

# SCIENTIFIC REPORTS



OPEN

## The reliability of estimating visual working memory capacity

Mengnuo Dai<sup>1,2</sup>, Yanju Li<sup>1,2</sup>, Shuoqiu Gan<sup>1</sup> & Feng Du<sup>1,2</sup>

The reliability of estimations of working memory capacity has not been thoroughly examined. The present study examined the test-retest reliability for working memory capacity as estimated in a lateralized change detection task, which is frequently used in studies involving electroencephalography. The test-retest correlations between K values for each set size in the two tests varied from 0.502 to 0.757, with test-retest correlations rising as set size increased. The results indicate that individual visual working memory capacity can be reliably estimated in a change detection task. Furthermore, test-retest reliability was higher when the two tests occurred at the same time of day than at different times of day.

Working memory is a temporary storage system under active attentional control<sup>1</sup>. The capacity of working memory is highly limited. For example, on average, people can only hold 3–4 items in visual working memory<sup>2–4</sup>. Although having a limited capacity, working memory has been shown to play an essential role in cognitive functions, such as fluid intelligence<sup>5</sup>, reasoning<sup>6</sup>, reading comprehension<sup>7,8</sup>, second language proficiency of adult learners<sup>9</sup>, and academic attainment<sup>10</sup>.

However, although research has involved an assumption that different kinds of estimation are all reliable measurements for working memory capacity, this has not been thoroughly examined. Initially, researchers focused on verbal working memory by using paradigms such as verbal recall tests<sup>8</sup>, the classic digit N-back task<sup>11</sup>, and the number counting task<sup>12</sup>. The test-retest reliability of working memory capacity estimation based on these tasks has been examined. For example, a study used verbal recall tasks to test 4–15-year-old children and retested them after 5–10 days. The results showed that the verbal recall tasks had an average test-retest reliability of 0.56 for non-word-syllables, 0.72 for words, 0.81 for digits, and 0.61 for sentences and counting numbers<sup>13</sup>. Another study showed that the test-retest reliability of working memory capacity in a spatial N-back paradigm varies from 0.493 to 0.857, depending on the N and whether accuracy or reaction time of the task is calculated<sup>14</sup>.

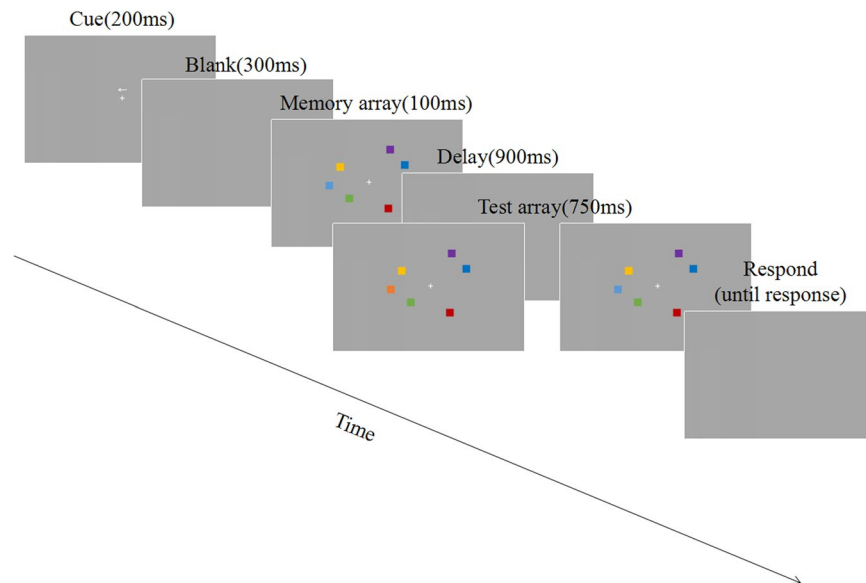
Studies on visual working memory (VWM) have relied on the change detection task to quantify VWM capacity<sup>2,3</sup>. Participants usually show a limitation of visual working memory capacity at about 3–4 objects' worth of information<sup>4</sup>, which has been linked with visual search<sup>15</sup> and multiple-object tracking performance<sup>16</sup>. A study also revealed that individuals with higher VWM capacity are more efficient at excluding unnecessary items during task performance<sup>17</sup>.

Since researchers use an estimation of individual VWM capacity to account for individual differences in other cognitive functions, it is essential to evaluate the reliability of VWM capacity estimation based on the change detection task. However, only a few studies have examined the reliability of VWM capacity estimation. For instance, Johnson *et al.*<sup>18</sup> examined the test-retest reliability of VWM capacity estimation with a separation of 1.5 years between the two tests. They found a test-retest correlation of 0.77. However, since their study only sampled 31 schizophrenia patients and 14 healthy participants, the representativeness of the result is questionable. Thus, whether VWM capacity estimation is reliable across tests remains essentially unknown.

