Behavior Research Methods
2008, 40 (1), 46-54
doi: 10.3758/BRM.40.1.46

External validation of the computerized, group administrable adaptation of the "operation span task"

JOSÉ L. PARDO-VÁZQUEZ AND JOSÉ FERNÁNDEZ-REY
Universidad de Santiago de Compostela, Santiago de Compostela, Spain

One of the most widely used tasks for measuring working memory capacity is the *operation span task* (OSPAN; Turner & Engle, 1989). This task has almost always been applied individually, and stimuli presentation is controlled by the experimenter. Recently, De Neys, d'Ydewalle, Schaeken, and Vos (2002) improved the administration procedure by designing an automated, group-administrable version of the task (GOSPAN). They found GOSPAN to be reliable, and they also provided evidence on its validity (a significant positive correlation between GOSPAN and OSPAN scores). However, an external test of GOSPAN validity is still lacking. In this work, we present such a validation for the automated version, when the task is administered both individually (Experiment 1) and to groups (Experiment 2). There are abundant previous data on the relation between working memory capacity and reading comprehension. In this work, this relation is studied using an automated OSPAN version to measure working memory capacity. Given that our results are similar to those found using the original OSPAN, our data support the external validity of the automated version of the task. We also tested the reliability of the task and found high internal consistency in both experiments.

Although a number of different models and definitions have been proposed for working memory (WM; Miyake & Shah, 1999), it is generally accepted as being a limited capacity, complex mnemonic system for the simultaneous processing and storage of information (Andrade, 2001; Ashcraft, 2002; Baddeley & Hitch, 1974).

WM is one of the components of the cognitive system that has aroused the most interest in cognitive psychology and the neurosciences in general, fundamentally due to the abundance of empirical evidence that has been found regarding its implication in higher cognitive functions, such as reading comprehension (Bornkessel, Fiebach, & Friederici, 2004; Conway & Engle, 1996; Daneman & Carpenter, 1980; Daneman & Merikle, 1996; Engle, Cantor, & Carullo, 1992; Engle, Carullo, & Collins, 1991; Engle & Conway, 1998; Friedman & Miyake, 2004; Turner & Engle, 1989), reasoning (Capon, Handley, & Dennis, 2003; De Neys, 2006; De Neys & Dieussaert, 2005; De Neys, Schaeken, & d'Ydewalle, 2005; De Neys & Verschueren, 2006; Kyllonen & Christal, 1990), writing (Hoskyn & Swanson, 2003), and learning (Kyllonen & Stephens, 1990; Unsworth & Engle, 2005). To a great degree, this evidence originates from the study of individual differences, which has been carried out by following the correlational and the quasiexperimental strategies. The first is based on the estimation of the correlations between WM capacity (WMC) and performance on different cognitive tasks: If WM is involved in carrying out a cognitive activity, a positive correlation between WMC and performance of that activity will be observed (Capon et al., 2003; Colom, Flores-Mendoza, & Rebollo, 2003; Conway & Engle, 1996; Dougherty & Hunter, 2003; Engle et al., 1992; Gilhooly, Wynn, Phillips, Logie, & Della Sala, 2002; Kane et al., 2004). The quasiexperimental strategy is based on the comparison of extreme groups in terms of WMC in various cognitive tasks: If WM is involved in the execution of a cognitive task, significant differences between participants with high and low WMC will be observed in the performance of this activity (Bleckley, Durso, Crutchfield, Engle, & Khanna, 2003; Brumback, Low, Gratton, & Fabiani, 2005; Bunting, Conway, & Heitz, 2004; Conway & Engle, 1994; De Neys et al., 2005; Kane & Engle, 2000; Rosen & Engle, 1997; Unsworth & Engle, 2005; Unsworth, Schrock, & Engle, 2004; Watson, Bunting, Poole, & Conway, 2005).

A number of complex span tasks have been designed for measuring WMC; these tasks include a storage component and a processing one, so that both of the WM functions proposed by Baddeley and Hitch (1974) are involved in their execution.

The first task of this type, known as the reading span test, was proposed by Daneman and Carpenter (1980). For the processing component, a set of sentences is given, which the participants must read and verify; for the storage component, at the end of the set, they are asked to remember the last word of each of the sentences. The sets are made up of a variable number of sentences (normally two to six), and several sets of each size are shown. WM span

J. L. Pardo-Vázquez, josepar@usc.es

or WMC is defined as the amount of information recalled. The reading span test has been highly useful in exploring the role of WM in different cognitive functions (e.g., Capon et al., 2003; Daneman & Merikle, 1996). Nevertheless, this task has been criticized, both theoretically and due to its psychometric characteristics. When participants' performance on cognitive tasks that rely on verbal processing are predicted, a correlation between the reading span test and these tasks should be interpreted carefully. The fact that the predictive task (reading span test) involves many of the operations required to carry out the predicted tasks (e.g., reading comprehension, verbal reasoning, etc.) could inflate correlations, and this has given rise to significant doubts as to the origin of these correlations (Waters & Caplan, 1996). This is an especially relevant limitation, given that a lot of human cognitive functions rely, at least in part, on verbal processing. Regarding the psychometric characteristics of the reading span test, Waters and Caplan reported low temporal stability (test–retest) and low internal consistency (low correlation between the scores obtained for each of the presentations of the sets of each size), and they concluded that this task does not supply reliable data on WMC. However, the reading span test has proved to be reliable in other experiments (e.g., Engle, Tuholski, Laughlin, & Conway, 1999; Kane et al., 2004). A few years after the publication of Daneman and Carpenter's study, Turner and Engle (1989) designed a new complex span task denominated the *operation span task* (OSPAN),[1] which involves remembering words (storage component) while solving arithmetic operations (processing component).

