**Title**

# MIS – A new scoring method for the operation span task that accounts for Math, remembered Items and Sequence.

**Abbreviated title**

MIS – New scoring system for the Ospan task.

**Authors**

Lammert, M.[1,2] *, & Morys F.[1,3] *, Hartmann H.[1,4], Janssen, L.K.[1,2], Horstmann A.[1,2,4,5]

**Affiliations**

[1] Department of Neurology, Max Planck Institute for Human Cognitive and Brain Sciences, 04103 Leipzig, Germany

[2] Leipzig University Medical Centre, IFB Adiposity Diseases, 04103 Leipzig, Germany

[3] Montreal Neurological Institute, McGill University, Montreal, QC, Canada

[4] Leipzig University Medical Centre, Collaborative Research Centre 1052-A5, 04103 Leipzig, Germany

[5] Department of Psychology and Logopedics, Faculty of Medicine, University of Helsinki, 00290 Helsinki, Finland

* Both authors contributed equally to this work

**Corresponding author**

Annette Horstmann, PhD

horstmann@cbs.mpg.de

Department of Neurology, Max Planck Institute for Human Cognitive and Brain Sciences,

Stephanstrasse 1a

04103 Leipzig, Germany

**Authors report no conflict of interest.**

## Acknowledgments

The authors would like to thank Johanna Westbrook and Scott Walter from Macquarie University, Australia, Matt Meier from Western Carolina University, North Carolina, USA, and Christopher Draheim from Georgia Institute of Technology, Georgia, USA for sharing their operation span task data and providing feedback to the manuscript.

## Financial support

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Projektnummer 209933838 – SFB 1052 - A05 (HH, ML, AH); and Federal Ministry of Education and Research (BMBF), Germany FKZ: 01E01501 (FM, ML, LKJ, AH).

Abstract

The operation span task is a well-validated measure of the executive component of working memory. Previous scoring systems of this task focus predominantly on the span part of the task, while the distractor – math task – serves as an exclusion criterion for test assessment only. Here, we propose a new Math-Item-Sequence (MIS) system to score performance on the Ospan based on both the span and math part. This new system provides three main improvements: 1) it eliminates the need to introduce arbitrary exclusion thresholds based on performance on the distractor task; 2) it takes into account remembered letters, and their relative position in the sequence separately; 3) it considers performance on the math task in the scoring of the Ospan task as a downweighing factor. In 6 independent samples we show that MIS score correlates highly with previously recommended scoring methods, suggesting that it measures the same underlying concepts. We also show that internal consistency of MIS is very good and comparable to or higher than the previous methods. We argue that MIS could be used in all samples, but might be of particular interest for small samples, where exclusions of participants are especially costly.

## 1. Introduction

Working memory (WM) can be measured with a variety of tasks that are usually divided into simple span and complex span tasks (Unsworth & Engle, 2007). An example of the simple span task is a digit span task, which requires individuals to remember and reproduce a sequence of numbers. On the other hand, the operation span (Ospan) task – an example of a complex span task – uses simple mathematical equations as distractors between each of the letters of the sequence that is to be remembered (Turner & Engle, 1989; Unsworth, Heitz, Schrock, & Engle, 2005). Individuals must correctly solve the distractor task while keeping in mind the presented sequence of letters. The letters must then be recalled in the correct order.

