Measuring bilingual working memory capacity of professional Auslan/English interpreters: a comparison of two scoring methods

Jihong Wang
Macquarie University, Australia
Jemina Napier
Heriot-Watt University, United Kingdom

Abstract

The evaluation of working memory capacity (WMC) in signed language interpreters represents a noticeable research gap in both cognitive psychology and interpreting studies. This study compared two scoring methods – total items and proportion items – for an English listening span task and an Auslan (Australian Sign Language) working memory (WM) span task, which were administered to 31 professional Auslan/English interpreters. Given the small sample size, results reveal that the total items measure was marginally better than the proportion items measure in terms of psychometric properties. When used for statistical analyses of the interpreters' bilingual WMC, the two scoring methods yielded the same result pattern occasionally, but they also produced discrepant outcomes at times. Unlike the proportion items measure, the total items measure did not reveal statistically significant results. The total items measure was chosen as the final scoring method for this study only. These findings indicate that researchers need to be aware of methodological issues when they create and score WM span tasks.

Introduction

Working memory (WM) is a multi-component system involving temporary storage and active manipulation of information in the service of complex cognitive activities (Baddeley 2003). WM is likely to be involved in spoken and signed language interpreting, both higher-order cognitive tasks. Verbal working mem-

ory capacity (WMC) is predominantly measured by WM span tasks such as the reading span (Daneman/Carpenter 1980), listening span (ibid.), operation span (Turner/Engle 1989), and counting span tasks (Case *et al*. 1982). Despite their extensive use, there is no standard approach to creating the test materials as well as administering and scoring the tasks (Conway *et al*. 2005; Friedman/Miyake 2005; Köpke/Signorelli 2012; St Clair-Thompson/Sykes 2010).

In a listening span task (Daneman/Carpenter 1980), participants are typically instructed to listen to sets of sentences, verify whether each sentence makes sense (the processing component), and at the same time remember the final word of each sentence (the storage component) for later recall. In a reading span task (ibid.), however, participants are often required to read a series of sentences aloud, judge whether each sentence is semantically meaningful, and recall the last word of each sentence. A set of three sentences in these tasks may take the form of the following example:

> She was looking across the lobby at a man in a suit.
> The man opened the door to pick up the rain.
> The student put all the articles on the same topic into a file.

In regard to recall of the sentence-final words ("suit", "rain", and "file"), if participants recalled "suit" and "file", should they receive a score of 2 (they correctly recalled 2 words – total items), a score of 0.67 (they successfully recalled 2 out of 3 words – proportion items), a score of 0 (they did not correctly recall the whole set of 3 words – correct sets items), or else? Uncertainty about the best scoring method for WM span tasks constitutes an identifiable gap in our knowledge.

A large number of studies have explored spoken language interpreters' WMC. Some studies found that professional interpreters significantly outperformed student interpreters and/or non-interpreters on a reading span task (Christoffels *et al*. 2006; Padilla *et al*. 1995; Signorelli *et al*. 2012). Other studies, however, revealed that professional interpreters performed similarly to student interpreters and/or non-interpreters on a listening span task (Köpke/Nespoulous 2006; Liu *et al*. 2004; Timarová 2007). Although recall order (serial recall versus free recall) of to-be-remembered items may be responsible for the contradictory results (Köpke/Signorelli 2012), scoring methods and other methodological issues are also likely to be at play.

Signed language interpreters' WMC is a significant research gap in cognitive psychology and interpreting studies. The present study has two aims: (i) to compare the total items with the proportion items for an English listening span task and an Auslan (Australian Sign Language) WM span task, in order to determine whether one scoring method is psychometrically more favourable than the other; and (ii) to ascertain whether the two scoring methods produce the same result pattern for statistical analyses of professional Auslan/English interpreters' bilingual WMC. The study will shed new light on the measurement of WMC in both spoken and signed language interpreters. In order to contextualize the research design of this study, an overview of the methodological differences in verbal WM span tasks is presented.

1. Overview

Verbal WM span tasks (especially listening span tasks and reading span tasks) in both psychological literature and interpreting studies literature differ in test materials, administration procedure, and scoring methods.

1.1 Test materials

Differences in test materials are mainly about test sentences and to-be-remembered items. A survey of the literature reveals that the number of test sentences in WM span tasks typically ranges from 42 (Christoffels *et al.* 2006) to 100 (Daneman/Hannon 2001). Waters/Caplan (1996) found a significantly smaller reading span for complex test sentences than for simple test sentences, suggesting that the difficulty level of test stimuli may have an impact on WMC. The to-be-remembered items vary from sentence-final words (Daneman/Carpenter 1980; Liu *et al.* 2004), unrelated words after test sentences (La Pointe/Engle 1990), to irrelevant letters following test sentences (Unsworth *et al.* 2009). Additionally, some studies have revealed the *word length effect* on WMC[1] (La Pointe/Engle 1990), the *phonological similarity effect* on WMC[2] (Lobley *et al.* 2005), and the *word frequency effect* on WMC[3] (Engle *et al.* 1990). These findings indicate that test sentences and the to-be-remembered items in verbal WM span tasks require careful control.

