Nuria Carriedo, M. Rosa Elosúa, Juan A. García-Madruga
Working Memory, Text Comprehension, and Propositional Reasoning: A New Semantic Anaphora WM Test
Universidad Complutense de Madrid
España

Available in: http://www.redalyc.org/articulo.oa?id=17217456003

The Spanish Journal of Psychology,
ISSN (Printed Version): 1138-7416
psyjour@sis.ucm.es
Universidad Complutense de Madrid
España
The aim of this study was to present a new working memory test following the line of work started by García-Madruga et al. (2007) and to examine its relation to reading comprehension and propositional reasoning measures. In that study we designed a new working memory span test –based on Daneman & Carpenter’s (1980) Reading Span Test (RST)– in which the processing task called for an inferential decision –to resolve a pronominal anaphora based on Morpho-Syntactic cues– and had people recall the result of this inference. In the current study, besides the RST and the Morpho-Syntactic Anaphora test, we presented a new Semantic Anaphora measure. In order to check the validity of this new Working Memory (WM) task, we used the same reasoning task used in the previous study as well as a new reading comprehension test. The results show the tight relationship amongst working memory, reading comprehension and reasoning, and confirm the validity of the new WM measure.

Keywords: working memory, reading comprehension, reasoning.

El propósito de este trabajo es presentar una nueva prueba de memoria operativa y examinar su relación con la comprensión lectora y el razonamiento, en la línea comenzada por el estudio de García-Madruga et al. (2007). En ese estudio se diseñó una nueva medida de amplitud de memoria operativa –basada en Reading Span Test (RST) de Daneman & Carpenter (1980)– en la que la tarea de procesamiento exigía la realización de una inferencia –resolver una anáfora pronominal a partir de los rasgos Morfosintácticos– y los participantes debían recordar el resultado de esta inferencia. Además del RST y la prueba de Anáforas Morfosintácticas, en el presente estudio presentamos una nueva medida de Anáforas Semánticas. Para comprobar la validez de esta nueva tarea de Memoria Operativa (MO) hemos utilizado la misma tarea de razonamiento que en el estudio anterior, así como una nueva tarea de comprensión lectora. Los resultados muestran la estrecha relación entre memoria operativa, comprensión lectora y razonamiento, y confirman la validez de la nueva medida de MO.

Palabras clave: memoria operativa, comprensión lectora, razonamiento.

This research was supported by the following grants: Spanish Ministerio de Educación y Ciencia BSO2003- 02103, SEJ2004-07398/PSIC, PSI2008-00754/PSIC and CONSOLIDER-INGENIO 2010 CSD2008-00048. We would thank teachers and students from Instituto Veritas for their participation in this study and to Eduardo Vidal-Abarca for providing us their Comprehension Strategy Test. We wish also to thank Francisco Gutiérrez, and Óscar Vila for their helpful ideas and comments.

Correspondence concerning this article should be addressed to Juan A. García-Madruga. Dpto. Psicología Evolutiva y de la Educación. U.N.E.D. C/ Juan del Rosal, 10. 28040. Madrid. (Spain). Pone: +34-913986244. Fax.: +34-913987951. E-mail: jmadruga@psi.uned.es
The aim of this study was to present a new working memory test (based on the Daneman & Carpenter's Reading Span Task) and to examine its relationship to reasoning and reading comprehension measures. This new task focuses on increasing the attentional demands of the Central Executive, just like in the line of work started by Garcia-Madruga, Gutiérrez, Carriedo, Luzón and Vila (2005, 2007); (Gutiérrez, García-Madruga, Carriedo, Vila, & Luzón, 2005).

Working Memory is considered a central component of cognition. However, as Barret, Tugade, and Engle (2004) have pointed out, an operational definition of working memory is easier than a conceptual one. From an operational point of view, the consensus is that working memory capacity is the number of items that can be recalled during a complex working memory task. From a conceptual perspective, there is no general agreement about the definition of Working Memory capacity, namely because there are diverse theories that emphasize different aspects of working memory (see Miyake & Shah, 1999). Nevertheless, there is no question that one of the most influential is the Multiple Component Model proposed by Baddeley and Hitch (1974; Baddeley, 2000).

According to this theoretical model, the working memory system is composed of two domain-specific storage structures or slave systems (the phonological loop and the visuo-spatial sketchpad), an episodic buffer that links two prior components with long term memory, and a central executive. The central executive is the main component of the working memory system. It not only has to coordinate the other components, but it is also in charge of the attentional control of information. That is, it has to focus and switch attention, to activate representations, to inhibit automatic processes and to discard irrelevant information (see Baddeley, 2000). Thus, the model attributes an active role to the central executive for processing information, but also to the recognized structural limitations of working memory concerning both capacity and permanence of information in the system.

This model explains not only a huge range of data concerning memory functioning, but the limited capacity of the central executive has also been used as a basis for explaining the individual differences observed in different high-level cognitive tasks (e.g., Richardson, 1996). This central role in cognition has led many authors to propose that complex span tasks -those that simultaneously imply processing and storage- are better at assessing WM capacity than traditional short-term memory span tasks, which require simple maintenance and recall of information such as digits, letters, or words.

