Pergamon

Learning and Individual Differences
13 (2002) 37–55

Learning and Individual Differences

On the correlation between working memory capacity and performance on intelligence tests

Tom Verguts*, Paul De Boeck

K.U. Leuven, Louvain, Belgium

Received 17 December 2001; received in revised form 5 February 2002; accepted 19 February 2002

Abstract

A ubiquitous finding in intelligence research is that there is a substantial correlation between working memory (WM) capacity and general (fluid) intelligence tests (e.g., [Intelligence 14 (1990) 389]). The standard explanation for this correlation is as follows: People with high WM capacity can keep in memory many elements and are therefore good at storing subresults needed within an item. We argue that another factor may be partly responsible for this correlation, namely, that people with a high WM capacity can store many solution principles *over items*. Two experiments (with N=42 and N=52, respectively) are conducted that validate this alternative explanation in two particular tests, the Raven Advanced Progressive Matrices Test (RPM) [Raven, J. C. (1965). *Advanced progressive matrices, set II*. New York: Psychological Corporation], and a number series test constructed by ourselves, but resembling standard number series intelligence tests (e.g., [J Educ Psychol 75 (1983) 603]).

1. Introduction

The role of working memory (WM) capacity is an important issue in intelligence research. Many authors (e.g., Babcock, 1994; Kyllonen & Christal, 1990; Larson & Saccuzzo, 1989) obtained moderately high correlations between a number of intelligence tests and WM capacity. A recent overview of this research from a developmental point of view is provided in Fry and Hale (2000).

* Corresponding author. Department of Experimental Psychology, Ghent University, H. Dunantlaan 2, B-9000 Ghent, Belgium.

E-mail address: tom.verguts@rug.ac.be (T. Verguts).

1041-6080/02/$ – see front matter
PII: S1041-6080(02)00049-3

One surprising finding is that short-term memory (STM) is separate from WM. A typical STM task consists of trying to remember a set of elements (e.g., words, numbers). On the other hand, a WM task consists of remembering elements while some other task (e.g., answering questions or solving simple arithmetic) is being performed. It turns out that (a) STM and WM are weakly correlated; (b) STM and intelligence are weakly correlated, whereas WM and intelligence are more strongly correlated; and (c) if the variance common to STM and WM is partialled out, STM does not correlate with intelligence, but WM does (e.g., Engle, Kane, & Tuholski, 1999). It is concluded by these authors that STM is a component of WM.

A next question is then why WM and intelligence tests are related. It seems obvious that WM capacity helps because it allows storage of information. Indeed, Conway and Engle (1996) and Engle, Cantor, and Carullo (1992) have shown that the reason a typical WM task (e.g., Daneman & Carpenter's, 1980 test: see below) is correlated with intelligence tests, is not because people with high WM are efficient in the processing component of the WM task (and hence have time left to rehearse the words to be remembered). If this processing efficiency is either statistically or experimentally controlled for, the correlation between WM capacity and intelligence test performance was still substantial.

However, what kind of information storage causes the correlation between WM capacity and intelligence test performance? This is not entirely clear and no direct tests of this problem seem to be reported in the literature. To quote Fry and Hale (2000):

> [T]he general issue of what types of information need to [be] kept accessible while engaged in what types of reasoning remains an important topic for future research. (p. 24)

Still, some authors have made suggestions along these lines, and the following explicit proposal was given by Carpenter, Just, and Shell (1990) in the context of the Raven Advanced Progressive Matrices Test (RPM; Raven, 1965) (see also Embretson, 1995; Wickelgren, 1997). In every item of the test, a number of (different) (sub)results have to be found and applied. This implies that a few subresults have to be stored during the period that the item is being solved. Persons with a large WM capacity can store more partial results and, hence, will have a higher probability of solving an item (see Carpenter et al.'s, 1990 "goal management" factor). Therefore, WM capacity and RPM performance are positively correlated. A similar claim was made by Just and Carpenter (1992) in the context of sentence comprehension: People with high WM capacity will be good in storing the results of parsings of early parts of a sentence, allowing them to obtain a correct interpretation when the end of the sentence is reached. Many authors seem to either explicitly or implicitly adhere to a view similar to this one. For example, O'Reilly, Braver, and Cohen (1999) write:

> Typically, complex tasks involve the temporally extended coordination of multiple steps of processing, often in novel combinations and situations, and the storage of intermediate products of computation, subgoals, and so on. Active memory together with the controlled encoding and retrieval of [hippocampal] memories can be used to retain the intermediate results of these processing steps for subsequent use. (p. 402)

A similar claim is made by Shah and Miyake (1996):

> Comprehending a complex sentence, mentally rotating an unfamiliar geometric figure, and solving a difficult reasoning problem, for example, all critically hinge on the person's ability to store various intermediate products of a computation while simultaneously processing new information. (p. 4)

These two quotes are consistent with the idea that WM capacity is needed to store and manipulate temporary subresults while solving an item of an intelligence test.

We do not wish to call this idea into question, but we suspect that there may be more to WM capacity than just storing partial results per item. Specifically, we propose that people with a high WM capacity will be better in storing rules or solution principles *over items*. We think of WM as a pool of resources that can be assigned to different elements in memory, thereby making them active and available for processing. Over time, the activation "leaks" and the memory elements become inactivated (see Cantor and Engle, 1993; Just and Carpenter, 1992 for similar views). If a high amount of WM is available, decay of available, earlier used elements will be slow (or slowly reach a fixed point), so once a useful rule or solution principle is found, people with high WM capacity will have a higher probability of retaining the (correct) rule on subsequent items and, hence, will have a higher probability of solving the item.

