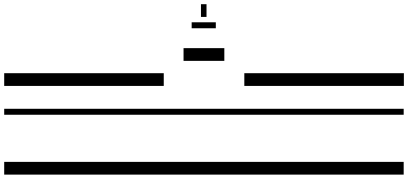


Westfälische
Wilhelms-Universität
Münster

Psychologie

Reasoning ability and working memory capacity in children

Inaugural-Dissertation
zur Erlangung des Doktorgrades
der
Philosophischen Fakultät
der
Westfälischen Wilhelms-Universität
zu Münster (Westf.)

vorgelegt von

Jörg-Tobias Kuhn

aus Münster (Westf.)

Dezember 2009

Tag der mündlichen Prüfung: 22.01.2010

Dekan: Prof. Dr. Pietsch
Referent: Prof. Dr. Holling
Korreferent: Prof. Dr. Dr. h. c. Schäfer

To Penelope
New roads emerge, but well-known paths always remain

Contents

1 Introduction

From its beginning, psychological science has been concerned with measuring and understanding mental processes and structures. In this context, one aspect of the human mind is especially noteworthy: The ability to maintain information in an active and accessible state, while simultaneously processing selective new information. Working memory (WM) is the term that cognitive psychologists use to describe this ability. One fundamental characteristic of WM is that it has a limited capacity. That is, only a limited amount of information can be kept in a readily accessible state. Converging evidence suggests that on average, about four separate pieces of information can be stored in WM by healthy adults (Cowan, 2001). Importantly, individuals differ in their WM capacity, such that some individuals are generally better than others at performing demanding cognitive tasks, such as complex learning, reading comprehension, or mathematical problem solving. The importance of WM for higher cognition has been soundly established within the experimental, neuropsychological, and individual differences research literature (e.g., Ackerman, Beier, & Boyle, 2005; Kane & Engle, 2002; Kyllonen & Christal, 1990). Although WM and general intelligence are highly related, they can still be separated on empirical and theoretical grounds (Ackerman et al., 2005; Blair, 2006).

WM is also highly related to reasoning. In all theories of intelligence structure, reasoning represents a key construct. In the definition of intelligence by Spearman (1923), educing correlates and relations, which is best reflected in reasoning measures, is of central importance. Further, reasoning measures, of all cognitive ability tests, have been shown to exhibit the highest loadings on the g factor, which is regarded as the core construct of human cognitive abilities (Carroll, 1993). Reasoning tasks, however, come in a wide variety of forms. Wilhelm (2005) offers several classification aspects of reasoning, including formal operational requirement (e.g., deductive vs. inductive reasoning), task content (e.g., numerical or verbal content), instantiation of reasoning problems (e.g., abstract vs. concrete) and vulnerability to differential strategy use. Wilhelm (2005) could show that a latent variable model based on different task content factors, all load-

ing on a higher-order g factor, provided satisfactory fit, whereas deductive and inductive reasoning factors could not be differentiated.

Although several studies with numerous measures of WM and reasoning facets have been conducted in adult populations, providing insight into the relationships of WM with reasoning, processing speed, short-term memory (STM), and other relevant factors (e.g., Colom, Abad, Quiroga, Shih, & Flores-Mendoza, 2008; Kane et al., 2004; Krumm et al., 2009), much less activity has been devoted to investigating these relationships in children. In addition, those studies conducted with children often opted for a specific research paradigm, which was either experimental or differential. Ever since Cronbach (1957), it has been argued that a combination of both approaches is necessary to obtain a more complete picture of underlying processes and structures. This thesis therefore is concerned with investigating the relationship between WM and reasoning in children using latent variable models that allow for a combination of experimental and differential perspectives.

1.1 Models of WM

Among the numerous theories of WM (cf. Miyake & Shah, 1999), one of the most influential and widely-recognized models is the multiple-component model suggested by Baddeley and colleagues (e.g., Baddeley, 1986; Baddeley & Hitch, 1974). WM, in this model, is assumed to have two temporary memory systems, the phonological loop for speech-based information and the visuo-spatial sketchpad for spatial information. These two components are used to maintain memory traces in an active state using rehearsal. In addition, the multiple-component model postulates a central executive which functions as a control and regulation instance in the WM system. The central executive is not involved in temporary storage, but rather coordinates the activity of the slave systems by processes such as activation of long-term memory traces or switching attention. Further, the central executive is responsible for manipulating material held in the temporary storage systems. The multiple-component model has been repeatedly shown to explain a wide range of experimental and individual differences data, both in children and adult studies, although some modifications were lately introduced (Baddeley, 2000). One of the key aspects of the multiple-component model is the differentiation of WM into functionally different modules, along with the separation of storage and processing components.

In contrast, the WM model introduced by Engle and colleagues (e.g., Engle, Tuholski, Laughlin, & Conway, 1999; Kane, Conway, Bleckley, & Engle, 2001) is more process-oriented than the multiple-component model. A central aspect of the perspective held by Engle et al. is the assumption that a domain-free, limited-capacity resource, controlled attention, lies at the heart of WM. These authors assume, similar to the multiple-component model, that there are domain-specific codes and maintenance processes (STM), whereas the ability to simultaneously store and process information is domain-general. As shown in Engle et al. (1999), using structural equation modeling, a WM factor was related to a reasoning factor even when controlling for short-term memory (i.e., storage) variance, whereas STM was unrelated to reasoning once controlled attention was partialled. Hence, controlled attention is seen as the key resource for higher intellectual functioning.

This view was recently challenged by Colom, Rebollo, Abad, and Shih (2006), who reanalyzed several data sets to shed light on the relevance of STM in predicting intellectual abilities. These authors found that when conceptualizing controlled attention as a residual factor, capturing only variance in WM tasks, and when STM was modeled as a factor capturing variance in both WM and short-term memory tasks, both WM and STM were relevant for predicting intelligence. According to Colom et al. (2006, p. 167), "both measures [WM and STM] share something in common that could produce their *association* with cognitive ability measures". Later, Unsworth and Engle (2007a) suggested a framework that differentiates between primary and secondary memory instead of WM and STM. Primary memory serves to keep a number of separate representations active for processing by continuously allocating attention. Access to contents stored in secondary memory, however, requires a cue-dependent search process. In contrast to STM tasks of the same list length, WM tasks measure secondary memory rather than primary memory. Hence, low WM capacity can stem from two sources, an inability to distinctly maintain representations in primary memory, or rather an inability to effectively search for stored representations in secondary memory. As soon as search processes in secondary memory are required, for example, by using a STM task with high list length, the relationship with higher cognitive functioning increases substantially (Unsworth & Engle, 2007b).

Another model of WM, the embedded-processes model, was suggested by Cowan (e.g., Cowan et al., 2005). In this view, WM consists of hierarchically arranged entities: Long-term memory, the subset of activated long-term memory, and the subset of activated long-memory that is in the focus of attention. The

focus of attention is capacity-limited (Cowan, 2001). As Cowan et al. (2005) mention, in addition to the voluntary control of attention, the scope of attention is a decisive factor. In the view of Cowan, simultaneous storage and processing is not required to build a WM task; rather, the *number* of elements that can be held in the focus of attention, that is, the scope of attention, is of interest. Indeed, Cowan et al. (2005) could show that tasks measuring the scope of attention, i.e., tasks without a storage component, were highly correlated with higher intellectual functioning, as predicted by the embedded-process model. Similar results as those reported by Cowan et al. (2005) were obtained by Oberauer, Süß, Wilhelm, and Wittmann (2008) and Krumm et al. (2009). Oberauer (2002) elaborated on Cowan's view, assuming that the ability to form temporary bindings between pieces of information is the key factor for the importance of WM in higher cognition.

To summarize, WM has generally been conceptualized as a limited-capacity system for the simultaneous storage and manipulation of information. However, the theories described above make different predictions pertaining to the essence of WM. Whereas the multiple-component model and the view of controlled attention postulate that simultaneous storage and processing are of paramount importance for measuring WM capacity, the embedded-processes model assumes that it is the scope of attention that is the most decisive factor. Further, whereas STM and WM are perceived as different constructs in some theories (multiple-component model, controlled attention model), others do not clearly differentiate between these two (Cowan et al., 2005; Unsworth & Engle, 2007a). More research is needed in this area to clarify these issues.

1.2 Reasoning and WM in children

A large variety of reasoning tests exists within the literature. Factor-analyzing a wide range of published studies, Carroll (1993) suggested that reasoning consists of three different factors, induction, deduction, and quantitative reasoning. However, Carroll (1993) mentions several possible objections to this structure of reasoning, such as the fact that many reasoning tests involve language, quantitative, or spatial skills to an unknown degree. Wilhelm (2005) found that a model focusing on task content factors (verbal, figural, or spatial material) was more successful in explaining the pattern of correlations between reasoning tests than other models focusing on process-related differentiations.

Only relatively few studies have investigated the relationship between WM and reasoning in children. For example, de Jong and Das-Smaal (1995) investigated the relationship of WM, fluid intelligence (verbal and figural reasoning), scholastic achievement, and processing speed in a sample of N = 2,222 9-year-old students. These authors showed that WM and fluid intelligence correlated at r = .66 in younger children. In another study (de Jonge & de Jong, 1996), children from fourth to sixth grade were investigated, in which WM and reasoning factors correlated at r = .49 at the latent level. A similar result was reported by Swanson (2008), who found that fluid intelligence and WM factors correlated at r = .54, and that WM, in contrast to STM, remained a significant predictor of fluid intelligence even when the effect of age, processing speed, and other constructs was statistically controlled for. Tillman, Nyberg, and Bohlin (2008), testing children from a broader age range (6 to 13 years old), reported correlations of reasoning with verbal WM and visuospatial WM r = .35 and r = .28, respectively. In Hutton and Towse (2001), who investigated children 8 to 11 years old, WM and STM showed similar correlations with number skills (r = .33 vs. r = .38) and fluid intelligence (r = .36 vs. r = .35), respectively. Results from other studies with children fall into a similar range (Andersson, 2008; Bayliss, Jarrold, Gunn, & Baddeley, 2003). All of these studies investigated relationships between constructs based on sum scores of single tests, using either zero-order correlations, multiple regression, or structural equation modeling. This can give valuable insights into the nomothetic framework of cognitive abilities in children, and more such "macroscopic" research with multiple measures of WM, STM, and reasoning is clearly necessary in samples of children in order to establish firm results.

However, often no systematic design of reasoning or WM items, based on prior cognitive theory, was implemented and statistically analyzed. Hence, postulated cognitive processes during item solving were usually not explicitly tested or modeled, which lies at the heart of construct validity (Borsboom, Mellenbergh, & van Heerden, 2004). In this thesis, an attempt was made to statistically model reasoning processes at the level of single items, using item response theory (IRT) models. Test design was connected with established theories of reasoning and cognitive complexity, allowing a systematic evaluation of these theories, and providing new insights into reasoning processes in children.

1.3 Outline

As mentioned above, there is a paucity of results on the intricate relationships of WM and STM with reasoning and other facets of intelligence in children. Chapter 2 reports results from a macroscopic study in which several research questions are pursued. First of all, the question of whether WM is domain-specific or domain-general has not yet been settled, especially with respect to children. Secondly, it is yet unclear whether classical measures of WM are measurement invariant across age. Thirdly, contradictory predictions of WM theories exist with respect to the structure of WM tasks: Is storage and processing really necessary, or do WM tasks without a processing component load on the same factor as complex span tasks? Finally, the relative contributions of WM and STM to fluid and crystallized intelligence are evaluated.

Chapter 3 focuses on a systematic analysis of algebraic reasoning in children using a bivariate mixed IRT model. A computer-based algebra test was systematically designed such that an assessment of the effects of relevant cognitive processes was possible. For example, the effects of storing intermediate results during algebraic computations was analyzed. An IRT model developed by Klein Entink, Kuhn, Hornke, and Fox (2009) was utilized to simultaneously analyze accuracy and response time data in order to obtain a broader picture of solution processes. The model developed by Klein Entink et al. (2009) was extended by including person-level covariates as well as theoretically interesting cross-level interactions between item and person characteristics.

In Chapter 4, a theory of cognitive complexity, relational complexity theory (Halford, Wilson, & Phillips, 1998), was evaluated. A figural reasoning test was designed based on prior complexity specifications. In order to assess basic tenets of the theory, a host of increasingly complex IRT models was used, each shedding light on the hypotheses of interest. Further, the effect of WM on basic reasoning processes was modeled at the item level, thus providing a more detailed and fine-grained picture of the cognitive processes involved.

This thesis concludes with a discussion of results and some suggestions for future research.

2 Controlled attention and storage, after all? An investigation of the relationship between working memory, short-term memory, and intelligence in children

Summary. Working memory (WM) has received considerable attention in psychological research, a core finding being a close relationship between WM and measures of complex cognition. However, only a limited amount of studies investigated this relationship in samples of children. This study explored the contribution of storage-and-processing tasks (WM), a measure of the scope of attention based on Cowan et al. [Cowan, N., Elliott, E. M., Saults, J. S., Morey, C. C., Mattox, S., Hismjatullina, A., et al. (2005). On the capacity of attention: Its estimation and its role in working memory and cognitive aptitudes. Cognitive Psychology, 51, 42–100.], and short-term memory (STM) tasks of supraspan length to 275 (8 to 13 years old) school children's fluid and crystallized intelligence. The results showed that a two-factor structure of memory, consisting of a WM (storage-and-processing as well as scope of attention tasks) and STM, was comparable across age groups. WM was a strong predictor of fluid (Gf) and crystallized (Gc) intelligence both when modeled separately and when modeled as a residual factor controlling for STM variance. Further, STM interacted with age and was unrelated to Gf in children older than 11 years, whereas the effect of WM on Gc was consistently mediated by STM. The results suggest that STM and WM are separable but highly-related constructs, secondary-memory processes (e.g., search and retrieval) along with controlled attention are hallmark predictors of intelligence in children, and STM effects on Gf are moderated by age.

2.1 Introduction

Working memory (WM) has been commonly referred to as a processing resource of limited capacity that enables the storage and simultaneous manipulation of information (Baddeley, 1986; Engle, Tuholski, Laughlin, & Conway, 1999). That is, WM has traditionally been measured with *complex span* tasks that consist of a storage and a processing component (Oberauer, 2005c). In contrast, short-term memory (STM) is viewed as the ability to keep a limited amount of information in a passive storage without processing the same or additional information (Just & Carpenter, 1992; Unsworth & Engle, 2007b). STM is usually measured by *simple span* tasks. Numerous studies have provided evidence for a close relationship between WM and intelligence, scholastic achievement, and learning (e.g.,

Bayliss, Jarrold, Gunn, & Baddeley, 2003; Colom, Abad, Quiroga, Shih, & Flores-Mendoza, 2008; de Jong & Das-Smaal, 1995; Engle et al., 1999; Hitch, Towse, & Hutton, 2001; Gathercole, Lamont, & Alloway, 2006; Kuhn & Holling, 2009; Kyllonen & Christal, 1990; Krumm, Ziegler, & Buehner, 2008; Süß, Oberauer, Wittmann, Wilhelm, & Schulze, 2002; Swanson, 2008), whereas evidence for the predictive power of STM for the latter constructs is less unanimous (Ackerman, Beier, & Boyle, 2005; Unsworth & Engle, 2007b).

Recently, several authors suggested that the measurement of WM need not be confined to complex span tasks, but that tasks without a processing component can measure WM capacity as well (Cowan et al., 2005; Haarmann, Davelaar, & Usher, 2003; Oberauer, 1993). For example, Cowan et al. (2005) argue that the prevention of rehearsal and grouping processes is the central aspect of successful WM measures, which does not necessarily imply a processing task. Cowan et al. (2005) could show that complex span tasks as well as tasks tapping the scope of attention (i.e., tasks without a processing component) both loaded on a single WM factor which correlated highly with a latent intelligence factor, g (r = .78). However, the relationship of WM and g with STM at the latent level was not investigated by these authors.

2.1.1 The distinction between WM and STM

Starting with Daneman and Carpenter (1980), several studies have shown that in adult participants, complex span tasks correlate more highly with measures of higher order cognition than simple span tasks (e.g., Conway, Cowan, Bunting, Therriault, & Minkoff, 2002; Daneman & Merikle, 1996; Engle et al., 1999). For example, although Engle et al. (1999) reported a strong correlation between STM and WM factors (r = .68), a two-factor model fit their data better than a one-factor model. Furthermore, they found that after controlling for STM variance in the WM factor, the WM residual was still significantly correlated with measures of fluid intelligence. In contrast, the STM residual was no longer significantly related to fluid intelligence when controlling for WM variance. According to Engle et al. (1999), the residual WM variance corresponds to *controlled attention*, a central executive component which is closely related to fluid intelligence. A similar result was obtained by Conway et al. (2002), who reported a standardized path coefficient of .60 between a residual WM factor and fluid intelligence, but no significant correlation between STM and fluid intelligence. In a related study with 7-

year-old children, STM measures predicted reading comprehension or arithmetic ability only when they were entered into a regression equation prior to WM tasks (Leather & Henry, 1994).

However, several other studies have found that simple span tasks correlate substantially and nearly as well as complex span tasks with higher order cognition (Colom, Rebollo, Abad, & Shih, 2006; Colom et al., 2008; Kane et al., 2004; Mogle, Lovett, Stawski, & Sliwinski, 2008; Shah & Miyake, 1996; Tillman, Nyberg, & Bohlin, 2008; Unsworth & Engle, 2006). Recently, Ackerman et al. (2005) reported a meta-analytic correlation of g and a WM factor of r = .50, whereas the correlation was r = .49 for g and a STM factor. Further, Kane et al. (2004) report a correlation of r = .54 between a spatial simple span factor and fluid intelligence which was slightly higher than the latent correlation of fluid intelligence with executive attention (cf. Miyake, Friedman, Rettinger, Shah, & Hegarty, 2001). This finding is in line with the results by Colom et al. (2008) who could show that WM was not a significant predictor of g when conceptualized as a residual factor beyond STM.

An interesting finding was reported by Unsworth and Engle (2006) who found that whereas complex span tasks of all list lengths correlated substantially with fluid intelligence, the same was true for simple span tasks only when the number of items to be retained exceeded four. Interestingly, humans can only keep up to four entities in mind at the same time (Cowan, 2001). This leads to the differentiation of *primary memory* (PM) from *secondary memory* (SM; cf. James, 1890). The purpose of PM is to maintain a distinct number of separate representations active for ongoing processing by continuously allocating attention (Unsworth & Engle, 2007a). In contrast, items that have been displaced from PM must be retrieved from SM, which requires a cue-dependent effortful search process that is vulnerable to interference (Lustig, May, & Hasher, 2001; Oberauer & Lewandowsky, 2008). SM can be measured, for example, by having subjects learn lists of supraspan length, whereas measures of PM rather capture the scope of immediate memory without activating storage or retrieval processes. In short, PM refers to a maintenance component of memory, whereas SM refers to search and retrieval processes (Unsworth & Engle, 2007a). However, it should be noted that simple and complex span tasks both measure PM and SM, albeit to a differing degree. As stated by Unsworth and Engle (2006, p. 70), "The main difference is that the majority of items in complex spans are displaced from primary memory and must be retrieved from secondary memory, whereas for simple spans many

items can be recalled from primary memory". It can therefore plausibly be assumed that in order to tap SM in complex and simple span tasks to a comparable degree, tasks of supraspan length should be used to measure STM.

Based on these ideas, Mogle et al. (2008) recently investigated to which degree PM, SM, WM, and processing speed predicted fluid intelligence. In a sequence of nested structural equation models, Mogle et al. (2008) found that WM played no significant role in predicting intelligence when tasks measuring SM were in the model. These authors provide evidence that SM and to a lesser degree, PM are sufficient predictors for fluid intelligence, rendering both WM and processing speed insignificant. Further, processing speed could be dropped as an insignificant predictor in the case of the full model as well (cf. Kane, Poole, Tuholski, & Engle, 2006). These results lead to a theoretically parsimonious and powerful explanation for the varying correlations of STM with fluid intelligence across the literature: STM tasks are especially good predictors of higher level cognition when performance is measured using supraspan lists and the role of rehearsal is reduced, that is, STM tasks are good predictors to the degree they capture SM. In line with these results, Maybery and Do (2003) found that verbal and spatial supraspan STM tasks substantially correlated with mathematical ability in a sample of children (r = .50 and .53, respectively). Hence, STM tasks should predict higher level cognition to a similar degree as WM tasks, and a residual WM factor should remain an insignificant predictor. However, tasks that primarily capture PM should be predictive of intelligence as well albeit to a lesser degree than SM tasks (e.g., antisaccade tasks; Unsworth, Schrock, & Engle, 2004).

STM measures commonly applied in the literature are either in free-recall or recognition format. STM measures that require recognition can further be differentiated according to whether they measure recollection or familiarity (Oberauer, 2005b). Familiarity measures require subjects to indicate whether a target word or stimulus occurred in a previously-shown list to be learned. In contrast, recollection measures require bindings between objects to be remembered and specific cues (e.g., spatial cues), that is, target words or objects have to be remembered in context. As shown by Oberauer (2005b), recollection measures correlated highly with WM, whereas familiarity measures did not. Recognition measures of STM can therefore be considered good operationalizations of SM in case they require learning supraspan tasks and have a recall or recognition format.

2.1.2 WM and STM in children

Several possible explanations have been suggested for the higher WM capacity in older as compared to younger children, ranging from higher processing speed of diverse WM components (Bayliss, Jarrold, Baddeley, Gunn, & Leigh, 2005; Case, Kurland, & Goldberg, 1982; Hale, Bronik, & Fry, 1997; Towse, Hitch, & Hutton, 1998) over better inhibition of irrelevant information (Swanson & Howell, 2001) to more efficient or larger storage capabilities in older children (Bayliss et al., 2005; Cowan et al., 2005). The latter bear on the relationship of STM with WM in children, which is less unequivocal. For example, Hutton and Towse (2001) found that the correlations of WM and STM measures with measures of reading, number skills, and fluid intelligence, respectively, were of the same magnitude in a sample of children from 8 to 11 years. Further, STM and WM measures correlated at r = .76. In contrast, Alloway, Gathercole, and Pickering (2006) reported results that showed a very high correlation between WM and visuo-spatial STM factors at ages 4 to 6 (r = .97) that dropped to r = .71 at ages 9 to 11. That is, in the parlance of the WM model proposed by Baddeley (1986), tasks relating to the phonological loop (i.e., STM) and tasks relating to executive function (i.e., WM) are not clearly separable in young children. Evidence reported by Swanson (2008) supported a dissociation between STM and WM factors in children 6 to 9 years old, as a model consisting of separate WM and STM factors showed a good fit. A possible reason for these findings might reside in the fact that because the phonological loop is more developed in younger children relative to their executive system and because the memory span of younger children is shorter, possibly due to lack of rehearsal (Alloway et al., 2006; Flavell, Beach, & Chinsky, 1966; Riggs, McTaggart, Simpson, & Freeman, 2006), they are required to access SM resources earlier (i.e., in tasks with fewer items to be remembered) than older children. In the latter case, STM tasks should be good predictors of higher cognitive functioning and WM should not explain much additional variance. In older children, however, STM and WM can be seen as separable constructs, and because older children have higher STM spans and a better executive system, the importance of STM relative to WM can be expected to decrease. In line with these assumptions, Bayliss et al. (2003) found no significant relationship between a residual WM factor and scores on the Raven Progressive Matrices Test in a sample of children 7 to 9 years old. In contrast, however, Tillman et al. (2008) reported significant correlations between WM residuals (controlling for STM) and fluid intelligence, both for the verbal and visuo-spatial domain. Similarly, Swanson (2008) found that

when controlling for STM, a residual WM factor was still an important predictor for fluid intelligence. Apparently, results in samples of children have hitherto not provided unequivocal results.

2.1.3 Theories of WM

Several different conceptualizations of WM have been suggested in the literature. Apart from the long-standing model introduced by Baddeley (1986), two approaches have been influential in the literature. The first one views *controlled attention* as a central component of WM (e.g., Heitz, Unsworth, & Engle, 2005; Kane, Conway, Bleckley, & Engle, 2001). Controlled attention is defined as "an ability to effectively maintain stimulus, goal, or context information in an active, easily accessible state in the face of interference, to effectively inhibit goal-irrelevant stimuli or responses" (Kane et al., 2001, p. 180). Consequently, proponents of this approach define WM capacity as "an *ability* reflecting the extent to which an individual is able to control attention, particularly in situations involving interference from competing information, activated representations, or task demands" (Heitz et al., 2005, p. 64). In this view, complex span tasks mainly reflect the ability to control attention, and therefore often show higher correlations with measures of higher cognitive functioning than measures of different WM facets (e.g., Buehner, Krumm, & Pick, 2005).

Another recent model views WM as the activated traces in long-term memory. Specifically, Cowan (2005) argues that the *focus of attention*, which is limited to approximately four items (Cowan, 2001), corresponds to the current working memory contents. The focus of attention, consisting of long-term memory elements that are highly activated, forms the first layer in this model. A second layer is composed of moderately-activated elements in long-term memory that can be retrieved into the focus of attention. The scope of attention refers to the amount of information that can be kept in the focus of attention for immediate retrieval. Individuals differ in the size of their scope of attention (Cowan et al., 2005), which appears to be a relatively fixed parameter (Cowan, Chen, & Rouder, 2004; Oberauer, 2006; Oberauer & Bialkova, 2009; Scolari, Vogel, & Awh, 2008), although it is subject to developmental constraints and therefore is smaller in younger children than in older children or adults (Cowan et al., 2005; Cowan, Naveh-Benjamin, Kilb, & Saults, 2006).

Because in contrast to long-term memory, the focus of attention is capacity-

limited, it is crucial to obtain a measure of this capacity which forms a gateway of cognitive processing. However, according to Cowan et al. (2005), complex span tasks are not well-suited to measure the capacity of the focus of attention. This is because it is not entirely clear what complex span task scores actually represent. For example, subjects might not divide attention between the storage and the processing task but instead switch between these tasks (e.g., Hitch et al., 2001). In this case, the ability to switch attention would be a crucial factor. In addition, different scoring methods for complex span tasks have been suggested, often leading to substantially different conclusions concerning their relationship with STM and intelligence (e.g., Conway et al., 2005; Unsworth & Engle, 2007b). Further, it is possible that the degree of proactive interference by the time that long lists are being presented plays a decisive role and not a general WM capacity (Lustig et al., 2001). Finally, and perhaps most importantly, complex span task scores form a mixture of processing and storage components, and it is entirely plausible that subjects with a high storage capacity but a lower degree of controlled attention achieve the same score as subjects with a high ability to control attention but lower storage capability. Complex span tasks therefore share some interpretational problems with other dual tasks (Pashler, 1994).

According to Cowan et al. (2005), therefore, WM tasks that neither require a processing subtask nor allow for rehearsal can be good measures of the scope of attention. These authors used a visual array comparison task task based on Luck and Vogel (1997). In that task, subjects had to compare two successively-shown arrays of squares of different color, and to indicate whether a target square in the second array had changed its color or not. Several papers have demonstrated the utility of this paradigm for the determination of individual capacity limits in the scope of attention in both children and adults (Cowan, Fristoe, Elliott, Brunner, & Saults, 2006; Cowan, Naveh-Benjamin, et al., 2006; Vogel, McCollough, & Machizawa, 2005; Wheeler & Treisman, 2002). Further, the work by Cowan et al. (2005) lends support to the view that this task, along with other tasks capturing the focus of attention, is substantially correlated with intelligence.

How can tasks measuring the scope of attention be reconciled with the framework of PM and SM recently advanced by Unsworth and Engle (2007a)? Because the scope of attention refers to a limited amount of highly-activated long-term memory traces, one might suppose that the visual array comparison task described above measures primarily PM. However, according to Cowan et al. (2005), tasks measuring the scope of attention, including a visual array compari-

son task, load on a single factor together with complex span tasks. A related finding by Kane et al. (2004) supports this result, as these authors found a very high correlation (r = .89) between a spatial WM and a spatial STM factor (cf. Miyake et al., 2001). In a sample of 4- to 6-year old children, Alloway et al. (2006) even report a correlation of r = .97 between visuospatial STM and WM tasks. Because rehearsal is difficult in visual array comparison tasks, they further conceptually differ from classical simple span tasks. Hence, it is of theoretical interest to locate the scope of attention in latent variable models including both complex span tasks as well as simple span tasks of supraspan length. The scope of attention should be statistically separable from complex and simple span tasks in case it measures primarily PM.

In contrast, simple and complex spans capture both PM and SM, although to varying degrees (Unsworth & Engle, 2007a): Whereas simple supraspan tasks are good indicators of SM in the face of potential memory overload, complex span tasks are more concerned with the control of attention and inhibition of irrelevant (processing) information while storing items in SM. That is, in simple span tasks of supraspan length, memory content is displaced from PM due to information overload, whereas in complex span tasks, memory content is displaced to SM due to the necessity of handling a secondary processing task. Apart from the work by Mogle et al. (2008), most studies used classical simple span tasks to measure STM, thereby potentially lowering its impact on general cognitive functioning (Unsworth & Engle, 2007b). A core question in this context therefore pertains to the relationship between (a) the limit of the scope of attention, (b) the ability to conduct cue-directed retrieval and (c) the ability to control attention as well as inhibit irrelevant information, respectively, with general intelligence. Further, it remains unclear to which degree potential differences in these relationships are affected by age in children. The studies by Mogle et al. (2008) and Cowan et al. (2005) provide some first results concerning this issue. However, Mogle et al. (2008) investigated a highly-selected adult sample, whereas Cowan et al. (2005, Experiment 2) did not include any supraspan STM measure in their analysis. In this study, a sample of children will be investigated cross-sectionally to provide some answers to the aforementioned questions.

A final question pertains to the domain-specificity or generality of WM. Whereas the domain-specific WM model assumes different WM factors across content domains (e.g., verbal/numeric vs. spatial), the domain-general model postulates a single WM factor that spans across content domains. Several re-

searchers have found support for a domain-general WM model (e.g., Ackerman, Beier, & Boyle, 2002; Colom, Flores-Mendoza, & Rebollo, 2003; Conway et al., 2002; de Jonge & de Jong, 1996; Kane et al., 2004). However, some results support the notion of separate factors of WM that are highly correlated. For example, Jarvis and Gathercole (2003), testing the WM model suggested by Baddeley (1986) in a sample of 11 and 14-year-old children, found evidence for a verbal and non-verbal WM factor that were highly correlated (r = .53 in younger children, r = .60 in older children, respectively). In the study by Süß et al. (2002), a visuospatial and a verbal-numerical WM factor correlated at r = .80. In this context, Engle et al. (1999) suggested a hierarchical model of WM that assumes both a domain-general factor as well as domain-specific residual factors. It should be noted that numerous studies supporting the domain-specificity of WM used relatively homogeneous samples, which can result in overfactorization (Shah & Miyake, 1996). In contrast, samples representing a broader range of the population (e.g., Kane et al., 2004) rather found support for a more domain-general model of WM.

2.1.4 Reasoning, fluid intelligence, and crystallized intelligence

Since the early 20th century, intelligence as a psychological construct has been fractionated into different factors. Two recent theories of the structure of intelligence have been of key importance, the three-stratum-theory proposed by Carroll (1993) as well as the Cattell-Horn Gf-Gc theory (e.g., Horn & Blankson, 2005). These theories are very similar in nature and have recently been merged in the Cattell-Horn-Carroll (CHC) theory approach (McGrew, 2005). CHC theory assumes a hierarchically-structured model of intelligence where the g factor is located at the highest and most general level. Below the g factor, broad ability factors like Fluid reasoning (Gf) or Comprehension-knowledge (Gc) are located. Gf is commonly defined as the ability to reason under novel conditions, whereas Gc is related to academic achievement or cultural knowledge based on already learned knowledge (cf. Haavisto & Lehto, 2004). Both of these factors, especially Gf, have been shown to be core components of human intelligence (Marshalek, Lohman, & Snow, 1983; Undheim & Gustafsson, 1987). Although conceptually distinct, Gf and Gc are statistically often closely related, i.e., they could not be statistically separated in several analyses (Carroll, 1993).

Inductive and deductive reasoning are generally considered the hallmark indicators of Gf. Whereas in inductive reasoning, participants are supposed to

detect rules, patterns or similarities in test items and successfully apply these rules, deductive reasoning requires participants to reason from premises to conclusions that properly and necessarily follow from them.

Harman (1999) assumes people reason in an essentially nondeductive way, and utilize the same reasoning processes on both inductive and deductive reasoning problems. Taking a related approach, Johnson-Laird (1994) has extended the mental models account, which is usually applied to deductive problems, to a range of inductive problems. In addition, several researchers have proposed accounts that focus mainly on reasoning about inductive arguments, and have described deductively correct arguments as special cases (e.g., Osherson, Smith, Wilkie, Lopez, & Shafir, 1990).

In contrast, other researchers have emphasized a distinction between two different reasoning systems (Sloman, 1996; Stanovich, 1999). Such two-process accounts assume one system that is relatively fast but heavily influenced by context and associations, whereas the other is more deliberative, analytic, and rule-based. Although these two systems do not necessarily correspond directly to induction and deduction, it is quite plausible to assume that induction would depend more on the first system, whereas deduction is heavily affected by the second system. Recent neuropsychological evidence, based on brain imaging, gives support for two anatomically separate systems of reasoning (Goel, Gold, Kapur, & Houle, 1997; Parsons & Osherson, 2001). Interestingly, these two-process accounts of reasoning are conceptually mirrored in the familiarity and recollection systems in memory research, where only recollection is substantially related to WM (Oberauer, 2005a; Yonelinas, 2002).

Shye (1988), reanalyzing data by Colberg, Nester, and Trattner (1985), could show that rule-application (i.e., deductive reasoning) and rule-inference (i.e., inductive reasoning) were clearly separable in a MDS analysis. However, a subsequent study could not separate inductive from deductive reasoning using latent variable models (Wilhelm, 2005). Rather, a confirmatory factor analysis (CFA) model based on separated content factors (figural, numerical, verbal) provided the best fit. In contrast to these results, Heit and Rotello (2005) provided results supportive of separate systems for inductive and deductive reasoning. Although both inductive and deductive reasoning are clearly central to Gf, it is therefore an open question whether they can be factorially distinguished or not.

