means; standardized latent means can be interpreted in the metric of Cohen's *d* (Cohen, 1988) if a reference group identification approach is applied. We examined the conducted invariance tests by assessing changes in the CFI with values > .01 as an indication of model deterioration (Chen, 2007).

Results: Study 1

Descriptive Results

Table 1 shows the descriptive statistics for all WMC tests. All tests showed a broad range of trial difficulty with neither floor nor ceiling effects. Thus, the tests were well-parametrized for children attending 1st grade. As expected, test difficulty systematically increased with load in the binding-stimulus and binding-position, or with updating steps in the spatial updating task, respectively. The mean scores of the tests correlated positively ($r_{BS\sim BP} = .40$; $r_{BS\sim SU} = .49$; $r_{BP\sim SU} = .59$; all $p < .001$).

Measurement Models

Next, we estimated unidimensional measurement models separately for the three WMC tests. Four parcels with comparable means were computed and loaded on a common latent factor for each model. As Table 2 shows, model fit was good for all tests. However, the RMSEA for the measurement model of Binding Position was slightly out of bounds. This deviating result concerning model fit might be because the RMSEA tends to be over-sensitive for models with low degrees of freedom (Kenny et al., 2015; Shi et al., 2022). All general factors captured substantial shares of variance (all $p < .001$) and showed acceptable to good reliabilities (Table 2).

Subsequently, we tested two competing latent factor models. First, we tested a g-factor model with a single factor on which all 12 WMC indicators loaded. This model corresponds to the idea that individual tests play no role in individual differences. Second, we tested a correlated group factors model with three correlated factors based on the single tests (i.e., four indicators each). This model reflects the assumption that each test assesses a distinct construct but that these constructs are correlated. Importantly, a correlated group factors model with three indicators is statistically equivalent to a higher-order factor

Table 1. Descriptive statistics for child-contextualized working memory capacity tests

Construct	Test	Score	Trials	*M* (*SD*)	Skew.	Kurt.
WMC	Binding stimulus (BS)	Load2	6	0.79 (.17)	−0.87	0.82
		Load3	6	0.55 (.18)	0.12	−0.48
		Load4	6	0.38 (.14)	0.39	−0.04
		Total	18	0.53 (.11)	0.03	0.03
	Binding position (BP)	Load1	2	0.96 (.16)	−3.96	16.23
		Load2	6	0.87 (.15)	−1.27	1.44
		Load3	4	0.56 (.22)	0.12	−0.58
		Load4	4	0.37 (.37)	0.48	−0.25
		Total	16	0.60 (.13)	−0.19	−0.03
	Spatial updating (SU)	Load1, 0 updates	2	0.84 (.30)	−1.68	1.62
		Load2, 1 update	1	0.40 (.26)	0.37	−0.57
		Load2, 2 updates	1	0.56 (.44)	−0.24	−1.66
		Load2, 3 updates	1	0.45 (.43)	0.18	−1.60
		Load2, 4 updates	1	0.63 (.41)	−0.50	−1.36
		Load2, 5 updates	1	0.48 (.43)	0.09	−1.67
		Load2, 6 updates	1	0.47 (.45)	0.10	−1.75
		Load2, 7 updates	1	0.40 (.43)	0.41	−1.56
		Load2, 8 updates	1	0.35 (.41)	0.59	−1.25
		Load2, 9 updates	1	0.27 (.37)	0.96	−0.55
		Load2, 10 updates	1	0.39 (.43)	0.45	−1.50
		Load2, 11 updates	1	0.31 (.38)	0.74	−0.93
		Load2, 12 updates	1	0.25 (.36)	1.04	−0.30
		Total	14	0.46 (.24)	0.15	−0.90
Intelligence	Figural matrix reasoning		38	13.23 (5.56)	0.58	0.55
	Mathematical abilities		53	37.61 (9.54)	−1.37	2.33
	Phonological awareness		26	21.25 (4.13)	−0.91	0.36
	Vocabulary		13	10.11 (3.55)	−1.40	0.83

Note. N = 323 for WMC. N = 211 for Intelligence. skew. = skewness; kurt. = kurtosis.

Table 2. Unidimensional measurement models of the three child-contextualized WMC tests

Measurement model	Indicators	χ^2	*df*	*p*	CFI	RMSEA [90% CI]	SRMR	ω
1. Binding stimulus	4	3.56	2	.170	.984	.049 [.000, .117]	.024	.59
2. Binding position	4	9.73	2	.010	.974	.109 [.041, .193]	.025	.72
3. Spatial updating	4	0.89	2	.640	< 1.000	.000 [.000, .089]	.006	.83

Note. N = 323. For all measurement models, the dynamic fit indices indicated a sensitivity of < .5 for the proposed cutoffs, indicating that the simulated cutoffs were only little sensitive to misfit of the models.

model, which reflects the concept of a higher-level construct (WMC) explaining the correlations (i.e., common variance) amongst the first-order constructs. The g-factor model did not fit the data well: $\chi^2(54) = 156.31$, CFI = .904, RMSEA = .077, < fair dynamic fit, SRMR = .054. In contrast to that, the correlated factors model (statistically equivalent with a higher-order model) fitted the data well [$\chi^2(51) = 60.58$, CFI = .991, RMSEA = .024, close dynamic fit, SRMR = .031] and significantly better than the g-factor model [$\Delta\chi^2(3) = 99.99$, $p < .001$]. The higher-order factor had significant variance ($p < .001$), and its saturation was acceptable ($\omega_{WMC} = .79$). The loadings on the higher-order factor and for each first-order factor are displayed in Figure 2.

Gender Differences

We studied gender differences by comparing latent means across boys (n = 130) and girls (n = 147). As we were both

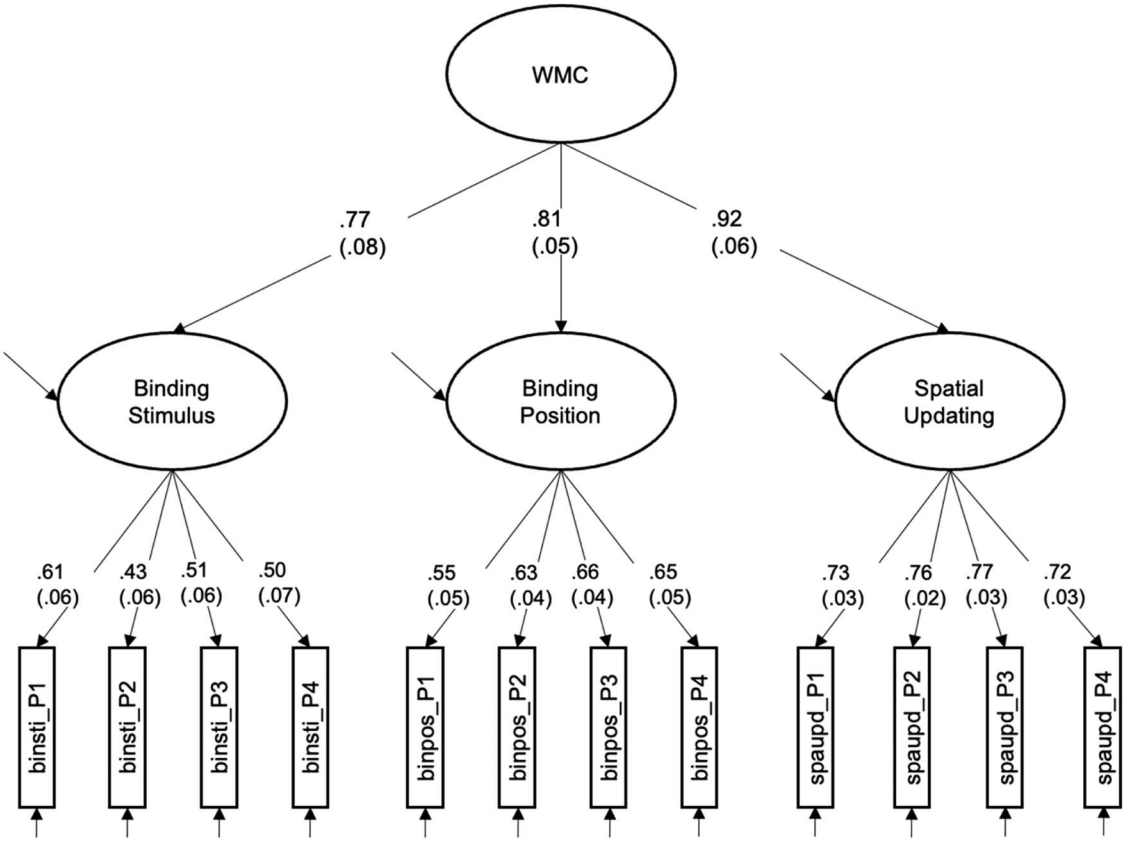

Figure 2. Higher-order model of working memory capacity (child-contextualized tests). *N* = 323; $\chi^2(51) = 60.58$, CFI = .991, RMSEA = .024, close dynamic fit, SRMR = .031. All parameters are standardized. Standard errors are given in parentheses.

interested in differences on the single task factors and on the higher-order factor WMC, we first established measurement invariance for the correlated factors model with three task factors and then included the second-order factor and additionally established measurement invariance for the higher-order structure of the model. For both types of model strict measurement invariance holds (see SM Table 3 in the supplementary materials). Due to that, means of all latent factors can be interpreted. As boys were the reference group, the latent mean estimates of the girls depict the gender differences (i.e., standardized mean differences; see parameter estimates in Table 3). The only factor for which substantial mean differences could be observed was the spatial updating factor. For two of the three tests (Binding Position and Binding Stimulus), no moderating effect of gender could be observed. In contrast to that, boys had a substantially higher mean in the Spatial Updating task compared to girls. The correlations between the three postulated factors were by and large similar across groups, although the correlation between Binding Stimulus and Binding Position was less nuanced in girls. Importantly, no moderating effect of gender could be observed for the higher-order factor of WMC. In addition to that, factor saturation for WMC in both groups was sufficient ($\omega_{boys} = .82$; $\omega_{girls} = .78$).