So, a second reason (apart from Carpenter et al.'s, 1990) why WM and RPM are correlated, we hypothesize, is that rules (possible solution principles) are retained over items. This prediction is in line with the result of Ferrara, Brown, and Campione (1986), who found that the number of hints a child needs in order to solve a set of related items is strongly related to her IQ score (on a different test). Another relevant result is that of Carlstedt, Gustafsson, and Ullstadius (2000). These authors administered an intelligence test with three types of items: Groups (find the odd-one-out in a sequence of five figures), Series (complete a sequence of four figures), and Bongard (find a feature connecting a set of figures; see Carlstedt et al., 2000 for more details). There were two experimental conditions. In the first condition, all similar items (of the same format) were presented consecutively (the homogeneous condition); in the second condition, items of different formats were alternated (the heterogeneous condition). Otherwise, the two tests contained exactly the same items. What they found, to their surprise, was that the homogeneous test had higher loadings on a general intelligence factor (constructed from marker intelligence tests) than the heterogeneous test. This was interpreted by the authors as indicating that, in homogeneous tests, people can profit from the rules, or solution strategies, they have used before, and thus a learning factor may appear. Since the same factor presumable appears in other intelligence tests, the homogeneous test and the (marker) intelligence tests show a higher correlation than the heterogeneous test and the marker intelligence tests. Both Ferrara et al.'s and Carlstedt et al.'s results take up the issue of intelligence as profiting from earlier items. We will come back to this link in the General Discussion section.

Note that the current proposal implies two things: First, people use the same rules throughout the test and become "primed" to use these rules. Second, the amount of priming is a factor of individual differences related to WM capacity.

What we will do in two experiments is construct intelligence tests (adapted versions of the RPM and a number series test, in Experiments 1 and 2, respectively) in which the number of

subresults to be stored within an item is relatively low in comparison with similar tests as, for example, the original RPM test. However, solution rules still need to be remembered over items. If WM only helps in storing subresults within an item, the standard account would predict that the correlation with WM capacity drops substantively in comparison with the correlation between WM capacity and standard intelligence tests. This is because the WM requirements within an item are now low, such that they can be fulfilled by any WM capacity within a reasonable range (our sample consists exclusively of students, who can be presumed to have WM capacities that are not extremely low). Therefore, since the range of performance in storing results within an item is strongly reduced, the correlation with WM capacity should be reduced. On the other hand, we would predict that the correlation does not decrease, since subresults still need to be stored over items, just as in, for example, the standard RPM test.

We will also check whether people indeed become primed to use the same solution principles over different items, as implicated by our hypothesis. This will be done by creating two conditions, in each of which different solution principles are used in an earlier phase. The question is then whether people who have seen items obeying one particular rule will be better on such items governed by that rule than people who have not seen this rule earlier (in Experiment 2), or have seen them a longer time ago (in Experiment 1).

2. Experiment 1

One intelligence test that is intensively used in intelligence research is the RPM. Different authors (e.g., Carpenter et al., 1990; Larson and Saccuzzo, 1989) have found correlations between performance on this test and WM capacity. Moreover, some authors have suggested that the RPM is the penultimate test to measure fluid intelligence (e.g., Marshalek, Lohman, & Snow, 1983).

In the present experiment, an adapted RPM test is administered to a group of participants (the adaptation we make is discussed in the next paragraph). Participants are divided in two conditions. People in each condition are given the same set of items, but in a different order. There are items of four rule types; each item can be solved by one (and only one) of these rules. In Condition 1, items of one type are presented one after the other and items of the remaining three types are presented intermixed. In Condition 2, the same set of items is presented but in a different order. Again items of one rule type (a different one) are presented one after the other, and the remaining ones are presented intermixed. According to our hypothesis, if people hold solution principles in WM, and the activation of these principles decays over time, people in each condition should find the items of the type that is presented consecutively easier than people who see these items intermixed between items of other rule types. Hence, we predict an interaction between condition and item type in response accuracy.

Most important, however, we investigate the relation between WM capacity and RPM performance. The WM test we use is constructed based on a popular test, namely, the WM span test developed by Daneman and Carpenter (1980). A recent review of the test and its use is given by Daneman and Merikle (1996). In their test, a set of sentences is presented to the participant; she is required to remember the last word of every sentence. The number of

sentences for which a participant can reliably remember all (or almost all) last words is her WM capacity. The idea is that, in a WM task, one should both remember and process elements at the same time (Just & Carpenter, 1992). We do not wish to assume that WM is a general (i.e., not domain-specific) factor. Therefore, we devise a WM test that is specific to the object/geometrical domain (see Smith & Jonides, 1997 for an overview of the evidence that object and verbal WM are separate systems; also Shah and Miyake, 1996). Nevertheless, the test should be as close as possible to the original Daneman and Carpenter (1980) test. How this is accomplished is detailed in the Method section.

Further, in our adapted RPM test, the amount of WM capacity required is held very low (within items). The traditional (e.g., Carpenter et al.'s, 1990) prediction would be that the correlation between RPM and WM drops strongly, since there are very few subresults to be stored within an (RPM) item. On the other hand, our prediction would be that, while the correlation may be dampened, it should still be substantial.

2.1. Method

2.1.1. Participants

$N=42$ persons participated in the experiment, $n=19$ in Condition 1, $n=23$ in Condition 2. All were first-year psychology students from the K.U. Leuven who participated for course credit.

2.1.2. RPM test

The first test is an adaptation of the RPM. A typical item is presented in Fig. 1. Participants are instructed to complete the matrix with one of the eight response alternatives in the lower part according to a logical rule. Only one rule is needed per item, with only one, or a few, rule instantiations (rule tokens) per item, where the number of rule instantiations refers to the number of times that a particular rule is used within an item. For example, if one rule is used twice in an item, there would be one rule and two rule instantiations involved in this item. This is in contrast with the real RPM, where many (up to five) rules should be used to solve an item. The test is presented on a computer. A response is chosen by clicking with a mouse pointer at one of the eight alternatives.