1.2 Administration procedure

Differences in the administration procedure of verbal WM span tasks mostly involve presentation order of test sentences, processing component, response modality, processing time, test termination point, recall order of the to-be-remembered items, and recall modality. These terms are explained briefly below. Regarding the presentation order of test sentences, shorter sets of sentences are often presented prior to increasingly longer sets of sentences (i.e. ascending order, see Daneman/Hannon 2001); but there are also variations in which the sets are randomized (Timarová 2007; Unsworth *et al.* 2009). As noted earlier, processing component refers to what participants are asked to do while retaining the to-be-remembered items. Processing component involves listening to test sentences and judging sentence meaningfulness (Liu *et al.* 2004, Timarová 2007), or reading aloud test sentences only (Christoffels *et al.* 2006; Friedman/Miyake 2004), or reading aloud test sentences and verifying sentence meaningfulness (Daneman/Hannon 2001). In regard to response modality, participants verify sentence meaningfulness either manually (Liu *et al.* 2004; Unsworth *et al.* 2009)

1 Short to-be-remembered words resulted in a larger WM span than long words.
2 To-be-remembered words that were phonologically distinct yielded a larger WM span than phonologically similar words.
3 To-be-remembered words with high frequency resulted in a larger WM span than low-frequency words.

or orally. Lobley *et al*. (2005) found a significantly larger listening span for manual response than for verbal response to sentence verification.

Processing time, namely the time interval between adjacent sentences for performing the processing task, varies across different studies (e.g. 1 second in Timarová 2007, 2 seconds in Stafford 2011). The processing time should be just sufficient for participants to carry out the processing task, as Friedman/Miyake (2004) found that allowing extra time for participants to rehearse the to-be-remembered words reduced the correlation between the participants' reading span and their reading comprehension. In relation to test termination point, some studies present all test sentences to participants (Daneman/Hannon 2001), while other studies discontinue testing when participants do not recall a majority of the sets at a particular span level (Lobley *et al*. 2005). It is preferable to present all test stimuli to participants, because this procedure not only allows them to showcase their full potential in performing a task but also permits researchers to use a wider range of scoring methods. As for recall order of the to-be-remembered items, participants are required to recall them in serial order (ibid.), or in completely random order (Alptekin/Erçetin 2009), or in other orders (Daneman/Hannon 2001; Friedman/Miyake 2004). With regard to recall modality, some studies instruct participants to recall the to-be-remembered items orally (Daneman/Carpenter 1980; Meredyth Daneman, personal communication, 8 July 2010), other studies manually (Liu *et al*. 2004; Unsworth *et al*. 2009). These differences in administration procedure need to be taken into consideration when researchers interpret previous findings or create WM span tasks.

1.3 Scoring methods

There is no consensus on how to score WM span tasks. Scoring methods typically include: *truncated span*, *correct sets items*, *total items*, and *proportion items* (see also Friedman/Miyake 2005). Traditionally, researchers mark WM span tasks using truncated span. This typically involves scoring for the highest span level at which recall is correct on a majority of sets, and giving half a credit if recall is correct on few sets at the subsequent span level (Daneman/Carpenter 1980; Liu *et al*. 2004; Padilla *et al*. 1995). Alternatively, the correct sets items measure involves summing up the items in only perfectly recalled sets (Conway *et al*. 2002). As illustrated by the example in the introduction of this article, the total items measure involves calculating the total number of correctly recalled items across all sets (Christoffels *et al*. 2006; Daneman/Hannon 2001). The proportion items measure, however, involves calculating the proportion of correctly recalled items for each set, and then computing the average proportion of correct recall across all sets (Kane *et al*. 2004).

Currently, the majority of researchers in cognitive psychology and cognitive science score WM span tasks using either the total items or the proportion items. This is because the total items and the proportion items are better than both the truncated span and the correct sets items in terms of normal distribution, internal reliability, test-retest reliability, and criterion-related validity (Conway *et*

al. 2005; Friedman/Miyake 2005; St Clair-Thompson/Sykes 2010). Nevertheless, the body of literature does not clearly stipulate how to choose between the total items and the proportion items. Friedman/Miyake (2005) found that WM span scores obtained using the total items and the proportion items are almost perfectly correlated, suggesting that there is a close relationship between the two scoring methods. Conway *et al*. (2005) found that the proportion items measure is slightly superior to the total items measure in terms of internal reliability. In contrast, Friedman/Miyake (2005) noted that the total items measure is marginally better than the proportion items measure in terms of internal reliability and test-retest reliability. According to Conway *et al*. (2005: 776), researchers may choose the proportion items measure ("partial-credit unit scoring") because "it follows established and sound procedures from psychometrics". Nonetheless, Friedman/Miyake (2005: 589) made the following comments:

> The total words [items] score may be preferable in that it is easier to compute and conceptually more direct (it simply counts up the number of words recalled, with no weighting for the levels at which the words were recalled). However, it may also make sense to penalize more for the forgetting of words at easier levels, as the proportion words [items] score does.

