A Standard Computerized Version of the Reading Span Test in Different Languages

Article *in* European Journal of Psychological Assessment · January 2008

DOI: 10.1027/1015-5759.24.1.35

CITATIONS
38

READS
4,823

4 authors, including:

Maurits van den Noort
Kyung Hee University
98 PUBLICATIONS **532** CITATIONS

SEE PROFILE

Peggy Margaretha, Petronella, Carolina Bosch
Radboud University
88 PUBLICATIONS **322** CITATIONS

SEE PROFILE

Kenneth Hugdahl
University of Bergen
523 PUBLICATIONS **16,903** CITATIONS

SEE PROFILE

A Standard Computerized Version of the Reading Span Test in Different Languages

Maurits van den Noort[1], Peggy Bosch[2], Marco Haverkort[3,4], and Kenneth Hugdahl[1]

[1]Department of Biological and Medical Psychology, University of Bergen, Norway
[2]Nijmegen Institute for Cognition and Information, Radboud University Nijmegen, The Netherlands,
[3]Department of Linguistics, Radboud University Nijmegen, The Netherlands
[4]Department of Linguistics, Boston University, USA

Abstract. The Reading Span Test (RST) is a verbal working-memory test. The original RST (Daneman & Carpenter, 1980), and derivatives of it, are being used increasingly as assessments of central executive functioning and for research on aging-associated cognitive decline (Whitney, Arnett, Driver, & Budd, 2001). Several versions have been made in order to further improve the test or to develop a version in a different language. However, all versions changed different things, making direct comparisons of the results with the RST between different research groups and across different languages impossible. This paper presents the results of testing a new standard computerized version of the RST in four languages (Dutch, English, German, and Norwegian). The new RST meets strict methodological criteria that are the same for all four language versions. A plausibility test, an abstract-concrete rating scale, and a pilot-study were conducted on native speakers to test the new RST. In addition, the internal and external reliability and the ecological validity of the new RST were tested. The results showed that the new RST is a suitable test to investigate verbal working memory. Finally, an important advantage of the new RST is that the different language versions make cross-linguistic comparisons of RST results possible.

Keywords: Reading Span Test, methodology, verbal working memory, reliability, validity

Introduction

One of the most widely used verbal working-memory tests is the Reading Span Test (RST) that was developed by Daneman and Carpenter (1980). The original RST, and derivatives of it, are being used increasingly as assessments of central executive functioning (Engle, Kane, & Tuholski, 1999). For example, the RST is a useful predictor of central executive functioning among populations with various neurological disorders including aphasia, Alzheimer's disease, and schizophrenia (e.g., Caspari, Parkinson, LaPointe, & Katz, 1998; Kempler, Almor, Tyler, Andersen, & MacDonald, 1998; Stone, Gabrieli, Stebbins, & Sullivan, 1998). In addition, there is considerable interest in whether the RST and analogs of it measure a basic ability that accounts for age-related declines in cognitive performance (e.g., Byrne, 1998; Salthouse & Kersten, 1993; Van den Noort, Haverkort, Bosch, & Hugdahl, 2006; Whitney et al., 2001).

In the original RST (Daneman & Carpenter, 1980), participants were instructed to read series of sentences aloud, while remembering the final word of each sentence in a particular series. The sentences were presented on cards, one sentence on each card, and at the end of a series a blank card appeared, which was the participant's cue for recall. It was the participant's task to recall all sentence-final words in the right order. The number of sentences in a series was gradually increased. The reading span was the maximum number of final words that were correctly recalled. The RST is a so-called "complex" verbal working-memory test since it taps both the storage and processing elements of working memory, whereas "simple" verbal working-memory tests only tap the storage element (Lobley, Baddeley, & Gathercole, 2005). In addition, Daneman and Carpenter (1980) developed a listening version of the RST. The listening span also requires the retention of sentence-final words, but the participant listened to, rather than read, lists of sentences.

In the last decades, several variants of the RST have been developed based on the original test (e.g., Desmette, Hupet, Schelstraete, & Van der Linden, 1995; Kondo & Osaka, 2004; LaPointe & Engle, 1990; Payne & Whitney, 2002; Whitney, Ritchie, & Clark, 1991; Whitney et al., 2001). In addition, new listening versions of the RST have been developed (Liu, Schallert, & Carroll, 2004). There are, however, several problems with the original RST (Daneman & Carpenter, 1980) and the later versions of the test. The main problems are: the method of sentence presentation, the psycholinguistic criteria of the stimulus material, the determination of the reading span, and translations of the RST in different languages. Some of the newer

DOI 10.1027/1015-5759.24.1.35

versions were improvements (see Daneman & Hannon, 2001; Hannon & Daneman, 2001), however, all versions changed different things (Whitney et al., 2001), making direct comparisons of the results with the RST between different research groups and across different languages impossible.

