PSICOLÓGICA

Measuring Verbal Working Memory Online in Research: Feasibility, Reliability, and Validity of two Implementations

Federico M. Gonzalez[1,2], Jonathan Marrujo[1], Magalí Martínez[1,2], Juan Pablo Barreyro[2,3], and Débora Burin[1,2]

[1] *Instituto de Investigaciones, Facultad de Psicología, Universidad de Buenos Aires, Buenos Aires, Argentina.*
[2] *Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Argentina.*
[3] *Centro Interdisciplinario de Investigaciones en Psicología Matemática y Experimental (CIIPME), Argentina.*

Abstract

Working memory (WM) tasks have been extensively studied with laboratory paper and pencil or computerized tasks; few studies have assessed the feasibility and psychometric properties of WM online tests for low-stakes research purposes. This study analyzed the online implementation of two verbal WM tasks, Letter-Number Sequencing and Running Span. Letter-Number Sequencing was tested in a supervised, online video conference setting in small groups. Running Span was assessed asynchronously and unsupervised. Both had adequate reliability. For Letter-Number Sequencing, approximately 25% of the sample had to be excluded following criteria for minimum performance and strategic behavior. In addition, performance did not follow the set size effect, a benchmark for WM tasks, and was systematically different from an in-person sample. In contrast, the remote online Running Span task displayed systematic lower performance with more to-be-remembered items (set size effect), with overall performance aligning closely with a similar in-person evaluation.

Keywords: working memory, online testing, remote testing, reliability, validity, set size effect.

Psicológica (2024)

DOI
10.20350/digitalCSIC/16342

Corresponding author
Federico M. Gonzalez
fmgonzalez@psi.uba.ar

Edited by
Tamara Giménez Fernández

Reviewed by
Cristina López Rojas, Miguel A. Pérez Sánchez

Introduction

Working memory tasks

Working memory (WM) is a cognitive system for the concurrent temporal maintenance and information processing required to execute complex cognitive tasks such as reasoning, comprehension, and learning (Baddeley, 2010; Miyake & Shah, 1999b; Oberauer et al., 2018). Among other robust results, there are distinct developmental and aging patterns in WM task performance (e.g. Alloway et al., 2017; Jaroslawska & Rhodes, 2019), and WM is strongly correlated with measures of fluid intelligence (Oberauer et al., 2018). Since the pioneering work of Baddeley and Hitch (1974), an extensive amount of research with normal and neurologically affected participants along the whole lifespan, computational approaches, and neuroimaging has led to the development of diverse theoretical models of working memory (e.g., Miyake & Shah, 1999a; Logie et al., 2020).

Two types of models have emerged (e.g., Rhodes et al., 2021): the multicomponent view (Baddeley, 2010; Baddeley et al., 2020) and the limited (attentional) resources view (Barrouillet & Camos, 2020; Cowan, 2017; Cowan et al., 2020; Kane et al., 2004; Oberauer, 2020). The multicomponent model conceptualizes WM as a system composed of interactive domain-specific and multimodal temporary memory stores subsystems, and control or executive processes (Baddeley et al., 2020; Logie, 2016). In these models, verbal working memory consists of a passive store plus active rehearsal processes, both verbally-based; this component is in charge of verbal inputs even with visual presentations (Baddeley & Hitch, 1974; Baddeley et al., 2020). Overall, WM capacity restrictions arise from interactions among limited modality specific and executive control components. On the other hand, other theories emphasize that a central limiting factor in WM is a limited (attentional) central resource. Kane, Cowan, Engle, and colleagues (Cowan, 2017; Cowan et al., 2020; Mashburn et al., 2020; Oberauer, 2020) propose that WM capacity is the ability of executive attention to maintain focus on, and disengage from, incoming information or temporarily activated long-term memory representations. For example, in a latent variable approach involving several memory and ability tests, Kane et al. (2004) found that WM tasks reflected a domain-general factor, and also shared variance with modality specific factors, reflecting verbal – although visually presented - and visuospatial span tasks.

While conceptual definitions of the working memory cognitive system are a matter of debate, the operational definition of individual differences in working memory "is fairly straightforward: It is the number of items that can be recalled during a complex working memory task. Complex working memory tasks have simultaneous storage (maintaining information in an active state for later recall) and processing (manipulating information for a current computation) components." (Feldman Barrett et al., 2004, p. 553). Complex span tasks can be implemented by a list of items interleaved with brief episodes of a distractor processing task, and serial recall order at the end, or by requiring the participant to manipulate those same items, and report them back, transformed. In these tasks, the number of to-be-remembered elements is systematically varied, so that the participant needs to immediately retain and report a particular amount or size of elements for a fixed set of items. The set size is increased until the participant fails all items. The set size effect on accuracy is a highly robust and general phenomenon (Oberauer et al., 2018): on every test of WM, accuracy declines as the set size increases. It is considered a benchmark for WM task validity (Oberauer et al., 2018).

Online testing

WM tasks have been extensively studied with laboratory paper and pencil or computerized tasks. However, online and remote studies are increasingly needed in psychology and neurocognitive research, due to their cost-effectiveness, participant reach, and other advantages in data collection

(Gosling & Mason, 2015; Reips, 2002). Online and remote testing can even be required by external circumstances preventing lab studies, such as COVID-19 pandemic restrictions, or hard to reach samples.

Evidence of adequate functioning in a lab context is not generalizable to an online and remote assessment scenario. In this sense, the International Test Commission (ITC, 2022, p. 8) has noted two threats to computerized testing validity: construct under-representation, "which means the test is not fully measuring what it intends to measure", and construct-irrelevant variance, "when item or test scores reflect factors the test was not intended to measure". Examples of threats to validity are differences in computer skills or knowledge, differences in the device in which the test is presented such as screen size, timing resolution, response input method, and other instrumental and contextual factors. The ITC recommends controlled and proctored administration for computerized assessment, to unify this type of variance.