Test-retest reliability is critical not only for measuring individual differences but for the reproducibility of psychological findings. For example, a large reproducibility project<sup>19</sup> examined the reproducibility of 100 experiments published in social and cognitive psychology journals. The results showed that only 38 of 100 experiments were rated to have replicated the original results. Thus, the test-retest reliability of VWM capacity estimation provides a baseline level of the reproducibility of VWM capacity estimation given that the subjects and task are the same in the two tests.

Some researchers have suggested that VWM capacity may not have a satisfactory internal consistency across different set sizes in a single test<sup>20</sup> because estimation of visual working memory capacity relies heavily on the set size (number of to-be-remembered items) of visual stimuli in the change detection task<sup>2,3</sup>. It has also been

<sup>1</sup>CAS Key Laboratory of Behavioral Science, Institute of Psychology, Chinese Academy of Sciences, Beijing, 100101, China. <sup>2</sup>Department of Psychology, University of Chinese Academy of Sciences, Beijing, 100049, China. Correspondence and requests for materials should be addressed to F.D. (email: [duf@psych.ac.cn](mailto:duf@psych.ac.cn))



**Figure 1.** The schematic illustration of events in a trial.

suggested that change detection tasks using a single-probe (only the possibly-changed square was tested) and using a whole-display-probe (all squares including the possibly-changed one were tested) yield different reliability<sup>20</sup>. However, studies have shown that there is a high correlation between different types of memory materials in the change detection task<sup>21</sup> and different working memory tasks<sup>22</sup>. Furthermore, Xu *et al.*<sup>23</sup> conducted one session of single-probe tests with 78 participants every day for 30 continuous days and the 31<sup>st</sup> session after a month. The correlation coefficients between adjacent sessions varied from 0.53 to 0.81, and importantly they were positively correlated with their session numbers, indicating an increased test-retest reliability as sessions grew in number. However, the test-retest reliability of VWM capacity estimation in a whole-display-probe task has not yet been investigated.

In addition to set size and repeated practice, other factors might also influence the test-retest reliability of VWM measures. For example, it is known that time of day affects people's ability to retrieve information from long-term memory<sup>24</sup>. A recent study showed that there are also circadian variations in the accuracy of visual working memory performance using a spatial working memory task. The variation of visual working memory performance showed a positive correlation with the rectal temperature by a 3-hour delay<sup>25</sup>. Moreover, the time of day effect can be modulated by participants' chronotype. For example, the performance of cognitive tasks is better when individuals are tested at their preferred time (e.g. the morning for morning-type people, and the evening for evening-type people)<sup>26</sup>. Additionally, some other daily functions and activities, such as glucose absorption<sup>27</sup>, exposure to computer screen light<sup>28</sup>, a single session of physical exercise<sup>29</sup> and so on, can also immediately influence cognitive performance. Therefore, it can be expected that time of day might also affect the estimation of visual working memory capacity.

In summary, some studies have examined the test-retest reliability of visual working memory capacity estimation<sup>18,23</sup>. However, the reliability of the lateralized whole-display-probe task has not yet been examined. Most electroencephalography (EEG) studies of visual working memory have used a "lateralized" variant of the whole-display-probe task, in which participants pay attention to stimuli in a hemi-field and ignore the distractors from the contra-lateral hemi-field. According to Pailian and Halberda's work<sup>20</sup>, reliabilities can be different in different visual working memory tasks. Thus, the present study aimed to examine the test-retest reliability of visual working memory capacity estimation with a lateralized whole-display-probe change detection task. Additionally, the present study also examined whether variation in testing time is an important source of test-retest variation for visual working memory capacity.

## Method

**Participants.** 96 subjects participated in this study (18~28 years old; 48 females). Before the beginning of the experiment, they signed the informed consent form and received monetary compensation. All of the participants had normal or corrected to normal vision. This study was approved by the Institutional Review Board of the Institute of Psychology, Chinese Academy of Sciences. Moreover, all experiments were performed following relevant guidelines and regulations.

**Apparatus, stimuli and procedure.** The present study adopted the change detection paradigm to measure visual working memory capacity<sup>2</sup>. The stimuli were presented on a 19 inch CRT display with a resolution of 1280 × 1024 pixels at a refresh rate of 75 Hz. Participants were seated 60 cm from the display with their heads rested on a chin-rest.

The events of a trial are illustrated in Fig. 1. In each trial an arrow was presented above a central fixation point for 200 ms, pointing to either the left or right to indicate a cued hemi-field. After a 300 ms blank interval, the first array of 4, 6, 8, 10, or 12 colored squares appeared on the gray background for 100 ms. The size of each square