The OSPAN task has two significant advantages over the reading span test. Since the processing component does not include sentence reading and comprehension, the overlap between the predicting and the predicted verbal tasks is significantly reduced. Thus, the OSPAN task is more suitable for studying the implication of WM in those cognitive functions that rely on verbal processing. Both the internal consistency and the temporal stability of OSPAN are higher than those of the reading span test (Klein & Fiss, 1999; Turner & Engle, 1989); hence, it can be considered to give a more reliable WMC index. Thanks to these advantages, OSPAN has become one of the most widely used WMC tasks (Beaman, 2004; Bleckley et al., 2003; Brumback et al., 2005; De Neys, d'Ydewalle, Schaeken, & Vos, 2002; Kane & Engle, 2000; Kane et al., 2004).

The OSPAN task requires the experimenter to control the presentation of stimuli, which gives rise to two important limitations: (1) The presence of the experimenter may influence the participants' performance, leading to a bias in the score, and (2) the administration procedure is highly demanding on the experimenter in terms of time and attention (Unsworth, Heitz, Schrock, & Engle, 2005). Because it is difficult to apply this task to groups (Unsworth et al., 2005), it has almost always been applied individually.

A version of the OSPAN task that, while conserving its advantages, does not require the continuous presence of the experimenter and may be conducted by the participants themselves would be of enormous benefit (De Neys et al., 2002; Unsworth et al., 2005). Furthermore, since the study of individual differences requires the use of very large samples, a significant amount of time would be saved if this task could be applied to groups, instead of individually (De Neys et al., 2002; Unsworth et al., 2005). Recently, an automated version of OSPAN has been published (Unsworth et al., 2005). This task seems to be a valid and reliable WMC task, but, as the authors admit, it is not strictly an OSPAN version, given that the storage component is different from that in the original version and that the task involves recognition, instead of recall. The relevance of these differences is uncertain and should be experimentally explored.

De Neys et al. (2002) designed GOSPAN, an automated (computerized), self-controllable, and group-administrable version of OSPAN. The procedure employed by De Neys et al. (2002) makes it possible to adjust the presentation time of the operation to the abilities of each participant.

Moreover, this procedure lends a certain degree of control over the use of strategies for rehearsing the words while the operations are being processed. The reaction times (RTs) to the operations are recorded, so that it is possible to identify and discard those participants who use part of the processing time to rehearse the words. The words are presented very briefly, so it is unlikely that the participants will be able to use this time to rehearse.

The procedure used by De Neys et al. (2002) also allows the experimenter to prevent participants from noting down the words as they are presented, because he or she is present while the groups of participants perform the task and can check that nobody notes down the words before being cued to do so. Thus, the GOSPAN task retains the majority of the advantages of the individual administration controlled by the experimenter, at the same time as it avoids its limitations.

The authors have found that the GOSPAN scores are reliable, with a higher internal consistency than the original, and have provided some results that support its validity: (1) They applied both GOSPAN and original OSPAN tasks to the same sample and found a significant positive correlation between both WMC scores ($r = .50$ and .70 when corrected for attenuation; De Neys et al., 2002), which would seem to signify that both tasks measure the same underlying construct; and (2) they studied the role of WM in a variety of reasoning tasks, using GOSPAN scores as a WMC index, finding that reasoning is related to GOSPAN performance (De Neys, 2006; De Neys & Dieussaert, 2005; De Neys et al., 2005; De Neys & Verschueren, 2006). Supposing that WM plays a role in reasoning task performance, these results may indicate that the GOSPAN task is a valid WMC measure. This assumption has been supported by previous studies showing a positive correlation between WMC and reasoning ability (e.g., Capon et al., 2003; Kyllonen & Christal, 1990). However, there are no data on the correlation between the traditional OSPAN task and the reasoning tasks used by De Neys and colleagues, so it is not possible to carry out a direct comparison between the automated version and the original task. To ensure that the automated, group-administrable version is a valid WMC task, an external validation allowing this comparison is necessary.

In this work we present SGOSPAN, a GOSPAN-based complex span task (employing Spanish[2] words for the storage component), with certain modifications with respect to the GOSPAN task in the application procedure and in the choice of materials. We test the reliability and validity of the task, applied both individually (Experiment 1) and to groups (Experiment 2). Its internal consistency is calculated as an indicator of reliability. An external validity test is carried out, thus complementing the evidence supplied by De Neys and colleagues (De Neys, 2006; De Neys & Dieussaert, 2005; De Neys et al., 2002; De Neys et al., 2005; De Neys & Verschueren, 2006): Given that numerous studies have shown the implication of the WM in reading comprehension, if SGOSPAN satisfactorily measures WMC, we would expect the scores in this task to be related to performance in a reading comprehension task (RCT).

The basic modification of the task that we propose is that, in SGOSPAN, participants do not know where they have to write down the words until the end of the set; thus, they are prevented from noting them down as they are presented. The aim of this feature is to allow individual administration without the presence of an experimenter. Furthermore, unlike in the GOSPAN task, in SGOSPAN the participants must read the operations and words while silently mouthing them; the aim of this instruction is to minimize the possibility of the participants' going over the words, given that mouthing the stimuli will interfere with word rehearsing (Beaman, 2004; Engle & Kane, 2004). SGOSPAN also has other minor modifications: (1) In the choice of material, we have attempted to maximize the equivalence between the stimuli, setting stricter criteria than those for GOSPAN in the construction of operations and controlling the words for imageability and familiarity, as well as the size and frequency of use; and (2) given that responding to the operations requires an association between the decision and one of the mouse buttons (correct–left-click, incorrect–right-click), a training block was introduced in which participants must solve operations alone and, thus, familiarize themselves with the type of answer they must supply (this training block has also been introduced in Unsworth et al., 2005).

EXPERIMENT 1

In the first experiment, we analyze the reliability and validity of SGOSPAN applied individually. The internal consistency of the task is used as an index of reliability. SGOSPAN will be considered reliable if its internal consistency is similar to that of previous versions of OSPAN.