Both simple and complex span tasks measure working memory and predict abilities such as language comprehension or fluid intelligence equally well (Colom, Rebollo, Abad, & Shih, 2006; Conway, Cowan, Bunting, Therriault, & Minkoff, 2002; Daneman & Merikle, 1996; Draheim, Harrison, Embretson, & Engle, 2018; Randall W. Engle, Kane, & Tuholski, 1999; R. W. Engle, Tuholski, Laughlin, & Conway, 1999; Kane et al., 2004; Shipstead, Harrison, & Engle, 2016; Turner & Engle, 1989; Unsworth & Engle, 2007; Unsworth, Heitz, & Engle, 2005). Nevertheless, complex span tasks are a preferred way to test WM, as they increase variability in results from all sequence lengths (Unsworth & Engle, 2007). This is predominantly due to the design of simple span tasks, which use sequences of increasing length and typically terminate after a sequence of a particular length is remembered incorrectly (Unsworth & Engle, 2007). In contrast, a modern and automated version of the Ospan task (Unsworth, Heitz, Schrock, et al., 2005) does not use a specific order of trials and does not terminate after an incorrectly remembered sequence. It therefore provides a measure of variability in all sequence lengths. Additionally, the Ospan task is one of the most popular complex span tasks due to its high reliability and validity (Conway et al., 2002; Draheim et al., 2018; Turner & Engle, 1989). Currently, there are two scoring systems of the Ospan task (and other complex span tasks): 1) an absolute system (ABS), where scores are assigned only when the entire sequence in each trial is remembered correctly; and 2) a partial credit scoring system, e.g. partial credit unit system (PCU), where partial scores can be assigned for correct recognition of parts of the entire sequence (Conway et al., 2005; Unsworth & Engle, 2007). PCU is generally preferred over the all-or-nothing ABS, because PCU increases variability in results from longer sequences, while ABS predominantly measures variability in short sequences, similarly to simple span tasks (Unsworth & Engle, 2007).

We have identified some aspects of the PCU systems where improvements can be made (Conway et al., 2005). Firstly, PCU requires an arbitrary threshold for participants' exclusions based on the distractor tasks (e.g., exclude participants whose overall distractor task performance was below 85%). Secondly, PCU does not differentiate between the items remembered and the order in which the items appear in the sequence – we propose that those two aspects should be treated independently. Thirdly, performance on the distractor task is only taken into account as a criterion for exclusion. In contrast, we propose that performance on the distractor task should be considered when calculating overall task performance, since participants with similar scores on the span part and different performance on the distractor task ought to have different scores. In this way, the performance in the mathematical task should correct the overall Ospan performance downward and 'punish' the participants for not paying enough attention to the distractor task. We created a new scoring system, which tackles those three improvement areas and takes into account three different contributions – performance on the distractor task (math, M), correctness of a remembered item (I) and correct placement of the item in the sequence (S). We further compared this MIS system with a previous PCU system in 6 independent samples. We claim that MIS is an improved way to analyse data from the Ospan task, especially in experiments with small sample sizes testing special populations, where exclusions are very costly. Further, we claim that MIS measures similar constructs to PCU, namely the executive component of working memory (Shipstead et al., 2016; Unsworth & Engle, 2007).

## 2. Materials and methods

### 2.1. Samples

We used 6 independent samples to test a new scoring system for the Ospan task and compare it with PCU. Samples were acquired from (Kane et al., 2016; Westbrook, Raban, Walter, & Douglas, 2018) with following sizes: 136, 128, 127, 68, 82, and 35 (Table 1).

### 2.2. Math-Item-Sequence (MIS) system

The proposed scoring system of the Ospan task consists of three parts: 1) math part (M) – performance on the distractor task, 2) item part (I) – correct identification of a remembered letter, and 3) sequence part (S) – longest correct sequence of letters. The MIS score differentiates between items and sequence parts, because, as we propose, remembering the letter and its place in the sequence reflects two different cognitive processes, namely the short-term memory and relative object placement. According to Conway's recommendations (2005), the MIS system is designed as a unit scoring (as opposed to load scoring) – individual items have the same contributions to full scores independent of sequence lengths. The following mathematical formula describes the MIS system for the $i^{th}$ trial for the $j^{th}$ participant:

$$MIS_{ij} = \frac{M_{cor} - 1}{M_{tot} - 1} \times \frac{1}{2}\left(\frac{I_{cor}}{I_{tot}} + \frac{S_{cor}}{S_{tot}}\right)$$