Given the ambiguity surrounding the best scoring method for WM span tasks, further investigation is warranted. The present study therefore compares the total items with the proportion items for an English listening span task and an Auslan WM span task.

2. Method

2.1 Participants

Participants were 31 professional level Auslan/English interpreters qualified by the National Accreditation Authority for Translators and Interpreters (NAATI)[4] in Australia. They included 14 native signers (11 female and 3 male; mean age 40,

4 In Australia, all signed language interpreters, spoken language interpreters, and translators are accredited through NAATI. Two accreditation levels are available for Auslan/English interpreters: Paraprofessional Interpreter and Professional Interpreter. Paraprofessional Interpreters typically undertake the interpretation of non-specialist dialogues. Professional Interpreter is the minimum level recommended by NAATI for professional interpreting work in most semi-specialized settings, such as banking, law, health, and social and community services. Although 932 Auslan/English interpreters have been accredited by NAATI since testing started in 1982, only 160 have received accreditation at professional level (Robert Foote, Accreditation Manager, NAATI, personal communication, 3 October 2012). It should be noted that not all of those who are professionally accredited are currently working as interpreters. Thus, the sample size of this study was considered to be good in relation to the actual population size.

SD = 14; with an average of 16 years of interpreting experience[5], *SD* = 9) and 17 non-native signers (16 female and 1 male; mean age 40, *SD* = 9; with an average of 14 years of interpreting experience, *SD* = 7). The native signers acquired Auslan from birth from their signing deaf parents and at the same time acquired English through interaction with the surrounding hearing population[6]. The non-native signers acquired English from birth from their hearing parents, and often had no family connections to the Australian Deaf Community, but started to learn Auslan at or after age 10 by receiving formal education in Auslan and/or associating with deaf signers through work or social networks.

2.2 Materials

2.2.1 English listening span task

This task required participants to listen to sets of English sentences, judge whether each sentence made sense (say "yes" if it made sense or "no" if not), at the same time remember the final word of each sentence, and at the end of each set utter all sentence-final words in serial order. The task consisted of four sets each of two, three, four, five, six, and seven unrelated sentences, 108 sentences in total. Each participant listened to four sets of two sentences, then four sets of three sentences, and so on in ascending order up to four sets of seven sentences, completing the task by listening to all 108 sentences. The time for verifying each sentence was one second. The time for recalling each sentence-final word was four seconds, which was observed to be sufficient for participants to complete their recall.

Test sentences varied from nine to 13 words and were easy to understand. Ninety-six sentences made sense and 12 sentences were purposefully nonsensical. Sentence-final words were monosyllabic nouns (e.g., "team", "card", "nurse", "phone") controlled for concreteness, frequency, length, and phonetic structure. All test sentences were recorded by a native English speaker, edited, and saved as an mp3 file.

2.2.2 Auslan working memory span task

This task followed the same structure and administration procedure as the English listening span task. It instructed participants to watch sets of Auslan sentences on a video, verify whether each sentence made sense (sign YES if it made sense or NO if not), at the same time memorize the final sign of each sentence, and at the end of each set reproduce all sentence-final signs in serial order.

All Auslan test sentences were created by a female deaf near-native signer[7].

5 Years of working as a paid interpreter, irrespective of NAATI accreditation.

6 It is important to note that not all hearing children with deaf parents may necessarily be native signers.

7 Due to the small population of deaf native signers, it was difficult to find a deaf native

She referred to Auslan dictionary resources to ensure that standard Auslan signs were used[8]. Test sentences varied from six to 10 Auslan signs and were easy to understand. Eighty-six sentences made sense and 22 sentences were purposefully nonsensical. Sentence-final signs in the Auslan WM span task were simple and commonly used, and were controlled for handshape, orientation, location, and movement. All test sentences were articulated by the deaf near-native signer, filmed, edited, and saved as an mp4 video.

2.3 Procedure

Participants were tested individually. After filling out a consent form and a demographic questionnaire, each participant completed the English listening span task and then the Auslan WM span task shown on a laptop computer. They received task instructions, participated in a practice session, and proceeded to the tasks. Participants were filmed during both tasks for later analysis.

2.4 Data scoring

Participants' WMC scores were obtained using the total items and the proportion items to validate previous findings. English WMC was abbreviated as E and Auslan WMC as A. Table 1 shows the scoring process, taking a randomly chosen participant as example.