A critical assessment of the different tests that are available suggests a number of improvements that are necessary for a standard computerized version of the RST (e.g., Daneman & Merikle, 1996):

1. The length of the sentences must be better controlled for. Daneman and Carpenter (1980) only looked at the number of words, but it would be better to control for the number of syllables (Desmette et al., 1995) and letters as well.
2. In the original RST (Daneman & Carpenter, 1980), no attention was paid to the length of the sentence-final words (e.g., LaPointe & Engle, 1990; Tehan, Hendry, & Kocinsky, 2001). There is an important difference between remembering a one syllable word or a four syllable word. Shorter words are better recalled in comparison with longer words (Baddeley, Thomson, & Buchanan, 1975).
3. Daneman and Carpenter did not control for the frequency of the sentence-final words. Frequent words are better recalled than infrequent words (Baddeley, 1997; Gregg, 1976).
4. In the original RST, no special attention was paid to the abstractness/concreteness of the sentence-final words (e.g., Kondo & Osaka, 2004). Yet, this distinction can be very important, because people can use two different memory resources (Baddeley, 1986). First, they can use the phonological loop: In other words, they remember the words by keeping the sound sequences active in their memory. Second, people can use a visual strategy. In Baddeley's model, this is called the "visual-spatial sketch pad." For example, if people have to remember the word "tree," they keep a visual picture of a tree active in this part of working memory and translate this picture back into a phonological representation of the word at recall.
5. Moreover, there are problems with the presentation of the material. Daneman and Carpenter (1980) used cards to present the sentences and they did not use any time restrictions during sentence presentation (Friedman & Miyake, 2004). One problem with this method is that participants can read the sentences more slowly to improve their recall (Saito & Miyake, 2003). When this type of strategy is used by participants, the RST is no longer a strict working memory test.
6. Furthermore, there are problems with the use of the RST in other languages. In most languages only translations of the original RST are used. Therefore, it is impossible to meet the original Daneman and Carpenter (1980) criteria (e.g., the old Dutch version, see Table 2). By using translations of the original RST, one does not control for the specific language differences between the original English version and the translation of the RST in word frequency, sentence length, etc.
7. Finally, Daneman and Carpenter (1980) determined the reading span by looking at the longest set of sentences of which participants could recall all the sentence-final words. This is based on only a few attempts and it is, therefore, too random an indication, which does not necessarily adequately reflect the participant's working-memory capacity (Desmette et al., 1995; Friedman & Miyake, 2004). It would be better to determine the reading span by looking at the total number of sentence-final words that participants can recall during the whole RST.

This paper presents the results of testing a new standard computerized version of the RST in four languages (Dutch, English, German, and Norwegian) which meets stricter methodological criteria than previous RSTs (e.g., Daneman & Hannon, 2001; Desmette et al., 1995; Hannon & Daneman, 2001; Kondo & Osaka, 2004). The main advantage of the new, standard computerized version of the RST is that it creates a possibility for direct comparison between different research groups and across languages (Van den Noort, Bosch, & Hugdahl, 2006). First, the stimulus materials of the new RST in Dutch, English, German, and Norwegian and the way it was tested will be discussed. Next, an experiment will be presented in which the new RST in Dutch, English, German, and Norwegian was tested. In this experiment, three different working memory tasks were used: digit span (forward and backward), letter-number ordering, and the new RST. In addition, both the internal and external reliability of the new RST will be estimated.

Materials and Methods

Construction of the Sentences

Reliability is an essential ingredient of any individual-differences test. To further improve the reliability of the new RST in Dutch, English, German, and Norwegian, 100 sentences were developed to meet strict methodological criteria (for examples see Appendix):

1. The length of the sentences was controlled for, ranging from 12 to 17 words, 20 to 22 syllables, and 55 to 73 letters. As a result, the sentences of the new RST (see Table 1) had a better controlled sentence length than the sentences that were used in other RSTs (see Table 2).
2. In contrast with the original RST, the number of syllables and the number of letters were controlled for, over the five series.
3. The frequency of the sentence-final words were controlled for, over the five series. The CELEX lexical database (Baayen, Piepenbrock, & Van Rijn, 1993) was used to determine the frequency of the sentence-final words for the Dutch, German, and English versions. To determine the frequency of the sentence-final words in Norwegian, the most common frequency book for Norwegian was used (Heggestad, 1982).