|                      |        | Set size2<br>N = 96 | Set size3<br>N = 96 | Set size4<br>N = 96 | Set size5<br>N = 96 | Set size6<br>N = 61 |
|----------------------|--------|---------------------|---------------------|---------------------|---------------------|---------------------|
| 1 <sup>st</sup> test | P(Hit) | 0.932 (0.064)       | 0.888 (0.102)       | 0.830 (0.125)       | 0.772 (0.166)       | 0.723 (0.157)       |
|                      | P(FA)  | 0.044 (0.046)       | 0.069 (0.068)       | 0.116 (0.091)       | 0.136 (0.113)       | 0.148 (0.109)       |
|                      | K      | 1.786 (0.211)       | 2.473 (0.443)       | 2.923 (0.705)       | 3.243 (1.154)       | 3.699 (1.225)       |
| 2 <sup>nd</sup> test | P(Hit) | 0.952 (0.062)       | 0.923 (0.089)       | 0.907 (0.072)       | 0.882 (0.129)       | 0.839 (0.142)       |
|                      | P(FA)  | 0.038 (0.102)       | 0.046 (0.096)       | 0.089 (0.114)       | 0.093 (0.112)       | 0.112 (0.098)       |
|                      | K      | 1.865 (0.184)       | 2.665 (0.425)       | 3.344 (0.534)       | 4.046 (0.913)       | 4.744 (1.009)       |

**Table 1.** The means and standard deviations of hit rates, false alarm rates, and K values.

was  $0.65^\circ \times 0.65^\circ$ , with a minimum distance of  $2^\circ$  between two squares. An equal number of colored squares were distributed in the left and right hemi-fields. Participants were instructed to only memorize the colored squares in the cued hemi-field. Then a blank screen appeared for 900 ms, which was followed by the second array of colored squares for 750 ms. The second array of squares was either identical to the first array or different from the first array in that one square in the cued hemi-field changed color. Then a blank screen remained until the participant responded by pressing corresponding keys. Since participants only had to memorize the squares in the cued hemi-field, the memory set size was 2, 3, 4, 5 or 6.

The first 35 participants were tested with memory set sizes of 2, 3, 4 or 5. However, the capacity estimate for two participants equaled 5, indicating that their working memory capacity might exceed the largest memory set size of 5. Thus, the largest memory set size was increased from 5 to 6. The other 61 subjects were tested with memory set sizes of 2, 3, 4, 5 or 6.

In the task, there were 100 trials for each set-size, and all trials were presented randomly. There was a 30-second break after every 100 trials. Each participant had to take the same test twice with a separation of 3 to 16 days.

**Working memory capacity measurement.** According to Pashler<sup>30</sup>, the capacity of visual working memory is measured by the K value, the formula of which is as below. Compared to the other commonly used method proposed by Cowan<sup>31</sup>, Pashler's method is more appropriate for the current paradigm<sup>32</sup>. The formula is as follows in equation 1:

$$K = \text{setsize} * (P(\text{hit}) - P(\text{FA})) \div (1 - P(\text{FA})) \quad (1)$$

where  $P(\text{Hit}) = \text{hits}/(\text{hits} + \text{misses})$ , and  $P(\text{FA}) = \text{false alarms}/(\text{false alarms} + \text{correct rejections})$ .

In addition to the K values of each set size, we also computed the average K value ( $K_{\text{mean}}$ ) and maximum K value ( $K_{\text{max}}$ ) across all set sizes as estimations for each participant's visual working memory capacity.

To assess the test-retest reliability of K estimates, correlations between K estimates of the two tests were calculated. The test-retest reliability was considered good when  $r \geq 0.71$ ; fair  $0.51 \leq r \leq 0.70$ ; weak  $0.31 \leq r \leq 0.50$ ; little or none  $r \leq 0.33$ .

**Analysis of the time of day effect.** The 96 subjects were divided equally into 2 groups based on the median time difference (48 minutes) between the two tests. The Same-time group (48 subjects, the time difference ranged from 0~44 minutes, mean = 13.23 min, sd = 10.02) and Different-time group (48 subjects, ranging from 52~688 minutes, mean = 217.44 min, sd = 162.95). The test-retest correlation coefficients of the two groups were compared by using Snedecor's method<sup>34</sup>. This method compares correlations of two independent samples on the same pair of variables by using the following formula 2:

$$Z = \frac{Z_1 - Z_2}{\frac{1}{(n_1 - 3) + (n_2 - 3)}} \quad (2)$$

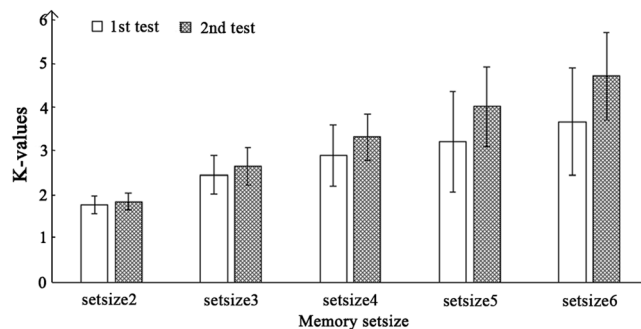
$Z_1$  and  $Z_2$  are Fisher's Zs<sup>35</sup> for correlation coefficients. Also,  $n_1$  and  $n_2$  are the sample sizes of the two independent samples.

## Results

**Descriptive statistics.** The hit rates, false alarm rates, and K values under each memory set size for the two tests are listed in Table 1.