The development of the complex span task has been highly influenced by the Reading Span Test (RST) (Daneman & Carpenter, 1980). In the RST, individuals are asked to read an increasing series of sentences out loud and then to recall the last word of each. Thus, this task requires individuals to simultaneously process and store information and provides a closer approximation to the demands of complex cognitive tasks such as reading comprehension. Moreover, since this task requires some kind of attentional control, it can be considered a measure of the central executive in working memory (see Engle & Oransky, 1999; Whitney, Arnett, Driver, & Budd, 2001).

According to this view, a fruitful line of research has emerged devoted to explaining individual differences in complex cognitive tasks as a function of working memory span. However, span measures have also received a lot of criticism, mainly due to the nature of the relationship between processing and storage (Friedman & Miyake, 2004); but also because the nature of the processing task is sometimes very close to the task used as criteria.

There are, however, some important reasons to disagree with this criticism. Among others, it is worth mentioning that a similar pattern of correlations with high-level cognitive processes was obtained when the processing task consisted of solving tasks such as: arithmetical operations (Operation Span Test; Oakhill, Yuill, & Parkin, 1986; Turner & Engle, 1989); counting (Counting Span Test; Case, Kurland, & Goldberg, 1982); speaking (Speaking Span Test; Daneman & Green, 1986); rotating objects (Spatial Span Test; Shah & Miyake, 1996) or resolving analogies or anaphoric inferences (Garcia-Madruga et al., 2005, 2007; Gutiérrez et al., 2005). But also because working memory span predicts a wide range of very different complex cognitive tasks such as reasoning (Garcia-Madruga et al., 2005, 2007; Kyllonen & Christal, 1990), reading comprehension (Daneman & Merikle, 1996; Garcia-Madruga, Gárate, Elosúa, Luque, & Gutiérrez, 1997), fluid intelligence (Engle, Kane, & Tuholski, 1999); and, especially, tasks that imply the acquisition of new information (Engle, 1996).

Central Executive and Reading Comprehension

According to the main theoretical models in both fields, both reading comprehension and reasoning require keeping a great amount of information active, updating models, and also suppressing information that becomes irrelevant. All of these processes are related to working memory capacity, but also to the attentional control mechanism attributed to the central executive (Radvansky & Copeland, 2001; Garcia-Madruga et al., 2007). Thus, both reading comprehension and reasoning should be sensitive to working memory individual differences.

The relationship between working memory span and reading comprehension has been well established in the literature (Daneman & Carpenter, 1980; Daneman & Merikle, 1996; Just & Carpenter, 1992). However, in spite of these previous results, some other studies have not found the predicted correlations between WM span and Reading comprehension (Baddeley, Logie, Nimmo-Smith, & Brereton, 1985; Caplan & Waters, 1999; Light & Anderson, 1985), or they have found correlations between complex
span tasks no higher than those obtained with simple span tasks (implicating only the storage component) (see Engle, Carullo & Collins, 1991). Along this line, Radvansky & Copeland (2004) pointed out that working memory span might be a good measure of lower levels of comprehension (e.g., text level), but may not be so good at higher levels of comprehension (e.g., mental models). It might be due to the nature of the processing task in the RST that only requires people to read aloud in order to remember the last word of the sentence, and thus, it might be done in an automated way that doesn’t require people to really understand the sentence.

**Central Executive and Deductive Reasoning**

Correlations between working memory span and deductive reasoning are not as consistent as they are with reading comprehension. Some studies have found interference using the double-task procedure in which the primary task is a reasoning task and the secondary task loads one of the WM components: the central executive, the phonological loop, or the visuospatial sketchpad. In general, these studies show that the central executive plays an important role in both syllogistic (e.g., Gilhooly, 1998; Gilhooly, Logie, & Wynn, 2003) and conditional reasoning (e.g., De Neys, Schaeken, & D’Ydewalle, 2005; Toms, Warris, & Ward, 1993).

However, the part the central executive plays in reasoning has not been confirmed in other lines of studies that have tried to correlate working memory span (measured with Reading Span Task or other similar tasks) and reasoning. Some of these studies have been criticized because of the type of working memory measures used (Bara, Buccaneley, & Johnson-Laird, 1995; Kyllonen & Christal, 1990). Others (Markovits, Doyon, & Simoneau, 2002) have obtained correlations contrary to that expected in conditional reasoning. However, other studies, have shown more optimistic results, such as the study carried out by Barrouillet and Lecas (1999; Barrouillet, 1996), where a high and positive correlation was found between WM span and conditional reasoning in children using Case et al. (1982) counting span task; by Copeland and Radvansky (2004), who found that working memory span correlates with syllogistic reasoning using the operation span test, and also by García-Madruga et al. (2005), who found correlations between two new working memory span measures—based on resolving analogies and pronominal anaphora—and propositional reasoning.

In summary, the relationship between working memory and reading comprehension on the one hand, and working memory and reasoning on the other, has elucidated the crucial role of the Central Executive. However, the relationship between working memory and reading comprehension has been criticized due to the nature of the processing task and also due to the similarity between the processing task and the criteria tasks. Likewise, the relationship between working memory and reasoning seems to be highly dependent on the kind of task used both to measure WM and reasoning, the kind of reasoning studied (spatial, propositional or syllogistic), and the strategies people use to cope with the task.