Controlled assessment is also suggested because of issues such as receiving help or using aids while taking the test, copying answers, taking photos or screen capturing, or taking written notes (ITC, 2022). These behaviors have been considered fraudulent or cheating (ITC, 2022). They can also reflect unintended (by test designers) solution strategies because they lack a supervisor making sure that the task instructions are properly understood and that the task is being executed according to pre-determined rules. The behavior might not intentional, but also leads to a loss of validity. In a preliminary observational report on adapting a computerized WM operation span task for web-based testing, Leidheiser et al. (2015) observed these potential threats to online testing validity, and others related to the user interface and technical issues.

The purpose of testing can be a relevant factor (ITC, 2022). When tests are employed for selection and/or placement, it is considered high stakes testing, and cheating and fairness are main concerns (Steger et al., 2020). In contrast, assessment for research purposes has been considered low stakes (Do, 2009). However, a meta-analysis (Steger et al., 2020) revealed significantly higher scores on cognitive tests in unproctored settings as compared to proctored test contexts, and suggested that cheating or solution strategies such as looking up information was similar in high and low stakes online testing. Another concern in research testing is the rising trend of psychological research employing crowdsourcing platforms such as Amazon Mechanical Turk, who pay participants according to task, which has increased fraudulent behaviors (Hauser et al., 2022; Webb & Tangney, 2023). Finally, another different context is clinical assessment, where the focus is on understanding the patient's case. For example, in neuropsychological assessment, the use of videoconferencing technology for individual remote testing has been recommended (Crivelli et al., 2022).

Online WM tests

Several studies have compared psychometric properties of lab versus online and remote WM tasks (Backx et al., 2020; Hicks et al., 2016; Kulikowski & Potasz-Kulikowska, 2016; Leong et al., 2022; Ruiz et al., 2019). Hicks et al. (2016) compared online versus lab test performance on different complex span tasks in visual and verbal modalities. To evaluate construct validity, in addition to WM capacity tasks they also measured fluid intelligence (Oberauer et al., 2018). Fifty-eight participants completed all lab and online tasks. Although they found that the online WM task predicted fluid intelligence, the prediction was worse than in-person. In a second, online only, study (100 Amazon Mechanical Turk workers), they found differences between visual and verbal WM associations with fluid intelligence. Participants with low fluid intelligence had high WM capacity in the verbal modality but low in the spatial tasks, and there was a very low association between online verbal WM capacity and fluid intelligence. This pattern led Hicks and colleagues to suspect that some of the participants from both studies had written down the WM letter stimuli. In two subsequent experiments (102 participants from their lab database, and 112 from Amazon Mechanical Turk) they reconsidered the stimuli employed to prevent copying and strategic behavior, adopting viso-spatial and operation span tasks. In this case they found the expected association with fluid intelligence (Oberauer et al., 2018). They concluded that WM online tasks were feasible; however, the lack of online control might lead to fraudulent or strategic behavior (e.g. writing down the letter stimuli).

Outside the United States, Kulikowski & Potasz-Kulikowska (2016) implemented an n-back task in an online and remote environment. Polish participants (169 college students) completed an nback task with 9 blocks of 3 n-levels (n1, n2, and n3, 3 times each). Split half Pearson's correlations indicated good reliability for overall accuracy and reaction times, as well as good reliability for reaction times, but not for accuracy measures. As for factorial validity, CFA fit indexes were best for a two-factor model that separated accuracy (n1+n2+n3) from reaction times (n2+n3). These results were considered consistent with previous research. In Germany, Ruiz et al. (2019) tested forty-three college students in lab and web-based versions of WM and declarative memory. The WM task was based on the operation span task with visuospatial symbols adapted from Hicks et al. (2016). They found high correlations between the lab and web based equivalent measures. It should be noted that reliability for the online WM task could not be computed and was not reported. In the UK, Backx et al. (2020) compared unsupervised web-based and lab-based administration of the Cambridge Neuropsychological Test Automated Battery (CANTAB), in a sample of 51 participants. They found low to moderate bivariate and intraclass correlations for various indices, and no significant differences between inperson versus online, but discrepancies in reaction times. Finally, Leong et al. (2022) tested 85 English speaking Singaporean healthy young adults (41 in the lab, 44 on an identical web-based platform) who completed ten cognitive tasks assessing neuropsychological executive function and learning (CANTAB, Inquisit, i-ABC). They implemented a supervised methodology, Remote Guided Testing, which simulated the lab setting via video conferencing. The experimenter provided technical support, as well as guidance on environment optimization (lightning, distractions) and comfort breaks. They found no significant differences in task performance across all measures, except for a vocabulary task (higher scores online). In contrast to previous studies, response times did not differ, which suggests supervised online assessment might be a good alternative for studies with time dependent tasks and high-stake contexts.

Overall, previous studies have shown the feasibility of online WM testing in North America and European based samples (except for a sample of English-speaking Singapore students). These studies have established validity by finding a high correlation between web and lab-based measures, and association between WM and related constructs. They have employed synchronous video conferencing and asynchronous unsupervised procedures. There are caveats, though: all studies had participant attrition in online testing, and there was evidence of possible cheating or strategic behavior in some participants, especially for verbal material. Moreover, sample size for in-person versus online comparisons has been small, between forty and fifty participants per group, except for the lab database versus Amazon Mechanical Turk comparison in Hicks et al. (2016). Such small samples are a limitation for psychometric analyses. Finally, while one of the proposed advantages of online testing is its ability to reach varied populations, except for an English-speaking Singaporean sample, most studies have been carried out in North America and European based samples, termed WEIRD samples for their narrow scope (Henrich et al., 2010).

The present studies

The objective of the present paper has been to analyze the feasibility, reliability, and validity of two WM tasks, Letter-Number Sequencing and Running Span, for online testing. We report two studies comparing results obtained in online testing with results in similar samples assessed in the lab. These latter are secondary data, already published (Barreyro et al., 2015a, Burin et al., 2021). Online tasks were designed to be as close as possible as the ones for which we had previous data, adaptations of Number Letter Sequencing (Barreyro et al., 2015b; Burin et al., 2021) and Running Span (Barreyro et al., 2015a) for research purposes. The online WM tasks were part of larger studies, where they were completed before other experimental tasks, not reported here.