There are theoretical reasons for considering that SGOSPAN will provide a valid index of WMC, since it has been designed with the principal features of complex span tasks in mind (Miyake, 2001). Moreover, De Neys and colleagues (De Neys, 2006; De Neys & Dieussaert, 2005; De Neys et al., 2002; De Neys et al., 2005; De Neys & Verschueren, 2006) have obtained some results that support the validity of the automated version (GOSPAN), but, as has been indicated, an external validation of the automated OSPAN version is still lacking. The aim of this experiment is to provide such a validation for the task when administered individually. In addition to the SGOSPAN task, participants are asked to carry out a reading comprehension test. Since there is abundant evidence in favor of WM's playing a role in this activity (Conway & Engle, 1996; Daneman & Carpenter, 1980; Daneman & Merikle, 1996; Engle et al., 1992; Engle et al., 1991; Turner & Engle, 1989), if SGOSPAN is to be considered a valid measurement of the WMC, the predictions are obvious: (1) There will be a significant positive correlation between the scores for this amplitude test and performance on the RCT, and (2) this correlation will be similar to those found in previous studies in which OSPAN or similar complex span tasks have been used.

Method

Participants

The initial sample for this experiment comprised 61 undergraduate students, between 18 and 25 years of age (M = 20.95, SD = 2.26).

Materials and Stimuli

SGOSPAN. Eighty-six arithmetical operations that comprised two terms (or *suboperations*) and a result were constructed. The first term (given in brackets) could be a multiplication or division operation for two integers between 1 and 9. The result of this first term was always a positive integer between 1 and 20, to which another positive integer between 1 and 9 (given outside the brackets) had to be added or subtracted. The result of the second term was never higher than 20. In half of the operations, the result proposed was correct [e.g., (8/8) + 2 = 3], and in the other half, it was incorrect [e.g., (8/8) + 2 = 4]; in the latter case, the difference between the correct result and the proposed one was never higher than 2. Ten correct and 10 incorrect operations were selected for use in the initial training phase, and the remaining 66 were assigned to the second training phase and to the experimental block.

Sixty-six high-frequency (M = 176.05, SD = 10.5; range, 50–941), medium-to-high imaginability (M = 5.58, SD = 0.82; range, 3–6.62), and medium-to-high familiarity (M = 5.48, SD = 0.76; range, 3.28–6.71) two-syllable Spanish words were used. Word selection was carried out in accordance with the norms proposed by Algarabel (1996).

Sets of two, three, four, five, and six operation–word strings were formed, including 3 sets of each size (a total of 15 sets for the experimental phase). These strings were constructed by randomly selecting 30 correct operations, 30 incorrect operations, and 60 words and then randomly matching each operation with a word. The operation–word strings were also assigned randomly to each of the sets.

For the second training phase, three additional sets of two strings were also included. The remaining six operations and six words were used to make up these sets, following the method described for the experimental block for constructing the strings and the sets. The stimuli were presented in black on a white background, using 20-point Arial font.

RCT. This task consisted of 15 brief texts and 4 statements relative to each of them (giving a total of 60 statements), which were taken from the Spanish adaptation of the GMA Verbal (GMA–V) Test, Medium and Advanced Level Assessment (Blinkhorn, 1985/1999). An example text and example statements were also included.

Procedure

Tasks were carried out in individual cabins, in one single session, and with an approximate duration of 1 h. Thirty-one participants performed the SGOSPAN task first, and 30 performed the RCT first. The tasks were presented using SuperLab Pro Version 2.0. Responses to operation verification in SGOSPAN and to the choice of alternatives in the RCT were recorded, along with the corresponding RTs, using the aforementioned program. Responses to the SGOSPAN storage component were recorded on response sheets that were specially designed for this task.

SGOSPAN. A simple arithmetic operation, comprising two sub-operations and a result that might be correct or incorrect, was presented on a computer monitor. The participants had to read the operation while silently mouthing it, resolve it, and verify the result that was proposed; if they considered it to be correct, they clicked the left mouse button or, in the opposite case, the right button. The operation disappeared immediately after the participants had responded, and a word was presented in its place for 800 msec. At the end of this period, another operation was presented and, after the participants had responded, a new word.

This succession of operation–word strings was repeated until a sound indicated the end of the set. The sound had a duration of 900 msec and was presented via two speakers located on either side of the monitor. On hearing the sound, the participants had to note down the words that they recalled from the last set presented on the response sheet, attempting to do so in the same order that they were given. The response sheet was composed of 16 boxes (4 rows × 4 columns), in 15 of which the participants had to note down the words. The participants did not know beforehand in which box they had to note down the words corresponding to the set that they were currently doing. Immediately after the sound signaling the end of the set, a representation of the response sheet was shown on-screen, and the box in which they had to write the words was indicated in blue. Once the participants had noted down all the recorded stimuli (there was no time limit), they had to click a mouse button to move on to the following set (see Figure 1).

Sets of different sizes were presented (two, three, four, five, and six operation–word strings), with three sets of each size being given (5 sizes × 3 sets = total of 15 sets). The sets were presented in random order, with the same order being used for all the participants. An instruction display was introduced before each set, reminding the participants that they had to silently mouth the stimuli.

Before starting the 15 experimental sets, the participants carried out two training blocks. In the first of them, only the task-processing component was carried out. Twenty training operations were sequentially presented on the computer screen. The participants had to read them while silently mouthing them, resolve them, and decide as quickly as possible whether the proposed result was correct or not. When they considered the result to be correct, they had to click the left mouse button; otherwise, they had to click the right button.