Correlations With Performance Indicators in Primary Schools

To examine the relationship of the higher-order factor WMC with the subscales of the performance indicators in primary schools, we first specified a correlated factors model in which each factor represents one of the tasks: figural matrix reasoning, mathematical abilities, phonological awareness, and vocabulary. Each factor was indicated by four parcels, and the model fitted the data well: $\chi^2(98) = 158.12$, CFI = .979, RMSEA = .054, close dynamic fit, SRMR = .042. Saturation of the group factors was very good (all $\omega \geq .85$). In a model with both the higher-order factor model of WMC and the correlated factors model [$\chi^2(337) = 482.43$, CFI = .962, RMSEA = .037, close dynamic fit, SRMR = .056], the correlations were $\rho_{WMC,\ Figural\ Reasoning} = .44$ (*SE* = 0.07), $\rho_{WMC,\ Mathematical\ Abilities} = .34$ (*SE* = 0.13), $\rho_{WMC,\ Phonological\ Awareness} = .48$ (*SE* = 0.10), $\rho_{WMC,\ Vocabulary} = .37$ (*SE* = 0.09), respectively.

Table 3. Parameter estimates of strict invariance model depicting three task factors (binding position, binding stimulus, updating position) and the higher-order factor (WMC)

	Unstandardized estimates (*SE*)			
Latent mean	Boys	Girls	95% CI	*d*
Mean (BinPos)	0	−0.02 (0.19)	[−.28, .32]	−0.02
Mean (BinSti)	0	0.02 (0.15)	[−.35, .31]	−0.02
Mean (SpaUpd)	0	−0.31 (0.14)	[−.64, −.04]	−0.34
Mean (WMC)	0	−0.23 (0.17)	[−.60, .10]	−0.25

	Standardized estimates (*SE*)			
Latent correlation	Boys	95% CI	Girls	95% CI
ρ (BinPos, BinSti)	0.71 (.14)	[.45, .98]	0.52 (.11)	[.28, .76]
ρ (BinPos, SpaUpd)	0.80 (.05)	[.71, .90]	0.72 (.07)	[.62, .81]
ρ (BinSti, SpaUpd)	0.74 (.07)	[.60, .89]	0.67 (.08)	[.51, .83]

Note. *d* = Cohen's *d*. The standardized mean differences depict the standardized means of the girls (boys as reference). Negative values of *d* indicate lower means for girls, whereas positive values of *d* indicate higher means for girls.

Discussion: Study 1

In Study 1, we developed child-contextualized WMC tests and provided evidence for their applicability and validity in a large sample of 1st graders. The descriptive statistics indicated a broad range of test difficulties and endorsed the adequacy of test parametrizations. In line with the binding hypothesis, the test difficulty systematically increased with increasing load and updating steps. Pragmatically, it was possible to administer the tests with up to six 1st-graders at a time while ensuring all children understood the instructions and diligently completed the test. We attribute this to the standardized audio instructions, the self-paced practice trials, and the appeal of working with tablets.

A higher-order factor model explained the data well, further supporting the validity of the tests as indicators of WMC. To investigate gender differences, we compared the latent means of the three subtests based on a correlated factors model and of the higher-order factor based on a second-order structure model between boys and girls after establishing measurement invariance. Two of the three tests (the two binding paradigms) were free of gender-associated mean differences, whereas boys had a higher mean on the spatial updating task. This aspect of the tests should be considered in future applications, because an inherent male advantage on one of the tests might bias rank-order comparisons in applied settings. Importantly, however, we observed no general male advantage when considering the overarching construct of WMC.

We examined the correlations between the higher-order factor of WMC and four tests of another test instrument tapping both crystallized and fluid abilities, which can be interpreted as precursory skills at school entry: two of the four tests (phonological awareness and vocabulary) arguably require cognitive performances that are akin to that of classical tests of crystallized intelligence (e.g., Schipolowski et al., 2014), whereas the other two (mathematical abilities and figural matrix reasoning) arguably require more fluid aspects of cognition. Hence, one expectation was that WMC should be associated more closely with tests primarily requiring fluid abilities than tests primarily requiring crystallized abilities. Contrary to this expectation, we found that the associations between WMC and all other test factors amounted to more or less the same height. However, the relations with mathematical abilities and figural matrix reasoning were slightly larger. On the one hand, this result can be clearly interpreted as evidence for the discriminant validity of the WMC factor. On the other hand, stronger correlations between WMC, mathematical abilities, and especially figural matrix reasoning could have been expected (Dirk & Schmiedek, 2016; Oberauer et al., 2005). Although we reported latent correlations and therefore accounted for measurement error, the height of the correlations might be attributable to the restriction of using a single test for each of the underlying constructs (i.e., mathematical abilities and reasoning ability), but also to the operationalization of the respective abilities the tests were supposed to tap. After all, it is best to measure reasoning ability with more than one task (Wilhelm, 2005).

In Study 2, we aimed to extend the validity findings to an older age range (i.e., adolescent students) and to demonstrate the equivalence of the new child-contextualized tests with conventional WMC tests. In this way, we aimed to make the tests more broadly applicable and, more importantly, to bridge the gap to measures established in research on the rest of the lifespan.

Methods and Materials: Study 2

Sample and Procedure

We conducted Study 2 with 379 children from 12 classes of a secondary school in southwestern Germany. After conducting a stepwise data cleaning procedure (see SM Table 2 in the supplementary materials for more details), the final analysis sample consisted of 368 children attending Grades 5 to 10, of which 46.5% were girls and which were, on average, 13.10 years old (SD = 1.84). SM Table 4, which can be found in the supplementary materials, provides more detailed information regarding the age distribution within class levels. The study was conducted according to the declaration of Helsinki. Informed consent from the parents was obtained prior to the study. Participation was voluntary and could be canceled at any time.

Test sessions took place during school hours. One class was tested at a time, supervised by trained administrators. Both the hardware (i.e., tablet computer) and software (i.e., Inquisit) were identical to Study 1. Each session lasted 90 min, with short breaks in between tests. To estimate all covariances between tests of interest with sufficient sample size, we implemented a planned-missingness design (Rhemtulla & Little, 2012, see SM Table 5 in the supplementary materials). We administered two types of WMC tests in a within-subjects design: either a child-contextualized version or a conventional version (i.e., a version with abstract stimuli). Due to time limitations, each class worked on five different WMC tests instead of all tests. The order of the tests was balanced across classes so that every test was administered on every position of the test sequence.The parallel test versions (i.e., child-contextualized vs. conventional versions) were (approximately) balanced across classes to control for learning effects and fatigue. For example, the child-contextualized version of the binding position test was administered equally often as the first, second, third, fourth, or fifth test, and the child-contextualized version was administered equally often before and after the conventional version of the test.

Measures

Working Memory Capacity

We administered the three WMC tests from Study 1. These child-contextualized test versions were identical in appearance (i.e., contextualized with monsters and corresponding cover stories per test) to the tests in Study 1. Additionally, we administered abstract test versions of the WMC tests, in which the child-contextualized stimuli were exchanged for abstract stimuli (i.e., geometric shapes and colors) as often used in conventional WMC tests, and there was no cover story (see Figure 3). We will refer to this test version as the conventional test version.

The tests' parameterizations (i.e., the applied load levels and hence difficulties) were adapted for subjects from Grades 5–7 and 8–10, respectively. The adaption of the tests was conducted through removing relatively easy trials from the tests and adding trials with higher loads and/or updating steps, to gain difficulty. Importantly, the parametrization for child-contextualized tests and conventional tests within the same age group (e.g., Grades 5–7) was identical, which allows for investigating test equivalence. Moreover, the age-specific versions of the WMC tests between age groups (i.e., child-contextualized in grades 5–7 vs. conventional versions in Grades 8–10) comprised a large set of common trials in all tests (14 trials for binding position, 15 trials for binding stimulus, and 10 trials for spatial updating, respectively), allowing to link the age-specific tests score psychometrically.