We explored two different testing procedures suggested by the literature review, one synchronous, and another asynchronous. Study 1 and Study 2, respectively, report the implementation and results for Number-Letter Sequencing and Running Span. In the first case, the task was performed in a supervised, online video conference setting, and in the second case, in an unsupervised asynchronous remote assessment. We analyzed reliability, item and set size (number of to-be-remembered elements per item) distributions, and compared them with data from previous in-person assessments. The research questions, for both studies, were:

1. Was the implementation feasible? Were major flaws found?

2. Does the online test have adequate reliability?

3. Are in-person and online set size scores equivalent?

4. Does online assessment difficulty increase as a function of set size?

An additional objective of these studies was to generalize this type of research out of WEIRD samples, in our case to Latin American participants.

Study 1

The objective of this study was to implement an online Number-Letter Sequencing test and compare its results with an in-person Letter Number Sequencing version for collective assessment, previously assessed in a similar sample (Burin et al., 2021). Given that working memory verbal tasks have routinely employed visual presentations, the in-person Letter-Number task was an adaptation for collective testing which included changing the original auditory stimuli to a visual presentation in a large screen (Barreyro et al., 2015b; for similar changes in this task see e.g. Emery et al., 2007; Macnamara & Conway, 2016). This adaptation was required because administering auditory stimuli for a group of participants poses challenges in terms of quality of emitted sound, site acoustics, and possible individual differences in hearing that cannot be accommodated in a collective session. In this vein, systematic comparisons between visual and auditory presentations of the Letter-Number Sequencing task have revealed that while visual presentations do not impair monolingual participants' performance, it does affect bilingual participants, suggesting the need for visual presentations for this case and other populations (Mielicki et al., 2018).

To overcome possible cheating or solutions strategies when employing verbal stimuli in the online case (e.g., Hicks et al., 2016) the assessment session was a supervised online video conference setting, in small groups (Backx et al., 2020; Crivelli et al., 2022; Leong et al., 2022). This synchronous assessment was intended to emulate the already validated in-person assessment, employing the same task, stimuli, and general procedure, with necessary adaptations. We analyzed reliability, set size effect, and set size accuracy compared to the previous similar in-person administration.

Method

Participants

Two first year psychology college students at a large public university in Argentina samples were included. The in-person sample is a reanalysis of data already collected and reported in the context of other measures (Burin et al., 2021). They were 238 students, 181 women, 57 men; *M* age = 22.71, *SD* =6.37. The online sample comprised 115 participants, 87 women, 28 men; *M* age = 25.39, *SD* = 7.63. Both samples took part of the study voluntarily, in exchange for partial credit in a general psychology course. In-person participants signed an informed consent form, and online participants clicked on an informed consent button. The study was approved by the institutional review board.

Materials and procedure

Letter Number Sequencing, In-person. The Letter Number Sequencing task from the WAIS III battery (Wechsler, 2003) was adapted for small groups assessment in research settings (Barreyro et al., 2015b). A series of letters and numbers were visually presented on a large screen, one at a time for 1.5 sec., followed by a recall cue. The participant had to write the sequence rearranging the stimuli, first the numbers, in ascending order, and then the letters, alphabetically ordered. The test shows three items for each set size, starting at set size 2, up to set size 8.

Instructions and stimuli were implemented in a Powerpoint presentation and projected on a large screen (approximately 225 cm x 1.35 m) to groups of 10-15 participants, who responded in a booklet. A researcher read aloud the instructions and practice items and made sure that all participants understood and followed instructions; in each session there was an additional research assistant monitoring the session. The task was implemented in Powerpoint given its cost, portability between different computers and screen types, possibility and ease of changing from autoplay to manual control, and duration of stimuli not requiring precise milliseconds accuracy. This implementation was similar to Shelton et al.'s (2007), who demonstrated validity of the procedure for WM task testing in small groups. More details can be found in Barreyro et al. (2015b).

Although all participants completed all items, scoring followed the standard procedure (Barreyro et al., 2015b; Wechsler, 2003): when the participant scored zero on a set size, no further answers were considered.

Letter Number Sequencing, Online. The same task as before was adapted for online, synchronous small groups assessment, in a Zoom video conference (https://zoom.us/). Data were collected when students had more than one year experience with video conferences and online learning tools due of pandemic lockdown.

Ten to fifteen participants took part in each session, in which a research assistant explained the task and response format and practiced how to respond. Before the test started, the research assistant helped students with possible technical difficulties. Given that the in-person previously validated assessment employed a Powerpoint presentation, and the videoconference environment allowed to employ the exact same presentation ("share " Zoom function), instructions and stimuli were the same Powerpoint presentation as in-person. The presentation was screen-shared during the videoconference. Response was implemented in a gamification app, Quizizz (https://quizizz.com), which advanced synchronically with the presentation. Participants downloaded the app to their cellphones and opened it while in the Zoom session, so that for each item, they saw item ID and had to enter their response on the app.

As in the in-person assessment, participants completed all items, but scoring followed a discontinuation rule (Barreyro et al., 2015b; Wechsler, 2003).

Data analyses

Data analyses were carried out in R 4.3.0 (R Core Team, 2023) and RStudio 2023.3.1 (RStudio Team, 2023). Cronbach's αs were estimated to determine internal consistency with the psych 2.1.6 package (Revelle, 2021). Descriptive statistics and visualizations were carried out with the tidyverse 2.0 package (Wickham et al., 2019). Generalized linear mixed models were built with lme4 1.1-32 (Brooks et al., 2017), mean contrasts with emmeans 1.8.5 (Length, 2021), and visualization with sjPlot_2.8.14 (Lüdecke, 2023).