**Test-retest improvement.** The K values were submitted to a repeated-measures ANOVA with memory set size and test-retest as two within-subject variables (see Fig. 2). Results showed a significant main effect of test-retest,  $F = 103.104$ ,  $p < 0.001$ ,  $\eta_p^2 = 0.632$ , with the larger K in the second test than in the first test. The main effect of memory set size was also significant,  $F = 220.563$ ,  $p < 0.001$ ,  $\eta_p^2 = 0.935$ , with larger K values as set size increased. The interaction was also significant,  $F = 48.566$ ,  $p < 0.001$ ,  $\eta_p^2 = 0.647$ . Further analysis showed that test-retest improvements were significant at all levels of memory set size (all  $ps < 0.05$ ).

The test-retest improvements of K values were possibly due to the enhancement of hit rate or the decline of the false-alarm rate. To address this, we undertook ANOVA analysis on both hit rate and false-alarm rate as for K values. Results showed that hit rate ( $F = 49.081$ ,  $p < 0.001$ ,  $\eta_p^2 = 0.454$ ) was improved in the second test compared with the first test. Also, the false-alarm rate in the second test ( $F = 9.822$ ,  $p = 0.003$ ,  $\eta_p^2 = 0.141$ ) was reduced compared with the first test.



**Figure 2.** The average K values for each set size in the two tests.

|                             | $K_{(T2\text{-set size}2)}$ | $K_{(T2\text{-set size}3)}$ | $K_{(T2\text{-set size}4)}$ | $K_{(T2\text{-set size}5)}$ | $K_{(T2\text{-set size}6)}$ | $K_{(T2\text{-mean})}$ | $K_{(T2\text{-max})}$ |
|-----------------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|------------------------|-----------------------|
| $K_{(T1\text{-set size}2)}$ | 0.505                       | 0.399                       | 0.386                       | 0.363                       | 0.421                       | 0.493                  | 0.432                 |
| $K_{(T1\text{-set size}3)}$ | 0.506                       | 0.502                       | 0.472                       | 0.497                       | 0.494                       | 0.590                  | 0.526                 |
| $K_{(T1\text{-set size}4)}$ | 0.332                       | 0.363                       | 0.572                       | 0.543                       | 0.544                       | 0.564                  | 0.536                 |
| $K_{(T1\text{-set size}5)}$ | 0.354                       | 0.339                       | 0.449                       | 0.647                       | 0.676                       | 0.592                  | 0.614                 |
| $K_{(T1\text{-set size}6)}$ | 0.393                       | 0.392                       | 0.498                       | 0.618                       | 0.757                       | 0.727                  | 0.768                 |
| $K_{(T1\text{-mean})}$      | 0.436                       | 0.438                       | 0.546                       | 0.633                       | 0.730                       | 0.704                  | 0.705                 |
| $K_{(T1\text{-max})}$       | 0.400                       | 0.393                       | 0.524                       | 0.625                       | 0.757                       | 0.691                  | 0.715                 |

**Table 2.** Pearson's correlations between Ks for each corresponding set size in the two tests. T1-set size 2 = memory set size 2 in the first test. Similarly, T2-set size3 = memory set size 3 in the second test. All correlations are significant at  $p \leq 0.01$ .

**Test-retest reliability.** Pearson product-moment correlations between the K values of the two tests under each set size are listed in Table 2. Results showed that at all set size levels, the K values in the second test were significantly correlated with the corresponding K values in the first test.

Also, as the memory set size increased, the test-retest reliability coefficients for each corresponding set size also showed a rising trend (Fig. 3). To examine whether this trend was significant, the modified Pearson-Filontest<sup>36</sup> as in equations (3) and (4), was used to compare test-retest reliability coefficients across set sizes. This test is to compare correlation coefficients within the same sample but with non-overlapping pairs of variables<sup>37</sup>.

$$Z_{PF} = \sqrt{\frac{n-3}{2}} \times \frac{Z_{AB} - Z_{XY}}{\sqrt{1 - \frac{kPF}{2(1-r_{AB}^2)(1-r_{XY}^2)}}} \quad (3)$$

where  $kPF$  is computed by

$$kPF = (r_{AX} - r_{BX} \times r_{AB}) \times (r_{BY} - r_{BX} \times r_{XY}) + (r_{AY} - r_{AX} \times r_{XY}) \\ \times (r_{BX} - r_{AX} \times r_{AB}) + (r_{AX} - r_{AY} \times r_{XY}) \times (r_{BY} - r_{AY} \times r_{AB}) \\ + (r_{AY} - r_{AB} \times r_{BY}) \times (r_{BX} - r_{BY} \times r_{XY}) \quad (4)$$