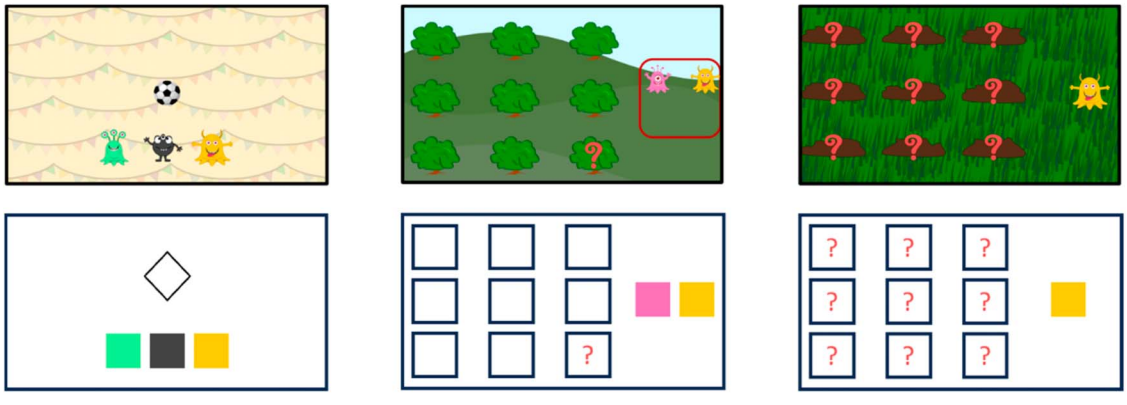

Figure 3. Illustration of structurally equivalent child-contextualized and conventional WMC tests. From left to right: stimulus-stimulus-binding, stimulus-position-binding, spatial updating.

Approval Ratings

After completing all tests, we asked all children to rate both the child-contextualized tests they worked on ("How would you rate the tests with the monsters?") and the conventional tests they worked on ("How would you rate the tests with the colored shapes?") on a five-point scale with distinct smileys indicating *very bad* to *very good*. These ratings were used to investigate whether the children had a preference regarding the contextualization of the tests.

Self-Reported Grades (German and Mathematics)

Grades of the participating children in German and mathematics were self-reported by participants. Students were asked to indicate their grades for each subject as reported in their last school year report. The possible values ranged from one to six, whereas one indicates *very good* and six indicates *poor* (c.f., the German grade system). The self-reported grades were recoded so that higher values indicate better performance, then averaged and standardized within single classes.

Fluid Intelligence

In Grades 9 and 10, we measured gf with short versions of the figural, verbal, and numerical subtests of the Berlin Test of Fluid and Crystallized Intelligence (BEFKI; see, e.g., Wilhelm et al., 2014). We used short versions due to the time limits of the study design. The figural content facet was measured with figural sequences, the numerical content facet with arithmetic text problems, and the verbal content facet with relational reasoning tests. Each subtest comprised eight items. The 16 verbal and numerical items were presented jointly with a time limit of 14 min, followed by the figural items with a time limit of 7 min. We used one score per content domain (proportion of correct responses) as indicator for the latent variable analysis.

School Achievement

For Grades 8 and 10 we derived the data from a standardized school achievement test [VERA-8 Vergleichsarbeiten (comparison tests)], which assesses skills in mathematics, German, and English (e.g., Maier & Kuper, 2012). These tests were implemented in German schools as part of a mandatory testing policy and refer to the educational standards of middle schools. The skills assessed include proficiency in mathematics, German orthography, German reading ability, English listening comprehension, and English reading ability. The criterion-referenced results are based on the educational standards, which are divided into six areas: standard 1a/1b corresponds to the lower minimum standard, Standard 2 corresponds to the minimum standard, Standard 3 and 4 correspond to normal standard, and normal standard plus, and standard five is the optimal educational standard. We used these data to test the validity of the WMC test battery.

Statistical Analyses

We estimated CFA models as described in Study 1. To bring the WMC scores from the tests for Grades 5–7 and 8–10 onto a common metric, we performed a concurrent calibration (Kolen & Brennan, 2014). Thereto, we estimated all trials jointly in a 2-parameter logistic (2PL) model and subsequently derived weighted likelihood estimator (WLE; Warm, 1989) scores for each subject and test. All item response theory analyses were performed with the R package *TAM* (Robitzsch et al., 2022). The WLE scores served as indicators in all CFA models.

Results: Study 2

We report detailed descriptive statistics of all applied tests in the supplementary materials (see SM Tables 6–9 in the supplementary materials). This also includes a comparison of test scores across test instantiations (i.e., child-contextualized vs. conventional; c.f., SM Figure 1 and SM Table 10 in the supplementary materials).

Independence of Instantiation of WMC Tests

We report a correlated factors model with two factors (Figure 4). One factor is indicated by WLE scores of the child-contextualized WMC tests, and the other is indicated by the WLE scores of the conventional WMC tests. The model fitted the data well (c.f., CFI and SRMR), but the RMSEA of the model was insufficient. This deviating result concerning model fit might be due to the fact that the RMSEA tends to be oversensitive for models with low degrees of freedom (Kenny et al., 2015). The correlation between both latent factors was not significantly different from unity (p = .121). Both factors captured substantial shares of variance (ps < .001) and showed sufficient reliabilities ($\omega_{\text{WMC Children}}$ = .74 and $\omega_{\text{WMC Conventional}}$ = .73). We allowed the test-specific residuals to correlate freely to account for test specificities. When controlling for age, the latent correlation did not change meaningfully (r = .902 vs. r = .897, with and without controlling for age, respectively). To ensure comprehensive reporting, we also present the results of a unidimensional measurement model [n = 238; $\chi^2(6)$ = 15.94, CFI = .990, RMSEA = .083, > fair dynamic fit, SRMR = .044], which – given the high correlation between

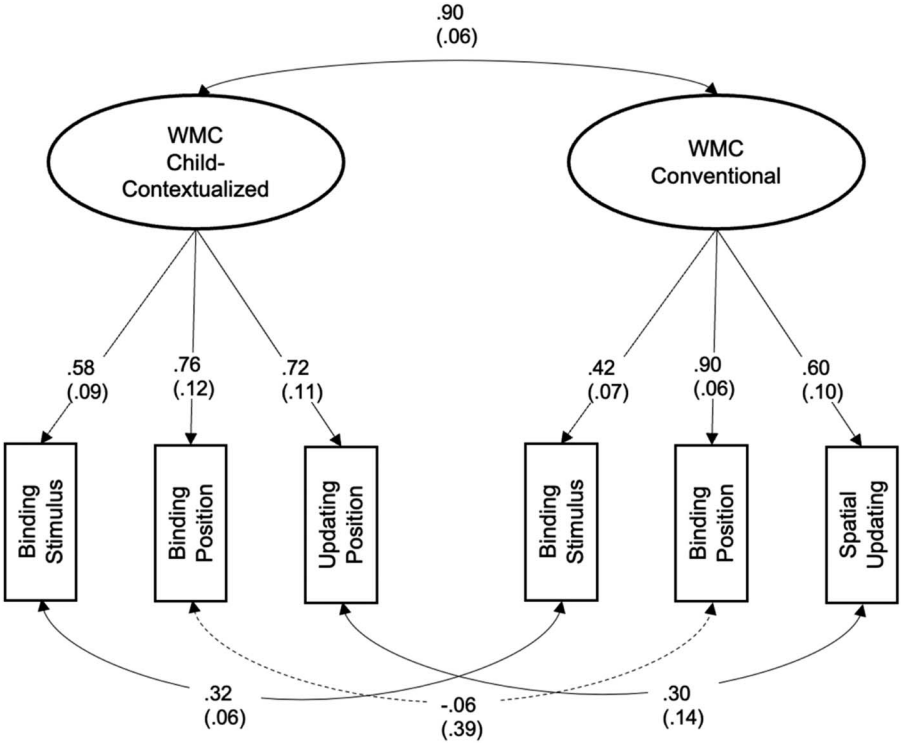

Figure 4. Correlated-factor model of working memory capacity based on child-contextualized and conventional tests. n = 238; $\chi^2(5)$ = 20.62, CFI = .979, RMSEA = .115, > fair dynamic fit, SRMR = .046. All parameters are standardized. Standard errors are given in parentheses. Dashed lines are non-significant (p > .05). For this model, observations from grades 5–8 were used (c.f., study design).

both specified factors – did not fit significantly worse [$\Delta\chi^2(1, n = 238) = .83$, $p = .362$].

The approval ratings of the child-contextualized tests were compared with the approval ratings of the conventional test versions from Grades 5 to 8. The approval ratings were not different with means of 3.18 and 3.27, respectively ($t(206) = -.98$, $p = .33$, $d = -0.07$, 95% CI [−.20, .07]). The approval ratings were correlated with one another ($r = .44$, $p < .001$), and the correlations between approval ratings and corresponding test scores ranged from $r = .04$ to .27.

Validity of WMC Tests

Validity was examined in a correlated factors model with two factors (Figure 5). One factor is indicated by the WLE scores of the conventional WMC tests, and the gf subscale scores indicate the other. The model fitted the data very well. The correlation between the two latent factors was significantly different from unity ($p = .015$). When controlling for age, the latent correlation did not change meaningfully ($r = .80$ vs. $r = .79$, with and without controlling for age, respectively). Both factors captured substantial shares of variance (all $p < .001$) and showed sufficient reliabilities ($\omega_{WMC} = .79$ and $\omega_{Gf} = .57$).