Data and analytical code for the study can be found at the Open Science Framework platform, data: https://osf.io/79gs5/, code: https://osf.io/aetxc/

Results

Was the implementation feasible? Were major flaws found?

First, we explored whether we could detect implementation problems or visible flaws in response patterns in the online sample. Given that during testing some participants expressed technical problems or trouble understanding the procedure, we identified the minimum score in the in-person sample (4 correct answers) to eliminate those online participants who did not meet this cutoff (e.g., number of correct items < 4). In addition, we identified and eliminated participants with all correct items in set size 5, but errors in set size 2 to 4 (2 cases), which raised suspicion of switching to writing down items (Hicks et al., 2016). Thus, the final sample comprised 86 participants, 74,78% of the initial sample.

Did the online test have adequate reliability?

Reliability in terms of internal consistency was adequate, both when taking into account each item's score, α Cronbach = .78, CI95% [.72 - .85], and when scoring for set size (number of correct items, 0 to 3), α Cronbach = .72, CI95% [.63 - .81].

Were in-person and online set size scores equivalent?

Score on each set size was the number of correct items (0-3). Table 1 shows descriptive statistics (mean score, SD) for in-person (left) and online (right) accuracy as a function of set size.

Table 1. *Descriptive Statistics (Mean, SD) for Accuracy Scores (Number of Correct Responses) for In-person (Left) and Online (Right) Letter Number Sequencing, per Set Size.*

	In-person		**Online**	
Set size	***M***	***SD***	***M***	***SD***
2	2.825	0.551	2.128	0.918
3	2.647	0.724	2.616	0.785
4	2.147	1.006	2.314	0.871
5	1.475	1.172	1.465	0.929
6	1.076	1.188	1.198	1.027
7	0.483	0.846	0.837	1.115
8	0.155	0.491	0.477	0.955

We tested the effects of Type of testing (In-Person, Online) and Set size (2 to 8) in a linear mixed model (Bates, 2010) with those two as fixed factors, participants as random factor, and accuracy as dependent variable, with sum contrasts for both factors, followed by the model's ANOVA with Satterthwaite's method. Post-hoc pairwise comparisons use Tukey HSD method, and Kenward-Roger method to estimate degrees of freedom. For the mixed model, *Marginal* R^2 = 0.5 and *Conditional* R^2 = 0.667. Modality did not have a significant effect, *F* (1,322) = 0.204, *MS* = 0.112, *p* = 0.652; there was a significant effect of Set size, *F* (6,1932) = 381.405, *MS* = 209.446, *p* < 0.001; and the Modality X Set size interaction, *F* (6,1932) = 14.123, *MS* = 7.756, *p* < 0.001. Post-hoc analyses of the interaction reveal higher scores for in-person for set size 2, *t* (1352) = -6.017, *p* < .001, higher scores for online for set size 7, *t* (1352) = 3.100, *p* = .002, and set size 8, *t* (1352) = 2.813, *p* = .005, all other differences were not significant. Figure 1 shows predicted accuracy for in-person and online scores for each set size.

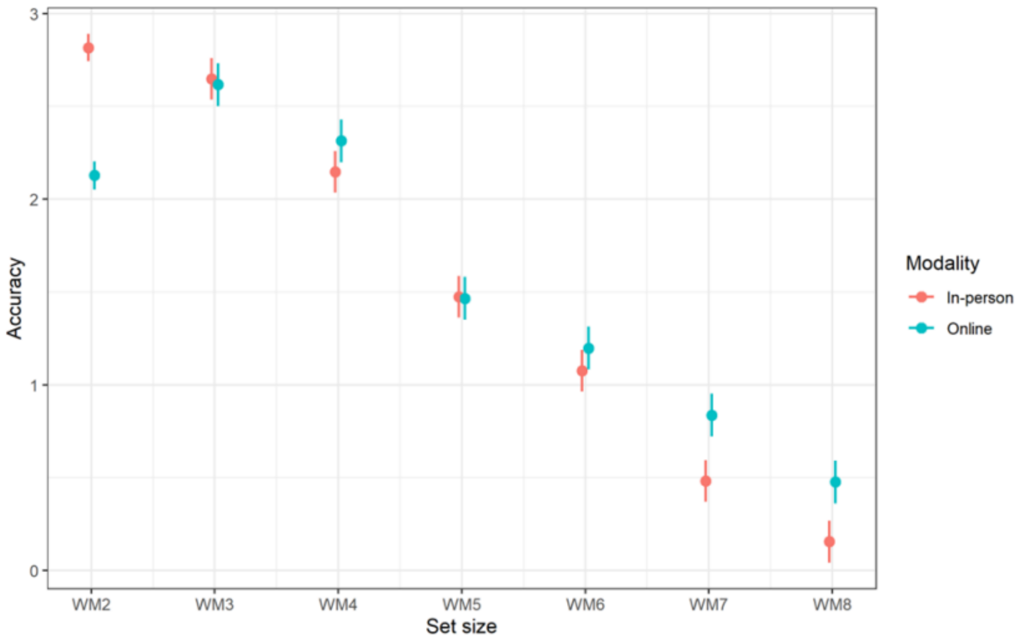

Figure 1. Predicted accuracy (scores and confidence intervals) by set size, for in-person and online Letter Number Sequencing.

Did online difficulty increase as a function of set size?

In the online sub-sample, we tested the effect of Set size (2 to 8) as fixed factor, participants as random factor, and accuracy as dependent variable, with a linear mixed model with Helmert contrast, followed by the model's ANOVA with Satterthwaite's method. Post-hoc pairwise comparisons use Tukey HSD method, and Kenward-Roger method to estimate degrees of freedom. For the generalized mixed model, *Marginal* R^2 = 0.380 and *Conditional* R^2 = 0.545. There was a significant effect of Set size, *F* (6,510) = 83.573, *MS* = 55.054, *p* < 0.001. Post-hoc analyses of the interaction reveal that set size 2 had significantly lower scores than set size 3, *t* (510) = -3.946, *p* < .001. In addition, there were no significant differences in scores between set size 3 and 4, *t* (510) = 2.443, *p* = .183, between set size 5 and 6, *t* (510)

= 2.161, *p* = 0.319, between set size 6 and 7, *t* (510) = 2.912, *p* = .057, or between set size 7 and 8, *t* (510) = 2.912, *p* = .057.