In our opinion, these criticisms and contradictory results point to the necessity of taking into account the nature of the processing task used in Working Memory Span tasks, especially in the RST. This hypothesis has been tested in previous studies (García-Madruga et al., 2005, 2007; Gutiérrez et al., 2005) devoted to designing new working memory span measures—based on RST— in which the processing task called for an inferential decision (to resolve an analogy or an anaphora) and in which people had to recall the result of an inference instead of the last word of a sentence (as they do in the RST). Previous results showed that these new measures increased the predictive power of working memory with regard to propositional reasoning. Nevertheless, there might be an element of circularity when using measures of WM to predict reasoning ability when the measure itself calls for some kind of inference. Although the pronominal anaphora problems used in the test are very easy—and quite different from the propositional reasoning inferences investigated—one of the purposes of the development of a new measure of WM is to avoid this possible circularity. This was resolved empirically in a previous study (see Elosúa, Carriedo, & García-Madruga, 2009) in two independent ways: (a) by selecting easy anaphoric references that could be resolved by more than 95% of participants; and (b) by testing in an independent study whether performance on the anaphoric resolution task on its own did not correlate with comprehension and reasoning measures. As the results of this previous study showed that performance in anaphoric resolution didn’t correlate with text comprehension and reasoning, differences that we did find could not be interpreted as due to the similarity between the tasks.

In this context, the first purpose of the present study is to develop new measures of WM in which the demands of the processing task—in terms of attentional control—were increased. In order to do that, we designed a different memory span test based on semantic anaphoric resolution. The difference between this new measure and the anaphora test used by García-Madruga et al. (2007) is that the processing task in the new test requires solving pronominal anaphors based on semantic cues instead of those based on Morpho-Syntactic ones. In both cases, participants have to read aloud a sequence of sentences, infer the word referred to by the anaphoric link, and then remember that word. Because semantic processing increases attentional control demands, we expected this new Semantic Anaphora test to increase the demands of processing in regards to RST, and to a lesser extent in regards to the Morpho-Syntactic Anaphora test. Thus, our new measure introduces some
modifications in the processing component (a deeper processing), whilst maintaining constant the storage component (see Elósúa et al., 2009).

The second purpose is to test the validity of the different span tests –RST and the two anaphoric span tests– as they relate to propositional reasoning (as in the previous study carried out by García-Madruga et al., 2005, 2007), and extend these to reading comprehension. In the previous study (see García-Madruga et al., 2007), the Morpho-Syntactic anaphora test showed significant correlations with reasoning measures, higher than RST, but all the correlations were small (the highest was .36). The new test should provide a good and probably better measure of the central executive’s capacity for reasoning since the new secondary task increases the demands of attentional control. We will try to test this hypothesis for reasoning and to extend it to reading comprehension.

Specifically, if we consider reasoning, positive correlations should occur between WM measures and those responses deriving from a greater cognitive load. On the other hand, we predicted negative correlations among WM measures and the responses yielded by superficial matching strategies. We also expect that the Semantic Anaphora measure of working memory span should yield higher and more reliable correlations with inferential performance than the standard reading span measure (RST), and the Morpho-Syntactic Anaphora measure.

In the case of reading comprehension, positive correlations should occur between WM and text comprehension measures, particularly in responses that demand inferential comprehension. We also expect that Semantic Anaphora measures of working memory should yield higher and more reliable correlations with text comprehension performance than with either the standard reading span measure (RST) or the Morpho-Syntactic Anaphora measure.

Method

Participants

Sixty-four high school students 15 to 18 years old from a middle class school in Madrid, participated in the experiment ($M = 16.45; SD = 0.75$; 29 women and 35 men). Participants had not received training in logic and had not been previously tested in any experiment on working memory and reasoning. They were randomly assigned to two groups. One group carried out the Morpho-Syntactic version of Anaphora Working Memory Span Test, and the other group did the semantic one. Moreover, both groups of participants carried out the Reading Span Test, a reading comprehension test (Strategy Comprehension Test; TEC, Vidal-Abarca et al., 2007) and a reasoning task.

Tasks and materials

Working memory tasks

The Reading Span Test (Daneman & Carpenter, 1980; Spanish version; see Elósúa, Gutiérrez, García-Madruga, Luque, & Gárate, 1996) was used.

Anaphora Span Tests. As in the preliminary versions (García-Madruga et al., 2005, 2007; Gutiérrez et al., 2005), people have to remember a word that is the referent of a simple pronominal anaphora, having only two possible words from which to choose. In order to manipulate different cognitive load implied by anaphoric resolution we used two different kinds of pronominal anaphora based either on Morpho-Syntactic cues –gender concordance— or on semantic cues. In both cases, people have to read a sentence, and then choose the correct response between two words. In the Morpho-Syntactic version both words are semantically appropriate, but only the correct word maintains the gender concordance with the pronoun (in bold) of the sentence. On the other hand, in the semantic version, both words maintain the gender concordance with the pronoun (in bold) of the sentence, but only one is semantically appropriate.

Here is an example of the same sentence both in the Morpho-Syntactic version and in the semantic one:

Morpho-Syntactic version:

Eladio encouraged her very much to interpret such a demanding role

Semantic version:

Eladio encouraged her a lot to interpret such a demanding role

In this case, both words –“career” and “actress”– are grammatically appropriate, because the pronoun “la” (“her” in English) is used in Spanish to refer both feminine things and people, but only the word “actress” is the correct response because it matches the meaning of the sentence. The foil is grammatically appropriate but semantically inappropriate.