Further, the magnitude of specific links between intelligence and memory

factors are of interest. Klauer, Stegmaier, and Meiser (1997), for example, utilized a dual-task paradigm to study disruptive effects of different secondary tasks on spatial and propositional reasoning, respectively, as primary tasks. These authors found that secondary tasks tapping the central executive (e.g., random number generation) were disruptive for both spatial and propositional reasoning. However, they found that the disruptive effect of a visual tracking task was only present for spatial reasoning but not for propositional reasoning. These results support the notion that reasoning (i.e., Gf) heavily depends on domain-free executive functioning, but less so on domain-specific storage. In contrast, because Gc represents academic achievement or cultural knowledge, it can be hypothesized that this factor depends on storage or maintenance of known, activated information to a much larger degree than on executive functioning (cf. Swanson, 2008). Further, the relationship between Gf and Gc can be assumed to be lower in older children because children's knowledge base tends to be standardized by school curricula (Schweizer & Koch, 2002). However, it can be expected that in children, STM generally plays a prominent role in predicting performance on intelligence tests, whereas in adults, WM is the decisive factor (Hutton & Towse, 2001).

2.1.5 Issues of measurement invariance

Although numerous studies have examined developmental differences in the structure of WM and STM in children, only few have investigated whether the hypothesized models exhibited measurement invariance (MI) across age groups. MI refers to the extent to which different test scores have the same meaning across groups of examinees (Gregorich, 2006). An investigation of MI can reveal whether a test is systematically biased against a specific subpopulation of participants, or whether an array of tests refers to the same latent variables, to the same degree, across groups. Research questions in this context might be, for example, whether a possible advantage of older children compared to younger ones on specific WM test scores is caused by a higher latent WM capacity or unrelated measurement artifacts (e.g., test sophistication). Assessing MI, therefore, helps to decide whether observed test scores can be attributed to latent (factor) scores or must be seen as being caused by unrelated sources (Wicherts, Dolan, & Hessen, 2005). Further, the results of fitting a CFA model to the whole sample under investigation can mask differences in the factor structures of the subgroups under investigation (Meredith & Teresi, 2006). Therefore, by investigating MI and using multiple-groups CFA, biased results might be avoided.

MI is usually evaluated by fitting a sequence of increasingly restrictive CFA models, with strict measurement invariance seen as the ideal for cross-group comparisons (Lubke & Dolan, 2003; Meredith, 1993). Strict measurement invariance requires equal factor loadings, latent intercepts, and error variances across groups. In order to compare latent means, however, weaker forms of measurement invariance are sometimes deemed acceptable (Thompson & Green, 2006). Latent mean comparisons commonly provide a higher statistical power than MANOVA comparisons based on manifest indicators (Yuan & Bentler, 2006). Further, effect sizes similar to those suggested by Cohen (1988) have been developed for latent mean comparisons (Hancock, 2001), which are especially helpful because classical goodness-of-fit indexes are highly sensitive to sample size in latent mean comparisons (Fan & Sivo, 2009).

From a conceptual point of view, MI analyses are highly attractive because they enable the researcher to constrain latent variable variances and covariances between the groups under investigation (Vandenberg & Lance, 2000). This speaks to the comparability of relations among latent variables. It should be investigated, for example, whether the relationship between WM and STM measures is the same in younger and older children, or whether it substantially differs (Alloway et al., 2006). MI analyses can easily implement and test such theoretical considerations.

Only few studies investigated the MI of hypothesized WM and STM models in children. Gathercole, Pickering, Ambridge, and Wearing (2004) investigated whether a WM model based on Baddeley (1986) comprising three latent factors (executive function, phonological loop, visuo-spatial sketchpad) was invariant across four age groups (6 to 15 years old) in children. They found that constraining selected factor loadings and covariances to be equal across age groups resulted in a well-fitting model. Further, Swanson (2008) investigated whether a model with one STM factor and two WM factors (verbal and visual) was invariant across two age groups (6 to 7 vs. 8 to 9 years old) in children. Swanson (2008) constrained factor loadings as well as covariances to be equal, and found a satisfactory fit. However, both studies only imposed relatively weak constraints on their respective latent factor models, leaving aside constraints on error covariances as well as latent intercepts. This constitutes a possibly biasing factor. A systematic investigation of MI (e.g., Vandenberg & Lance, 2000; Wicherts et al., 2004) of the memory structure in children differing in age therefore seems necessary.

2.1.6 Objectives

The present study pursues three main goals. Firstly, we want to shed light on the structure of WM, STM, and the scope of attention, respectively, as well as intelligence factors in a sample of children by means of latent variable models. Although several large-scale investigations have provided detailed results in adult samples (e.g., Colom et al., 2008; Engle et al., 1999; Kane et al., 2004; Süß et al., 2002), data for samples of children are still sparse. Secondly, it is our goal to clarify whether WM, STM, and the scope of attention are constant predictors of intelligence factors, or whether these constructs are differentially related. For example, while reasoning factors (Gf) might heavily depend on WM, it is conceivable that Gc is more dependent on STM (Swanson, 2008). And thirdly, we are concerned with the stability of cognitive structure across age; that is, we investigate MI of our measurement models across different age groups, and we check whether age interacts with one of these variables. We expect that in the sample investigated here, MI will hold across subgroups, while latent means will significantly differ, i.e. we expect no qualitative but quantitative (mean) differences across subgroups (Gathercole et al., 2004).

2.2 Method

2.2.1 Subjects

Two-hundred seventy-five children participated in this study, of whom 59 children visited primary school, whereas the remaining 216 children went to secondary schools in various regions of Germany. Mean age was 10;8 years (SD = 1.07, range: 8;0-13;4). 49.6% of the participants were female. Parental consent was obtained for all participants prior to testing. Few participants (n = 18) indicated German was not their first language, although all of these participants spoke German since they were 3 years old.

2.2.2 WM tasks

Three types of computer-based complex span tasks were used, one comprising verbal material, one with numerical and one with visuo-spatial material. The three tasks used were Verbal Span, Spatial Working Memory, and Computation

Span, respectively. The complex span tasks used here were based on Vock and Holling (2008). These authors modified several complex span tasks from the literature (e.g., Daneman & Carpenter, 1980; Oberauer, Süß, Schulze, Wilhelm, & Wittmann, 2000) concerning presentation times, item complexity, and instructions to make them more appropriate for children 8 to 13 years old. Thus, Vock and Holling (2008) were able to design a WM test battery with excellent psychometric properties for children.

Each complex span task commenced with between two to three simple practice tasks that provided subjects with immediate feedback. All subjects had to repeat the practice tasks until each had been solved correctly. This procedure was implemented to ensure that subjects fully understood the instructions before the testing phase began, and to become acquainted with the testing procedure.

At this point, a few words with respect to scoring of WM and STM tasks are appropriate. Several papers have investigated the effects of different scoring methods for WM and STM tasks on psychometric properties and relations to other constructs (Conway et al., 2005; Friedman & Miyake, 2005; Unsworth & Engle, 2007b). WM tasks generally comprise several items and subitems. For example, in this study, Computation Span comprised 10 items, each consisting of three to seven equations displayed on the screen (subitems). The subitems (in the case of Computation Span, the results of the equations) were the contents that had to be remembered. Whereas absolute scoring procedures require all subitems to be remembered correctly for an item to be scored 1, partial scoring methods compute the item score as the proportion of subitems remembered correctly. That is, a subject remembering three out of four items correctly on the Computation Span task would score 0 using absolute scoring and 0.75 using partial scoring. In partial scoring, the total score of a WM task corresponds to the mean of all partial scores across all items. It has been shown that partial scoring results in better psychometric properties and higher correlations with measures of fluid intelligence, presumably because much information especially from long list lengths is retained that is lost when using absolute scoring procedures (Unsworth & Engle, 2007b). In this study, therefore, the partial scoring procedure was used for all WM tasks.

Verbal Span (VS)

This WM task (Oberauer et al., 2000; Vock & Holling, 2008) consisted of two different parts pertaining to storage and processing. Participants first had to memorize a list of words presented simultaneously on the screen (presentation time 6 s). List length in this storage task varied between three to six words. Then, between two and three verbal decision tasks followed in which participants had to respond as quickly as possible. In these processing tasks, participants were supposed to decide which of four words displayed in each corner of the screen stood in a subconcept relation to the word shown in the center of the screen (e.g., "animal" - "lion"). Finally, participants had to reproduce the learned words in correct order. The task consisted of two practice items and 10 test items.

Spatial Working Memory (SWM)

Initially, this task was developed by Oberauer et al. (2000) as a spatial equivalent to the Reading Span task (Daneman & Carpenter, 1980). Participants had to memorize simple chessboard-like 3×3-patterns (storage task). However, the patterns had to be memorized in a rotated fashion, rotated either 90° clockwise or counterclockwise (processing task). That is, before the patterns were shown successively for 4 s each, an arrow indicated whether patterns had to be mentally rotated to the left or to the right. Finally, participants had to successively reproduce the memorized patterns into empty 3×3 matrices on the screen. The task consisted of 13 items with between one to four patterns. Three practice items preceded the testing phase.

Computation Span (CS)

Participants were sequentially shown a series of simple, single-digit equations that included either an addition or a subtraction (e.g., 4 + 3 = 8). Each equation was shown for 5 s. Approximately half of the equations were correct and half were incorrect. The processing task consisted in deciding whether the equation shown on screen was correct or incorrect. Further, all shown equation results had to be memorized irrespective of whether they were correct or not. After all equations had been shown, subjects were presented with an answer screen and successively clicked the to-be-remembered equation results. Each item consisted

of between three to seven equations, resulting in 10 test items. Two practice items were administered before the testing phase.

2.2.3 STM tasks

In order to measure STM capacity, we selected three subtests from the Berliner Intelligenzstruktur-Test für Jugendliche: Begabungs- und Hochbegabungsdiagnostik (BIS-HB; Jäger et al., 2005). The BIS-HB is based on a faceted, well-replicated model of intelligence comprising four operation facets (processing capacity, creativity, memory, and speed) as well as three content facets (verbal, figural, and numerical; see Süß & Beauducel, 2005). We selected one subtest that was representative for each content area of the memory facet. All subtests required participants to memorize stimuli lists of supraspan length, thereby heavily tapping SM (Unsworth & Engle, 2007b). One of the subtests (verbal content) was in free-recall format, whereas the other two (numerical and figural content) were recollection measures. Recollection measures show higher correlations with WM than familiarity measures (Oberauer, 2005a) and can be considered to be similar to free-recall measures.

Verbal STM

In order to measure verbal STM, the subtest "Meaningful text" (MT) was selected. In this test, participants were required to read a short text consisting of five simple sentences with numerous bits of information. Participants were allowed to memorize the text for one and a half minutes. After that, they had to turn the page and write down the answers to 22 specific questions pertaining to the text (e.g., "What was the name of the main character?"). The test score was the number of questions answered correctly. Participants were granted two minutes for answering the questions.

Figural STM

In this subtest, called "Firm logos" (FM), participants were supposed to memorize 20 firm logos, each of which had an individual border. One minute was allowed to learn the combination of firm logos and logo-specific borders. Afterwards, participants turned the page, where under each firm logo, now shown

without a border, four different border forms (distractors) were located. The goal was to select the correct border belonging to each firm logo. The testing phase was limited to 1.5 minutes.

Numerical STM

This subtest ("Pairs of numbers", PN) required subjects to memorize 12 pairs of numbers, the pairs being listed one below each other. One of the numbers had two digits, the other one three, both being separated by a dash (e.g., 12 - 237). After two minutes, participants had to turn the page and select out of five alternatives the correct three-digit number belonging to a previously-presented two-digit number. The two-digit numbers were presented in a different order than in the learning phase. The test score corresponded to the number of items remembered correctly. The time limit for the recollection part of this subtest was two minutes.

2.2.4 The scope of attention

In this study, the scope of attention was measured using a visual array comparison task (VACT; cf. Luck & Vogel, 1997; Cowan et al., 2005). In this computer-based task, participants first saw a red fixation cross for 500 ms in the center of a grey 4 $\times$ 4-matrix on the screen. After that, a visual array of four, six, or eight solid-colored, haphazardly-placed squares, representing set sizes of four (V4), six (V6), and eight (V8), respectively, was displayed within the matrix. Set sizes were randomly ordered across trials. The square colors used were red, blue, violet, green, yellow, black, and white. Care was taken that at least one color was displayed twice in each initial visual array such that subjects had to memorize both color and location of the squares (Cowan, Naveh-Benjamin, et al., 2006). On half of the trials, the first visual array was displayed for 250 ms and for 500 ms on the other half of the trials. This experimental condition was introduced because pretests had shown that 250 ms was too brief for some children and task conditions to encode the visual information. After a blank interval, a second visual array was displayed in which one of the squares was encircled. The participants then had to decide whether the color of the encircled square had changed in comparison to the first visual array or not. On 50% of the trials, the color of the encircled square had changed. The length of the interstimulus interval was

either 1 s, 2 s, or 4 s and equally distributed across trials to raise the difficulty of the task. However, as reported by Cornelissen and Greenlee (2000), in visual array comparison tasks even very complex stimuli show a half-life of about 3 s. In their study, hit rates and false alarm rates were affected by the length of the interstimulus interval, but they remained above chance level.

Three practice trials preceded 48 test trials, including an equal number of trials for each set size. In order to obtain capacity estimates k for each set size, we utilized the formula presented by Cowan (2001, p. 166), which provides relatively stable results across set sizes. The formula is $k = N * (H + CR - 1)$, where H = hit rate, CR = correct response rate and N = number of items (squares) presented. The capacity estimates of a small portion of participants was below 1 ($n = 9$ for set size 4, $n = 41$ for set size 6 and $n = 33$ for set size 8). These values were set to 1 as a lower bound for capacity.

2.2.5 Intelligence measures

In this study, eight measures tapping fluid and crystallized intelligence were used. Six measures were indicators of Gf, whereas two captured Gc. Gf was further decomposed into four tests measuring inductive reasoning and two tests measuring deductive reasoning. All Gf measures consisted of figural content, whereas Gc measures utilized verbal and numerical material, respectively. All eight tests used number-correct scoring.

Inductive reasoning

In order to measure inductive reasoning, the four subtests from the Grundintelligenztest Skala 2 (CFT 20; Weiß, 1998), a German adaptation of the Culture Fair Intelligence Test, Scale 2 (Cattell, 1973), were utilized. The CFT 20 is a paper-and-pencil test which provides high loadings on fluid intelligence (Cattell, 1968) and has good psychometric properties.

Series completion (SC). Participants were supposed to complete a series of three figural elements with a fourth one to be chosen from five distractors. This subtest consisted of 12 items. The time limit was four minutes.

Classifications (CL). Each item consisted of five different figural objects. One of these objects was unrelated to the four others and had to be selected as the cor-

rect answer. Four minutes were allowed to work on this subtest, which consisted of 14 items.

Matrices (MA). Items consisted of 2×2 or 3×3 matrices containing geometric figures. The bottom-right cell of each matrix was left empty. Participants had to select the correct answer that completed the matrix out of five distractors. Overall, 12 items were administered in three minutes.

Topologies (TP). On each item, a target square containing intersecting geometric elements as well as one or more black dots located within these elements was presented. To the right, five distractors containing similar geometric elements that were differently arrayed were shown. Participants had to select the distractor in which a black dot might be placed such that the relationship between the geometric elements and the black dot in the target square was conceptually preserved. This subtest comprised eight items and had to be completed within three minutes.

Deductive reasoning

As mentioned by Wilhelm (2005), only very few tests have been proposed in the literature to measure deductive reasoning with figural content. In this study, we developed two such measures based on the work by Birney, Halford, and Andrews (2006) and Bouwmeester, Vermunt, and Sijtsma (2007).

Latin Square Task (LST). The LST is based on relational complexity theory (Halford, Wilson, & Phillips, 1998), which holds that the number of relationships to be processed is the key component of the difficulty of cognitive processes. By providing a metric for cognitive complexity, it allows for a rule-based and theory-driven test design. Birney et al. (2006) first presented the LST as a cognitive test incorporating the principles of relational complexity theory. In the LST by Birney et al. (2006), an incomplete 4×4 Latin Square was shown to the participant. Some of the cells were filled with geometric figures, whereas some others were empty. One of the empty cells contained a question mark, and the participant had to select the correct geometric figure from a set of distractors. The only rule in the test is that no geometric figure can occur more than once in each row or column. In order to solve items from the LST, a varying number of intermediate steps has to be carried out, each time storing intermediate results in memory. However, the ability to adhere to the single pre-specified rule has to be kept in mind while solving the test.

The LST task designed in this study was administered by computer. It consisted of four practice items, including immediate feedback, and 24 test items. Participants could work on each test item for maximally one minute, after which the next item appeared. As a modification of the task version by Birney et al. (2006), the LST used here consisted of both 4×4 and 5×5 matrices, and it contained four contradictory items that could not be solved unequivocally. In the latter case, participants had to click a crossed question-mark which was displayed next to the other distractors located below the matrix on each item to indicate they had detected the contradiction.

Transitive Reasoning Task (TRT). This computer-administered test was designed based on the work by Bouwmeester et al. (2007) and could be solved by using transitive reasoning (i.e., reasoning of the sort $A > B, B > C \Rightarrow A > C$). In one condition, participants were first shown a box containing five adjacent bars of different color. Initially, only two of the bars were fully displayed such that their length was visible. The other three bars were partially covered such that their color, but not their length was visible. In the second step, one of the bars that had previously been fully visible was partially covered, whereas another bar that was initially covered could be seen in full. This sequence was continued until all bars had been seen in full length, although only two bars were shown in full at each step. After the final step, all bars were partially covered, and two arrows appeared below two of the bars, and an equal sign and a question mark appeared to the right of the box. The task of the participant was to indicate which one of the two bars was higher (arrows), whether the two bars were of equal length (equal sign), or the answer was unknown (question mark). Four different constellations of bars were tested sequentially on each item (i.e., the arrows appeared under four different combinations of bars). That is, each TRT item consisted of four subitems. Two of these subitems were measures of STM (i.e., the two bars had previously been shown simultaneously), whereas two others were measures of transitive reasoning (the two bars had not been shown simultaneously). In this study, only the transitive reasoning subitems were used.

Another difficulty factor manipulated was memory load, in that the bars were either ordered by length (e.g., largest bar to the left, smallest bar to the right) or not. It was assumed that bars in ordered position are easier to remember. Further, the presentation of two bars in each step of the task could be ordered or unordered, that is, the bars are not shown in an ordered sequence from largest to smallest, but in a disordered way. Again, disordered presentation was assumed

to render items more difficult due to WM taxation. In addition, on some items, only a box containing two bars was shown. The other three bars were hidden from view. This manipulation required participants to construct a mental model of the size of the bars without relying on the spatial cues provided in the other conditions (e.g., red bar to the right). Finally, some items contained six bars, as pretests had shown that items with three or four bars were too easy. All of these experimental conditions were distributed in a balanced way across the test.

Two practice tasks with immediate feedback preceded 18 test items, i.e. a maximum test score of 36 could be obtained. Each subitem was presented for 5 s. The time limit to answer each subitem was 5 s, after which the next subitem appeared, or the next TRT item began.

Crystallized intelligence

In order to measure Gc in both the verbal and numerical domain, we used two paper-and-pencil tests developed by Weiß (1998). Although both of these tests required some minimum amount of reasoning ability (i.e., Gf), the tests were primarily designed to capture knowledge. Weiß (1998) notes that whereas the subtests of the CFT 20 (described above) form a Gf factor, the two Gc tests described below both can be allocated to a distinct Gc factor.

Vocabulary test (VT). In this test, participants were presented 30 target words. Next to each target word, five distractors were given. Participants were supposed to select the distractor that had the same meaning as the target word. Overall, 30 test items were preceded by three practice items. The time limit for this test was 12 minutes.

Number series (NS). On each item of this test, participants were presented six numbers in a row, followed by an empty cell with a question mark. Out of five distractors, they had to select the correct answer which correctly continued the sequence of numbers presented. Only one- or two-digit numbers were utilized, and the complexity of the cognitive processes involved was low to medium such that the mastery of elementary arithmetic operations was in the focus of interest, whereas hierarchical, complex relations found in some number series (e.g., Holzman, Pellegrino, & Glaser, 1983) were avoided. The test consisted of 21 items that had to be solved in 16 minutes. Four practice items were administered before the testing phase.

2.2.6 Procedure

Participants were tested in groups of between 8 to 15 children on notebooks provided by the authors' university or in the computer pool of the school. Two separate testing sessions took place in the morning, each including two breaks. Overall testing time was 3.5h. All computer-based tests (WM, deductive reasoning, scope of attention) were administered in the first session, whereas all paper-and-pencil-tests (inductive reasoning, STM, Gc) were administered in the second session that took place approximately one week later. In the second session, participants further answered several demographic questions.

2.3 Results

There are three sections to the results. First, descriptive statistics and zero-order correlations are presented. Second, dimensionality and MI analyses using multiple-group CFAs were conducted. Last, a series of structural equation models (SEM) and hierarchical regression analyses were conducted to examine the relationship among the constructs under investigation.

Descriptive statistics are presented in Table 2.1. No internal consistencies for the different set sizes of the visual array comparison task are provided because a scoring algorithm based on signal detection theory was used (see Section 2.2.4). However, we computed construct reliabilities $\hat{H}$ (Hancock & Mueller, 2001) as well as bootstrapped empirical 95% confidence intervals based on models 2a and 2 in Table 2.3, indicating that the hypothesis of $\hat{H}$ = .70 representing adequate reliability could not be rejected for any of the latent variables. Reliabilities of WM tasks were good and comparably high (Beckmann, Holling, & Kuhn, 2007).

All measures were generally within the acceptable limits of skewness less than 3 and kurtosis less than 4 suggested by Kline (2005). The data were also screened for outliers, with univariate outliers being any data points outside of 3.5 standard deviations from the mean. Five values out of the 4,675 in the data set met this criterion and were replaced with values corresponding to $\pm$ 3.5 standard deviations as appropriate. One multivariate outlier with a Mahanalobis d^2 score (p < .001) was eliminated. Multivariate normality was determined by examining Mardia's multivariate skewness (Z = 7.36, p < .01) and kurtosis (Z = 3.91, p < .01), respectively. Based on these results, multivariate normality had to be rejected. Therefore, a scaled χ^2 statistic (SB-χ^2; Satorra & Bentler, 2001), based on

maximum likelihood estimation, was utilized in all CFA variable models. Using SB-χ^2, likelihood-ratio testing of nested models is feasible under non-normality conditions, and corrected standard errors for all parameter estimates can be computed.

Table 2.2 provides all zero-order correlations of the measures investigated in this study. Relationships among variables were mostly moderate to high, providing a good starting point for latent variable modeling. We proceeded by testing the dimensionality of memory as well as intelligence tests used in this study. With respect to memory, five different models were compared: (a) A single-factor model, assuming a unitary memory structure, (b) a model with two factors, WM and STM (with scope of attention loading on WM; cf. Cowan et al., 2005), (c) a two-factor model (WM and STM) with scope of attention loading on a STM factor, (d) a model comprising three distinct factors, WM, STM, and scope of attention, respectively, and (e) a model consisting of three factors, visuospatial WM (visual array comparison task and SWM), verbal WM (VS and CS), and STM, respectively. Concerning intelligence factors, we compared three models: (a) A single-factor model, (b) a model assuming two factors (crystallized intelligence vs. deductive/inductive intelligence) and (c) a three-factor model, assuming distinct factors for inductive and deductive reasoning, respectively. We compared all models using appropriate fit indexes. Likelihood-ratio testing is feasible under conditions of boundary constraints (Stoel, Garre, Dolan, & van den Wittenboer, 2006), i. e. when models differing in the number of latent factors are compared. It should be noted that the likelihood-ratio test should only be used when the base model shows a good fit (Yuan & Bentler, 2004) and that because the likelihood-ratio test tends to be a liberal criterion, additional information must be taken into account.

Fit indexes especially adequate in MI analyses (Cheung & Rensvold, 2002; Meade, Johnson, & Braddy, 2008) comprise the comparative fit index (CFI) and the noncentrality index (NCI; McDonald, 1989). Changes of more than .002 in the CFI in MI analyses commonly reflect a difference in model fit (Meade et al., 2008). In addition, the Bayesian information criterion (BIC), which penalizes for overparametrization, was used in this study (cf. Raftery, 1995). We further report the root mean square error of approximation (RMSEA), which is more robust under some conditions than incremental fit indexes (Beauducel & Wittmann, 2005).

Table 2.3 shows the results of a first full-sample analysis of memory and intelligence structure, respectively. As can be seen, a two-factor model (model

DM2) lumping complex span tasks and the visual array comparison task (scope of attention) together into one factor, and STM into another, fit significantly better than a one-factor solution (model DM1). This was not the case for a two-factor solution merging STM tasks and the scope of attention into one factor (model DM3), this model was therefore dropped. Finally, although a model with a three-factor task-specific structure (WM, scope of attention, STM tasks; model DM4) resulted in a slightly better model fit than the first two-factor model, an inspection of the 95% confidence interval of the correlation of a WM factor and a scope of attention factor revealed that the upper bound was 1.03. The same was true for a three-factor model assuming different factors for visuospatial and verbal WM (model DM5, upper bound of 95% confidence interval: 1.02). Hence, we rejected all three-factor models and retained a two-factor solution with a STM factor and a WM factor, the latter consisting of complex span tasks and scope of attention.

In addition, we checked the dimensionality of all tests measuring facets of intelligence. As seen in Table 2.3, a one-factor model (model DI1) showed a good fit to the data. However, a two-factor model, dissociating Gf and Gc (model DI2), resulted in an even better model fit. In this model, the correlation between Gf and Gc was high (r = .90), although the upper bound of the 95% confidence interval was .96. However, a three-factor solution (model DI3) did not significantly improve model-fit above the two-factor solution. The 95% confidence interval of the correlation between deductive and inductive reasoning (.75 – 1.02) provided evidence for the unity of deductive and inductive reasoning. We therefore retained a two-factor model of intelligence with highly correlated factors (Gf vs. Gc) for further analysis. At this point, it should be mentioned that this two-factor model is equivalent to a one-factor model with a residual covariance between the two Gc measures, i.e., Gc can also be conceptualized as a residual factor.

Table 2.1: Descriptive statistics and reliability estimates

Variable	M	SD	Skewness	Kurtosis	K-S[a]	α	CI α[b]	$\hat{H}$[c]	CI $\hat{H}$[d]	r_{tt} bat[e]
Working memory								.88[f]	.84 – .91	.94
1. VS	5.88	2.26	-.60**	-.32	.07**	.88	.87 – .90			
2. SWM	5.88	2.85	-.33*	-.64**	.07**	.85	.82 – .87			
3. CS	7.19	2.17	-1.48**	1.52**	.18**	.89	.86 – .91			
Short-term memory								.66	.56 – .76	.81
4. MT	5.12	2.63	.25	-.59**	.12**	.69	.61 – .77			
5. FM	7.75	3.65	.10	-.46	.08**	.74	.67 – .81			
6. PN	3.94	1.93	.15	-.44	.11**	.64	.57 – .72			
Scope of attention								.88[f]	.84 – .91	N/A
7. V4	3.10	.82	-.97**	.34	.20**	N/A	N/A			
8. V6	3.12	1.34	-.07	-.98**	.15**	N/A	N/A			
9. V8	2.94	1.60	.46**	-.78**	.17**	N/A	N/A			
Inductive reasoning								.80[g]	.75 – .84	.86
10. SC	9.43	2.15	-1.66**	3.50**	.20**	.75	.67 – .82			
11. CL	7.88	2.22	-.19	-.10	.11**	.61	.54 – .69			
12. MA	7.88	2.05	-1.02**	.94**	.18**	.70	.62 – .77			
13. TP	4.20	1.73	-.27	-.80**	.15**	.59	.52 – .66			
Deductive reasoning								.80[g]	.75 – .84	.78
14. LST	13.64	3.83	-.24	-.20	.09**	.70	.65 – .75			
15. TRT	25.61	3.97	-.54**	.43	.10**	.72	.63 – .81			
Crystallized intelligence								.81	.75 – .86	.94
16. VT	20.50	5.77	-1.30**	1.20**	.20**	.89	.86 – .91			
17. NS	13.09	5.28	-.50**	-.89**	.13**	.90	.89 – .92			

Note. [a] $Z(p)$ of Kolmogorov-Smirnov-test on normal distribution with correction of significance by Lilliefors. [b] 95% confidence interval for Cronbach's α, based on Maydeu-Olivares, Coffman, and Hartmann (2007). [c] Construct reliability, based on Hancock and Mueller (2001). [d] Bootstrapped 95% confidence interval of $\hat{H}$ (1,000 draws). [e] Test battery reliability (Cronbach's α), computed according to Lienert and Raatz (1998, p. 330). [f] WM and scope of attention formed a single factor (see below) and, hence, had the same construct reliability. [g] Inductive and deductive reasoning were statistically inseparable and had the same construct reliability (see text).
*$p < .05$. **$p < .01$.

Table 2.2: Zero-order correlations among study measures

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Working memory																		
1. VS	–																	
2. SWM	.53	–																
3. CS	.70	.59	–															
Short-term memory																		
4. MT	.37	.33	.36	–														
5. FM	.42	.44	.53	.42	–													
6. PN	.20	.20	.28	.22	.27	–												
Scope of attention																		
7. V4	.47	.46	.56	.23	.33	.23	–											
8. V6	.30	.32	.37	**.10**	.26	**.10**	.39	–										
9. V8	.25	.24	.28	.24	.32	**.12**	.21	.27	–									
Inductive reasoning																		
10. SC	.50	.51	.62	.34	.46	.26	.41	.26	.29	–								
11. CL	.35	.41	.40	.24	.35	.21	.24	.23	.20	.47	–							
12. MA	.43	.45	.52	.34	.36	.22	.35	.31	.25	.55	.42	–						
13. TP	.40	.49	.48	.27	.41	.19	.31	.30	.22	.40	.42	.41	–					
Deductive reasoning																		
14. LST	.34	.39	.39	.22	.30	.20	.29	.35	.24	.38	.27	.32	.33	–				
15. TRT	.47	.41	.52	.28	.41	.25	.47	.34	.33	.39	.32	.37	.38	.34	–			
Crystallized intelligence																		
16. VT	.64	.56	.70	.50	.52	.22	.47	.34	.28	.54	.43	.52	.52	.41	.45	–		
17. NS	.54	.54	.60	.38	.47	.27	.39	.34	.29	.46	.37	.46	.53	.38	.42	.66	–	
Age (in months)																		
18. Age	.50	.46	.56	.33	.47	**.07**	.27	.33	.27	.48	.36	.42	.39	.31	.31	.62	.54	–

Note. Correlations not meeting significance at the .05 level are boldfaced, all others are significant at $p < .01$.

Next, we outline the results obtained from a systematic MI analysis of the two-factor memory structure described above. The sample was split into children aged 8;0 to 10;11 years (n = 151) and children aged 11 years or older (n = 124). We chose this cut-off value for two reasons. Firstly, as noted by Alloway et al. (2006), developmental increases in STM span, in contrast to WM span, level off between 10 to 11 years of age. A systematic MI analysis using 11 years as a cut-off value, therefore, can shed light on the question whether this effect is based on a slowing of developmental growth or whether it represents a measurement artifact. Secondly, choosing 11 years as a cut-off value resulted in approximately equal group sample sizes as advocated by Kaplan and George (1995).

As can be gleaned from in Table 2.4, both configural and metric MI models showed a very good fit to the data. A strong MI model (model MI4) fit the data excellently as well, compared to the configural MI model. That is, factor loadings and latent intercepts did not vary substantially between younger and older children. However, a model assuming equal residual variances across groups (model MI3) exhibited a substantial drop in model fit. We therefore rejected this model and the nested strict MI model (model MI5) and proceeded by comparing a strong MI model assuming equal latent means in both groups (model MI6) to the strong MI model without such constraints. This drastically reduced model fit, implying substantial latent mean differences. A similar drop in model fit was observed for a model constraining factor variances and covariance across groups (model MI7). Hence, it appears that although loading patterns and latent intercepts are homogeneous between younger and older children, substantial differences both in latent ability as well as the relationship between latent abilities can be assumed. Figure 2.1 illustrates this. Both variances and covariance of WM and STM factors were of larger magnitude in younger children, a result that is in line with previous work finding high correlations between STM and WM in young children (e.g., Hutton & Towse, 2001).

In order to quantify the magnitude of latent mean differences, we further computed effect sizes (ES) according to Hancock (2001). The latter are comparable to Cohen's d. The resulting standardized latent mean differences were ES = .62 for STM and ES = 0.72 for WM, respectively, indicating a medium difference between age groups in favor of older children.

To illuminate the predictive power of WM and STM for both Gf and Gc while controlling for age differences, we fit a series of nested structural equation models. In these analyses, we utilized a MIMIC model (Muthén, 1989) instead

of multiple-group CFA, i.e., we included age (in months) directly as a covariate into the model. Because Gf and Gc were very highly correlated (r = .90 in model DI2 in Table 2.3), we allowed Gc tests to load both on a Gf factor and a Gc factor. Because Gc tests often require reasoning abilities to solve them (Carroll, 1993), we deemed this assumption to provide a more realistic picture of Gc.

The results of this analysis are shown in the upper portion of Table 2.5. A model with age (in months) as the sole predictor (model RC1) for both Gf and Gc exhibited a bad model fit. Introducing WM into the model (model RC2) substantially improved the results. Interest goes out to STM, which was a significant predictor of both Gf and Gc when introduced as the third predictor, resulting in a slightly better model fit (model RC3). Parameter estimates of model RC3 are illustrated in Figure 2.2.

As can be seen, WM predicted both Gf and Gc, whereas STM was only relevant for Gf, although it was marginally significant for Gc (p = .07). Age, in contrast, only mattered in the context of Gc, but was irrelevant for Gf when taking STM and WM into account. This could have been expected, because older children should have better knowledge due to more advanced school curricula. We proceeded by placing equality constraints on the unstandardized regression coefficients of WM and STM to Gf, which resulted in a significant deterioration in model fit, ΔSB-$\chi^2(1) = 10.33$, $p < .01$. WM therefore was a better predictor for Gf than STM. However, when constraining the paths from WM and STM to Gc to be equal, a nonsignificant likelihood-ratio test resulted, ΔSB-$\chi^2(1) = 1.29$, $p = .25$. The results presented here contradict evidence provided by Martínez and Colom (2009), who found that WM was no longer relevant for predicting Gc when Gf was statistically controlled for. These authors, however, used observed scores and tested a sample of university students, which can be considered relatively homogeneous in their acquired knowledge (Gc).