Four types of rules are used. The rules are chosen on the basis of the Carpenter et al. (1990) rule system, but in such a way that the rules are conceptually as different as possible. The resulting rules are rotation, progression, unique/common, and distribution of 3. There were 4 items of type rotation, 6 of type progression, 5 of item type unique/common, and 5 of item type distribution of 3, making 20 items altogether. The *rotation* rule entails that a figure is rotating over the different columns of the matrix. The second rule, *progression*, means that there is a steady progression (ascending or descending) in the number of elements; for example, the first figure of the matrix may contain one square, the second one two squares and the third one three. *Unique/common* captures two rules (which are aggregated because they are similar). "Unique" means the following: Element 3 is a combination of the previous two, in such a way that only the unique parts of Elements 1 and 2 are retained in the third element. The example of Fig. 1 is a "unique" item. The "common" rule is the opposite: Here, the

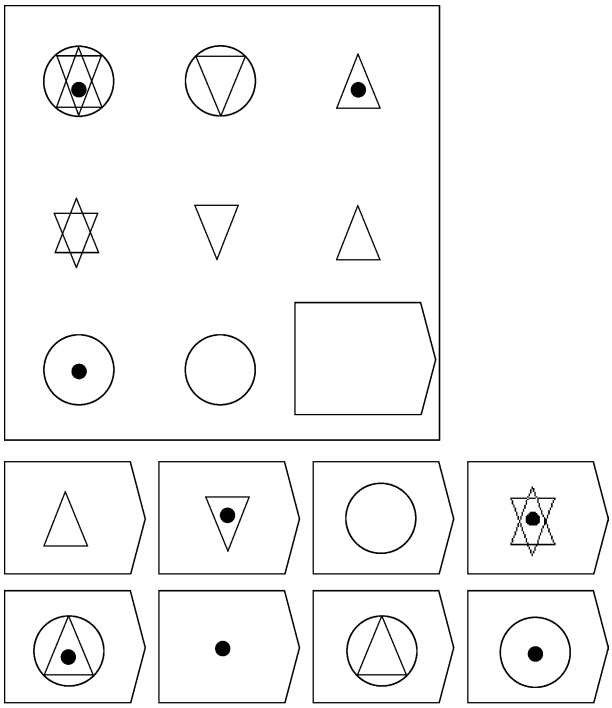

Fig. 1. An item of the adapted RPM task.

common (i.e., nonunique) parts of Elements 1 and 2 are retained in Element 3. Finally, *distribution of 3* is the rule where the same three figures are used in every row (but possibly in a different order). For example, the first row may consist of the elements "circle–square–triangle," the second one of the elements "square–triangle–circle" (i.e., the same elements in a different order), and the third one of the elements "circle–triangle–square."

As noted, we have tried to construct items where the WM requirements within an item are low. To check this, two observers were asked to score all items on a scale ranging from 0 (*very low WM capacity required*) to 10 (*very high WM capacity required*). The mean score (over items and observers) was 3.60, which suggests that we succeeded in our intention.

2.1.3. WM test

This is also a computerized test. Two series of five items are presented like the one in Fig. 2. Two principles are used to construct the items: Figures become bigger (as in the example in Fig. 2) and figures become darker. Participants are requested to complete the series: They are given four answer possibilities and are requested to choose one by clicking on it with a mouse pointer (see Fig. 2). This is clearly very easy; its difficulty is comparable to the process of reading a sentence, as is required in Daneman and Carpenter's (1980) task.

Also, participants should remember the last figure of every series, that is, the last figure before the question mark. (In the example of Fig. 2, they should remember the moderately

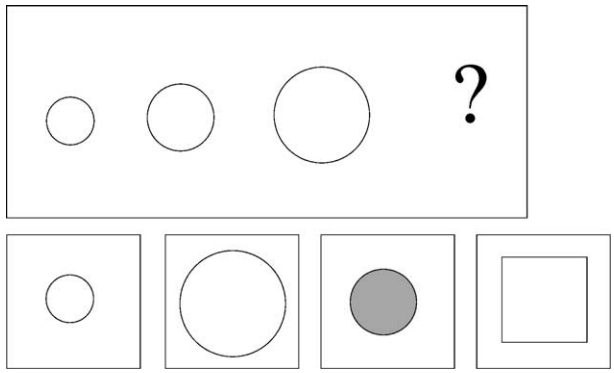

Fig. 2. An item of the WM task.

large circle.) This last process is comparable to remembering the last word of every sentence as is required in Daneman and Carpenter's (1980) task.

After each series of five such items, a recognition test is given. In this recognition test, 10 figures are shown sequentially, and the participant is asked whether the figure is in her remembered list or not (the participant should click a "Yes" or "No" button). Only 4 out of 10 items per recognition test refer to items seen earlier (i.e., to only 4 items, the answer should be "Yes"). The other 6 are filler items and do not correspond to items seen previously.

Since we assume WM capacity to put a limit on the number of items that can be remembered, it seems plausible that remembering one item has a serious impact on remembering other items. Indeed, concentrating attention on one particular item probably limits the extent to which other items can be attended to. Hence, there probably are item dependencies in this test, thus violating the statistical independence assumption of classical test theory (we return to this point later).

2.1.4. Procedure

Participants enter the experiment room and are seated in front of a Pentium computer. They are requested to guide themselves through the introduction to the WM test by clicking the mouse button. After the introduction, there is time to ask questions concerning the testing procedure. If everything is clear, the participant is left alone to perform the WM test. After the WM test, the experimenter enters the room again, and the same procedure is repeated for the RPM test.

2.2. Design and predictions

Two conditions are created. In the first condition, all items of type distribution of 3 are presented immediately following each other. Specifically, these items are presented as numbers 16, 17, 18, 19, and 20. Other types of items (e.g., progression items) are presented with gaps of 1 to 4, meaning that 1, 2, 3, or 4 items of other types appear between two consecutive items of a certain type. In Condition 2, items of type unique/common are presented as the items 16 to 20;

the distribution of 3 and unique/common item numbers of Condition 1 are simply replaced with one another. If rules are used and retained (in a gradually decaying manner) throughout the test, then participants in Condition 1 should find distribution of 3 items easier than participants in Condition 2 do; the reverse is predicted for unique/common items. The reason for this prediction is that when items of the same type follow each other, it is easier to retain the corresponding rule. Intervening items (items of other types) may interfere with retention of a specific rule. Hence, we predict an interaction effect between condition (1, 2) and rule type (distribution of 3, unique/common) with accuracy as the dependent variable.