They were then given a second training block, made up of three sets of two operation–word strings. Once they had concluded these sets, the participants had to verify that they had understood the task by comparing their responses with the training template. When the participants had suitably completed this training block, they started with the experimental block. The WMC score for each participant was the sum of the words remembered in those sets that were recalled completely and in the correct order (absolute span, ABSPAN). The average RT was also calculated, along with the percentage of correct responses to the operations (OPRT and HITS, respectively).

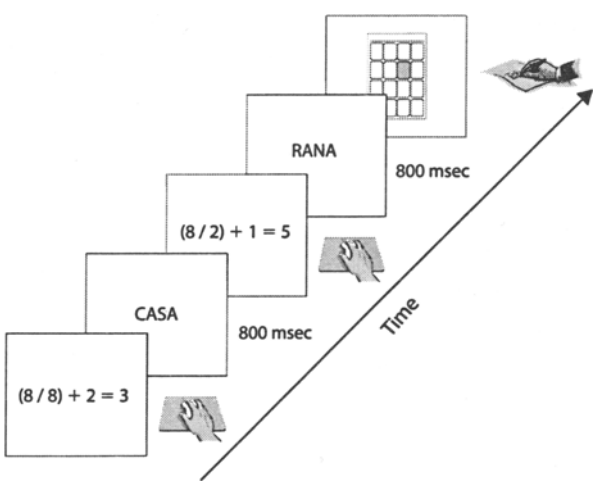

Figure 1. Sequence of events that occurred when a set of two operation–word strings was presented.

RCT. The instructions included in the GMA–V, adapted to the administration procedure, were presented on-screen. Before starting with the 15 experimental texts, the participants read a training text and gave the four corresponding truth judgments. In the example statements, but not in the experimental ones, feedback was given on the suitability of the alternative chosen.

Each text was shown on the computer screen for a maximum of 90 sec. If the participants finished reading the text before the end of this period, they could press any key. After 90 sec (or after a key had been pressed), a statement related to the content of the text appeared on-screen. The information required to make a judgment regarding the statement could appear in the text either explicitly or implicitly, due to which, in some cases, the participants had to make inferences on the content. Three responses were supplied with the statement, and the keys corresponding to each alternative were indicated: A–true, S–false, and D–no way of knowing. After the participants had made their choice, the remaining three statements were then presented sequentially. There was no time limit for responding, and once the participants had verified the four statements, they went on to the following text. The score for the RCT (COMP) was the sum of the correct truth judgments made.

Results

Description of the Definitive Sample

In order to be included in the analysis, the participants had to comply with two previously established conditions: (1) to correctly resolve 80% of the SGOSPAN operations and (2) to resolve these operations, on average, in a time that was lower than the sample mean plus 2.5 standard deviations. One participant was excluded (1.64% of the initial sample) after having exceeded the maximum time permitted for the SGOSPAN operations (M = 4,915 msec, SD = 1,673; OPRT maximum = 9,098 msec). The definitive sample comprised 60 participants, between 18 and 25 years of age (M = 20.92, SD = 2.29).

SGOSPAN

The mean ABSPAN score was 26.47 (SD = 9.41; range, 10–56). Applying a χ^2 test, no significant differences were found between the ABSPAN score distribution and the normal distribution [$\chi^2(3) = 4.73$, $p > .1$]. The average RT for the SGOSPAN operations was 4,839 msec (SD = 1,578), and the mean for correct responses was 95.31% (SD = 3.09%). It may have been the case that some of the participants obtaining the highest span scores used part of the operation-processing time to rehearse the words. In order to rule out this possibility, we calculated the correlation between OPRT and ABSPAN: If rehearsing played a significant role in the WMC scores, we should find a significant, positive correlation between the two variables. This was not the case with our results (r = .18, $p > .1$). It may also have been the case that some of the participants made less of an effort in the processing component, neglecting the resolution of the operations; by eliminating those individuals that did not correctly resolve 80% of the operations, we avoided, to a certain degree, the

use of this strategy. Nevertheless, even with those who resolved the majority of the operations, there may have been performance-related differences in the processing component. If this were the case, we would expect to find a negative correlation between the span scores (ABSPAN) and the percentage of operations resolved (HITS). This did not occur in our data, since we found a positive correlation that approached significance ($r = .23, p = .08$).

SGOSPAN Reliability

Three span subscores were calculated for each participant, corresponding to the first, second, and third presentations of the sets of each size. The Cronbach's alpha coefficient for these three measurements was .73.

SGOSPAN–Reading-Comprehension Correlation

The Pearson product–moment correlation was calculated between WM span (ABSPAN) and performance on the RCT (COMP), and a significant positive correlation was found ($r = .31, p < .02, n = 60$).

Discussion

ABSPAN scores do not seem to have depended on the use of rehearsing strategies while operations were processed (the participants achieving the highest ABSPAN scores did not spend more time in resolving the operations). Neither do these scores seem to have been related to lower performance for the processing component (the participants achieving the highest ABSPAN scores did not make more errors).

The mean span score found with SGOSPAN (26.47) is lower than that found by De Neys et al. (2002) with GOSPAN (31.33) and is higher than habitual mean scores found using OSPAN (e.g., 21.55; Beaman, 2004). These differences may be related to the fact that in SGOSPAN, participants are asked to read while silently mouthing both operations and words. In OSPAN, participants must read aloud, whereas in GOSPAN, they have to do so silently and without mouthing the stimuli. It has been pointed out that reading aloud can interfere with word recall (Beaman, 2004; De Neys et al., 2002; Engle & Kane, 2004); thus, it is not surprising that OSPAN scores were lower than GOSPAN and SGOSPAN ones. As De Neys et al. (2002) pointed out, it is important to bear in mind that the recall advantage related to silent reading affects all participants (both low- and high-span ones) in a similar way, so it does not affect the validity of WMC scores (De Neys et al., 2002). This point has been confirmed by Beaman, who administered two OSPAN tasks: the original one and a new version in which participants had to read the stimuli silently. He compared the scores obtained with both tasks and found that the span scores obtained for the silent reading version were higher than those obtained for the original task (30 and 21.55, respectively). He also found a high positive correlation between the scores obtained with the two OSPAN tasks. Although we did not test this point, we have no reason to expect the mouthing of the stimuli to affect SGOSPAN validity. Our data seem to show that, in SGOSPAN, mouthing operations and words affects recall more than does silent reading in GOSPAN, but less than does reading aloud in OSPAN.

