Mem Cogn (2014) 42:448–463
DOI 10.3758/s13421-013-0366-x

The relational integration task explains fluid reasoning above and beyond other working memory tasks

Adam Chuderski

Published online: 25 September 2013

Abstract This study aimed to evaluate how well fluid reasoning can be predicted by a task that involves the monitoring of patterns of stimuli. This task is believed to measure the effectiveness of relational integration—the process that binds mental representations into more complex relational structures. In Experiments 1 and 2, the task was indeed validated as a proper measure of relational integration, since participants' performance depended on the number of bindings that had to be constructed in the diverse conditions of the task, whereas neither the number of objects to be bound nor the amount of elicited interference could affect this performance. In Experiment3, by means of structural equation modeling and variance partitioning, the relation integration task was found to be the strongest predictor of fluid reasoning, explaining variance above and beyond the amounts accounted for by four other kinds of well-established working memory tasks.

Keywords Relational integration · Working memory · Fluid reasoning

Fluid reasoning (fluid intelligence, Gf), most often assessed with matrix problems or visual analogies (Snow, Kyllonen, and Marshalek 1984), has been assumed to be the core component ability in most of the influential models of human intelligence (see McGrew 2009). Because of the fact that measures of working memory capacity (WMC) appear to most strongly predict Gf, for the last 20 years most researchers' views (e.g., Cowan 2001; Kane et al. 2007a; Oberauer et al. 2007) have converged on the idea that Gf primarily relies on working memory (WM)—a mechanism for active maintenance and transformation of a limited amount of information crucial for the current task (Cowan 2001).

However, several different methods have been developed to tap WMC. A classic method, which was derived from the research on short-term memory (STM), involves tasks that require memorizing a set of several stimuli, and then either recalling that set (i.e., recall/span tasks) or deciding whether a subsequent stimulus was or was not drawn from it (e.g., the Sternberg task or the change detection paradigm). Despite early skepticism regarding the plausibility of STM tasks as both WMC measures and Gf predictors (e.g., Engle, Tuholski, Laughlin, and Conway 1999), more recent studies have suggested that proper versions of these tasks (i.e., excluding mnemotechniques like chunking and phonological rehearsal) can be very useful tools in WM and Gf research. It has been found that the number of items that people can successfully maintain in WM predicts a substantial part of variance in Gf (e.g., Colom, Abad, Quiroga, Shih, and Flores-Mendoza 2008; Unsworth and Engle 2007b).

Another class of paradigmatic WM tests, called *complex span tasks*, combine the maintenance of several stimuli for later recall with a number of simple manual decisions, and appear especially popular in psychometric research (see Conway et al. 2005; Unsworth and Engle 2007b). Because these tasks also predict various measures of executive control, like error rates in the antisaccade task (Unsworth, Shrock, and Engle 2004) and lapses of attention in the psychomotor vigilance task (Unsworth, Redick, Lakey, and Young 2010), some investigators (e.g., Burgess, Gray, Conway, and Braver 2011; Kane et al. 2007a) have proposed that the performance in both complex span tasks and Gf tests depends primarily on the effectiveness of domain-general control over attention. In consequence, tasks that do not require any memorization, but instead impose a strong load on

A. Chuderski
Cognitive Science Department, Institute of Philosophy, and Institute of Psychology, Jagiellonian University, Krakow, Poland

A. Chuderski (✉)
Institute of Philosophy, Jagiellonian University, Grodzka 52, 31-044 Krakow, Poland
e-mail: adam.chuderski@gmail.com

executive processes, like the antisaccade task, have also been used with some success as predictors of Gf (Unsworth et al. 2010; Unsworth and Spillers 2010; Unsworth, Spillers, and Brewer 2009).

In cognitive neuroscience, the so called *n*-back task has gained much popularity as a WMC measure (see Owen, McMillan, Laird, and Bullmore 2005). This task requires continuous memory updating and simple decision-making about whether the current stimulus in a stream of stimuli does or does not match the stimulus presented *n* stimuli back. Although some researchers have suggested that *n*-back tasks measure a somewhat different aspect of WM than do complex span tasks (Kane et al. 2007b), more recent studies have shown that latent variables based on scores from the former and the latter measures are statistically indistinguishable (Schmiedek, Hildebrandt, Lövdén, Wilhelm, and Lindenberger 2009).

In contrast to the aforementioned views about the proper measurement of WMC, Oberauer et al. (2007) proposed that the driving force of strong WMC–Gf correlations is neither the sheer storage of information, even in the context of processing, nor the executive control of that storage. In line with analogous theorizing (Halford, Wilson, and Phillips 1998; Hummel and Holyoak 2003; Waltz et al. 1999), Oberauer et al. (2007; Oberauer, Süß, Wilhelm, and Wittmann 2008) proposed that the fundamental mechanism that determines both WMC and fluid reasoning is the human capacity to set and maintain the flexible, temporary bindings between chunks held in WM, or between them and their respective positions within some mental structure. For instance, these positions can constitute concrete coordinates like serial positions during recall, or they can be abstract placeholders in some schema required in a reasoning task (so-called *role-filler bindings*). Due to temporary bindings, a person is able to integrate elementary relations into novel arbitrary relational structures. Creating such structures is the essence of relational thinking—thinking driven by the way objects are assigned to certain roles in situations, and not by objects' intrinsic features.

For the purpose of measurement of the effectiveness of relational integration, Oberauer and colleagues (Oberauer, Süß, Schulze, Wilhelm, and Wittmann 2000; Oberauer et al. 2008) have developed versions of a so-called *relation-monitoring task* (henceforth called the *relation integration task*). In such a task, a participant observes a constantly changing pattern of stimuli that is available perceptually (no need for storage in WM), and detects stimuli matching a simple rule. For example, the task may consist of the presentation of a three-by-three matrix of words, and may require the pressing of a button if and only if three words in a row, column, or diagonal line rhyme. Other versions require three numbers that end with the same digit to be found, or recognizing four dots that form a square within a pattern of several dots. A few studies showed that the latent variables loaded by the relation integration tasks are at least as strong predictors of fluid reasoning as are complex spans (Buehner, Krumm, and Pick 2005; Buehner, Krumm, Ziegler, and Pluecken 2006; Krumm et al. 2009; Oberauer et al. 2008; Süß, Oberauer, Wittmann, Wilhelm, and Schulze 2002), and much better predictors than both STM and executive control tasks (Chuderski, Taraday, Nęcka, and Smoleń 2012).

However, two questions regarding the relational integration hypothesis seem to have so far gained too little attention. First, no research has been conducted on the properties of relation integration tasks. Although *n*-back (e.g., Kane et al. 2007b; Schmiedek et al. 2009) and complex span (e.g., Conway et al. 2005; Unsworth and Engle 2007a, 2007b) tasks have undergone substantial examination, testing the ways in which the parameters of these tasks influence various indices of task performance, as well as the strength of their correlations with Gf, no similar questions have been posed with regard to the measures of relational integration.

Second, relation integration tasks have been compared to competing Gf predictors only to a limited extent. For example, Oberauer et al. (2008; see also Buehner et al. 2005; Buehner et al. 2006; Süß et al. 2002) tested their relation-monitoring tasks against complex span tasks and task-switching tests, and found that the predictive power of relation monitoring was comparable to the power of the former tasks, but much better than the power of the latter tasks. However, this study was undermined by the fact that the task-switching paradigm has been assessed as a poor measure of executive control (e.g., Logan and Bundesen 2003). Krumm et al. (2009) compared the relation integration task to the wide range of measures of storage capacity (including both simple and complex spans), sustained attention, and mental speed. They found that relation integration predicts Gf even when all of the latter variables are accounted for, but they failed to obtain a homogeneous inhibition/attention control factor using three different executive control tasks (the Stroop, antisaccade, and stop-signal tasks). Chuderski et al. (2012), in their Study 2, showed that relational integration seemed to be a better Gf predictor than the scope and the control of attention (the latter factor being validly measured with the Stroop and antisaccade tasks), though this study did not include complex span tasks. Thus, the comprehensive evaluation of the predictive power of all four widely used types of WM measures described above (i.e., STM, complex span, attention control, and *n*-back tasks) with regard to Gf, in comparison to relational integration, seems to call for more data.

Consequently, the goals of the present research were to study a novel version of the relation integration task, in order to examine (a) whether the level of performance in this task can be determined by the need to construct and/or integrate bindings among mental representations, and not only by the need to select and maintain those very representations; (b) to what extent such performance can be explained by the amount of executive control that may be required for coping with interference present in the trials of this task; (c) whether versions of the

relation integration task differing in difficulty would or would not differ in predictive power with regard to reasoning; and (d) whether this task might or might not be a better Gf predictor than other WM tasks. The body of data cited above suggests that relational integration could be a highly fruitful line of investigation pertaining to WM and its links to Gf. Thus, the tasks developed in order to tap relational integration need especially careful examination. Answering the questions above could lead cognitive science to a better understanding of the crucial mechanisms and constraints of WM, and could also shed some new light on the causes of the close links between WM and general intelligence.

Rationale and general method

What is so crucial about finding three matching items or four symmetrically located dots that in previous studies has made the relation-monitoring task such a powerful WMC measure? Three explanations seem likely.