The second and most important prediction we make is that the correlation between the WM score and the RPM score is significantly larger than zero, and similar to comparable correlations reported in the literature. This prediction is the most important because it directly contrasts the two views on the role of WM capacity. Indeed, in the traditional view, WM and RPM scores are correlated because many subresults should be stored within an item. Hence, for a reasoning test with very few rule instantiations (like the one of the present paper), the correlation should drop to zero according to the traditional view. On the other hand, since solution principles still need to be remembered over items, we would predict that the correlation remains substantial.

2.3. Results

2.3.1. Descriptive measures

The mean and standard deviation of the adapted RPM test are 13.93 and 2.72, respectively. The Spearman–Brown corrected split-half reliability equals .68.

For the WM test, the mean and standard deviation are 14.14 and 2.02, respectively. The Spearman-Brown corrected split-half reliability equals .22. This is low, but it should be noted that reliability is not a useful measure in the present context (see Appendix A). We view WM as a limited, person-specific capacity ("resource pool") that can be used to store and process items (Just & Carpenter, 1992). Therefore, if a person concentrates on the first set of items, she will have less WM available for storing the remaining items. Similarly, one person may choose to concentrate attention on the square elements only, or the dark elements only, or maybe a random subset of items, since storing all items is too difficult. This would result in negative interdependencies between items, violating the assumptions of classical test theory[1] and making the concept of reliability less useful.

[1] In fact, interitem correlations range from −.49 to .53 with a mean of .02 in our data. However, it is difficult to describe the exact pattern of item interdependencies, since each person may follow his or her own idiosyncratic strategy to store items. If negative interdependencies occur, the standard assumption of classical test theory (error correlates only with itself, Lord & Novick, 1968, p. 56), no longer holds. Therefore, the classical result that validity is lower than the root of the test's reliability (e.g., Lord & Novick, 1968, p. 69) should no longer hold. In fact, it is possible that the test's estimate of reliability is zero or negative and the validity of the test is high. We illustrate this phenomenon in Appendix A with a simulation study. For a similar reason, low reliabilities are typically obtained in the TAT test (Atkinson, Bongort, & Price, 1977; Reuman, 1982; Tuerlinckx, De Boeck, & Lens, 2002).

Table 1
Correlations between WM capacity and RPM test, and between WM capacity and number series test for Experiments 1 and 2, respectively (one-sided *P* values within parentheses)

	Condition 1	Condition 2	All participants
Experiment 1			
Total WM score	.54 (.009)	.44 (.019)	.47 (<.001)
Positive item score	.61 (.003)	.51 (.007)	.54 (<.001)
Experiment 2	.58 (.001)	.44 (.010)	.50 (<.001)

2.3.2. WM–RPM correlation

The correlations between the total WM score and RPM score for each condition separately and over both conditions are reported in the third row of Table 1 (one-sided *P* values are shown within parentheses). As can be seen, these correlations are moderately high and statistically significant.

As may be recalled, in the recollection phase of the WM test, the participant is to respond "Yes" to some figures shown there (meaning, "Yes, I have seen this figure earlier in the test"). These items seem to be the most relevant items in the WM score, since they refer to figures that are to be remembered by the participant (whereas the other items contain figures that the participant has not seen). If only these "positive" items are incorporated into the WM score, Table 1 (fourth row; positive item score) shows that the correlation increases.

We now compare these correlations with correlations obtained in similar settings reported in the literature. Only correlations between WM tests and the advanced RPM test (on which our own test is based), administered on comparable populations (university students) are mentioned here. These are .50, .55, .59, and .77, reported by Larson, Merritt, and Williams (1988), Larson and Saccuzzo (1989), Babcock (1994), and Carpenter et al. (1990), respectively. Hence, the correlations we find are of a magnitude similar to comparable correlations reported in the literature.

2.4. Experimental effect

The Condition × Item Type mean proportions correct are given in Table 2. One can see that, as expected, people in Condition 1 are better on distribution of 3 items (relative to Condition 2), while people in Condition 2 are better on unique/common items (relative to Condition 1). However, the interaction is only marginally significant at level .05, $F(1,40)= 3.17$, $P=.08$.

Table 2
Condition×Group data, Experiment 1

	Unique/common items	Distribution of three items
Condition 1	.58	.56
Condition 2	.70	.51

2.5. Discussion

We have succeeded in finding a correlation between WM capacity and an intelligence test with low within-item WM requirements. However, some aspects of our design were not optimal. Our WM procedure may have been problematic in that the correctness on the series completion multiple choice question was not taken into account. This is a problem because some people may have chosen not to pay attention to the question and simply remember the figure, thus boosting their WM score (Salthouse, 1991). However, because the answers to these questions were not recorded, this could not be checked.

We also note that the reliabilities were low (for both tests, although the problem was most pronounced for the WM test). One possible reason was already mentioned earlier for the WM test (negative item dependencies). Another possible reason could have been that both tests were rather short.

One other problem with our results is that, although the interaction was in the expected direction, it was not statistically significant. Two possible problems that may have caused this are the following. First, the items we used may have allowed different ways of solving the items correctly. Hence, people may have been solving the items in ways other than we intended them to do, possibly weakening the amount of relevant solution rule priming. Second, since different people solve a different number of items correctly, it is difficult to control the exact amount of priming each participant receives. To replicate our findings in another domain and avoid the problems noted here, the second experiment is now presented.