Step 1: Decide whether recall of a sentence-final word/sign was correct		
Span level	English listening span task	Auslan WM span task
4 sets at span level 2	2 + 2 + 2 + 2	2 + 2 + 2 + 2
4 sets at span level 3	3 + 3 + 2 + 3	3 + 3 + 3 + 3
4 sets at span level 4	3 + 3 + 3 + 3	3 + 3 + 1 + 4
4 sets at span level 5	4 + 5 + 4 + 3	5 + 5 + 4 + 5
4 sets at span level 6	4 + 6 + 3 + 4	6 + 3 + 4 + 4
4 sets at span level 7	5 + 4 + 4 + 6	6 + 4 + 6 + 5

signer who was available to create Auslan test sentences for this study. The female deaf near-native signer was selected because she was a highly fluent Auslan signer and had worked as an Auslan model for the Auslan Signbank (http://www.auslan.org.au/), an online Auslan dictionary.

8 At least two native signers checked all Auslan test sentences before filming. During filming, a native signer monitored the deaf near-native signer to ensure that the sentences looked as natural as possible.

Step 2: Apply two scoring methods		
Scoring methods	English WMC	Auslan WMC
Total items (E1, A1)	2 + 2 + 2 + 2 + 3 + 3 + 2 + 3 + 3 + 3 + 3 + 3 + 4 + 5 + 4 + 3 + 4 + 6 + 3 + 4 + 5 + 4 + 4 + 6 E1 = 83	2 + 2 + 2 + 2 + 3 + 3 + 3 + 3 + 3 + 3 + 1 + 4 + 5 + 5 + 4 + 5 + 6 + 3 + 4 + 4 + 6 + 4 + 6 + 5 A1 = 88
Proportion items (E2, A2)	((2 + 2 + 2 + 2)/2 + (3 + 3 + 2 + 3)/3 + (3 + 3 + 3 + 3)/4 + (4 + 5 + 4 + 3)/5 + (4 + 6 + 3 + 4)/6 + (5 + 4 + 4 + 6)/7)/24 E2 = 0.81	((2 + 2 + 2 + 2)/2 + (3 + 3 + 3 + 3)/3 + (3 + 3 + 1 + 4)/4 + (5 + 5 + 4 + 5)/5 + (6 + 3 + 4 + 4)/6 + (6 + 4 + 6 + 5)/7)/24 A2 = 0.85

Table 1. Scoring process as illustrated by a participant's data

- Total items (E1, A1): the total number of correctly recalled items across all sets, with the maximum possible score being 108. The number of correctly recalled items for each set was calculated, and then added up across all 24 sets. For example, if participants recalled 3 out of 6 words in a set of 6 sentences (at span level 6), they received 3 points for that set. Since this method rewarded every correctly recalled item, it picked up subtle individual differences.
- Proportion items (E2, A2): the average proportional recall for each set, with the maximum possible score being 1.00. The proportion of correctly recalled items was calculated for each set (e.g. if participants recalled 4 out of 5 words in a set of 5 sentences at span level 5, they received 0.80 for that set), and then the proportional recall for all 24 sets was averaged. Although this scoring method also rewarded every item correctly recalled, it gave different weightings to one item at span level 2 (a weighting of 0.50) and another item at span level 7 (a weighting of 0.14). Hence, forgetting items at low span levels would result in lower scores than failing to recall the same number of items at high span levels. For example, forgetting one word in a set of two sentences resulted in a score of 0.50 for that set while forgetting one word in a set of seven sentences resulted in a score of 0.86 for that set.

3. Results

Both scoring methods were examined in terms of psychometric properties, and then used for statistical analyses of participants' bilingual WMC.

3.1 Psychometric properties

3.1.1 Correlations between the two scoring methods

For all 31 participants' English WMC, the total items (E1) correlated almost perfectly with the proportion items (E2), $r = 0.99$, $N = 31$, $p < 0.001$, $r^2 = 0.98$. Removal of the outlier (see Table 2) did not alter that result pattern. Moreover, for all 31 participants' Auslan WMC, the total items (A1) also correlated nearly perfectly with the proportion items (A2), $r = 0.995$, $N = 31$, $p < 0.001$, $r^2 = 0.99$. As in Friedman/Miyake's (2005) study, the virtually perfect correlations indicate that the total items measure is closely related to the proportion items measure.

3.1.2 Normal distribution

Table 2 below provides descriptive statistics for the two scoring methods.