Scores for each individual Ospan trial are summed into a total MIS score, $MIS_j = \sum_i \; MIS_{ij}$. The first part of the score describes performance on the distractor math task. Here, a ratio of correct responses ($M_{cor}$) to total math tasks ($M_{tot}$) for each trial is calculated. However, the first math task in each trial is never taken into account in the MIS score, because it has no influence on the overall task performance, as it is presented before any letters are shown. For that reason, the total number of math tasks in each trial is always decreased by one, and the number of correct responses is decreased by one even if the answer to the first math task of a given trial was correct. The second part of the score consists of the sum of item and sequence scoring ($\frac{I_{cor}}{I_{tot}}$ and $\frac{S_{cor}}{S_{tot}}$, respectively) and reflects the span task. Here, similar to the math score, a ratio of correct responses to total numbers in the sequence is calculated and summed. This sum is further multiplied by ½ (average of item and sequence scoring) to equalize its weight with respect to the math score. Additionally, through the multiplication of math score by item and sequence scores, performance on the math task downweighs performance on the span task. This ensures that participants who focussed only on the span task, but not on the math task, achieve lower scores.

Further, MIS scoring creates the possibility to calculate and analyse math, item and sequence scores separately ($MIS_{math}$, $MIS_{item}$, $MIS_{sequence}$).

### 2.3. Data analysis

Data analysis was performed in R v.3.4.3 (RCoreTeam, 2015) within RStudio (RStudioTeam, 2016). We calculated partial credit unit (PCU) scores according to (Conway et al., 2005), where for each trial the number of correct letters in correct places is summed up and divided by the length of the sequence. For PCU scores we excluded individuals whose performance on the math task was below 85%. We also calculated $PCU_{all}$ scores without this exclusion threshold for further correlation analysis with MIS, where no such threshold needs to be introduced. We calculated the MIS scores according to the previously presented equation. Further, we estimated internal consistency of the MIS method and all subscores using Cronbach's alpha, which describes whether different trials measure the same variable (Cronbach, 1951). In case of the Ospan task, different trials are trials with different sequence lengths, and the measured variable is the executive aspect of working memory (Klein & Fiss, 1999). We therefore divided the task into five types of trials including sequences with 3 to 7 elements. Each type comprised three trials (Unsworth, Heitz, Schrock, et al., 2005). For each participant we sorted trials according to their length and entered them into a matrix, where each row corresponded to one participant, and each column corresponded to one trial. We used Cronbach function within psych package (Revelle, 2018) to calculate Cronbach's alpha.

As a next step, we investigated whether MIS reflects between-subject variability and between-subject differences similarly to PCU (with 85% exclusion threshold based on math performance). To test this, we used Spearman's correlation. The hypothesis behind this analysis was that previous scores reflect WM capacity well, and that the MIS system should highly correlate with these scores. Correlations between PCU and $PCU_{all}$ were of similar magnitude to MIS and PCU, and since correlations between PCU and $PCU_{all}$ were high, we decided to only use $PCU_{all}$ for further analysis. We investigated how $MIS_{math}$, $MIS_{item}$ and $MIS_{sequence}$ correlate with $PCU_{all}$. Here, we hypothesised that $MIS_{item}$ and $MIS_{sequence}$ will correlate better with $PCU_{all}$ than $MIS_{math}$, because $MIS_{math}$ is not taken into consideration within $PCU_{all}$.

Lastly, in an exploratory fashion we tested relationships within MIS by correlating all MIS subscores with MIS and with each other. This was done to see how MIS subscores contribute to MIS and how they are related to each other.

## 3. Results

For descriptive statistics on Ospan scores in all samples see Table 1.