3. Experiment 2

A second type of task that is often used in intelligence research is a number series task (e.g., Carlstedt et al., 2000; Holzman, Pellegrino, & Glaser, 1983; Lefevre and Bisanz, 1986). In this experiment, we administer two tasks, a verbal WM task and a number series task. In the latter, participants are shown number series of the form "3 5 7 9 11 13" and are instructed to find the rule describing the series. The number series always consists of six numbers. Four solution rules are used: addition, Fibonacci, interpolation, and multiplication. An example of the *addition* type is "3 4 6 9 13 18." Here, the increment between two consecutive items is itself incremented by one (so, the sequence of increments is +1, +2, +3, +4, +5). The increment between two successive increments is always equal to one, but items can differ in starting number and initial increment. In items of the *Fibonacci* type, each number is the sum of the previous two. The first two numbers are chosen arbitrarily, with the second number larger than the first. An example is "2 4 6 10 16 26." In items of the *interpolation* type, two sequences are intermixed, each one with a constant increment rule. An example is "2 5 6 8 10 11," which is a mix of the sequences "2 6 10 ..." and "5 8 11 ...." In items of the *multiplication* type, the increment is always multiplied by a constant value. An example is "1 3 7 15 31 63." The multiplier is always equal to 2.

These rules are not unique and in fact, an infinite number of rules can validly describe each item (Korossy, 1998). However, we believe these four rules to be the simplest ones that

adequately describe the items. Since we asked participants to *describe* the rule that they used in this task (rather than complete the series in a logical manner, which is the standard procedure; see Method section below), it is possible to investigate this. It turned out that none of our participants ever used a *valid* rule other than the ones mentioned here.

As in the previous experiment, the number of WM requirements within an item was low. This can be justified on the basis of two aspects of the current task: A low number of placekeepers (to be explained in a moment) is required within an item, and no application of the rule that was found is needed. Each aspect will now be discussed consecutively. First, Holzman et al. (1983) introduced a measure of WM requirement within number series items, namely, the *number of placekeepers* an item requires. This refers to the number of variables needed to describe the rule involved in the item. This number can be assessed in a relatively straightforward manner. Holzman et al. devised a formal coding system for this type of rules: The number of variables that is required within this system to describe a rule can be treated as the number of placekeepers. For example, the addition rule could be described in their system as $[M_1, +N_1(M_1), +1(N_1)]$, meaning "start with M_1, add N_1 to M_1, add 1 to N_1," and thus requires two placekeepers (M_1 and N_1). Holzman et al. devised rules where 0, 1, 2, or 3 placekeepers were required. They found that the number of placekeepers that is required is a strong predictor of item difficulty: The authors report, for different subject groups, correlations between number of placekeepers and item difficulty above .70.

Each of our four rules (addition, Fibonacci, interpolation, and multiplication) can be described in Holzman et al.'s (1983) system as a rule where exactly two placekeepers are required. In our opinion, this is as low as possible without making the item extremely easy: Items with 0 or 1 placekeepers are very easy and, thus, pose no challenge (e.g., typical items with 0 or 1 placekeepers would be "3 3 3 3 3 3"... and "3 5 7 9 11 13," respectively). This is the first reason why we assume that the WM load within an item is low in our newly constructed test.

A second reason is that part of the WM load in solving number series items results from correctly applying the rule that is found. Since this requirement is no longer imposed (people have to state the rule they are using, rather than actually applying it), this source of memory load is removed. Hence, we think we have succeeded in again making the WM requirements within an item as low as possible. Further evidence for this assertion will be presented in the Results section (subsection Number size), where we look at the effect of the size of the numbers that participants have to operate with while finding the correct rule.

3.1. Method

3.1.1. Participants

Fifty-five people participated, either for course credit or a small monetary reward. Of these, 27 participated in Condition 1, 28 in Condition 2. They were first-year psychology students receiving credit for their participation.

3.1.2. Procedure

Two tasks are administered, with about 1 week in between. The first task is a close analogue of the WM task described by Shah and Miyake (1996) (which is itself an adaptation

of the WM task of Daneman & Carpenter, 1980). Since the intelligence test used is no longer geometric but verbal/quantitative in this experiment, a verbal version of the WM task will now be administered (see Oberauer, Süß, Schulze, Wilhelm, & Wittmann, 2000, for evidence that verbal and quantitative information tap the same WM source). Specifically, in this task, a sentence is presented on the screen, which has to be judged to be correct or incorrect. The level of difficulty of judging a sentence is very low. For example, a correct sentence is "Robert de Niro is an actor." An incorrect sentence is "Cancer is caused by kissing." After this sentence, a word is presented for 800 milliseconds: This is the word that participants are instructed to remember. After n such sentences, participants are required to write down all words they have remembered. The number n increases throughout the test. First, five items with n=2 are presented, then five items with n=3, and so on, up to n=5. Hence, to obtain a perfect score, the participant should remember $5\times(2+3+4+5)=70$ words (but not, of course, all at once). All words to be remembered are two-syllable nouns. Every participant completed all items of this test.

In the WM task, a word is scored as correct if it is recalled *and* the sentence preceding that word is answered *correctly* and *in time*. The time limit to judge the correctness of a sentence is 5 seconds. The rationale for requiring the corresponding sentence to be correct, is that otherwise people can simply remember the words without paying attention to the sentences (see Discussion, Experiment 1). The rationale for requiring that the sentence be answered in time is that otherwise people can take time to internally rehearse the words to be remembered. If people would pay no attention to the sentence, or would internally rehearse the words, our task would become an STM task, rather than a WM task (see discussion earlier). By correcting the WM score with the answer to the sentences, we try to make sure that the task is a real WM task rather than an STM task (see also Salthouse, 1991). People are informed that they have to answer fast, and they are given a warning sign on every sentence on which they are too slow.