Table 2.3: Full-sample CFA dimensionality results

Facet	Model	NF	*df*	SB-χ^2	*p*	ΔSB-χ^2	Compare	CFI	NCI	RMSEA	BIC
Memory	DM1	1	27	55.11	.00	–		.960	.950	.062	9757
	DM2	2[a]	26	37.94	.06	16.45**	DM1	.983	.979	.041	9741
	DM3	2[b]	26	50.50	.00	2.90*	DM1	.965	.956	.042	9753
	DM4	3[c]	24	27.38	.16	5.66*	DM2	.990	.988	.032	9737
	DM5	3[d]	24	31.28	.15	6.69*	DM2	.990	.987	.033	9739
Intelligence	DI1	1	20	33.76	.03	–		.980	.975	.050	10488
	DI2	2	19	22.92	.24	12.79**	DI1	.994	.993	.027	10479
	DI3	3	17	19.68	.29	3.54	DI2	.996	.993	.024	10481

Note. NF = Number of factors, ΔSB-χ^2 = Corrected likelihood-ratio test (Stoel et al., 2006); CFI = Comparative fit index, NCI = Noncentrality index, RMSEA = Root mean square error of approximation, BIC = Sample-size adjusted Bayesian information criterion. [a]Model assuming a WM factor (including scope of attention) and a STM factor. [b]Model assuming a WM factor and a STM factor (including scope of attention). [c]Model assuming three factors: WM, STM, and scope of attention. [d]Model assuming three factors: Visuospatial WM, Verbal WM, and STM.
*$p < .05$. **$p < .01$.

Table 2.4: Fit indexes for MI analysis of two-factor memory structure

Model[a]	EQC[b]	*df*	SB-χ^2	Compare	Δdf	ΔSB-χ^2	CFI	NCI	RMSEA	BIC
MI1	–	52	57.38	–			.990	.990	.027	9682
MI2	$\boldsymbol{\Lambda}$	59	60.46	2 vs. 1	7	3.13	.997	.997	.013	9667
MI3	$\boldsymbol{\Lambda}, \boldsymbol{\Theta}$	68	78.55	3 vs. 2	9	18.57*	.981	.981	.034	9663
MI4	$\boldsymbol{\Lambda}, \boldsymbol{\nu}$	66	72.82	4 vs. 2	7	11.95	.988	.988	.027	9663
MI5	$\boldsymbol{\Lambda}, \boldsymbol{\Theta}, \boldsymbol{\nu}$	75	90.59	5 vs. 3	7	11.84	.972	.972	.039	9659
MI6	$\boldsymbol{\Lambda}, \boldsymbol{\nu}, \boldsymbol{\alpha}$	63	122.18**	6 vs. 4	2	54.90**	.901	.906	.076	9709
MI7	$\boldsymbol{\Lambda}, \boldsymbol{\nu}, \boldsymbol{\Psi}$	69	116.42**	7 vs. 4	3	61.14**	.914	.917	.071	9699

Note. [a]MI1 = Configural MI, MI2 = Metric MI, MI3 = Equal residual variances, MI4 = Strong MI, MI5 = Strict MI, MI6 = Strong MI with latent means fixed, MI7 = Strong MI with fixed factor variances and covariance. [b]Equality constraints across groups.
*$p < .05$. **$p < .01$.

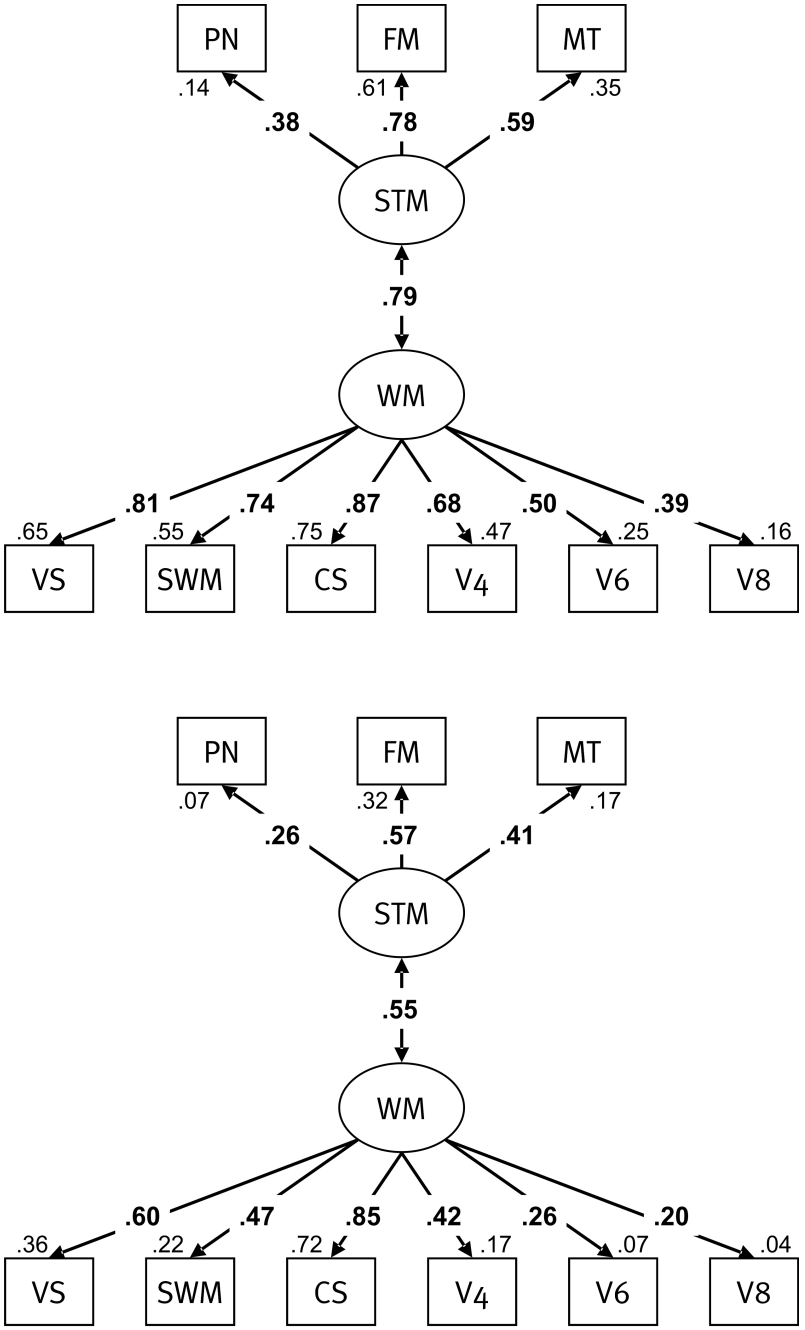


Figure 2.1: Standardized solution (strong MI, model MI4 in Table 2.4) for the two-factor memory structure in younger (upper panel) and older children (lower panel). For abbreviations, see Table 2.1. Unstandardized estimate, younger children (standard error; critical ratio): WM ↔ STM: 4.55 (.62, 7.39). Unstandardized estimate, older children: WM ↔ STM: .983 (.340, 2.89).

In the last two models in Table 2.5, we specified a bifactor structure with respect to WM and STM (e.g., Gustafsson & Balke, 1993). In this context, Colom et al. (2006) reanalyzed several key studies to investigate whether WM, when defined as nested in and orthogonal to STM, still has predictive value for cognitive abilities (e.g., Gf). We follow their approach here and specify WM as a residual factor uncorrelated with STM, i.e. all WM tasks load on both a WM and a STM factor, whereas STM tasks only load on the STM factor. Thus, storage of information was captured by the STM factor, whereas controlled attention (or anything that is measured by complex span tasks above storage) was captured by the residual WM factor. Figure 2.3 presents the results of model RN2, showing that a residual WM factor still had explanatory power for both Gf and Gc, although less so than in model RC3. It was found that constraining the paths from STM and WM to Gf to be equal did not result in a loss of model fit, ΔSB-$\chi^2(1) = .07$, $p = .80$. The same result occurred when constraining the paths of WM and STM to Gc to be of the same magnitude, ΔSB-$\chi^2(1) = .07$, $p = .79$. In model RN2, therefore, STM and a residual WM factor were of similar importance in predicting both Gf and Gc.

In order to cross-validate these findings, we computed a series of hierarchical regression analyses (for a rationale, see Luo, Thompson, & Detterman, 2006). Scores on all indicators pertaining to specific constructs were z-standardized and aggregated. Further, we were interested in the interaction terms of age with WM and STM, respectively. Product terms pertaining to interaction effects were z-standardized after computation of the product. By analyzing the statistical significance of these interaction terms, it was possible to check whether the regression slopes of WM and STM changed or remained constant across age. As illustrated by Table 2.6, we also computed residuals for both WM and Gc. These residual variables were used in additional analyses to check whether WM with STM partialed still affected the criterion variables, and whether controlling for all Gf variance in Gc substantially changed the results.

Table 2.5: Fit indexes for nested structural equation models

Model	Predictors	*df*	SB-χ^2	Δdf	Compare	ΔSB-χ^2	CFI	NCI	RMSEA	BIC
Correlated predictors										
RC1	Age	129	431.05**		–		.848	.576	.092	22083
RC2	Age, WM	127	198.25**	2	RC2 vs. RC1	370.11**	.964	.879	.045	21854
RC3	Age, WM, STM	125	189.70**	2	RC3 vs. RC2	7.93*	.968	.889	.043	21850
Nested predictors										
RN1	Age, WM[a]	122	325.48**		–		.898	.691	.078	21993
RN2	Age, WM[a], STM	120	179.08**	2	RN2 vs. RN1	153.93**	.970	.897	.042	21851

Note.[a]WM was specified as a residual factor in this model (see text).
*$p < .05$. **$p < .01$.

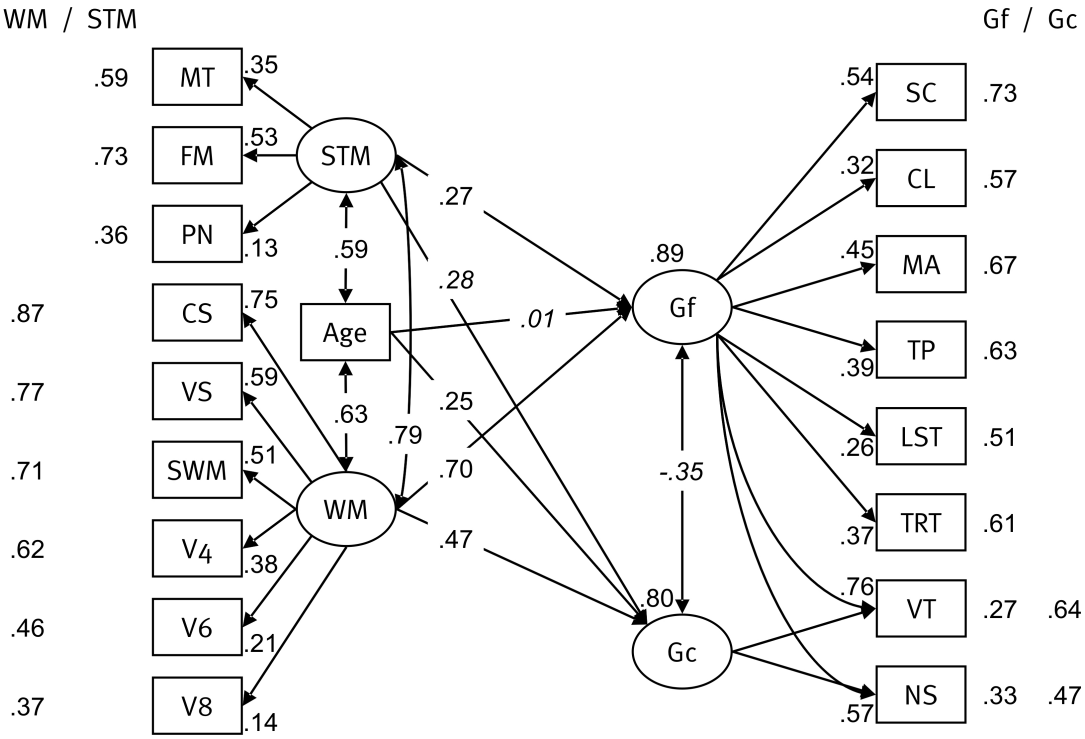


Figure 2.2: Standardized solution of structural equation model predicting Gf and Gc from WM, STM, and age in months (model RC3 in Table 2.5). Factor loadings are given to the left (WM, STM) or right (Gf, Gc) of each indicator. Estimates in italics are statistically nonsignificant. Unstandardized regression weight estimates (standard error; critical ratio): WM → Gf: .64 (.10, 6.32); WM → Gc: 1.00 (.34, 2.97); STM → Gf: .16 (.07, 2.52); STM → Gc: .40 (.22, 1.76); Age → Gf: .00 (.01, .21); Age → Gc: .08 (.02, 3.62).

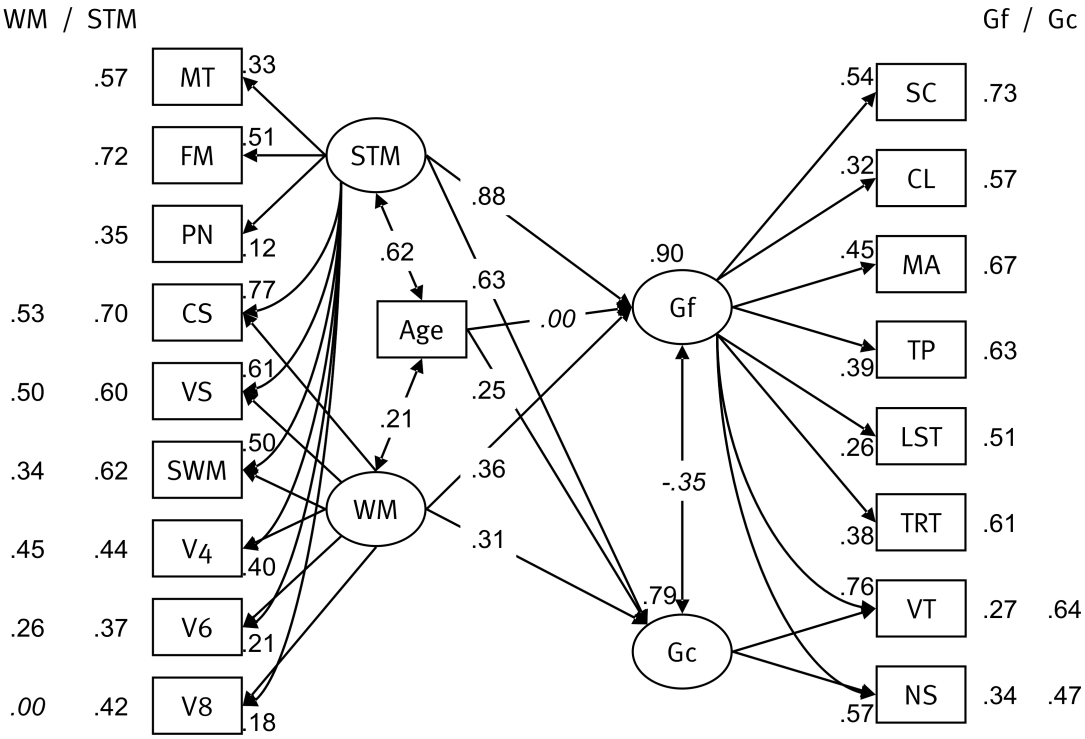


Figure 2.3: Standardized solution of structural equation model predicting Gf and Gc from WM, STM, and age in months (model RN2 in Table 2.5). Factor loadings are given to the left (WM, STM) or right (Gf, Gc) of each indicator. Estimates in italics are statistically nonsignificant. Unstandardized regression weight estimates (standard error; critical ratio): WM → Gf: .49 (.12, 4.26); WM → Gc: 1.00 (.33, 3.04); STM → Gf: .53 (.08, 7.11); STM → Gc: .90 (.17, 5.43); Age → Gf: .00 (.00, -.08); Age → Gc: .08 (.02, 3.63).

All results of hierarchical regression analyses are provided in Table 2.7. The largest variance inflation factor observed overall was 2.3 (WM in model HR4, Gf), indicating that multicollinearity did not affect the results to a significant degree. With respect to Gf, age was a significant predictor only when WM was not taken into account, whereas both STM and WM predicted Gf in all models. Interestingly, the effect of STM was moderated by age, indicating that STM was less important for predicting Gf in older children than in younger children.

This finding is illustrated by a simple slopes analysis (Aiken & West, 1991) in Figure 2.4, indicating that STM was a significant predictor for children aged 9;8 years (b = 1.60, t = 5.77, p < .01) and 10;8 years (b = .78, t = 4.27, p < .01), respectively, but not for children aged 11;8 years (b = -.04, t = -.15, p = .88). A computation of the Johnson-Neyman regions of significance (cf. Preacher, Curran, & Bauer, 2006) revealed that STM was a significant positive predictor for Gf for children up to approximately 11 years old (boundaries of Johnson-Neyman regions of significance for standardized age variable: .44 and 2.20). A supplementary analysis (not shown), using the residual of WM regressed on STM instead of WM unpartialed, in general provided similar results. Importantly, the residual of WM, controlling for STM, remained a statistically significant predictor for Gf in models HR4 and HR5.

Table 2.6: Intercorrelations among aggregated variables

Variable	1	2	3	4	5	6	7	8	9
1. Age	–								
2. STM	.44**	–							
3. STM×Age[a]	-.26**	-.31**	–						
4. WM	.59**	.59**	-.44**	–					
5. WM×Age[b]	-.51**	-.38**	.65**	-.58**	–				
6. WMres[c]	.41**	.00	-.32**	.81**	-.45**	–			
7. Gf	.52**	.59**	-.48**	.77**	-.53**	.53**	–		
8. Gc	.64**	.61**	-.41**	.76**	-.54**	.50**	.70**	–	
9. Gcres[d]	.39**	.28**	-.09	.31**	-.24**	.18**	.00	.71**	–

Note. [a]Interaction term of STM with age. [b]Interaction term of WM with age. [c]Residual of WM regressed on STM. [d]Residual of Gc regressed on Gf.
**p < .01.

Table 2.7: Hierarchical regression analyses of Gf, Gc, and Gc residual on age, STM, and WM

		Gf					Gc					Gcres				
Model	Variables	β	R^2	ΔR^2	95% CI[a]	pr^2†	β	R^2	ΔR^2	95% CI	pr^2	β	R^2	ΔR^2	95% CI	pr^2
HR1	Age	.52**	.27	.27**	.18 – .37	.27	.64**	.41	.41**	.31 – .50	.41	.38**	.14	.14**	.08 – .22	.14
HR2	Age	.32**	.43	.16**	.10 – .22	.08	.46**	.54	.13**	.08 – .19	.17	.32**	.16	.02*	.00 – .05	.08
	STM	.44**				.16	.41**				.13	.14*				.01
HR3	Age	.28**	.51	.08**	.04 – .14	.06	.43**	.57	.03**	.01 – .07	.15	.33**	.16	.00	.00 – .02	.08
	STM	.38**				.11	.37**				.10	.15*				.02
	STM×Age	-.30**				.08	-.18**				.03	.04				.00
HR4	Age	.07	.65	.14**	.09 – .20	.00	.26**	.67	.10**	.05 – .15	.04	.30**	.16	.00	.00 – .03	.06
	STM	.18**				.02	.21**				.03	.12				.01
	STM×Age	-.18**				.02	-.07				.00	.06				.00
	WM	.55**				.14	.46**				.10	.08				.00
HR5	Age	.07	.65	.00	.00 – .01	.00	.25**	.67	.00	.00 – .01	.04	.28**	.17	.00	.00 – .02	.05
	STM	.18**				.02	.21**				.03	.12				.01
	STM×Age	-.18**				.02	-.06				.00	.09				.00
	WM	.55**				.14	.45**				.09	.07				.00
	WM×Age	.02				.00	-.03				.00	-.06				.00
		Model significance														
HR1		$F(1,273) = 100.45$**					$F(1,273) = 189.04$**					$F(1,273) = 45.93$**				
HR2		$F(2,272) = 101.70$**					$F(2,272) = 161.47$**					$F(2,272) = 25.71$**				
HR3		$F(3,271) = 92.72$**					$F(3,271) = 120.66$**					$F(3,271) = 17.30$**				
HR4		$F(4,270) = 125.11$**					$F(4,270) = 136.99$**					$F(4,270) = 13.22$**				
HR5		$F(5,269) = 99.76$**					$F(5,269) = 109.42$**					$F(5,269) = 10.67$**				

Note. [a]Bootstrapped 95% confidence interval (1,000 draws) of ΔR^2 (Algina, Keselman, & Penfield, 2007). †Squared part correlation, representing the unique contribution of each predictor.
*$p < .05$. **$p < .01$.

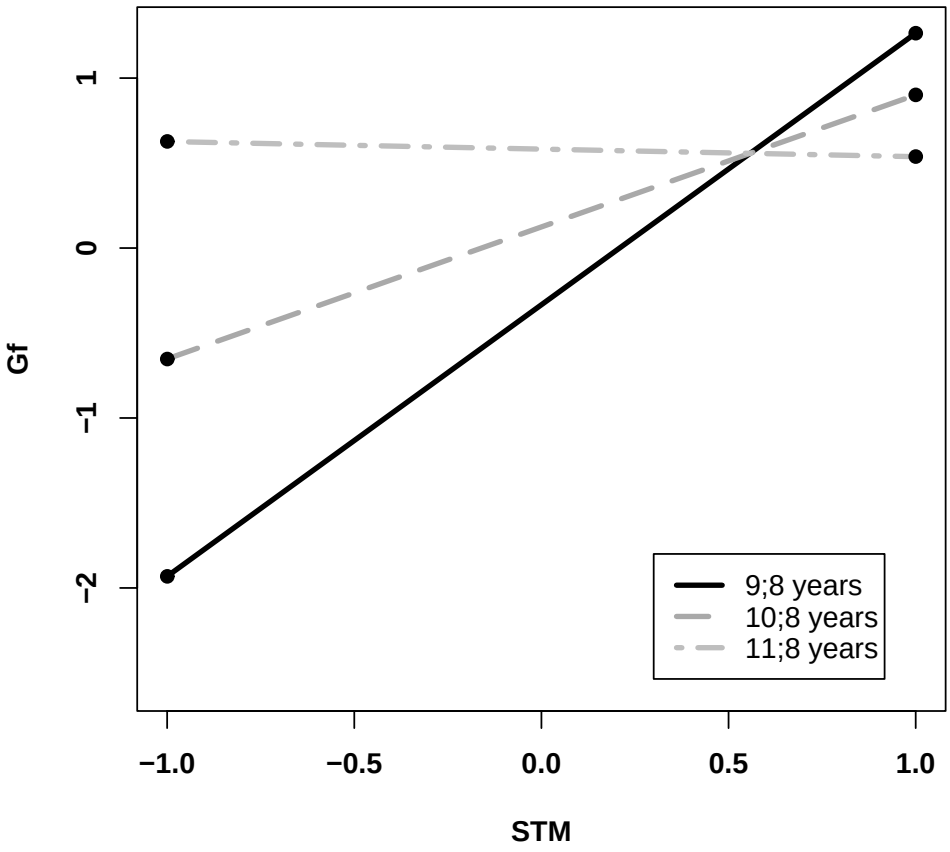


Figure 2.4: Mean plot illustrating the interaction of STM and age

With respect to Gc, it should be mentioned that in the full model (model HR5), only WM, STM, and age were significant predictors. This finding is similar to the results of the structural equation models above. Hence, as expected, age predicted Gc but not Gf after WM and STM were entered into the model. However, in model HR5, no substantial interaction effect occurred, underscoring the fact that neither STM nor WM lost predictive power in older children for Gc. However, things looked differently when focusing on the residual of Gc when controlling for Gf. Here, age was the dominant predictor, and STM only marginally contributed to the improvement of the model, while WM was entirely irrelevant when entered into the model at a later stage. This finding slightly differs from the results obtained in the structural equation models above, in which Gc tests were modeled as loading on both Gf and Gc. The regression model spec-

ified for Gcres, therefore, was an extremely strict test. However, in contrast to the findings reported by other authors (e.g., Martínez & Colom, 2009), the findings in Table 2.6 show that Gcres still substantially correlated with WM and STM. Hence, even when controlling for all Gf variance in Gc, WM and STM are substantially related to Gc. The same was true for WMres, which exhibited substantial correlations with Gf and Gc. For both Gc and Gcres, similar results were obtained when using WMres instead of WM.

It should be noted here that the results obtained in model HR5 for Gf, Gc, and Gcres can be cast in the framework of a moderated mediation analysis (cf. Preacher, Rucker, & Hayes, 2007). That is, we might be interested in assessing the indirect effect that WM exerts on Gf via STM, for example, and whether this effect is partly moderated by age. It would be theoretically plausible that the importance of STM as a predictor for intellectual abilities like Gf diminishes with age, but remains high for knowledge (i.e., Gc; Swanson, 2008). We therefore computed the indirect effects of WM on Gf, Gc, and Gcres, respectively, with STM as a mediator. In a first step, a mediator model predicting STM from WM and age was computed, where both WM (b = .49, p < .01) and age (b = .15, p < .05) were significant predictors, whereas they did not interact (b = -.02, p = .74). The indirect effect is the product of the path from WM to STM and the path from STM to the cognitive ability under investigation (e.g., Gf). In the case that the effect of WM on cognitive abilities is partly transmitted by STM, the indirect effect should be statistically significant. In order to evaluate whether this indirect effect is moderated by another variable (e.g., age), the indirect effect is evaluated using specific values of the moderator (usually, the mean and $\pm$ 1 standard deviations). Table 2.8 illustrates that whereas the indirect path WM $\rightarrow$ STM $\rightarrow$ Gc is always relevant in predicting Gc irrespective of age, indicating that WM affects Gc by storage capabilities in all children, the indirect path WM $\rightarrow$ STM $\rightarrow$ Gf is insignificant for older children (compare Figure 2.4). That is, in older children, WM does not affect Gf indirectly through storage capabilities, whereas in younger children, STM is a partial mediator.

2.4 Discussion

In this study, we pursued three main goals, namely, a thorough investigation of the structure of memory and intelligence in children, an analysis of stability of memory across age groups, and an investigation of age effects on the different

Table 2.8: Analysis of indirect effects in moderated mediation

Variable	z_{age}	$a_1 * b_1$	SE	Z	p
Gf	-1	.78	.16	4.78	.00
	0	.38	.10	3.71	.00
	1	.00	.13	.01	.99
Gc	-1	.23	.07	3.34	.00
	0	.18	.05	3.73	.00
	1	.13	.06	2.02	.04
Gcres	-1	.00	.07	-.06	.95
	0	.06	.05	1.37	.17
	1	.13	.07	1.82	.07

Note. z_{age} = Values of z-standardized moderator variable age; $a_1 * b_1$ = Indirect path coefficient WM → STM → complex cognition (Gf, Gc, Gcres); SE = Bootstrapped standard error of indirect effect (1,000 draws; Preacher et al., 2007).

relations between memory and intelligence facets. Key findings of this study pertain to a relatively stable structure of WM and STM across age as well as differing relations of WM and STM with fluid and crystallized intelligence that were partly moderated by age. We address these findings in turn.

Similar to Swanson (2008), our findings indicate that WM and STM are separable factors in children, although the factors were more closely related in younger than in older children. This finding is in line with work by Alloway et al. (2006) who found very high correlations between STM and WM in young children that decreased with age (cf. Gathercole et al., 2004). This evidence supports the notion of an incremental differentiation of STM and WM. For example, Engle et al. (1999) suggested that because rehearsal and chunking capabilities are less well developed in younger children, an executive component of working memory must be activated earlier in STM tasks, thus making STM and WM tasks more similar in younger children. In fact, we found that STM and WM factors correlated at r = .79 in children younger than 10 years, whereas the same factors correlated at r = .55 in older children. Further, MI analyses revealed that there was no substantial qualitative differences in the memory structure of younger and older children, although latent mean differences of medium size were observed. The strong MI model showing adequate fit here supports the notion that these differences were based on developmental constraints and did not represent

a measurement artifact. However, the idea that WM and STM can be merged into a single factor (Hutton & Towse, 2001) had to be rejected. The fact that WM and STM are dissociable is supported by numerous results from the behavioral and neuroimaging domains (Engle et al., 1999; Fletcher & Henson, 2001). However, although STM and WM measures were structurally and factorially different here, we assume that both mainly capture SM, which has recently been suggested as the key variable in higher cognitive functioning (Mogle et al., 2008; Unsworth & Engle, 2007b). Tasks differed in structure, though, possibly affecting intelligence measures as discussed below.

In line with Cowan et al. (2005), we found evidence for the fact that the scope of attention is closely related to WM and, in fact, had to be merged into a single factor with complex span measures. That is, although the task structure was very different from the traditional complex span tasks (i.e., no processing task, no overt rehearsal), the visual array comparison task captured variance that was closely related to traditional measures of WM. Hence, we could not find support for the assumption that this task is a measure of PM, which was clearly separable from both WM and SM in earlier work (Mogle et al., 2008). Our results are in agreement with evidence reported by other authors (e.g., Alloway et al., 2006; Bayliss et al., 2003; Miyake et al., 2001), who found that visuo-spatial STM tasks without a processing component highly correlated with WM, possibly due to the need for a higher involvement of executive functioning in visuo-spatial STM tasks than in verbal STM tasks. For example, Bayliss et al. (2003) showed that a visuo-spatial STM task (Corsi task) correlated at r = .62 with fluid intelligence, whereas digit span as a verbal-numerical STM measure only correlated at r = .17.

Our results further support a domain-general model of WM in that a distinction between visuospatial and verbal-numerical WM factors was not tenable, in contrast to other findings based on samples of children (e.g., Jarvis & Gathercole, 2003; Tillman et al., 2008). However, the sample reported in Jarvis and Gathercole (2003) was relatively small, whereas Tillman et al. (2008) used only two WM tasks that showed a substantial zero-order correlation (r = .57). Because we used structurally heterogeneous WM tasks, in contrast to other studies (Kane et al., 2004; Tillman et al., 2008), our findings strongly support a domain-general factor, in that the generality of WM extends to dual-task and single-task measures of WM. However, more research is clearly needed with respect to this issue.

Concerning the separability of inductive and deductive reasoning, it was

found that a single factor was sufficient for representing Gf. Hence, in accordance with results obtained by Wilhelm (2005), a factorial separation of deductive and inductive reasoning is not tenable in children. This supports the assumption by Harman (1999) and others that similar reasoning processes are utilized for deductive and inductive reasoning. With respect to the relation of Gf and Gc, a high correlation (r = .90) was found, despite the fact that the Gc tests used in this study used verbal and numerical content, whereas all Gf tests were figural. The magnitude of this correlation, which contrasts other findings (e.g., a Gf-Gc correlation of r = .35 in Haavisto & Lehto, 2004), was unexpected, although it is not uncommon that Gf and Gc are very closely related (Carroll, 1993). The Gc tests utilized here obviously required reasoning to a significant degree, which was underlined by results obtained in the structural equation models predicting Gf and Gc from WM, STM, and age. In this model, we specified Gc tests to additionally load on Gf, thus taking the hybrid character of these tests into account, and controlling for Gf variance in Gc.

Unsurprisingly, structural equation modeling revealed that both WM and to a lesser degree STM, when conceptualized as distinct but correlated factors, were important in predicting Gf performance, resulting in an R^2 of .89 at the latent level. Age, in contrast, was irrelevant once these variables had been taken into account, supporting the notion that additional relevant psychological variables (e.g., processing speed) that increase with age were not necessary to explain differences in Gf. These findings were supported by hierarchical regression analyses, but they contrast evidence reported by Swanson (2008, Model 3, p. 597), who found that age continued to be a predictor of Gf even when STM, WM, and other factors (e.g., processing speed) were taken into account. However, processing speed was not related to Gf in Swanson (2008) when used in concert with WM and STM, although it has been shown to be substantially related to WM in some studies with children (e.g., Bayliss et al., 2003; Fry & Hale, 1996). The fact that age was insignificant once WM and STM were taken into account in our study, however, could be interpreted in a way such that developmental differences in processing speed, as they would manifest in the age variable in our context, may not be substantial for predicting complex cognition (Mogle et al., 2008).

The statistically significant interaction between STM and age and the Johnson-Neyman regions of significance indicated that STM was not a substantial predictor of Gf from approximately 11 years on, although WM remained central irrespective of age. A complementary moderated mediation analysis revealed that

STM worked as a mediator of WM capacity on Gf in children younger than 11 years, that is, storage abilities partly explain the effect of WM on Gf in younger, but not older children. STM span grows during childhood (Chuah & Maybery, 1999; Logie & Pearson, 1997), and whereas STM tasks can be assumed to require executive attention in younger children because their storage capabilities are lower, possibly due to less-developed rehearsal processes (Baddeley, 1986; Henry & Millar, 1993), older children might have differentiated executive and storage processes to a larger degree, rendering the effect of storage capabilities on reasoning ability insignificant. In line with this assumption, Alloway et al. (2006) noted that developmental increases in STM span level off between 10 to 11 years of age, whereas WM span is still subject to developmental differences at that age.

In addition, in contrast to Mogle et al. (2008) and others (cf. Unsworth & Engle, 2007b), but in line with recent evidence obtained in another sample of children (Tillman et al., 2008), we found that WM capacity remained a substantial predictor for both Gf and Gc when it was conceptualized as a residual factor, although it did not differ from STM in predictive power. This was the case even though the STM measures utilized in this study, due to consisting of supraspan list lengths, represented SM, and despite the fact that we used a partial scoring algorithm for all complex span tasks, which has been shown to render a WM residual factor insignificant (Unsworth & Engle, 2007b). In addition, similar to Maybery and Do (2003), we did not use classical simple span tasks as a measure of STM, but rather measures of supraspan list length, thereby excluding short lists. In combination, this should have made it even harder for a residual WM factor to attain statistical significance in predicting higher cognitive functioning above a common storage factor.