First, Oberauer et al. (2008) may be right, and the performance in this task may primarily reflect the binding and/or integration of a few pieces of information in WM. For example, the task may require integrating "item X is bound to the upper-left location," "… is also bound to the central-left location," and "… is also bound to the bottom-left location" into one relational structure that encodes all of the information: "upper-left, central-left, and bottom-left locations form a column containing an item X." Once such a structure is integrated, it may be relatively simple to detect the match of the current pattern to the predefined relation, whereas without that structure in mind, one can easily miss the match. So, under the relation integration hypothesis, the increase in the number of bindings will make the relation-monitoring task more difficult, and people who represent and/or integrate bindings better will score higher on that task.

Second, the integration of relations may primarily require the maintenance of prospective objects (e.g., those that partially satisfy the relation) within the scope of attention, while perceptually scanning the stimulus pattern for the remaining objects. So, before a cognitive system can relate all those objects, it may have to transfer their representations from perception to the WM's scope of attention. Thus, the load caused by the number of objects to be related, and not necessarily the need to establish bindings among them, may determine the difficulty of the relation integration task. Since the variants of the task used so far have required the relating of three (syllables, digits) or four (dots) objects, which were values around the mean scope of attention in the population (Cowan 2001; Luck and Vogel 1997), a certain portion of the population may possess an insufficient scope of less than three or four slots to perform the task without error. The remaining portion may have enough capacity, and thus the relation integration task would correlate with other WM tasks, as well as with fluid intelligence tests.

Finally, the relation-monitoring task involves the presentation of a relatively complex pattern of stimuli (e.g., 27 digits in the number version of the task) that need to be scanned, and in some cases selected, while at the same time other stimuli may be competing for selection (e.g., three identical digits not placed in one row, column, or diagonal line). In consequence, the task may require a substantial amount of attention control: to carry out scanning, selection, interference resolution, and inhibition. According to the executive-attention theory of WM and Gf (Kane et al. 2007a), these very requirements may cause significant correlations among this task and WM and Gf tests.

In order to discriminate between the possibilities presented above, I modified the original word and number relation integration tasks in such a way that I was able to manipulate (a)the number of bindings to be integrated, (b)the number of objects to be related, and (c)the level of interference imposed by the stimuli in the task. As a result, which factor(s) had a significant impact on response accuracy in the modified relation integration tasks (Exps. 1–3) could be observed.

Experiment 1

Method

Participants

A total of 112 people, randomly assigned to groups, participated (73 women, 39 men; mean age = 24.4 years, *SD* = 5.2, range 19–45 years). All of them were recruited via publicly accessible social networking websites. For their participation, each person received the equivalent of €5 in Polish zloty. Participants were tested in groups of several people.

Materials and procedure

Modified, no-memory versions of the alphanumeric monitoring task introduced by Oberauer et al. (2000) were used. The task consisted of the presentation of a continuous sequence of symbol patterns (trials). Each trial included a 3 × 3 array (approx. 6 × 6 cm in size) of three-symbol strings (each string approx. 1.5 × 1.2 cm in size). In one version of the task (50 trials preceded by five training trials), the strings in the sequence contained three letters out of a set of ten consonants. In the subsequent version (also 50 trials, no training) they were three-digit numbers. In one ("three-same") group of participants, they were asked to detect whether three strings ending with the same letter/digit were located in one row or column. In another ("three-different") group, they were required to respond if three strings ending with three different letters/digits were located in one row or column (note that both three-object conditions of the

modified task, in contrast to Oberauer et al.'s 2000, original task, did not require participants to monitor the diagonal lines for the target patterns). In the third ("five-same") group, participants had to find five identical ending letters/digits forming a cross, a "letter T" pattern, or a "letter T" pattern rotated by 90, 180, or 270 deg. Finally, the fourth ("five-different") group was instructed to find five different ending letters/digits placed in one of the five patterns above.

In each group, half of the arrays matched a predefined rule, requiring the user to press the space button. Exactly three or five strings fulfilled the rule, in the three- or five-object conditions, respectively, and all other letters/digits in an array were randomly chosen, with the constraint that they could not match the rule. In the other half of the arrays, no patterns matched the rule. Participants were instructed to respond only to trials that included the target relation, and to withhold their responses in all other (i.e., no-relation) trials. So, responses given in the latter trials were interpreted as false alarm errors. In order to minimize the influence on the results of either processing speed or visual search efficiency, in each trial 5.5 s (plus a 0.1-s blink separating the subsequent arrays) were allowed. Also, in order to decrease the amount of information changed from trial to trial, one to four strings (at random) in each subsequent array were the same as in the preceding array (i.e., not all stimuli changed from trial to trial). Examples of the arrays in each task condition are presented in Fig. 1.

The four groups constituted the 2 × 2 experimental design, with two factors manipulated between participants: the number of target objects (either three or five) and the type of target relation (either same or different objects). The dependent variable was the mean accuracy in the relation trials minus the mean false alarm rate in the no-relation trials (see Snodgrass and Corwin 1988).

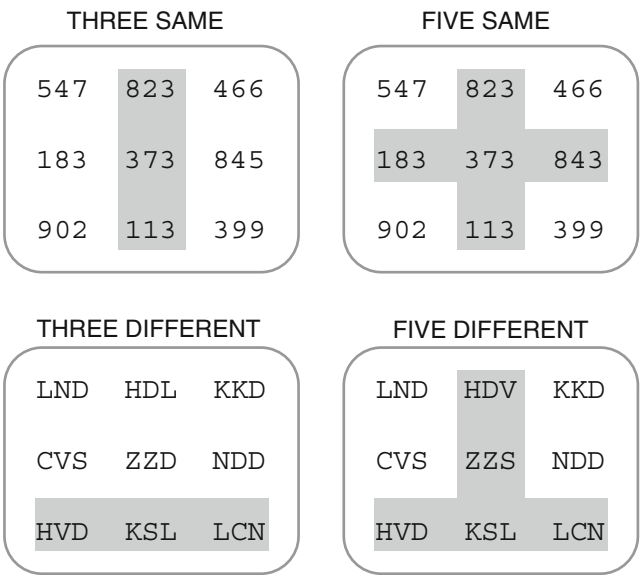

Fig. 1 Sample target trials of the relation-monitoring task (as used in Exp. 1), for each condition with regard to the number of strings to be detected (three or five) and the predefined rule to be followed (find the same or different end letters/digits). Gray boxes (absent in the real task) indicate target strings. The top row includes samples of the number variant of the task, whereas the bottom row contains instances of the letter variant

Hypotheses

By introducing the three- and five-different conditions, I aimed to increase the number of bindings between the ending letters/digits (the target objects). I assumed that in the three- and five-same conditions, target objects can be compared incrementally, and only two bindings might be sufficient to encode the eventual relational structure. That is, in these trials, participants had to detect that the first and second objects were the same and to construct a respective item–item binding (e.g., "two Xs in an *n*th row"), and then to compare the next object with the information that was encoded by the binding ("is a third object also X"?). In contrast, this incremental processing did not seem possible in the three- or five-different trials, because there was no common letter/digit, and one binding might have to be constructed for each pair of target objects (e.g., "a first object in a row is different from a second one," "a second object is different from a third one," but also "a third object is different from a first one"). Such a construction should result in a load of three bindings in the three-different trials, and as many as ten bindings in the five-different trials (i.e., the number of two-element combinations of a five-element set). Alternatively, if in the three- and five-different trials participants dealt with the task by forming item–context bindings, specifically by binding all stimuli to the respective positions in the row/column or the cross/T pattern, the latter pattern would resulted in as many as five simultaneous bindings. In either case (i.e., either item–item or item–context bindings), the number of necessary bindings in the five-different condition (i.e., either ten or five bindings) exceeded the WMC of most participants. In consequence, the relational-integration hypothesis predicts that increasing the number of objects that underlie the target relation would not influence accuracy when the same objects have to be looked for (i.e., always only two bindings are necessary), but would drastically decrease accuracy when different objects have to be found (i.e., because more objects result in more bindings). In contrast, if the bare number of target objects that have to be simultaneously attended to in order to be compared determines the difficulty of the relation integration task (see Oberauer et al. 2008, p. 650), accuracy would decrease in similar ways in both the five-same and five-different conditions, in comparison to the three-same and three-different conditions, respectively.

Results and discussion

First, I verified that no significant difference in the ratio of false alarms to all errors (i.e., the beta parameter) could be found between groups, all *p*s > .6. Then, a 2 × 2 ANOVA in

accuracy indicated that both main effects were significant—that is, the effect of three (M = .56) versus five (M = .45) objects, $F(1, 108) = 7.77$, $p = .006$, $\eta^2 = .07$, and of the same (M = .73) versus different (M = .27) objects, $F(1, 108) = 144.91$, $p < .001$, $\eta^2 = .57$. Most importantly, however, the pattern of results was explained by the two-way interaction between the factors, $F(1, 108) = 6.84$, $p = .010$, $\eta^2 = .06$, which indicated that the former factor had no influence when the same objects were detected (both Ms = .73), $F(1, 108) = 0.018$, $p = .892$, $\eta^2 = .0003$, but drastically decreased performance when different objects were searched for (M = .38 vs. M = .18, in the cases of three and five objects, respectively), $F(1, 108) = 12.02$, $p = .001$, $\eta^2 = .23$.