In order to ensure that the anaphora problems were really easy and to insure that the difficulty of each problem
was similar, we first carried out a normative study in which we asked a similar sample of participants to solve the same list of anaphora problems (see Elosúa et al. 2009). The percentage of correct responses for anaphora problems was really quite high, and very similar for both versions: 98% for Morpho-Syntactic problems and 97% for semantic ones. Therefore, we may assert that the difficulty of anaphora problems was minimal and similar for each one of the versions.

Both anaphoric span tasks consist of 42 inference problems that shift through three series of different levels of 2, 3, 4 and 5 problems each. After reading the instructions, participants trained by solving 3 trials of 2 problems each, and then began the experiment by doing a series of 2, 3, 4, and 5 sentences. Level 6 was removed because a pilot study showed that no participant was able to reach a consistent performance on level 5. The order in which the sentence and familiarity of the word-solutions appeared was also controlled (see Elosúa et al., 2009).

Comprehension task. The Strategy Comprehension Test (TEC, Vidal-Abarca et al., 2007) was used. This test is made up of two expository texts of 548 and 469 words each. After reading each text, participants have to answer 10 comprehension questions referred to as: explicit ideas, macro-ideas, anaphoric inferences, and inferences based on knowledge. The reliability of the test is .80 and its construct validity with regard to another comprehension test in Spanish (PROLEC-SE, Ramos & Cueto, 2000) is .72.

Reasoning task. As in the previous studies (García-Madruga et al., 2005, 2007) we used three sets of problems, one based on ‘if A then B’ conditionals, one based on ‘not A unless B’ conditionals, and one based on ‘A or B, or both’ disjunctions. Each set contained 12 problems. We used neutral content with information about people and locations (e.g., Óscar is in Granada or Nuria is in Sevilla, or both). Each problem consisted of a propositional premise (either conditional or disjunctive) and a categorical premise corresponding to one of the four logical possibilities: A (Modus Ponens), B (Affirmation of the Consequent), not-A (Denial of Antecedent) and not-B (Modus Tollens). Participants had to generate their own conclusions. The problems were in the participant’s native Spanish and referred to common Spanish proper names and well-known Spanish cities (see an example in Table 1).

### Procedure

First, participants were tested collectively in reading comprehension (TEC) and reasoning tasks. One week later, participants individually carried out the Reading Span Test and one of the versions of the Anaphoric Working Memory Test. The order of the working memory tasks was counterbalanced.

On the comprehension task people had to read each text for 9 minutes. After reading each text, they had 6 minutes to answer multiple choice comprehension questions.

On the reasoning task people had to write down the conclusion that necessarily follows, taking into account the propositional statement and the categorical premises. People did it at their own pace with no time limit.

Working Memory tasks were presented on a PC computer controlled by the E-prime 1.1. software program. The experimenter controlled the pace of presentation of the sentences or problems on the computer and also recorded the responses given by each participant (the series of words that he/she had to remember). The task ended at the level at which the participant failed on the recall task consecutively in the three series.

### Scoring procedure

Working memory tasks:

The scoring procedure of WM tests was developed by Elosúa, García-Madruga, Gutiérrez, Luque, and Gárata (1997) for RST. This procedure scores as the level at which participants are able to reach a minimal consistent performance (as an integer number), and adds decimals whenever correct performance increases at the same or at higher levels. In each of the three series in each level, participant performance can be (a) correct (accurate words, correct order), (b) half correct (accurate words, incorrect order), and (c) incorrect. The minimum consistent performance at each level is reached when a participant performs at least half of the maximum performance level, that is, either three series of words half correct, or one series of words correct, one half correct, and one incorrect.

Every performance better than the minimum consistent performance at the same or higher levels was scored by the addition of decimals. On the same level, each supplementary correct response would add two decimal points and each supplementary half-correct response one decimal point. For instance, minimum correct performance on the fourth level

---

**Table 1**

The three assertions and the four categorical premises used in the reasoning task

<table>
<thead>
<tr>
<th>Assertions</th>
</tr>
</thead>
<tbody>
<tr>
<td>If Óscar is in Granada then Nuria is in Sevilla.</td>
</tr>
<tr>
<td>Óscar is not in Granada unless Nuria is in Sevilla.</td>
</tr>
<tr>
<td>Óscar is in Granada or Nuria is in Sevilla, or both.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Categorical premises</th>
</tr>
</thead>
<tbody>
<tr>
<td>Óscar is in Granada   (MP)</td>
</tr>
<tr>
<td>Nuria is in Sevilla   (AC)</td>
</tr>
<tr>
<td>Óscar is not in Granada (DA)</td>
</tr>
<tr>
<td>Nuria is not in Sevilla (MT)</td>
</tr>
</tbody>
</table>

Note: MP = Modus Ponens AC = Affirmation of the Consequent DA = Denial of the Antecedent MT = Modus Tollens.
is 4, and maximum performance on the fourth level is 4.3; if a participant remembers accurately and in the correct order the three series of four words (level 4), his score will be 4.3.

On a higher level, a supplementary correct response would add 5 decimal points and a supplementary half-correct response 4 decimal points. For example, if the previous participant also remembers correctly a series of five words (level 5), his scoring will be 4.3 + .5 = 4.8. (see Gutiérrez et al., 2005; García-Madruga et al., 2007).