What does the residual WM factor represent, then? If we regard our STM measures as capacity measures of SM, a residual WM factor might be conceptualized as controlled attention, which implies the inhibition of irrelevant information as well as the activation of relevant information in the context of interfering stimuli (cf. Engle et al., 1999). Apparently, this factor plays a major role for higher cognitive functioning in children. In order to predict Gf in children, therefore, both storage and controlled attention are equally relevant, as evidenced by a similar magnitude of path coefficients. This finding is consistent with the position of Hasher, Lustig, and Zacks (2007), who propose that inhibitory control is the key factor in WM. In line with this reasoning, Swanson (2008) reported a strong

relation between inhibition and WM as well as intelligence measures. The lack of interaction between age and WM in explaining cognitive abilities, therefore, reveals that controlled attention is a consistently crucial factor for higher cognitive abilities in children.

Different results were obtained with respect to crystallized intelligence. Structural equation modeling results support the notion of WM and STM as important predictors for Gc. However, age remained substantially related to Gc in all analyses. This is consistent with our expectation that verbal and numerical processing knowledge are tied to school curricula, which should result in higher Gc scores with age irrespective of cognitive abilities. Of note, STM mediated the effect of WM on Gc across all age groups, indicating that the ability to store and retrieve learned information is of key importance for crystallized intelligence across all age groups. This is also consistent with the result that when Gc is statistically purified of all Gf variance, only age mattered as a predictor in model HR5, although STM added slightly to the fold when being the only additional predictor. However, substantial zero-order correlations between Gcres, WM, and STM remained. We therefore could not replicate the findings reported by Martínez and Colom (2009) who found that WM was unrelated to a Gc residual controlling for Gf. In children, however, both storage and controlled attention, therefore, appear to be related to a "purified" knowledge factor.

Overall, several key findings can be extracted from this study. Firstly, a general WM factor, including structurally heterogeneous tasks with and without a processing component, was found, which could be statistically separated from a STM factor, resulting in a two-factor structure of memory in children. This was the case even though STM tasks consisted of lists of supraspan length representing SM, although WM and STM were highly correlated. Secondly, when comparing younger and older children with respect to the measurement structure of these memory tasks, it was found that strong MI was tenable, although WM and STM were more highly correlated in younger than in older children and substantial latent mean differences were observed. Thirdly, deductive and inductive reasoning were inseparable in children, whereas Gf and Gc were separable but highly related. And finally, controlled attention, when conceptualized as a residual WM factor controlling for STM, was found to be substantially related to both Gf and Gc, whereas Gf did not depend on STM performance in children older than 11 years.

2.4.1 Limitations

Like all studies, this work has several limitations. Firstly, we slightly modified the visual array comparison task compared to Cowan et al. (2005), especially with respect to the interstimulus intervals. Longer interstimulus intervals generally result in the decay of stimuli (Barrouillet, Bernardin, Portrat, Vergauwe, & Camos, 2007), thus countervailing maintenance or rehearsal activities and potentially activating SM processes, especially in larger set sizes of the visual array comparison task. The task version used here therefore might reflect a heterogeneous measure, capturing both the scope of attention and SM. Although our results are in line with Cowan et al. (2005), therefore, the lack of separability between the scope of attention and WM might have been exacerbated by the task structure of the visual array comparison task. Consequently, we did not include a relatively pure measure of PM in this study, which would have been helpful in gauging whether WM would then lose its predictive power for higher cognitive functioning (Mogle et al., 2008). Future research should integrate measures of SM and PM, both with respect to simple and complex span tasks, and systematically manipulate factors that are known to affect the degree to which PM and SM are captured (e.g., list length, serial position, inhibiton; Unsworth & Engle, 2007b) in order to provide a clearer picture of which combination of elementary cognitive processes are responsible for the high relation between SM, PM, and complex cognition.

Observed power is often computed in a post-hoc fashion to gauge the probability of finding effects that are present in the data. We did not compute observed power here, however, because no additional information is gained from doing so (Hoenig & Heisey, 2001). This has to do with the fact that observed power is closely related to the obtained p value from the test of the null hypothesis in structural equation modeling (cf. MacCallum, Browne, & Cai, 2006). However, it can generally be said that in the case of MI analyses, large sample sizes are preferable because they provide more precise results and higher statistical power (Meade et al., 2008). In this study, sample size was relatively small, therefore small violations of MI might not have been detected due to a lack of power. However, we found substantial differences in the factor variances and covariance as well as latent mean differences of moderate size between age groups, indicating that power was sufficient to detect notable effects.

Further, we confined our analysis of interaction terms to aggregated ob-

served variables. Although the analysis of interactions on the level of latent variables is becoming increasingly common (cf. Marsh, Wen, & Hau, 2004), this topic remains an open field for research with several unresolved issues. In addition, due to the complexity of the models specified (e.g., nested factor structures), we analyzed interactions on the level of manifest indicators exclusively. Again, this might have lowered power due to the fact that measurement error is not explicitly taken into account; however, the effects found in the hierarchical regression models were in line with theoretical considerations.

3 Children's performance on equations and its relationship with working memory, intelligence, and facets of processing speed

Summary. Numerous studies on individual differences in mathematical abilities have shown that working memory substantially affects arithmetic performance. In this study, we extended this research to algebra problem solving. A total of 376 8- to 13-year old children were administered algebra problems and measures of working memory, intelligence, arithmetic ability, literacy, and processing speed. Further, the effects of number size and memory load were investigated. A new modeling approach was utilized to simultaneously analyze effects of item- and person-level predictors on algebra ability and solution speed, respectively. At the item level, number size showed no effect, memory load substantially affected item difficulty and time intensity. At the person level, working memory remained a substantial predictor for algebra ability even in the face of arithmetic ability, literacy, and age. Results do not support the notion of a domain-general working memory model, whereas access to information held in working memory during problem solving appears to be a key predictor for intellectual functioning.

3.1 Introduction

Working memory (WM), although separable from general intelligence, has been shown to be one of the core processes of human intellectual functioning for a wide range of tasks (Ackerman, Beier, & Boyle, 2005; Kyllonen & Christal, 1990). Consequently, WM plays a central role for mathematical abilities as well. For example, several studies have shown that arithmetic calculation skills are strongly related to WM in children (Berg, 2008; Bull & Scerif, 2001; Rasmussen & Bisanz, 2005) as well as adults (Fürst & Hitch, 2000; Heathcote, 1994; Logie, Gilhooly, & Wynn, 1994). However, only few studies have investigated the relationship between WM and performance on equations while taking additional cognitive resources into account. The goal of the current study, therefore, was to examine the relationship of algebra achievement with WM, intelligence and facets of processing speed in a sample of children, and to combine this individual differences approach with an experimentally-designed test in order to evaluate theories pertaining to algebra performance.

3.1.1 WM and attentional control: Domain-general or domain-specific?

Different models of WM have been established in the literature (cf. Conway, Jarrold, Kane, Miyake, & Towse, 2007; Miyake & Shah, 1999). Generally, WM is regarded as a set of modules or processes that allow the storage of information as well as the simultaneous manipulation of the same or other information. The latter element distinguishes WM from more basic forms of memory such as short-term memory.

The first and still most widely-researched WM model was introduced by Baddeley and Hitch (1974). This structural model consists of two components for temporarily storing information, the phonological loop and the visuo-spatial sketchpad, respectively. Whereas the phonological loop stores speech-based information and is sometimes referred to as "verbal WM", the visuo-spatial sketchpad is used to maintain visual-spatial information. A third component, the central executive, represents a coordinating entity that directs information towards the relevant subsystems, inhibits irrelevant information, and coordinates activity. The central executive is assumed to have no storage capacity. Baddeley (2000) later added a new subsystem, the episodic buffer, which is proposed to handle some problems of the multiple-component model, such as how verbal information can be stored during articulatory suppression.

Another important conceptualization of WM was suggested by Engle and coworkers (e.g., Engle, Tuholski, Laughlin, & Conway, 1999). It differs from the model suggested by Baddeley and Hitch (1974) in several ways. Firstly, it is more process-oriented than structural, i.e. it defines WM in terms of relevant processes and not modules. Secondly, this WM model is mainly concerned with *controlled attention*, which these authors define as follows (Kane, Conway, Bleckley, & Engle, 2001, p. 180):

> By "controlled attention" we generally mean an executive control capability; that is, an ability to effectively maintain stimulus, goal, or context information in an active, easily accessible state in the face of interference, to effectively inhibit goal-irrelevant stimuli or responses, or both.

Controlled attention, i.e., WM capacity, which roughly corresponds to the central executive in the model by Baddeley and Hitch (1974), is usually measured by complex span tasks like reading span or computation span. These tasks gen-

erally require subjects to maintain information in an active state while simultaneously processing other information. Although not pure measures of controlled attention, they are "reasonably good measures of a domain-general attentional capability" (Kane, Conway, Hambrick, & Engle, 2007, p. 24). Hence, the executive control capability is described as a central, domain-general resource relevant for a variety of complex cognitive tasks, whereas short-term memory is assumed to tap rather domain-specific storage resources (cf. Kane et al., 2004). In this view, complex span tasks (e.g., reading span), as operationalizations of controlled attention, should be able to predict performance in complex cognitive tasks from different domains (e.g., arithmetic). However, Shah and Miyake (1996) provided evidence for a dissociation of verbal and visuo-spatial controlled attention. These authors only report weak relationships between reading span and a psychometric test of spatial visualization ability, contrary to what could have been expected. Some further studies, both in children and adolescents (Jarvis & Gathercole, 2003; Leather & Henry, 1994; Mackintosh & Bennett, 2003; Tillman, Nyberg, & Bohlin, 2008) as well as adults (Süß, Oberauer, Wittmann, Wilhelm, & Schulze, 2002), found that verbal and visuo-spatial WM factors could be separated as well, indicating domain-specific controlled attention. Other results, again in children (Alloway, Gathercole, & Pickering, 2006; Hitch, Towse, & Hutton, 2001) as well as adults (Engle et al., 1999; Conway, Cowan, Bunting, Therriault, & Minkoff, 2002; Kane et al., 2004), found that WM tasks could be considered as forming a single, general factor. The question of whether controlled attention is domain-general or domain-specific therefore has not yet been settled.

3.1.2 WM and arithmetic calculation

Numerous studies have shown that WM capacity and arithmetic calculation are closely related, both in children and adults (cf. DeStefano & LeFevre, 2004). Relating this finding to specific arithmetic operations, WM capacity is required for addition (Adams & Hitch, 1997), subtraction (Seyler, Kirk, & Ashcraft, 2003), multiplication (Seitz & Schumann-Hengsteler, 2000) as well as division (Imbo & Vandierendonck, 2007). In a series of studies, Bull and colleagues (Bull & Johnston, 1997; Bull, Johnston, & Roy, 1999; Bull & Scerif, 2001) could show that the central executive reliably predicted arithmetic ability, but that neither the phonological loop nor the visuo-spatial sketchpad were directly related to arithmetic performance in children when controlling for reading ability or age. Fürst and

Hitch (2000) found that articulatory suppression, which supposedly blocks the phonological loop, does not affect on arithmetic calculation in adults when the problem was constantly visible. In another study (K. Lee, Ng, Ng, & Lim, 2004), the phonological loop and the visuo-spatial sketchpad failed to affect performance in algebraic word problems directly as well. Also, a recent meta-analysis indicates that children with mathematics difficulties can be differentiated from unimpaired counterparts mainly by worse performance on complex span tasks with verbal content (Swanson & Jerman, 2006). In line with these results, Leather and Henry (1994) provided evidence for a small effect of short-term memory for predicting arithmetic performance when entered into a regression model first, which failed to reach significance once complex span tasks were taken into account. Several studies using dual-task methodology generally support this evidence. For example, Logie et al. (1994), comparing performance on visually-presented addition problems in conditions with secondary tasks tapping components of the WM model by Baddeley and Hitch (1974), found that the strongest disturbance was exerted by tapping the central executive during mental calculation, whereas tapping the phonological loop and the visuo-spatial sketchpad had a much smaller effect. This result is in line with evidence presented by Thomas, Zoelch, Seitz-Stein, and Schumann-Hengsteler (2006) who found that in a dual-task paradigm, only taxing the central executive lead to deterioration in math performance in children (cf. De Rammelaere, Stuyven, & Vandierendonck, 1999). Taking age effects into account, Floyd, Evans, and McGrew (2003) in a large-scale study presented results that support the highest relationship with executive components of WM and arithmetic abilities in children 11 to 13 years old. Finally, Berg (2008) reported that complex span tasks with verbal and visuo-spatial content independently contributed to arithmetic performance in children 9 to 12 years old, even when controlling for age, reading, short-term memory and processing speed. In one of the few studies lacking an effect of WM, Mayes, Calhoun, Bixler, and Zimmerman (2009) report findings indicating that arithmetic ability was related to IQ, graphomotor speed, and visuo-motor integration in a sample of children from kindergarten through fifth grade, but not to WM. However, WM was only conceptualized using a digit span task by these authors.

In contrast, several other studies report effects of short-term memory on arithmetic calculation. For example, in the study by Andersson (2008), digit span was significantly related to arithmetic calculation in a sample of children 9 to 10 years old, although it was not related to solving arithmetical equations. In addi-

tion, Berg (2008) presents evidence for a relationship between short-term memory and arithmetic calculation when controlling for age and reading ability, although no information is provided on how this relationship changed when complex span tasks were taken into account. Fuchs et al. (2006) investigated the relationship of WM and arithmetic ability while controlling for variables as phonological decoding, attention, and processing speed, and found that WM was neither related to simple (e.g., 3 + 2) nor complex (e.g., 35 + 29) mental addition, although these authors did not take complex span tasks with visuo-spatial content into account. Rasmussen and Bisanz (2005) report a strong relationship between arithmetic calculation and the visuo-spatial sketchpad, but they only investigated pre-school children and grade 1 students. However, especially in younger children, short-term memory and WM capacity are closely related. For example, in a sample of 7- to 9-year-old children, Bayliss, Jarrold, Gunn, and Baddeley (2003) found that when controlling for short-term memory variance in WM, WM capacity did not correlate with fluid intelligence. In adults, the opposite was found (Engle et al., 1999), i.e., a WM residual was strongly related to fluid intelligence. This possibly indicates that controlled attention is used in short-term memory tasks by younger children, and that WM and short-term memory processes are more similar in children than adults (cf. Hutton & Towse, 2001). Overall, in older children, the phonological loop and the visuo-spatial sketchpad only appear to play a minor role for arithmetic calculation, whereas controlled attention or facets of the central executive appear to be of paramount importance.

One of the most robust findings in research on mental arithmetic is the *problem-size effect*, i.e. mental calculations become slower and more error prone with larger numbers (e.g., 7×8) than with smaller ones (e.g., 2×3; Ashcraft, 1992; Campbell & Graham, 1985; LeFevre, Sadesky, & Bisanz, 1996). Several studies have maintained that the reason for this finding lies with the more frequent utilization of non-retrieval strategies in mental arithmetic with larger numbers. For example, Penner-Wilger, Leth-Steensen, and LeFevre (2002) analyzed response times from simple and more complex multiplication problems using the ex-Gaussian distributional model. These authors found that in Chinese students who primarily utilized retrieval strategies, the problem-size effect showed only in the mean of the normal component of response times (μ), representing less efficient retrieval in calculation problems with larger numbers. In contrast, in Canadian students, the problem-size effect was related to both μ and τ, the latter representing the mean of the exponential component. A recent study (Schmiedek,

Oberauer, Wilhelm, Süß, & Wittmann, 2007) has revealed that τ is closely related to the involvement of higher cognitive functions (WM), i.e., the group of Canadian students in the study by Penner-Wilger et al. (2002) used additional procedures differing from mere retrieval in problems with larger numbers. This finding is consistent with results reported by Hecht (2002), who, in a dual-task study, reported greater interference of a central executive load in more difficult problems in the case that non-retrieval strategies (e.g., transformation, counting) were involved (cf. LeFevre et al., 1996). This result is also supported by neurophysiological evidence (Jost, Hennighausen, & Rösler, 2004) and additional results that show an interaction trend between problem size and response latencies (Lemaire, Abdi, & Fayol, 1996). Further, from an individual differences perspective, higher WM capacity has been shown to be related to more frequent and efficient utilization of retrieval in mental arithmetic (Barrouillet & Lépine, 2005; Imbo & Vandierendonck, 2008). Hence, participants with higher WM capacity should be more prone to use a retrieval strategy in mental arithmetic and therefore should exhibit shorter response times than participants with lower WM capacity, whereas accuracy should be less unaffected.

3.1.3 WM and algebraic performance

Equations are mathematical statements that use the equal sign to indicate that two mathematical expressions are (or are defined to be) equivalent. In addition to basic arithmetic skills, solving an equation correctly requires switching between operational and structural views of mathematical expressions (Sfard & Linchevski, 1994). That is, the relationships between variables and numbers in an equation must be analyzed first, before the necessary arithmetic computations are carried out to provide a solution. Hence, the process of solving equations can be considered a generalized form of arithmetic computation in that it does not involve specific numbers, but rather variables and functions (Carraher, Schliemann, Brizuela, & Earnest, 2006). In addition to being familiar with arithmetic computation, solving equations as well as algebraic understanding therefore require the ability to cognitively represent the relationship of equation entities correctly (Humberstone & Reeve, 2008).

An important prerequisite for solving equations is the correct understanding of the equal sign (Knuth, Stephens, McNeil, & Alibali, 2006). Whereas a relational understanding of the equal sign (both sides of the equation must be

equivalent) allows to solve equations correctly, children often have incorrect conceptions of the equal sign. In their change-resistance account of algebraic dysfunctionalities in children, McNeil and Alibali (2005) mention several operational patterns that prevent children from correctly solving equations. For example, during arithmetic training in school, pre-algebraic children learn that the equal sign (e.g., 2 + 3 + 4 + 5 = ?) means "the total". Thus, they are unable to correctly solve the equation 7 + 4 + 5 = 7 + ? and provide the solution 23, corresponding to the sum of all numbers presented. McNeil and Alibali (2005) could successfully show that early-learned arithmetic procedures and operational patterns can hinder the development of algebraic thinking in children, and that undergraduate students primed with these operational patterns performed worse when solving equations. Thus, under some circumstances, prior arithmetic knowledge can actually deteriorate performance in solving equations. Nevertheless, algebraic thinking can be taught efficiently, and it can be practiced well. Neuropsychological evidence (Qin et al., 2004) has shown that pre-algebra students were nearly equivalent in solution accuracy to adults after 5 days of training, and they exhibited less activity in pre-frontal and parietal brain regions after practice, indicating a reduced involvement of higher cognitive functions.

Several studies have investigated the relationship of WM with achievement in higher-level domains of mathematics, although very few have investigated the relations between cognitive resources and algebra alone. Whereas in Engle et al. (1999), WM was related to performance on the Scholastic Aptitude Test - Mathematics (SAT-M), although short-term memory was not, Rohde and Thompson (2007) found that when controlling for fluid intelligence, processing speed, vocabulary and spatial ability, WM was not related to college students' SAT-M performance. Reuhkala (2001) showed that high school students' performance on national maths exams were correlated with spatial-short term memory (cf. Holmes, Adams, & Hamilton, 2008; St. Clair-Thompson & Gathercole, 2006) and mental rotation ability, but not with a reading span task. However, no complex span tasks using numerical or spatial content were utilized in this study. In the work conducted by Tolar, Lederberg, and Fletcher (2009), investigating a sample of college students, the correlations of an algebra equations test with four verbal-numerical complex span tasks were relatively low and largely insignificant (r = .06 – .14), whereas the relationship of spatial ability as well as numerical fluency with algebraic achievement was substantial. Using structural equation modeling, these authors could show that WM was related to algebra achievement only in-

directly by spatial visualization and arithmetic ability. Again, these authors did not use a complex span task with spatial content. A recent study by Andersson (2008), investigating the ability to solve simple arithmetic equations (e.g., ? + 25 = 500) in third-to fourth grade school children, found that controlling for fluid intelligence and age, several complex span tasks (counting span, visual-matrix span, verbal fluency) were significantly related to performance, whereas short-term memory was not. In summary, results on the relationship between algebraic performance and WM capacity, especially in children, are not unequivocal and clearly necessitate further research.

Some studies have investigated the effect of concurrent memory load on algebra performance. For example, Anderson, Reder, and Lebiere (1996) studied the effect of concurrent memory load (2, 4, or 6 items) on solving equations. In the first condition, the memory load was irrelevant to the equation-solving task, whereas in the second condition variables in the equations had to be substituted by the first or second digit from the memory set. Although the size of the memory set affected problem-solving latency and accuracy in both conditions, the effect was larger in the substitution condition. In Oberauer, Demmrich, Mayr, and Kliegl (2001), three conditions of solving equations were compared: No memory load, memory load with irrelevant items not accessed by working memory, and memory load with items that had to be accessed during solving the equations. In the latter condition, the processing task required access to contents stored in working memory, whereas in the condition with irrelevant items, no access to the stored items during processing was required. Oberauer et al. (2001) indeed found that in the access condition, mean response times as well as proportion of correct responses were much lower than in the irrelevant load or no load conditions, which did not differ. The authors concluded that the results are difficult to reconcile with a resource-sharing account between processing and storage in WM (e.g., Just & Carpenter, 1992), because this model assumes no difference between whether content stored in WM is accessed or not. Rather, they hypothesized that cross-talk between items held and accessed in WM result in more errors and slower responses. According to these authors, storing intermediate results in memory and accessing them later during equation solving should result in a general drop of performance, both with respect to processing speed as well as response accuracy.

3.1.4 Processing speed as a contributory process

The speed of processing information within WM is crucial for the efficiency with which cognitive operations can be carried out. For example, Case, Kurland, and Goldberg (1982) reported a linear relationship between counting speed and counting span in children 6 to 12 years old, indicating that faster information processing is related to higher WM capacity. The authors interpreted this finding as strong evidence that a higher processing efficiency results in less cognitive resources required and, hence, the ability to store more information. Building on this finding, Bayliss, Jarrold, Baddeley, Gunn, and Leigh (2005) showed that when processing speed as well as storage ability were taken into account, age was no longer a relevant predictor for complex span tasks in a sample of children aged 6 to 10 years. Also, Fry and Hale (1996) found that age differences in WM were largely accounted for by age-related changes in processing speed. From these findings, it can be concluded that processing speed plays a major role in cognitive development.

Several studies have investigated the role of processing speed in arithmetic calculation. In a group of 7-year-old children, Bull and Johnston (1997) found that processing speed was the strongest predictor of arithmetic performance, even when controlling for short-term memory, speech rate, and item identification. These authors did not measure WM capacity, however. Similarly, Barrouillet and Lépine (2005) showed elementary school children with higher WM capacity were also more likely to use fact retrieval in simple addition and, thus, were able to solve simple addition tasks more quickly. However, in several studies, processing speed was unrelated to arithmetic calculation when additional cognitive variables were taken into account. For example, in Swanson and Beebe-Frankenberger (2004), who investigated children from the first to third grade, only reading, age, and WM contributed unique variance to math calculation. A similar result was obtained in Berg (2008), where processing speed failed to predict arithmetic calculation once age and reading were taken into account. Hitch et al. (2001) note that WM had a considerably larger effect on basic arithmetic skills than processing speed in children 9 to 11 years old.

One of the reasons for this conflicting state of affairs might be the conceptualization of processing speed. Often, mean or median reaction times, mostly of correct trials only, were used as equivalent to processing speed. This approach, however, has some disadvantages. Firstly, by focusing on correct trials only, infor-

mation is lost. Secondly, this approach ignores the speed-accuracy trade-off, i.e., the strong inverse relationship between response speed and response accuracy (e.g. Wickelgren, 1977). Finally, mean or median reaction times are not motivated by psychological theory. Hence, it is unclear whether longer reaction times reflect slower information processing, slower motor processes (e.g., in pressing answer buttons), or a more conservative criterion until an answer is provided. Evidence based on simple reaction times, therefore, can be misleading.

Diffusion models, first introduced by Ratcliff (1978), can overcome these shortcomings. Diffusion models can be used to analyze response times from two-choice tasks (cf. Wagenmakers, 2009). By taking the reaction time distributions of both correct and error responses as well as accuracy into account, diffusion models allow a detailed analysis of facets of processing speed. A simplified version, the EZ-diffusion model (Wagenmakers, van der Maas, & Grasman, 2007), focuses on three central parameters that can be related to underlying psychological processes, mean drift rate (ν), boundary separation (a) as well as nondecision time (T_{er}). The drift rate quantifies the "ease" of information processing in a two-choice task. It is high in the case that decisions are fast and accurate, whereas it is low in the case of slower, more error-prone decisions. The drift rate ν is generally assumed to lie out of participants' control. In a recent study relating diffusion model parameters to WM, Schmiedek et al. (2007), using structural equation modeling, found that ν correlated at r = .68 with a WM factor and at r = .79 with a reasoning factor. Boundary separation, in contrast, is assumed to be under subjective control. It reflects response caution, i.e., for a participant who is carefully trying to avoid erroneous answers, the boundaries are set widely apart, which results in higher accuracy but slower response times. Finally, the nondecision time parameter T_{er} incorporates encoding or response (motor) processes that are unrelated to the decision process. The EZ-diffusion model is conceptually similar to signal-detection theory, although it additionally integrates information from response times and accuracy.

Recent advances in statistical modeling have further resulted in approaches that allow the simultaneous modeling of ability and processing speed in computerized psychometric tests. For example, van der Linden (2007) developed an item response theory (IRT) model that, in addition to classically estimating person ability and item parameters (e.g., difficulty), allows the estimation of person speed and item parameters such as time intensity. Thus, by dissociating time intensity and person speed, the model circumvents the problem of directly equating

person speed with reaction times, which is problematic in case that items differ in the amount of information processing necessary. In contrast, the model assumes fixed time intensities for each item, but assumes that person speed is a random variable that varies individually. Hence, it becomes possible to assess the relationship of unbiased person parameters that relate to ability and speed, respectively. Building on this approach, Klein Entink, Kuhn, Hornke, and Fox (2009) developed a model that further allows to estimate the effects of specific item properties (e.g., one-digit vs. two-digit numbers) on item parameters such as difficulty or time intensity. That is, the model by Klein Entink et al. (2009) provides the possibility to relate the cognitive processes required to solve the item to item difficulty and time intensity, respectively, and to assess to which degree the postulated cognitive model structure fits empirically.

3.1.5 Purpose of the present study

In review, only few studies have investigated the role of working memory in children's arithmetic or algebraic performance, several challenges therefore remain. Firstly, in predicting arithmetic or algebraic performance, many studies have treated WM as a unitary system in the past. However, it remains unclear whether WM can be considered a domain-general construct or not. Hence, complex span tasks from different content domains and their unique effect on solving equations should be taken into account. In the case of a domain-general WM model, all tasks should predict algebra performance, regardless whether verbal, numerical, or spatial content is used. Secondly, the effect of facets of processing speed on algebra performance is of interest. Especially, the relationship of parameters estimated with the diffusion model to equation solving needs to be investigated. Finally, the effect of to-be-accessed stimuli stored in WM in solving equations as well as problem size on both speed and accuracy in solving equations will be analyzed.

In order to rule out alternative explanations, several contributory processes or abilities should be controlled for. Age is known to be an important general factor (e.g., Kail & Park, 1994), and we were interested in measures that accounted for variation when entered later. Given the causal relationship between IQ and achievement (Watkins, Lei, & Canivez, 2007), it is also critical to take IQ into account when determining predictors of achievement. Further, the role of arithmetic skills in equation solving is of interest. Because algebra is assumedly

cognitively demanding, WM should be of importance even after controlling for arithmetic skills. Finally, reading-related abilities must be taken into consideration, because they show a strong relationship with arithmetic calculation (e.g., Hecht, Torgesen, Wagner, & Rashotte, 2001; Swanson & Beebe-Frankenberger, 2004), possibly because some arithmetic strategies afford verbal processing. Recently, Lonigan et al. (2009) found that vocabulary tests are highly correlated with phonological awareness (r = .73 – .74), one of the key variables in language acquisition. Vocabulary tests therefore appear good candidates for capturing central reading-related abilities.

3.2 Method

3.2.1 Subjects

Three-hundred seventy-six children from three higher-track secondary schools in various regions of Germany participated in this study. Mean age was 11;3 years (SD = 0.97, range: 8;9-13;8). 37.4% of the participants were female. Parental consent was obtained for all participants prior to testing. Few participants (n = 13) indicated German was not their first language, although all of these participants spoke German since they were 3 years old.

3.2.2 Measures

WM Tasks

Three types of computer-based complex span tasks were used, one comprising verbal material, one with numerical and one with visuo-spatial material. The three tasks used were Verbal Span, Spatial Working Memory, and Computation Span, respectively. The complex span tasks utilized in this study are based on Vock and Holling (2008). These authors adapted several complex span tasks from the literature (e.g., Daneman & Carpenter, 1980; Oberauer, Süß, Schulze, Wilhelm, & Wittmann, 2000) to be appropriate for children 8 to 13 years old. Each complex span task first presented between two to three simple practice tasks to subjects, including immediate feedback. All subjects had to repeat the practice tasks until each had been solved correctly.

WM tasks generally comprise several items and subitems. For example, in this study, computation span comprised 10 items, each consisting of three to seven equations displayed on the screen (subitems). The subitems (in the case of computation span, the results of the equations) were the contents that had to be remembered. Recently, it has been shown that partial scoring (i.e., computing the sum of proportion of correctly-solved subitems for each item) results in better psychometric properties and higher correlations with measures of fluid intelligence than other scoring procedures, presumably because more information is retained (Unsworth & Engle, 2007b). We therefore used partial scoring for all WM tasks.

Verbal Span This WM task (Oberauer et al., 2000; Vock & Holling, 2008) had two different parts. Participants first had to memorize a list of words presented simultaneously on the screen (presentation time 6 s). List length in this storage task varied between three to six words. Then, between two and three verbal decision tasks followed in which participants had to respond as quickly as possible. In these processing tasks, participants had to decide which of four words displayed in each corner of the screen stood in a subconcept relation to the word shown in the center of the screen (e.g., "animal" - "lion"). Finally, participants were supposed to reproduce the learned words in correct order. The task consisted of two practice items and 10 test items.

Spatial Working Memory Participants had to memorize simple chessboard-like 3×3-patterns (storage task). However, the patterns had to be stored in a rotated fashion, rotated either 90° clockwise or counterclockwise (processing task). That is, before the patterns were shown successively for 4 s each, an arrow indicated whether patterns had to be mentally rotated to the left or to the right. Finally, participants had to successively reproduce the memorized patterns into empty 3×3 matrices on the screen. The task consisted of 13 items with between one to four patterns. Three practice items preceded the testing phase.

Computation Span In this task, participants were sequentially shown a series of simple, single-digit equations that included either an addition or a subtraction (e.g., 4 + 3 = 8). Each equation was shown for 5 s. Approximately half of the equations were correct and half were incorrect. The processing task consisted in deciding whether the equation shown on screen was correct or incorrect. Further, all shown equation results had to be memorized irrespective of whether they were correct or not. Finally, subjects were presented with an answer screen and successively clicked the to-be-remembered equation results. Each item con-

sisted of between three to seven items, resulting in 10 test items. Two practice items were administered before the testing phase.

Fluid intelligence

In order to determine fluid intelligence (IQ), the short form of the Grundintelligenztest Skala 2 (CFT 20; Weiß, 1998), a German adaptation of the Culture Fair Intelligence Test, Scale 2 (Cattell, 1973), was utilized. In one school (n = 125), the revised form of this test, the CFT 20-R (Weiß, 2006), was utilized[1]. CFT 20 and CFT 20-R are paper-and-pencil tests which provide high loadings on fluid intelligence (Cattell, 1968) and have good psychometric properties. They consist of four different subtests: Series completion, Classifications, Matrices and Topologies. Overall testing time, including instructions for each subtest, was approximately 20 minutes.

Arithmetic skills

We used a test consisting of relatively simple number series that is part of the CFT 20 (Weiß, 1998) to measure arithmetic skills. On each item of this test, participants were presented six numbers in a row, followed by an empty cell with a question mark. Out of five distractors, they had to select the correct answer which correctly continued the sequence of numbers presented. Only one- or two-digit numbers were utilized, and the complexity of the cognitive processes involved was low to medium such that the mastery of all elementary arithmetic operations was in the focus of interest. The test consisted of 21 items that had to be solved in 16 minutes. Four practice items were administered before testing began.

Vocabulary test

In this test, which is also part of the CFT 20, participants were presented 30 target words. Next to each target word, five distractors were given. Participants had to select the distractor that had the same meaning as the target word. Overall, 30

[1] The CFT 20-R contains some additional items and a new norm sample. In order to make these measures comparable, we used the aggregated scores from the WM tasks to equate IQs. This did not substantially affect results reported later, therefore, raw IQs from each test form were retained.

test items were preceded by three practice items. The time limit for this test was 12 minutes.

Mental speed

In this study, a two-choice task with visual content, the visual array comparison task (VACT; cf. Luck & Vogel, 1997), was chosen. In this computer-based task, participants first saw a red fixation cross for 500 ms in the center of a grey 4 $\times$ 4-matrix on the screen. After that, a visual array of four, six, or eight solid-colored, haphazardly-placed squares, representing set sizes of four, six, and eight, respectively, was displayed within the matrix. Set sizes were randomly ordered across trials. The square colors used were red, blue, violet, green, yellow, black, and white. Care was taken that at least one color was displayed twice in each initial visual array such that subjects had to memorize both color and location of the squares (Cowan, Naveh-Benjamin, Kilb, & Saults, 2006). On half of the trials, the first visual array was displayed for 250 ms and for 500 ms on the other half of the trials. After a blank interval, a second visual array was displayed in which one of the squares was encircled. The participants then had to decide whether the color of the encircled square had changed in comparison to the first visual array or not. On 50% of the trials, the color of the encircled square had changed. The length of the interstimulus interval was either 1 s, 2 s, or 4 s and equally distributed across trials to raise the difficulty of the task. However in visual array comparison tasks even very complex stimuli show a half-life of about 3 s (Cornelissen & Greenlee, 2000). Three practice trials with feedback preceded 48 test trials, including an equal number of trials for each set size.