Table 1. The mean number of syllables of the sentences and the mean number of syllables, letters, and the mean frequency of the final words per series for the Dutch, English, German, and Norwegian RST

	Syllables sentences	Syllables final words	Letters final words	Frequency final words
Dutch RST				
Series 1	21.0[a] (0.9)	1.9[a] (0.7)	6.0[a] (1.6)	2502[a] (3466)
Series 2	21.4[a] (0.9)	1.8[a] (0.8)	5.7[a] (1.8)	2426[a] (3063)
Series 3	21.3[a] (0.9)	1.7[a] (0.6)	5.6[a] (1.6)	2442[a] (2103)
Series 4	21.2[a] (0.9)	2.1[a] (0.9)	6.5[a] (1.6)	2430[a] (2340)
Series 5	21.6[a] (0.6)	2.0[a] (0.7)	6.2[a] (1.5)	2433[a] (2870)
English RST				
Series 1	21.2[a] (0.8)	1.9[a] (0.6)	6.1[a] (1.3)	2869[a] (2193)
Series 2	21.2[a] (0.8)	1.8[a] (0.7)	6.0[a] (1.4)	2810[a] (2166)
Series 3	21.3[a] (0.7)	1.9[a] (0.7)	6.2[a] (1.4)	2885[a] (1988)
Series 4	21.1[a] (0.8)	1.9[a] (0.7)	6.1[a] (1.5)	2924[a] (2746)
Series 5	21.1[a] (0.8)	2.0[a] (0.8)	6.3[a] (1.6)	2792[a] (1837)
German RST				
Series 1	21.0[a] (0.7)	1.9[a] (0.8)	5.9[a] (1.9)	2470[a] (2196)
Series 2	21.0[a] (0.8)	2.0[a] (0.6)	6.0[a] (1.4)	2478[a] (2511)
Series 3	21.0[a] (0.8)	1.9[a] (0.7)	5.8[a] (1.4)	2491[a] (2546)
Series 4	21.3[a] (0.8)	1.8[a] (0.7)	5.6[a] (1.7)	2440[a] (2496)
Series 5	20.7[a] (0.9)	1.8[a] (0.7)	5.8[a] (1.7)	2465[a] (2702)
Norwegian RST				
Series 1	21.5[a] (1.0)	2.2[a] (0.8)	5.9[a] (1.4)	1012[a] (1792)
Series 2	21.2[a] (0.9)	2.1[a] (0.7)	6.1[a] (1.8)	920[a] (2103)
Series 3	21.5[a] (0.9)	2.2[a] (0.9)	6.5[a] (2.1)	1027[a] (1676)
Series 4	21.4[a] (0.8)	2.0[a] (0.8)	5.9[a] (1.5)	914[a] (1358)
Series 5	21.4[a] (0.8)	2.1[a] (0.8)	6.0[a] (1.7)	1004[a] (1499)

Note. Different letters indicate significant difference at $p < .05$ level.

Table 2. The mean number of syllables and letters per series of the original RST compared to a Dutch and French version of the test

RST	Original[1]	Dutch[2]	French[3]
Syllables			
Series 1	21.6 (2.5)	24.1 (3.6)	21.3 (1.7)
Series 2	22.1 (2.1)	22.8 (2.8)	22.3 (1.5)
Series 3	21.7 (2.2)	24.2 (3.4)	22.1 (1.7)
Series 4	22.5 (2.1)	23.3 (3.0)	21.5 (1.8)
Series 5			22.3 (1.0)
Letters			
Series 1	66.1 (5.8)	71.2 (8.6)	62.6 (4.1)
Series 2	67.0 (5.1)	69.3 (8.0)	64.8 (2.6)
Series 3	68.0 (5.9)	73.3 (9.3)	65.3 (4.4)
Series 4	69.5 (4.5)	68.1 (9.4)	65.2 (2.4)
Series 5			64.9 (4.0)

Note. [1] = The original RST (Daneman & Carpenter, 1980). [2] = The old version of the Dutch RST which was a translation of the original RST (Hoeks, unpublished). [3] = French RST (Desmette et al., 1995).