**Comprehension task**

As stated before, the comprehension test includes items referring to explicit main ideas, macro-ideas, anaphoric inferences, and inferences based on knowledge that occurs in the text. In order to analyze the relationship between working memory and comprehension more deeply, apart from the overall correct responses, we derived two additional indexes. One of them we refer to as “inferential comprehension” and the other as “textual comprehension,” defined as follows:

1. Inferential comprehension: this includes items that refer to inferences (anaphoric and based on knowledge)1.
2. Textual comprehension: this includes items that refer to the identification of explicit ideas in the text and macro-ideas, that is, the product of macrostructure elaboration.

**Reasoning task**

For ‘If A then B’ conditionals, the correct responses are the two valid inferences for Modus Ponens (‘A’, then ‘B’) and Modus Tollens (‘not-B’, then ‘not-A’), whereas for the Affirmation of Consequent (‘B’, then ‘A’) and the Negation of Antecedent (‘not-A’, then ‘not-B’) the correct responses are ‘No conclusion’. On the contrary, as previous studies have also shown, “not-A unless B” conditionals are usually interpreted as biconditionals (see Dancygier, 2002; Montolío, 1999), especially when their contents refer to threats and warnings (see Fillenbaum, 1986; Carriedo, García-Madruga, Gutiérrez y Moreno-Ríos (1999); García-Madruga et al., 2008). Therefore, we considered responses yielded by this biconditional reading of unless as correct; that is, the four kinds of inferences will be considered correct. Previous studies have shown that when people reason with “unless” conditionals, some make an odd “asymmetric” response due to a superficial “matching” strategy (see, Carriedo et al. (1999); García-Madruga, Moreno-Rios, Carriedo, Gutiérrez, & Johnson-Laird, 2001; García-Madruga, Carriedo, Moreno-Rios, Gutiérrez, & Schaeken, 2008; Schaeken, García-Madruga, & D’Ydewalle, 1997). Given, say, premises of the DA form: ‘not-A unless B’, ‘not-A’, they respond: ‘B’. Similarly, given premises of the AC form: ‘not-A unless B’, ‘B’, they respond: ‘not-A’.

An inclusive disjunction such as ‘A or B, or both’ requires the construction of three explicit models, as follows (see Table 1):

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>not-B</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>B</td>
<td>not-A</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Three kind of responses to ‘A or B, or both’ are particularly interesting for us (see a more detailed analysis in Garcia-Madruga et al., 2007): (a) After receiving the two affirmative categorical premises (A and B), some people may give an invalid ‘symmetric’ conclusion by applying some kind of superficial matching strategy. In this way, from ‘A’ they answer ‘B’ and from ‘B’ they answer ‘A’. (b) After receiving the same two affirmative categorical premises (A and B), some other people may give the valid correct responses to these inferences: “There is no valid conclusion” that demands the construction of two of the above models. (c) For the two negative categorical premises: ‘not-A’ and ‘not-B’, some reasoners would probably be able to construct the three models in order to reach the valid asymmetric conclusions: ‘B’ from ‘not-A’, and ‘A’ from ‘not-B’. Taking into account these previous results and studies (e.g., García-Madruga et al., 2007), we categorized responses given by participants as follows:

1) Correct responses:

- If then: Symmetric responses for MP (A therefore B) and MT (not-B therefore not-A) and “no conclusion” responses for AC and DA.
- Unless: Symmetric responses for the four inferences, MP (A therefore B), MT (not-B therefore not-A), AC (B therefore A) and DA (not-A therefore not-B).
- Disjunctions: “No conclusion” responses for categorical premises A and B; Asymmetric responses for not-A (not-A therefore B) and not-B (not-B therefore A).

2) Superficial erroneous responses:

- Unless: DA (not-A therefore B) and AC (B

---

1. The comprehension test includes items referred to as anaphoric inferences. However, these anaphors are not pronominal ones, and thus they differ from those used in the working memory tests. In order to avoid a criticism of circularity we analyze data with and without these items. Results did not vary.
therefore not-A) asymmetric responses.
– Disjunctions: Symmetric responses for A (A, therefore B) and B (B, therefore A).

Results and discussion

Descriptive analysis

Working Memory tasks
Table 2 shows descriptive results for Working Memory tasks.

There were no significant differences between two groups in RST scores (T-test: -.481; p = .632). Participant’s performance on diverse WM tests was significantly different. The mean score in the Reading Span Test ($M = 3.72; SD = 0.72$ words) was reliably higher than on both the Morpho-Syntactic Anaphora Test ($M = 3.39; SD = 0.70$ words; $T$-test: 2.75; $p < .005$) and the Semantic Anaphora Test ($M = 3.28; SD = 0.66$ words; $T$-test: 4.0; $p < .0002$). However, the difference between the two anaphor tests was not reliable ($T$-test: .66; $p = .51$).

Comprehension task

Two participants were excluded in this analysis because they performed at a rate clearly inferior to the mean performance exhibited by the rest of participants. As we can see in Table 3, the comprehension test was quite easy for participants. In fact means scores for each comprehension index is above 70% of correct responses. There were no significant differences between two groups in overall comprehension scores ($T$-test: -.056; $p = .955$).