We first screened for RT outliers, which were iteratively identified based on individual RT distributions. For each participant, RTs smaller than 500 ms and larger than 4 individual standard deviations were excluded from further analysis. This criterion was chosen because it represented an acceptable tradeoff between deleting RTs clearly away from the rest of the RT distribution and keeping the shape of the distribution intact. On average, less than 1 RT per participant was excluded. In order to estimate individual parameters ν, a, and T_{er} in the EZ-diffusion model, the proportion of correct responses, the mean of correct response times, and the variance of correct response times are required. We used the formulas provided by Wagenmakers et al. (2007) to calculate EZ-diffusion model parameters directly in closed form.

Algebra test

A new algebra test to model the effects of memory load and number size in solving equations was designed. A detailed instruction on how to solve algebraic equations was provided first. Special care was taken to promote a relational understanding of the equal sign, and how to conduct arithmetic transformations to maintain equality on both sides of the equation. All basic arithmetic operations (addition, subtraction, multiplication, division) were distributed across items ($n_{Addition}$ = 14, $n_{Substraction}$ = 13, $n_{Multiplication}$ = 12, $n_{Division}$ = 8) and had to be carried out correctly in different combinations.

We manipulated memory load by designing items in which intermediate results had to be stored in memory to obtain the correct solution. Memory load varied between 0 to 2 items. A typical test item with a memory load of 2 was

- $B + 14 = 3 \times C$
- $A - 16 = B$
- $D = C + B - 10$
- $A = 35$
- $D =?$

In this example, three equations needed to be solved successively, and two intermediate results had to be repeatedly accessed WM in order to provide the correct solution. Twelve items had a memory load of 0, seven had a memory load of 1, and three had a memory load of 2, respectively. Items with memory load of 1 or larger required for solving the equations in descending order of complexity, such that subjects had to solve for the most difficult unknown first (in the example, D). In the next step, they were supposed to solve for the next unknown with a lower memory load (in the example, C), until items without memory load were reached (in the example, solving for B).

The second factor that was manipulated was number size. Concerning multiplication, problems with both operands > 5 were seen as large, whereas they were seen as small if both were $\leq$ 5. If one operand was smaller than 5 and the other larger, problem size was classified according to Campbell and Tarling (1996). Ties (e.g., 9 + 9, 4 $\times$ 4) were avoided. Analogous classification schemes for addition, subtraction, and division were used (e.g., Imbo & Vandierendonck,

2008). We then computed an average number size indicator across all operations necessary to solve each item, with a possible range from 0 to 1. After 5 practice items with feedback, the testing phase comprising 22 items commenced. There was no time-limit, i.e., the algebra test utilized here was a power test.

3.2.3 A Model for Response Accuracies and Response Times

To address the research questions outlined in the previous section, a modeling framework is needed that models the children's arithmetic ability and their processing speed simultaneously. Subsequently, it should be possible to relate person level covariates (e.g., measures of WM) to ability and speed as well as controlling for item characteristics.

van der Linden (2007) proposed a model that jointly models a person's ability and speed level on a test that uses separate measurement models for ability and speed, respectively. At a higher level, a population model for the person parameters (ability and speed) is deployed to take account of the possible dependencies between the person parameters. This framework was further developed by Klein Entink, Fox, and van der Linden (2009) and Klein Entink et al. (2009), who extended it to allow for explanatory variables on the person and item side, respectively. Therefore, this model is well suited for the current research. Figure 3.1 gives a schematic representation of the model. Below, the model is described in more detail. For an extensive discussion of the joint modeling of response accuracies and RTs on psychometric tests, the reader is referred to van der Linden (in press).

Measurement Models for Accuracy and Speed

The left oval in Figure 3.1 denotes the measurement model for arithmetic ability. The ability level of a child is represented by θ, while a and b are parameters that describe the characteristics of an item, denoting its discriminative capacity between persons of different ability levels and its difficulty level, respectively. Together, these parameters describe the observed variability in the responses $\mathbf{Y}$ over items and persons. Mathematically, the probability that person $i = 1, \ldots, N$ answers item $k = 1, \ldots, K$ correctly ($Y_{ik} = 1$), is assumed to follow the two-

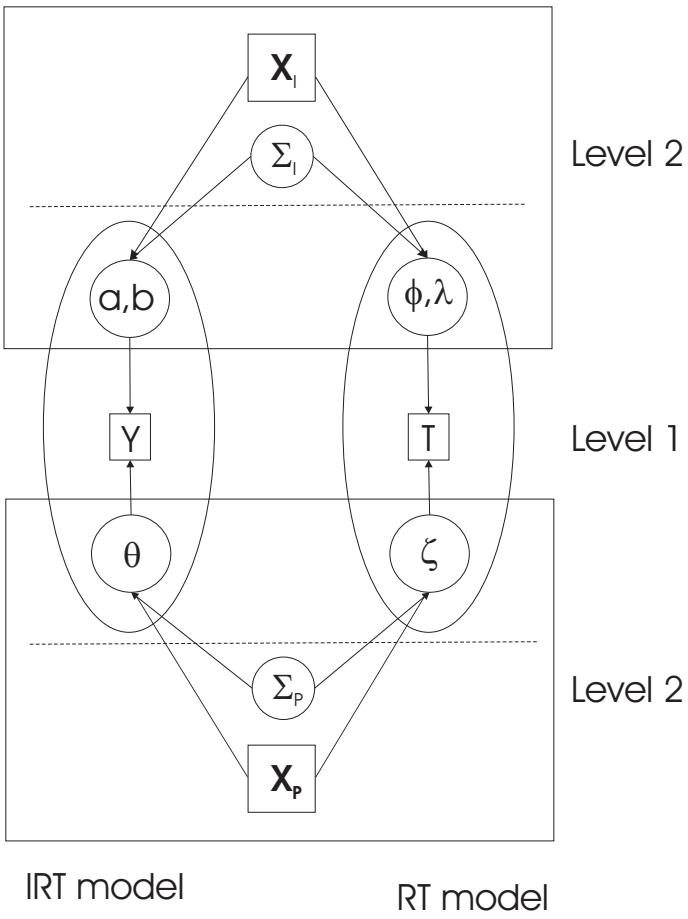


Figure 3.1: Modeling framework

parameter normal ogive IRT model:

$$P(Y_{ik} = 1|\theta_i, a_k, b_k) = \Phi(a_k\theta_i - b_k), \tag{3.1}$$

where θ_i denotes the ability parameter of test taker i and a_k and b_k denote the discrimination and difficulty parameters of item k, respectively. $\Phi(\cdot)$ denotes the cumulative normal distribution function.

Two characteristics of this model are important to mention: First, it is assumed that the probability of a correct response increases with ability. This is represented in the subfigure on the left in Figure 3.2 where the expectation of a correct response $E(Y)$ is plotted against ability θ for two items that differ in their difficulty b. Second, it is assumed that the ability level of the person explains all associations between the observed responses to different items. This is known as the *local independence* assumption in the IRT literature.

The oval on the right hand side in Figure 3.1 denotes the measurement model for speed. This model for the response times $\mathbf{T}$ has a similar parameter structure as the IRT model. It is assumed that a person works with a constant speed ζ during the test. This assumption is plausible in the case of the algebra test as it has no time-limit. The speed parameter is the equivalent of the ability parameter. Like ability, speed is assumed to be the underlying construct for the RTs and, conditional on speed, the RTs on a set of items are assumed to be conditionally independent. The item parameters λ and ϕ account for differences in time intensity and discriminative ability of the items, respectively. That is, λ models the expected RT on an item and thereby allows for differences in the time consumingness of items. Since RTs are bounded by 0 and have a skewed distribution, it is assumed that the log-response time T_{ik} of person i on item k follows a normal model according to:

$$T_{ik} = -\phi_k \zeta_i + \lambda_k + \epsilon_{\zeta_{ik}}, \tag{3.2}$$

where $\epsilon_{\zeta_{ik}} \sim N(0, \sigma_k^2)$ models the residual variance. Note that the minus sign reflects that persons working at a higher speed have a lower expected RT. On the right hand side of Figure 3.2, the expected RT as a function of speed is plotted for two items with different time intensities.

These two measurement models form the basis of the modeling framework used in this paper. The dependencies between the responses and response times are modeled at a second level.

Level 2 Model for the Item Parameters

A possible source of covariation between responses and RTs are the items in the test. For instance, an item that requires multiple processing steps to obtain the solution can make that item both relatively difficult and time consuming for the test takers. As a result, we would expect a positive correlation between b and λ. To model such dependencies, the vector of item parameters $\boldsymbol{\xi}_k = (a_k, b_k, \phi_k, \lambda_k)$ is assumed to follow a multivariate normal distribution (MVN),

$$\boldsymbol{\xi}_k \sim N(\boldsymbol{\mu}_I, \boldsymbol{\Sigma}_I) \tag{3.3}$$

where $\mathbf{\Sigma}_I$ specifies the covariance structure of the item parameters:

$$\mathbf{\Sigma}_I = \begin{bmatrix} \sigma_a^2 & \sigma_{ab} & \sigma_{a\phi} & \sigma_{a\lambda} \\ \sigma_{ab} & \sigma_b^2 & \sigma_{b\phi} & \sigma_{b\lambda} \\ \sigma_{a\phi} & \sigma_{b\phi} & \sigma_\phi^2 & \sigma_{\phi\lambda} \\ \sigma_{a\lambda} & \sigma_{b\lambda} & \sigma_{\phi\lambda} & \sigma_\lambda^2 \end{bmatrix}. \tag{3.4}$$

In Figure 3.1 this structural model on the item parameters is depicted by the square on the upper side of the figure, symbolizing the possible dependencies between the item parameters. Furthermore, it is possible to explain (a proportion of the) variance in the item characteristics as a function of known design features of the items. In the figure, these design features are represented as the item covariate matrix $\mathbf{X}_I$.

More specifically, in our application the following design features were known: number size (NM), memory load (ML), and repetition (RP). These item covariates may contain useful information why certain items are more difficult or time intensive. Therefore, following Klein Entink et al. (2009), the time intensity and difficulty of the items are modeled as a function of these covariates:

$$b_k = \gamma_{b0} + NM_k\gamma_{b1} + ML_k\gamma_{b2} + RP_k\gamma_{b3} + NM_k * ML_k\gamma_{b4} + e_{b_k} \tag{3.5}$$

$$\lambda_k = \gamma_{\lambda 0} + NM_k\gamma_{\lambda 1} + ML_k\gamma_{\lambda 2} + RP_k\gamma_{\lambda 3} + NM_k * ML_k\gamma_{\lambda 4} + e_{\lambda_k}, \tag{3.6}$$

where $\boldsymbol{\gamma}$ denotes the matrix of regression effects and the error terms are assumed to follow a MVN distribution, together with the residuals of a and ϕ, with covariance matrix $\mathbf{\Sigma}_I$ as given in Equation 3.3.

Level 2 Model for the Person Parameters

The main research question is how covariates like WM, intelligence and facets of processing speed of children relate to their performance in solving a series of mathematical equations. Their performance on the mathematical equations is measured from the responses and RTs on the arithmetic test, and represented by their ability and speed levels, θ, ζ, respectively. To relate the latter to the measures of working memory, intelligence and processing speed, a multivariate regression model similar to the model for difficulty and time intensity above is developed. This multivariate model that allows for explaining variability in the ability and speed levels of test takers as a function of covariates was proposed by

Klein Entink et al. (2009).

The simplest model that describes dependencies between ability and speed is to assume a bivariate normal distribution:

$$(\theta_i, \zeta_i) = \boldsymbol{\mu}_P + \boldsymbol{e}_P, \boldsymbol{e}_P \sim N(\mathbf{0}, \boldsymbol{\Sigma}_P), \tag{3.7}$$

where $\boldsymbol{\mu}_P = (\mu_\theta, \mu_\zeta)$ and the covariance structure is specified by:

$$\boldsymbol{\Sigma}_P = \begin{bmatrix} \sigma_\theta^2 & \sigma_{\theta\zeta} \\ \sigma_{\theta\zeta} & \sigma_\zeta^2 \end{bmatrix}, \tag{3.8}$$

where $\sigma_{\theta\zeta}$ models the covariance between ability and speed. Like the possible dependencies between item characteristics, also the dependencies between ability and speed of test takers can be a source of covariation between the responses and RTs. It is important to model these different sources of covariation and to separate the item effects from the person effects (van der Linden, in press).

In Figure 3.1, the structural model for ability and speed is represented by the lower square. The (residual) covariance between θ, ζ is modeled by $\boldsymbol{\Sigma}_P$. The matrix $\mathbf{X}_P$ contains person level covariates that might explain a proportion of the variance in ability and speed. For instance, two such covariates could be the age of the children (AGE) and a measure of their working memory capacity (WM). Then, the following model for ability and speed of child i would be obtained:

$$\theta_i = \gamma_{\theta 0} + AGE_i\gamma_{\theta 1} + WM_i\gamma_{\theta 2} + e_{\theta_i} \tag{3.9}$$

$$\zeta_i = \gamma_{\zeta 0} + AGE_i\gamma_{\zeta 1} + WM_i\gamma_{\zeta 2} + e_{\zeta_i}, \tag{3.10}$$

where $\boldsymbol{\gamma}_\theta, \boldsymbol{\gamma}_\zeta$ are the regression effects on ability and speed, respectively, and the errors follow a multivariate normal distribution with mean 0 and the covariance matrix given by $\boldsymbol{\Sigma}_P$.

Thereby, the full modeling framework as represented in Figure 3.1 thus allows us to measure the ability and speed of the children on the algebra test from their responses and RTs and to relate these to measures of intelligence, processing speed or other background information like age. In addition, effects of experimental variables at the item level, as described above, can be evaluated simultaneously. In the next section the estimation of the model and the testing of hypotheses is discussed.

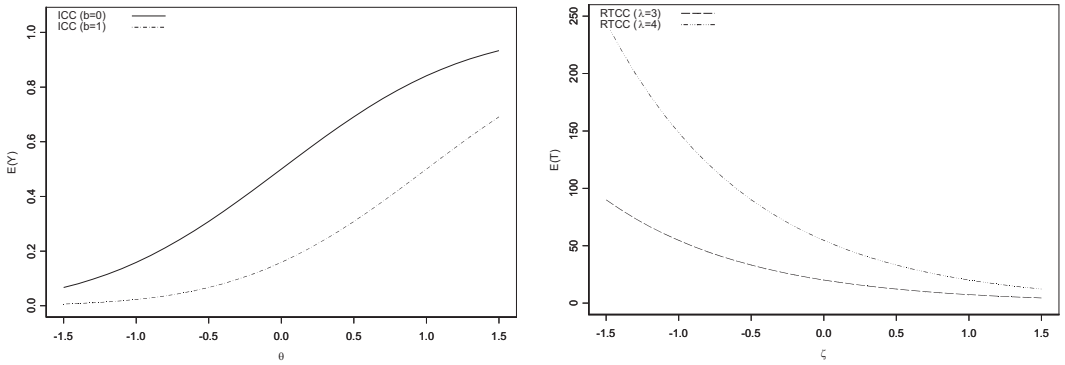


Figure 3.2: ICC and RTCC curves for two items with differing difficulty and time intensity, where $a = \phi = 1$

3.2.4 Statistical Inference

This section briefly discusses the estimation procedures and some statistical tests to assess model fit and to be able to test hypotheses of interest. Since the full model is a complex multivariate multilevel structure with many parameters, estimation and testing of the model is all performed within the Bayesian statistical framework. The Bayesian approach facilitates the use of Markov Chain Monte Carlo (MCMC) methods, which use simulation-based algorithms to obtain estimates of the model parameters. It is beyond the scope of this paper to go into the technical details of these algorithms. Therefore, we refer the interested reader to the papers by van der Linden (2007); Fox, Klein Entink, and van der Linden (2007) and Klein Entink et al. (2009) for the specifics regarding the model used here. Software to estimate the models is available as a package for use in the statistical environment *R* on the website of the second author. However, before we discuss some aspects of model selection and model fit, we will briefly discuss the principles of the Bayesian approach and MCMC for statistical inferences. For a more thorough introduction to Bayesian statistics, we refer the reader to Gelman, Carlin, Stern, and Rubin (2004).

3.2.5 MCMC algorithm

In the Bayesian approach, a model parameter is a random variable with a probability distribution. Statistical inference focuses on the marginal posterior distribution of the parameters of interest. The posterior distribution of the model parameters are obtained by first specifying a prior distribution that reflects the prior

uncertainty of the researcher about the parameters before seeing the data. Subsequently, when data about these parameters are gathered the prior distribution is updated and the posterior distribution of all model parameters is obtained. Inferences can then simply focus on the marginal distribution of the model parameters of interest, ignoring possible nuisance parameters. The latter requires that these nuisance parameters are integrated out, which is difficult analytically for complex models with many parameters. To circumvent this problem, MCMC methods are used to approximate the full posterior distribution by obtaining draws from a density that is proportional to the posterior. MCMC algorithms construct a Markov chain with the joint posterior distribution of the model parameters as its equilibrium distribution. More specifically, a complex multivariate distribution from which it is hard to sample is broken down into smaller univariate distributions, conditional on the other model parameters, from which sampling is straightforward. After providing the algorithm with arbitrary starting values for all parameters, it alternates between the conditional distributions for M iterations. For our model, the algorithm proceeds like:

1. Generate starting values for all parameters
2. For iteration m, draw $\theta^{(m)}$ from $p(\theta|\zeta^{(m-1)}, \mathbf{y}, \mathbf{t}, \mathbf{x_P},$ (*all other parameters*))
3. Draw ζ from $p(\zeta^{(m)}|\theta^{(m)}, \mathbf{y}, \mathbf{t}, \mathbf{x_P},$ (*all other parameters*))
4. Obtain draws for all other model parameters
5. Repeat step 2-5 until M draws from the joint posterior have been obtained

When M draws of the algorithm have been obtained, first, stationarity tests have to be performed to check the convergence of the algorithm. To do so, the BOA package for use in *R* or *SPLUS* can be used to evaluate several statistical tests that give an indication about convergence of the MCMC chain (Smith, 2007). Since the algorithm is provided with arbitrary starting values, a burn-in period of the algorithm is estimated and these samples are discarded. After this burn-in, stationarity of the chain is assumed and from the remaining samples of the chain, summary statistics of the model parameters can be obtained for inferences.

Model Assessment

To check some model assumptions and for model selection purposes, we briefly introduce the evaluation of test statistics in the Bayesian framework.

Since parameters have a probability distribution, statistical inferences are based on the posterior distribution given the observed data. An interval summary of the posterior is the Highest Posterior Density (HPD) region of level $(1-\alpha)$. For instance, a 95% HPD region contains (1) 95% of the density region and (2) parameter values inside the interval have higher probability than parameter values outside the interval. This property can be used for regression coefficients to evaluate if the parameter value 0 is contained in the level $(1-\alpha)$ HPD region or not.

The appropriateness of a statistical model can be assessed by means of posterior predictive checks. The general principle is to compare replications of the data, drawn from the posterior distribution under the hypothesized model, with the observed data by means of an appropriate test statistic (for example, mean squared error). When using MCMC algorithms for estimation, these statistics can be obtained as a by product of the algorithm, simply by drawing a new dataset under the model after each iteration of the algorithm and comparing it to the observed data. For more details, see Gelman et al. (2004) and Gelman, Meng, and Stern (1996). We used the following statistics to assess the appropriateness of our model:

- An Odds Ratio statistic to test for local independence in the IRT model, as proposed by Sinharay (2005),
- An observed score statistic, that gives an impression of overall model fit by comparing the observed sum scores of the test takers with their replicated sum scores under the model (Sinharay, 2005; Sinharay, Johnson, & Stern, 2006),
- A Bayesian residual analysis for the RT model (van der Linden & Guo, in press), that assesses the discrepancy between observed and expected RTs under the model.

For model comparison, non-nested models can be compared by means of the Deviance Information Criterium (DIC), which is a deviance statistic with a penalty term for model complexity (Spiegelhalter, Best, Carlin, & van der Linde, 2002). The DIC is useful to evaluate the discrimination parameters in the IRT and RT model, since these enter the model as a product with ability $(a\theta)$ and speed $(-\phi\zeta)$.

A Bayes factor (Kass & Raftery, 1995; Klugkist, Laudy, & Hoijtink, 2005) can be used to test a model M_1 against another model M_0 for the observed data $\mathbf{y}$. The

Bayes factor is defined as the ratio of the marginal likelihoods of these models:

$$BF = \frac{p(\mathbf{y}|M_0)}{p(\mathbf{y}|M_1)}. \tag{3.11}$$

Since the Bayes factor weighs the two models against each other, a value near one means that both models are equally likely. A value of 3 or greater is considered to be strong evidence in favor of the null model, while on the contrary a value near zero favors the larger model as the best explanation for the data (Kass & Raftery, 1995). We used the Bayes factor to test several nested regression models of covariates on θ, ζ and on b, λ.

Furthermore, the Bayesian R^2 statistic as proposed by Gelman and Pardoe (2006) was used to assess the proportion of explained variance in ability and speed, and item difficulty and time intensity, by the person and item level covariates. An R^2 value near 1 means that the covariates explain almost all observed variability in the parameters, while a value near 0 means that the variance in the model parameters almost equals the error variance. In contrast to the R^2 computed in classical linear regression, the Bayesian R^2 can actually be larger if less predictors are in the model, because predictors without any relevance for the criterion add noise to the Bayesian R^2, thus reducing its magnitude.

3.3 Results

First, descriptive statistics and results of dimensionality and model analyses of the algebra test are presented. Second, the effects of item-level predictors (memory load, number size) on algebraic performance are analyzed within the methodological framework described above. Third, we analyzed the results of person-level predictors on both algebraic ability and speed.

Descriptive statistics and reliabilities are presented in Table 3.1. For the EZ-diffusion model parameters, no internal consistencies could be computed. As can be seen from the IQ data, the sample investigated here was above average in cognitive ability, reflecting the fact that all students came from the highest track of the education system. We proceeded with a dimensionality analysis of the algebra test. We used NOHARM (Fraser & McDonald, 1988) to investigate accuracy data, which is one of the best statistical procedures to assess unidimensionality (Finch & Habing, 2007). NOHARM revealed a good fit to unidimensionality, with

Tanaka's goodness-of-fit index at .95 (Tanaka, 1993).

Table 3.1: Descriptive statistics and reliability estimates

Variable	M	SD	Skewness	Kurtosis	K-S[a]	α
Working memory						
1. Verbal Span	6.93	1.73	-.43	-.27	.05*	.81
2. Spatial WM	7.12	2.25	-.22	-.33	.04	.73
3. Computation Span	8.47	1.30	-1.76	5.48	.13**	.78
Fluid intelligence						
4. IQ CFT 20[b]	117	12.55	-.05	-.23	.05	.91
5. IQ CFT 20-R[c]	120	13.94	-.10	.15	.08*	.82
Arithmetic skills						
6. Number series[d]	60	8.10	.00	-.18	.07**	.88
Reading-related abilities						
7. Vocabulary test[d]	59	6.99	.10	.00	.07**	.85
Mental speed						
8. ν	.06	.02	.35	-.06	.06**	N/A
9. a	.27	.07	.33	-.29	.04	N/A
10. T_{er}	1.01	.37	-.34	.31	.08**	N/A
Algebra performance						
11. Algebra test	16.94	3.76	-.67	-.10	.12**	.82

Note. [a] $Z(p)$ of Kolmogorov-Smirnov-test on normal distribution with correction of significance by Lilliefors. [b] n = 221. [c] n = 155. [d] T-values.
*p < .05. **p < .01.

In the next step, we fit four different bivariate IRT models to the algebra test, and used the DIC to select the one that fit the data best (Gelman et al., 2004). Table 3.2 provides a summary of the results. As indicated by the DIC, Model 4, assuming two item parameters for both accuracy data and response times, respectively, clearly fit the data best. All further analyses are therefore based on this model.

The MCMC algorithm was run for 15,000 iterations and its output was analyzed using the BOA package. Aspects like autocorrelation of the chains, Geweke's Z-statistic for stability of the chain, Heidelberg's stationarity test and Gelman's convergence diagnostic were assessed. The results suggested that stability of the MCMC chain was reached for all parameters after 300 iterations. We decided to discard the first 5,000 iterations and base our inferences on the last 10,000 MCMC samples. This applies to all analyses reported below.

The posterior model fit checks were based on 2,000 replicated data sets under the model. We evaluated the Bayesian residuals for the RT model for each

Table 3.2: Deviance summaries for the measurement models

Model	Parametrization[a]	DIC
1	1PNO, 1PRT	21227.31
2	2PNO, 1PRT	21017.25
3	1PNO, 2PRT	20783.74
4	2PNO, 2PRT	20565.64

Note. [a]1PNO = 1-parameter normal ogive model, 2PNO = 2-parameter normal ogive model, 1PRT = 1-parameter RT model, 2PRT = 2-parameter RT model.

item graphically, using quantile-quantile plots of the observed residuals against their expected values under the model. These plots did not suggest any serious flaws. The fit to the response data was acceptable, too. Only for a few possible item combinations did the odds-ratio statistic point at a violation of local independence (a Bayesian p-value $< .025$ or $> .975$). The observed sum score statistic suggested that the replicated data sets under the model reflected the observed data. From Figure 3.3 in the Appendix it can be seen that the model slightly underpredicted the number of people who answered 6 and 8 items correctly, but in general described the data well.

Table 3.4 summarizes all variance-covariance estimates of Model 4. Ability and speed parameters were modestly correlated (r = .24), indicating that good algebra problem solvers tended to work faster. Further, time intensity and difficulty parameters showed a substantial relationship (r = .64), i.e., more difficult items required more time to be solved. In addition, whereas more difficult algebra items had a higher discriminatory power with respect to ability, the same was not true for more time-intensive items concerning speed.

We proceeded by investigating the effects of item-level predictors on both item difficulty and time intensity, respectively. Three predictors were of interest: Memory load, number size, and repetition. Repetition referred to items with memory load, in which a result had to be retrieved that had been computed in a prior step. We hypothesized that both memory load and number size would raise time intensity, whereas only memory load, but not number size would affect item difficulty. Repetition was expected to lower both item difficulty and time intensity. Finally, we investigated the interaction between number size and memory load, expecting a larger effect of memory load for larger numbers.

Table 3.3: Estimated proportions of explained variance (in b, λ) for models M_{I0} - M_{I2}

Model	$R^2(b)$	$R^2(\lambda)$
M_{I0}	0.00	0.00
M_{I1}	0.18	0.52
M_{I2}	0.32	0.61

Note. M_{I0} = Empty model (no predictors), M_{I1} = Memory load, number size, repetition, and memory load × number size as predictors, M_{I2} = Memory load as predictor.

Table 3.4: Estimated covariance components and correlations, Model 4

Variance components		EAP[a]	SD	r
$\mathbf{\Sigma}_P$				
	Σ_{11}	1.00	-	1.00
	Σ_{12}	0.08	0.02	0.24
	Σ_{22}	0.10	0.01	1.00
$\mathbf{\Sigma}_I$				
	Σ_{11}	0.18	0.06	1.00
	Σ_{12}	0.20	0.08	0.79
	Σ_{13}	0.13	0.05	0.74
	Σ_{14}	0.06	0.06	0.23
	Σ_{22}	0.36	0.12	1.00
	Σ_{23}	0.13	0.07	0.52
	Σ_{24}	0.25	0.06	0.64
	Σ_{33}	0.16	0.08	1.00
	Σ_{34}	-.03	0.06	-.09
	Σ_{44}	0.40	0.13	1.00

Note. Σ_P = Variance-covariance parameters on person level, Σ_I = Variance-covariance parameters on item level. [a]Expected a posteriori parameter estimate.

Table 3.3 provides an overview of the proportions of explained variance in both time intensity and item difficulty. Only memory load had a substantial effect, which it exerted on both time intensity and item difficulty. Hence, taking memory load into account, there was no effect of number size. Further, there

was no difference between smaller and larger numbers to be stored in WM. In addition, repetition had no effect on item difficulty or time intensity, suggesting that intermediate results were not simply retrieved from memory. Parameter estimates of model M_{I2} are summarized in Table 3.5, underscoring the substantial effect of memory load on both item difficulty and time intensity.

Table 3.5: Estimated effects and .95 HPD regions for model M_{I2}

	Effect	EAP	SD	.95 HPD[a]
Discrimination (a)	γ_{100} (intercept)	0.71	0.10	[0.52,0.92]
Difficulty (b)	γ_{200} (intercept)	-1.21	0.14	[-1.48,-.94]
	γ_{201} (Memory load)	0.59	0.17	[0.26,0.92]
Time Discrimination (ϕ)	γ_{300} (intercept)	1.09	0.10	[0.89,1.27]
Time intensity (λ)	γ_{400} (intercept)	9.33	0.12	[9.10,9.57]
	γ_{401} (Memory load)	0.98	0.19	[0.61,1.35]

Note. [a]95% Highest posterior density intervals of parameter estimates.

Having established that only at the item-level, only memory load was substantially related to ability and speed in solving equations, we proceeded by investigating the effect of person-level predictors. A sequence of hierarchical regression models was fitted to the data, relating all person-level variables to both algebra ability and speed, respectively. Results are shown in Table 3.6. More variation in ability than speed was explained when all predictors were taken into account. Of note, WM variables still exerted a substantial effect on algebra ability when all other person-level variables were taken into account, whereas the effect was much smaller for speed in solving equations. In contrast, parameters from the diffusion model appeared more important in predicting speed in solving equations than algebra ability.

Several variables were unrelated to algebra performance or speed, respectively, in model M_{P5}. The ability parameter in model M_{P5} was unaffected by IQ, ν, and Verbal span, whereas the speed parameter was unaffected by Spatial WM and Verbal Span. We therefore excluded these variables from further analysis and estimated a more parsimonious model (M_{P6}). The comparison of M_{P6} and M_{P5} yielded a Bayes factor of $BF = \exp(13.8)$, providing strong evidence in favor of model M_{P6}. Table 3.7 provides an overview of parameter estimates in model M_{P6}. Arithmetic ability, as could have been expected, was a strong predictor of algebraic reasoning. Surprisingly, IQ did not affect when all other variables were

Table 3.6: Estimated proportions of explained variance in θ, ζ (models M_{P0} - M_{P6})

Model	$R^2(\theta)$	$R^2(\zeta)$
M_{P0}	0.00	0.00
M_{P1}	0.04	0.10
M_{P2}	0.15	0.17
M_{P3}	0.16	0.22
M_{P4}	0.37	0.31
M_{P5}	0.41	0.32
M_{P6}	0.41	0.32

Note. Predictors in the models: M_{P0} = Null model, M_{P1} = Age, M_{P2} = Age + IQ, M_{P3} = Age + IQ + ν + a + T_{er}, M_{P4} = Age + IQ + ν + a + T_{er} + number series + vocabulary test, M_{P5} = Age + IQ + ν + a + T_{er} + number series + vocabulary test + Computation Span + Verbal span + Spatial WM, M_{P6} = As M_{P5} but with predictor selections specific for ability and speed.

taken into account. It did affect speed of algebra problem solving, however, with age being the strongest predictor.

Additionally, we checked whether the single item-level predictor, memory load, interacted with WM capacity (both Computation Span and Spatial WM) in predicting algebra ability. Substantial interaction effects would indicate that high-WM subjects process memory load during solving equations differently. However, no substantial interaction effects were found, supporting the notion of similar processing in high- and low-WM participants, respectively.

3.4 Discussion

Building on a dearth of research on arithmetic calculation and algebra performance in children, this study sought to illuminate the relative contributions of WM, intelligence, and components of processing speed in children's algebra performance as well as speed. Results provide further evidence for the role of WM and related cognitive processes in solving equations. Results suggested three important findings. First, WM had a substantial effect on algebra ability, even when controlling for age, reading-related abilities, arithmetic ability, and facets of processing speed. Second, two WM tasks contributed unique variance to algebra ability, contradicting the assumption of a domain-general WM model. Third, only memory load affected item difficulty and time intensity, whereas number size had no effect. All points will be addressed in turn.

Table 3.7: Estimated standardized effects for model M_{P6}

Model	Effect	EAP	SD	.95 HPD
Ability (θ)	γ_{100} (Intercept)	0.00	-	-
	γ_{101} (Age)	0.26	0.07	[0.13, 0.39]
	γ_{102} (IQ)	0	-	-
	γ_{103} (Vocabulary test)	0.16	0.07	[0.03,0.29]
	γ_{104} (Number series)	0.50	0.07	[0.35,0.64]
	γ_{105} (ν)	0	-	-
	γ_{106} (a)	0.22	0.10	[0.02,0.42]
	γ_{107} (T_{er})	0.23	0.10	[0.03,0.44]
	γ_{108} (Computation Span)	0.25	0.07	[0.11,0.39]
	γ_{109} (Spatial WM)	0.24	0.07	[0.11,0.37]
	γ_{110} (Verbal Span)	0	-	-
Speed (ζ)	γ_{200} (Intercept)	0.00	-	-
	γ_{201} (Age)	0.10	0.02	[0.07,0.13]
	γ_{202} (IQ)	0.04	0.02	[0.01,0.07]
	γ_{203} (Vocabulary test)	0.04	0.02	[0.00,0.07]
	γ_{204} (Number series)	0.06	0.02	[0.03,0.10]
	γ_{205} (ν)	0.03	0.02	[0.00,0.07]
	γ_{206} (a)	-.06	0.03	[-.12,-.01]
	γ_{207} (T_{er})	-.20	0.03	[-.16,-.05]
	γ_{208} (Computation Span)	0.04	0.02	[0.01,0.08]
	γ_{209} (Spatial WM)	0	-	-
	γ_{210} (Verbal Span)	0	-	-

Note. Both intercepts ($\gamma_{100}, \gamma_{200}$) were fixed to 0 to identify the model.