The internal consistency of SGOSPAN ($\alpha = .73$) is similar to that of GOSPAN ($\alpha = .74$; De Neys et al., 2002) and even higher than the .69 that Engle et al. (1999) reported for the OSPAN task. We thus consider SGOSPAN to be a task that is at least as reliable as the previous versions of OSPAN.

The predictions related to the validity of the task did hold true: (1) A significant positive correlation (.31) was found between ABSPAN and performance on the RCT, and (2) the degree of correlation observed lies within the range of correlations reported in other studies in which tasks similar to SGOSPAN were used (between .30 and .48, according to the extensive meta-analysis published by Daneman & Merikle, 1996). Thus, the results support the validity of SGOSPAN as a WMC task.

The results of this study show that, applied individually, SGOSPAN supplies a reliable, valid index of WMC. In this sense, we believe that SGOSPAN is an interesting contribution, since, unlike the majority of previous versions, the presence of the experimenter while the participant is carrying out the task is not required. Nevertheless, given that the study of individual differences requires the evaluation of a large number of participants, an external validation of SGOSPAN administered to groups should be an especially interesting improvement.

EXPERIMENT 2

In this experiment, we analyze the reliability and validity of SGOSPAN administered to groups, once again taking the task's internal consistency and its capacity to predict performance on a reading comprehension test as the criteria for reliability and validity, respectively.

Besides SGOSPAN, we used an RCT similar to the one used by Daneman and Carpenter (1980) to analyze the role of WM in the recall of facts and pronominal references presented in a text. Recalling facts reflects the comprehension of the whole text, whereas the pronominal referents subscore clearly implies both WM components (maintain the referent while reading the text). We decided to design a different comprehension test from the one used in Experiment 1, since the 15 texts were excessively demanding on the participants and this could lead to fatigue-related biases. This new task is denominated RCT′, in order to distinguish it from the one used in Experiment 1. If SGOSPAN is valid when administered to groups, the predictions are the same as those in Experiment 1: (1) There will be a significant positive correlation between the SGOSPAN scores and performance on the RCT, and (2) this correlation will be similar to those found in previous studies.

Method

Participants

The tasks were carried out by a total of 149 undergraduate students, between 19 and 33 years of age ($M = 20.54$, $SD = 2.31$). None of them had participated in Experiment 1.

Materials and Stimuli

For the SGOSPAN, the 86 operations from Experiment 1 were used, and, adhering to the same criteria, 10 correct and 10 incorrect

operations were added to the training block. The 66 words from Experiment 1 were also used for the storage component.

For RCT′, 11 short texts relating short stories were written. In the last phrase of each text, a pronoun referring to a previously presented name was included. Four questions were drawn up for each text. The first of them asked about the pronominal referent. The other three referred to information explicitly included in the text.

Procedure

The tasks were carried out in groups of 18 or 19 participants, in one single session with an approximate duration of 1 h. Half of the groups performed the SGOSPAN task first, and the other half performed RCT′ first. Both tasks were presented by computer using SuperLab Pro Version 2.0. Responses to operation verification in SGOSPAN, along with the corresponding RTs, were recorded with the aforementioned program. Responses to the SGOSPAN storage component and to RCT′ were recorded on response sheets.

The SGOSPAN administration procedure was similar to that used in Experiment 1, although the number of operations presented in the initial training phase was modified (40 instead of 20) and the representation of the response sheet was used as notification that the test was finished (the sound used in Experiment 1 was eliminated). The ABSPAN score for each participant was calculated as the index of WMC. Two comprehension indices were calculated for RCT′: recall of the pronominal referents (PRON) and recall of the facts presented in the text (FACTS).

Results

Exclusion Criteria and Description of the Definitive Sample

The same exclusion criteria as those used in Experiment 1 were employed. A total of 10 participants (6.71% of the initial sample) were excluded: 4 for failing to reach the 80% right-answer threshold and 6 for having exceeded the maximum time allowed for operations (M = 4,470 msec, SD = 1,141 msec; maximum RT = 7,998 msec). The definitive sample comprised 139 participants, between 19 and 33 years of age (M = 20.5, SD = 2.21).

SGOSPAN

The mean ABSPAN score was 28.21 (SD = 9.94; range, 9–55). No significant differences were found between the ABSPAN score distribution and the normal distribution [$\chi^2(9) = 14.45, p > .1$]. The mean RT for the SGOSPAN operations was 4,264 msec (SD = 1,055). As in Experiment 1, the correlation between ABSPAN and OPRT was calculated, and this was not significant ($r = -.07, p > .1$), which suggests that the procedure did not allow individuals to use processing time to rehearse their responses. Furthermore, the ABSPAN–HITS correlation was positive and significant ($r = .22, p < .01$), which would seem to indicate that the high WMC scores were not due to reduced effort in the correct resolution of operations.

Reliability of SGOSPAN

Three absolute span subscores were calculated, corresponding to the first, second, and third presentations of the sets of each size. The Cronbach's alpha coefficient for these three measurements was .69.

Correlational Strategy

ABSPAN correlated positively with the recall of facts presented in the texts ($r = .28, p < .002, n = 139$). Nevertheless, no significant correlation was found between ABSPAN and PRON ($r = .10, p > .1, n = 139$).