Reasoning task

The same two participants excluded in reading comprehension were excluded in this analysis because they performed at a rate below 20% of the mean performance exhibited by the rest of participants. There were no significant differences between the two groups in overall correct responses ($T$-test: .918; $p = .362$). Table 4 presents the results for the three sorts of statements.

Results on reasoning tasks corroborated those obtained in previous studies (see for example, Garcia-Madruga et al., 2007).

Correlation analyses

As expected, Pearson correlations among RST and the other two WM measures were positive and significant: Reading Span and Morpho-Syntactic Anaphora Span: ($r = .68; p < .003$); Reading Span and Semantic Anaphora Span: ($r = .50; p < .003$).

On the comprehension measures (see Table 5), as expected, we found positive correlations between RST and all the comprehension measures, but these are only significant for the overall comprehension index. In the case of the Working Memory Semantic Anaphora measure,
### Table 4
*The percentages of responses for both conditional statements and inclusive disjunctions, as a function of inference type (N= 62)*

<table>
<thead>
<tr>
<th>Categorical Premise</th>
<th>A (MP)</th>
<th>B (AC)</th>
<th>not-A (DA)</th>
<th>not-B (MT)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Symmetric responses</td>
<td>95</td>
<td>92</td>
<td>69</td>
<td>69</td>
</tr>
<tr>
<td>No conclusion</td>
<td>2</td>
<td>6</td>
<td>30</td>
<td>20</td>
</tr>
<tr>
<td>Asymmetric responses</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>Other</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>If A then B</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Symmetric responses</td>
<td>87</td>
<td>84</td>
<td>69</td>
<td>80</td>
</tr>
<tr>
<td>No conclusion</td>
<td>3</td>
<td>5</td>
<td>16</td>
<td>7</td>
</tr>
<tr>
<td>Asymmetric responses</td>
<td>7</td>
<td>9</td>
<td>12</td>
<td>11</td>
</tr>
<tr>
<td>Other</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Not-A unless B</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Symmetric responses</td>
<td>19</td>
<td>19</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>No conclusion</td>
<td>64</td>
<td>61</td>
<td>52</td>
<td>56</td>
</tr>
<tr>
<td>Asymmetric responses</td>
<td>9</td>
<td>13</td>
<td>37</td>
<td>33</td>
</tr>
<tr>
<td>Other</td>
<td>8</td>
<td>7</td>
<td>5</td>
<td>4</td>
</tr>
</tbody>
</table>

Note: MP = Modus Ponens AC = Affirmation of the Consequent DA = Denial of the Antecedent MT = Modus Tollens. Correct responses are in bold.

### Table 5
*Correlations between different working memory and reading comprehension measures*

<table>
<thead>
<tr>
<th>Measures of WM</th>
<th>Inferential comprehension</th>
<th>Textual comprehension</th>
<th>Overall comprehension</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading span (RST) N= 62</td>
<td>.19</td>
<td>.19</td>
<td>.22*</td>
</tr>
<tr>
<td>Morpho-Syntactic Anaphora N=32</td>
<td>-.00</td>
<td>.04</td>
<td>-.01</td>
</tr>
<tr>
<td>Semantic Anaphora N=32</td>
<td>.37*</td>
<td>.30*</td>
<td>.39*</td>
</tr>
</tbody>
</table>

* p < .05

### Table 6
*Correlations between different working memory and reasoning measures*

<table>
<thead>
<tr>
<th>Measures of WM</th>
<th>Correct If then</th>
<th>Correct Unless</th>
<th>Correct Or</th>
<th>Overall Correct</th>
<th>Superficial Erroneous</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading Span (RST) N = 62</td>
<td>.23*</td>
<td>.07</td>
<td>.30*</td>
<td>.31**</td>
<td>-.22*</td>
</tr>
<tr>
<td>Morpho-Syntactic Anaphora. N = 30</td>
<td>.24</td>
<td>.05</td>
<td>.39*</td>
<td>.42**</td>
<td>-.29</td>
</tr>
<tr>
<td>Semantic Anaphora N = 32</td>
<td>.28</td>
<td>.25</td>
<td>.49**</td>
<td>50**</td>
<td>-.33*</td>
</tr>
</tbody>
</table>

** p < .01; * p < .05
the correlations with the diverse comprehension measures were all significant and higher for the three reading comprehension measures. Unexpectedly, the Morpho-Syntactic anaphora measure didn’t correlate with any of the comprehension measures. A possible explanation might be that this working memory task taps into some kind of superficial factors.

With regard to reasoning performance, as expected, we obtained positive significant correlations among all WM measures and correct “or”, as well as among all WM measures and overall correct responses for the three types of sentences: “if”, “unless” and “or” (see Table 6). Likewise, negative correlations were found between WM measures and “superficial erroneous responses,” which were yielded by superficial “matching” strategies that do not load WM. These correlations were significant in the case of RST and Semantic Anaphora and fell short of the standards levels of significance in the case of the Morpho-Syntactic anaphora test. Moreover, the correlations found for reasoning measures were higher for the Semantic Anaphora span test than for the Morpho-Syntactic anaphora span and the RST.