*Table 1 Descriptive statistics of the Ospan task scores*

| | Sample 1 (n = 136) | | Sample 2 (n=128) | | Sample 3 (n=127) | | Sample 4 (n=68) | | Sample 5 (n=82) | | Sample 6 (n=35) | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Median | Range | Median | Range | Median | Range | Median | Range | Median | Range | Median | Range |
| MIS | 10.55 | 0.26-14.83 | 10.33 | 2.04-14.11 | 10.68 | 2.64-14.53 | 11.17 | 3.72-13.93 | 10.52 | 1.55-14.22 | 12.14 | 4.07-14.83 |
| $MIS_{item}$ | 12.71 | 0.57 – 15.00 | 12.46 | 3.64 – 14.86 | 12.73 | 3.39 – 15.00 | 13.3 | 5.8 – 15.00 | 12.86 | 2.76 – 15.00 | 13.60 | 7.09 – 15.00 |
| $MIS_{sequence}$ | 10.65 | 0.00 – 15.00 | 10.52 | 1.33 – 14.86 | 10.65 | 1.10 – 14.86 | 11.19 | 3.02 – 15.00 | 10.74 | 0.90 - 1500 | 12.12 | 1.60 – 15.00 |
| $MIS_{math}$ | 13.85 | 9.95 – 15.00 | 13.83 | 7.90 – 15.00 | 13.80 | 10.87 – 15.00 | 13.73 | 11.45 – 15.00 | 13.86 | 11.10 – 15.00 | 14.30 | 11.82 – 15.00 |
| PCU | 11.04 | 0.00 – 15.00 | 10.79 | 2.03 – 14.86 | 10.85 | 2.80 – 14.86 | 11.65 | 3.10 – 15.00 | 10.73 | 1.01 – 15.00 | 12.32 | 2.81 – 15.00 |

*MIS – math-item-sequence, PCU – partial credit unit*

### 3.1. MIS internal consistency

In all samples, value of Cronbach's alpha for MIS was above 0.8 (Table 2), indicating good internal consistency. This means that each trial of the Ospan task measured by MIS reflects the same concept. Values of internal consistency for MIS subscores were similarly high, with the exception of $MIS_{math}$ (Table 2).

*Table 2 Internal consistency in tested samples*

| | Sample 1 | | Sample 2 | | Sample 3 | | Sample 4 | | Sample 5 | | Sample 6 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | α | CI | α | CI | α | CI | α | CI | α | CI | α | CI |
| MIS | 0.87 | 0.83-0.90 | 0.83 | 0.78-0.87 | 0.86 | 0.83-0.90 | 0.85 | 0.79-0.90 | 0.87 | 0.83-0.91 | 0.88 | 0.82-0.94 |
| $MIS_{item}$ | 0.89 | 0.87-0.92 | 0.86 | 0.83 - 0.9 | 0.88 | 0.85-0.91 | 0.87 | 0.82-0.91 | 0.9 | 0.87-0.93 | 0.88 | 0.82-0.94 |
| $MIS_{sequence}$ | 0.85 | 0.82-0.89 | 0.81 | 0.76-0.86 | 0.86 | 0.83-0.9 | 0.85 | 0.8-0.9 | 0.86 | 0.82-0.9 | 0.88 | 0.82-0.94 |
| $MIS_{math}$ | 0.59 | 0.49-0.69 | 0.61 | 0.51-0.71 | 0.53 | 0.41-0.65 | 0.33 | 0.1-0.56 | 0.51 | 0.36-0.67 | 0.44 | 0.18-0.69 |

*CI – confidence intervals; MIS – math-item-sequence*

### 3.2. Relationship between MIS and PCU

We first investigated correlations between MIS, PCU, and $PCU_{all}$. Correlations of these measures for all samples were above 0.900 (Table 3), indicating that MIS, $PCU_{all}$ and PCU are similar.

Further, as hypothesised, for all samples we found high correlations between $MIS_{item}$ and $PCU_{all}$, $MIS_{sequence}$ and $PCU_{all}$, and lower but still highly significant correlations between $MIS_{math}$ and $PCU_{all}$ (Table 3, Figure 1).