Discussion

We implemented a synchronous video conference assessment of verbal WM, an adaptation of Number Letter Sequencing, for testing small groups in research, with college students. We found that 25% of the sample had to be discarded because of visible procedural flaws, such as technical problems with the implementation or with the testing procedure, and detectable cheating and/or solution strategies, e.g., much better scores in the more difficult items compared to the lower set size. In addition, when comparing results with a similar task and sample tested in the lab, performance was significantly lower for the smallest set size (set size 2), and higher for the more difficult set sizes (7 and 8). Moreover, in online testing, performance was lower for set size 2 compared to set size 3, against the set size effect (Oberauer et al, 2018).

In line with problems observed before, worse performance in set size 2 might reflect problems adjusting to the testing procedure. Also, better performance in set sizes 7 and 8 compared to in-person arises suspicion of strategic behavior, such as writing down, copying, or capturing the items in some way (Hicks et al., 2016). Regarding the reliability and validity of the online task, although the test showed adequate reliability, there was not a pattern of significant differences between set sizes, which suggests that performance did not follow the expected set size effect.

Our results differ from Leong et al. (2022) who successfully implemented a supervised video conferencing methodology. However, in their case, testing was one-on- one, and the experimenter provided procedural guidance, technical support, and even comfort breaks. In contrast, this was a small groups assessment, which seems to have been ineffective to provide the guidance and support needed.

In addition, given the technical and procedural issues, differences in task duration, for example, in stimuli duration time, cannot be ruled out. In conclusion, we do not recommend synchronous video conference assessment of verbal WM for testing small groups of college participants in research. This type of implementation needs more initial attention to technical and procedural issues at the beginning, more practice items, and possible proctoring when taking the test (which might introduce even more technical and procedural problems) or one-on-one video conferences (Leong et al., 2022).

Study 2

Given the shortcomings of the WM implementations in Study 1, we sought another way to test verbal WM online and remote for research purposes. The objective of this study was to implement an online and remote version of a verbal WM task previously validated in a similar population, the Running Span test (Barreyro et al., 2015a). The Running Span task is fast paced and has an unpredictable component, which can help preventing solving strategies or other "gaming" behaviors. In addition, to overcome procedural problems with the synchronous group procedure, the assessment session was designed as an asynchronous online remote test (Hicks et al., 2016; Kulikowski & Potasz-Kulikowska; 2016; Ruiz et al., 2019). We report the task design, implementation, reliability, set size effect, and set size accuracy comparisons with a previous similar in-person administration.

Method

Participants

Two first year psychology college students' samples, at a large public university in Argentina, were included. The in-person sample, 106 students (62 women, 44 men; *M* age = 21.56, *SD* =3.87) has been already reported (Barreyro et al., 2015a), and is included here for comparisons with the online sample. The online sample comprised 289 participants, 248 after applying exclusion criteria, 219 women, 29 men; *M* age = 23.48, *SD* = 6.58. Both samples took part of the study voluntarily, in exchange for partial credit in a general psychology course. In-person participants signed an informed consent form, and online participants clicked on an informed consent button. The study was approved by an institutional review board.

Materials and procedure

Running Span, In-person. The Running Span task was adapted for small groups assessment (Barreyro et al., 2015a). The task presents, one at a time, a string of letters of variable length, and the participant had to remember the final letters (last 2, 3, 4, 5 or 6, depending on the item set size). Each item contained, randomly, 1 to 3 more letters than those to be remembered. There were three items for each set size, and they were presented in ascending order with the instruction to remember a particular number of letters (e.g. to remember the last two, three, four…). Stimuli were shown for 500 ms and appeared sequentially, followed by a cue to remember the stimuli.

As in the first case, the task was implemented in Powerpoint given the already mentioned benefits: low cost, easy to launch in different computers and screens, easy to control, precise milliseconds accuracy not needed. This implementation was similar to Shelton et al.'s (2007), with a similar task. More details can be found in Barreyro et al. (2015a).

In-person testing was similar to Study 1: participants were tested in small groups (10-15 participants), watched a Powerpoint presentation on a large screen, and responded writing in response sheets. The assessment session was conducted by a researcher who read aloud the instructions, practice items, and monitored comprehension and compliance, with an additional researcher.

Also similar to Study 1, all participants completed all items, but scoring followed a discontinuation rule if the participant scored zero on a set size.

Running Span, Online. The Running Span task was implemented as an online, remote task. The same stimuli as the in-person test were employed. The task instructions appeared first, followed by two sets of practice items of two and three letter widths each. Participants received one or two practice items for set size 2 and 3. After practice, the participant started the test by pressing enter; at the beginning of each set size (three items) a sign instructed how many items they had to remember. When the cue to remember appeared, participants entered their response with the keyboard, after which the next item appeared. Given that it was impossible to employ Powerpoint for remote and online testing, the task was implemented with PsyToolkit version 3.3.2 (Stoet, 2010, 2017). Participants received a link to the task, and carried it out remotely, at their home. The task was not enabled for cellphones or tablets.

As in all previous cases, all items were completed, but scoring followed a discontinuation rule: when the participant did not remember any item in the series correctly, no further responses were considered.

Data analyses

The same R environment and packages of Study 1 were employed.

Data and analytical code for the study can be found at the Open Science Framework platform,
data: https://osf.io/cdj4m/, code: https://osf.io/aetxc/

Results

Was the implementation feasible? Were major flaws found?