Since the accuracy observed in all groups nicely correlated with the supposed number of bindings that were necessary to fulfill each condition (i.e., two, two, three, and five/ten—the latter number depending on assuming either the item–context or the item–item interpretation of binding performance—in the three-same, five-same, three-different, and five-different conditions, respectively), the conclusion stating that the relation integration task indeed taps the difficulty of relational integration seems to be the most plausible. This conclusion is also consistent with the fact that in Experiment 1 a very strong effect emerged of the same- and different-object conditions, because between the former and latter conditions the number of to-be-integrated bindings seems to have increased by 100%, if we accept the item–context interpretation of binding performance (i.e., it increased from 2 to [3 + 5]/2 = 4 bindings, on average), or even by 225%, in light of the item–item interpretation (i.e., from 2 to [3 + 10]/2 = 6.5 bindings, on average). Importantly, the number of objects (either three or five) that supposedly needed to be maintained within the scope of attention had no influence on performance in the same-object condition.

Experiment 2

The difficulty of the relation integration task may also be linked to the amount of interference, for example the number of items in arrays that match only some conditions of a target rule (i.e., are all identical or all different, but not placed in end locations nor within a row/column/cross/T patttern). A positive relation between the similarity of targets and distractors, and the difficulty of processing targets is predicted, for instance, by Duncan and Humphreys's (1989) theory of visual search. This could partially explain the results of Experiment 1, as much more interference can be expected when detecting different objects (because there will be plenty of patterns of different objects, as each symbol is represented by only 10% of objects in the array) than when looking for the same objects (because only those 10% of identical objects will constitute distractors). If the experimentally controlled increase in the number of distractors (identical digits/letters placed in arrays, but not forming key relations) can decrease the hit rate in the relation trials (e.g., because the target relation cannot be filtered out) and/or increase the false alarm rate (e.g., because false alarms may be committed on the basis of partial matches), resulting in a negative effect on the overall accuracy, this might indicate that the relation integration task, at least to some extent, reflects not only relational integration, but also coping with interference. In Experiment 2, I aimed to test this possibility. Moreover, by using a within-subjects design with regard to the three-same versus five-same conditions, any influence of individual differences on the comparison of these two conditions could be eliminated.

Method

Participants

A total of 40 people participated (26 women, 14 men; mean age = 23.9 years, SD = 3.6, range 19–33 years). All other recruitment and testing conditions were the same as in Experiment 1.

Materials and procedure

The three-same and five-same conditions were applied to each participant, in a random order. Each condition contained 60 number and 60 letter trials, and each was preceded by ten number training trials. Apart from that, the task was identical to that from Experiment 1, with one exception, which constituted the key manipulation: In a random half of the arrays, all stimuli besides the three/five identical digits/letters in the relation trials were chosen randomly (the low-interference condition). In contrast, in the other half of the arrays (the high-interference condition; for an example, see Fig. 2), 12 stimuli in that set were identical (if it was the relation trial, they were also identical to a digit/letter in the target relation). Those additional identical stimuli, however, could not be placed in ending positions of one row/column or cross/T-pattern (depending on the condition). It was expected that if coping with interference is important for performance in the relation-monitoring task, the large number of fake (almost correct) target patterns should negatively affect accuracy, in comparison with the low-interference condition.

Results and discussion

Although the power of the experiment had been increased, the factor associated with the number of objects to be related—three (M = .79) versus five (M = .74)—was still not able to reach the adopted level of significance, $F(1, 39) = 3.75$, $p = .060$, $\eta^2 = .09$. I observed a trend toward a decrease in accuracy with an increasing number of objects ($\Delta = -.05$), which with an increased sample would probably turn into a

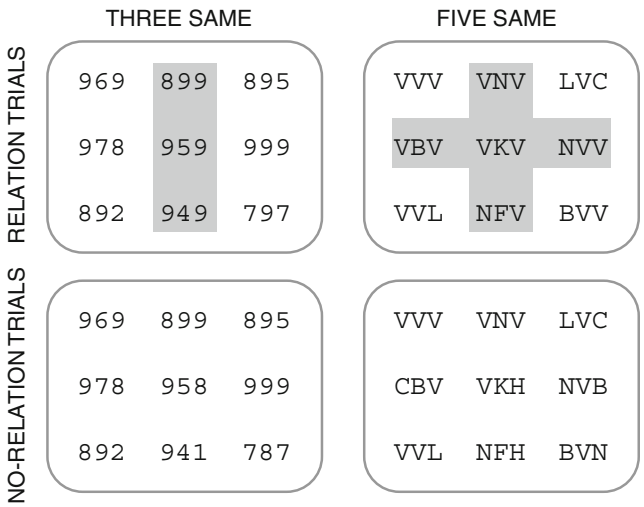

Fig. 2 Sample trials of the relation-monitoring task in the high-interference condition (used in Exp. 2), with regard to the number of strings to be detected (three or five) and the presence (relation trials) or absence (no-relation trials) of the target relation. Gray boxes (absent in the real task) indicate target strings in the relation trials. The left column includes samples of the number variant of the task, whereas the right column contains instances of the letter variant

significant effect, and which might result from a relatively more complex instruction in the five-object condition (i.e., the need to detect cross/T patterns instead of more uniform rows and columns). However, it did not even approach the dramatic decrease in the different-object condition of Experiment 1 (Δ = –.21). Regarding the interference factor, the low- (M = .76) and high- (M = .77) interference trials barely differed in accuracy, $F(1, 39) = 0.72$, $p = .40$, $\eta^2 = .02$.

Experiment 2 delivered two null effects. First, replicating the results of Experiment 1, accuracy did not substantially decrease with the increasing number of to-be-related objects. Second, the increased level of interference had no impact on accuracy. Taking into account the results of both experiments, it could be concluded that what primarily affects accuracy in the relation-monitoring task is the load imposed on the mechanisms responsible for the construction, maintenance, and/or integration of bindings among the target objects, which depends on the number of bindings that have to be dealt with in parallel. At the same time, neither the actual number of objects in the target patterns nor the number of distractors present within the patterns could affect that accuracy.

Experiment 3

Experiment 3 had two goals. First, by using a large sample of participants and each of the most widely used WMC measures, described in the introduction (i.e., complex span tasks, n-back tasks with highly effective lures, STM tasks, and antisaccade tasks), this experiment was aimed at evaluating the relative plausibility of the relation-monitoring task as a predictor of fluid reasoning, and also its validity as a WMC measure. Additionally, my other goal was a comparison between the strengths of the Gf correlations yielded by the five-same and three-different conditions, and with reference to the three-same condition, asking whether either increasing the number of objects or altering the type of the rule might change the predictive power of the monitoring task (the five-different condition was not applied, because of the floor effects observed in Exp. 1).

Method

Participants

A total of 243 people participated (142 women, 101 men; mean age = 24.3 years, SD = 5.0, range 18–45 years). For participation, each person received the equivalent of €15 in Polish zloty. All other recruitment and testing conditions were the same as in Experiments 1 and 2.

Materials and procedure

The study was administered in two sessions, each lasting 4 h (including proper intrasession breaks), both taking place on the same day (with a 1-h intersession break), and applied in a random order. One session consisted of the relation-monitoring task, applied first, as well as several other computerized WM tasks, including four complex span tasks, four STM tasks, three antisaccade tasks, and three n-back tasks (and also two Stroop tasks and a stop signal task, unreported here). The other WM tasks were administered in a fixed order, in such a way that one task from a particular class of tasks (i.e., complex span, STM, antisaccade, or n-back) was applied early in the session, one or two in the middle of the session, and one late in the session, in order to balance the amounts of automatization and tiredness/boredom regarding each class of tasks. Always, two subsequent tasks involved different types of materials (i.e., figural, spatial, letter, or number). In the other session, four intellectual ability tests were administered. The order of the tests was as follows: the paper-and-pencil relational discovery test (related to another project and unreported here), the computerized figural analogy test, Raven's matrices, and the paper-and-pencil figural analogy test.

Relation-monitoring task The three-same, five-same, and three-different conditions of the letter and number versions of the task were applied to each participant in the same manner as in Experiment 1 (meaning that the number trials preceded the letter trials in each condition), with two exceptions: The numbers of number/letter trials were decreased to 40/40 trials in each condition of the task, and the three-same condition was

presented first (as a baseline condition), whereas the order of the two remaining conditions was random. This change was introduced in order to eliminate the effects of the order of conditions on the comparison of the Gf correlations yielded by each of the two latter conditions, while minimizing the method variance related to the relation-monitoring task as a whole. The dependent variable, calculated separately for each of the three conditions of the task, was the same as in Experiments 1 and 2.

Complex span tasks Adapted versions of the operation span, reading span, and symmetry span tasks (Conway et al. 2005) were applied. Each task required participants to memorize a sequence of three to seven (i.e., set size) stimuli. Each stimulus was presented for 1.2 s apiece, out of nine possible stimuli for that task. Each stimulus was followed by a simple decision task, presented until a response was given, but for the maximum of 9 s. After two two-stimulus training trials, three trials for each set size (in increasing order) were presented in each complex span task. The operation span task required the memorization of letters while deciding with a mouse button whether an intermittent simple arithmetical equation (e.g., "2 × 3 – 1 = 5?") was or was not correct. The modified reading span task consisted of memorizing digits while checking whether letter strings (e.g., "EWZTE," "KTANY") began and ended with the same letter. The spatial span task involved memorizing locations of a red square in a 3 × 3 matrix while deciding which of two presented bars was larger (the difference was always 25%). Also, a figural span task was applied, but due to the use of the same material—geometric figures—as in my Gf tests, its scores were not included in the present study (in order not to attribute the correlation between complex span and Gf to the shared modality).