**Regression analyses**

In order to determine the predictive power of anaphoric working memory tasks, simple regression analyses were conducted on main comprehension and reasoning scores. Results showed that the semantic anaphora working memory score was the best predictor of both reasoning, which explains 25% of the variance for “overall correct”, $F(1,30) = 9.81; p < .05$, and reading comprehension, which explains 15% of the variance for overall comprehension, $F(1,30) = 5.21; p < .05$; see Tables 7 and 8). Thus, regression results corroborate the validity of the Semantic Anaphora measure, corroborating our hypothesis regarding the incremental attentional control of the Semantic Anaphora Span Test, and consequently, the more expected reliable correlations with inferential performance than the standard reading span measure (RST).

**General Discussion**

The main aim of this paper was to present a new Semantic Anaphora working memory test and examine, in comparison with other WM measures (RST and Morpho-Syntactic Anaphora), its relationship to reading comprehension and reasoning measures following the work by García-Madruga et al. (2007). In that study we designed a new working memory span measure for reasoning – based on RST – in which the processing task calls for an inferential decision (to resolve a pronominal anaphora based on Morpho-Syntactic cues) and people have to recall the result of an inference instead of the last word of a sentence (as they have to do in the RST). Previous results

<table>
<thead>
<tr>
<th>Table 7</th>
<th>Regression analysis: Comprehension measures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predictors</td>
<td>Overall comprehension</td>
</tr>
<tr>
<td></td>
<td>$R^2$</td>
</tr>
<tr>
<td>RST ($N = 60$)</td>
<td>.05</td>
</tr>
<tr>
<td>M-S Anaphora Test ($N = 29$)</td>
<td>.008</td>
</tr>
<tr>
<td>Semantic Anaphora Test ($N = 31$)</td>
<td>.15</td>
</tr>
<tr>
<td>$^*$ $p &lt; .05$</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 8</th>
<th>Regression analysis: Overall correct responses in Reasoning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predictors</td>
<td>Overall correct</td>
</tr>
<tr>
<td></td>
<td>$R^2$</td>
</tr>
<tr>
<td>RST ($N = 60$)</td>
<td>.09</td>
</tr>
<tr>
<td>Morphosyntactic Anaphora Test ($N = 29$)</td>
<td>.18</td>
</tr>
<tr>
<td>Semantic Anaphora Test ($N = 30$)</td>
<td>.25</td>
</tr>
<tr>
<td>$^*$ $p &lt; .05$; ** $p &lt; .01$</td>
<td></td>
</tr>
</tbody>
</table>
showed that this new measure increased the predictive power of working memory relative to propositional reasoning. In the present study, we refined the anaphoric measures trying to avoid a possible element of circularity by increasing the involvement of the processing task that manipulates the demands of attentional control. In order to do that, we designed a different memory span test based on the solving of pronominal anaphora based on semantic cues. Thus, our new measure introduces some modifications to the processing component, whilst maintaining constant the storage component (see Elósúa et al., 2009). Since the new secondary task increases the demands of attentional control, we expected that the new Semantic test will provide a good measure of the central executive’s capacity for reasoning and reading comprehension, probably better than the RST and the Morpho-Syntactic Anaphora test.

Results obtained confirm our hypothesis. First, as means show in Table 2, the new Semantic Anaphora Working Memory Span Tests is reliably more difficult than RST, although the difference with the other Anaphora measure is not reliable. These results confirm the difference in difficulty between the Reading Span Test and the Morpho-Syntactic Anaphora Test also found by García-Madruga et al. (2007), and they are extended to the Semantic Anaphora Test. The work of solving the pronominal-anaphora and selecting the word that has to be remembered on the Reading Span Test increases WM load and reduces the resources for storing the result. Our new WM Anaphora measure increases the attentional control demands of the three classical functions of EC: updating, inhibition and shifting, with respect to RST. In comparison with RST, solving semantic anaphora forces participants to update their semantic knowledge, to inhibit and discard alternative word solutions, and to shift between the selection and the recall task. In comparison with the Morpho-Syntactic Anaphora test, the Semantic test requires deep semantic processing that the Morpho-Syntactic Anaphora test does not; that is, it increases the updating work. Nevertheless, the three measures share the basic characteristic of being WM measures and, as expected, the inter-correlations between them are high and significant, particularly between RST and Morpho-Syntactic Anaphora. Nevertheless, the pattern of correlations observed with regard to reading comprehension and reasoning measures varied from one anaphora measure to the other.

Let’s first compare the correlations among RST and the Semantic Anaphora Working Memory Span Test with comprehension and reasoning tasks. As expected, the correlations between both measures of working memory and reading comprehension—particularly with inferential and overall comprehension—are positive and significant. However, what is especially interesting is that the correlations between the Semantic Anaphora Span Test and the different comprehension indices are higher and more significant than the correlations observed with RST. Moreover, this result is confirmed by regression analyses: the Semantic Anaphora Span Test alone is able to explain 15% of the variance of overall comprehension. These results are important because they confirm that this new semantic measure increases the demands of attentional control with regard to RST and reading comprehension. These results do not entirely converge with those by Radvansky and Copeland’s (2004). On the one hand, our results on RST confirm Radvansky and Copeland’s (2004) regarding the difficulty of WM measures in predicting deep level comprehension. On the other hand, our results seem to show that the Semantic Anaphora test may be more implicated in deep level comprehension: regression analyses show that Semantic Anaphora is the only WM measure that predicts inferential comprehension and overall comprehension measures.