4. The concreteness of the sentence-final words over the series was controlled for.
5. All sentences have a maximum presentation time of 6.5 seconds.
6. The 100 sentences in all language versions were evaluated using the same psychometric characteristics.
7. Within the sets, the sentences and sentence-final words were controlled for semantic relations as much as possible. Numerous observations have continued to indicate semantic influences on verbal working memory. These include, for instance, the incorrect inclusion of close associates in recalling lists of semantically related words (e.g., Kittler, Krinsky-McHale, & Devenny, 2004). To minimize the semantic influences on verbal working memory in the new RST, three native speakers independently rated all sentences and sentence-final words for their semantic relations and changes in sentence order were made to control for semantic relations within the sets, and over the five series.

A plausibility test and an abstract-concrete rating task were used to test the stimulus material of the new RST (see below).

Plausibility Test

A total of 120 students participated in a plausibility test. They were all native speakers of Dutch, English, German, or Norwegian. The participants rated the plausibility of 120 sentences in their native languages. Of these sentences, 100 sentences were experimental and 20 sentences were distractors. The distractor sentences were used to check if the participants conducted the plausibility test properly. The participants rated the sentences on a 5-point scale, ranging from 1 (*not plausible at all*) to 5 (*very plausible*). Experimental sentences below a mean of 3.5 were replaced by more plausible sentences. These new sentences were tested for plausibility by 20 native speakers on the same 5-point scale. As a result, all of the sentences included in the test have been rated as plausible (above a mean of 3.5).

Abstract-Concrete Rating Scale

A total of 120 students filled out an abstract-concrete rating scale. All participants were native speakers of Dutch, English, German, or Norwegian. They rated 100 words (the sentence-final words) on their degree of concreteness or abstractness in their native languages. All words were rated on a 6-point scale, ranging from 1 (*concrete*) to 6 (*abstract*). In the final version of the RST, the mean concreteness/abstractness scores of the 20 sentence-final words did not differ between the five series. Moreover, there were no significant differences in mean concreteness/abstractness scores between the Dutch, English, German, and Norwegian RST.

Pilot Study

After the construction and testing of the 100 sentences, a pilot study on 40 students (10 native speakers of Dutch, English, German, and Norwegian, respectively) was conducted to further test the material of the new RST in Dutch, English, German, and Norwegian. During the pilot study, the participants received the 100 experimental sentences of the new RST in a random order. The participants had to read the sentences aloud and then they had to press the "Enter" key as soon as possible after reading the sentences. The mean reading time per sentence was recorded. The analysis of the reading times showed no significant differences in mean reading time between the five series of 20 sentences in the Dutch, English, German, or Norwegian RST, supporting the methodology of the new RST. Participants read the sentences in the five different series equally fast, suggesting that there were no differences in difficulty between the series. The mean reading time of the 100 sentences was: for the pilot study in Dutch 4044 ms (SD = 249.18), for English 4042 ms (SD = 195.1), for German 4097 ms (SD = 267.3), and for Norwegian 4061 ms (SD = 173.7).

Computerized Version

The new RST was programmed in E-Prime (Schneider, Eschman, & Zuccolotto, 2002). The RST started with an instruction on the computer screen and two exercise trials, after which the experimental material was presented. The instruction emphasized the importance of reading the sentences as fast as possible while reading for content. To further stress the importance of reading for content, we informed the participants that questions about the sentences would be asked at the end of the RST. After reading all the sentences, participants had to answer 10 questions about the content of several particular sentences. As in the original RST (Daneman & Carpenter, 1980), the sentences were presented in different set sizes (2, 3, 4, 5, or 6 sentences). There were two main reasons for using different set sizes in the new RST. On the one hand, small sets of, for instance, 2 or 3 sentences, could result in "ceiling effects." Almost all participants will perform at or near the maximum, making the test useless for measuring individual differences in verbal working-memory capacity. On the other hand, large sets of, for instance, 5 or 6 sentences, could lead to problems as well. Participants could get the impression that remembering 5 or 6 sentence-final words is the "standard" and since most participants will not be able to remember 5 or 6 sentence-final words this could lead to stress and a worse performance on the RST. Contrary to the original Daneman and Carpenter (1980) version, the different set sizes were presented in random order. As a result, participants did not know how many sentences would be presented in a set and, therefore, could not anticipate. The participants' task was to read all sentences aloud. When a participant had finished reading a sentence, he/she pressed the space bar, after which another sentence appeared on the screen. If the participant could not finish the sentence within 6.5 seconds, the computer automatically presented the next sentence. When a participant had completed all the sentences of a set, the word "recall" was presented. At that point, participants had to recall the last word of each sentence in the set. The order of recall was free. This is important since free recall gives important information on possible primacy- and recency effects (Baddeley, 1997). The participants completed 100 sentences in total, divided over five different series of 2, 3, 4, 5, or 6 sentences.