The response procedure in each task consisted of the presentation of as many 3 × 3 matrices as was a particular set size, in the center of the computer screen, from left to right. Each matrix contained the same set of all nine possible stimuli for a given task. A participant was required to point with the mouse at those stimuli that had been presented in a sequence, in the correct order. Participants had no time limit for responding. Only a choice that matched both the identity and ordinal position of a given stimulus was taken as the correct answer. The dependent variable for each complex span task was the proportion of correctly pointed-out stimuli to all stimuli presented in the task.

STM tasks I used a modified change detection paradigm (Luck and Vogel 1997), in three tasks: the letter, number, and color versions. A figural version was also applied, but it was omitted for the same reason as the figural complex span task. Each of the 60 trials of the task (plus two training trials) consisted of a virtual 4 × 4 array filled with a few stimuli (i.e., only some cells in the array were filled). The stimuli were ten Greek symbols (e.g., α, β, χ, etc.), the digits 0 to 9, or squares in ten sufficiently distinctive colors. Each stimulus was approximately 2 × 2 cm in size. The number of stimuli within the array could be five, seven, or nine items. The array was presented for a period equal to the number of its items, multiplied by 200 ms, and then followed by a black square mask of the same size as the array, presented for 1.2 s. In a random 50% of the trials, the second array was identical to the first, whereas in the remaining trials, both arrays differed by exactly one item at one position, which was always a new item (not an item from another position). If the arrays differed, the new item was highlighted by a square red border. If they were identical, a random item was highlighted. The task was to press one of two response keys, depending on whether the highlighted item differed or not in the two arrays. The second array was shown until a response was given or until 4 s had elapsed. The trials were self-paced. The score on this task was the estimated sheer capacity of the STM buffer (the k value; Cowan 2001), calculated as the difference between the proportions of correct responses for arrays with one item changed and incorrect responses for unchanged arrays, multiplied by the set size. The total score was the mean k in the task.

N-back tasks The stimuli in the three 4-back tasks, adapted from Chuderski and Nęcka (2012), were 16 consonants, two-digit numbers, or figures (though the latter were the same materials used in the Gf tests, I retained that task in order to avoid calculating the n-back latent variable on the basis of only two n-back tasks), each approximately 2.5 × 2.5 cm in size, presented for 1.2 s plus a 0.6-s mask. A total of 80 stimuli were presented serially to the participants in each session. Two sessions were used in each task, preceded by the 40-stimulus training. Each session included eight 4-back target repetitions of stimuli and eight 1-back lure repetitions. No other stimuli could be repeated in a time window of ten stimuli. Participants were instructed to respond to 4-back repetitions and to suppress responses to all other repetitions as well as responses to nonrepeating items. The dependent variable for each task was the mean accuracy for target repetitions minus the mean false alarm rate for lures (Snodgrass and Corwin 1988).

Antisaccade tasks Each antisaccade task consisted of five training and 40 test trials. In order to increase the load on attention control, the tasks were slightly modified in comparison to the most commonly applied version (e.g., Unsworth et al. 2004): A participant's eyes could be directed to three locations (instead of one), and each test trial consisted of four events. First, a cue presented for 1.5 s informed that a target would be presented in the top, middle, or bottom of the side opposite to a flashing square (e.g., "Look at the bottom corner

opposite the flashing square," in Polish). Next, a fixation point was presented at the center of the screen for 1–2 s. Then, a rapidly flashing black square (3 cm in size) was shown in the middle of the left or right side of the screen, about 16 cm from the fixation point, for 0.15 s. Finally, depending on the task, a small dark gray arrow pointing left, down, or right (the spatial version), a digit 1, 2, or 3 (the number version), or a string "left," "down," or "right" (the letter version) was presented in the location opposite to the square for only 0.2 s and was then replaced by a mask. The visual angle from both the square and the arrow/digit/string to the fixation point was 14º. The task was to look away from the flashing square, to detect the direction of the arrow or the identity of the digit/string, and to press the key associated with the stimulus. The trials were self-paced. The dependent variable in each task was mean accuracy.

Raven's Advanced Progressive Matrices The test (Raven, Court, and Raven 1983) consists of 36 items that include a 3 × 3 matrix of figural patterns that is missing the bottom-right pattern, and eight response options, which are the patterns that can potentially match the missing one. The participant's task was to discover the rules that govern the distribution of patterns and to apply them to response options in order to choose the one and only right pattern. Sixty minutes were allowed for the test. The score was the total number of correctly answered items.

Figural analogy test This test (Orzechowski and Chuderski 2007) includes 36 figural analogies in the form "A is to B as C is to X," in which A, B, and C are types of relatively simple patterns of figures, A is related to B according to two, three, four, or five latent rules (e.g., symmetry, rotation, change in size, color, thickness, number of objects, etc.), and X is an empty space. The task is to choose one figure from a choice of four that relates to figure C as B relates to A. The administration time was 45 min. As with Raven's matrices, the total number of correct answers was taken as the score.

Computerized figural analogy test This test is a computerized and substantially modified version of the paper-and-pencil analogy test. The test includes 48 figural analogies in the form "A is to B as C is to X," in which A, B, and C are types of relatively complex patterns of figures, each including either five or eight figures (depending on the test item). In each item, A is related to B according to two to eight latent rules (rotation, change in location, color, thickness, filling, etc.), and X has to be selected by clicking with a mouse on one of seven alternative answer patterns. The one and only pattern that should be chosen is the one that relates to pattern C as B relates to A. After two training items, the participants were allowed up to 4 min to solve each test item. The total number of correct answers was taken as the score.

Results

Descriptive statistics and reliabilities for all tasks analyzed are presented in Table 1. All measures had proper distributions and acceptable reliability. It is worth noting that the accuracy in the five-same condition of the relation-monitoring task was significantly higher than that in the three-different condition, $t(242) = 21.63$, $p < .001$, Cohen's $d = 1.39$, replicating the results of Experiment 1 with the use of a much larger sample. This result again supported the predictions of the relational integration hypothesis, suggesting that the former condition yields higher accuracy because it requires integrating only two bindings, whereas the latter involves three bindings, although the former requires attending to five objects, whereas the latter requires attention to only three objects.

Table 2 includes the matrix of correlations among all variables. First, with the use of partial correlations, I tested whether either of two control variables, age and level of anxiety—the latter assessed with two administrations (applied in the middle of the first and second sessions) of the Polish adaptation of the State–Trait Anxiety Inventory questionnaire (Spielberger, Gorsuch, and Lushene 1970)—influences the pattern of correlations between WM tasks and Gf scores. It was found that neither variable significantly influenced any correlation coefficient between a WM task and a Gf test (the largest $\Delta r = .05$, n.s.). Second, virtually no differences were found in the Gf correlations between the three conditions of the relation integration task. The respective values, which were calculated for a factor loaded by three Gf tests (eigenvalue = 2.12) and the scores in the three-same, five-same, and three-different conditions, were $r = .479$, .430, and .471, respectively (no difference between any pair of these correlations was significant).

Next, I turned to latent-variable analysis. The fit of all of the models described below was evaluated by four indices of fit (for the justification of the criterion values, see Hu and Bentler 1999; Kline 1998): the ratio of the chi-squared (χ^2) statistic to the number of degrees of freedom (χ^2/df; should not surpass 2.0), Bentler's comparative fit index (CFI; should exceed .90), the root-mean square error of approximation (RMSEA; should not exceed .08) and its 90% confidence interval, and the standardized root-mean square residual (SRMR; should be less than .08). A measurement model (see Fig. 3)—which tested whether the latent variables representing particular classes of WM tasks (i.e., the relational integration, complex span, STM span, *n*-back, and antisaccade variables, each loaded by three scores) were easily discriminable—had a good fit, $N = 243$, $\chi^2(80) = 149.88$, $\chi^2/df = 1.87$, CFI = .962, RMSEA = .057 [.042–.072], SRMR = .041. All latent variables significantly correlated, but each variable was clearly distinguishable from all other variables, as setting any correlation coefficient to unity significantly decreased the the fit of the model, with the minimum decrease in the goodness of fit equaling $\Delta\chi^2(1) = 27.40$, $p < .001$.