The acceptance ratings of the conventional WMC tests were compared with the acceptance ratings of the gf scales from Grades 9 and 10. The approval ratings of the conventional WMC tests were slightly better than the approval ratings of the gf scales with means of 3.15 and 2.90, respectively ($t(128) = -2.31$, $p = .02$, $d = -0.20$, 95% CI [−.38, −.03]).

Next, we report the correlations of the newly developed WMC tests with a pooled score of self-reported school grades for German and mathematics (i.e., GPA). To this end, we computed the correlations gradewise and then integrated them by means of Fisher's Z transformation. Moreover, we report the association of the WMC test battery with school achievement as indicated by the results of the VERA-8 test battery. The correlations can be found in Table 4. Please note that the cells of the covariance matrix are unequally distributed due to the planned missingness design. We report all correlations where less than 25% of the total observations were missing to avoid biased estimates.

Discussion: Study 2

The aim of Study 2 was threefold. First, we aimed to provide evidence for the applicability of our tests in subjects attending 5th to 10th grade. Second, we tested if identically parametrized WMC tests that only differed in superficial contextualization were interchangeable. With

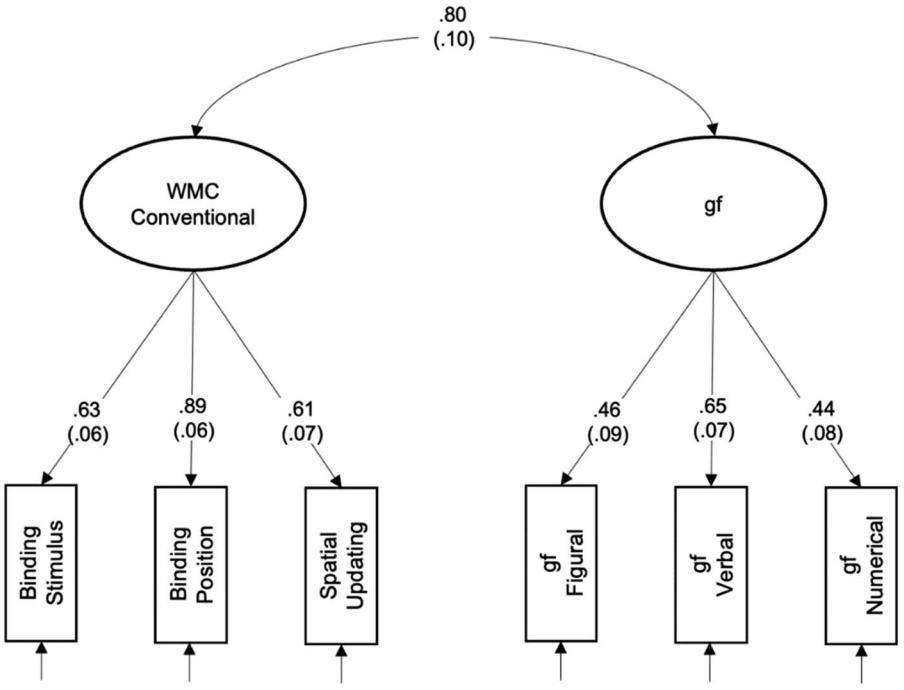

Figure 5. Correlated factors model of WMC and gf. n = 130; $\chi^2(8)$ = 7.89, CFI = 1.00, RMSEA = <.001, SRMR = .032. All parameters are standardized. Standard errors are given in parentheses. For this model, observations from Grades 9 and 10 were used (c.f., study design).

this, we tested if it is possible to switch seamlessly from child-contextualized to conventional tests. Through that, we investigated whether the tests developed in Study 1 can be used in older samples with adjusted parameterization and whether the contextualization affects the measurement of WMC. Third, we investigated the convergent validity of the WMC tests by investigating their correlation with fluid intelligence, self-reported school grades, and standardized school achievement tests to provide preliminary evidence of the nomological net of the tests.

The descriptive analyses indicated that all tests were well-parametrized, exhibiting a broad range of difficulties. In line with the results from Study 1 and the binding hypothesis, the difficulty predictably increased with increasing load levels for all administered age versions (i.e., Grades 5–7 and 8–10). Thus, the tests developed in Study 1 can readily be applied in older child samples (and adolescents) with simple adaptations of the parametrization, which we present in more detail in the supplementary materials. From a pragmatic perspective, we showed that simultaneously administering the tests to entire school classes is feasible.

To investigate the equivalence of child-contextualized and conventional versions of our WMC tests, we estimated

Table 4. Pairwise manifest correlations and corresponding 95% confidence intervals for Grades 5–7 and 8–10 between scores of WMC tests, gf, and self-reported grades (GPA), and school achievement (VERA scores)

Test context		Measures	1	2	3	4	5	6	7
Child-contextualized	1	WMC 5–7							
	2	WMC 8–10	—						
Conventional	3	WMC 5–7	.62 [.53, .70] n = 192	—					
	4	WMC 8–10	—	.68 [.48, .81] n = 45	—				
	5	gf score	—	—	—	.49 [.34, .61] n = 124			
	6	GPA	.33 [.20, .44] n = 153	—	.29 [.17, .39] n = 153	.21 [.04, .37] n = 170	.34 [.18, .48] n = 122		
	7	School achievement	—	—	—	.22 [−.02, .43] n = 69	.46 [.10, .72] n = 27	.17 [−.07, .40] n = 66	

Note. Self-reported GPA was recoded so that higher values indicate better performance.

a model in which a latent factor indicated by the three child-contextualized WMC tests correlated freely with a latent factor indicated by three conventional but equally parametrized WMC tests. As expected, the correlation was very high and did not differ significantly from unity. This finding endorses an interpretation of the equivalence of the test versions and, hence, their interchangeability. The interchangeability of surface characteristics is well-studied in the literature on automatic item generation (e.g., Gierl & Haladyna, 2013) but has received little (explicit) attention in the WMC literature. The equivalence of test contextualization is significant because a continuous parametrization of WMC tests across several age groups helps with comparability across the life span through linking and irrelevance of presented stimuli is beneficial for that. Although the approval ratings between child-contextualized and more conventional test versions did not differ substantially in the present study (i.e., participants did not prefer one version over the other), and the correlations with single test scores were relatively small (i.e., ratings are not merely a representation of how well participants performed), it might be worthwhile to investigate the potential effects of stimulus interchangeability in even younger children.

Regarding the convergent validity of our newly developed WMC tests, we provided evidence for the predicted very high correlation with gf. As well-established meta-analytically (Kane et al., 2005; Oberauer et al., 2005), a high correlation between any WMC test and any gf test can serve as a benchmark in test development for WMC tests (Oberauer et al., 2018). Our finding aligns with the literature and supports our proposal that WMC can serve as an alternative to gf. In the present design, we could not test for the correlation between child-contextualized WMC tests and gf. Given the strong correlations between child-contextualized and conventional WMC tests, we predict that similarly high correlations will be found.

The manifest correlations between the single WMC scores derived from single test batteries (i.e., Grades 5–7 vs. Grades 8–10, and child-contextualized vs. conventional) showed that the height of the associations with self-reported GPA is highly similar for both child-contextualized and for conventional WMC tests. This is in line with the above-reported finding and assumption that test materials (i.e., stimuli) can be easily swapped without substantial loss of predictive power. Additionally, the correlations with self-reported GPA were in the same ball park as the correlation between self-reported GPA and gf. Regarding the association with school achievement, the height of the correlations with WMC were clearly lower than with gf. We attribute this difference to the somewhat stronger emphasis on gc-related abilities in the measurement of school achievement because the gf test battery requires reading and comprehending single items. In contrast, the WMC test battery is language-free.

General Discussion

The impetus for the present research was twofold: First, intelligence tests for children often do not meet basic psychometric standards (Cronbach, 1949), and many test formats are not well-suited for younger children due to their lack of verbal and numerical literacy. Second, the discontinuity of test formats in assessing intelligence in children versus older individuals limits a continuous investigation from preschool to old age. Building on a comprehensive theoretical and empirical basis (Oberauer, 2019; Oberauer et al., 2005), we propose that a multivariate measure of WMC can serve as a valid indicator of general cognitive functioning and that it overcomes extant problems with the lifespan assessment of cognitive ability.

Overall, both studies provided evidence for good psychometric properties of all three WMC tests in 1st graders and also 5th to 10th graders. In the second study, child-contextualized and conventional versions of the WMC tests delivered interchangeable results. The conventional instantiation was correlated very strongly with gf. Therefore, we argue that this newly developed WMC test battery can be used as a measure of general cognitive functioning in young children and adolescents. Given the tablet-based administration mode the tests are easy to administer and can be used in group settings. They are free to use and can be readily adapted for various age levels so far up to adolescent 10th graders. Hence, they offer the possibility for both cross-sectional and longitudinal research. The following sections will recapitulate the key findings and discuss their implications and future directions.