Table 1
Descriptive Statistics for Low-Span and High-Span Groups

	Low Span		High Span	
	M	*SD*	*M*	*SD*
ABSPAN**	16.42	3.35	41.54	6.70
HITS (%)*	93.33	4.37	95.76	4.30
OPRT (msec)	4,456	1,040	4,352	1,283
PRON*	6.76	1.77	7.63	1.35
FACTS**	19.61	4.61	22.40	3.81

Note—Mean scores (with standard deviations) in absolute span (ABSPAN), reaction time (OPRT), and percentage of correct responses (HITS) for the operations, and reading comprehension for pronouns (PRON) and facts (FACTS). *$p < .05$ and **$p < .01$, t tests.

Quasiexperimental Strategy

We found no significant correlation between ABSPAN and PRON; we followed the quasiexperimental strategy to test whether this methodology could show differences in reading comprehension between participants with high and low WMC. Two groups were formed according to the ABSPAN scores: (1) *high span*, those participants whose score was in the highest quartile of the sample (n = 35; mean ABSPAN score = 41.54, SD = 6.7), and (2) *low span*, those participants whose score was in the lowest quartile of the sample (n = 33; mean ABSPAN score = 16.42, SD = 3.35). By means of t tests, the scores of these groups for each of the variables studied were compared (see Table 1). High-span participants obtained significantly higher scores than did low-span ones on both reading comprehension indices: FACTS [22.40 vs. 19.61; $t(66) = -2.73, p < .01$] and PRON [7.63 vs. 6.76; $t(66) = -2.29, p < .05$] (see Figure 2), as well as on the percentage of correct responses in the SGOSPAN processing component [95.76 vs. 93.33; $t(66) = -2.31$, $p < .05$]. No significant differences were found regarding the OPRT.

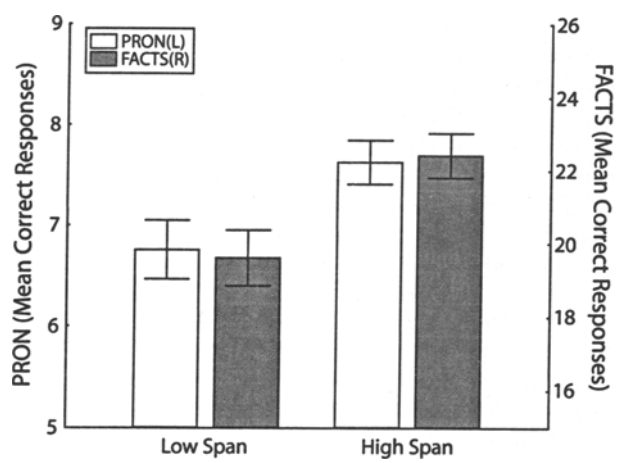

Figure 2. Direct scores on both reading comprehension indices (FACTS and PRON) for high- and low-span groups.

Discussion

The internal consistency of SGOSPAN administered to groups ($\alpha = .69$), although lower than that observed in the individual application (see Experiment 1), is identical to that found when the original version is used (Engle et al., 1999), due to which we believe that SGOSPAN is a reliable task when performed in groups.

Span scores obtained when SGOSPAN is administered to groups do not depend either on the differential use of rehearsing strategies during the presentation of operations (no correlation was found between ABSPAN and OPRT, and the RTs of the groups of high and low spans were practically identical) or on reduced effort in the correct resolution of operations (we found a positive correlation, instead of a negative one, between ABSPAN and HITS, and the high-span group obtained significantly higher percentages of correct responses in the operations). These data confirm the tendency observed in Experiment 1 supporting the notion that individuals with higher WMC, besides recalling more words while they process other information, carry out the processing more efficiently when they are asked to store other information at the same time. In any case, on average, both groups correctly realized 90% of the operations, which indicates that the processing requirements were easily satisfied by both low- and high-span participants.

As in Experiment 1, the mean SGOSPAN score for the entire sample falls between mean GOSPAN and OSPAN scores. Once again, reading instructions for the SGOSPAN stimuli seem to explain this pattern of results.

With regard to the validity of SGOSPAN when administered to groups, the correlation between WM span (ABSPAN) and the recall of facts (FACTS) was significant, its magnitude ($r = .28$) being slightly lower than the correlations found in Experiment 1 ($r = .31$) and in other studies in which complex span tasks (similar to SGOSPAN) have been used (between .30 and .48; Daneman & Merikle, 1996). On the other hand, we expected to find a significant correlation between ABSPAN and the second comprehension index used (PRON); our results did not confirm this.

Both the low correlation between ABSPAN and FACTS and the lack of a significant correlation between ABSPAN and PRON may be due to a lower validity for group-administered SGOSPAN with respect to similar tasks.

Nevertheless, it is also possible that some of the characteristics of RCT′ may have decisively influenced these results. First, the texts were written with the aim of being easy to read and to understand, using simple grammatical structures and general content about situations that would be familiar to the participants. Furthermore, the questions referred to information appearing explicitly in the texts, so that the participants did not need to infer anything from them. This feature is highly relevant, since WM is fundamental for drawing inferences from a linguistic message (Calvo, 2001; Moran & Gillon, 2005). Given that the implication of WM in linguistic comprehension depends, to a large extent, on the difficulty of the linguistic message (Engle & Conway, 1998; Engle & Kane, 2004), we consider that WM may be less relevant for implementing RCT′ than for carrying out other, more complex comprehension tasks, such as the RCT. Hence, these characteristics would explain the lower correlation between ABSPAN and FACTS and, at least in part, the absence of any correlation between ABSPAN and PRON. Furthermore, the latter result could be related to the position of the pronoun within the text, since it always appeared in the last sentence; after a small number of texts, the participants may have noticed its position and may have read back over the text in search of the corresponding reference. Thus, the scores for the PRON variable may reflect not only the text comprehension process, but also the use of strategies by the participants.