Concerning the number series task, a test of 50 items is constructed, with items of the four types intermixed. This test is presented twice in identical order. Hence, 100 items are presented. The different rule types are introduced gradually. In items 1 to 10, the addition and Fibonacci rules are used. In items 11 to 20, the interpolation rule is added to these two. In items 21 to 30, the addition and interpolation rules are used. In items 31 to 40, the multiplication rule is added to these two. Finally, in items 41 to 50 all four rule types are used. The exact same sequence is repeated for items 51–100. In total, there are 34, 24, 28, and 14 items for types addition, Fibonacci, interpolation, and multiplication, respectively. Since some people did not complete the number series test (because one hour had passed), the number series score is defined as the number of successes divided by the number of items attempted.

A time limit of 25 seconds per item is imposed in the number series task. Participants worked until either the task was completed or one hour had passed, whichever came first.

3.2. Design and predictions

Two conditions are created, differing only in the explanation that is given in the introduction to the number series test. In the first condition, rule interpolation is shortly

explained in the introduction of the test. In the second condition, rule multiplication is shortly explained in the introduction. We predict people in Condition 1 to be relatively better on interpolation items, and participants in Condition 2 to be relatively better on multiplication items. Hence, we predict an interaction between condition and item type (interpolation, multiplication) with accuracy as the dependent variable. For the correlational effect, the prediction is similar to Experiment 1.

3.3. Results

One person was removed from Condition 1 since no number series data were collected for this person. Two persons were removed from Condition 2, one because the WM task was incorrectly filled in, and one because the number series test was incorrectly filled in.

3.3.1. Descriptive measures

For the number series test, the mean accuracy is .65 with standard deviation .13 (the test score is rescaled here to a 0–1 scale because not every participant attempted the same number of items; see earlier). The Spearman–Brown corrected split-half reliability equals .75.

For the WM test, the mean and standard deviation are 56.35 and 7.14, respectively. The Spearman–Brown corrected split-half reliability equals .82.

3.3.2. Correlations

We first investigate whether there is a correlation between the WM score and the number series score. The seventh row of Table 1 shows the correlations between the WM score and number series score, separately for every condition. For example, in Condition 1, the correlation of the WM score with total number series score equals .58. Within parentheses, we display one-sided *P* values. This table shows that there is a reasonably strong correlation between the WM score and the number series score. It can be noted that the correlations are in line with the ones obtained in Experiment 1, and also with other correlations reported in the literature (see earlier references and Kyllonen & Christal, 1990, p. 403).

3.3.3. Experimental effects

We now turn to the effect of hints in the introduction of the test. Table 3 shows mean accuracies separately for each solution principle and for all four types aggregated (column "Total"). The results are reported per condition. Standard deviations are shown within parentheses. It can be seen that the experimental manipulation (prior experience) is

Table 3
Mean accuracies (standard deviations within parentheses) per item type and condition in the number series task, Experiment 2

	Total	Addition	Fibonacci	Interpolation	Multiplication
All items					
Condition 1	.67 (.16)	.87 (.14)	.75 (.22)	.40 (.38)	.58 (.27)
Condition 2	.62 (.10)	.92 (.07)	.75 (.21)	.07 (.19)	.72 (.22)

successful: Accuracies on items of rule interpolation are higher in Condition 1 than in Condition 2. The reverse holds for items of rule multiplication. Hence, the expected interaction is found. Taking only the results of items of types interpolation and multiplication (the ones that are of primary interest), we find a main effect of rule type, $F(1,50)=67.15$, $P<.001$, and the expected interaction, $F(1,50)=19.30$, $P<.001$.

On the other hand, rules that are not presented in the introduction (addition, Fibonacci), are not influenced by the experimental manipulation. There is a significant effect of item type, in that the addition rule is easier than the Fibonacci rule, $F(1,50)=33.34$, $P<.001$, but neither the condition nor the interaction effect are significant (both Ps $> .2$).

3.3.4. Number size

One might object that another source of difficulty in solving these items is the size of the numbers that are used in a number series item. Hence, items may be difficult (WM-dependent) because they require juggling with large numbers. To test this, we define *low-number items* as items where the sum of the numbers used in that item is smaller than the mean sum of numbers (e.g., for "1 2 4 7 11," the sum equals 1+2+4+7+11=25) for the rule class (e.g., Fibonacci) to which the item belongs. Analogously, define *high-number items* to be items where the sum is higher than this mean. Then, if this alternative account is correct, and high-number items require more WM resources than low-number items, low-number items should be easier than high-number items. Table 4 displays the mean accuracies of both low-number and high-number items for the four item types. There is no consistent effect of low-number items being easier found here, contradicting this number size account.

3.4. Discussion

As before, we obtained a strong correlation between WM capacity and the cognitive task (in this case, number series), although we have argued that WM requirements within an item were low. The reasons were that (a) only two placekeepers were required within an item, and (b) the rule found by a participant did not have to be applied to generate new numbers of the series. Further, it turned out that the size of the numbers within an item did not influence the

Table 4
Accuracies for low-number and high-number items, Experiment 2

		Condition 1	Condition 2
Addition	low	.87	.92
	high	.88	.93
Fibonacci	low	.78	.79
	high	.73	.70
Interpolation	low	.38	.07
	high	.41	.07
Multiplication	low	.58	.72
	high	.83[a]	.72

low=low-number items; high=high-number items.

[a] One person not included since no high-number multiplication items were attempted.

difficulty of the item. Finally, in contrast to the previous experiment, the experimental interaction was statistically significant.

4. General discussion

It was shown that, for the RPM test, individual differences in WM are relevant also if the WM load per item is low. The same pattern was found in a number series test that was very similar to tests commonly used in the literature. The correlations we obtained were of comparable magnitude to correlations reported in the literature. It was suggested that the reason for this correlation is that persons with high WM capacity are more efficient in storing rules over items than are persons with low WM capacity.