Reading Span

In contrast to the original RST study (Daneman & Carpenter, 1980), the total number of remembered words was determined, instead of the maximum number of remembered words per series. These criteria were used because the total number of remembered words gives us more specific information about the performance of the participants over the whole RST (e.g., Desmette et al., 1995; Friedman & Miyake, 2004).

Table 3. The mean total number of remembered words and mean reaction times per series for the Dutch, English, German, and Norwegian version of the new RST

	Dutch RST		English RST	
	Mean Reading span	Mean Reaction time	Mean Readings span	Mean Reaction time
Series 1	13.0[a] (1.5)	4013.5[a] (230.3)	13.2[a] (1.4)	4039.3[a] (220.3)
Series 2	13.2[a] (1.5)	4028.2[a] (136.0)	13.0[a] (1.4)	4031.3[a] (127.2)
Series 3	13.5[a] (1.9)	4053.0[a] (147.6)	13.0[a] (1.3)	4035.9[a] (154.8)
Series 4	12.9[a] (1.7)	4017.0[a] (170.1)	13.4[a] (1.9)	4032.3[a] (186.2)
Series 5	13.4[a] (1.6)	4026.4[a] (222.3)	13.2[a] (1.6)	4044.8[a] (227.0)
	German RST		Norwegian RST	
	Mean Reading span	Mean Reaction time	Mean Readings span	Mean Reaction time
Series 1	12.9[a] (1.9)	4129.4[a] (365.6)	13.1[a] (1.5)	4033.2[a] (156.7)
Series 2	13.3[a] (1.5)	4091.7[a] (322.4)	13.0[a] (1.7)	4046.2[a] (144.3)
Series 3	13.0[a] (1.7)	4132.7[a] (405.8)	13.3[a] (1.9)	4039.5[a] (184.1)
Series 4	13.0[a] (1.5)	4144.1[a] (475.8)	13.4[a] (1.9)	4022.2[a] (148.1)
Series 5	13.1[a] (1.8)	4151.6[a] (374.2)	13.0[a] (1.9)	4057.8[a] (202.0)

Note. Different letters indicate significant difference at $p < .05$ level.

Experiment

An experiment was conducted in which the new standard computerized RST was tested on native speakers of Dutch, English, German, or Norwegian. In the experiment, we compared the new RST with two other working memory tasks: digit span (forward and backward) and letter-number ordering. The digit span forward, the digit span backward, and the letter-number ordering task were all taken from the Wechsler Adult Intelligence Scale (WAIS-III) (Wechsler, 2000).

In total, 160 undergraduate students participated in the study after giving their written informed consent according to institutional guidelines. There were 40 native speakers of Dutch, English, German, and Norwegian, respectively. The average age of the participants was 23.12 years (SD = 3.44), ranging from 18 to 32 years, and there were no significant differences in mean age between the four groups of native speakers. In addition, we controlled for gender; there were 20 men and 20 women in each group. An honorarium was given for participation.

The participants were tested individually in an experimental room. After receiving a general instruction in their native language, a detailed instruction was given for each separate task. The participants completed three different experimental tasks in random order: The digit span (forward and backward), letter-number ordering, and the new RST. After 2 weeks, all participants completed the new RST for a second time to investigate the test-retest reliability. The time interval of 2 weeks was chosen to avoid an overestimation of the reliability because of a learning/practice effect. After the second testing, participants received feedback about the aim of the study.

Since the underlying methodology in all languages is the same, no significant differences in mean number of remembered words and in mean reaction time were expected between the RSTs in Dutch, English, German, and Norwegian. Moreover, no significant differences in the mean number of remembered words and in the mean reaction time between the five series were expected within the four RSTs since the methodological criteria of the five series were the same. In addition, it was expected that the new RST would only correlate with letter-number ordering since it is a more complex verbal working-memory task, and not with digit span (forward and backward), which is a simple verbal working-memory task. Because of the restricted sentence presentation time, a lower reading span was expected compared to the original RST (Daneman & Carpenter, 1980), since the participants have less time to process the information.

Results

One-way analysis of variance (ANOVA)s and independent t-tests, with a Bonferroni (Huberty & Morris, 1989) adjustment of the α level (.05), were conducted to determine if the five series were comparable and whether differences in reading span scores between the Dutch, English, German, and Norwegian RST could be found. As Table 3 shows, no significant differences were found. There were no differences in the total number of remembered words between the five series for all language versions, suggesting that there were no differences in difficulty in remembering the sentence-final words. Moreover, there were no differences in the total number of remembered words between the four language versions. Native speakers of Dutch, English, Ger-

man, and Norwegian performed the same on the four different language versions. The mean number of recalled words on all RSTs was 65.64 (*SD* = 5.82), ranging from 48 to 76. Note that the maximum score on the RST is 100.