All participants' English WMC (N = 31)						
English WMC	Mean (*SD*)	Skewness	Kurtosis	*p* for Shapiro-Wilk	Outlier	Cronbach's α
E1	74.58 (13.21)	-0.53	0.84	0.34	1	0.931
E2	0.76 (0.11)	-1.02	2.44	0.06	1	0.928
30 participants' English WMC (N = 30, without the above outlier)						
English WMC	Mean (*SD*)	Skewness	Kurtosis	*p* for Shapiro-Wilk	Outlier	Cronbach's α
E1	75.83 (11.41)	0.12	-0.68	0.39	0	0.911
E2	0.77 (0.09)	-0.004	-0.76	0.56	0	0.899
All participants' Auslan WMC (N = 31)						
Auslan WMC	Mean (*SD*)	Skewness	Kurtosis	*p* for Shapiro-Wilk	Outlier	Cronbach's α
A1	70.65 (20.62)	-0.20	-0.90	0.49	0	0.962
A2	0.71 (0.17)	-0.38	-0.77	0.28	0	0.952
Note. Total items (E1, A1). Proportion items (E2, A2). Standard error for Skewness = 0.42. Standard error for Kurtosis = 0.82.						

Table 2. Descriptive statistics and reliability estimates for the two scoring methods

– If the absolute values (ignoring the sign at the front) of skewness and kurtosis are closer to zero, the distribution characteristics are more favourable. The absolute values of skewness and kurtosis for the total items (E1, A1) were consistently below 1; whereas those for the proportion items (E2, A2) were occasionally above 1 (see E2 for all 31 participants). These results indicate that in this study the total items measure appeared to be slightly superior to the proportion items measure in terms of distribution characteristics.

– Regarding the Shapiro-Wilk Test of Normality, non-significant results ($p > 0.05$) as shown in Table 2 indicate normal distribution. Both scoring methods, therefore, were deemed satisfactory in terms of normality.
– Each scoring method caused very few outliers. An outlier was defined as a "data point" that extended more than 2.50 standard deviations (*SD*) from the mean (*M*). The only outlier of E1 was a native signer with a score of 37. The outlier of E2 was the same participant with a score of 0.40.

3.1.3 Internal reliability

Table 2 also provides the internal reliability estimate (Cronbach's α) for each scoring method. Cronbach's α was calculated with the scores summed across levels for the first set at each span level, the second set at each span level, and so on through to the last (the fourth) set at each span level (see also Friedman/Miyake 2005; St Clair-Thompson/Sykes 2010). Although both scoring methods exhibited remarkably high internal reliability in this study, the total items measure (E1, A1) appeared to be marginally more satisfactory than the proportion items measure (E2, A2).

3.2 Effects of three variables on working memory capacity

A 2x2x2 three-way mixed between-within subjects analysis of variance (ANOVA) was conducted to examine the effects of age of signed language acquisition (native signer, non-native signer), test language (English, Auslan), and scoring method (total items, proportion items) on WMC.

There was no significant main effect for age of signed language acquisition, $F(1, 29) = 2.10$, $p = 0.16$. In other words, if we ignore whether the test language was English or Auslan as well as the type of scoring method used, the native signers were similar to the non-native signers in WMC.

Likewise, there was no significant main effect for test language, $F(1, 29) = 2.27$, $p = 0.14$. Namely, if we ignore the type of scoring method used, all 31 participants' English WMC was comparable to their Auslan WMC.

Nonetheless, there was a significant main effect for scoring method, $F(1, 29) = 656.54$, $p < 0.001$, partial $\eta^2 = 0.96$ (a very large effect size). That is, if we ignore whether participants were native signers or non-native signers as well as whether the test language was English or Auslan, WMC using the total items measure was significantly larger than WMC using the proportion items measure. This finding is not surprising, because the maximum possible WMC score using the total items was 108 while the highest possible WMC score using the proportion items was 1.00.

The interaction effect between test language and age of signed language acquisition was not statistically significant, $F(1, 29) = 0.35$, $p = 0.56$. This means that the effect of different test languages on WMC was the same for the native signers and the non-native signers.

There was no significant interaction between scoring method and age of signed language acquisition, $F(1, 29) = 2.10$, $p = 0.16$. Specifically, the effect of different scoring methods on WMC was the same for the native signers and the non-native signers.

There was no significant interaction between test language and scoring method, $F(1, 29) = 2.23$, $p = 0.15$. To be specific, the effect of different test languages on WMC was the same for the total items and the proportion items.

Finally, there was no significant interaction between age of signed language acquisition, test language, and scoring method, $F(1, 29) = 0.36$, $p = 0.56$. In other words, the above interaction effect of test language and age of signed language acquisition was the same for the total items and the proportion items. It should be noted that removal of the outlier (see Table 2) did not change any of the result patterns.

3.3 Impact of two scoring methods on the overall results

In addition to the aforementioned main effects and interaction effects, we conducted the following *t*-tests to explore simple effects. We compared the native signers with the non-native signers in terms of WMC, and contrasted each group's English WMC with their Auslan WMC, so as to investigate whether both scoring methods produce the same result pattern for the current study.