Table 1 Descriptive statistics and reliabilities for the measures used in Experiment 3

Task	*M*	*SD*	Range	Skew	Kurtosis	Reliability
Relat. integr.: 3-same	0.75	0.17	0.00–1.00	–1.37	2.18	.85
Relat. integr.: 5-same	0.69	0.18	0.00–1.00	–0.97	1.54	.83
Relat. integr.: 3-different	0.41	0.23	–0.17–0.93	–0.33	–0.50	.82
Operation span	0.68	0.18	0.05–0.99	–0.72	0.20	.87
Reading span	0.78	0.16	0.09–1.00	–0.76	2.20	.86
Spatial span	0.52	0.18	0.05–0.97	–1.30	0.03	.85
Letter STM	2.44	1.42	–1.33–6.40	0.17	–0.17	.73
Number STM	4.63	1.43	–0.63–7.00	–1.07	1.05	.83
Color STM	2.84	1.41	–0.87–6.00	–0.34	–0.39	.71
Letter 4-back	0.24	0.40	–0.93–1.00	–0.39	–0.55	.91
Number 4-back	0.19	0.28	–0.81–0.80	–0.94	1.30	.84
Figural 4-back	0.18	0.35	–0.87–0.94	–0.52	–0.19	.89
Letter antisaccade	0.58	0.26	0.00–1.00	–0.73	–0.67	.94
Number antisaccade	0.48	0.23	0.00–1.00	0.04	–0.81	.90
Arrow antisaccade	0.59	0.24	0.05–1.00	–0.26	–1.02	.92
Raven	21.77	6.79	3–35	–0.48	–0.14	.88
Paper analogies	22.12	6.69	6–35	–0.24	–0.80	.86
Computerized analogies	21.81	11.83	0–48	0.40	–0.84	.93

$N = 243$ for all tasks. Relat. integr. = relational integration task. STM = short-term memory task. Reliability = Cronbach's alpha.

The correlations between the calculated WM latent variables and the fluid-reasoning variable, the latter loaded by three ability tests, are presented in Table 3. I used structural equation modeling to predict the latter variable by all of the former variables (see Fig. 4), $N = 243$, $\chi^2(120) = 201.94$, $\chi^2/df = 1.68$, CFI = .963, RMSEA = .051 [.038–.064], SRMR = .041. The only two significant predictors were the relational integration and complex span variables, and the former yielded a numerically stronger regression path than did the latter, $\Delta r = .118$—though that difference was not significant, as

Table 2 Correlation matrix of measures used in Experiment 3

Task	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1. Relat. integr.: 3-same	–																
2. Relat. integr.: 5-same	.53	–															
3. Relat. integr.: 3-different	.54	.52	–														
4. Operation span	.47	.39	.46	–													
5. Reading span	.40	.35	.38	.70	–												
6. Spatial span	.41	.35	.31	.57	.51	–											
7. Letter STM	.38	.21	.25	.38	.34	.35	–										
8. Number STM	.43	.25	.31	.38	.45	.40	.42	–									
9. Color STM	.47	.39	.33	.39	.33	.44	.40	.43	–								
10. Letter 4-back	.32	.34	.32	.39	.35	.36	.26	.24	.37	–							
11. Number 4-back	.21	.28	.27	.22	.21	.12[a]	.28	.17	.26	.50	–						
12. Figural 4-back	.33	.35	.35	.31	.26	.34	.22	.23	.32	.70	.55	–					
13. Letter antisaccade	.49	.43	.43	.44	.51	.45	.35	.38	.41	.40	.24	.36	–				
14. Number antisaccade	.38	.39	.43	.43	.35	.40	.37	.32	.36	.30	.17	.29	.70	–			
15. Arrow antisaccade	.47	.40	.40	.50	.45	.52	.37	.36	.39	.42	.19	.36	.82	.81	–		
16. Raven	.50	.42	.43	.45	.40	.47	.32	.40	.46	.29	.22	.32	.42	.46	.46	–	
17. Paper analogies	.40	.34	.36	.47	.37	.43	.24	.25	.35	.21	.11[b]	.23	.37	.38	.36	.67	–
18. Computerized analogies	.32	.33	.41	.41	.31	.38	.22	.18	.29	.19	.17	.23	.31	.30	.32	.50	.51

$N = 243$. Relat. integr. = relational integration task. STM = short-term memory task. All ps < .05, except [a] $p = .062$ and [b] $p = .085$.

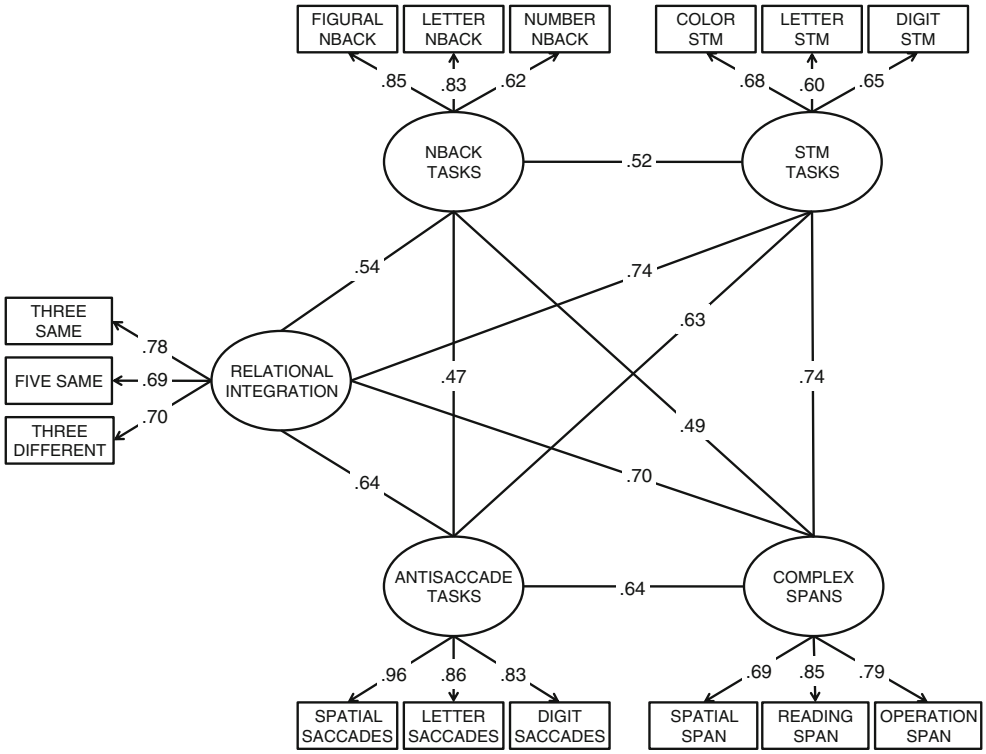

Fig. 3 The measurement model, including latent variables (ovals) representing variance shared by triples of the relational integration tasks, complex span tasks, short-term memory (STM) tasks, *n*-back tasks, and antisaccade tasks. Boxes represent manifest variables (particular tasks). Values between the ovals and boxes represent relevant standardized factor loadings (all *p*s < .001). Values between the ovals represent relevant path coefficients among the latent variables (all *p*s < .001)

was indicated by the model that constrained both path coefficients to be the same, $\Delta\chi^2(1) = 0.38$, $p = .462$.

However, due to the high amount of multicollinearity among the two significant predictors, the comparison above was not the best method to assess these predictors' relative strengths. So, I calculated two models with only one predictor (either the relational integration or complex span variable), two predicted variables (fluid reasoning and the other WM variable), and with the disturbance terms of the predicted variables left free to correlate. Such an analysis could answer the question of whether the relational integration variable can account for additional variance in fluid reasoning (i.e.,

Table 3 Correlations between latent variables representing classes of WM tasks and the fluid reasoning variable, in Experiment 3

WM latent variables	*r*
Relational integration tasks	.707
Complex span tasks	.669
Short-term memory tasks	.656
N-back tasks	.390
Antisaccade tasks	.549

N = 243.

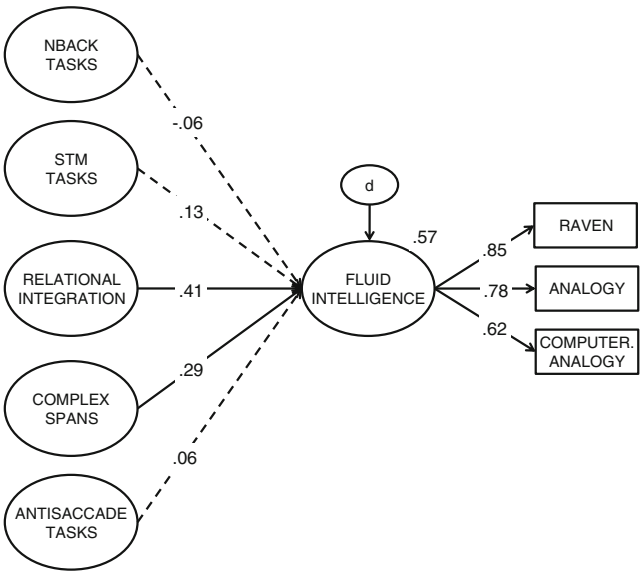

Fig. 4 Structural equation model in which the fluid intelligence endogenous latent variable, loaded by scores on three intelligence tests, is predicted by five exogenous latent variables representing relational integration, complex span, short term memory (STM), *n*-back, and antisaccade tasks. The small oval represents a disturbance term. All standardized factor loadings were significant (*p*s < .001). Solid lines between the ovals represent *p*s < .001, whereas dashed lines depict *p*s > .07