Similar to other studies (e.g., Andersson, 2008), we found that complex span tasks, except for Verbal Span, substantially predicted algebra ability, even when numerous control variables like arithmetic ability, reading-related abilities, IQ, age, and facets of processing speed were taken into consideration. Similar to K. Lee et al. (2004), 4% of total variation could be uniquely attributed to WM capacity. Complex span tasks predicted algebra performance more strongly than reading-related abilities, in contrast to other results (K. Lee et al., 2004). A possible reason for this finding might be that we did not use word problems here. Although processing speed has been described as a basic function of WM (Case et al., 1982), and of cognitive processing in general (Fry & Hale, 1996), a key parameter of EZ-diffusion model reflecting speed of information processing, drift rate ν, did not substantially contribute to algebra performance, in contrast to results reported earlier by Schmiedek et al. (2007) who found large correlations

between ν and reasoning. One of the reasons for this finding might reside in the fact that the algebra test utilized here was given without a time-limit, thus possibly eliminating effects of test speededness. However, ν was related to the speed of solving algebra problems. Interestingly, therefore, a higher quality of information processing predicted solution speed, but not solution quality in algebraic reasoning. Further, studies based on younger samples (e.g., mean age 89 months; Bull & Johnston, 1997) often found a substantial effect of processing speed on arithmetic performance, in contrast to studies based on older samples of children (Berg, 2008), suggesting that speed of processing plays a larger role in samples of younger children (Hecht et al., 2001). We found, however, that IQ did not affect algebra ability in the full model. This surprising finding contrasts results reported in several studies (e.g., Andersson, 2008; K. Lee et al., 2004), and might be due to the fact that the IQ test utilized here here had a relatively strict time-limit, in contrast to the other tests used here. This assumption is underlined by the fact that IQ predicted algebra solution speed. However, Alloway (2009) reports findings from a longitudinal study indicating that WM, along with domain-specific knowledge, but not IQ, predict subsequent learning in children with learning disabilities. More research is needed to clarify this issue.

The fact that two WM tasks independently predicted algebra ability is difficult to reconcile with the assumption that a single domain-general WM system governs cognitive processes. Rather, it is in line with results from studies painting a more differentiated picture of verbal and visuo-spatial controlled attention (e.g., Berg, 2008; Hitch et al., 2001; Shah & Miyake, 1996), showing that WM facets are differentially related to cognitive functioning. Similar to Leather and Henry (1994), the results support the conclusion that the WM tasks used here capture both domain-free and domain-specific processes. However, the precise role of each WM task in algebra ability requires further research. Logie et al. (1994) report evidence that in arithmetic calculation, verbal WM serves to retain intermediate results, whereas visual-spatial WM is involved in encoding visually-presented parts of the problem. The current study offers no insight into such specialized roles. However, it was found that there was no interaction between memory load at the item level and WM capacity, suggesting that high-WM subjects did not use qualitatively different solution strategies in algebra items with high memory load than low-WM subjects (e.g., more retrieval in high-WM subjects; Barrouillet & Lépine, 2005), and that WM capacity was involved equally in high-load and low-load algebra items.

At the item level, we found that number size neither affected item difficulty nor time intensity. This finding differs from other results reported in the literature (e.g., Ashcraft, 1992; LeFevre et al., 1996), and might be related to the fact that the algebra test in this study, for practical reasons, comprised only 22 items, thus not covering number size effects only selectively, and that the task format was considerably more complex than in other studies. Future studies using both an experimental as well as an individual-differences approach would benefit from a more complete estimation of the effect of number size. However, a substantial effect of memory load on both item difficulty and time intensity was found, underlining the effects of Oberauer et al. (2001) that access to WM load during problem solving substantially affects algebra ability as well as solution speed. Like in the paper by Oberauer et al. (2001), these findings speak against a single-resource model of WM, supporting the view that cross-talk between competing memory elements when the processing task requires access to the contents of working memory is a key factor for intellectual functioning. A limitation of the current study, however, is that we did not use a control condition in which irrelevant memory load was used.

Overall, a key finding of this study is that WM plays a crucial role in algebra problem solving, even when taking prior arithmetic experience, IQ, and other parameters of interest into account. From an applied perspective, it would therefore be useful to take WM tasks into account when predicting algebra ability, and when doing high-stakes examinations. Although arithmetic ability certainly is a core predictor, and amenable to educational intervention, WM capacity can be substantially raised by training as well (Klingberg et al., 2005; Thorell, Lindqvist, Nutley, Bohlin, & Klingberg, 2009). Future interventions intended for low performers in mathematics might take this finding into account.

3.5 Appendix

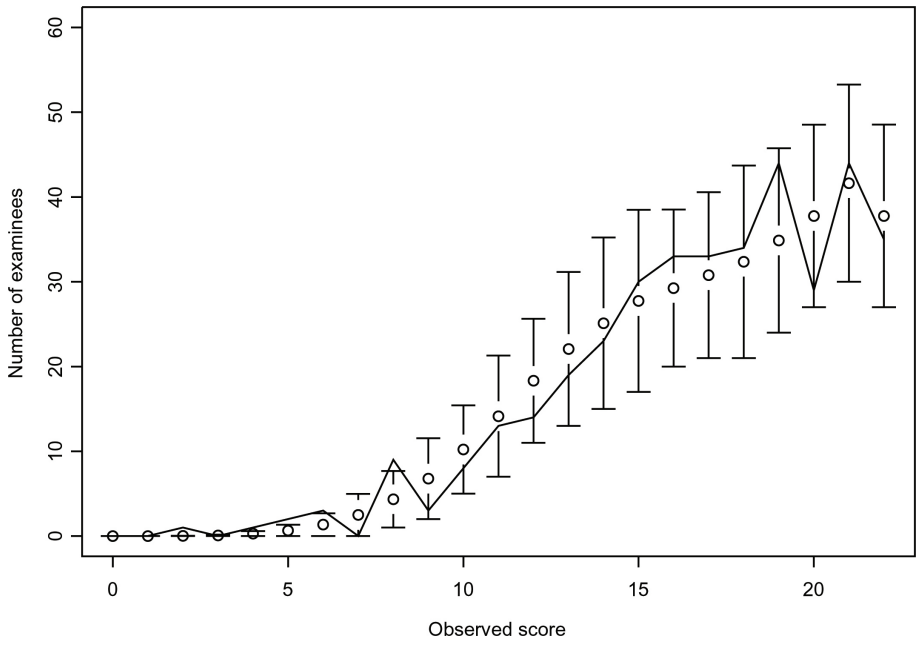


Figure 3.3: Observed sum scores (line) and model predicted sum scores (dots with .95 HPD regions)

4 Cognitive complexity and working memory in children: An investigation using the Latin Square Task

Summary. Relational complexity (RC) theory assumes that task complexity is defined by the number of distinct elements that must be simultaneously represented while solving a complex task. In this study, a measure of deductive reasoning, the Latin Square Task (LST), was systematically designed to assess core assumptions of RC. The LST task is conceptually simple in that it requires understanding of only a single rule. In line with RC theory, we found in a sample of children (N = 557, 8-13 years) that relational complexity, along with memory load, were core predictors of item difficulty, whereas the effect of chunk size was negligible. Further, using linear logistic test models with random effects, large interindividual differences in complex (quaternary) processing were found, suggesting age-related constraints in processing. Finally, the effect of quaternary processing and memory load were moderated by spatial working memory capacity. Results are discussed with respect to systematic investigations of reasoning ability.

4.1 Introduction

A plethora of research highlights the fact that reasoning is a core ability of fluid intelligence (cf. Carroll, 1993). Contemporary research on reasoning has focused on so-called dual-process theories. In these theories, an implicit, intuitive, heuristic system of reasoning is contrasted with a rule-based, analytical, explicit system (Sloman, 1996; Stanovich, 1999). In individual differences research, the interest lies mainly with the second system. Individual differences in reasoning ability have often been attributed to restrictions in the ability to create and manipulate mental representations, i.e., to working memory capacity (WM). As has been abundantly shown, WM resides at the heart of all higher cognitive functions (Ackerman, Beier, & Boyle, 2005; Kyllonen & Christal, 1990).

One important question in this context is, what makes a reasoning problem difficult? In other words, can the cognitive complexity of a reasoning problem be defined and quantified in advance, based on a strong theory? Several studies utilizing numerous indicators of reasoning, WM, and general intelligence have been published (e.g., Colom, Abad, Quiroga, Shih, & Flores-Mendoza, 2008; Krumm et al., 2009; Oberauer, Süß, Wilhelm, & Wittmann, 2008), but these offer only a

macroscopic perspective and do not focus on cognitive processes during solving single reasoning problems. In addition, they offer no insight into how cognitive complexity may be defined. In order to evaluate cognitive complexity, one would therefore need a carefully-designed reasoning test based on a strong theory of cognitive complexity. As mentioned by Krumm et al. (2009, p. 361), "the experimental manipulation of task requirements seems to be a promising next step".

Several theories of cognitive complexity, and of modeling thought processes in reasoning, have been suggested. Halford, Wilson, and Phillips (1998) proposed a theory of relational complexity (RC), which allows to define task complexity independent of domain, and which offers a metric to quantify cognitive complexity. In this study, our goal was to evaluate the cognitive complexity of a figural reasoning test, the Latin Square Task (Birney, Halford, & Andrews, 2006), as predicted by RC. Further, the role of WM in specific reasoning processes was assessed. Finally, using random effects item response theory (IRT) models (de Boeck, 2008), hypotheses with respect to individual differences in reasoning and cognitive complexity were evaluated.

4.1.1 Cognitive complexity, relational complexity, and RC theory

Cognitive complexity has been defined in different ways in the literature. For example, in Marshalek, Lohman, and Snow (1983), complexity was defined as the degree to which a task loaded on the g-factor (cf. Spilsbury, Stankov, & Roberts, 1990), whereas Vernon and Jensen (1984) defined complexity based on the response time required to solve a task. These earlier approaches are purely empirical in that the establish complexity in a post-hoc fashion without a prior cognitive theory.

Other approaches have described cognitive complexity with respect to the number of distinct elements or type of element relations in a task (Carpenter, Just, & Shell, 1990; Holzman, Pellegrino, & Glaser, 1983; Primi, 2001). In these approaches, reasoning ability is limited by WM. WM is often regarded as a capacity-limited system of information processing (e.g., Just & Carpenter, 1992). Based on such a view, more difficult items in a psychometric test require more WM capacity, because more elements have to be stored and manipulated simultaneously, and interindividual differences in reasoning therefore can be based on differences in WM capacity. However, this definition of complexity does not necessarily bear

on the depth of processing required to solve a task, because increasing storage demands of an otherwise simple task will result in a higher task difficulty, but not necessarily a higher complexity.

Although storage in the context of processing, as captured by classical WM tasks, appears to play an important role in reasoning ability, it is not necessarily the most important factor. One important facet of cognitive architecture is the ability of relational integration, i.e., the ability to build structural relations between elements and thus create structural subrelations (Waltz et al., 1999). These elements can be presented visually or held in memory. As could be shown by both Oberauer et al. (2008) and Krumm et al. (2009), relational integration, as captured by a battery of tasks, is a decisive factor in predicting reasoning ability, above the traditional storage and processing component of WM. Oberauer et al. (2008) hypothesize that reasoning tasks such as series completion require people to construct a representation of the relations between elements of the series, which has to be transferred to a later segment of the series in order to generate the next element. To construct new relational representations, elements must be bound to each other in a new representation. Hence, a limit on the number of bindings that can be upheld simultaneously posits a limit on the complexity of new relational representations that can be processed. That is, even WM tasks without a storage component can be powerful predictors of reasoning ability, although storage and processing as well as coordination factors are substantially correlated but dissociable (Oberauer et al., 2008).

Similarly, Halford et al. (1998) argued that it is not the amount of information per se, but the complexity between the pieces of information that have to be processed which is subject to capacity limitations. RC theory is based on two axioms. Axiom 1 pertains to the complexity of a cognitive process, which is represented by the number of interacting variables that must be processed in parallel to correctly carry out that process. Axiom 2 is concerned with the processing complexity of a task, which is "the number of interacting variables that must be represented in parallel to perform the most complex process involved in the task, using the least demanding strategy available to humans for that task" (Halford et al., 1998, p. 805). The complexity of a relation $R(a_1, a_2, \ldots, a_n)$ is generally determined by the number of arguments n involved. More arguments allow more complex relations. For example, a binary relation would be larger-than(elephant, mouse), in which two arguments are related to each other. In contrast, a unary relation is simpler, representing a categorization, such as dog(Fido).

Ternary relations have three arguments, such as addition(2, 3, 5). More elements allow more complex relations to be established, which according to Halford et al. (1998) requires more WM capacity to represent them. Halford, Cowan, and Andrews (2007) argue that WM and reasoning are also conceptually similar because in WM, elements are bound to a temporary coordinate system, which is closely related to relational representations in reasoning.

RC theory suggests that two cognitive strategies are available to reduce the cognitive demand of processing such relations. The first is conceptual chunking, which recodes a high-dimensional relation into a lower-dimensional one. As mentioned by Birney et al. (2006), velocity can either be processed as ratio(distance, time, velocity), which represents a ternary relation, or as a unary relation, velocity(60km/h). In the latter case, information concerning distance and time are disregarded. Conceptual chunking, therefore, reduces processing demand at the cost of loss of information. A second strategy to reduce processing demand is segmentation, which entails breaking down a complex task into several steps of lower complexity. That is, in the case of a reasoning test, for example, a stepwise solution process is triggered, and only relations of elements in the current solution step are analyzed. Other relations are inaccessible at the time. Complexity of the task is then determined as the complexity of the most complex step, according to RC theory.

Relational complexity has been shown to affect the difficulty of deductions (Birney et al., 2006; Holling, Bertling, Zeuch, & Kuhn, in press; N. Y. L. Lee, Goodwin, & Johnson-Laird, 2008) and to play a role in cognitive development (Andrews & Halford, 2002). However, relational complexity as defined by RC theory hinges upon identical strategies being used by persons solving the same reasoning task (Sweller, 1998). That is, cognitive tasks used to assess RC theory must be carefully designed to avoid ambiguity, despite the fact that differential strategy use does not affect the relationship between WM and higher cognitive functioning (Turley-Ames & Whitfield, 2003). Further, prior knowledge differences have to be taken into account. They can best be minimized by using reasoning tasks with figural content. Reasoning measures with figural content have additionally been shown to be the best measures of fluid intelligence (Undheim & Gustafsson, 1987).

The Latin Square Task (LST), an innovative deductive reasoning measure, was developed by Birney et al. (2006) in order to operationalize complexity levels as defined by RC theory. Although deductive and inductive reasoning are

sometimes seen as being differentiable (Colberg, Nester, & Trattner, 1985), recent research has shown that inductive and deductive reasoning cannot be separated at the latent level (Wilhelm, 2005). Birney et al. (2006) found that in this task, complexity level as defined by RC theory explained 64% of variance in item difficulty, showing that complexity as defined by RC theory is a powerful and theoretically sound predictor of item difficulty. The study of these authors, however, is preliminary. First, they did not include any other constructs of interest in their analysis (e.g., WM). Second, they did not use IRT models that allow detailed insights into variation of complexity level difficulty across individuals, or interactions of complexity levels with person covariates like WM. Third, some of the items in Birney et al. (2006) had to be reclassified concerning their complexity level. The results reported by these authors, therefore, should be considered preliminary and require further investigation.

4.1.2 Modeling cognitive complexity using IRT models

Modeling cognitive complexity requires using IRT models, because classical test theory is focused on test scores. One of the central IRT models for assessing the difficulty of item components or rules is the linear-logistic test model (LLTM) proposed by Fischer (1973). In order to estimate this model, however, Rasch-scalability must be given. The LLTM can be used to test hypotheses with respect to item component difficulty, which are grounded in prior theory.

In the LLTM, item difficulty is decomposed such that the probability of person i to answer item k correctly is provided by

$$P(Y_{ik} = 1|\theta_i, \gamma_j, q_{jk}) = \frac{exp(\theta_i - \sum_{j=0}^{J} q_{jk}\gamma_j)}{1 + exp(\theta_i - \sum_{j=0}^{J} q_{jk}\gamma_j),} \tag{4.1}$$

where $\theta_i =$ person ability with $\theta_i \sim N(0, \sigma_\theta^2)$, $\gamma_j =$ difficulty of item component j and $q_{jk} =$ dummy-variable indicating whether component j is present in item k or not. $j = 0$ indicates an intercept term for scaling the IRT model.

The LLTM is a restrictive model that practically always shows inferior fit to the Rasch model, due to the usually smaller number of item-related parameters ($J < K$). Items consisting of identical components are fixed to an identical item difficulty. To overcome this strong restriction, an item-related random

effect can be introduced, resulting in an LLTM with a random item effect (RE-LLTM; Janssen, Schepers, & Peres, 2004; van den Noortgate, de Boeck, & Meulders, 2003),

$$P(Y_{ik} = 1|\theta_i, \gamma_j, q_{jk}) = \frac{exp(\theta_i - \sum_{j=0}^{J} q_{jk}\gamma_j + \epsilon_k)}{1 + exp(\theta_i - \sum_{j=0}^{J} q_{jk}\gamma_j + \epsilon_k),} \tag{4.2}$$

with $\epsilon_k \sim N(\sum_{j=1}^{J} q_{kj}\gamma_j, \sigma_\epsilon^2)$. That is, items with an identical configuration have an expected value of $\sum_{j=1}^{J} q_{kj}\gamma_j$, but random variation is possible, and captured by the variance σ_ϵ^2. Therefore, identical items can be seen as clones stemming from the same item family (Glas & van der Linden, 2003). σ_ϵ^2 can be regarded as the residual variance when regressing item difficulties in the Rasch model on the predictors q_{jk}. The RE-LLTM therefore is helpful in assessing the explanatory power of the cognitive model under investigation. It can be compared against a version of the Rasch model that allows both random item and person effects (de Boeck, 2008).

An additional extension of the LLTM is helpful to assess whether item component difficulties vary across subjects. This might be an indication of differing solution strategies. Using the random-weights LLTM (RW-LLTM; Rijmen & de Boeck, 2002), it is possible to estimate variance components with respect to item components. It corresponds to a logistic regression model with random slopes. The RW-LLTM is defined as

$$P(Y_{ik} = 1|\theta_{il}, \gamma_j, b_{lk}, q_{jk}) = \frac{exp(\sum_{l=0}^{L} b_{lk}\theta_{il} - \sum_{j=0}^{J} q_{jk}\gamma_j)}{1 + exp(\sum_{l=0}^{L} b_{lk}\theta_{il} - \sum_{j=0}^{J} q_{jk}\gamma_j),} \tag{4.3}$$

with $\theta_{il} \sim N(0, \sigma_{\theta_l}^2)$, where $\sigma_{\theta_0}^2 = \sigma_\theta^2$ as in the standard LLTM. This model is a multidimensional extension of the LLTM because it assumes additional component-specific person abilities θ_{il}. By estimating their variance components, an assessment of the homogeneity of solution processes is principally possible.

A third extension of the LLTM takes person-level covariates into account in order to explain differences at the level of the individual. In this model, the person ability θ_i is written as $\sum_{m=1}^{M} w_{im}\vartheta_m + \epsilon_i$, where w_{im} is the value of person i on some property m (e.g., a WM task score), ϑ_m is the fixed regression weight of person property m, and ϵ_i is the remaining person effect after controlling for differences due to properties $m(m = 1, \ldots, M)$, with $\epsilon_i \sim N(0, \sigma_\epsilon^2)$. The latent

regression LLTM (LR-LLTM; Zwinderman, 1991) has the form

$$P(Y_{ik}=1|\vartheta_m,\gamma_j,\epsilon_i,w_{im},q_{jk})=\frac{exp(\sum_{m=1}^{M}w_{im}\vartheta_m+\epsilon_i-\sum_{j=0}^{J}q_{jk}\gamma_j)}{1+exp(\sum_{m=1}^{M}w_{im}\vartheta_m+\epsilon_i-\sum_{j=0}^{J}q_{jk}\gamma_j).} \quad (4.4)$$

Finally, it is often of interest to investigate person-by-item-interactions more closely, which corresponds to an analysis of differential item functioning (DIF; Meulders & Xie, 2004). In the context of the LLTM, it is of interest whether item component difficulties interact with person properties, which is called differential facet functioning (DFF; Engelhard, 1992). In DFF, a more explanatory investigation of component difficulties due to specific person properties (e.g., WM capacity) is feasible. Often, DFF is assessed assuming a fixed effect assumption, although random DFF conceptions have been suggested (Meulders & Xie, 2004). A DFF model can be conceptualized as an extension of the LR-LLTM (excluding main effects of person properties) as

$$P(Y_{ik}=1|\theta_i,\gamma_j,\delta_j,w_{im},q_{jk})=\frac{exp(\theta_i-\sum_{j=0}^{J}q_{jk}\gamma_j+\sum_{j=1}^{J}\delta_j q_{jk}w_{im})}{1+exp(\theta_i-\sum_{j=0}^{J}q_{jk}\gamma_j+\sum_{j=1}^{J}\delta_j q_{jk}w_{im}),} \quad (4.5)$$

where all δ_j capture interactions of item component difficulties with person properties, respectively. This model, then, allows to assess whether difficulties of item components vary with person properties, indicating possible qualitative differences in cognitive processing.

The models outlined above can be combined in various ways. For example, RW-LLTM, LR-LLTM, and DFF can be combined to yield something analogous to a model with intercepts and slopes as outcomes in the classical multilevel literature (Snijders & Bosker, 1999). Each model allows a close examination of prior hypotheses within an IRT framework.

4.1.3 Purpose of the present study

The purpose the present study was to corroborate and advance the results obtained by Birney et al. (2006). Our goal was to design a LST with unambiguous items for children. We wanted to assess the effects of relational complexity

and number of steps on item difficulty, using the IRT models described above. Hence, we were able to investigate whether higher relational complexity raised item difficulty, whether interindividual variation with respect to complexity level difficulty could be observed, and whether theoretically relevant constructs such as WM affected the ability to provide a solution or whether they moderated the effects of relational complexity.

4.2 Method

4.2.1 Subjects

Five-hundred fifty-seven children participated in this study, of whom 130 children visited primary school, whereas the remaining 447 children went to secondary schools in various regions of Germany. Mean age was 10;8 years (SD = 1.05, range: 8;0-13;4). 49% of the participants were female. Parental consent was obtained for all participants prior to testing. Few participants (n = 21) indicated German was not their first language, although all of these participants spoke German since they were 3 years old.

4.2.2 Measures

The Latin Square Task

We designed a LST based on Birney et al. (2006). A Latin Square is based on the ancient puzzle in which each element in a square occurs in each row and column only once, as shown in Figure 4.1 in the lower right panel. The principle of Latin Squares can be used to systematically design test items of differing relational complexity. In the LST, subjects are required to correctly deduce the content of a pre-specified cell by utilizing this simple principle.

In the upper left panel of Figure 4.1, for example, subjects must perform *binary processing* in order to solve the item correctly. One of the cells in the LST contains a question mark. The content of this cell must be deduced using the single rule that each element below the LST (except the question mark) must occur once in each row and column. In this case, the second row contains three different elements. These can be conceptually chunked into one set, because a differentiation of these elements from each other is not necessary. However, it is clear that

the element in the cell with the question mark must differ from the chunked set. In order to solve this item correctly, therefore, two sets of elements must be represented, the complete set of elements, given below the LST, as well as the given set of elements in the second row. The difference between these two sets corresponds to the solution of the item. Binary processing has a relational complexity of 2.

Ternary processing (upper right panel in Figure 4.1) requires integration of information from both a row and a column. In the example given, the solution can be deduced by taking both the second column as well as the fourth row into account. The intersection of these two must not contain any element present in either the respective row or column. Because elements in the second column are not independent of any elements in the intersecting rows, they cannot be represented in a single chunk together (Birney et al., 2006). However, elements in the lowest row can be chunked, because their relation does not have to be considered for solving the item. Therefore, in this example of ternary processing, three distinct sets of elements must be cognitively represented: The full set of elements, the two elements in the lowest row, and the element in the second column. Ternary processing has a relational complexity of 3.

Quaternary processing requires integration of elements across multiple rows or columns. In the example given in Figure 4.1 (lower left panel), binary or ternary solution strategies do not produce a unique solution. Here, the distribution of the circle of rows and columns must be taken into consideration. As can be seen, the circle occurs in the second and third row and column, respectively. It is impossible that a circle could be put into the lower right cell of the LST, because this cell is already occupied. Therefore, the only possibility to put a circle into the rightmost column is the cell with the question mark. That is, a subject must take into consideration all possible elements in the rightmost column while taking into consideration elements in all rows. This entails representing four pieces of information. Here, relations between the elements must be considered, and therefore, conceptual chunking is not possible. Quaternary processing has a relational complexity of 4.

The complexity of solution steps in LST items can therefore be systematically manipulated by specifying the relations of elements within the LST. In addition, the number of steps to arrive at a solution can be manipulated. That is, items can be constructed in which the content of empty nontarget cells must be resolved in an itermediate step and stored in memory before the target cell is approachable. By introducing intermediate steps, serial processing is invoked, and

memory load is introduced.

We developed 16 LST items of differing relational complexity. Complexity of each item was determined, according to RC theory, by the most complex step in each item. Four items were binary, eight items were ternary and four items were quaternary. Pretests had shown that it is difficult for younger children to solve complex LST items with a high memory load. Therefore, nine items had no memory load (one step), five items had a memory load of one intermediate result (two steps) and two items had a memory load of two (three steps).

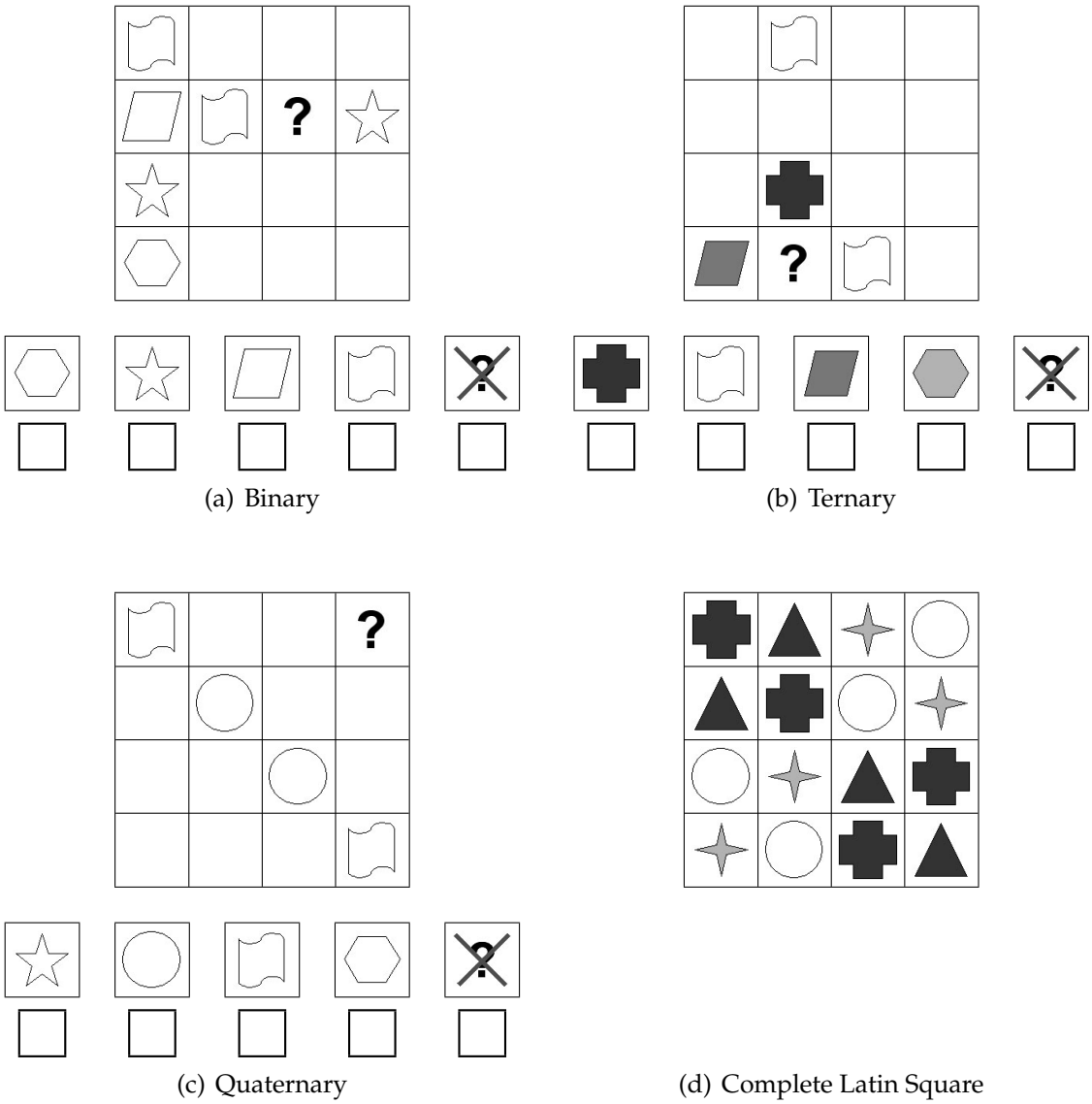
We further introduced an additional manipulation to check whether chunk size affected item difficulty. In order to do so, Latin Squares of two size formats, 4×4 and 5×5, were introduced. Based on RC theory, the size of chunks should not play any role, in contrast to the complexity of element or set relations. Eight items of each size were distributed across the test.

The LST task designed here was administered by computer. After a detailed instruction explaining the Latin Square principle, four practice items with feedback had to be solved. The test was given without a time limit such that subjects could work at their own pace.

WM tasks

WM tasks usually consist of several items and subitems. For example, in this study, computation span comprised 10 items, each consisting of three to seven equations displayed on the screen (subitems). The subitems (in the case of computation span, the results of the equations) were the contents that had to be remembered. Recently, it has been shown that partial scoring (i.e., computing the sum of proportion of correctly-solved subitems for each item) results in better psychometric properties and higher correlations with measures of fluid intelligence than other scoring procedures, presumably because more information is retained (Unsworth & Engle, 2007b). We therefore used partial scoring for all WM tasks.

Verbal Span (VS) This WM task, based on Oberauer, Süß, Schulze, Wilhelm, and Wittmann (2000) and Vock and Holling (2008), consisted of two different parts. Participants first had to memorize a list of words presented on the screen (presentation time 6 s). List length in this storage task varied between three to six words. Then, between two and three verbal decision tasks followed in which



(a) Binary (b) Ternary

(c) Quaternary (d) Complete Latin Square

Figure 4.1: Examples of different Latin Square Task items and a complete Latin Square

participants had to respond as quickly as possible. In these processing tasks, participants had to decide which of four words displayed in each corner of the screen stood in a subconcept relation to the word shown in the center of the screen (e.g., ”animal” - ”lion”). Finally, participants were supposed to reproduce the learned words in correct order. The task consisted of two practice items and 10 test items.

Spatial Working Memory (SWM) Participants had to memorize one or several simple chessboard-like 3 $\times$ 3-patterns (storage task). However, the patterns

had to be stored in a rotated fashion, rotated either 90° clockwise or counterclockwise (processing task). That is, before the patterns were shown successively for 4 s each, an arrow indicated whether patterns had to be mentally rotated to the left or to the right. Finally, participants had to successively reproduce the memorized patterns into empty 3×3 matrices on the screen. The task consisted of 13 items with between one to four patterns. Three practice items preceded the testing phase.

Computation Span (CS) In this task, participants were sequentially shown a series of simple, single-digit equations that included either an addition or a subtraction (e.g., 4 + 3 = 8). Each equation was shown for 5 s. Approximately half of the equations were correct and half were incorrect. The processing task consisted in deciding whether the equation shown on screen was correct or incorrect. Further, all shown equation results had to be memorized irrespective of whether they were correct or not. Finally, subjects were presented with an answer screen and successively clicked the to-be-remembered equation results. Each item consisted of between three to seven items, resulting in 10 test items. Two practice items were administered before the testing phase.

Fluid intelligence

In order to determine fluid intelligence (IQ), the short form of the Grundintelligenztest Skala 2 (CFT 20; Weiß, 1998), a German adaptation of the Culture Fair Intelligence Test, Scale 2 (Cattell, 1973), was utilized. The CFT 20 is a paper-and-pencil test which provides high loadings on fluid intelligence (Cattell, 1968) and has good psychometric properties. It consists of four different subtests: Series completion, Classifications, Matrices and Topologies. Between two and three practice items were given before each subtest commenced. Overall testing time, including instructions for each subtest, was approximately 23 minutes. IQ was seen as a control variable here, which should be taken into account whenever higher cognitive processing is analyzed (Mayes, Calhoun, Bixler, & Zimmerman, 2009).

4.3 Results

There are three sections to the results. First, descriptive statistics and reliabilities of study measures are reported. Second, Rasch analysis results concerning the

LST are reported. Third, we used the IRT models described above to assess effects of relational complexity and memory load as well as WM, age, and IQ on LST performance.

As shown in Table 4.1, reliabilities were mostly satisfactory, although the internal consistency of the LST was not as high. However, ANOVA reliability (Kerlinger, 1973) of the LST was .76, suggesting sufficient measurement precision for subsequent IRT analyses.

Table 4.1: Descriptive statistics and reliability estimates

Variable	M	SD	Skewness	Kurtosis	K-S[a]	α
Reasoning						
1. Latin Square Task	9.54	2.98	-.11	-.21	.08**	.71
Working memory						
2. Verbal Span	5.91	2.14	-.63	-.10	.07*	.86
3. Spatial WM	5.77	2.74	-.24	-.62	.04*	.82
4. Computation Span	7.11	2.07	-1.30	1.39	.14**	.86
Fluid intelligence						
5. IQ CFT 20	112	14.26	-.20	.23	.05*	.91

Note. [a] $Z(p)$ of Kolmogorov-Smirnov-test on normal distribution with correction of significance by Lilliefors.
*$p < .05$. **$p < .01$.