WMC as a (Better) Measure of gf

Gf is typically considered the most prototypical form of intelligence and presumably the best single indicator of general intelligence (Carroll, 1993; Gustafsson, 1984; Kan et al., 2011). Despite this reputation, gf tests have several weaknesses. First, developing psychometrically sound gf tests remains more of an art than a science (Kyllonen & Christal, 1990). Although progress has been made in the automatic rule-based development of reasoning tasks (e.g., Gühne et al., 2020; Koch et al., 2022; Loe et al.,

2018), there is still controversy about what actually determines difficulty. Unfortunately, there seem to be several such determinants, and they vary across tests. In contrast, the very limited number of task features affecting WMC tasks' difficulty is well established (e.g., Oberauer et al., 2018). Their systematic variation allows for a reliable prediction of task difficulty, which greatly facilitates item development (including automatic item generation) and is critical for establishing construct validity. Second, most gf tests presume a minimum level of verbal literacy, numerical literacy, or general knowledge, e.g., in number series, verbal analogies, or quantitative reasoning. While this presumption is tenable in cases where literacy or knowledge holds minimal sway over one's performance in gf tasks, such an assumption may not hold for individuals with low or heterogeneous levels of education, particularly within cohorts of young children. Third, assuming that WMC is the limiting factor for reasoning ability (e.g., Oberauer, Süß, et al., 2008; Wilhelm, 2005) it is only sensible to take advantage of the above-described merits and test WMC instead of gf. Desirable attributes of WMC tests are that they should be easy to instruct, and not be susceptible to contamination from other ability constructs (e.g., mental speed). Moreover, it should be easy to generate new items with predictable difficulty for such tests through a limited number of relevant attributes all of which can be manipulated. Lastly, they should be invariant across moderator variables such as gender, age, or socioeconomic status.

Measuring WMC From Childhood to Adolescence

The present studies aimed to address challenges in measuring young children's cognitive ability by developing a newly compiled WMC test battery. Study 1 showed that the newly developed tests functioned well in 1st graders, and Study 2 supports the notion that adjusted test compilation delivers adequate tests for adolescents. Presumably, it is possible to adapt the tasks for even younger children (e.g., preschool), older adults, or participants with cognitive impairment. Pilot studies currently underway show that, in individual test settings, the tasks work well with preschool children – although this is probably the lower limit of feasibility. At the other end of the age spectrum, we propose that the test battery should also work, although empirical studies supporting this idea are not yet available. By providing a set of identical tasks for all age ranges, comprising a substantial number of linking items for robust test linking and equating (see Kolen & Brennan, 2014), we aim to build on the current work and extend our findings to even older age samples.

One prerequisite for the life span application was demonstrating the equivalence of the child-contextualized tasks with the conventional tasks. To our knowledge, the interchangeability of instantiations, also termed incidentals in the literature on automatic item generation (Irvine et al., 2002), has not been studied in WMC tests before. Study 2 provides evidence that the measurement of the underlying trait (i.e., WMC) remains unaffected by changes in the appearance of stimuli as long as the cognitive processes necessary to perform the tasks (i.e., building, maintaining, and updating bindings) remain identical. The correlation between child-contextualized stimuli (i.e., monsters with a cover story) and more conventional stimuli (i.e., colored rectangles) was very high and not statistically different from unity. Arguably, the observed deviation from unity is due to problems in balancing the groups in the within-subject design rather than substantial differences in WMC measurement.

Extending the Present Test Battery

Besides (decontextualized) reasoning ability, educational aspects are unquestionably part of individual differences in maximal cognitive effort. In consensual intelligence models, they are subsumed under the factor crystallized intelligence (gc) and reflect the cumulative outcome of learning processes (Cattell, 1987; Horn & Blankson, 2005; Schipolowski et al., 2014). Gc reflects verbal skills, such as oral and written receptive and productive abilities (Carroll, 1993), and declarative knowledge of all kinds (Cattell, 1987). In early school years, for example, Gc includes pragmatic abilities such as an elementary understanding of the number line, vocabulary, or syntax, and becomes increasingly complex and specific as education and idiosyncratic life experiences increase (e.g., vocational knowledge; Ackerman, 2000).

While developing the current WMC tests, it was a priority to make them as independent of education and prior knowledge as possible. When describing intelligence along a continuum from maximally decontextualized to maximally contextualized abilities, gc represents the opposite pole of WMC (Wilhelm & Schroeders, 2019). We argue that one should measure cognitive abilities as distinctly as possible to maximize (incremental) diagnostic information about the individuals under study (Carroll, 1993). However, individual differences in education-related abilities can be substantial even before school entry (Autorengruppe Bildungsberichterstattung [Educational Reporting Author Group], 2020) and influence subsequent educational achievement (Postlethwaite, 2011). Furthermore, crystallized abilities change fundamentally differently over the life

course than more mechanical abilities such as reasoning ability or mental speed (e.g., Baltes et al., 1999).

Consequently, we argue that gc tests could complement the highly decontextualized WMC measurement in meaningful ways. For example, children from families with low educational or low socio-economical backgrounds are likely to perform worse in gc-type competencies (e.g., Anum, 2022) but fundamentally have the cognitive potential for a successful educational career. In the early stages of development, low SES can negatively impact the relationship between gf and academic skills, which are often akin to gc. Although these negative effects can be mitigated over time through education and learning experiences (Peng et al., 2019), we argue that investigating such discrepancies or synergies requires distinct measurements of contextualized and decontextualized abilities.

Limitations and Future Directions

While our study provides valuable insights into the development and validation of a new test battery for measuring WMC in children and adolescents, there are several limitations that should be acknowledged. First, the current length of the test battery may be a disadvantage for its practical application in educational settings, depending on the sample it would be used on. However, the desire to measure highly general and essential abilities should go hand in hand with the commitment to provide adequate testing time to do so. Second, our study adheres to a specific tradition of understanding WMC, that is, through the lens of the binding hypothesis of working memory. Alternative measurement approaches are conceivable and could offer additional construct converage, although binding tests of WMC were shown to be well-representative of the construct overall (Wilhelm et al., 2013). Third, in Study 1, the absence of a broad gf test battery limited our ability to demonstrate convergent validity. Although we provided convincing evidence of shared variance with gf in Study 2, it is an empirical question whether or not these relations also hold in a sample of younger children (i.e., 1st graders). Fourth, the outcome measure of self-reported grades in Study 2 is susceptible to biases such as social desirability and recall inaccuracies, which may have weakened correlations with WMC. The VERA-8 test, although a standardized test instrument, is usually administered under conditions that might ensure consistent effort from all participants, potentially impacting the reliability of the data. Although these tests are supposed to be obligatory, many schools do not enforce standardized test situations when administering these tests. These factors could have affected the predictive validity of our findings, and future studies might address these issues to strengthen the preliminary evidence base we provided. Fifth, in Study 2, the sample that was drawn from a single school, thus limiting generalizability of our findings to the broader population of students in similar age groups. With respect to convergent validity we report, presumably range restricted samples should deliver correlational strength at the lower end of what should be found in unrestricted samples.

The cross-sectional nature of our studies leaves open the empirical question of whether the test battery introduced here offers incremental utility for education-related outcomes. Finally, the lack of administering an additional test battery for gc-related abilities in our study prevented us from examining incremental predictive validities that would have been interesting to investigate.

Despite these limitations, we believe our newly developed tests hold significant potential for further research and possibly even cross-national research. The audio instructions can easily be translated and exchanged, and the tablet-based offline administration facilitates the distribution of test materials in areas with limited technological infrastructure. Moreover, the tests can also be administered on stationary computers, making them versatile tools for a variety of research settings. For applications such as diagnosing intellectual giftedness, clinical disorders, or school readiness, comprehensive norm bases will be essential. While age- or year-based norms might suffice for older children and adults, continuous norms will be particularly important for young children due to their rapid cognitive development (e.g., Lenhard et al., 2019).

Conclusion

Taken together, we developed a multivariate WMC test battery for children and adolescent students, allowing for a valid assessment of cognitive ability across a broad age range. The tests can easily be administered in group settings from 1st grade onwards. Child-contextualized and conventional tests are interchangeable, laying the groundwork for extensions. The test battery is freely available for interested researchers.