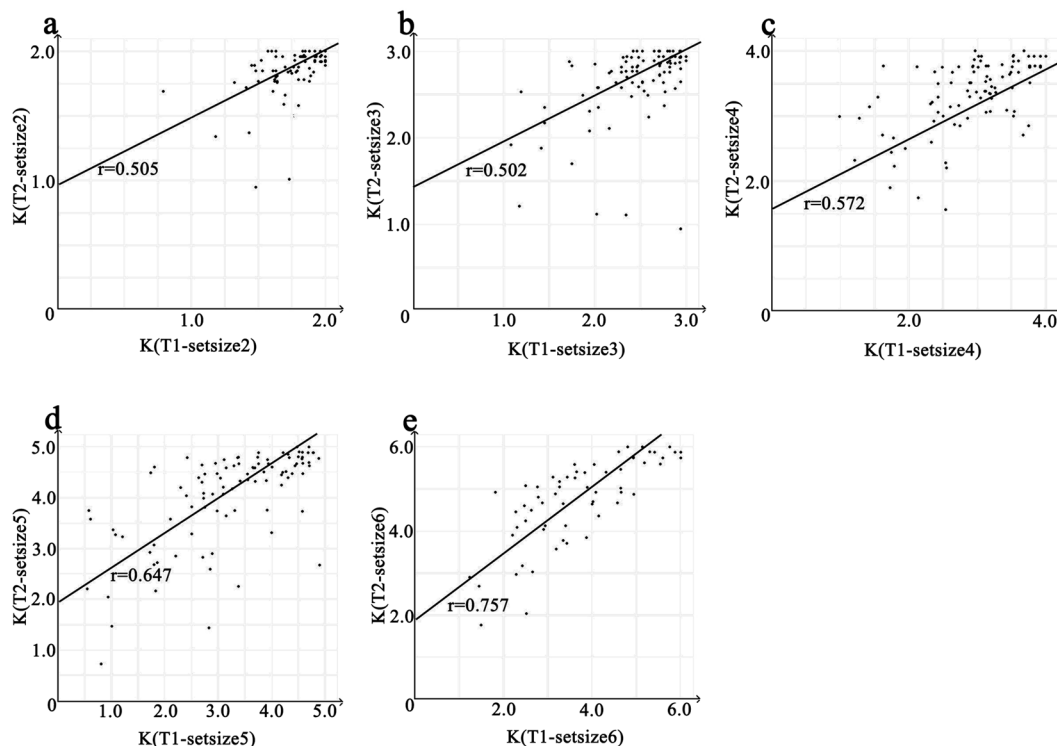
A, B, X, Y represent the two non-overlapping pairs of variables.  $Z_{AB}$  and  $Z_{XY}$  are the Fisher-z transformed result of  $r_{AB}$  and  $r_{XY}$ .

Results showed that only the test-retest reliability coefficients for set size 6 were significantly higher than those for set size 2 (Tables 3 and 4). The sample size was 61 in each pair comparison including set size 6, and 96 in those not including set size 6.

**The internal consistency across memory set sizes within a single test.** The within-test consistency across set sizes refers to the correlations between K values for different memory set sizes within a single test. The correlations between K values for different memory set sizes in the first test are listed in Table 5. All of the correlations were medium or high.

We also examined the within-test consistency across set sizes in the second test. The correlations between K values for different memory set sizes in the second test are listed in Table 6. All of the correlations were medium or high.

**The time of day effect on test-retest reliability.** The test-retest reliability coefficients for the two groups (the same time of day v.s. different time of day) are listed in Table 7. We examined whether participants who took the two tests at the same time of day showed a higher test-retest reliability than participants who undertook the two tests at different times of day. As Table 7 shows, in the K(setsize3), K(setsize5), K(mean) and K(max) conditions, the test-retest reliability coefficients for the same time of day group were significantly higher than those for the different time of day group.



**Figure 3.** The test-retest correlations between Ks for each set size in the two tests. Panel a illustrates the correlations between K for set size 2 in the two tests. Panel b illustrates the correlations between K for set size 3 in the two tests. Panel c illustrates correlations between K for set size 4 in the two tests. Panel d illustrates correlations between K for set size 5 in the two tests. Panel e illustrates correlations between K for set size 6 in the two tests.

|                                | $r_{\text{set size}3} = 0.502$ | $r_{\text{set size}4} = 0.572$ | $r_{\text{set size}5} = 0.647$ |
|--------------------------------|--------------------------------|--------------------------------|--------------------------------|
| $r_{\text{set size}2} = 0.505$ | Z = 0.029                      | Z = -0.064                     | Z = -1.488                     |
|                                | p = 0.976                      | p = 0.506                      | p = 0.137                      |
|                                | N = 96                         | N = 96                         | N = 96                         |
| $r_{\text{set size}3} = 0.502$ |                                | Z = -0.708                     | Z = -1.541                     |
|                                |                                | p = 0.479                      | p = 0.123                      |
|                                |                                | N = 96                         | N = 96                         |
| $r_{\text{set size}4} = 0.572$ |                                |                                | Z = -0.851                     |
|                                |                                |                                | p = 0.395                      |
|                                |                                |                                | N = 96                         |