3.3.1 Native signers versus non-native signers in terms of working memory capacity

According to an independent-samples *t*-test (see Table 3 below), both the total items (E1) and the proportion items (E2) showed that the native signers were significantly outperformed by the non-native signers on the English listening span task, with both *p* values (0.041 and 0.046) below 0.05.

English WMC	Native signers (N = 14)	Non-native signers (N = 17)	*t*	*df*	*p*	η^2
	Mean (*SD*)	Mean (*SD*)				
Total items (E1)	69.29 (13.41)	78.94 (11.66)	-2.14	29	0.041	0.14
Proportion items (E2)	0.71 (0.12)	0.79 (0.09)	-2.09	29	0.046	0.13
Note. Both *p* values were two-tailed. Bold type indicated $p < 0.05$. The effect size η^2 was calculated when there was a significant difference.						

Table 3. Native signers versus non-native signers in terms of English WMC

Regarding participants' English WMC, a sensitivity analysis was then undertaken to examine the impact of outliers on the above result pattern. When the only outlier (see Table 2) was removed, the total items (E1) revealed no significant difference between the native signers (M = 71.77, SD = 10.07) and the non-native signers (M = 78.94, SD = 11.66) in English WMC, $t(28) = -1.77$, $p = 0.09$ (two-tailed). When the outlier was excluded, the proportion items (E2) also revealed no significant difference between the native signers (M = 0.74, SD = 0.08) and the non-native signers (M = 0.79, SD = 0.09) in English WMC, $t(28) = -1.71$, $p = 0.10$ (two-tailed).

In relation to participants' Auslan WMC, according to an independent-samples t-test in Table 4 below, the total items (A1) showed that the native signers were similar to the non-native signers in Auslan WMC, $t(29) = -0.89$, $p = 0.38$. The proportion items (A2) demonstrated the same result pattern, with the p value (0.35) larger than 0.05.

Auslan WMC	Native signers (N = 14)	Non-native signers (N = 17)	t	df	p	η^2
	Mean (*SD*)	Mean (*SD*)				
Total items (A1)	67.00 (20.08)	73.65 (21.18)	-0.89	29	0.38	-
Proportion items (A2)	0.68 (0.17)	0.73 (0.17)	-0.94	29	0.35	-
Note. Both p values were two-tailed. The effect size η^2 was calculated when there was a significant difference.						

Table 4. Native signers versus non-native signers in terms of Auslan WMC

3.3.2 English working memory capacity versus Auslan working memory capacity

Table 5 below summarizes the results of a paired-samples t-test for a comparison between the native signers' English WMC and their Auslan WMC. The total items revealed no significant difference between the native signers' English WMC and their Auslan WMC, $t(13) = 0.52$, $p = 0.61$. The proportion items produced the same outcome, with the p value (0.33) larger than 0.05. These results remained unchanged when the outlier (see Table 2) was excluded.

Scoring methods	English WMC (N = 14)	Auslan WMC (N = 14)	*t*	*df*	*p*	η^2
	Mean (SD)	Mean (SD)				
Total items (E1, A1)	69.29 (13.41)	67.00 (20.08)	0.52	13	0.61	-
Proportion items (E2, A2)	0.71 (0.12)	0.68 (0.17)	1.02	13	0.33	-
Note. Both *p* values were two-tailed. The effect size η^2 was calculated when there was a significant difference.						

Table 5. Native signers' English WMC versus their Auslan WMC

Table 6 shows the outcomes of a paired-samples *t*-test for a comparison between the non-native signers' English WMC and their Auslan WMC. The total items indicated that the non-native signers' English WMC was comparable to their Auslan WMC, $t(16) = 1.87$, $p = 0.08$. In contrast, the proportion items revealed that the non-native signers' English WMC was significantly better than their Auslan WMC, $t(16) = 2.38$, $p = 0.03$, $\eta^2 = 0.26$.

Scoring methods	English WMC (N = 17)	Auslan WMC (N = 17)	*t*	*df*	*p*	η^2
	Mean (*SD*)	Mean (*SD*)				
Total items (E1, A1)	78.94 (11.66)	73.65 (21.18)	1.87	16	0.08	-
Proportion items (E2, A2)	0.79 (0.09)	0.73 (0.17)	2.38	16	**0.03**	0.26
Note. Both *p* values were two-tailed. Bold type indicated $p < 0.05$. The effect size η^2 was calculated when there was a significant difference.						