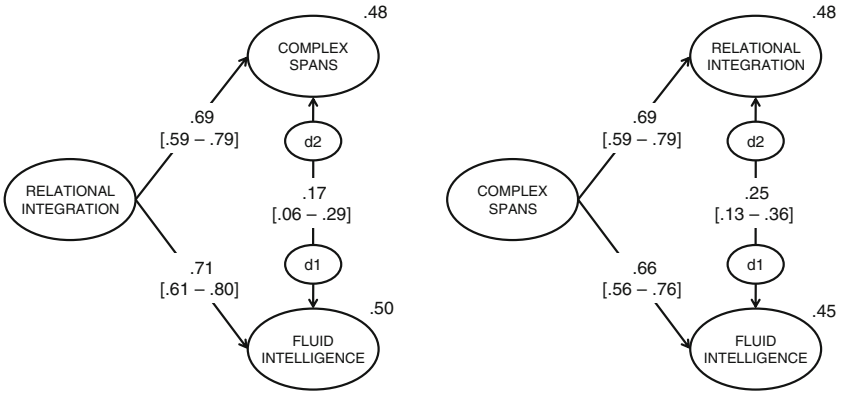

Fig. 5 Left panel: Structural equation model in which the fluid intelligence and complex span endogenous latent variables are predicted by the relational integration latent variable. All standardized factor loadings were significant, as were path coefficients among the latent variables (ps < .001). Right panel: Analogous structural equation model in which the fluid intelligence and relation integration variables are predicted by the complex span variable

whether the correlation between respective disturbances was significant) above and beyond the variance shared by complex spans and Gf. Both models had a very good fit, N = 243, $\chi^2(24) = 34.61$, $\chi^2/df = 1.44$, CFI = .988, RMSEA = .043 [.0–.072], SRMR = .036, and are presented in Fig. 5. The relational integration explained twice as much Gf variance (6.0%) that was not explained by the complex span variable as the complex span variable explained (3.0%), with regard to the Gf variance left unexplained by relational integration. However, both amounts of unexplained variance were relatively small, and the difference between them was not significant.

When the complex spans were substituted with STM tasks in the models from Fig. 5, the disturbances of the STM span variable and the Gf variable correlated only marginally (r = .129, p = .041), whereas the correlation between the disturbances of the relational integration and Gf variables was similar (r = .207, p = .002) to when they were predicted by the complex span variable (the total amount of Gf variance explained by the model equaled R^2 = .49). When scores on either the antisaccade or n-back tasks were used instead of complex spans, the correlation between their latent variables and Gf disturbances was no longer significant (the amount of Gf variance explained was R^2 = .50).

In another analysis, I calculated a broad WM storage factor, composed of all of the STM and complex span measures, and allowed it also to load on the relation integration tasks. Furthermore, I allowed the attention control factor (calculated from the antisaccade tasks) to load on the latter tasks. Such an analysis aimed at testing whether relation integration contributed to Gf because of its overlap with storage and/or control aspects of WM, or whether it constituted an independent contribution after both of these elements had been controlled for (for an analogous model including the storage, mental speed, and relation integration [coordination] variables, see Krumm et al. 2009, Fig. 3). I excluded the n-back tasks from that analysis, because their interpretation as storage versus control tests was less clear, as their scores need to be based on both hit and lure trials (in order to account for a decision bias that affects hit vs. lure performance in that task). The model, depicted in Fig. 6, N = 243, $\chi^2(80) = 185.08$, $\chi^2/df = 2.31$, CFI = .943, RMSEA = .075 [.061–.089], SRMR = .044, showed that the relation integration task significantly predicted Gf (r = .237, p = .001), even if the variance common to tasks meant to tap storage as well as to indices intended to capture control was partialed out. It is worth noting that relation integration tasks' loads on the control latent variable were weak (two nonsignificant and one marginal), suggesting, in line with the results of Experiment 2, that these tasks involve little attention control.

Finally, using the family of nested structural equation models, I estimated the amounts of variance explained separately by the relational integration (RI), complex span (CS), antisaccade (AS), and STM variables, as well as the amounts predicted jointly by two (e.g., RI + CS), three (e.g., RI + CS + STM), and all four (i.e., RI + CS + STM + AS) of the variables (see Table 4). Then, by means of the technique of variance partitioning (see Chuah and Maybery 1999), for each variable I estimated the contribution to Gf variance that was unique to that variable or that was shared by it with other variables (for the sake of clarity, and due to its minimal contribution to Gf, I excluded the n-back variable from that analysis). For example, the part of variance unique to the relational integration variable was computed by subtracting the amount of Gf variance predicted by all four variables minus the amount of Gf variance predicted by the complex span, antisaccade, and STM variables. The variance shared by the four variables was the remaining Gf variance after both unique variances and the variances shared by pairs and triples of variables were subtracted from the amount of Gf variance predicted by the four variables.

The results of variance partitioning are presented in Fig. 7 (note that only amounts of shared variances that were

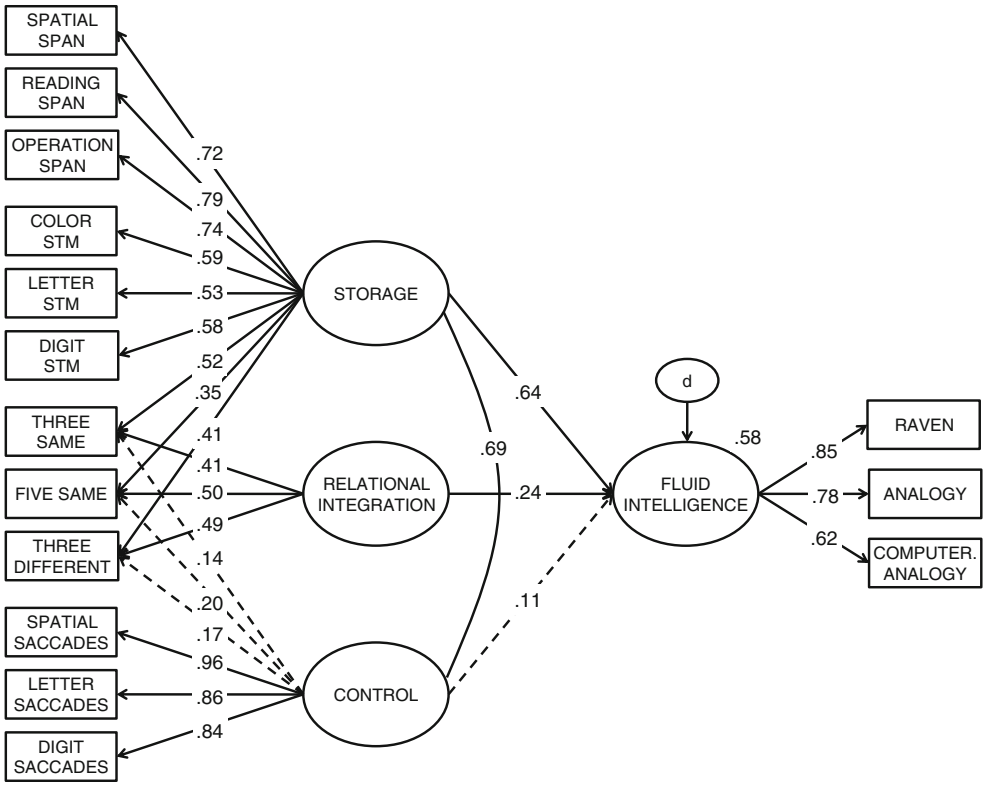

Fig. 6 Structural equation model in which the fluid intelligence endogenous latent variable, loaded by scores on three intelligence tests, is predicted by three exogenous latent variables representing relational integration, storage (complex span and STM), and control (antisaccade) tasks. Relational integration tasks were allowed to load on both the storage and control latent variables. The small oval represents a disturbance term. All standardized factor loadings were significant (ps < .01), apart from the loadings of the relational integration tasks on the control variable. Solid lines between ovals represent ps < .01, whereas the dashed line depicts p > .10

Table 4 Amounts of Gf variance explained by all possible combinations of predictors used in Experiment 3

Predictors	R^2
RI, CS, STM, AS	.565
RI, CS, STM	.563
RI, CS, AS	.561
RI, STM, AS	.536
CS, STM, AS	.506
RI, CS	.556
RI, STM	.529
RI, AS	.514
CS, STM	.495
CS, AS	.473
STM, AS	.459
RI	.498
CS	.442
STM	.429
AS	.300

N = 243. RI = relational integration tasks. CS = complex span tasks. STM = short-term memory tasks. AS = antisaccade tasks.

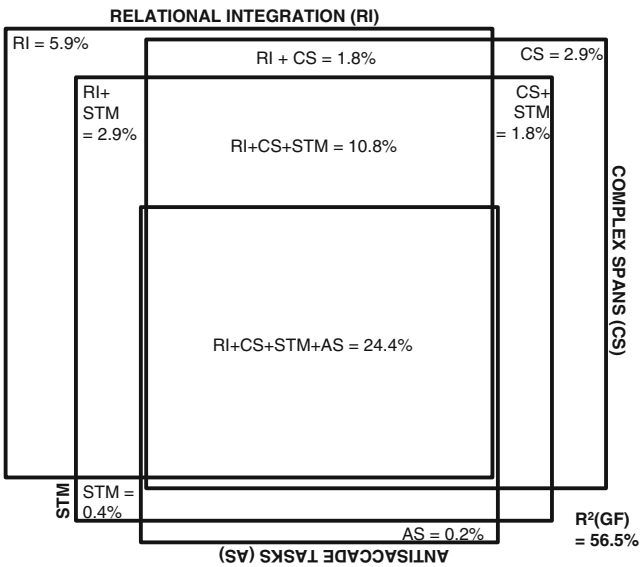

Fig. 7 Diagram indicating the amounts of the shared and unique variance in fluid intelligence accounted for by the relation integration (RI), complex span (CS), short-term memory (STM), and antisaccade (AS) tasks. The numbers are based on the regressions from Table 4

significant—i.e., 1.8% or larger—are included in this figure). Out of the 56.5% of Gf variance accounted for by all predictors, STM (0.4%) and AS (0.2%) uniquely contributed negligible amounts of Gf variance, whereas CS (2.9%) and RI (5.9%) uniquely explained larger amounts of Gf variance. The largest part of the explained Gf variance (24.4%) was jointly accounted for by all four predictors, whereas another 10.8% was shared among RI, CS, and STM. Small but significant amounts of variance were shared between RI and STM (2.9%), RI and CS (1.8%), and CS and STM (1.8%), whereas all other amounts were not significant (0.3%–1.6%).