First, we explored whether we could detect problems with the task procedure or visible flaws in response patterns in the online sample. Given the asynchronous and remote testing, to screen for technical and procedural problems, we eliminated those who had accessed the task but did not complete it. In addition, we identified the minimum score in the in-person sample (1 correct answer) to eliminate those online participants who did not meet this cutoff (e.g., number of correct items < 1, N = 3). Finally, there were no participants with higher correct scores for bigger set sizes as compared to their performance in lower set sizes. Thus, the final sample comprised 248 participants, 85,81 % of the initial sample.

Did the online test have adequate reliability?

Reliability in terms of internal consistency was adequate, *α* Cronbach = .81, CI95% [.78 - .84].

Were in-person and online set size scores equivalent?

Score on each set size is the number of correct items (0-3). Table 2 shows descriptive statistics (mean, SD) for in-person (left) and online (right) accuracy as a function of set size.

Table 2. *Descriptive Statistics (Mean, SD) for Accuracy Scores (Number of Correct Responses) for In-person (Left) and Online (Right) Running Span, per Set Size.*

	In-person		**Online**	
Set size	***M***	***SD***	***M***	***SD***
2	2.821	0.432	2.677	0.629
3	2.057	0.984	1.932	1.020
4	1.104	1.179	0.614	0.911
5	0.349	0.718	0.243	0.699
6	0.085	0.368	0.124	0.548

We tested the effects of Type of testing (In-Person, Online) and Set size (2 to 6) in a linear mixed model (Bates, 2010) with those two as fixed factors, participants as random factor, and accuracy as dependent variable, with sum contrasts for both factors, followed by the model's ANOVA with Satterthwaite's method. Post-hoc pairwise comparisons use Tukey HSD method, and Kenward-Roger method to estimate degrees of freedom. For the mixed model, *Marginal* R^2 = 0.627 and *Conditional* R^2= 0.754. There were significant effects of Set size, *F* (4,1420) = 944.846, *MS* = 384.93, $p < 0.001$, decreasing scores with increased Set size; of Modality, *F* (1,355) = 6.697, *MS* = 2.83, $p = 0.008$, higher scores for In-person; and of Modality X Set size interaction, *F* (4,1420) = 6.989, *MS* = 2.85, $p < 0.001$. However, regarding Modality, post-hoc analyses of the interaction revealed only higher scores for in- person for set size 4, *t* (1209) = 5.378, $p < .001$; all other differences were not significant. Figure 2 shows predicted accuracy for in-person and online scores for each set size.

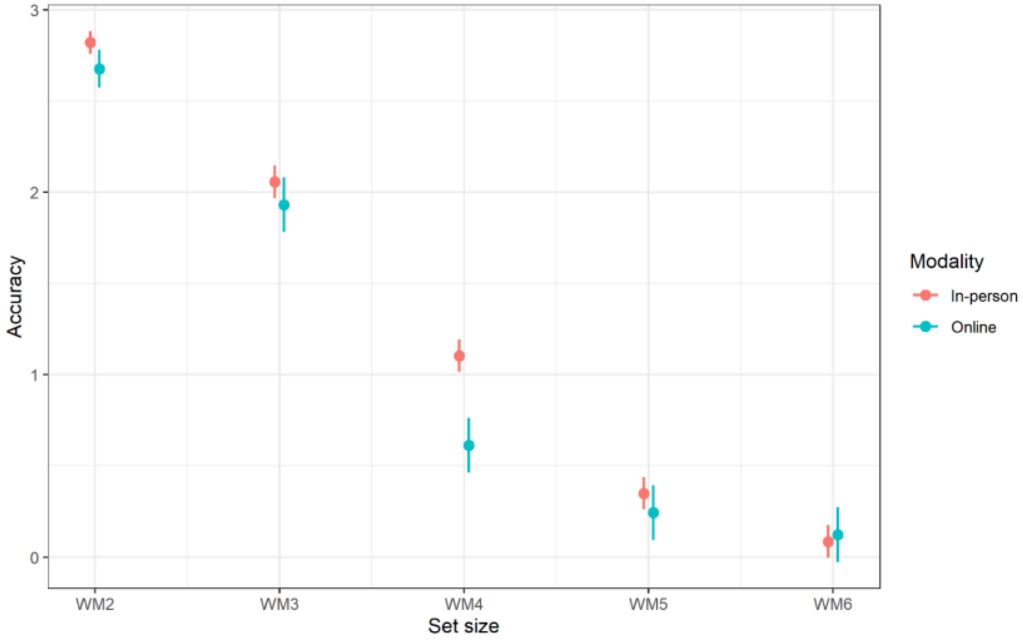

Figure 2. *Predicted accuracy (scores and confidence intervals) by set size, for in-person and online Running Span.*

Did online difficulty increase as a function of set size?

In the online sub-sample, we tested the effect of Set size (2 to 6) as fixed factor, participants as random factor, and accuracy as dependent variable, with a linear mixed model with Helmert contrast, followed by the model’s ANOVA with Satterthwaite's method. Post-hoc pairwise comparisons use Tukey HSD method, and Kenward-Roger method to estimate degrees of freedom. For the mixed model, *Marginal* R^2 = 0.626 and *Conditional* R^2 = 0.759. There was a significant effect of Set size, *F* (4,1000) = 814.82, *MS* = 320.24, *p* < 0.001. Post-hoc analyses reveal significant differences, favoring higher set sizes, between set size 2 and 3, *t* (1000) = 13.313, *p* < .001, set size 3 and 4, *t* (1000) = 23.565, *p* < .001, set size 4 and 5, *t* (1000) = 6.621, *p* < .001, but not between set size 5 and 6, *t* (1000) = 2.136, *p* = .206.

Discussion

We implemented an online verbal WM task, the Running Span, as an asynchronous online remote test (Hicks et al., 2016; Kulikowski & Potasz-Kulikowska; 2016; Ruiz et al., 2019) and compared it with a similar test employed in-lab previously. We found that 14% of the sample had to be discarded because they did not input any answer, or had zero correct answers, and did not find obvious response patterns compatible with cheating or strategic behavior (e.g., much better scores in the more difficult items).