A recent view on WM is that it consists of the activated portion of long-term memory elements (e.g., Cantor and Engle, 1993; Engle et al., 1999; Just & Carpenter, 1992). We adhere to this view, and applied it in this paper to the relation between WM capacity and intelligence by noting that elements can become activated during problem solving not only *within*, but also *between* items. As noted in the Introduction, recent research that is consistent with our view is that of Carlstedt et al. (2000). These authors found that the involvement of general intelligence is higher for homogeneous tests than it is for heterogeneous tests; that is, tests in which items of a particular rule type are grouped in the test (homogeneous tests) correlate more highly with general intelligence than tests in which items adhering to different solution principles are randomly intermixed (heterogeneous tests). As suggested by Carlstedt et al., this may be because the relevant rules are activated over items in homogeneous tests, but not in heterogeneous ones. Of course, our tests also used the intermixed (heterogeneous) format, but the crucial thing is not so much whether a test is homogeneous versus heterogeneous, but rather whether WM differences can lead to different levels of rule retention (and hence to differences in performance). In this sense, our test (and interpretation) is similar to theirs.

This brings us to a link between two popular concepts in modern intelligence research, namely learning and WM capacity. Early negative findings by Woodrow (1946) seem to have discouraged later researchers from further investigating the relationship between learning and intelligence. However, many authors have noted problematic aspects of Woodrow's investigation, notably (a) the use of gain scores to assess learning, which have problematic psychometric properties (Ackerman, 1987; Cronbach & Furby, 1970), (b) not taking into account the initial level of performance (Ackerman, 1987), and (c) the use of very simple tasks to assess learning potential (Ferrara et al., 1986).

Interest in the relationship between learning and intelligence has recently revived, and gain scores for assessing learning have been abandoned in favor of other measures that circumvent their problems (e.g., Embretson, 1991; Ferrara et al., 1986; Ferretti and Butterfield, 1992). Also, the work of Ackerman (1987, 1988) has been concerned with the link between learning and intelligence, in which he investigates sources of variance at different time points during skill acquisition. Although we have not explicitly studied learning as a source of individual differences, we did study WM capacity and interpreted the role of WM capacity in the present

context as the capacity to remember solution rules over items. In this sense, we have indirectly reconfirmed the link between intelligence and learning.

In our opinion, this view on the mutual relationships between these three concepts (WM capacity, learning, intelligence) can also elucidate the *process* of how people solve intelligence tests. One suggestion made by the present research is that an overlooked aspect of the process of solving intelligence tests, is that the same rules are used repeatedly over items. This process was already studied in more detail by Verguts, Maris, and De Boeck (in press), and the present paper looked at this process from the point of view of individual differences. Examining just how important it is relative to other sources of individual differences stands out as an issue of future research.

Acknowledgments

Part of this work was done while the first author was a postdoctoral researcher with the Fund of Scientific Research–Flanders (Belgium). We thank Frank Rijmen for his useful comments on a previous version of the paper.

Appendix A.

We will show that low reliabilities do not necessarily imply low validities if classical test theory assumptions are violated. Suppose W_{pi} denotes the test score of person p on item i. The second test score, which is used to validate the first one, is denoted T. Person p achieves a total score of T_p on this second test. We assume that the same true score, or latent ability, θ_p generates both W and T scores. That is, we assume $W_{pi}=\theta_p+C_{pi}+\varepsilon_{pi}$, and $T_p=\theta_p+\varepsilon'_p$. The variables ε_{pi} and ε'_p are assumed to be normally distributed with mean 0 and standard deviation 1 and do not correlate with any other variable (other than with itself). The variable C_{pi} is used to introduce negative item interdependencies between the discrepancies $W_{pi}-\theta_p$. If all $C_{pi}=0$, the standard classical test theory assumptions hold (Lord & Novick, 1968, p. 56). If some $C_{pi}\neq 0$ but $\sum_i C_{pi} = 0$ for all p, the reliability of the test scores W is affected but the validity is not.

Suppose there are 100 persons (N=100), 20 items (I=20) and latent abilities are sampled from a normal distribution with mean 0 and a standard deviation of 3 ($\theta\sim N(0, 9)$). Further, suppose everyone concentrates either on the first part of the test (the first 10 items) or the second part of the test (the last 10 items). This can be formalized as follows. For persons concentrating on the first part of the test, $C_{p1}=C_{p2}=\ldots=C_{p10}=C>0$, while $C_{p11}=C_{p12}=\ldots=C_{p20}=-C<0$. The probability of concentrating on the first or the second part is equal to .50. Simulation results for different values of C are presented in Table A1. We show, for different values of C, the correlation (reliability estimate) between Parts 1 and 2 of the test ($r(W_{\text{part1}}, W_{\text{part2}})$), and between the even and odd parts of the test, the even part corresponding to the items 2, 4, 6 and so on, the odd part to the items 1, 3, 5 and so on. The results show that, if $C=0$, validity ($r(W, T)$) is lower than the root of the reliability, as it should be. (Table A1 shows

Table A1
Reliability and validity, simulation results (1)

C	$r(W_{part1}, W_{part2})$	$r(W_{even}, W_{odd})$	$r(W, T)$
0	.99	.99	.92
1	.81	.99	.95
2	.46	.99	.95
3	.06	.99	.95
4	−.28	.99	.95
5	−.44	.99	.94

that the validity is lower than the reliability; hence it is also lower than the root of the reliability.) However, as soon as negative interdependencies are introduced, the reliability dampens. The reliability based on the odd/even partition does not suffer from this effect, since each variable used in its calculation contains items with positive and negative discrepancies $W_{pi}-\theta_p$. On the other hand, if the reliability is calculated based on the Part 1/Part 2 split, the reliability decreases dramatically. This does not affect the validity of the test, as is shown in the last column of Table A1.