**Table 3.** Comparisons between test-retest reliability coefficients for each set size.

|                                | $r_{\text{set size}2} = 0.512$ | $r_{\text{set size}3} = 0.670$ | $r_{\text{set size}4} = 0.614$ | $r_{\text{set size}5} = 0.708$ |
|--------------------------------|--------------------------------|--------------------------------|--------------------------------|--------------------------------|
| $r_{\text{set size}6} = 0.757$ | Z = -2.307                     | Z = -0.966                     | Z = -1.489                     | Z = -0.577                     |
|                                | p = 0.021                      | p = 0.334                      | p = 0.137                      | p = 0.564                      |
|                                | N = 61                         | N = 61                         | N = 61                         | N = 61                         |

**Table 4.** Comparisons between test-retest reliability coefficients for each set size.

### Discussion

The present study examined the test-retest reliability of VWM capacity estimation by using the change detection paradigm. The test-retest correlations between K values for corresponding set sizes in the two tests were comparable to those in the verbal recall tasks, as found by Gathercole, *et al.*<sup>13</sup>, varying from fair to good. The results also showed a trend that the test-retest reliability increased as the set size became larger. In addition, we found that the test-retest reliability of VWM capacity estimation was higher when the two tests occurred at the same time of day than at different times of day.

|                             | $K_{(T1\text{-set size}3)}$ | $K_{(T1\text{-set size}4)}$ | $K_{(T1\text{-set size}5)}$ | $K_{(T1\text{-set size}6)}$ |
|-----------------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|
| $K_{(T1\text{-set size}2)}$ | 0.710                       | 0.562                       | 0.526                       | 0.559                       |
| $K_{(T1\text{-set size}3)}$ |                             | 0.744                       | 0.728                       | 0.618                       |
| $K_{(T1\text{-set size}4)}$ |                             |                             | 0.803                       | 0.696                       |
| $K_{(T1\text{-set size}5)}$ |                             |                             |                             | 0.793                       |

**Table 5.** Pearson's correlations between Ks for each set size in the first test. All correlations are significant at  $p \leq 0.005$ .

|                             | $K_{(T2\text{-set size}3)}$ | $K_{(T2\text{-set size}4)}$ | $K_{(T2\text{-set size}5)}$ | $K_{(T2\text{-set size}6)}$ |
|-----------------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|
| $K_{(T2\text{-set size}2)}$ | 0.722                       | 0.498                       | 0.448                       | 0.427                       |
| $K_{(T2\text{-set size}3)}$ |                             | 0.593                       | 0.551                       | 0.431                       |
| $K_{(T2\text{-set size}4)}$ |                             |                             | 0.710                       | 0.512                       |
| $K_{(T2\text{-set size}5)}$ |                             |                             |                             | 0.780                       |
| $K_{(T2\text{-set size}6)}$ |                             |                             |                             |                             |

**Table 6.** Pearson's correlations between Ks for each set size in the second test. All correlations are significant at  $p \leq 0.01$ .

|                               |   | $K_{(set\ size\ 2)}$ | $K_{(set\ size\ 3)}$ | $K_{(set\ size\ 4)}$ | $K_{(set\ size\ 5)}$ | $K_{(set\ size\ 6)}$ | $K_{(mean)}$ | $K_{(max)}$ |
|-------------------------------|---|----------------------|----------------------|----------------------|----------------------|----------------------|--------------|-------------|
| Same-time of day              | r | 0.654                | 0.749                | 0.590                | 0.795                | 0.814                | 0.842        | 0.847       |
|                               | N | 48                   | 48                   | 48                   | 48                   | 37                   | 48           | 48          |
| Different-time of day         | r | 0.467                | 0.283                | 0.598                | 0.534                | 0.640                | 0.602        | 0.581       |
|                               | N | 48                   | 48                   | 48                   | 48                   | 24                   | 48           | 48          |
| Comparison between two groups | z | 1.390                | 3.224                | -0.059               | 2.320                | 1.371                | 2.522        | 2.758       |
|                               | p | 0.190                | 0.001                | 0.953                | 0.020                | 0.170                | 0.012        | 0.006       |

**Table 7.** Test-retest reliabilities for the same-time group and the different-time group.

The test-retest reliability of VWM capacity estimation (K values) was reasonably good in the present study, indicating that K values can be a relatively reliable index of individual VWM capacity. Secondly, a trend was found showing that the larger the memory set size, the higher the corresponding test-retest reliability for the K values. This trend was possibly due to a ceiling effect at set sizes 2 and 3 (see Fig. 2a for set size 2 and Fig. 2b for set size 3), which were lower than many participants' true VWM capacity<sup>4</sup>. This ceiling effect made it difficult to detect individual differences among participants. Thus, it is important for future work to use a memory set size that is at least larger than 4 when estimating individual VWM capacity.