The online test showed high reliability, and difficulty increased as a function of set size as expected, except for the last set size, maybe due to a ceiling effect for this sample. Compared to the same items tested in-person, in small groups, performance in the online and remote implementation was equivalent except for set size 4. Future research might explore whether this is a result of this particular study, or a phenomenon of online testing. Overall, this study shows the viability of using the Running Span in online and remote testing.

General discussion

This paper reported the implementation and results of two verbal WM tests for online and remote assessment in psychological and cognitive research. The Number-Letter Sequencing task was implemented in a supervised, online video conference setting (Backx et al., 2020; Crivelli et al., 2022; Leong et al., 2022), and the Running Span task, in an unsupervised asynchronous remote assessment (Hicks et al., 2016; Kulikowski & Potasz-Kulikowska; 2016; Ruiz et al., 2019).

For the online Number Letter Sequencing test we found that a sizable proportion of the sample had to be discarded initially, scores for set size 2 were higher than for set size 3, and there was not a pattern of growing difficulty according to set sizes. This latter goes against the set size effect, a benchmark of WM tasks (Oberauer et al., 2018) and so it is another evidence against the validity of the test. In addition, compared to the in-person similar implementation, scores were systematically different. In particular, scores for the more difficult higher set sizes, were significantly higher for the online test. This could be due to a strategy of writing down or copying the stimuli. In conclusion, these results do not support using this task and implementation, at least for low-stakes research purposes (ITC, 2022; Leong et al., 2022). The results are also in line with Steger et al. (2020) who concluded that cheating or strategic behavior was similar in high and low stakes online testing.

In addition, in online testing, given the lack of controlled conditions, participants' fatigue, boredom, or general lack of engagement in the task would also give rise to omissions or errors. Although the videoconference setting tried to emulate the controlled setting interpersonal factors, it could be the case that remote participants were less engaged than in-person. In conclusion, this type of implementation might require intensive technical and procedural guidance at the start, and more practice items; in addition, it would be advisable to test individually with camera on, or proctoring software. It should be noted that individual results could also be different an individual and auditory administration.

On the other hand, we found that the remote online Running Span task showed high reliability, an expected set size effect, and performance was generally equivalent to the in-person assessment. Thus, this study showed the viability of the Running Span task for research in relatively low-stakes assessment. Previous research had also validated testing with similar fast paced tasks, such as the nback (Kulikowski & Potasz-Kulikowska, 2016), and also with visuo-spatial stimuli (Hicks et al., 2016; Ruiz et al., 2019), given that both factors seem to prevent strategies to "game the test", and also could prevent boredom or disengagement. However, online research needs attention to possible constructirrelevant variance stemming from online implementation, such as those considered in these studies (e.g., test accessed but not attempted, minimum performance expected), and researchers would need to specify exclusion criteria related to these sources.

One of the limitations of the present research is that we compared new data with preexisting inperson data, as different from other studies (Backx et al., 2020; Hicks et al., 2016; Kulikowski & Potasz-Kulikowska; 2016; Ruiz et al., 2019). On one hand, our secondary data approach has allowed for larger samples than previous studies with direct comparisons; small previous samples are also a severe limitation. The sensitivity analyses in the Appendix show that these samples were enough to detect small effects of modality (0.10 out of 3 points in performance) and small to medium differences between consecutive set sizes (0.4 out of 3 points in performance) (see Appendix for rationale and simulations). In addition, the Appendix also pinpoints the need perform power simulations taking into account the full model to be explored, and not just paired comparisons or simple effects. Future research can implement an ideal design to directly compare modalities with random assignment to conditions in large samples; the Appendix provides a basis for power considerations.

Moreover, our results help to generalize psychological and neurocognitive research beyond the controlled lab environment in North America and Europe; future research could also take these generalization issues into account.

Acknowledgments

This research was supported by Secretaría de Ciencia y Técnica, Universidad de Buenos Aires grants UBACYT 20020190100077BA, and PIDAE 2020-3942.

Conflict of interest

The authors declare no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Disclaimer

Formatting, following the templates provided by Psicológica, spelling, grammar-checking and correct referencing are the sole responsibility of the authors.

Supplementary material

Data and analytical code for the study can be found at the Open Science Framework platform. Data study 1: https://osf.io/79gs5/, data study 2: https://osf.io/cdj4m/, code: https://osf.io/aetxc/

References

Alloway, T. P., Moulder, R., Horton, J. C., Leedy, A., Archibald, L. M. D., Burin, D., Injoque-Ricle, I., Passolunghi, M. C., & Dos Santos, F. H. (2017). Is it a small world after all? Investigating the theoretical structure of working memory cross-nationally. *Journal of Cognition and Culture, 17*(3-4), 331-353. https://doi.org/10.1163/15685373-12340010

Backx, R., Skirrow, C., Dente, P., Barnett, J. H., & Cormack, F. K. (2020). Comparing web-based and lab-based cognitive assessment using the cambridge neuropsychological test automated battery: a within-subjects counterbalanced study. *Journal of Medical Internet Research, 22*(8), e16792. https://doi.org/10.2196/16792

Baddeley, A. (2010). Working memory. *Current Biology, 20*(4), R136-R140. https://doi.org/10.1016/j.cub.2009.12.014

Baddeley, A. D., & Hitch, G. (1974). Working memory. In G.H. Bower (Ed.), *The psychology of learning and motivation: advances in research and theory* (pp. 47-89). Academic Press.

Baddeley, A., Hitch, G., & Allen, R. (2020). A multicomponent model of working memory. In R. Logie, V. Camos, & N. Cowan (Eds.), *Working memory: state of the science* (pp. 10-43). Oxford University Press. https://doi.org/10.1093/oso/9780198842286.003.0002

Barreyro, J.P.; Injoque-Ricle, R.; Formoso, J. y Burin, D. (2015a). Validez y confiabilidad de la prueba Running Memory Span. *Revista Argentina de Ciencias del Comportamiento, 7*(3), 26-331. https://revistas.unc.edu.ar/index.php/racc/article/view/11509

Barreyro, J. P.; Injoque-Ricle, I.; González, J. M.; Burin; D. I. (2015b). Estudio acerca de las propiedades psicométricas de pruebas clásicas de memoria de trabajo para tomas en grupos. *Anuario de Investigaciones, 22*(2), 283-288.