Furthermore, one-way ANOVAs and independent *t*-tests, with a Bonferroni (Huberty & Morris, 1989) adjustment of the α level (.05), were conducted to determine if the five series of the RST produced equivalent reaction time results and if differences in mean reaction time could be found between the Dutch, English, German, and Norwegian versions. In all analyses, the reaction times below two standard deviations from the mean were removed. The rationale behind this decision was that, if a participant read a sentence extremely fast or extremely slow in comparison with other sentences, one cannot be sure that this sentence was read in a normal way. Because of the maximum presentation time of 6.5 seconds, there were no extremely long reaction times. As Table 3 shows, no significant differences were found between all the series. Participants read the sentences in these series equally fast, suggesting that there were no differences in difficulty. These results are in line with the results of the pilot study. In addition, no significant differences in mean reaction time between the Dutch, English, German, and Norwegian RST were found.

A Pearson correlation was conducted between the scores of the native speakers on the RST, the digit span (forward and backward), and letter-number ordering. A significant correlation between RST and letter-number ordering was found for the Dutch, English, German, and Norwegian native speakers, which was in line with our expectations (r = .58, $p < .01$). No correlation between the digit span (forward and backward) and the RST was found.

Finally, the assessment of the instrument's internal consistency resulted in a Cronbach's α of .92. In addition, the test-retest reliability between the reading span scores on the first measurement and the second measurement was estimated. The Spearman-Brown coefficient α was .88.

Discussion

One of the most widely used verbal working-memory tests is the RST that was developed by Daneman and Carpenter (1980). In recent decades, several variants of the test have been developed based on the original RST (e.g., Desmette et al., 1995; Kondo & Osaka, 2004; LaPointe & Engle, 1990; Payne & Whitney, 2002; Whitney et al., 1991, 2001) to further improve the test or to use the RST in different languages. All versions changed different things (Whitney et al., 2001), making direct comparisons of the results with the RST between different research groups and across different languages impossible. Therefore, a new standard computerized version of the RST was developed in Dutch, English, German, and Norwegian. In this paper, the new RST was presented and tested (note that more language versions, based on the same methodological criteria, are under construction).

The new computerized RST appears to be an improvement over the Daneman and Carpenter (1980) and/or later versions of the RST (e.g., Desmette et al., 1995; Kondo & Osaka, 2004). The psychometric characteristics of the new RST are better than previous versions. The new RST has been developed in different languages, making cross-linguistic comparisons of RST performance possible. The results of the experiment that was presented in this paper show that the internal reliability and the test-retest reliability of the new RST are high. The reading span results and the reaction time results seem to further support the methodology of the new RST. No differences in total number of remembered words and in mean reaction time between the five series were found for each language version. These results are in line with the results of the pilot study. Importantly, no significant differences in the mean number of remembered words and in the mean reaction time between the four RSTs were found, showing that the performance scores on the different language versions are comparable.

According to our hypothesis, the mean reading span was lower than the original reading span reported by Daneman and Carpenter (1980). In the new RST, the presentation time is restricted, therefore, participants have less time to remember the sentence-final words and it is almost impossible for the participants to read the sentences more slowly to improve their recall (e.g., Desmette et al., 1995; Friedman & Miyake, 2004). In other words, there is almost no time for strategies that are not related to working memory.

In this study, three different working memory tasks were compared: digit span (forward and backward), letter-number ordering, and the new RST. It was expected that the new RST would only correlate with letter-number ordering since it is a more complex verbal working-memory task and not with digit span (forward and backward), which is a simple verbal working-memory task. In line with our expectations, only a significant correlation between the RST and the letter-number ordering was found. No correlation between the digit span (forward and backward) and the RST was found.

It is obvious that it is not only important to develop a RST in which different language versions produce the same span size (in absolute terms); it must also be able to predict performance on real-world cognitive and language tasks. Therefore, in order to test the ecological validity of the new RST, 40 native speakers of Dutch completed two additional tasks when they came back for their second testing. After completing the RST for the second time, they also completed a relative clause task and a sentence reading task (for a detailed description see Van den Noort, Bosch, & Hugdahl, 2007). High correlations between the reading span and performance scores on the relative clause task and the sentence reading task were expected since all tasks tap language processing. In line with our expectations, high correlations between the RST, the relative clause task, and the sentence reading task were found. However, more research on the ecological validity of the new RST needs to be done.