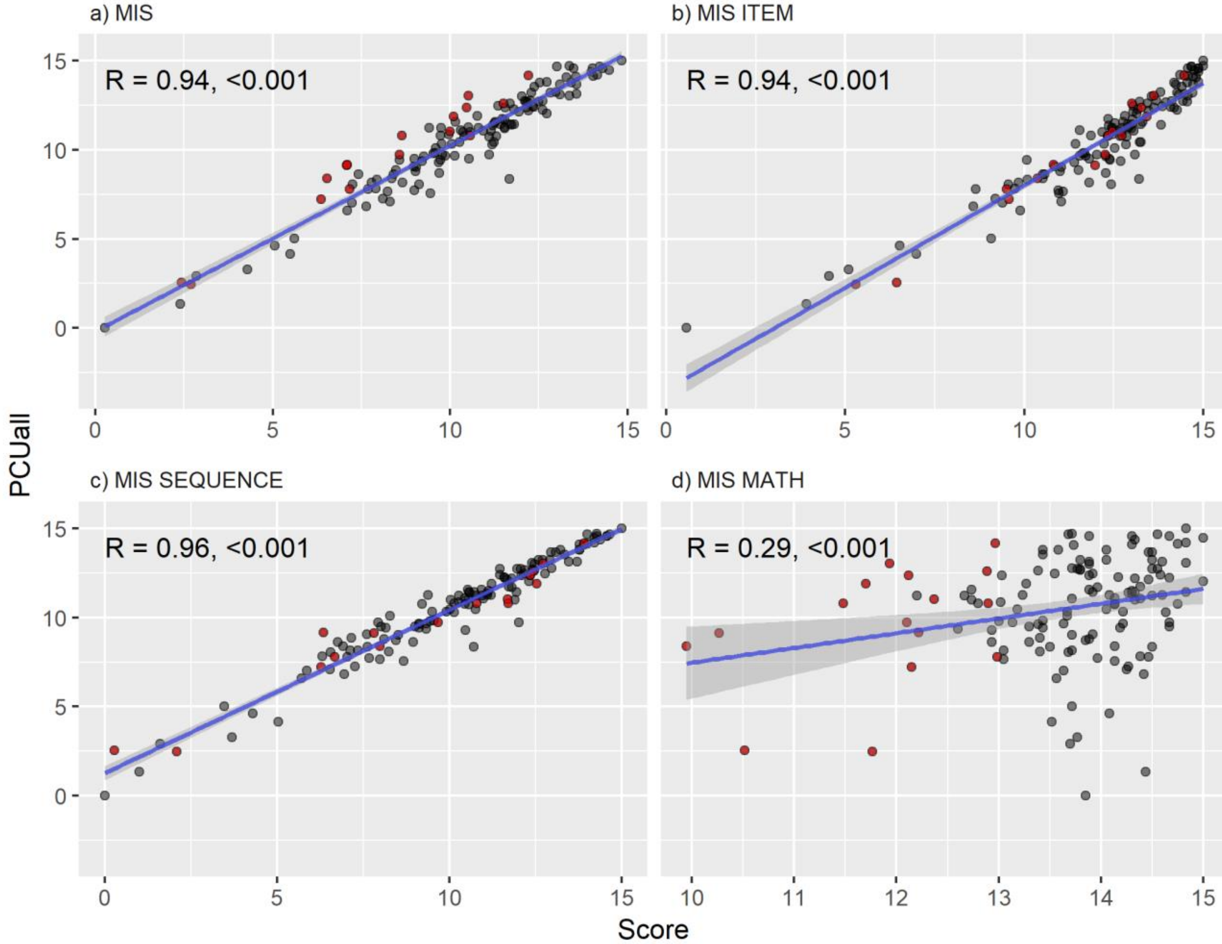


*Figure 1 Relationship between $PCU_{all}$ and MIS (a), $MIS_{item}$ (b), $MIS_{sequence}$ (c) and $MIS_{math}$ (d) scores, respectively. Red datapoints represent participants who would be excluded under PCU's exclusion criterion – below 85% performance on the distractor task. Blue lines indicate best fit, and shaded areas indicates 95% confidence intervals.*

### 3.3. Within MIS correlations

As expected, we found significant correlations between MIS and $MIS_{math}$, MIS and $MIS_{item}$, and MIS and $MIS_{sequence}$ (Table 3). Lower correlations of $MIS_{math}$ and MIS indicate lower contribution of this subscore to the full score, compared to the $MIS_{item}$ and $MIS_{sequence}$ scores. Further, our analyses show significant correlations between all MIS subscores (Table 3).