With regard to the results of the quasiexperimental strategy, we found evidence supporting the validity of SGOSPAN when administered to groups, since the high-span group obtained significantly higher scores than did the low-span group in both comprehension indices. The fact that the result obtained for the PRON variable with the quasiexperimental strategy differed so greatly from the correlational result could be due to the homogeneity of the sample used: All the participants were university students of very similar ages, and most of them had average WMC. Among these participants, differences regarding WMC would not be very relevant, and their performance of the comprehension task would be greatly influenced by other variables, such as motivation or reading habits. Nevertheless, among participants with extreme ABSPAN scores, the WMC would play a fundamental role when individual differences in comprehension were explained. This possibility was explored by means of the ABSPAN–PRON and ABSPAN–FACTS correlations, including only those participants from the first and fourth quartiles. Both correlations were notably higher with respect to those calculated over the entire sample ($r = .20, p > .1, n = 66$, and $r = .35$, $p < .005$, $n = 66$, respectively), although the ABSPAN–PRON correlation was not statistically significant.

GENERAL CONCLUSIONS

SGOSPAN's internal consistency data, similar to those found with GOSPAN (De Neys et al., 2002) and OSPAN (Engle et al., 1999; Turner & Engle, 1989), show that it is a reliable task, applied both individually and to groups. On the other hand, for the individual (Experiment 1) and the group (Experiment 2) applications, almost all predictions regarding the validity of the tasks were confirmed. That is to say, we found a positive correlation between SGOSPAN and reading comprehension that falls within the range of those found in previous studies. Moreover, following the quasiexperimental strategy, we found that the high-span group obtained significantly higher scores in the RCT than did the low-span one. These results provide us with an external validation for SGOSPAN and, together with those obtained by De Neys and colleagues (De Neys, 2006; De Neys & Dieussaert, 2005; De Neys et al., 2002; De Neys et al., 2005; De Neys & Verschueren, 2006), seem to indicate that an automated, group-administrable version of

OSPAN can be used to measure WMC, while maintaining the reliability and validity of the original version.

Some procedural features of SGOSPAN are relevant and merit some discussion. First, almost all the participants achieved high accuracy rates in solving the SGOSPAN operations. This result suggests that, after the first training phase, the association between each mouse button and the corresponding decision was well established. Thus, it supports the use of the mouse buttons for the participants to indicate their decisions. Second, the method used for instructing the participants where to write down the words seems to be suitable. This method solves one of the major problems in the development of a self-administered version of the OSPAN task.

AUTHOR NOTE

We thank Wim De Neys for his valuable help and for critical readings of previous versions of the article. Correspondence concerning this article should be addressed to J. L. Pardo-Vázquez, Facultad de Medicina (Dpto. de Fisiología), Universidad de Santiago de Compostela, C\ San Francisco, nº 1, C.P. 15701, Santiago de Compostela, A Coruña, Spain (e-mail: josepar@usc.es).

REFERENCES

ALGARABEL, S. (1996). Índices de interés psicolingüístico de 1.917 palabras castellanas. *Cognitiva*, **8**, 43-88.

ANDRADE, J. (2001). *Working memory in perspective*. Hove, U.K.: Psychology Press.

ASHCRAFT, M. H. (2002). *Cognition* (3rd ed.). Upper Saddle River, NJ: Prentice-Hall.

BADDELEY, A. D., & HITCH, G. J. (1974). Working memory. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 8, pp. 47-90). New York: Academic Press.

BEAMAN, C. P. (2004). The irrelevant sound phenomenon revisited: What role for working memory capacity? *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **30**, 1106-1118.

BLECKLEY, M. K., DURSO, F. T., CRUTCHFIELD, J. M., ENGLE, R. W., & KHANNA, M. M. (2003). Individual differences in working memory capacity predict visual attention allocation. *Psychonomic Bulletin & Review*, **10**, 884-889.

BLINKHORN, S. F. (1999). *GMA: Evaluación de grado medio y alto*. Madrid: TEA Ediciones. (Original work published 1985)

BORNKESSEL, I. D., FIEBACH, C. J., & FRIEDERICI, A. D. (2004). On the cost of syntactic ambiguity in human language comprehension: An individual differences approach. *Cognitive Brain Research*, **21**, 11-21.

BRUMBACK, C. R., LOW, K. A., GRATTON, G., & FABIANI, M. (2005). Putting things into perspective: Individual differences in working-memory span and the integration of information. *Experimental Psychology*, **52**, 21-30.

BUNTING, M. F., CONWAY, A. R. A., & HEITZ, R. P. (2004). Individual differences in the fan effect and working memory capacity. *Journal of Memory & Language*, **51**, 604-622.

CALVO, M. G. (2001). Working memory and inferences: Evidence from eye fixations during reading. *Memory*, **9**, 365-381.

CAPON, A., HANDLEY, S. J., & DENNIS, I. (2003). Working memory and reasoning: An individual differences perspective. *Thinking & Reasoning*, **9**, 203-244.

COLOM, R., FLORES-MENDOZA, C., & REBOLLO, I. (2003). Working memory and intelligence. *Personality & Individual Differences*, **34**, 33-39.

CONWAY, A. R. A., & ENGLE, R. W. (1994). Working memory and retrieval: A resource-dependent inhibition model. *Journal of Experimental Psychology: General*, **123**, 354-373.

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.

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

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

DE NEYS, W. (2006). Dual processing in reasoning: Two systems but one reasoner. *Psychological Science*, **17**, 428-433.

DE NEYS, W., & DIEUSSAERT, K. (2005). Individual differences in rational thinking time. In B. G. Bara, L. W. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the 27th Annual Conference of the Cognitive Science Society* (pp. 577-582). Mahwah, NJ: Erlbaum.

DE NEYS, W., D'YDEWALLE, G., SCHAEKEN, W., & VOS, G. (2002). A Dutch, computerized, and group administrable adaptation of the operation span test. *Psychologica Belgica*, **42**, 177-190.