The new RST still has some restrictions. In order to get a more accurate reading span score, participants had to complete the whole test. Therefore, motivational problems cannot be excluded. During the debriefing, some people reported concentration problems (13% of the participants) that occurred after three series. However, no indications for concentration problems were found in the behavioral data. Another limitation of our results is that a number of reaction times could not be registered because the maximal presentation time of 6.5 seconds was exceeded (3%). On the other hand, an individually-based presentation time cannot be used, because differences in reaction times can also be the result of different individual memory strategies (e.g., Desmette et al., 1995; Friedman & Miyake, 2004).

In future research it would be interesting to conduct eye-movement studies (e.g., Traxler, Williams, Blozis, & Morris, 2005) to investigate whether real sentence reading is different from sentence reading in the RST, while remembering the sentence-final words. Moreover, it would be interesting to investigate the possible clinical applications of the new RST (Van den Bos & Spelberg, 1993; Van den Noort, Haverkort et al., 2006). Previous studies with the RST have already shown that the RST could be used for clinical applications (e.g., Byrne, 1998; Caspari et al., 1998; Kempler et al., 1998; Salthouse & Kersten, 1993; Stone et al., 1998).

To conclude, an important advantage of the new RST is that the different language versions are based on the same methodological criteria, making cross-linguistic comparisons of RST results easier. As a result, the new RST is an excellent verbal working-memory test for bilingual and multilingual research (Van den Noort, Bosch et al., 2006). In addition, the new RST can be used in first language research on working memory issues (e.g., Nieuwland & Van Berkum, 2006).

References

Baayen, R.H., Piepenbrock, R., & Van Rijn, H. (1993). *The CELEX lexical database*. Philadelphia, PA: Linguistic Data Consortium.

Baddeley, A.D. (1986). *Working memory*. New York: Oxford University Press.

Baddeley, A.D. (1997). *Human memory: Theory and practice*. Hove, UK: Psychology Press.

Baddeley, A.D., Thomson, N., & Buchanan, M. (1975). Word length and the structure of short-term memory. *Journal of Verbal Learning and Verbal Behavior, 14*, 575–589.

Byrne, M.D. (1998). Taking a computational approach to aging: The SPAN theory of working memory. *Psychology and Aging, 13*, 309–322.

Caspari, I., Parkinson, S.R., LaPointe, L.L., & Katz, R.C. (1998). Working memory and aphasia. *Brain and Cognition, 37*, 205–223.

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

Daneman, M., & Hannon, B. (2001). Using working memory theory to investigate the construct validity of multiple-choice reading comprehension tests such as the SAT. *Journal of Experimental Psychology: General, 130*, 208–223.

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

Desmette, D., Hupet, M., Schelstraete, M.A., & Van der Linden, M. (1995). Adaptation en langue française du "Reading Span Test" de Daneman et Carpenter [A French adaptation of Daneman and Carpenter's "Reading Span Test"]. *L'Annee Psychologique, 95*, 459–482.

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

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

Gregg, V. (1976). Word frequency, recognition, and recall. In J. Brown (Ed.), *Recognition and recall*. Chichester, UK: Wiley and Sons.

Hannon, B., & Daneman, M. (2001). A new tool for measuring and understanding individual differences in the component processes of reading comprehension. *Journal of Educational Psychology, 93*, 103–128.

Heggestad, K. (1982). *Norsk frekvensordbok* [Norwegian frequency dictionary]. Bergen, Norway: Universitetsforlaget.

Huberty, C.J., & Morris, J.D. (1989). Multivariate analysis versus multiple univariate analyses. *Psychological Bulletin, 105*, 302–308.

Kempler, D., Almor, A., Tyler, L.K., Andersen, E.S., & MacDonald, M.C. (1998). Sentence comprehension deficits in Alzheimer's disease: A comparison of off-line and on-line processing. *Brain and Language, 64*, 297–316.

Kittler, P., Krinsky-McHale, S.J., & Devenny, D.A. (2004). Semantic and phonological loop effects on verbal working memory in middle-age adults with mental retardation. *American Journal on Mental Retardation, 109*, 467–480.

Kondo, H., & Osaka, N. (2004). Effect of concreteness of target word on verbal working memory: An evaluation using Japanese version of reading span test. *Japanese Journal of Psychology, 71*, 51–56.

LaPointe, L.B., & Engle, R.W. (1990). Simple and complex word spans as measures of working memory capacity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*, 1118–1133.