Table 6. Non-native signers' English WMC versus their Auslan WMC

4. Discussion

The primary goal of this study was to compare the total items with the proportion items in terms of psychometric properties. Our results demonstrate that both the total items and the proportion items are satisfactory psychometrically. This finding may be explained by the fact that both scoring methods capture subtle individual differences and include additional information from high span levels. The additional information from high span levels may reflect the involvement of both the central executive (a cognitive system that controls attention)

and secondary memory (controlled search and retrieval processes) in WM span tasks (St Clair-Thompson/Sykes 2010). More importantly, our results appear to show that the total items measure was marginally superior to the proportion items measure in terms of both normal distribution and internal reliability, supporting Friedman/Miyake's (2005) results but contradicting Conway *et al.*'s (2005) findings.

The second goal of this study was to explore whether the total items and the proportion items yield the same result pattern for statistical analyses of the professional Auslan/English interpreters' bilingual WMC. The study found that sometimes the total items and the proportion items produced consistent result patterns. Both scoring methods showed that the native signers were similar to the non-native signers not only in English WMC but also in Auslan WMC. Further, both scoring methods revealed that the native signers' English WMC was similar to their Auslan WMC. These findings may be due to the small sample size of the present study. These findings provide additional evidence that using different scoring methods may result in the same outcome (see also Friedman/Miyake 2004; La Pointe/Engle 1990; Turner/Engle 1989).

Another interesting finding was that the total items and the proportion items occasionally produced contradictory result patterns. The total items showed that the non-native signers' English WMC was similar to their Auslan WMC, suggesting that their bilingual WMC is domain-general (independent of the test language). In contrast, the proportion items revealed that the non-native signers' English WMC was significantly larger than their Auslan WMC, suggesting that their bilingual WMC is domain-specific (dependent on the test language). These findings indicate that different scoring methods may provide some degree of freedom for researchers attempting to identify and report statistical significance.

We recommend that researchers use different scoring methods to mark WM span tasks and then select the most appropriate scoring method to report their results. A choice among various scoring methods depends on the sample size, the psychometric properties of the actual WMC data, the perspectives defended by the researchers, and other factors. Therefore, the decision on the best scoring method for WM span tasks remains open-ended and should be made on a case-by-case basis. For the current study only, we made an executive decision to choose the total items rather than the proportion items for the following four reasons:

- With regard to the WMC data of this study, the total items measure appears to be slightly better than the proportion items measure in terms of both distribution characteristics and internal reliability.
- High span levels are more challenging than low span levels; perfect recall of a set of 7 words therefore deserves more credit than correct recall of a set of 2 words. The total items measure gives 7 points to the former and 2 points to the latter, whereas the proportion items measure gives the same score of 1 to each case.
- The total items measure was practically easier to calculate and conceptually more direct (see also Friedman/Miyake 2005).

– The total items measure did not involve rounding up decimals, whereas the proportion items measure constantly required rounding up decimals and thus, might have excluded individual difference information.

Like some similar studies (e.g. Friedman/Miyake 2005), the present study revealed that outliers at times influenced the overall outcomes. The removal of the only outlier eliminated the significant differences between the native signers and the non-native signers in English WMC. This finding suggests that it is important to examine the impact of outliers on the overall results, especially for small sample sizes.

Limitations of this study need to be acknowledged. The sample size of 31 participants is very small for the analysis of psychometric properties; consequently, our results regarding the psychometric properties should be interpreted with caution. The small sample size may also limit the statistical power of our results from the ANOVA and *t*-tests. The low complexity of test sentences in both WM span tasks may shift up participants' WMC scores and thus influence the results, as Waters/Caplan (1996) found that simple test sentences resulted in a considerably larger WM span than complex test sentences. The to-be-remembered signs in the Auslan WM span task need to be more carefully controlled to minimize sign variants in the Auslan WMC data. Articulation length, part of speech, and phonological similarity of the to-be-remembered signs and their spoken language translation equivalents are also worthy of consideration. The proportion of semantically meaningful test sentences in each WM span task should be 50%. Different set sizes in each WM span task should be administered in random order rather than in ascending order. To minimize order effects, the presentation order of the two WM span tasks should be counterbalanced.

5. Conclusion

This study has compared two scoring methods – the total items and the proportion items – for an English listening span task and an Auslan WM span task. Professional Auslan/English interpreters consisting of native signers and non-native signers completed the two WM span tasks. Our results appear to show that the total items measure was slightly more favourable than the proportion items measure in terms of psychometric properties, although both scoring methods were highly satisfactory.

Moreover, the study found that the two scoring methods produced consistent outcomes occasionally, whereas at other times they yielded different conclusions. Both scoring methods revealed that the native signers were similar to the non-native signers not only in English WMC, but also in Auslan WMC. Further, both scoring methods showed that the native signers' English WMC was as good as their Auslan WMC. Nevertheless, the two scoring methods produced discrepant results for a comparison between the non-native signers' English WMC and their Auslan WMC – the proportion items produced statistically significant results while the total items did not. For this study only, we selected the total items over the proportion items as the final scoring method. The study demonstrates

that researchers should be mindful of which scoring method they use, as it may have an impact on statistical significance and the overall conclusion.