Discussion

The present experiment indicated that the relational integration task is a plausible measure of WMC, which shares much of its variance with two other standard WMC measures—the STM and complex span tasks. Also, the study showed that the relational integration task strongly predicts fluid reasoning, and its different versions can account for comparable amounts of Gf variance. The relation integration task appeared to be a numerically better predictor of fluid reasoning than was the complex span task; the former task tapped a few percent of variance that could not be accounted for by the latter task. The complex span task is believed to primarily reflect the attentional control component of WM, so the relation integration task seems to predict fluid reasoning above and beyond what can be predicted by referring to attentional control. This conclusion is supported by the fact that both the *n*-back task including lures and the antisaccade task, which both probably tap attention control, were poor predictors of Gf, when compared to the relation-monitoring task.

Furthermore, the relation integration task also accounted for significant additional Gf variance when variance related to the storage aspect of WM tasks, which was aimed to reflect the individual scope of attention, was controlled for. Thus, it seems that the relational integration task taps processes that are important for fluid reasoning, but that constitute something that is beyond the sheer ability to maintain as much information as possible within the scope of attention.

General discussion

To sum up the results of Experiments 1–3, in Experiments 1 and 3 the relational integration task was shown to primarily measure the process of constructing, maintaining, and integrating the flexible, temporary bindings in WM, as the performance in this task was strongly affected by the postulated number of bindings that had to be maintained and/or integrated (but not the number of objects possibly maintained in the scope of attention nor by the amount of interference present during visual scanning for objects). Experiment3 supported the view that the measure of relational integration is a plausible WM task, as it shares a large amount of variance with other well-established measures of WMC, and it is a better predictor of fluid reasoning than are any of these measures. Such a high predictive power was achieved even though the relation integration task neither requires any memorizing of stimuli nor involves significant executive control.

The presented study may help in finding an answer to the following question: What do the well-established WM tasks really measure? It is noteworthy that the complex span as well as indices of both the scope (i.e., the STM task) and the control (i.e., the antisaccade task) of attention uniquely or jointly accounted for only 6.7% of the variance in fluid reasoning, whereas the remaining 43.9% of variance explained by these measures was shared with relational integration. However, in light of the results of Experiments 1 and 2, it is unlikely that the 43.9% of variance that was shared by the relation-monitoring task, standard WM task, and reasoning tests was shared due to common requirements of these tasks that would pertain to either the scope or control of attention (or both), because it was demonstrated that performance in the relation-monitoring task depended on factors loading neither the scope nor the control of attention. Thus, it does not seem that accuracy in that task reflects the effectiveness of attentional mechanisms to a large extent.

On the contrary, the present results seem to be consistent with the view (Oberauer et al. 2007,2008) proposing that scores on standard WM tasks, like STM tasks or complex spans, may at least partly reflect the need to construct, maintain, and/or integrate the temporary bindings between memorized stimuli and some other contextual information. Taking as an example the complex span task, it obviously requires binding of the current stimulus to its serial position, and maintaining such a binding for all items and the whole duration of the trial, since a properly memorized stimulus that is recalled at an incorrect position will constitute an error. However, such requirements for processing bindings may be even more important. For instance, during recall, participants may have to bind the already-recalled items to the tag noting that they had been recalled, in order to inhibit recalling them again (Farrell and Lewandowsky 2012). Similarly, still unrecalled items may have to be bound to the tag indicating that they are waiting for recall. Moreover, in the STM tasks, stimuli in the memory set may not only be linked to context, but also bound together, as suggested by the interactive effects in visual STM tasks (i.e., tasks that were also used in the present study). These effects, strongly implied by recent research on the structured nature of both WM and long-term memory representations (for a review, see Brady et al. 2011), are caused by the fact that along with the individual items, participants also encode in WM the whole structure describing these items (e.g., visual layout, including perceptual grouping, as well as ensemble statistics for items).

So, the presented results, as well as some influential computational models of WM tasks (e.g., Murre, Wolters, and Raffone 2006; O'Reilly et al. 2003), together suggest that WM tasks, like complex spans or change detection tasks, may to some extent rely on constructing and integrating proper bindings in WM, and not only on the sheer maintenance of particular items, nor on the control of access to them.

Of course, the present results in no way allow the conclusion that relational integration is the only process that is captured by the well-established WM tasks. In fact, the extant research on WM surely tells us that active maintenance (Unsworth and Engle 2007a), attentional control over interference (Unsworth and Spillers 2010; Unsworth and Engle 2007a), and control over retrievals from secondary memory (Mogle, Lovett, Stawski, and Sliwinski 2008) are all important components of WM that substantially assist with coping in situations requiring memorizing and transformation of perceptually unavailable stimuli. However, though performance in WM may rely on multiple mechanisms and processes (see Conway, Getz, Macnamara, and Engel de Abreu 2011), the link between relatively simple WM tasks and much more complex abstract-reasoning tests may be primarily driven by the relational integration component of WM. Although this link cannot be driven by relational integration exclusively, as in the present study the remaining WM tasks were able to independently account for an additional 6.7% variance in Gf, and almost half of the Gf variance was unexplained by WM, still, relational integration may determine the most important part of that link.

First, the scope of attention (i.e., STM) tasks was not able to uniquely account for any significant variance in Gf when the alternative measures of WM were controlled. Its potential contribution to fluid reasoning was almost entirely captured by the relation-monitoring tasks, which—as was demonstrated by Experiments 1 and 2—do not seem to rely on any maintenance of information within the scope of attention. Moreover, in light of recent empirical evidence (see Brady et al. 2011) and the results of computational simulations (Chuderski, Andrelczyk, and Smoleń 2013; Raffone and Wolters 2001), it is unlikely that scores on these tasks reflect only the number of items that can be maintained within attention, and not the relational structure describing the corresponding pattern governing the stimuli, which consists of multiple bindings between those items.

Second, though the complex span tasks are often interpreted as primarily tapping executive attention ability (Kane et al. 2007a), the alternative interpretation that assumes that they are in a substantial part measures of relational integration is fully consistent with the present data. For example, one argument for interpreting complex span in terms of executive attention ability is based on the fact that it correlates with attention control tests (e.g., Unsworth et al. 2004; Unsworth et al. 2009). However, in the present study the latent-level correlation between complex span tasks and antisaccade tasks (r = .64) was weaker than the correlation between complex span tasks and the relational integration variables (r = .70).

One potential limitation of the present study concerns the identification of exactly what low-level cognitive mechanism(s) involved in processing relational structures drive(s) the link between scores on relational integration tasks and reasoning. Is the integration of arbitrary bindings the crucial process, or is maybe the very construction and maintenance of (as many as possible) temporary bindings the key factor (so that the requirement to integrate them is less important)? In the latter sense (as was noticed by a reviewer), the relation integration task may in fact primarily rely on abstract pattern detection/recognition—specifically, on comparing the patterns of stimuli to one of several predefined patterns held in mind. However, it seems unlikely that the performance in the relation integration task can be reduced to sheer detection of abstract patterns. Although the possession of predefined patterns (e.g., rows "XX1 XX1 XX1," "XX2 XX2 XX2," etc.) by participants may be at least imaginable in the three-same condition, in the different-object conditions so many possible patterns would fulfill the target relation that their encoding and simultaneous testing by participants seems psychologically implausible. The selection, binding, and (most probably) integration of the most promising information "on the fly" may be the most effective way to cope with the relation integration task. Nevertheless, the task used in this study did not allow for orthogonal manipulation of the need for binding construction versus the need for binding integration, and a question about which of them was primarily responsible for the task's strong relationship with fluid intelligence cannot be decisively answered solely on the basis of this study.

However, one argument for the hypothesis that binding construction does not suffice for that relationship, and that the integration of bindings may be necessary, is the fact that the simple comparison of stimuli (e.g., processing binary relations, in contrast to processing more complex relations) neither constitutes a cognitive load strong enough for young healthy adults nor yields substantial interindividual variation (Halford, Baker, McCredden, and Bain 2005; Viskontas, Holyoak, and Knowlton 2005; Waltz et al. 1999). Although a recent study by Wilhelm, Hildebrandt, and Oberauer (2013) provided evidence that a task that requires primarily the memorizing of bindings (i.e., encoding associations between pairs of stimuli from two domains, and then identifying a paired associate, given the respective stimulus) is an excellent WM measure and strong Gf predictor, it required memorizing several (2–6) bindings in parallel, and thus it might have involved the integration of all presented associations into a larger mental structure, which surely helped participants to deal with that task. It seems that answering a question about which aspect of relational integration tasks (e.g., either binding construction/maintenance or binding integration, or both)

is primarily responsible for their excellent characteristics as WM measures and Gf predictors requires future research (as well as novel tasks).