We analyzed Rasch scalability of the LST test by investigating classical goodness of fit tests as suggested by Andersen (1973) and Martin-Löf (1973). We started by computing the Andersen test, which basically analyzes whether item difficulties in the Rasch model are comparable across subgroups. Three splitting criteria were used to build subgroups: Gender, age (younger than 10;8 years vs. older) and mean score (below vs. above). The Andersen test suggested that person homogeneity was given with respect to gender ($\chi^2(15) = 21.34$, $p = .12$) but not with respect to age ($\chi^2(15) = 26.06$, $p = .04$) or mean score ($\chi^2(15) = 42.05$, $p = .00$). Two criteria were used to form item subgroups for the Martin-Löf test, odd item numbers vs. even item numbers and ternary items vs. binary and quaternary items. Neither the odd-even comparison ($\chi^2(63) = 53.05$, $p = .80$) nor the comparison of subgroups based on item structure ($\chi^2(63) = 74.49$, $p = .15$) rejected item homogeneity.

We then computed the Q-index introduced by Rost and von Davier (1994). This statistic, which is based on the log-likelihood of the observed data pattern,

is an item-fit measure, which allows to assess overfit or underfit of single items to the Rasch model. As shown in Table 4.2, except for one item, which showed underfit, all items showed an acceptable fit to the Rasch model. This item was excluded from further analyses.

Table 4.2: Design, Rasch difficulties, and fit statistics of the LST items

Item	RC[a]	ML[b]	Size	β_i^c	SE	Q^d	p
1	2	0	4×4	-3.02	.23	.15	.75
2	2	0	5×5	-2.12	.17	.24	.36
3	3	0	4×4	-1.22	.13	.24	.22
4	3	0	5×5	-2.26	.17	.13	.86
5	2	1	4×4	-.20	.10	.16	.73
6	3	1	5×5	.23	.10	.19	.43
7	2	1	5×5	-.04	.10	.18	.57
8	3	1	4×4	.66	.10	.21	.24
9	3	0	5×5	-1.26	.13	.19	.61
10	3	2	4×4	1.08	.10	.27	.00
11	3	2	5×5	.93	.10	.21	.18
12	3	1	4×4	1.56	.10	.16	.70
13	4	0	5×5	1.30	.10	.17	.60
14	4	0	5×5	2.47	.13	.17	.62
15	4	0	4×4	.67	.10	.16	.68
16	4	0	4×4	1.22	.10	.16	.77

Note. [a]Level of relational complexity (2 = binary, 3 = ternary, 4 = quaternary). [b]Memory load. [c]Rasch item difficulty. [d]$Q-$statistic.

Prior to model estimation, we used the marginal modelling approach described in Balázs, Hidegkuti, and de Boeck (2006), which allows to detect covariate-specific heterogeneity in the data. This is important to determine whether local dependencies exist, and it is helpful concerning the parsimonious specification of random slopes in the RW-LLTM. We utilized an alternating logistic regression algorithm (Carey, Zeger, & Diggle, 1993) to assess heterogeneity. We found that a significant degree of heterogeneity was related to quaternary processing ($\alpha_2 =$ .20, $p < .01$), but not to the other item covariates. Hence, only one random slope (quaternary) was allowed in the RW-LLTM.

In the following, all IRT models described above to assess cognitive complexity were computed. As a general overview, it can be gleaned from Table 4.3 that the Rasch model, along with IRT models allowing an item-specific random

Table 4.3: Model fit statistics and model comparisons

Nr.	Model	$\ell\ell^a$	*df*	AIC	BIC	Compare with	$\Delta\chi^{2b}$	Δdf	p
1	Rasch	-3703.78	17	7442	7559				
2	RE-Rasch[c]	-3755.54	3	7517	7538				
3	LLTM	-3938.20	6	7888	7929	Rasch	468.84	11	.00
4	RE-LLTM	-3742.23	7	7498	7547	RE-Rasch	26.63	4	.00
5	RE-LLTM					LLTM	391.95	$0/1^d$	.00
6	RW-LLTM	-3926.73	7	7867	7916	LLTM	22.94	$0/1^d$	.00
7	LR-LLTM	-3879.82	11	7782	7857	LLTM	116.77	5	.00
8	DFF	-3912.74	9	7843	7905	LLTM	50.93	3	.00

Note. [a]Log-likelihood of model. [b]Likelihood-ratio test statistic. [c]Rasch model with random item effects, based on de Boeck (2008). [d]In case of boundary conditions, a mixture of two χ^2-distributions (with differing *df*) had to be used (Stoel, Garre, Dolan, & van den Wittenboer, 2006).

effect, showed the best fit. As could have been expected, the classical LLTM exhibited the worst model fit due to its severe restrictions. The largest improvement in fit concerning LLTM variants without an item random component was achieved by allowing person covariates (LR-LLTM). Comparing the Rasch model with random item effects with the RE-LLTM, we used the reduction in item variance due to item predictors to determine model quality. The four item-level predictors (ternary, quaternary, memory load, size) reduced item variance from 2.25 in the Rasch model with item random effects to 0.41 in the RE-LLTM. Hence, item variance was reduced by 82% by taking these four predictors into account, or differently stated, 82% of item variance was explained. Relational complexity alone, in a LLTM including only dummy-coded ternary and quaternary as predictors (binary served as a baseline category), 42% of item variance were explained, whereas memory load alone reduced item variance by 14%.

Table 4.4 summarizes parameter estimates of all IRT models. In all LLTM variants, with the exception of the RE-LLTM due to its larger standard errors (de Boeck, 2008), all predictors except for size were statistically significant. Apparently, in line with RC theory, chunk size is not decisive for cognitive complexity. However, as expected, the ability to process complex relations, especially quaternary processing, plays a substantial role in reasoning. Further, memory load affected item difficulty substantially as well, indicating that segmentation, along with the ability to keep intermediate results in mind, is strongly related to deductive reasoning.

Table 4.4: Parameter estimates of IRT models

	LLTM		RE-LLTM		RW-LLTM		LR-LLTM		DFF	
Parameter	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE
Item predictors										
Ternary[a]	-.60**	.08	-.67	.42	-.60**	.08	-.60**	.08	-.60**	.08
Quaternary[a]	-3.07**	.10	-3.50**	.49	-3.14**	.11	-3.07**	.10	-3.04**	.09
ML[b]	-1.34**	.05	-1.56**	.27	-1.33**	.05	-1.34**	.05	-1.33**	.05
Size	.06	.06	.00	.33	.07	.06	.06	.06	.06	.06
Person predictors										
Age							.01*	.00		
VS							.04	.03		
SWM							.10**	.02		
CS							.05	.03		
IQ							.01**	.00		
Ternary × SWM									.03	.03
Quaternary × SWM									.06*	.03
ML × SWM									.06**	.01
Variance components										
σ^2_θ	.85**	.09	1.00**	.11	.84**	.09	.57**	.07	.66**	.08
σ^2_ϵ			.41**	.15						
$\sigma^{2c}_{\theta_1}$					.78**	.22				

Note.[a]The intercept was fixed to 0. Dummy-coding was used for ternary and quaternary, Hence, binary items without memory load served as a reference criterion. [b]Memory load, with values of 0 (no memory load) or 1 (memory load of 1 or 2). [c]Variance component pertaining to random slope of quaternary predictor.
*$p < .05$. **$p < .01$.

We found that there were large interindividual differences in the difficulty to process quaternary relations, as exhibited by the large random slope variance in the RW-LLTM, which was nearly as large as the variance of the overall person ability parameter. This result is difficult to reconcile with the assumption that individuals used homogeneous strategies with respect to quaternary processing, which should have resulted in a nonsignificant random slope variance. Further, in the LR-LLTM, age and IQ were related to reasoning ability, although of all WM tasks only the spatial WM task was related to solution accuracy.

In order to compute a DFF model, we first centered the SWM task and then computed the product terms with ternary, quaternary, and memory load, respectively. Nonsignificant predictors (size, VS, CS) were omitted. We found that higher spatial WM was especially beneficial in items with higher relational complexity (i.e., quaternary) and in items requiring the storage of intermediate results. That is, a higher WM capacity, as evidenced by the spatial WM task, was especially valuable in solving more complex or memory-demanding tasks.

4.4 Discussion

This study sought to investigate the impact of facets of relational complexity on deductive reasoning, and to provide insight into the effect of WM, IQ, and age on reasoning ability. Using different explanatory IRT models, several new insights could be gained. We could replicate findings by Birney et al. (2006), showing that relational complexity level was a core predictor of item difficulty. Application of the RE-LLTM showed that a large portion of item variance (82 %) could be explained by the item covariates, especially by relational complexity level. These results are in line with RC theory, as well as with the theory of relational integration by Oberauer et al. (2008). Especially quaternary items were highly difficult, requiring the integration of four different information pieces into a temporary representation. It has been shown that the upper limit of information processing in humans is met when four distinct elements must be simultaneously represented (Cowan, 2001). Therefore, the high difficulty of quaternary items relative to ternary items is not surprising. Chunk size, on the other hand, did not play a role, as shown by the fact that item size was an insignificant predictor of item difficulty. RC theory predicts that the number of relations to be represented simultaneously, not the size of elements or element sets, is of key importance. Both results were supported by our analysis.

Memory load affected item difficulty substantially. Because in items with memory load, intermediate results had to be retrieved in the next step, a strong effect of memory load could have been expected. Oberauer, Demmrich, Mayr, and Kliegl (2001) showed that whereas irrelevant memory load only had a small effect in an equation solving task, a strong effect of memory load could be observed if intermediate results had to be accessed by WM during subsequent solution steps, that is, when intermediate results were no "passive load". In the LST task utilized here, intermediate results necessarily had to be stored and used in the next step. Hence, the effect of memory load was substantial, although not as large as the effect of quaternary processing.

The RW-LLTM revealed that there were large interindividual differences with respect to the difficulty of quaternary processing, as indicated by a substantial slope variance. Normative data have shown that processing quaternary relations occurs at median age 11 (Andrews, Halford, Bunch, Bowden, & Jones, 2003), which is close to the mean age in the sample investigated here. One of the reasons for the substantial random slope variation of quaternary processing, therefore, could reside in the fact that younger children were not able to process quaternary relations, whereas older children could do so. In line with this assumption, age, along with IQ, were related to deductive reasoning ability in the LR-LLTM, along with spatial WM. No other WM task was relevant for solving the LST items. This is not surprising, as reasoning tasks with figural content often show high correlations with spatial WM factors (Kane et al., 2004). Binding LST elements to temporal relational representations, further, at least implicitly requires some spatial representation of the item structure. Interestingly, spatial WM capacity was especially beneficial for quaternary items or items with memory load, as evidenced by a DFF model investigating interaction effects. That is, in addition to an overall advantage in solving deductive reasoning tasks, having a high spatial WM capacity is especially helpful in building complex relational representations and storing intermediate results. The effect was small, however. No interaction effect of spatial WM with ternary processing was found, indicating that less complex relations do not require proportional more WM capacity to be processed.

To summarize, we could replicate and extend findings reported by Birney et al. (2006) concerning a reasoning measure whose items were based on RC theory. Relational complexity level was a substantial predictor of item difficulty. Huge interindividual differences in quaternary processing were found, possibly due to

developmental constraints. Finally, along with age and IQ, deductive reasoning was related to spatial WM, especially to more complex and memory-demanding processes. LST items can be designed in a very systematic and rule-based way, combining a strong prior theory based on cognitive psychology with growing empirical support. Therefore, this test format offers good possibilities for a more systematic investigation of reasoning and its relationship with WM and the level of single cognitive processes.

4.4.1 Limitations

This study had several limitations that shall be outlined below. Firstly, we did not include a WM measure of relational integration or coordination (Oberauer et al., 2008). This would have been helpful in disentangling effects relational complexity and memory load. It could be hypothesized that relational integration is more strongly related to relational complexity, whereas storage-and-processing tasks of WM capture memory load variance.

Secondly, although the task format allows for a systematic item design, it introduces several constraints. For example, it is not feasible to design an item with quaternary processing that contains additional processing steps, because in this case, the item is rendered ambiguous, i.e., it can be solved with a simpler strategy. The restrictions imposed by the task format can lead to some dependencies of item predictors, which cannot vary freely and independently of each other. These constraints must be taken into consideration when testing specific hypotheses is of interest.

Thirdly, we investigated a sample of children here, thus rendering comparability to prior studies using the Latin Square Task difficult. Children differ from adults in their processing capacities (e.g., Andrews & Halford, 2002), and the results presented here might therefore be qualitatively different from results that would have been obtained with an adult sample.

5 Epilogue

As the overwhelming portion of research literature shows, WM and reasoning are closely related. This finding has found support on both empirical and theoretical grounds and is not new (e.g., Ackerman, Beier, & Boyle, 2005; Blair, 2006; Kyllonen & Christal, 1990). However, only few studies have investigated the role of WM in reasoning processes in samples of children. The goal of this thesis, therefore, was to shed light on reasoning processes in children, using both a "macroscopic" structural equation modeling perspective at the level of construct covariation and a "microscopic" IRT perspective at the level of item solution processes. A special focus is placed on the role of WM in reasoning.

As mentioned in the Introduction, several theories of WM exist that provide contradictory hypotheses with respect to the structure and breadth of WM. One of the goals of the study in Chapter 2, therefore, was to investigate whether a cognitive task measuring the scope of attention, i.e., a task that does not have a processing component, operates as a WM or STM task. We found that the scope of attention os related to WM and not STM, in line with results by Cowan et al. (2005). Further, the structure of WM and STM is stable across age, as a partial strong measurement invariance model could not be rejected. hence, no qualitative differences in the cognitive structure of younger and older children could be detected. Further, a domain-general model of WM showed the best fit to the data. This finding gives strong support to the assumption of domain generality, because structurally heterogeneous WM tasks were used. One important result was that the role of STM in predicting Gf declined with age, with STM becoming an insignificant predictor from approximately 11 years on. No interaction between age and WM was found, indicating that WM is a consistently important predictor for intellectual performance.

In Chapter 3, algebraic reasoning of children was investigated using a bivariate IRT model to assess algebra ability and solution speed simultaneously. We found that number size did not affect algebra problem difficulty, whereas memory load had a strong effect. Hence, the ability to keep intermediate results accessible in WM during algebra problem solving is of key importance. As expected,

WM remained a substantial predictor of algebra performance, even when taking IQ, age, and domain-specific knowledge (number series) into account. Speed of processing, as embodied by the drift rate in the diffusion model, was unrelated to algebra performance, but correlated with algebra processing speed, possibly due to the fact that the algebra test used was a pure power test. Further, no interaction between WM and memory load was found, providing some indirect support for the assumption that high-WM students did not utilize different solution strategies than low-WM students. Because two WM tasks differentially predicted algebra performance, the domain-generality of WM was disputed here. This finding stands in contrast to those of Chapter 2. However, different analysis procedures were utilized, structural equation modeling and IRT modeling. The differing results might be a method-related artifact. Further research using samples of children are required.

Results of a systematic analysis of cognitive complexity are discussed in Chapter 4. A figural reasoning task was designed based on relational complexity theory (Halford, Wilson, & Phillips, 1998). Relational complexity has been suggested as a universal metric of task complexity. In line with prior results, using increasingly complex IRT models, relational complexity was found to be a core predictor of item difficulty in reasoning, along with memory load. We further found large differences in processing complex (quaternary) items. This result was interpreted as an age-related processing constraint. Finally, spatial WM interacted with quaternary processing and memory load, suggesting a specific advantage of high-WM students in building complex mental representations.

Future research could address some limitations of the studies presented here. First of all, it has recently been described that relational integration is a core facet of WM (Oberauer, Süß, Wilhelm, & Wittmann, 2008). Tasks capturing relational integration have, to our knowledge, not been systematically applied to children. Relational integration is of key importance for reasoning ability. It would be especially fruitful to combine complex span tasks with relational integration tasks to analyze reasoning tasks based on relational complexity theory. Hence, it becomes possible to disentangle the effects of controlled attention, as captured by complex span tasks, from effects of relational integration. Further, a pressing issue is the investigation of differential utilization of processing strategies in reasoning tasks. For example, it would be possible to utilize a mixture LLTM when the different solution strategies are known in advance (Mislevy & Verhelst, 1990), or some other form of constrained latent class modeling like cog-

nitive diagnostic modeling (Junker & Sijtsma, 2001). Finally, a systematic IRT-based scaling of WM tasks offers great potential for the future as well, as most WM scoring schemes do not take item dependencies into account. Local item dependencies could be analyzed by developing testlet models for WM tasks (Wang & Wilson, 2005), or by supplementing LLTM variants with additional random effects (Ip, Smits, & de Boeck, 2009). Cognitive psychology would generally benefit from a closer integration with psychometric modeling.

References

Ackerman, P. L., Beier, M. E., & Boyle, M. O. (2002). Individual differences in working memory within a nomological network of cognitive and perceptual speed abilities. *Journal of Experimental Psychology: General*, *131*, 567–589.

Ackerman, P. L., Beier, M. E., & Boyle, M. O. (2005). Working memory and intelligence: The same or different constructs? *Psychological Bulletin*, *131*, 30–60.

Adams, J. W., & Hitch, G. J. (1997). Working memory and children's mental addition. *Journal of Experimental Child Psychology*, *67*, 21–38.

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Thousand Oaks, CA: Sage.

Algina, J., Keselman, H. J., & Penfield, R. D. (2007). Confidence intervals for an effect size measure in multiple linear regression. *Educational and Psychological Measurement*, *67*, 207–218.

Alloway, T. P. (2009). Working memory, but not IQ, predicts subsequent learning in children with learning difficulties. *European Journal of Psychological Assessment*, *25*, 92–98.

Alloway, T. P., Gathercole, S. E., & Pickering, S. J. (2006). Verbal and visuospatial short-term and working memory in children: Are they separable? *Child Development*, *77*, 1698–1716.

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika*, *38*, 123–140.

Anderson, J. R., Reder, L. M., & Lebiere, C. (1996). Working memory: Activation limitations on retrieval. *Cognitive Psychology*, *30*, 221–256.

Andersson, U. (2008). Working memory as a predictor of written arithmetical skills in children: The importance of central executive functions. *British Journal of Educational Psychology*, *78*, 181–203.

Andrews, G., & Halford, G. S. (2002). A cognitive complexity metric applied to cognitive development. *Cognitive Psychology*, *45*, 153–219.

Andrews, G., Halford, G. S., Bunch, K. M., Bowden, D., & Jones, T. (2003). Theory of mind and relational complexity. *Child Development*, *74*, 1476–1499.

Ashcraft, M. H. (1992). Cognitive arithmetic: A review of data and theory. *Cognition*, *44*, 75–106.

Baddeley, A. D. (1986). *Working memory*. Oxford: Clarendon Press.

Baddeley, A. D. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences*, *4*, 417–423.

Baddeley, A. D., & Hitch, G. J. (1974). Working memory. In G. A. Bower (Ed.), *Recent advances in learning and motivation* (pp. 47–90). New York: Academic Press.

Balázs, K., Hidegkuti, I., & de Boeck, P. (2006). Detecting heterogeneity in logistic regression models. *Applied Psychological Measurement*, *30*, 322–344.

Barrouillet, P., Bernardin, S., Portrat, S., Vergauwe, E., & Camos, V. (2007). Time and cognitive load in working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*, 570–585.

Barrouillet, P., & Lépine, R. (2005). Working memory and children's use of retrieval to solve addition problems. *Journal of Experimental Child Psychology*, *91*, 183–204.

Bayliss, D. M., Jarrold, C., Baddeley, A. D., Gunn, D. M., & Leigh, E. (2005). Mapping the developmental constraints on working memory span performance. *Developmental Psychology*, *41*, 579–597.

Bayliss, D. M., Jarrold, C., Gunn, D. M., & Baddeley, A. D. (2003). The complexities of complex span: Explaining individual differences in working memory in children and adults. *Journal of Experimental Psychology: General*, *132*, 71–92.

Beauducel, A., & Wittmann, W. W. (2005). Simulation study on fit indexes in CFA based on data with slightly distorted simple structure. *Structural Equation Modeling*, *12*, 41–75.

Beckmann, B., Holling, H., & Kuhn, J.-T. (2007). Reliability of verbal–numerical working memory tasks. *Personality and Individual Differences*, *43*, 703–714.

Berg, D. H. (2008). Working memory and arithmetic calculation in children: The contributory roles of processing speed, short-term memory, and reading. *Journal of Experimental Child Psychology*, *99*, 288–308.

Birney, D. P., Halford, G. S., & Andrews, G. (2006). Measuring the influence of complexity on relational reasoning: The development of the Latin Square Task. *Educational and Psychological Measurement*, *66*, 146–171.

Blair, C. (2006). How similar are fluid cognition and general intelligence? A developmental neuroscience perspective on fluid cognition as an aspect of human cognitive ability. *Behavioral and Brain Sciences*, *29*, 109–160.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*, 1061–1071.

Bouwmeester, S., Vermunt, J. K., & Sijtsma, K. (2007). Development and individual differences in transitive reasoning: A fuzzy trace theory approach. *Developmental Review*, *27*, 41–74.

Buehner, M., Krumm, S., & Pick, M. (2005). Reasoning = working memory $\neq$ attention. *Intelligence*, *33*, 251–272.

Bull, R., & Johnston, R. S. (1997). Children's arithmetical difficulties: Contributions from processing speed, item identification, and short-term memory. *Journal of Experimental Child Psychology*, *65*, 1–24.

Bull, R., Johnston, R. S., & Roy, J. A. (1999). Exploring the roles of the visual-spatial sketch pad and central executive in children's arithmetical skills: Views from cognition and developmental neuropsychology. *Developmental Neuropsychology*, *15*, 421–442.

Bull, R., & Scerif, G. (2001). Executive functioning as a predictor of children's mathematics ability: Inhibition, switching, and working memory. *Developmental Neuropsychology*, *19*, 273–293.

Campbell, J. I., & Graham, D. J. (1985). Mental multiplication skill: Structure, process, and acquisition. *Canadian Journal of Psychology/Revue canadienne de psychologie*, *39*, 338–366.

Campbell, J. I., & Tarling, D. P. M. (1996). Retrieval processes in arithmetic production and verification. *Memory & Cognition*, *24*, 156–172.

Carey, V. J., Zeger, S. L., & Diggle, P. J. (1993). Modelling multivariate binary data with alternating logistic regressions. *Biometrika*, *80*, 517–526.

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices test. *Psychological Review*, *97*, 404–431.

Carraher, D. W., Schliemann, A. D., Brizuela, B. M., & Earnest, D. (2006). Arithmetic and algebra in early mathematics education. *Journal for Research in Mathematics Education*, *37*, 87–115.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge, MA: Cambridge University Press.

Case, R., Kurland, D. M., & Goldberg, J. (1982). Operational efficiency and the growth of short-term memory span. *Journal of Experimental Child Psychology*, *33*, 386–404.

Cattell, R. B. (1968). Are IQ-tests intelligent? *Psychology Today*, *2*, 56–62.

Cattell, R. B. (1973). *Measuring intelligence with the Culture Fair Tests*. Champaign,

IL: Institute for Personality and Ability Testing.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*, 233–255.

Chuah, Y. M. L., & Maybery, M. T. (1999). Verbal and spatial short-term memory: Common sources of developmental change? *Journal of Experimental Child Psychology*, *73*, 7–44.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Colberg, M., Nester, M. A., & Trattner, S. M. (1985). Convergence of the inductive and deductive models in the measurement of reasoning abilities. *Journal of Applied Psychology*, *70*, 681–694.

Colom, R., Abad, F. J., Quiroga, M. A., Shih, P. C., & Flores-Mendoza, C. (2008). Working memory and intelligence are highly related constructs, but why? *Intelligence*, *36*, 584–606.

Colom, R., Flores-Mendoza, C., & Rebollo, I. (2003). Working memory and intelligence. *Personality and Individual Differences*, *34*, 33–39.

Colom, R., Rebollo, I., Abad, F. J., & Shih, P. C. (2006). Complex span tasks, simple span tasks, and cognitive abilities: A reanalysis of key studies. *Memory & Cognition*, *34*, 158–171.

Conway, A. R. A., Cowan, N., Bunting, M. F., Therriault, D. J., & Minkoff, S. R. B. (2002). A latent variable analysis of working memory capacity, short-term memory capacity, processing speed, and general fluid intelligence. *Intelligence*, *30*, 163–183.

Conway, A. R. A., Jarrold, C., Kane, M. J., Miyake, A., & Towse, J. N. (Eds.). (2007). *Variation in working memory*. Oxford: 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*, 769–786.

Cornelissen, F. W., & Greenlee, M. W. (2000). Visual memory for random block patterns defined by luminance and color contrast. *Vision Research*, *40*, 287–299.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*, 87–185.

Cowan, N. (2005). *Working memory capacity*. New York: Psychology Press.

Cowan, N., Chen, Z., & Rouder, J. N. (2004). Constant capacity in an immediate serial-recall task: A logical sequel to Miller (1956). *Psychological Science*, *15*, 634–640.

Cowan, N., Elliott, E. M., Saults, J. S., Morey, C. C., Mattox, S., Hismjatullina, A., et al. (2005). On the capacity of attention: Its estimation and its role in working memory and cognitive aptitudes. *Cognitive Psychology*, *51*, 42–100.

Cowan, N., Fristoe, N. M., Elliott, E. M., Brunner, R. P., & Saults, J. S. (2006). Scope of attention, control of attention, and intelligence in children and adults. *Memory & Cognition*, *34*, 1754–1768.

Cowan, N., Naveh-Benjamin, M., Kilb, A., & Saults, J. S. (2006). Life-span development of visual working memory: When is feature binding difficult? *Developmental Psychology*, *42*, 1089–1102.

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist*, *12*, 671–684.

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning & Verbal Behavior*, *19*, 450–466.

Daneman, M., & Merikle, P. M. (1996). Working memory and language comprehension: A meta-analysis. *Psychonomic Bulletin & Review*, *3*, 422–433.

de Boeck, P. (2008). Random item IRT models. *Psychometrika*, *73*, 533–559.

de Jong, P. F., & Das-Smaal, E. A. (1995). Attention and intelligence: The validity of the star counting test. *Journal of Educational Psychology*, *87*, 80–92.

de Jonge, P., & de Jong, P. (1996). Working memory, intelligence and reading ability in children. *Personality and Individual Differences*, *21*, 1007–1020.

De Rammelaere, S., Stuyven, E., & Vandierendonck, A. (1999). The contribution of working memory resources in the verification of simple mental arithmetic sums. *Psychological Research/Psychologische Forschung*, *62*, 72–77.

DeStefano, D., & LeFevre, J.-A. (2004). The role of working memory in mental arithmetic. *European Journal of Cognitive Psychology*, *16*, 353–386.

Engelhard, G. J. (1992). The measurement of writing ability with a many-faceted Rasch model. *Applied Measurement in Education*, *5*, 171–191.

Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. A. (1999). Working memory, short-term memory, and fluid intelligence: A latent-variable approach. *Journal of Experimental Psychology: General*, *128*, 309–331.

Fan, X., & Sivo, S. A. (2009). Using $\triangle$goodness-of-fit indexes in assessing mean structure invariance. *Structural Equation Modeling*, *16*, 54–69.

Finch, H., & Habing, B. (2007). Performance of DIMTEST- and NOHARM-based statistics for testing unidimensionality. *Applied Psychological Measurement*, *31*, 292–307.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, *37*, 359–374.

Flavell, J. H., Beach, D. H., & Chinsky, J. M. (1966). Spontaneous verbal rehearsal in a memory task as a function of age. *Child Development*, *37*, 283–299.

Fletcher, P. C., & Henson, R. N. A. (2001). Frontal lobes and human memory: Insights from functional neuroimaging. *Brain*, *124*, 849–881.

Floyd, R. G., Evans, J. J., & McGrew, K. S. (2003). Relations between measures of Cattell-Horn-Carroll (CHC) cognitive abilities and mathematics achievement across the school-age years. *Psychology in the Schools*, *40*, 155–171.

Fox, J.-P., Klein Entink, R. H., & van der Linden, W. J. (2007). Modeling of responses and response times with the package cirt. *Journal of Statistical Software*, *20*, 1–14.

Fraser, C., & McDonald, R. P. (1988). NOHARM: Least squares item factor analysis. *Multivariate Behavioral Research*, *23*, 267–269.

Friedman, N. P., & Miyake, A. (2005). Comparison of four scoring methods for the reading span test. *Behavior Research Methods*, *37*, 581–590.

Fürst, A. J., & Hitch, G. J. (2000). Separate roles for executive and phonological components of working memory in mental arithmetic. *Memory & Cognition*, *28*, 774–782.

Fry, A. F., & Hale, S. (1996). Processing speed, working memory, and fluid intelligence: Evidence for a developmental cascade. *Psychological Science*, *7*, 237–241.

Fuchs, L. S., Fuchs, D., Compton, D. L., Powell, S. R., Seethaler, P. M., Capizzi, A. M., et al. (2006). The cognitive correlates of third-grade skill in arithmetic, algorithmic computation, and arithmetic word problems. *Journal of Educational Psychology*, *98*, 29–43.

Gathercole, S. E., Lamont, E., & Alloway, T. P. (2006). Working memory in the classroom. In S. Pickering (Ed.), *Working memory in the classroom* (pp. 220–238). Oxford: Academic Press.

Gathercole, S. E., Pickering, S. J., Ambridge, B., & Wearing, H. (2004). The structure of working memory from 4 to 15 years of age. *Developmental Psychology*, *40*, 177–190.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.). Boca Raton, FL: Chapmann & Hall/CRC.

Gelman, A., Meng, X.-L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, *6*, 733–807.

Gelman, A., & Pardoe, I. (2006). Bayesian measures of explained variance and pooling in multilevel (hierarchical) models. *Technometrics*, *48*, 241–251.

Glas, C. A. W., & van der Linden, W. J. (2003). Computerized adaptive testing

with item cloning. *Applied Psychological Measurement*, *27*, 247–261.

Goel, V., Gold, B., Kapur, S., & Houle, S. (1997). The seats of reason: A localization study of deductive and inductive reasoning using PET (O15) blood flow technique. *NeuroReport*, *8*, 1305–1310.

Gregorich, S. E. (2006). Do self-report instruments allow meaningful comparisons across diverse population groups? Testing measurement invariance using the confirmatory factor analysis framework. *Medical Care*, *44*, 78–94.

Gustafsson, J.-E., & Balke, G. (1993). General and specific abilities as predictors of school achievement. *Multivariate Behavioral Research*, *28*, 407–434.

Haarmann, H. J., Davelaar, E. J., & Usher, M. (2003). Individual differences in short-term memory capacity and reading comprehension. *Journal of Memory and Language*, *48*, 320–345.

Haavisto, M.-L., & Lehto, J. E. (2004). Fluid/spatial and crystallized intelligence in relation to domain-specific working memory: A latent-variable approach. *Learning and Individual Differences*, *15*, 1–21.

Hale, S., Bronik, M. D., & Fry, A. F. (1997). Verbal and spatial working memory in school-aged children: Developmental differences in susceptibility to interference. *Developmental Psychology*, *33*, 364–371.

Halford, G. S., Cowan, N., & Andrews, G. (2007). Separating cognitive capacity from knowledge: A new hypothesis. *Trends in Cognitive Sciences*, *11*, 236–242.

Halford, G. S., Wilson, W. H., & Phillips, S. (1998). Processing capacity defined by relational complexity: Implications for comparative, developmental, and cognitive psychology. *Behavioral and Brain Sciences*, *21*, 803–864.

Hancock, G. R. (2001). Effect size, power, and sample size determination for structured means modeling and MIMIC approaches to between-groups hypothesis testing of means on a single latent construct. *Psychometrika*, *66*, 373–388.

Hancock, G. R., & Mueller, R. O. (2001). Rethinking construct reliability within latent variable systems. In *Structural equation modeling: Present and future. A Festschrift in honor of Karl Jöreskog* (pp. 195–216). Lincolnwood, IL: Scientific Software International.

Harman, G. (1999). *Reasoning, meaning, and mind*. Oxford: Oxford University Press.

Hasher, L., Lustig, C., & Zacks, R. (2007). Inhibitory meachanisms and the control of attention. In A. C. A. Conway, C. Jarrold, M. J. Kane, A. Miyake, & J. N. Towse (Eds.), *Variation in working memory* (pp. 227–249). New York: Oxford

University Press.

Heathcote, D. (1994). The role of visuo-spatial working memory in the mental addition of multi-digit addends. *Cahiers de Psychologie Cognitive/Current Psychology of Cognition*, *13*, 207–245.

Hecht, S. A. (2002). Counting on working memory in simple arithmetic when counting is used for problem solving. *Memory & Cognition*, *30*, 447–455.

Hecht, S. A., Torgesen, J. K., Wagner, R. K., & Rashotte, C. A. (2001). The relations between phonological processing abilities and emerging individual differences in mathematical computational skills: A longitudinal study from second to fifth grades. *Journal of Experimental Child Psychology*, *79*, 192–227.

Heit, E., & Rotello, C. M. (2005). Are there two kinds of reasoning? In B. G. Bara, L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the twenty-seventh annual conference of the Cognitive Science Society* (pp. 923–928). Mahwah, NJ: Erlbaum.

Heitz, R. P., Unsworth, N., & Engle, R. W. (2005). Working memory capacity, attention control, and fluid intelligence. In O. Wilhelm & R. W. Engle (Eds.), *Handbook of understanding and measuring intelligence* (pp. 61–77). Thousand Oaks, CA: Sage.

Henry, L. A., & Millar, S. (1993). Why does memory span improve with age? A review of the evidence for two current hypotheses. *European Journal of Cognitive Psychology*, *5*, 241–287.

Hitch, G. J., Towse, J. N., & Hutton, U. (2001). What limits children's working memory span? Theoretical accounts and applications for scholastic development. *Journal of Experimental Psychology: General*, *130*, 184–198.

Hoenig, J. M., & Heisey, D. M. (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. *The American Statistician*, *55*, 19–24.

Holling, H., Bertling, J., Zeuch, N., & Kuhn, J.-T. (in press). Automatische Itemgenerierung [automatic item generation]. In F. Preckel, W. Schneider, & H. Holling (Eds.), *Jahrbuch der pädagogisch-psychologischen Diagnostik: Tests und Trends, Band Hochbegabung.* Göttingen, Germany: Hogrefe.

Holmes, J., Adams, J. W., & Hamilton, C. J. (2008). The relationship between visuospatial sketchpad capacity and children's mathematical skills. *European Journal of Cognitive Psychology*, *20*, 272–289.