Liu, M., Schallert, D.L., & Carroll, P.J. (2004). Working memory and expertise in simultaneous interpreting. *Interpreting, 6*(1), 19–42.

Lobley, K.J., Baddeley, A.D., & Gathercole, S.E. (2005). Phonological similarity effects in verbal complex span. *The Quarterly Journal of Experimental Psychology, 58A*, 1462–1478.

Nieuwland, M.S., & Van Berkum, J.J.A. (2006). Individual differences and contextual bias in pronoun resolution: Evidence from ERPs. *Brain Research, 1118*, 155–167.

Payne, J.S., & Whitney, P.J. (2002). Developing L2 oral proficiency through synchronous CMC: Output, working memory, and interlanguage development. *CALICO Journal, 20(1)*, 7–32.

Saito, S., & Miyake, A. (2003). On the nature of forgetting and the processing-storage relationship in reading span performance. *Journal of Memory and Language, 50*, 425–443.

Salthouse, T.A., & Kersten, A.W. (1993). Decomposing adult age differences in symbol arithmetic. *Memory and Cognition, 21*, 699–710.

Schneider, W., Eschman, A., & Zuccolotto, A. (2002). *E-Prime reference guide*. Pittsburgh, PA: Psychology Software Tools Inc.

Stone, M., Gabrieli, J.D.E., Stebbins, G.T., & Sullivan, E.V. (1998). Working and strategic memory deficits in schizophrenia. *Neuropsychology, 12*, 278–288.

Tehan, G., Hendry, L., & Kocinsky, D. (2001). Word length and phonological similarity effects in simple, complex, and delayed serial recall tasks: Implications for working memory. *Memory, 9*, 333–348.

Traxler, M.J., Williams, R.S., Blozis, S.A., & Morris, R.K. (2005). Working memory, animacy, and verb class in the processing of relative clauses. *Journal of Memory and Language, 53*, 204–224.

Van den Bos, K.P., & Spelberg, H.C.L. (1993). Reading comprehension and related skills in 9-year-old normal and poor readers. In R.M. Joshi & C.K. Leong (Eds.), *Reading disabilities: Diagnosis and component processes* (pp. 319–334). Dordrecht, NL: Kluwer.

Van den Noort, M.W.M.L., Bosch, M.P.C., & Hugdahl, K. (2006). Foreign language proficiency and working memory capacity. *European Psychologist, 11*, 289–296.

Van den Noort, M.W.M.L., Bosch, M.P.C., & Hugdahl, K. (2007). Multilingual processing of relative clauses. In F. Mansouri (Ed.), *Second language acquisition research: Theory construction and testing* (pp. 225–246). Newcastle, UK: Cambridge Scholars Press.

Van den Noort, M.W.M.L., Haverkort, M., Bosch, M.P.C., & Hugdahl, K. (2006). Is there a decline in verbal working memory over age? A study with the new standard computerized reading span test. *Europe's Journal of Psychology, 2(6)*.

Wechsler, D. (2000). *WAIS-III Nederlandstalige bewerking: Afname- en scoringshandleiding* [WAIS-III Dutch version: User manual]. Lisse, The Netherlands: Swets and Zeitlinger.

Whitney, P., Arnett, P.A., Driver, A., & Budd, D. (2001). Measuring central executive functioning: What's in a reading span? *Brain and Cognition, 45*, 1–14.

Whitney, P., Ritchie, B.G., & Clark, M.B. (1991). Working memory capacity and the use of inferences in text comprehension. *Discourse Processes, 14*, 133–145.

Maurits van den Noort

Department of Biological and Medical Psychology
University of Bergen
Jonas Lies vei 91
N-5009 Bergen
Norway
Tel. +47 555 86205
Fax +47 555 89872
E-mail maurits.noort@psybp.uib.no

Appendix

Two example sentences from the Dutch, English, German, and Norwegian RST.

Dutch RST:
Een wat onhandig geschreven artikel werd het begin van een publiek schandaal.
Het kind had een besmettelijke oorinfectie en moest hiervoor naar de dokter.

English RST:
His parents couldn't understand why he wanted a tattoo on his right shoulder.
The director was very popular, until the employees heard about his affair.

German RST:
Sie bekam Geld von ihrer Mutter, weil sie so gut geholfen hatte mit dem Umzug.
Der alte Professor sah die Einladung und freute sich schon auf den Kongress.

Norwegian RST:
En mann ble i Oslo tingsrett i går dømt til tretten års fengsel for smugling.
Utrolig, men i dag gikk den nye platen rett inn på en andreplass på lista.