Nevertheless, though the identification of the cognitive mechanisms underlying the strong link between WMC and fluid reasoning is beyond the explanatory power of one study, the present work replicated and greatly extended results (Oberauer et al. 2000; Oberauer et al. 2007; Oberauer et al. 2008) suggesting that such mechanisms may rely on the valid construction and active maintenance of temporary bindings between representations crucial for the task at hand, or on their proper integration into more complex relational structures, or some mixture of both of these processes. Such a binding/integration may allow for coping with abstract, novel, and arbitrary situations, of which WM tasks and Gf tests are the most prominent examples. As such, the present work seems to constitute important progress in understanding the key constraints of WM, as well as in explaining the nature of human fluid intelligence.

In the latter regard, the study provided new evidence that clearly supports theories (e.g., Halford et al. 1998; Hummel and Holyoak 2003; Oberauer et al. 2007; Waltz et al. 1999) that have proposed that human intelligence may reflect the domain-general ability to construct higher-level relational structures that bind a certain number of more atomic representations (e.g., perceptual or memorial), are extremely flexible, and can be effectively abstracted from any intrinsic features of the low-level representations. This line of research—explaining intelligence as the ability to conduct role-based relational reasoning based on the processing of relational roles explicitly and separately from (perceptual or semantic) features of entities that fill these roles, and that involves coding the bindings of entities to their specific roles—is in a way a revival of Spearman's (1927) classical idea of the eduction of relations. This relational-reasoning account has recently gained substantial attention within psychology, and seems to be a very fruitful framework for future studies on fluid reasoning.

Author note This research was supported by Grant No. N106 417140, funded by the National Science Centre of Poland. Thanks to Krzysztof Cipora, Dominika Czajak, Maciej Taraday, and Jolanta Wójcik for their help in conducting the study, and to Mike Timberlake for correcting the text.

References

Brady, T.F., Konkle, T., Alvarez, G.A. (2011). A review of visual memory capacity: Beyond individual items and toward structured representations. *Journal of Vision*, *11*(5):4, 1–34. doi:10.1167/11.5.4

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

Buehner, M., Krumm, S., Ziegler, M., & Pluecken, T. (2006). Cognitive abilities and their interplay: Reasoning, crystallized intelligence, working memory components, and sustained attention. *Journal of Individual Differences, 27,* 57–72.

Burgess, G. C., Gray, J. R., Conway, A. R. A., & Braver, T. S. (2011). Neural mechanisms of interference control underlie the relationship between fluid intelligence and memory span. *Journal of Experimental Psychology. General, 140,* 674–692.

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.

Chuderski, A., Andrelczyk, K., & Smoleń, T. (2013). An oscillatory model of individual differences in working memory capacity and relational integration. *Cognitive Systems Research, 24,* 87–95.

Chuderski, A., & Nęcka, E. (2012). The contribution of working memory to fluid reasoning: Capacity, control, or both? *Journal of Experimental Psychology. Learning, Memory, and Cognition, 38,* 1689–1710. doi:10.1037/a0028465

Chuderski, A., Taraday, M., Nęcka, E., & Smoleń, T. (2012). Storage capacity explains fluid intelligence while executive control does not. *Intelligence, 40,* 278–295.

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. doi:10.1016/j.intell.2008.01.002

Conway, A. R. A., Getz, S. J., Macnamara, B., & Engel de Abreu, P. M. J. (2011). Working memory and fluid intelligence: A multi-mechanism view. In R. J. Sternberg & S. B. Kaufman (Eds.), *The Cambridge handbook of intelligence* (pp. 394–418). Cambridge, UK: Cambridge 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. doi:10.3758/BF03196772

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24,* 87–114. doi:10.1017/S0140525X01003922

Duncan, J., & Humphreys, G. W. (1989). Visual search and stimulus similarity. *Psychological Review, 96,* 433–458. doi:10.1037/0033-295X.96.3.433

Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. A. (1999). Working memory, short-term memory, and general fluid intelligence: A latent-variable approach. *Journal of Experimental Psychology. General, 128,* 309–331. doi:10.1037/0096-3445.128.3.309

Farrell, S., & Lewandowsky, S. (2012). Response suppression contributes to recency in serial recall. *Memory & Cognition, 40,* 1070–1080. doi:10.3758/s13421-012-0212-6

Halford, G. S., Baker, R., McCredden, J. E., & Bain, J. D. (2005). How many variables can humans process? *Psychological Science, 16,* 70–76.

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.

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

Hummel, J. E., & Holyoak, K. J. (2003). A symbolic–connectionist theory of relational inference and generalization. *Psychological Review, 110,* 220–264. doi:10.1037/0033-295X.110.2.220

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

Kane, M. J., Conway, A. R. A., Miura, T. K., & Colflesh, G. J. H. (2007b). Working memory, attention control, and the *n*-back task: A question of construct validity. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 33,* 615–622. doi:10.1037/0278-7393.33.3.615

Kline, R. B. (1998). *Principles and practice of structural equation modeling*. New York, NY: Guilford Press.

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.

Logan, G. D., & Bundesen, C. (2003). Clever homunculus: Is there an endogenous act of control in the explicit task-cuing procedure? *Journal of Experimental Psychology. Human Perception and Performance, 29,* 575–599. doi:10.1037/0096-1523.29.3.575

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature, 390,* 279–281. doi:10.1038/36846

McGrew, K. S. (2009). CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. *Intelligence, 37,* 1–10.

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.

Murre, J. M. J., Wolters, G., & Raffone, A. (2006). Binding in working memory and long-term memory: Towards an integrated model. In H. D. Zimmer, A. Mecklinger, & U. Lindenberger (Eds.), *Binding in human memory: A neurocognitive perspective* (pp. 221–250). Oxford, UK: Oxford University Press.

O'Reilly, R. C., Busby, R., & Soto, R. (2003). Three forms of binding and their neural substrates: Alternatives to neural synchrony. In A. Cleeremans (Ed.), *The unity of consciousness*. Oxford, UK: Oxford University Press.

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., & Sander, N. (2007). Individual differences in working memory capacity and reasoning ability. In A. R. A. Conway, C. Jarrold, M. J. Kane, A. Miyake, & J. N. Towse (Eds.), *Variation in working memory* (pp. 49–75). Oxford, UK: Oxford University Press.

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

Orzechowski, J., & Chuderski, A. (2007). *Test analogii obrazkowych* [*A figural analogies test*]. Unpublished manuscript, Jagiellonian University, Krakow, Poland

Owen, A. M., McMillan, K. M., Laird, A. R., & Bullmore, E. (2005). *N*-back working memory paradigm: A meta-analysis of normative functional neuroimaging studies. *Human Brain Mapping, 25,* 46–59. doi:10.1002/hbm.20131

Raffone, A., & Wolters, G. (2001). A cortical mechanism for binding in visual memory. *Journal of Cognitive Neuroscience, 13,* 766–785.

Raven, J. C., Court, J. H., & Raven, J. (1983). *Manual for Raven's Progressive Matrices and Vocabulary Scales*. London, UK: H. K. Lewis.

Schmiedek, F., Hildebrandt, A., Lövdén, M., Wilhelm, O., & Lindenberger, U. (2009). Complex span versus updating tasks of working memory: The gap is not that big. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 35,* 1089–1096. doi:10.1037/a0015730

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology. General, 117,* 34–50. doi:10.1037/0096-3445.117.1.34

Snow, R. E., Kyllonen, P. C., & Marshalek, B. (1984). The topography of ability and learning correlations. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence* (Vol. 2, pp. 47–103). Hillsdale, NJ: Erlbaum.

Spearman, C. (1927). *The abilities of man: Their nature and measurement*. London: Macmillan.

Spielberger, C. D., Gorsuch, R. L., & Lushene, R. E. (1970). *Manual for the State–Trait Anxiety Inventory*. Palo Alto, CA: Consulting Psychologists Press.

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.

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. doi:10.1037/0033-295X.114.1.104

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. doi:10.1037/0033-2909.133.6.1038

Unsworth, N., Redick, T. S., Lakey, C. E., & Young, D. L. (2010). Lapses in sustained attention and their relation to executive control and fluid abilities: An individual differences investigation. *Intelligence, 38,* 111–122.

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, & Cognition, 30,* 1302–1321. doi:10.1037/0278-7393.30.6.1302

Unsworth, N., & Spillers, G. J. (2010). Working memory capacity: Attention, memory, or both? A direct test of the dual-component model. *Journal of Memory and Language, 62,* 392–406.

Unsworth, N., Spillers, G. J., & Brewer, G. (2009). Examining the relations among working memory capacity, attention control, and fluid intelligence from a dual-component framework. *Psychology Science Quarterly, 51,* 388–402.

Viskontas, I. V., Holyoak, K. J., & Knowlton, B. J. (2005). Relational integration in older adults. *Thinking & Reasoning, 11,* 390–410.

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

Wilhelm, O., Hildebrandt, A., & Oberauer, K. (2013). What is working memory, and how can we measure it? *Frontiers in Psychology, 4,* 433.