The same pattern of correlations was observed in relation to reasoning measures. Both RST and the Semantic Anaphora Span Test showed the same pattern of correlations. The two measures correlate positive and significantly with correct reasoning responses. Moreover, both working memory tests correlate negative and significantly with superficial erroneous responses, which are automatic and do not consume cognitive resources (see Gilhooly, 1998). But interestingly, as expected, the Semantic Anaphora Span Test correlated with reasoning to a greater degree than the classical Reading Span Test did, and explains—with its own—25% of the variance in overall correct reasoning, whereas RST explains only 9% of the variance. Thus, the Semantic Anaphora Working Memory Test seems to be more sensitive to predicting reasoning performance than RST, especially for those responses that require the building, integration, and manipulation of semantic representations and hence, demand more cognitive resources (particularly in the control and supervisory functions of the central executive). These results are in accordance with previous results obtained by García-Madruga et al., (2005, 2007; Gutiérrez et al., 2005).

However, the case of the Morpho-Syntactic Anaphora Span Test is different and deserves special attention. Contrary to expected, this measure did not correlate with comprehension, although it did with reasoning in a lesser but similar way to the Semantic Anaphora Span Test: positive and significant correlations with overall correct responses and negative correlations with superficial erroneous responses.

A possible explanation for these results is that our participants carried out the processing task enclosed in the Morpho-Syntactic Span Test in a highly superficial way, by just matching the pronoun of the sentence, in bold, with the grammatical gender of the word chosen, without considering the meaning of the phrases. This superficial nature of pronominal anaphora resolution in the Morpho-Syntactic test might explain why this measure does not correlate with comprehension measures, which are in effect indices of deeper understanding. In other words, the updating component of this task is probably reduced...
to the grammar knowledge of the participants without any comprehension of the meaning of phrases or updating of semantic knowledge. This hypothesis does not affect the ability of the Morpho-Syntactic working memory task to predict reasoning performance, since reasoning is affected not only by Type 2 explicit, semantic and controlled processes but also by Type 1 implicit, superficial and automated processes (Evans, 2006, 2008; García-Madruga et al., 2007; Stanovich, 1999). Obviously, as the comprehension measures correlate with RST, albeit in a small yet significant way, this explanation requires that the processing part of the RST be done more deeply than it is in the Morpho-Syntactic Anaphora test. This is not unlikely since the task of participants in RST requires only reading aloud a short sentence, and this is done by most of readers with some access to the meaning of the sentences. In any case, this hypothesis is in need of further empirical work.

We might propose another argument to explain how two similar working memory span tests, as measured with two anaphora measures, show such a different pattern in their correlation with comprehension. Although we think that additional research is needed to solve this question, a preliminary answer might be—as Baddeley (2007, p. 3) suggests— that working memory span is influenced both by the phonological loop and by the central executive, but also that its capacity to predict complex cognitive task relies only on its function as central executive. The similar difficulty the two anaphora measures exhibit does not mean that both measures load the CE in a similar way. In the semantic anaphora task, the work of updating semantic knowledge in order to solve the pronominal anaphora and select the correct word clearly demands more attentional resources than does repeatedly updating the Morpho-Syntactic automated knowledge when matching the gender of the pronoun and the noun.

A main result from our study is that the increase of attentional control in the processing task (as in the semantic anaphora span test) allows one to explain performance in both reading comprehension and reasoning better. This result appears to contradict those obtained by Lépine, Barrouillet and Camos (2005) showing that the complexity of the processing task and the strategies people use do not contribute to the predictive value of the working memory task. The comparison between the predictive power of the semantic anaphora span test and RST allows us to conclude that the more demanding the processing task is, the better it explains performance—not only in reasoning but also in deeper comprehension. In other words, our results do not question that the fundamental capacity of WM is attentional, as these authors argue; it just shows that attention (demands of processing) can be manipulated by different means: by controlling time—as they did— or by manipulating the difficulty of task, as we have done.

Taken together, we think that these results suggest that even though performance on complex span tasks is common in that its measurement points to a domain-free ability to control attention (as suggested by Barrett, Tugade, & Engle, 2004; Conway, Cowan, Bunting, Therriault, & Minkoff, 2002; Kane & Engle, 2002; Lépine et al., 2005), it may be also influenced by domain specific processing competencies. In this same vein, Just and Carpenter (2002, p. 172) proposed that it is likely that working memory is a large set of processing resources, only a subset of which is used for a given task domain.

The most relevant result of this study is that our new working memory measure, the Semantic Anaphora measure, correlates higher and more reliably with correct text comprehension and reasoning, and superficial reasoning responses. This finding is enhanced by regression analyses that show the Semantic Anaphora was the best predictor of both reading comprehension and reasoning. Our results follow the same line of work in other studies (Cain, Oakhill, & Bryant, 2004; Cain, Oakhill, & Lemmon, 2004; Daneman & Hannon, 2001; Saito & Miyake, 2004) that highlight the necessity of treating the working memory measure more in depth. One way of doing so is by designing new measures that increase the attentional demands of the Central Executive, and as our results confirm, the Semantic Anaphora Span Test achieves this aim.

References


Cain, K., Oakhill, J., & Bryant, P. (2004). Children’s reading comprehension ability: Concurrent prediction by working memory, verbal ability, and component skills. Journal of