*Table 3 Correlation of Ospan scores for all samples*

| | Sample 1 | | Sample 2 | | Sample 3 | | Sample 4 | | Sample 5 | | Sample 6 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ρ | p-value | ρ | p-value | ρ | p-value | ρ | p-value | ρ | p-value | ρ | p-value |
| MIS * PCU | 0.955 | <0.001 | 0.947 | <0.001 | 0.952 | <0.001 | 0.943 | <0.001 | 0.95 | <0.001 | 0.959 | <0.001 |
| MIS * $PCU_{all}$ | 0.94 | <0.001 | 0.937 | <0.001 | 0.941 | <0.001 | 0.948 | <0.001 | 0.942 | <0.001 | 0.962 | <0.001 |
| $MIS_{item}$ * $PCU_{all}$ | 0.943 | <0.001 | 0.942 | <0.001 | 0.955 | <0.001 | 0.9 | <0.001 | 0.94 | <0.001 | 0.956 | <0.001 |
| $MIS_{sequence}$ * $PCU_{all}$ | 0.964 | <0.001 | 0.955 | <0.001 | 0.976 | <0.001 | 0.974 | <0.001 | 0.964 | <0.001 | 0.967 | <0.001 |
| $MIS_{math}$ * $PCU_{all}$ | 0.286 | <0.001 | 0.263 | 0.003 | 0.271 | 0.002 | 0.294 | 0.015 | 0.364 | <0.001 | 0.441 | 0.008 |
| MIS * $MIS_{item}$ | 0.95 | <0.001 | 0.937 | <0.001 | 0.933 | <0.001 | 0.93 | <0.001 | 0.942 | <0.001 | 0.957 | <0.001 |
| MIS * $MIS_{sequence}$ | 0.939 | <0.001 | 0.952 | <0.001 | 0.944 | <0.001 | 0.947 | <0.001 | 0.947 | <0.001 | 0.984 | <0.001 |
| MIS * $MIS_{math}$ | 0.493 | <0.001 | 0.456 | <0.001 | 0.5 | <0.001 | 0.45 | <0.001 | 0.561 | <0.001 | 0.553 | <0.001 |
| $MIS_{item}$ * $MIS_{sequence}$ | 0.943 | <0.001 | 0.942 | <0.001 | 0.946 | <0.001 | 0.918 | <0.001 | 0.933 | <0.001 | 0.954 | <0.001 |
| $MIS_{math}$ * $MIS_{item}$ | 0.289 | <0.001 | 0.229 | 0.009 | 0.243 | 0.006 | 0.202 | 0.098 | 0.361 | <0.001 | 0.401 | 0.017 |
| $MIS_{math}$ * $MIS_{sequence}$ | 0.241 | 0.005 | 0.248 | 0.005 | 0.269 | 0.002 | 0.23 | 0.060 | 0.361 | <0.001 | 0.465 | 0.005 |

*N – sample size, ρ - Spearman's correlation coefficient*

## 4. Discussion

According to Conway, scoring of the Ospan task is a neglected topic (Conway et al., 2005). We have therefore come up with a new scoring system that introduces three improvements in comparison to previously used scoring systems: 1) it does not require an arbitrary exclusion threshold on the distractor task, 2) it differentiates between two independent aspects of the span task – memory of the identity of a letter, and memory of its place in the sequence, and 3) it takes into account performance on the distractor task for scoring of the span task. In 6 independent samples we show that internal consistency of MIS is good, and that MIS highly correlates with previous scoring systems, thus preserving between-subject variability in task performance. In addition, we show that the individual subscales of MIS have good internal consistency and correlate with previous scoring systems.