DE NEYS, W., SCHAEKEN, W., & D'YDEWALLE, G. (2005). Working memory and everyday conditional reasoning: Retrieval and inhibition of stored counterexamples. *Thinking & Reasoning*, **11**, 349-381.

DE NEYS, W., & VERSCHUEREN, N. (2006). Working memory capacity and a notorious brain teaser: The case of the Monty Hall dilemma. *Experimental Psychology*, **53**, 123-131.

DOUGHERTY, M. R. P., & HUNTER, J. (2003). Probability judgment and subadditivity: The role of working memory capacity and constraining retrieval. *Memory & Cognition*, **31**, 968-982.

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, & Cognition*, **18**, 972-992.

ENGLE, R. W., CARULLO, J. J., & COLLINS, K. W. (1991). Individual differences in the role of working memory in comprehension and following directions. *Journal of Educational Research*, **84**, 253-262.

ENGLE, R. W., & CONWAY, A. R. A. (1998). Comprehension and working memory. In R. H. Logie & K. J. Gilhooly (Eds.), *Working memory and thinking* (pp. 67-91). Hillsdale, NJ: Erlbaum.

ENGLE, R. W., & KANE, M. J. (2004). Executive attention, working memory capacity, and a two-factor theory of cognitive control. In B. H. Ross (Ed.), *The psychology of learning and motivation* (Vol. 44, pp. 145-199). New York: Elsevier.

ENGLE, R. W., TUHOLSKI, S. W., LAUGHLIN, J. E., & CONWAY, A. R. A. (1999). Working memory, short-term memory, and general fluid intelligence: A latent-variable approach. *Journal of Experimental Psychology: General*, **128**, 309-331.

FRIEDMAN, N. P., & MIYAKE, A. (2004). The reading span test and its predictive power for reading comprehension ability. *Journal of Memory & Language*, **51**, 136-158.

GILHOOLY, K. J., WYNN, V., PHILLIPS, L. H., LOGIE, R. H., & DELLA SALA, S. (2002). Visuo-spatial and verbal working memory in the five-disc Tower of London task: An individual differences approach. *Thinking & Reasoning*, **8**, 165-178.

HOSKYN, M., & SWANSON, H. L. (2003). The relationship between working memory and writing in younger and older adults. *Reading & Writing*, **16**, 759-784.

KANE, M. J., & ENGLE, R. W. (2000). Working-memory capacity, proactive interference, and divided attention: Limits on long-term memory retrieval. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **26**, 336-358.

KANE, M. J., HAMBRICK, D. Z., TUHOLSKI, S. W., WILHELM, O., PAYNE, T. W., & ENGLE, R. W. (2004). The generality of working memory capacity: A latent-variable approach to verbal and visuospatial memory span and reasoning. *Journal of Experimental Psychology: General*, **133**, 189-217.

KLEIN, K., & FISS, W. H. (1999). The reliability and stability of the Turner and Engle working memory task. *Behavior Research Methods, Instruments, & Computers*, **31**, 429-432.

KYLLONEN, P. C., & CHRISTAL, R. E. (1990). Reasoning ability is (little more than) working-memory capacity? *Intelligence*, **14**, 389-433.

KYLLONEN, P. C., & STEPHENS, D. L. (1990). Cognitive abilities as determinants of success in acquiring logic skill. *Learning & Individual Differences*, **2**, 129-160.

MIYAKE, A. (2001). Individual differences in working memory: Introduction to the special section. *Journal of Experimental Psychology: General*, **130**, 163-168.

MIYAKE, A., & SHAH, P. (1999). Toward unified theories of working memory: Emerging general consensus, unresolved theoretical issues and future research directions. In A. Miyake & P. Shah (Eds.), *Models*

of working memory: Mechanisms of active maintenance and executive control (pp. 442-481). New York: Cambridge University Press.

MORAN, C., & GILLON, G. (2005). Inference comprehension of adolescents with traumatic brain injury: A working memory hypothesis. *Brain Injury*, **19**, 743-751.

ROSEN, V. M., & ENGLE, R. W. (1997). The role of working memory capacity in retrieval. *Journal of Experimental Psychology: General*, **126**, 211-227.

TURNER, M. L., & ENGLE, R. W. (1989). Is working memory capacity task dependent? *Journal of Memory & Language*, **28**, 127-154.

UNSWORTH, N., & ENGLE, R. W. (2005). Individual differences in working memory capacity and learning: Evidence from the serial reaction time task. *Memory & Cognition*, **33**, 213-220.

UNSWORTH, N., HEITZ, R. P., SCHROCK, J. C., & ENGLE, R. W. (2005). An automated version of the operation span task. *Behavior Research Methods*, **37**, 498-505.

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

WATERS, G. S., & CAPLAN, D. (1996). The measurement of verbal working memory capacity and its relation to reading comprehension. *Quarterly Journal of Experimental Psychology*, **49A**, 51-79.

WATSON, J. M., BUNTING, M. F., POOLE, B. J., & CONWAY, A. R. A. (2005). Individual differences in susceptibility to false memory in the Deese–Roediger–McDermott paradigm. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **31**, 76-85.

NOTES

1. The name of the task is related to the processing component (arithmetic operation resolution), and it has been used with different storage components (e.g., word or digit recall; Turner & Engle, 1989). Better results have been obtained combining operation resolution with word recall, and this combination has usually been used and has been assumed as the standard procedure, as can be seen in the description recently written by one of the authors of the task (Unsworth, Heitz, Schrock, & Engle, 2005).

2. In the present work, Spanish words have been employed; nevertheless, the application procedure of the task allows any language to be used, with the only condition being that of selecting a set of words similar with regard to different variables (see the description of the materials).

(Manuscript received November 20, 2006;
revision accepted for publication March 28, 2007.)