Further research is needed to compare a range of scoring methods for various WM span tasks, using large samples to achieve strong statistical power. To help researchers decide on the most appropriate scoring method, future studies may recruit two different groups with known properties and then test whether a specific scoring method detects the properties correctly. It would also be of interest to ascertain whether different WM span tasks in the literature measure the same WMC construct. In addition, large-scale empirical studies of signed language interpreting are needed to cross-reference with spoken language interpreting studies.

Acknowledgments

This article presents selected findings from Jihong Wang's PhD thesis entitled "Working Memory and Signed Language Interpreting" submitted to Macquarie University in 2013. We would like to thank all professional Auslan/English interpreters for participating in this study. We are very grateful to Trevor Johnston, Andy Carmichael, and Della Goswell for sharing with us their views on Auslan sign variation.

References

Alptekin C. / Erçetin G. (2009) "Assessing the relationship of working memory to L2 reading: does the nature of comprehension process and reading span task make a difference?", *System* 37/4, 627-639.

Baddeley A. D. (2003) "Working memory and language: an overview", *Journal of Communication Disorders* 36/3, 189-208.

Case R. / Kurland D. M. / Goldberg J. (1982) "Operational efficiency and the growth of short-term memory span", *Journal of Experimental Child Psychology* 33, 386-404.

Christoffels I. K. / De Groot A. M. B. / Kroll J. F. (2006) "Memory and language skills in simultaneous interpreters: the role of expertise and language proficiency", *Journal of Memory and Language* 54/3, 324-345.

Conway A. R. A. / Cowan N. / Bunting M. F. / Therriault D. J. / Minkoff S. R. B. (2002) "A latent variable analysis of working memory capacity, short-term memory capacity, processing speed, and general fluid intelligence", *Intelligence* 30/2, 163-183.

Conway A. R. A. / Kane M. J. / Bunting M. F. / Hambrick D. Z. / Wilhelm O. / Engle R. W. (2005) "Working memory span tasks: a methodological review and user's guide", *Psychonomic Bulletin and Review* 12/5, 769-786.

Daneman M. / Carpenter P. A. (1980) "Individual differences in working memory and reading", *Journal of Verbal Learning and Verbal Behavior* 19/4, 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* 130/2, 208-223.
Engle R. W. / Nations J. K. / Cantor J. (1990) "Is 'working memory capacity' just another name for word knowledge?", *Journal of Educational Psychology* 82/4, 799-804.
Friedman N. P. / Miyake A. (2004) "The reading span test and its predictive power for reading comprehension ability", *Journal of Memory and Language* 51/1, 136-158.
Friedman N. P. / Miyake A. (2005) "Comparison of four scoring methods for the reading span test", *Behavior Research Methods* 37/4, 581-590.
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* 133/2, 189-217.
Köpke B. / Nespoulous J.-L. (2006) "Working memory performance in expert and novice interpreters", *Interpreting* 8/1, 1-23.
Köpke B. / Signorelli T. M. (2012) "Methodological aspects of working memory assessment in simultaneous interpreters", *International Journal of Bilingualism* 16/2, 183-197.
La Pointe 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* 16/6, 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/8, 1462-1478.
Padilla P. / Bajo M. T. / Cañas J. J. / Padilla F. (1995) "Cognitive processes of memory in simultaneous interpretation", in J. Tommola (ed.), *Topics in Interpreting Research*, Turku, University of Turku, 61-71.
Signorelli T. M. / Haarmann H. J. / Obler L. K. (2012) "Working memory in simultaneous interpreters: effects of task and age", *International Journal of Bilingualism* 16/2, 198-212.
Stafford C. A. (2011) "Bilingualism and enhanced attention in early adulthood", *International Journal of Bilingual Education and Bilingualism* 14/1, 1-22.
St Clair-Thompson H. L. / Sykes S. (2010) "Scoring methods and the predictive ability of working memory tasks", *Behavior Research Methods* 42/4, 969-975.
Timarová Š. (2007) *Measuring Working Memory in Interpreters*, unpublished DEA (pre-doctoral) thesis, University of Geneva.
Turner M. L. / Engle R. W. (1989) "Is working memory capacity task dependent?", *Journal of Memory and Language* 28/2, 127-154.
Unsworth N. / Redick T. S. / Heitz R. P. / Broadway J. M. / Engle R. W. (2009) "Complex working memory span tasks and higher-order cognition:

a latent-variable analysis of the relationship between processing and storage", *Memory* 17/6, 635-654.
Wang J. (2013) *Working Memory and Signed Language Interpreting*, unpublished doctoral dissertation, Macquarie University.
Waters G. S. / Caplan D. (1996) "The measurement of verbal working memory capacity and its relation to reading comprehension", *The Quarterly Journal of Experimental Psychology* 49A/1, 51-79.