References

Ackerman, P. L. (2000). Domain-specific knowledge as the "dark matter" of adult intelligence: Gf/Gc, personality and interest correlates. *Journal of Gerontology*, *55*(2), 69–84. https://doi.org/10.1093/geronb/55.2.P69

Ahmed, S. F., Ellis, A., Ward, K. P., Chaku, N., & Davis-Kean, P. E. (2022). Working memory development from early childhood to

adolescence using two nationally representative samples. *Developmental Psychology*, *58*(10), 1962–1973. https://doi.org/10.1037/dev0001396

Alloway, T. P., & Alloway, R. G. (2010). Investigating the predictive roles of working memory and IQ in academic attainment. *Journal of Experimental Child Psychology*, *106*(1), 20–29. https://doi.org/10.1016/j.jecp.2009.11.003

Anum, A. (2022). Does socio-economic status have different impact on fluid and crystallized abilities? Comparing scores on Raven's progressive matrices, Kaufman assessment battery for Children II Story Completion and Kilifi Naming Test Among Children in Ghana. *Frontiers in Psychology*, *13*, Article 880005. https://doi.org/10.3389/fpsyg.2022.880005

Arendasy, M. (2005). Automatic generation of Rasch-calibrated items: Figural matrices test GEOM and Endless-Loops test EC. *International Journal of Testing*, *5*(3), 197–224. https://doi.org/10.1207/s15327574ijt0503_2

Autorengruppe Bildungsberichterstattung [Educational Reporting Author Group]. (2020). *Bildung in Deutschland 2020: Ein indikatorengestützter Bericht mit einer Analyse zu Bildung in einer digitalisierten Welt* [Education in Germany 2020: An indicator-based report with an analysis of education in a digitalized world]. wbv Media. https://www.bildungsbericht.de/de/bildungsberichte-seit-2006/bildungsbericht-2020/pdf-dateien-2020/bildungsbericht-2020-barrierefrei.pdf

Baltes, P. B., Staudinger, U. M., & Lindenberger, U. (1999). Lifespan psychology: Theory and application to intellectual functioning. *Annual Review of Psychology*, *50*(1), 471–507. https://doi.org/10.1146/annurev.psych.50.1.471

Bäuerlein, K., Beinicke, A., Schorr, M., & Schneider, W. (2021). *Fähigkeitsindikatoren Primarschule (FIPS). Ein digitales Testverfahren zur Erfassung der Lernausgangslage und der Lernentwicklung in der 1. Klasse* [Ability Indicators Primary School (FIPS). A digital measurement instrument for assessing the learning situation and learning development in the 1st grade]. Hogrefe.

Becker, N., Preckel, F., Karbach, J., Raffel, N., & Spinath, F. M. (2015). Die Matrizenkonstruktionsaufgabe: Validierung eines distraktorfreien Aufgabenformats zur Vorgabe figuraler Matrizen [The matrix construction task: Validation of a distractor-free task format for the specification of figural matrices]. *Diagnostica*, *61*(1), 22–33. https://doi.org/10.1026/0012-1924/a000111

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*(2), 238–246. https://doi.org/10.1037/0033-2909.107.2.238

Blum, D., Holling, H., Galibert, M. S., & Forthmann, B. (2016). Task difficulty prediction of figural analogies. *Intelligence*, *56*, 72–81. https://doi.org/10.1016/j.intell.2016.03.001

Blume, F., Irmer, A., Dirk, J., & Schmiedek, F. (2022). Day-to-day variation in students' academic success: The role of self-regulation, working memory, and achievement goals. *Developmental Science*, *25*(6), Article e13301. https://doi.org/10.1111/desc.13301

Bradley, R. H., & Corwyn, R. F. (2002). Socioeconomic status and child development. *Annual Review of Psychology*, *53*, 371–399. https://doi.org/10.1146/annurev.psych.53.100901.135233

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research*, *21*(2), 230–258. https://doi.org/10.1177/0049124192021002005

Brunner, M., Nagy, G., & Wilhelm, O. (2012). A tutorial on hierarchically structured constructs: Hierarchically structured constructs. *Journal of Personality*, *80*(4), 796–846. https://doi.org/10.1111/j.1467-6494.2011.00749.x

Cabbage, K., Brinkley, S., Gray, S., Alt, M., Cowan, N., Green, S., Kuo, T., & Hogan, T. P. (2017). Assessing working memory in children: The Comprehensive Assessment Battery for Children – Working Memory (CABC-WM). *Journal of Visualized Experiments*, *124*, Article 55121. https://doi.org/10.3791/55121

Cattell, R. B. (1971). *Abilities: Their structure, growth, and action.* Houghton Mifflin Company.

Cattell, R. B. (1987). *Intelligence: Its structure, growth and action*. North Holland.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge University Press. https://doi.org/10.1017/CBO9780511571312

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(3), 464–504. https://doi.org/10.1080/10705510701301834

Cheng, C., & Kibbe, M. M. (2022). Development of updating in working memory in 4–7-year-old children. *Developmental Psychology*, *58*(5), 902–912. https://doi.org/10.1037/dev0001337

Cirino, P. T., Ahmed, Y., Miciak, J., Taylor, W. P., Gerst, E. H., & Barnes, M. A. (2018). A framework for executive function in the late elementary years. *Neuropsychology*, *32*(2), 176–189. https://doi.org/10.1037/neu0000427

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.n). L. Erlbaum Associates.

Conway, A. R. A., Jarrold, C., Kane, M. J., Miyake, A., & Towse, J. (2008). Variation in working memory: An introduction. In A. R. A. Conway, C. Jarrold, M. Kane, A. Miyake, & J. Towse (Eds.), *Variation in working memory* (1st ed., pp. 3–18). Oxford University Press.

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

Cowan, N., AuBuchon, A. M., Gilchrist, A. L., Ricker, T. J., & Saults, J. S. (2011). Age differences in visual working memory capacity: Not based on encoding limitations: Differences in working memory capacity. *Developmental Science*, *14*(5), 1066–1074. https://doi.org/10.1111/j.1467-7687.2011.01060.x

Cronbach, L. J. (1949). *Essentials of psychological testing*. Harper & Brothers, Publishers.

Demetriou, A., Spanoudis, G., Shayer, M., van der Ven, S., Brydges, C. R., Kroesbergen, E., Podjarny, G., & Swanson, H. L. (2014). Relations between speed, working memory, and intelligence from preschool to adulthood: Structural equation modeling of 14 studies. *Intelligence*, *46*, 107–121. https://doi.org/10.1016/j.intell.2014.05.013

Dirk, J., & Schmiedek, F. (2016). Fluctuations in elementary school children's working memory performance in the school context. *Journal of Educational Psychology*, *108*(5), 722–739. https://doi.org/10.1037/edu0000076

DiStefano, C., & Zhang, T. (2022). A primer for using multilevel confirmatory factor analysis models in educational research. In M. S. Khine (Ed.), *Methodology for multilevel modeling in educational research* (pp. 11–28). Springer Singapore. https://doi.org/10.1007/978-981-16-9142-3_2

Dixon, L. Q., Wu, S., & Daraghmeh, A. (2012). Profiles in bilingualism: Factors influencing kindergartners' language proficiency. *Early Childhood Education Journal*, *40*(1), 25–34. https://doi.org/10.1007/s10643-011-0491-8

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., Pagani, L. S., Feinstein, L., Engel, M., Brooks-Gunn, J., Sexton, H., Duckworth, K., & Japel, C. (2007). School readiness and later achievement. *Developmental Psychology*, *43*(6), 1428–1446. https://doi.org/10.1037/0012-1649.43.6.1428

Elliott, J., & Shepherd, P. (2006). Cohort profile: 1970 British birth cohort (BCS70). *International Journal of Epidemiology*, *35*(4), 836–843. https://doi.org/10.1093/ije/dyl174

Enders, C. K. (2010). *Applied missing data analysis*. Guilford Press.

Engle, R. W. (2002). Working memory capacity as executive attention. *Current Directions in Psychological Science*, *11*(1), 19–23. 10.1111/1467-8721.00160

Esser, G., Wyschkorn, A., & Ballaschk, K. (2021). *Basisdiagnostik Umschriebener Entwicklungsstörungen im Grundschulalter* [*Basic diagnostics of specific developmental disorders in elementary school*]. Hogrefe.

Galeano-Keiner, E. M., Neubauer, A. B., Irmer, A., & Schmiedek, F. (2022). Daily fluctuations in children's working memory accuracy and precision: Variability at multiple time scales and links to daily sleep behavior and fluid intelligence. *Cognitive Development*, *64*, Article 101260. https://doi.org/10.1016/j.cogdev.2022.101260

Gierl, M. J., & Haladyna, T. M. (Eds.). (2013). *Automatic item generation: Theory and practice*. Routledge.

Goecke, B., Schmitz, F., & Wilhelm, O. (2021). Binding costs in processing efficiency as determinants of cognitive ability. *Journal of Intelligence*, *9*(2), Article 18. https://doi.org/10.3390/jintelligence9020018

Gühne, D., Doebler, P., Condon, D. M., Luo, F., & Sun, L. (2020). Validity and reliability of automatically generated propositional reasoning items. *European Journal of Psychological Assessment*, *37*(4), 325–339. https://doi.org/10.1027/1015-5759/a000616

Goecke, B., Zimny, L., Hartung, J., Wilhelm, O., & Golle, J. (2024). *WMC Kids* [Open data]. https://osf.io/evfsd/

Goldhammer, F. (2015). Measuring ability, speed, or both? Challenges, psychometric solutions, and what can be gained from experimental control. *Measurement: Interdisciplinary Research and Perspectives*, *13*(3–4), 133–164. https://doi.org/10.1080/15366367.2015.1100020

Goldhammer, F., & Klein Entink, R. H. (2011). Speed of reasoning and its relation to reasoning ability. *Intelligence*, *39*(2–3), 108–119. https://doi.org/10.1016/j.intell.2011.02.001

Gottfredson, L. S., & Deary, I. J. (2004). Intelligence predicts health and longevity, but why? *Current Directions in Psychological Science*, *13*(1), 1–4. https://doi.org/10.1111/j.0963-7214.2004.01301001.x

Gray, S., Green, S., Alt, M., Hogan, T., Kuo, T., Brinkley, S., & Cowan, N. (2017). The structure of working memory in young children and its relation to intelligence. *Journal of Memory and Language*, *92*, 183–201. https://doi.org/10.1016/j.jml.2016.06.004

Groskurth, K., Bluemke, M., & Lechner, C. M. (2023). Why we need to abandon fixed cutoffs for goodness-of-fit indices: An extensive simulation and possible solutions. *Behavior Research Methods*, *56*, 3891–3914. https://doi.org/10.3758/s13428-023-02193-3

Gustafsson, J.-E. (1984). A unifying model for the structure of intellectual abilities. *Intelligence*, *8*(3), 179–203. https://doi.org/10.1016/0160-2896(84)90008-4

Gustafsson, J.-E., & Wolff, U. (2015). Measuring fluid intelligence at age four. *Intelligence*, *50*, 175–185. https://doi.org/10.1016/j.intell.2015.04.008

Hartung, J., Goecke, B., Schroeders, U., Schmitz, F., & Wilhelm, O. (2022). Latin square tasks: A multi-study evaluation. *Intelligence*, *94*, Article 101683. https://doi.org/10.1016/j.intell.2022.101683

Horn, J. L., & Blankson, N. (2005). Foundations for better understanding of cognitive abilities. In Flanagan, D. P. & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 41–68). Guilford Press.