Cronbach's alpha, a measure of internal consistency was above 0.80 for each sample tested in our study, indicating that all MIS items measure the same concept. In comparison, internal consistency for previous methods were similar or lower, e.g. 0.81 for PCU (Conway et al., 2005), or 0.78 for ABS (Unsworth, Heitz, Schrock, et al., 2005). Further, value of Cronbach's alpha for item and sequence subscales was above 0.80, also indicating good internal consistency. In the case of math subscale, where internal consistency varied in values (0.33-0.61), Cronbach's alpha was calculated using a ratio of correct mathematical responses for each trial. Since the mathematical tasks are randomly assigned to trials and participants, Cronbach's alpha compared results of participants who solved different mathematical equations. Even though these equations are designed to have similar difficulty, small differences in this respect might result in smaller and variable internal consistency of the $MIS_{math}$ measure.

We further tested correlations between MIS and PCU scores calculated in the same samples. High correlations were expected, as we compared scoring systems reflecting similar underlying constructs and relying on the same data. In all tested samples correlations were highly significant and correlation coefficients were above 0.93, indicating that MIS and PCU indeed measure similar concepts, and that MIS preserves rank order of subjects of PCU. This is important, as PCU is generally the recommended method to score the Ospan task, and is validated as a working memory measure (Conway et al., 2005).

As hypothesised, correlations between the span parts of MIS ($MIS_{item}$ and $MIS_{sequence}$), and PCU were high and indicated similarities between the two systems. This was expected, as PCU only takes into account the span part of the Ospan task. We observed a lower correlation between PCU and $MIS_{math}$ – which is due to the fact that PCU does not consider performance on the distractor task. This analysis shows that individuals with low math performance can have high scores on PCU. Here, low MIS score can be interpreted in two ways: first, some participants performed well on the span task just because they did not focus on the distractor task, which facilitated the span task. Second, participants focused on the distractor task, but could not perform the simple calculations well for various reasons (math skills, time pressure etc.), yet were still able to perform well on the span part. This, however, is unlikely, since the calculations in the Ospan task are very simple. In the first case, one might argue for exclusion of those individuals, as is done in PCU, because the task does not measure memory performance with an ongoing distractor task. However, in MIS, lower math score will influence span part, resulting in lower overall MIS scores. Admittedly, in the second case one should not downgrade performance on the span part of the task, however, we believe that this is still a better solution than exclusion of participants. Since it is impossible to tell which of those two possible mechanisms took place, we argue that excluding participants with low math performance would not be a good practice and could lead to unnecessary decreases in sample sizes. This might be especially problematic in studies on clinical populations, where participants are difficult to recruit. This concern is alleviated by including $MIS_{math}$ in MIS which constitutes a correction tool to 'punish' participants who performed poorly on the distractor task. One suggestion for future studies using Ospan tasks is to include a baseline test of mathematical abilities. This would allow researchers to compare Ospan mathematical tasks performance with participants' baseline abilities. One could then apply the downweighing of Ospan scores by math scores only if participants' Ospan math performance is lower than their baseline math performance.

To investigate how individual MIS subscores differentially contribute to the overall MIS score, we performed correlations between those measures. We found high correlations (above 0.93) between $MIS_{item}$, $MIS_{sequence}$, and MIS, and a lower correlation (0.26) between $MIS_{math}$ and MIS. This suggests that there is a high contribution of $MIS_{item}$ and $MIS_{sequence}$ to MIS, meaning that MIS reflects mainly span components measured by remembering the item and its place in the sequence. On the other hand, scores on the distractor task do not correlate as well with MIS, indicating that participants who performed poorly on the distractor task could perform well on the span part, and *vice versa*. Including $MIS_{math}$ is an important aspect of MIS, because participants who performed equally well on the span task, but differently on the distractor task will obtain a different score. This allows differentiating between participants who performed a simple span task (while ignoring the distractor task) and those who performed the intended complex task.