In this study, the performance of the second test was better than that of the first one, indicating a practice effect with a 3~16 day interval (the time between 2 tests) between the two tests. This finding is inconsistent with Johnson *et al.*'s<sup>18</sup> research, which found no practice effect for visual working memory tests with a 1.5-year interval. It is possible that the practice effect waned with the greater interval between the two tests. However, the critical point of time at which practice effects disappear has yet to be studied. Thus, it is important for future work to allow for sufficient practice before formal measurement because the K value based on a short version of the change detection task might result in underestimation of VWM capacity. This conclusion is similar to that from Xu and colleagues' work, which found that adjacent-session correlation increased along with session number within 31 sessions<sup>23</sup>. However, more factors able to influence retest reliability should be discussed. It is not practical to test research subjects dozens of times in every experiment.

Reasonably good correlations were found between different set sizes in the same test, in contrast to the quite low correlations between different set sizes in a previous study<sup>20</sup>. This discrepancy might be due to methodological differences between the studies. First, the included set sizes were different. The current research used memory set sizes of 2, 3, 4, 5, and 6, while Pailian and Halberda<sup>20</sup> tested memory set sizes of 2, 4 and 8. The larger step size of the memory set sizes in the Pailian and Halberda<sup>20</sup> study might account for the lower correlations between different set sizes. Second, a larger sample was included in the current research (96 participants) compared with the 14 participants in Pailian and Halberda's<sup>20</sup> study.

Test-retest reliability was higher when the two tests occurred at the same time of day than at different times of day. This time of day effect was significant for the mean and maximum values of Ks across set sizes, and also for the values of K for set sizes 3 and 5. According to a meta-analysis in Schmidt and colleagues' review, cognitive performance in a phonological working memory task<sup>25</sup>, visual working memory task<sup>25</sup> and visual selective attention<sup>38</sup> can be influenced by circadian rhythm<sup>26</sup>. Also cognitive performance can be affected by daily activities<sup>27-29</sup>. Therefore, at different times of day, participants' performance varied in different directions, causing a reduced correlation of the test results, while tests taken at the same time of day had higher retest reliability. However, it is worth noting that the time of day effect on test-retest reliability was not significant for set sizes of 2, 4 and 6. There



are several possible explanations for this variation. First, the K value for a specific set size might not be sufficiently reliable, especially for small set sizes. Second, time of day was a rough correlate of circadian rhythms and daily activities, which might also add variation to the results. Finally, the relatively smaller sample in set size 6 might explain the absence of a time of day effect. Thus, the time of day effect needs further examination.

In conclusion, with appropriate control, the estimation of VWM capacity based on the change detection task is highly reliable. The average or maximum K values across multiple set sizes are also reliable. Results from the current study also indicate that future research on measuring individual working memory capacity should use set sizes larger than 4, and should allow for sufficient practice before formal measurement.