Houwen, S., Kamphorst, E., van der Veer, G., & Cantell, M. (2019). Identifying patterns of motor performance, executive functioning, and verbal ability in preschool children: A latent profile analysis. *Research in Developmental Disabilities*, *84*, 3–15. https://doi.org/10.1016/j.ridd.2018.04.002

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Inquisit 6 (Version 6.5.2). (2022). [Computer software]. Millisecond Software. https://www.millisecond.com

Irvine, S. H., Kyllonen, P. C., Air Force Human Resources Laboratory, & Educational Testing Service (Eds.). (2002). *Item generation for test development*. L. Erlbaum Associates.

Johnson-Laird, P. N. (1980). Mental models in cognitive science. *Cognitive Science*, *4*(1), 71–115. https://doi.org/10.1207/s15516709cog0401_4

Kan, K.-J., Kievit, R. A., Dolan, C., & van der Maas, H. (2011). On the interpretation of the CHC factor Gc. *Intelligence*, *39*(5), 292–302. https://doi.org/10.1016/j.intell.2011.05.003

Kane, M. J., Hambrick, D. Z., & Conway, A. R. A. (2005). Working memory capacity and fluid intelligence are strongly related constructs: Comment on Ackerman, Beier, and Boyle (2005). *Psychological Bulletin*, *131*(1), 66–71. https://doi.org/10.1037/0033-2909.131.1.66

Kenny, D. A., Kaniskan, B., & McCoach, D. B. (2015). The performance of RMSEA in models with small degrees of freedom. *Sociological Methods & Research*, *44*(3), 486–507. https://doi.org/10.1177/0049124114543236

Koch, M., Spinath, F. M., Greiff, S., & Becker, N. (2022). Development and validation of the open matrices item bank. *Journal of Intelligence*, *10*(3), 1–10. https://doi.org/10.3390/jintelligence10030041

Kolen, M. J., & Brennan, R. L. (Eds.). (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). Springer. https://doi.org/10.1007/978-1-4939-0317-7

Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (little more than) working-memory capacity? *Intelligence*, *14*(4), 389–433. https://doi.org/10.1016/S0160-2896(05)80012-1

Lan, X., Legare, C. H., Ponitz, C. C., Li, S., & Morrison, F. J. (2011). Investigating the links between the subcomponents of executive function and academic achievement: A cross-cultural analysis of Chinese and American preschoolers. *Journal of Experimental Child Psychology*, *108*(3), 677–692. https://doi.org/10.1016/j.jecp.2010.11.001

Larson, G. E., Merritt, C. R., & Williams, S. E. (1988). Information processing and intelligence: Some implications of task complexity. *Intelligence*, *12*(2), 131–147. https://doi.org/10.1016/0160-2896(88)90012-8

Lee, V. E., & Burkam, D. T. (2002). *Inequality at the starting gate: Social background differences in achievement as children begin school*. Economic Policy Institute.

Lenhard, A., Lenhard, W., & Gary, S. (2019). Continuous norming of psychometric tests: A simulation study of parametric and semi-parametric approaches. *PLoS ONE*, *14*(9), Article e0222279. https://doi.org/10.1371/journal.pone.0222279

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

Loe, B. S., & Rust, J. (2019). The Perceptual Maze Test revisited: Evaluating the difficulty of automatically generated mazes. *Assessment*, *26*(8), 1524–1539. https://doi.org/10.1177/1073191117746501

Loe, B. S., Sun, L., Simonfy, F., & Doebler, P. (2018). Evaluating an automated number series item generator using linear logistic test models. *Journal of Intelligence*, *6*(2), Article 20. https://doi.org/10.3390/jintelligence6020020

Lösche, P., Wiley, J., & Hasselhorn, M. (2015). How knowing the rules affects solving the Raven advanced progressive matrices test. *Intelligence*, *48*, 58–75. https://doi.org/10.1016/j.intell.2014.10.004

Maier, U., & Kuper, H. (2012). Vergleichsarbeiten als Instrumente der Qualitätsentwicklung an Schulen [Comparative work as an instrument for quality development in schools]. *DDS–Die Deutsche Schule*, *104*(1), 88–99. https://doi.org/10.25656/01:25723

McDonald, R. P. (1999). *Test theory: A unified treatment*. Erlbaum.

McNeish, D., & Wolf, M. G. (2023). *Direct discrepancy dynamic fit index cutoffs for arbitrary covariance structure models* [Preprints]. PsyArXiv. https://doi.org/10.31234/osf.io/4r9fq

Miyake, A., & Friedman, N. P. (2012). The nature and organization of individual differences in executive functions: Four general conclusions. *Current Directions in Psychological Science*, *21*(1), 8–14. https://doi.org/10.1177/0963721411429458

Neubauer, A. B., Dirk, J., & Schmiedek, F. (2019). Momentary working memory performance is coupled with different dimensions of affect for different children: A mixture model analysis of ambulatory assessment data. *Developmental Psychology*, *55*(4), 754–766. https://doi.org/10.1037/dev0000668

Oberauer, K. (2005). Binding and inhibition in working memory: Individual and age differences in short-term recognition. *Journal of Experimental Psychology: General*, *134*(3), 368–387. https://doi.org/10.1037/0096-3445.134.3.368

Oberauer, K. (2019). Working memory capacity limits memory for bindings. *Journal of Cognition*, *2*(1), Article 40. https://doi.org/10.5334/joc.86

Oberauer, K., Lewandowsky, S., Awh, E., Brown, G. D. A., Conway, A., Cowan, N., Donkin, C., Farrell, S., Hitch, G. J., Hurlstone, M. J., Ma, W. J., Morey, C. C., Nee, D. E., Schweppe, J., Vergauwe, E., & Ward, G. (2018). Benchmarks for models of short-term and working memory. *Psychological Bulletin*, *144*(9), 885–958. https://doi.org/10.1037/bul0000153

Oberauer, K., Schulze, R., Wilhelm, O., & Süß, H.-M. (2005). Working memory and intelligence – their correlation and their relation: Comment on Ackerman, Beier, and Boyle (2005). *Psychological Bulletin*, *131*(1), 61–65. https://doi.org/10.1037/0033-2909.131.1.61

Oberauer, K., Süß, H.-M., Wilhelm, O., & Sander, N. (2008). Individual differences in working memory capacity and reasoning ability. In A. R. A. Conway, C. Jarrold, M. Kane, A. Miyake, & J. Towse (Eds.), *Variation in working memory* (pp. 49–75). Oxford University Press.

Oberauer, K., Süß, H.-M., Wilhelm, O., & Wittmann, W. W. (2008). Which working memory functions predict intelligence? *Intelligence*, *36*(6), 641–652. https://doi.org/10.1016/j.intell.2008.01.007

Panesi, S., Bandettini, A., Traverso, L., & Morra, S. (2022). On the Relation between the development of working memory updating and working memory capacity in preschoolers. *Journal of Intelligence*, *10*(1), Article 5. https://doi.org/10.3390/jintelligence10010005

Peng, P., & Fuchs, D. (2017). A randomized control trial of working memory training with and without strategy instruction: Effects on young children's working memory and comprehension. *Journal of Learning Disabilities*, *50*(1), 62–80. https://doi.org/10.1177/0022219415594609

Peng, P., & Kievit, R. A. (2020). The development of academic achievement and cognitive abilities: A bidirectional perspective. *Child Development Perspectives*, *14*(1), 15–20. https://doi.org/10.1111/cdep.12352

Peng, P., Wang, T., Wang, C., & Lin, X. (2019). A meta-analysis on the relation between fluid intelligence and reading/mathematics: Effects of tasks, age, and social economics status. *Psychological Bulletin*, *145*(2), 189–236. https://doi.org/10.1037/bul0000182

Phillips, H., & Rabbitt, P. M. A. (1995). Lmpulsivity and speed-accuracy strategies in intelligence test performance. *Intelligence*, *21*(1), 13–29. doi:10.1016/0160-2896(95)90036-5