Overall, MIS is an Ospan task scoring method with a number of advantages over PCU. Firstly, there is no need for exclusion of participants based on performance, which is especially important in clinical studies, where sample sizes are small and participants' exclusions can be very costly. Such exclusions might also lead to biased samples or limit the generalisability of potential findings, so they should be avoided, which is what MIS aims at. Secondly, MIS differentiates between remembering an item and remembering its place in the sequence. We suggest that these pieces of information can be remembered independently, and therefore should be taken into account separately, even if they highly correlate with each other. Lastly, including math score in MIS ensures that participants who did not focus on the distractor task and performed only the span part will have lower overall scores for the complex task.

Colom, R., Rebollo, I., Abad, F. J., & Shih, P. C. (2006). Complex span tasks, simple span tasks, and cognitive abilities: A reanalysis of key studies. *Memory & Cognition, 34*(1), 158-171. doi:10.3758/BF03193395

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. doi:10.1016/S0160-2896(01)00096-4

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 & Review, 12*(5), 769-786. doi:10.3758/BF03196772

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297-334. doi:10.1007/BF02310555

Daneman, M., & Merikle, P. M. (1996). Working memory and language comprehension: A meta-analysis. *Psychonomic Bulletin & Review, 3*(4), 422-433. doi:10.3758/BF03214546

Draheim, C., Harrison, T. L., Embretson, S. E., & Engle, R. W. (2018). What item response theory can tell us about the complex span tasks. *Psychological Assessment, 30*(1), 116-129. doi:10.1037/pas0000444

Engle, R. W., Kane, M. J., & Tuholski, S. W. (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 *Models of working memory: Mechanisms of active maintenance and executive control* (pp. 102-134). New York, NY, US: Cambridge University Press.

Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. (1999). Working memory, short-term memory, and general fluid intelligence: a latent-variable approach. *Journal of Experimental Psychology. General, 128*(3), 309-331.

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. doi:10.1037/0096-3445.133.2.189

Kane, M. J., Meier, M. E., Smeekens, B. A., Gross, G. M., Chun, C. A., Silvia, P. J., & Kwapil, T. R. (2016). Individual differences in the executive control of attention, memory, and thought, and their associations with schizotypy. *Journal of Experimental Psychology. General, 145*(8), 1017-1048. doi:10.1037/xge0000184

Klein, K., & Fiss, W. H. (1999). The reliability and stability of the turner and Engle working memory task. *Behavior Research Methods, Instruments, & Computers, 31*(3), 429-432. doi:10.3758/BF03200722

RCoreTeam. (2015). R: A Language and Environment for Statistical Computing. Retrieved from www.R-project.org

Revelle, W. (2018). psych: Procedures for Personality and Psychological Research. In.

RStudioTeam. (2016). RStudio: Integrated Development for R. Retrieved from www.rstudio.com

Shipstead, Z., Harrison, T. L., & Engle, R. W. (2016). Working Memory Capacity and Fluid Intelligence: Maintenance and Disengagement. *Perspectives on Psychological Science, 11*(6), 771-799. doi:10.1177/1745691616650647

Turner, M. L., & Engle, R. W. (1989). Is working memory capacity task dependent? *Journal of Memory and Language, 28*(2), 127-154. doi:10.1016/0749-596X(89)90040-5

Unsworth, N., & Engle, R. W. (2007). On the division of short-term and working memory: An examination of simple and complex span and their relation to higher order abilities. *Psychological Bulletin, 133*(6), 1038-1066. doi:10.1037/0033-2909.133.6.1038

Unsworth, N., Heitz, R. P., & Engle, R. W. (2005). Working Memory Capacity in Hot and Cold Cognition. In *Cognitive limitations in aging and psychopathology* (pp. 19-43). New York, NY, US: Cambridge University Press.

Unsworth, N., Heitz, R. P., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. *Behavior Research Methods, 37*(3), 498-505. doi:10.3758/BF03192720

Westbrook, J. I., Raban, M. Z., Walter, S. R., & Douglas, H. (2018). Task errors by emergency physicians are associated with interruptions, multitasking, fatigue and working memory capacity: a prospective, direct observation study. *BMJ quality & safety, 27*(8), 655-663. doi:10.1136/bmjqs-2017-007333