For the case just discussed, one can calculate the reliability based on the odd/even partition, and give a plausible reason why the Part 1/Part 2 split-up is inappropriate. However, even this is not always possible, namely, when different people use different strategies to divide their attention. To be specific, suppose that some people follow the strategy to remember odd items (1, 3, 5, . . .), other people remember even items (2, 4, 6, . . .) and still other people make use of one the two strategies discussed above (concentrate on Part 1 or Part 2 of the test). The two new C patterns are (C, $-C$, C, $-C$, . . ., $-C$), ($-C$, C, $-C$, C, . . ., C) for odd and even preferences, respectively. Suppose each preference (odd, even, Part 1, Part 2) occurs with a probability of .25. The corresponding simulation results are presented in Table A2. One may note that the Part 1/Part 2 reliability suffers less than before (since there are fewer people choosing the Part 1 or Part 2 strategy), but the effect is that the other reliability estimate (based on the odd/even split) lowers as well. However, there is still no effect on the validity of the test. These findings corroborate the statement made in the text that in a test situation like the present one, where negative item dependencies occur, the reliability coefficient is not a very useful measure.

Table A2
Reliability and validity, simulation results (2)

C	$r(W_{part1}, W_{part2})$	$r(W_{even}, W_{odd})$	$r(W, T)$
0	.99	.99	.95
1	.92	.91	.97
2	.62	.68	.94
3	.39	.43	.95
4	−.01	−.01	.94
5	−.11	−.09	.95

References

Ackerman, P. L. (1987). Individual differences in skill learning: an integration of psychometric and information processing perspectives. *Psychological Bulletin*, *102*, 3–27.

Ackerman, P. L. (1988). Determinants of individual differences during skill acquisition: cognitive abilities and information processing. *Journal of Experimental Psychology: General*, *117*, 288–318.

Atkinson, J. W., Bongort, K., & Prince, L. H. (1977). Explorations using computer simulation to comprehend thematic apperceptive measurement of motivation. *Motivation and Emotion*, *1*, 1–27.

Babcock, R. L. (1994). Analysis of adult age differences on the Raven's advanced progressive matrices test. *Psychology and Aging*, *9*, 303–314.

Cantor, J., & Engle, R. W. Working-memory capacity as long-term memory activation: An individual differences approach. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 19*, 1101–1114.

Carlstedt, B., Gustafsson, J.-E., & Ullstadius, E. (2000). Item sequencing effects on the measurement of fluid intelligence. *Intelligence*, *28*, 145–160.

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

Conway, A. R. A., & Engle, R. W. (1996). Individual differences in working memory capacity: more evidence for a general capacity theory. *Memory*, *4*, 577–590.

Cronbach, L. J., & Furby, L. (1970). How we should measure "change"— or should we? *Psychological Bulletin*, *70*, 68–80.

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

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

Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika*, *56*, 495–515.

Embretson, S. E. (1995). The role of working memory capacity and general control processes. *Intelligence*, *20*, 175–186.

Engle, R. W., Cantor, J., & Carullo, J. J. (1992). Individual differences in working memory and comprehension: a test of four hypotheses. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*, 972–992.

Engle, R. W., Kane, M. J., & Tuholski, S. W. (1999). Individual differences in working memory capacity and what they tell us about controlled attention, general fluid intelligence, and functions of the prefrontal cortex. In A. Miyake, & P. Shah (Eds.), *Models of working memory* (pp. 102–134). New York: Cambridge University Press.

Ferrara, R. A., Brown, A. L., & Campione, J. C. (1986). Children's learning and transfer of inductive reasoning rules: studies of proximal development. *Child Development*, *57*, 1087–1099.

Ferretti, R. P., & Butterfield, E. C. (1992). Intelligence-related differences in the learning, maintenance, and transfer of problem-solving strategies. *Intelligence*, *16*, 207–223.

Fry, A. F., & Hale, S. (2000). Relationships among processing speed, working memory, and fluid intelligence. *Biological Psychology*, *54*, 1–34.

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

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

Korossy, K. (1998). Solvability and uniqueness of linear-recursive number sequence tasks. *Methods of Psychological Research*, *3* (http://www.mpr-online.de).

Kyllonen, P., & Christal, R. (1990). Reasoning ability is (little more than) WM capacity?! *Intelligence*, *14*, 389–434.

Larson, G. E., Merritt, C. R., & Williams, S. E. (1988). Information processing and intelligence: some implications of task complexity. *Intelligence*, *12*, 131–147.

Larson, G. E., & Saccuzzo, D. P. (1989). Cognitive correlates of general intelligence: toward a process theory of *g*. *Intelligence*, *13*, 5–31.

LeFevre, J., & Bisanz, J. (1986). A cognitive analysis of number-series problems: Sources of individual differences in performance. *Memory & Cognition*, *14*, 287–298.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

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

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.

O'Reilly, R. C., Braver, T. S., & Cohen, J. D. (1999). A biologically based computational model of working memory. In A. Miyake, & P. Shah (Eds.), *Models of working memory* (pp. 375–411). New York: Cambridge University Press.

Raven, J. C. (1965). *Advanced progressive matrices, set II*. New York: Psychological Corporation.

Reuman, D. A. (1982). Ipsative behavioral variability and the quality of thematic apperceptive measurement of the achievement motive. *Journal of Personality and Social Psychology*, *43*, 1098–1110.

Salthouse, T. A. (1991). Mediation of adult age differences in cognition by reductions in working memory and speed of processing. *Psychological Science*, *2*, 179–183.

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.

Smith, E. E., & Jonides, J. (1997). Working memory: a view from neuroimaging. *Cognitive Psychology*, *33*, 2–42.

Tuerlinckx, F., De Boeck, P., & Lens, W. (2002). Measuring needs with the Thematic Apperception Test: A psychometric study. *Journal of Personality and Social Psychology*, *82*, 448–461.

Verguts, T., Maris, E., De Boeck, P. (in press). A dynamic model for rule induction tasks. *Journal of Mathematical Psychology*.

Wickelgren, I. (1997). Working memory linked to intelligence. *Science*, *275*, 1581.

Woodrow, H. (1946). The ability to learn. *Psychological Review*, *53*, 147–158.