Holzman, T. G., Pellegrino, J. W., & Glaser, R. (1983). Cognitive variables in series completion. *Journal of Educational Psychology*, *75*, 603–618.

Horn, J. L., & Blankson, N. (2005). Foundations for better understanding of cognitive abilities. In D. P. Flanagan & P. I. Harrison (Eds.), *Contemporary*

intellectual assessment (pp. 41–68). New York: Guilford.

Humberstone, J., & Reeve, R. A. (2008). Profiles of algebraic competence. *Learning and Instruction*, *18*, 354–367.

Hutton, U. M. Z., & Towse, J. N. (2001). Short-term memory and working memory as indices of children's cognitive skills. *Memory*, *9*, 383–394.

Imbo, I., & Vandierendonck, A. (2007). Do multiplication and division strategies rely on executive and phonological working memory resources? *Memory & Cognition*, *35*, 1759–1771.

Imbo, I., & Vandierendonck, A. (2008). Effects of problem size, operation, and working-memory span on simple-arithmetic strategies: Differences between children and adults? *Psychological Research/Psychologische Forschung*, *72*, 331–346.

Ip, E. H., Smits, D. J. M., & de Boeck, P. (2009). Locally dependent linear logistic test model with person covariates. *Applied Psychological Measurement*, *33*, 555–569.

James, W. (1890). *Principles of psychology*. New York: Holt.

Janssen, R., Schepers, J., & Peres, D. (2004). Models with item and item group predictors. In P. de Boeck & M. Wilson (Eds.), *Explanatory item response models* (pp. 189–212). New York: Springer.

Jarvis, H. L., & Gathercole, S. E. (2003). Verbal and non-verbal working memory and achievements on National Curriculum tests at 11 and 14 years of age. *Educational and Child Psychology*, *20*, 123–140.

Jäger, A. O., Holling, H., Preckel, F., Schulze, R., Vock, M., Süß, H.-M., et al. (2005). *Berliner Intelligenzstruktur-Test für Jugendliche: Begabungs- und Hochbegabungsdiagnostik (BISHB) [Berlin Structure-of-Intelligence test for Youth: Assessment of giftedness]*. Göttingen, Germany: Hogrefe.

Johnson-Laird, P. N. (1994). Mental models and probabilistic thinking. *Cognition*, *50*, 189–209.

Jost, K., Hennighausen, E., & Rösler, F. (2004). Comparing arithmetic and semantic fact retrieval: Effects of problem size and sentence constraint on event-related brain potentials. *Psychophysiology*, *41*, 46–59.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, *25*, 258–272.

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, *99*, 122–149.

Kail, R., & Park, Y.-S. (1994). Processing time, articulation time, and memory

span. *Journal of Experimental Child Psychology*, *57*, 281–291.

Kane, M. J., Conway, A. R. A., Bleckley, M. K., & Engle, R. W. (2001). A controlled-attention view of working memory capacity. *Journal of Experimental Psychology: General*, *130*, 169–183.

Kane, M. J., Conway, A. R. A., Hambrick, D. Z., & Engle, R. W. (2007). Variation in working memory capacity as variation in executive attention and control. In A. C. A. Conway, C. Jarrold, M. J. Kane, A. Miyake, & J. N. Towse (Eds.), *Variation in working memory* (pp. 21–48). Oxford: Oxford University Press.

Kane, M. J., & Engle, R. W. (2002). The role of prefrontal cortex in working-memory capacity, executive attention, and general fluid intelligence: An individual-differences perspective. *Psychonomic Bulletin & Review*, *9*, 637–671.

Kane, M. J., Hambrick, D. Z., Tuholski, S. W., Wilhelm, O., Payne, T. W., & Engle, R. W. (2004). The generality of working memory capacity: A latent-variable approach to verbal and visuospatial memory span and reasoning. *Journal of Experimental Psychology: General*, *133*, 189–217.

Kane, M. J., Poole, B. J., Tuholski, S. W., & Engle, R. W. (2006). Working memory capacity and the top-down control of visual search: Exploring the boundaries of 'executive attention'. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*, 749–777.

Kaplan, D., & George, R. (1995). A study of the power associated with testing factor mean differences under violations of factorial invariance. *Structural Equation Modeling*, *2*, 101–118.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795.

Kerlinger, F. N. (1973). *Foundations of behavioral research*. New York: Rhinehart, Holt & Winston.

Klauer, K. C., Stegmaier, R., & Meiser, T. (1997). Working memory involvement in propositional and spatial reasoning. *Thinking & Reasoning*, *3*, 9–47.

Klein Entink, R. H., Fox, J.-P., & van der Linden, W. J. (2009). A multivariate multilevel approach to the modeling of accuracy and speed of test takers. *Psychometrika*, *74*, 21-48.

Klein Entink, R. H., Kuhn, J.-T., Hornke, L. F., & Fox, J.-P. (2009). Evaluating cognitive theory: A joint modelling approach using responses and response times. *Psychological Methods*, *14*, 54–75.

Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York: Guilford.

Klingberg, T., Fernell, E., Olesen, P. J., Johnson, M., Gustafsson, P., Dahlström, K., et al. (2005). Computerized training of working memory in children with ADHD – a randomized, controlled trial. *Journal of the American Academy of Child & Adolescent Psychiatry*, *44*, 177–186.

Klugkist, I., Laudy, O., & Hoijtink, H. (2005). Inequality constrained analysis of variance: A Bayesian approach. *Psychological Methods*, *10*, 477–493.

Knuth, E. J., Stephens, A. C., McNeil, N. M., & Alibali, M. W. (2006). Does understanding the equal sign matter? evidence from solving equations. *Journal for Research in Mathematics Education*, *37*, 297–312.

Krumm, S., Schmidt-Atzert, L., Buehner, M., Ziegler, M., Michalczyk, K., & Arrow, K. (2009). Storage and non-storage components of working memory predicting reasoning: A simultaneous examination of a wide range of ability factors. *Intelligence*, *37*, 347–364.

Krumm, S., Ziegler, M., & Buehner, M. (2008). Reasoning and working memory as predictors of school grades. *Learning and Individual Differences*, *18*, 248–257.

Kuhn, J.-T., & Holling, H. (2009). Gender, reasoning ability, and scholastic achievement: A multilevel mediation analysis. *Learning and Individual Differences*, *19*, 22–233.

Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (little more than) working memory capacity?! *Intelligence*, *14*, 389–433.

Leather, C. V., & Henry, L. A. (1994). Working memory span and phonological awareness tasks as predictors of early reading ability. *Journal of Experimental Child Psychology*, *58*, 88–111.

Lee, K., Ng, S.-F., Ng, E.-L., & Lim, Z.-Y. (2004). Working memory and literacy as predictors of performance on algebraic word problems. *Journal of Experimental Child Psychology*, *89*, 140–158.

Lee, N. Y. L., Goodwin, G. P., & Johnson-Laird, P. M. (2008). The psychological puzzle of Sudoku. *Thinking & Reasoning*, *14*, 342–364.

LeFevre, J.-A., Sadesky, G. S., & Bisanz, J. (1996). Selection of procedures in mental addition: Reassessing the problem size effect in adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 216–230.

Lemaire, P., Abdi, H., & Fayol, M. (1996). The role of working memory resources in simple cognitive arithmetic. *European Journal of Cognitive Psychology*, *8*, 73–103.

Lienert, G. A., & Raatz, U. (1998). *Testaufbau und Testanalyse [test design and test analysis]* (6th ed.). Weinheim, Germany: Psychologie Verlags Union.

Logie, R. H., Gilhooly, K. J., & Wynn, V. (1994). Counting on working memory in

arithmetic problem solving. *Memory & Cognition*, *22*, 395 –410.

Logie, R. H., & Pearson, D. G. (1997). The inner eye and the inner scribe of visuo-spatial working memory: Evidence from developmental fractionation. *European Journal of Cognitive Psychology*, *9*, 241–257.

Lonigan, C. J., Anthony, J. L., Phillips, B. M., Purpura, D. J., Wilson, S. B., & McQueen, J. D. (2009). The nature of preschool phonological processing abilities and their relations to vocabulary, general cognitive abilities, and print knowledge. *Journal of Educational Psychology*, *101*, 345–358.

Lubke, G. H., & Dolan, C. V. (2003). Can unequal residual variances across groups mask measurement invariance in the common factor model? *Structural Equation Modeling*, *10*, 175–192.

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*, 279–281.

Luo, D., Thompson, L. A., & Detterman, D. K. (2006). The criterion validity of tasks of basic cognitive processes. *Intelligence*, *34*, 79–120.

Lustig, C., May, C. P., & Hasher, L. (2001). Working memory span and the role of proactive interference. *Journal of Experimental Psychology: General*, *130*, 199–207.

MacCallum, R. C., Browne, M. W., & Cai, L. (2006). Testing differences between nested covariance structure models: Power analysis and null hypotheses. *Psychological Methods*, *11*, 19–35.

Mackintosh, N. J., & Bennett, E. S. (2003). The fractionation of working memory maps onto different components of intelligence. *Intelligence*, *31*, 519–531.

Marsh, H. W., Wen, Z., & Hau, K.-T. (2004). Structural equation models of latent interactions: Evaluation of alternative estimation strategies and indicator construction. *Psychological Methods*, *9*, 275–200.

Marshalek, B., Lohman, D. F., & Snow, R. E. (1983). The complexity continuum in the radex and hierarchical models of intelligence. *Intelligence*, *7*, 107–127.

Martin-Löf, P. (1973). *Statistica modeller: Anteckningar från seminarier lasåret 1969-70 utarbetade av Rolf Sundberg. [Statistial models: Notes from seminars 1969-1970, prepared by Rolf Sundberg].* (Tech. Rep.). Stockholm: Institutet för Försäkringsmatematik och Matematisk Statistik vid Stockholms Universitet.

Martínez, K., & Colom, R. (2009). Working memory capacity and processing efficiency predict fluid but not crystallized and spatial intelligence: Evidence supporting the neural noise hypothesis. *Personality and Individual Differences*, *46*, 281–286.

Maybery, M. T., & Do, N. (2003). Relationships between facets of working memory and performance on a curriculum-based mathematics test in children. *Educational and Child Psychology*, *20*, 77–92.

Maydeu-Olivares, A., Coffman, D. L., & Hartmann, W. M. (2007). Asymptotically distribution-free (ADF) interval estimation of coefficient alpha. *Psychological Methods*, *12*, 157–176.

Mayes, S. D., Calhoun, S. L., Bixler, E. O., & Zimmerman, D. N. (2009). IQ and neuropsychological predictors of academic achievement. *Learning and Individual Differences*, *19*, 238–241.

McDonald, R. P. (1989). An index of goodness-of-fit based on noncentrality. *Journal of Classification*, *6*, 97–103.

McGrew, K. (2005). The Cattell-Horn-Carroll theory of cognitive abilities: Past, present, and future. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment* (pp. 136–181). New York: Guilford.

McNeil, N. M., & Alibali, M. W. (2005). Why won't you change your mind? Knowledge of operational patterns hinders learning and performance on equations. *Child Development*, *76*, 883–899.

Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology*, *93*, 568–592.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*, 525–543.

Meredith, W., & Teresi, J. A. (2006). An essay on measurement and factorial invariance. *Medical Care*, *44*, 69–77.

Meulders, M., & Xie, Y. (2004). Person-by-item predictors. In P. de Boeck & M. Wilson (Eds.), *Explanatory item response models* (pp. 213–240). New York: Springer.

Mislevy, R. J., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika*, *55*, 195–215.

Miyake, A., Friedman, N. P., Rettinger, D. A., Shah, P., & Hegarty, M. (2001). How are visuospatial working memory, executive functioning, and spatial abilities related? A latent-variable analysis. *Journal of Experimental Psychology: General*, *130*, 621–640.

Miyake, A., & Shah, P. (Eds.). (1999). *Models of working memory*. Cambridge: Cambridge University Press.

Mogle, J. A., Lovett, B. J., Stawski, R. S., & Sliwinski, M. J. (2008). What's so special about working memory? An examination of the relationships among

working memory, secondary memory, and fluid intelligence. *Psychological Science*, *19*, 1071–1077.

Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika*, *54*, 557–585.

Oberauer, K. (1993). Die Koordination kognitiver Operationen: Eine Studie zum Zusammenhang von 'working memory' und Intelligenz [the coordination of cognitive operations: A study on the relationship between 'working memory' and intelligence]. *Zeitschrift für Psychologie*, *201*, 57–81.

Oberauer, K. (2002). Access to information in working memory: Exploring the focus of attention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 411–422.

Oberauer, K. (2005a). Binding and inibition in working memory: Individual and age differences in short-term recognition. *Journal of Experimental Psychology: General*, *134*, 368–387.

Oberauer, K. (2005b). Control of the contents of working memory - A comparison of two paradigms and two age groups. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 714–728.

Oberauer, K. (2005c). The measurement of working memory capacity. In O. Wilhelm & R. W. Engle (Eds.), *Handbook of measuring and understanding intelligence* (pp. 393–407). Thousand Oaks, CA: Sage.

Oberauer, K. (2006). Is the focus of attention in working memory expanded through practice? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*, 197–214.

Oberauer, K., & Bialkova, S. (2009). Accessing information in working memory: Can the focus of attention grasp two elements at the same time? *Journal of Experimental Psychology: General*, *138*, 64–87.

Oberauer, K., Demmrich, A., Mayr, U., & Kliegl, R. (2001). Dissociating retention and access in working memory: An age-comparative study of mental arithmetic. *Memory & Cognition*, *29*, 18–33.

Oberauer, K., & Lewandowsky, S. (2008). Forgetting in immediate serial recall: Decay, temporal distinctiveness, or interference? *Psychological Review*, *115*, 544–576.

Oberauer, K., Süß, H.-M., Schulze, R., Wilhelm, O., & Wittmann, W. W. (2000). Working memory capacity - facets of a cognitive ability construct. *Personality and Individual Differences*, *29*, 1017–1045.

Oberauer, K., Süß, H.-M., Wilhelm, O., & Wittmann, W. W. (2008). Which working memory functions predict intelligence? *Intelligence*, *36*, 641–652.

Osherson, D. N., Smith, E. E., Wilkie, O., Lopez, A., & Shafir, E. (1990). Category-based induction. *Psychological Review*, *97*, 185–200.

Parsons, L. M., & Osherson, D. (2001). New evidence for distinct right and left brain systems for deductive versus probabilistic reasoning. *Cerebral Cortex*, *11*, 954–965.

Pashler, H. (1994). Dual-task interference in simple tasks: Data and theory. *Psychological Bulletin*, *116*, 220–244.

Penner-Wilger, M., Leth-Steensen, C., & LeFevre, J.-A. (2002). Decomposing the problem-size effect: A comparison of response time distributions across cultures. *Memory & Cognition*, *30*, 1160–1167.

Preacher, K. J., Curran, P. J., & Bauer, D. J. (2006). Computational tools for probing interaction effects in multiple linear regression, multilevel modeling, and latent curve analysis. *Journal of Educational and Behavioral Statistics*, *31*, 437–448.

Preacher, K. J., Rucker, D. D., & Hayes, A. F. (2007). Addressing moderated mediation hypotheses: Theory, methods, and prescriptions. *Multivariate Behavioral Research*, *42*, 185–227.

Primi, R. (2001). Complexity of geometric inductive reasoning tasks: Contribution to the understanding of fluid intelligence. *Intelligence*, *30*, 41–70.

Qin, Y., Careter, C. S., Silk, E. M., Stenger, V. A., Fissell, K., Goode, A., et al. (2004). The change of the brain activation patterns as children learn algebra equation solving. *Proceedings of the National Academy of Sciences*, *101*, 5686–5691.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, *25*, 111–163.

Rasmussen, C., & Bisanz, J. (2005). Representation and working memory in early arithmetic. *Journal of Experimental Child Psychology*, *91*, 137–157.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*, 59–108.

Reuhkala, M. (2001). Mathematical skills in ninth-graders: Relationship with visuo-spatial abilities and working memory. *Educational Psychology*, *21*, 387–399.

Riggs, K. J., McTaggart, J., Simpson, A., & Freeman, R. P. J. (2006). Changes in the capacity of visual working memory in 5- to 10-year-olds. *Journal of Experimental Child Psychology*, *95*, 18–26.

Rijmen, F., & de Boeck, P. (2002). The random weights linear logistic test model. *Applied Psychological Measurement*, *26*, 271–285.

Rohde, T. E., & Thompson, L. A. (2007). Predicting academic achievement with

cognitive ability. *Intelligence*, *35*, 83–92.

Rost, J., & von Davier, M. (1994). A conditional item fit index for Rasch models. *Applied Psychological Measurement*, *18*, 171–182.

Süß, H.-M., & Beauducel, A. (2005). Faceted models of intelligence. In O. Wilhelm & R. W. Engle (Eds.), *Handbook of measuring and understanding intelligence* (pp. 313–311). Thousand Oaks, CA: Sage.

Süß, H.-M., Oberauer, K., Wittmann, W. W., Wilhelm, O., & Schulze, R. (2002). Working-memory capacity explains reasoning ability - and a little bit more. *Intelligence*, *30*, 261–288.

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, *66*, 507–514.

Schmiedek, F., Oberauer, K., Wilhelm, O., Süß, H.-M., & Wittmann, W. W. (2007). Individual differences in components of reaction time distributions and their relations to working memory and intelligence. *Journal of Experimental Psychology: General*, *136*, 414–429.

Schweizer, K., & Koch, W. (2002). A revision of Cattell's investment theory: Cognitive properties influence learning. *Learning and Individual Differences*, *13*, 57–82.

Scolari, M., Vogel, E. K., & Awh, E. (2008). Perceptual expertise enhances the resolution but not the number of representations in working memory. *Psychonomic Bulletin & Review*, *15*, 215–222.

Seitz, K., & Schumann-Hengsteler, R. (2000). Mental multiplication and working memory. *European Journal of Cognitive Psychology*, *12*, 552–570.

Seyler, D. J., Kirk, E. P., & Ashcraft, M. H. (2003). Elementary subtraction. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 1339–1352.

Sfard, A., & Linchevski, L. (1994). The gains and the pitfalls of reification–the case of algebra. *Educational Studies in Mathematics*, *26*, 191–228.

Shah, P., & Miyake, A. (1996). The separability of working memory resources for spatial thinking and language processing: An individual differences approach. *Journal of Experimental Psychology: General*, *125*, 4–27.

Shye, S. (1988). Inductive and deductive reasoning: A structural reanalysis of ability tests. *Journal of Applied Psychology*, *73*, 308–311.

Sinharay, S. (2005). Assessing fit of unidimensional item response theory models using a Bayesian approach. *Journal of Educational Measurement*, *42*, 375–394.

Sinharay, S., Johnson, M. S., & Stern, H. S. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement*, *30*, 298–321.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, *119*, 3–22.

Smith, B. J. (2007). BOA: An R package for MCMC output convergence assessment and posterior inference. *Journal of Statistical Software*, *21*, 1–37.

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London: Sage.

Spearman, C. (1923). *The nature of 'intelligence' and the principles of cognition*. London: Macmillan.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B*, *64*, 583–639.

Spilsbury, G., Stankov, L., & Roberts, R. (1990). The effect of a test's difficulty on its correlation with intelligence. *Personality and Individual Differences*, *11*, 1069–1077.

Stanovich, K. E. (1999). *Who is rational? Studies of individual differences in reasoning*. Mahwah, NJ: Erlbaum.

St. Clair-Thompson, H. L., & Gathercole, S. E. (2006). Executive functions and achievements in school: Shifting, updating, inhibition, and working memory. *Quarterly Journal of Experimental Psychology*, *59*, 745–759.

Stoel, R. D., Garre, F. G., Dolan, C., & van den Wittenboer, G. (2006). On the likelihood ratio test in structural equation modeling when parameters are subject to boundary constraints. *Psychological Methods*, *11*, 439–455.

Swanson, H. L. (2008). Working memory and intelligence in children: What develops? *Journal of Educational Psychology*, *100*, 581–602.

Swanson, H. L., & Beebe-Frankenberger, M. (2004). The relationship between working memory and mathematical problem solving in children at risk and not at risk for serious math difficulties. *Journal of Educational Psychology*, *96*, 471–491.

Swanson, H. L., & Howell, M. (2001). Working memory, short-term memory, and speech rate as predictors of children's learning. *Journal of Educational Psychology*, *93*, 720–734.

Swanson, H. L., & Jerman, O. (2006). Math disabilities: A selective meta-analysis of the literature. *Review of Educational Research*, *76*, 249–274.

Sweller, J. (1998). Can we measure working memory capacity without contamination from knowledge held in long-term memory? *Behavioral and Brain Sciences*, *24*, 845–846.

Tanaka, J. S. (1993). Multifaceted conceptions of fit in structural equation models.

In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 10–40). Newbury Park, CA: Sage.

Thomas, J., Zoelch, C., Seitz-Stein, K., & Schumann-Hengsteler, R. (2006). Phonologische und zentral-exekutive Arbeitsgedächtnisprozesse bei der mentalen Addition und Multiplikation bei Grundschulkindern [phonological and central-executive working memory processes in mental addition and multiplication in elementary school children]. *Psychologie in Erziehung und Unterricht*, *53*, 275–290.

Thompson, M. S., & Green, S. B. (2006). Evaluating between-group differences in latent variable means. In G. R. Hancock & G. R. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 119–169). Greenwich, CT: Information Age Publishing.

Thorell, L. B., Lindqvist, S., Nutley, S. B., Bohlin, G., & Klingberg, T. (2009). Training and transfer effects of executive functions in preschool children. *Developmental Science*, *12*, 106–113.

Tillman, C., Nyberg, L., & Bohlin, G. (2008). Working memory components and intelligence in children. *Intelligence*, *36*, 394–402.

Tolar, T. D., Lederberg, A. R., & Fletcher, J. M. (2009). A structural model of algebra achievement: Computational fluency and spatial visualisation as mediators of the effect of working memory on algebra achievement. *Educational Psychology*, *29*, 239–266.

Towse, J. N., Hitch, G. J., & Hutton, U. M. Z. (1998). A reevaluation of working memory capacity in children. *Journal of Memory and Language*, *39*, 195–217.

Turley-Ames, K. J., & Whitfield, M. M. (2003). Strategy training and working memory task performance. *Journal of Memory and Language*, *49*, 446–468.

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*, 149–171.

Unsworth, N., & Engle, R. W. (2006). Simple and complex memory spans and their relation to fluid abilities: Evidence from list-length effects. *Journal of Memory and Language*, *54*, 68–80.

Unsworth, N., & Engle, R. W. (2007a). The nature of individual differences in working memory capacity: Active maintenance in primary memory and controlled search from secondary memory. *Psychological Review*, *114*, 104–132.

Unsworth, N., & Engle, R. W. (2007b). On the division of short-term and working memory: An examination of simple and complex span and their relation to

higher order abilities. *Psychological Bulletin*, *133*, 1038–1066.

Unsworth, N., Schrock, J. C., & Engle, R. W. (2004). Working memory capacity and the antisaccade task: Individual differences in voluntary saccade control. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*, 1302–1321.

van den Noortgate, W., de Boeck, P., & Meulders, M. (2003). Cross-classification multilevel logistic models in psychometrics. *Journal of Educational and Behavioral Statistics*, *28*, 369–386.

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*, 287–308.

van der Linden, W. J. (in press). Conceptual issues in response-time modeling. *Journal of Educational Measurement*.

van der Linden, W. J., & Guo, F. (in press). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika*.

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*, 4–69.

Vernon, P. A., & Jensen, A. R. (1984). Individual and group differences in intelligence and speed of information processing. *Personality and Individual Differences*, *5*, 411–423.

Vock, M., & Holling, H. (2008). The measurement of visuo-spatial and verbal-numerical working memory: Development of IRT-based scales. *Intelligence*, *36*, 161–182.

Vogel, E. K., McCollough, A. W., & Machizawa, M. G. (2005). Neural measures reveal individual differences in controlling access to working memory. *Nature*, *438*, 500–503.

Wagenmakers, E.-J. (2009). Methodological and empirical developments for the ratcliff diffusion model of response times and accuracy. *European Journal of Cognitive Psychology*, *21*, 641–671.

Wagenmakers, E.-J., van der Maas, H. L. J., & Grasman, R. P. P. P. (2007). An EZ-diffusion model for response time and accuracy. *Psychonomic Bulletin & Review*, *14*, 3–22.

Waltz, J. A., Knowlton, B. J., Holyoak, K. J., Boone, K. B., Mishkin, F. S., Menezes Santos, M. de, et al. (1999). A system for relational reasoning in human prefrontal cortex. *Psychological Science*, *10*, 119–125.

Wang, W.-C., & Wilson, M. (2005). The Rasch testlet model. *Applied Psychological Measurement*, *29*, 126–149.

Watkins, M. W., Lei, P.-W., & Canivez, G. L. (2007). Psychometric intelligence and achievement: A cross-lagged panel analysis. *Intelligence*, *35*, 59–68.

Weiß, R. H. (1998). *Grundintelligenztest Skala 2 (CFT 20) mit Wortschatztest (WS) und Zahlenfolgentest (ZF) [Basic intelligence test, scale 2 with vocabulary test and number series]* (4th ed.). Göttingen, Germany: Hogrefe.

Weiß, R. H. (2006). *Grundintelligenztest Skala 2 - Revision (CFT 20-R) [Basic intelligence test, scale 2 - revised]*. Göttingen, Germany: Hogrefe.

Wheeler, M. E., & Treisman, A. M. (2002). Binding in short-term visual memory. *Journal of Experimental Psychology: General*, *131*, 48–64.

Wicherts, J. M., Dolan, C. V., & Hessen, D. J. (2005). Stereotype threat and group differences in test performance: A question of measurement invariance. *Journal of Personality and Social Psychology*, *89*, 696–716.

Wicherts, J. M., Dolan, C. V., Hessen, D. J., Oosterveld, P., van Baal, G. C. M., Boomsma, D. I., et al. (2004). Are intelligence tests measurement invariant over time? Investigating the nature of the Flynn effect. *Intelligence*, *32*, 509–537.

Wickelgren, W. A. (1977). Speed-accuracy tradeoff and information processing dynamics. *Acta Psychologica*, *41*, 67–85.

Wilhelm, O. (2005). Measuring reasoning ability. In O. Wilhelm & R. W. Engle (Eds.), *Handbook of understanding and measuring intelligence* (pp. 373–392). Thousand Oaks, CA: Sage.

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language*, *46*, 441–517.

Yuan, K.-H., & Bentler, P. M. (2004). On chi-square difference and z tests in mean and covariance structure analysis when the base model is misspecified. *Educational and Psychological Measurement*, *64*, 737–757.

Yuan, K.-H., & Bentler, P. M. (2006). Mean comparison: Manifest variable versus latent variable. *Psychometrika*, *71*, 139–159.

Zwinderman, A. H. (1991). A generalized Rasch model for manifest predictors. *Psychometrika*, *56*, 589–600.

Summary in German / Zusammenfassung

Die Psychologie hat sich seit ihren Anfängen mit der systematischen Analyse von Denkprozessen auseinandergesetzt. Das schlussfolgernde Denken nimmt in nahezu allen Intelligenztheorien einen zentralen Raum ein. Ein wichtiges Konstrukt der Kognitionspsychologie, das Arbeitsgedächtnis, hat sich bei komplexen Denkprozessen als ein kritischer Faktor erwiesen. Zahlreiche Studien belegen einen engen Zusammenhang von Arbeitsgedächtniskapazität und Intelligenz sowie schlussfolgerndem Denken (vgl. Ackerman, Beier, & Boyle, 2005), auch wenn diese Konstrukte sowohl empirisch als auch konzeptuell unterscheidbar sind (Blair, 2006). Das Arbeitsgedächtnis kann beschrieben werden als eine kapazitätsbegrenzte kognitive Ressource, die das simultane Speichern und Verarbeiten von Informationen ermöglicht.

Totz der häufig berichteten hohen Zusammenhänge von Arbeitsgedächtniskapazität und schlussfolgerndem Denken existieren in der Literatur noch etliche offene Fragen, insbesondere im Hinblick auf eine Charakterisierung kognitiver Prozesse beim schlussfolgernden Denken sowie insbesondere deren Wechselwirkung mit Arbeitsgedächtnisprozessen. Des Weiteren wurden überwiegend Prozesse bei erwachsenen Probanden untersucht, welche aber nicht unbedingt auf Kinder generalisierbar sind. In der vorliegenden Arbeit soll daher anhand einer detaillierten Analyse von Testaufgaben und Reaktionszeitdaten zum schlussfolgernden Denken und Arbeitsgedächtnis mittels Strukturgleichungsmodellen sowie Modellen der Item Response Theory einer Beantwortung dieser Fragen nachgegangen werden.

Berichtet werden Ergebnisse aus drei separaten Studien. Die erste Studie analysierte Zusammenhänge von Arbeitsgedächtnis, Kurzzeitgedächtnis, sowie fluider Intelligenz und kristalliner Intelligenz. Es konnte zunächst gezeigt werden, dass das Arbeitsgedächtnis sich vom Kurzzeitgedächtnis separieren lässt, sowie dass diese latente Faktorstruktur messäquivalent hinsichtlich jüngerer und älterer Kinder ist. Weiterhin konnte mittels Strukturgleichungsmodellen nicht zwischen deduktiven und induktiven schlussfolgernden Denkaufgaben unterschieden werden. Die Bedeutung des Kurzzeitgedächtnisses zur Vorhersage von

Intelligenzleistungen nahm mit zunehmendem Alter der Kinder ab, so dass ab ca. 11 Jahren allein die Arbeitsgedächtniskapazität für die Vorhersage von Intelligenzleistungen von Bedeutung war. Wurde die Arbeitsgedächtniskapazität im Sinne von Colom, Rebollo, Abad, and Shih (2006) als Residualfaktor spezifiziert, blieb sie ein entscheidender Prädiktor zur Vorhersage von Intelligenzleistungen.

In einer zweiten Studie wurde untersucht, inwieweit algebraisches Denken bei Kindern durch unterschiedliche kognitive Faktoren beeinflusst wird. Dazu wurde ein bivariates IRT-Modell mit Zufallseffekten spezifiziert (Klein Entink, Kuhn, Hornke, & Fox, 2009). Es zeigte sich, dass eine konkurrente Gedächtnisbelastung während des Lösens von Algebraaufgaben zu einer Erhöhung der Aufgabenschwierigkeit führt (vgl. Oberauer, Demmrich, Mayr, & Kliegl, 2001), während die Größe der zu verarbeitenden Zahlen (*number size effect*) keine Auswirkungen hatte. Eine simultane Analyse der Reaktionszeizdaten zeigte ähnliche Resultate im Hinblick auf die Zeitintensität der Algebraaufgaben. Die kognitive Verarbeitungsgeschwindigkeit wurde mittels zentraler Parameter des EZ-Diffusionsmodells (Wagenmakers, van der Maas, & Grasman, 2007) operationalisiert. Es zeigte sich, im Gegensatz zu anderen Arbeiten, kein Zusammenhang der "drift rate" mit der algebraischen Leistungsfähigkeit, wohl aber mit der Testbearbeitungsgeschwindigkeit.

In einer dritten Studie wurde untersucht, inwieweit sich ein Testverfahren zum schlussfolgernden Denken (Lateinische Quadrate) basierend auf der Theorie der relationalen Komplexität (Halford, Wilson, & Phillips, 1998) konzipieren lässt, und ob zentrale Aussagen dieser Theorie als zutreffend angesehen werden können. Erwartungsgemäß war die relationale Komplexität ein entscheidender Prädiktor für die Aufgabenschwierigkeit, ebenso wie eine konkurrente Gedächtnisbelastung. Die Größe der zu verarbeitenden Chunks war im Gegensatz dazu vernachlässigbar. Mittels komplexer IRT-Modelle konnte nachgewiesen werden, dass insbesondere im Hinblick auf die komplexesten Verarbeitungsschritte große interindividuelle Unterschiede bestehen. In einer abschließenden Diskussion werden die Ergebnisse kurz reflektiert sowie ein Ausblick auf potenzielle zukünftige Forschungsaktivitäten gegeben.

Acknowledgements

First of all, I would like to thank Prof. Dr. Heinz Holling for giving me the opportunity to write this thesis, for his continuing support, and confidence in this project.

I am also grateful to Prof. Dr. Dr. h. c. Bernd Schäfer, for acting as an assessor for this dissertation, and for giving me the opportunity to extend my horizon by working in the field of social psychology.

Further, the following individuals deserve a special thank you:

Dr. Rinke Klein Entink, for another fruitful scientific cooperation.

Prof. Dr. Ralf Schulze, for enabling me to use his notebooks for test administrations.

PD Dr. Günther Gediga, for programming the working memory tasks.

Jonas Bertling, for some fruitful discussions on thesis-related topics.

Nina Zeuch, Bernadette Gold, Marlene Pacharra, Kathrin Gediga, and many research interns and student assistants for their help in data collection, coding, programming, and other helpful activities. Marlene Pacharra deserves additional praise for thorough proofreading, all errors that remain are entirely my own.

I would like to thank my family, especially my parents Klaus and Dorothea Kuhn, for their continuous help and support. Penelope often cheered me up at times when the work was hard-going. One day, not far in the future, we will read books again, in the evening, on a regular basis. I want to thank Eva Nonhoff-Kuhn for her support over the last years.

Finally, I thank Eva-Maria Schiller. Our life changed forever since the sunny miracle days of Bamberg, and every day with you is filled with wonder. Our journey continues together - West of the Moon, East of the Sun.

Curriculum Vitae

Name:	Jörg-Tobias Kuhn
Born:	April 9th, 1977, in Münster (Westf.)
1997-2004	Study of Psychology at the Westfälische Wilhelms-Universität Münster and Rheinisch-Westfälische Technische Hochschule Aachen
2004	Graduation Thesis: Analysen zum Zeitverhalten bei computergestützten Intelligenztests Supervisors: Prof. Dr. Lutz F. Hornke and Prof. Dr. Will Spijkers
2004-2009	PhD Student at the Westfälische Wilhelms-Universität Münster Supervisors: Prof. Dr. Heinz Holling and Prof. Dr. Dr. h. c. Bernd Schäfer