Postlethwaite, B. E. (2011). *Fluid ability, crystallized ability, and performance across multiple domains: A meta-analysis* [Unpublished doctoral dissertation]. University of Iowa. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.829.963&rep=rep1&type=pdf

Primi, R. (2001). Complexity of geometric inductive reasoning tasks Contribution to the understanding of fluid intelligence. *Intelligence*, *30*(1), 41–70. doi:10.1016/S0160-2896(01)00067-8

R Core Team. (2024). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Redick, T. S., Broadway, J. M., Meier, M. E., Kuriakose, P. S., Unsworth, N., Kane, M. J., & Engle, R. W. (2012). Measuring working memory capacity with automated complex span tasks. *European Journal of Psychological Assessment*, *28*(3), 164–171. https://doi.org/10.1027/1015-5759/a000123

Renner, E., Somai, R. S., Van der Stigchel, S., Campbell, C., Kean, D., & Caldwell, C. A. (2021). Adaptation of the missing scan task to a touchscreen format for assessing working memory capacity in children. *Infant and Child Development*, *30*(6), Article e2277. https://doi.org/10.1002/icd.2277

Revelle, W. (2024). *psych: Procedures for psychological, psychometric, and personality research (R package Version 2.4.3)* [Computer software]. https://CRAN.R-project.org/package=psych

Reynolds, M. R., Niileksela, C. R., Gignac, G. E., & Sevillano, C. N. (2022). Working memory capacity development through childhood: A longitudinal analysis. *Developmental Psychology*, *58*(7), 1254–1263. https://doi.org/10.1037/dev0001360

Rhemtulla, M., & Little, T. D. (2012). Planned missing data designs for research in cognitive development. *Journal of Cognition and Development*, *13*(4), 425–438. https://doi.org/10.1080/15248372.2012.717340

Robitzsch, A., Kiefer, T., & Wu, M. (2022). *TAM: Test analysis modules (Version R package version 4.1-4)* [Computer software]. https://CRAN.R-project.org/package=TAM

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

Roth, B., Becker, N., Romeyke, S., Schäfer, S., Domnick, F., & Spinath, F. M. (2015). Intelligence and school grades: A meta-analysis. *Intelligence*, *53*, 118–137. https://doi.org/10.1016/j.intell.2015.09.002

Sánchez-Izquierdo, M., Fernández-Ballesteros, R., Valeriano-Lorenzo, E. L., & Botella, J. (2023). Intelligence and life expectancy in late adulthood: A meta-analysis. *Intelligence*, *98*, Article 101738. https://doi.org/10.1016/j.intell.2023.101738

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*(2), 147–177. https://doi.org/10.1037/1082-989X.7.2.147

Schipolowski, S., Wilhelm, O., & Schroeders, U. (2014). On the nature of crystallized intelligence: The relationship between verbal ability and factual knowledge. *Intelligence*, *46*, 156–168. https://doi.org/10.1016/j.intell.2014.05.014

Schroeders, U., & Gnambs, T. (2020). Degrees of freedom in multigroup confirmatory factor analyses: Are models of measurement invariance testing correctly specified? *European Journal of Psychological Assessment*, *36*(1), 105–113. https://doi.org/10.1027/1015-5759/a000500

Shi, D., DiStefano, C., Maydeu-Olivares, A., & Lee, T. (2022). Evaluating SEM model fit with small degrees of freedom. *Multivariate Behavioral Research*, *57*(2–3), 179–207. https://doi.org/10.1080/00273171.2020.1868965

Spinath, B., Spinath, F. M., Harlaar, N., & Plomin, R. (2006). Predicting school achievement from general cognitive ability, self-perceived ability, and intrinsic value. *Intelligence*, *34*(4), 363–374. https://doi.org/10.1016/j.intell.2005.11.004

Tymms, P., Merrell, C., & Hawker, D. (2014). *Performance indicators in primary schools: A comparison of performance on entry to school and the progress made in the first year in England and four other jurisdictions: Research Report*. Department of Education, UK. https://www.gov.uk/government/publications/performance-indicators-in-primary-schools

Undheim, J. O., & Gustafsson, J.-E. (1987). The Hierarchical organization of cognitive abilities: Restoring general intelligence through the use of linear structural relations (LISREL). *Multivariate Behavioral Research*, *22*(2), 149–171. https://doi.org/10.1207/s15327906mbr2202_2

Unsworth, N., & Engle, R. W. (2007). The nature of individual differences in working memory capacity: Active maintenance in primary memory and controlled search from secondary memory. *Psychological Review*, *114*(1), 104–132. https://doi.org/10.1037/0033-295X.114.1.104

Unsworth, N., Heitz, R. P., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. *Behavior Research Methods*, *37*(3), 498–505. https://doi.org/10.3758/BF03192720

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*(1), 4–69. https://doi.org/10.1177/109442810031002

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*(3), 427–450. https://doi.org/10.1007/BF02294627

Watrin, L., Hülür, G., & Wilhelm, O. (2022). Training working memory for two years – no evidence of transfer to intelligence. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *48*(5), 717–733. https://doi.org/10.1037/xlm0001135

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., . . . Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43), Article 1686. https://doi.org/10.21105/joss.01686

Wilhelm, O. (2005). Measuring reasoning ability. In O. Wilhelm & R. W. Engle (Eds.), *Handbook of understanding and measuring intelligence* (pp. 373–392). Sage. https://doi.org/10.4135/9781452233529.n21

Wilhelm, O., Hildebrandt, A., & Oberauer, K. (2013). What is working memory capacity, and how can we measure it? *Frontiers in Psychology*, *4*, Article 433. https://doi.org/10.3389/fpsyg.2013.00433

Wilhelm, O., & Schroeders, U. (2019). Intelligence. In R. Sternberg & J. Funke (Eds.), *The psychology of human thought: An introduction* (pp. 257–277). Heidelberg University Publishing.

Wilhelm, O., Schroeders, U., & Schipolowski, S. (2014). *Berliner Test zur Erfassung fluider und kristalliner Intelligenz für die 8. Bis 10. Jahrgangsstufe (BEFKI 8-10)* [Berlin Test for the Assessment of Fluid and Crystallized Intelligence for Grades 8 to 10 (BEFKI 8-10)]. Hogrefe.

Wolf, M. G., & McNeish, D. (2023). dynamic: An R package for deriving dynamic fit index cutoffs for factor analysis. *Multivariate Behavioral Research*, *58*(1), 189–194. https://doi.org/10.1080/00273171.2022.2163476

Zhang, X., Räsänen, P., Koponen, T., Aunola, K., Lerkkanen, M.-K., & Nurmi, J.-E. (2017). Knowing, applying, and reasoning about arithmetic: Roles of domain-general and numerical skills in multiple domains of arithmetic learning. *Developmental Psychology*, *53*(12), 2304–2318. https://doi.org/10.1037/dev0000432

History
Received May 10, 2024
Revision received September 19, 2024
Accepted October 9, 2024
Published online November 26, 2024

Section: Intelligence

Acknowledgments
The authors thank numerous student assistants for supporting this work, and especially Annika Wilke. We would also like to thank the schools for their support in data collections, with a special thank you to Dr. Martin Böhnisch.

Conflict of Interest
The authors have no conflicts of interest to disclose.

Publication Ethics
The authors confirm that the work conforms to Standard 8 of the American Psychological Association's Ethical Principles of Psychologists and Code of Conduct.

Study 1: The anonymity of participants was guaranteed, and the study was conducted in accordance with the Declaration of Helsinki and was approved by a local ethics committee. Participation was voluntary and could be canceled at any time.

Study 2. The study was conducted according to the declaration of Helsinki. Informed consent from the parents was obtained prior to the study. Participation was voluntary and could be canceled at any time.

Open Science
Open Analytic Code and Open Materials: Newly developed tests, materials, data and code, and supplementary materials are available at https://osf.io/evfsd/ and https://github.com/luc-w/wmc-ulkabe.

Open Data: The authors confirm that there is sufficient information for an independent researcher to reproduce all of the reported results, including codebook if relevant (Goecke et al., 2024).

Open Materials: The authors confirm that there is sufficient information for an independent researcher to reproduce all of the reported methodology (Goecke et al., 2024).

Open Analytic Code: The authors confirm that all the scripts, code, and outputs needed to reproduce the reported analyses are provided (Goecke et al., 2024).

Preregistration: Study 1 was not preregistered.

Funding
The work reported herein was supported by grants from the Hector Foundation II. The authors acknowledge support from the Open Access Fund of the University of Tübingen.

ORCID
Benjamin Goecke
https://orcid.org/0000-0002-3050-1848
Luc Zimny
https://orcid.org/0000-0003-4343-3781
Johanna Hartung
https://orcid.org/0000-0002-6392-4468
Patrick Lösche
https://orcid.org/0000-0001-9346-4223
Jessika Golle
https://orcid.org/0000-0002-8507-7079
Oliver Wilhelm
https://orcid.org/0000-0001-7980-1166

Benjamin Goecke
Hector Research Institute of Education Sciences and Psychology
University of Tübingen
Walter-Simon-Straße 12
72072 Tübingen
Germany
academ@benjamin-goecke.de