## References

1. Baddeley, A. *Working memory, thought, and action*. Vol. 45 (OUP Oxford, 2007).
2. Luck, S. J. & Vogel, E. K. The capacity of visual working memory for features and conjunctions. *Nature* **390**, 279–281 (1997).
3. Luck, S. J. & Vogel, E. K. Visual working memory capacity: from psychophysics and neurobiology to individual differences. *Trends in cognitive sciences* **17**, 391–400 (2013).
4. Vogel, E. K. & Awh, E. How to exploit diversity for scientific gain: Using individual differences to constrain cognitive theory. *Current Directions in Psychological Science* **17**, 171–176 (2008).
5. Fukuda, K., Vogel, E., Mayr, U. & Awh, E. Quantity, not quality: The relationship between fluid intelligence and working memory capacity. *Psychonomic Bulletin & Review* **17**, 673–679 (2010).
6. Kyllonen, P. C. & Christal, R. E. Reasoning ability is (little more than) working-memory capacity?! *Intelligence* **14**, 389–433 (1990).
7. Caplan, D. & Waters, G. S. Verbal working memory and sentence comprehension. *Behavioral and Brain Sciences* **22**, 77–94 (1999).
8. Daneman, M. & Carpenter, P. A. Individual differences in working memory and reading. *Journal of verbal learning and verbal behavior* **19**, 450–466 (1980).
9. Miyake, A. & Friedman, N. P. Individual differences in second language proficiency: Working memory as language aptitude. *Foreign language learning: Psycholinguistic studies on training and retention*, 339–364 (1998).
10. Alloway, T. P. & Alloway, R. G. Investigating the predictive roles of working memory and IQ in academic attainment. *Journal of experimental child psychology* **106**, 20–29 (2010).
11. Moore, M. E. & Ross, B. M. Context effects in running memory. *Psychological Reports* **12**, 451–465 (1963).
12. Petrides, M., Alivisatos, B., Meyer, E. & Evans, A. C. Functional activation of the human frontal cortex during the performance of verbal working memory tasks. *Proceedings of the National Academy of Sciences* **90**, 878–882 (1993).
13. Gathercole, S. E., Pickering, S. J., Ambridge, B. & Wearing, H. The structure of working memory from 4 to 15 years of age. *Developmental psychology* **40**, 177 (2004).
14. Hockey, A. & Geffen, G. The concurrent validity and test–retest reliability of a visuospatial working memory task. *Intelligence* **32**, 591–605 (2004).
15. Luria, R. & Vogel, E. K. Visual search demands dictate reliance on working memory storage. *Journal of Neuroscience* **31**, 6199–6207 (2011).
16. Drew, T., Horowitz, T. S., Wolfe, J. M. & Vogel, E. K. Delineating the neural signatures of tracking spatial position and working memory during attentive tracking. *Journal of Neuroscience* **31**, 659–668 (2011).
17. Vogel, E. K., McCollough, A. W. & Machizawa, M. G. Neural measures reveal individual differences in controlling access to working memory. *Nature* **438**, 500–503 (2005).
18. Johnson, M. K. *et al.* The relationship between working memory capacity and broad measures of cognitive ability in healthy adults and people with schizophrenia. *Neuropsychology* **27**, 220 (2013).
19. Open Science Collaboration. Estimating the reproducibility of psychological science. *Science* **349**, aac4716 (2015).
20. Pailian, H. & Halberda, J. The reliability and internal consistency of one-shot and flicker change detection for measuring individual differences in visual working memory capacity. *Memory & Cognition* **43**, 397–420 (2015).
21. Awh, E., Barton, B. & Vogel, E. K. Visual Working Memory Represents a Fixed Number of Items Regardless of Complexity. *Psychological Science* **18**, 622–628 (2007).
22. Unsworth, N., Fukuda, K., Awh, E. & Vogel, E. K. Working memory and fluid intelligence: Capacity, attention control, and secondary memory retrieval. *Cognitive psychology* **71**, 1–26 (2014).
23. Xu, Z., Adam, K. C. S., Fang, X., & Vogel, E. K. The reliability and stability of visual working memory capacity. *Behavior Research Methods*, 1–13 (2017).
24. Folkard, S. & Monk, T. H. Circadian rhythms in human memory. *British Journal of Psychology* **71**, 295–307 (1980).
25. Ramírez, C. *et al.* Circadian rhythms in phonological and visuospatial storage components of working memory. *Biological Rhythm Research* **37**, 433–441 (2006).
26. Schmidt, C., Collette, F., Cajochen, C. & Peigneux, P. A time to think: circadian rhythms in human cognition. *Cognitive Neuropsychology* **24**(7), 755–789 (2007).
27. Sünram-Lea, S. I., Foster, J. K., Durlach, P. & Perez, C. Glucose facilitation of cognitive performance in healthy young adults: examination of the influence of fast-duration, time of day and pre-consumption plasma glucose levels. *Psychopharmacology* **157**(1), 46–54 (2001).
28. Cajochen, C. *et al.* Evening exposure to a light-emitting diodes (LED)-backlit computer screen affects circadian physiology and cognitive performance. *Journal of applied physiology* **110**(5), 1432–1438 (2011).
29. Chang, Y. K., Labban, J. D., Gapin, J. I. & Etnier, J. L. The effects of acute exercise on cognitive performance: a meta-analysis. *Brain research* **1453**, 87–101 (2012).
30. Pashler, H. Familiarity and visual change detection. *Perception & Psychophysics* **44**, 369–378 (1988).
31. Cowan, N. Metatheory of storage capacity limits. *Behavioral and brain sciences* **24**(1), 154–176 (2001).
32. Roudner, J. N., Morey, R. D., Morey, C. C. & Cowan, N. How to measure working memory capacity in the change detection paradigm. *Psychonomic Bulletin & Review* **18**(2), 324–330 (2011).
33. Basaran, S., Guzel, R., Seydaoglu, G. & Guler-Uysal, F. Validity, reliability, and comparison of the WOMAC osteoarthritis index and Lequesnealgofunctional index in Turkish patients with hip or knee osteoarthritis. *Clinical rheumatology* **29**, 749–756 (2010).
34. Snedecor, G. W. & Cochran, W. G. Statistical methods. *The Iowa State University*, 185–188 (1980).
35. Fisher, R. A. Statistical methods for research workers. *Genesis Publishing Pvt Ltd* (1925).
36. Raghunathan, T. E., Rosenthal, R. & Rubin, D. B. Comparing correlated but nonoverlapping correlations. *Psychological Methods* **1**, 178–183 (1996).
37. Zou, G. Y. Toward using confidence intervals to compare correlations. *Psychological Methods* **12**(4), 399 (2007).
38. Bonnefond, A., Rohmer, O., Hoefl, A., Muzet, A. & Tassi, P. Interaction of age with time of day and mental load in different cognitive tasks. *Perceptual and motor skills* **96**(3\_suppl), 1223–1236 (2003).

## Acknowledgements

This study was supported by grants from the National Natural Science Foundation of China (31470982), and the scientific foundation of the Institute of Psychology, Chinese Academy of Sciences (Y4CX033008).

### Author Contributions

Study conceived by F. Du and M. Dai. M. Dai, Y. Li and S. Gan collected and analyzed data under F. Du's supervision. F. Du and M. Dai wrote the manuscript.

### Additional Information

**Competing Interests:** The authors declare no competing interests.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2019