Barrouillet, P., & Camos, V. (2020). The time-based resource-sharing model of working memory. In R. Logie, V. Camos, & N. Cowan (Eds.), *Working memory: state of the science* (pp. 85-115). Oxford University Press. https://doi.org/10.1093/oso/9780198842286.003.0004

Bates, D. (2010). *lme4: Mixed-effects modeling with R*. https://lme4.r-forge.r-project.org/book/front.pdf

Brooks, M., Kristensen, K., van Benthem, K., Magnusson, A., Berg, C., Nielsen, A., Skaug, H., Maechler, M., & Bolker, B. (2017). glmmTMB balance speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal, 9*(2), 378-400. https://doi.org/10.3929/ethz-b-000240890.

Cowan, N. (2017). The many faces of working memory and short-term storage. *Psychonomic Bulletin & Review, 24*(4), 1158-1170. https://doi.org/10.3758/s13423-016-1191-6

Cowan, N., Morey, C. C., & Naveh-Benjamin, M. (2020). An embedded-processes approach to working memory: How is it distinct from other approaches, and to what ends? In R. Logie, V. Camos, & N. Cowan (Eds.), *Working memory: state of the science* (pp. 44-84). Oxford University Press. https://doi.org/10.1093/oso/9780198842286.003.0003

Crivelli, L., Quiroz, Y. T., Calandri, I. L., Martin, M. E., Velilla, L. M., Cusicanqui, M. I., Yglesias, F. C., Llibre-Rodríguez, J. J., Armele, M., Román, F., Barceló, E., Dechent, C., Carello, M. A., Olavarría, L., Yassuda, M. S., Custodio, N., Dansilio, S., Sosa, A. L., Acosta, D. M., … Allegri, R. F. (2022). Working group recommendations for the practice of teleneuropsychology in latin america. *Archives of Clinical Neuropsychology, 37*(3), 553-567. https://doi.org/10.1093/arclin/acab080

Do, B.-R. (2009). Research on unproctored internet testing. *Industrial and Organizational Psychology, 2*, 49–51. https://doi.org/10.1111/j.1754-9434.2008.01107.x

Emery, L., Myerson, J., & Hale, S. (2007). Age differences in item manipulation span: The case of letter-number sequencing*. Psychology and Aging, 22*(1), 75–83. https://doi.org/10.1037/0882- 7974.22.1.75

Feldman Barrett, L., Tugade, M. M., & Engle, R. W. (2004). Individual differences in working memory capacity and dual-process theories of the mind. *Psychological Bulletin, 130*(4), 553-573. https://doi.org/10.1037/0033-2909.130.4.553

Gosling, S. D., & Mason, W. (2015). Internet research in psychology. *Annual Review of Psychology, 66*, 877–902. https://doi.org/10.1146/annurev-psych-010814-015321

Hauser, D.J., Moss, A.J., Rosenzweig, C., Jaffe, S. N., Robinson, J. & Litman, L. (2022). Evaluating CloudResearch's Approved Group as a solution for problematic data quality on MTurk. *Behavior Research Methods* (online). https://doi.org/10.3758/s13428-022-01999-x

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not WEIRD. *Nature, 466*(7302), 29-29. https://doi.org/10.1038/466029a

Hicks, K. L., Foster, J. L., & Engle, R. W. (2016). Measuring working memory capacity on the web with the online working memory lab (the OWL). *Journal of Applied Research in Memory and Cognition, 5*(4), 478-489. https://doi.org/10.1016/j.jarmac.2016.07.010

International Test Commision (2022). *Guidelines for technology-based assessment (Draft).* https://www.intestcom.org/upload/media-library/tba-guidelines-3-14-2022-draftnumbered-1647343978NGDYR.pdf

Jaroslawska, A. J., & Rhodes, S. (2019). Adult age differences in the effects of processing on storage in working memory: A meta-analysis. *Psychology and Aging, 34*(4), 512–530. https://doi.org/10.1037/pag0000358

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. https://doi.org/10.1037/0096-3445.133.2.189

Kulikowski, K., & Potasz-Kulikowska, K. (2016). Can we measure working memory via the internet? the reliability and factorial validity of an online n-back task. *Polish Psychological Bulletin, 47*(1), 51-61. http://dx.doi.org/10.1515/ppb-2016-0006

Leidheiser, W., Branyon, J., Baldwin, N., Pak, R., & McLaughlin, A. (2015). Lessons learned in adapting a lab-based measure of working memory capacity for the web. Proceedings of the Human Factors and *Ergonomics Society Annual Meeting, 59*(1), 756-760. https://doi.org/10.1177/1541931215591235

Length, R. (2021). *emmeans: Estimated marginal means, aka least-squares means*. https://CRAN.Rproject.org/package=emmeans

Leong, V., Raheel, K., Sim, J. Y., Kacker, K., Karlaftis, V. M., Vassiliu, C., Kalaivanan, K., Chen, S. H. A., Robbins, T. W., Sahakian, B. J., & Kourtzi, Z. (2022). *A new remote guided method for supervised web-based cognitive testing to ensure high-quality data: development and usability study. Journal of Medical Internet Research, 24*(1), e28368. https://doi.org/10.2196/28368

Logie, R., Camos, V., & Cowan, N. (Eds.) (2020). *Working memory: the state of the science*. Oxford University Press. https://doi.org/10.1093/oso/9780198842286.001.0001

Logie, R. H. (2016). Retiring the central executive. *Quarterly Journal of Experimental Psychology, 69*(10), 2093-2109. https://doi.org/10.1080/17470218.2015.1136657

Lüdecke D (2023). *sjPlot: Data Visualization for Statistics in Social Science. R package version 2.8.15,* https://CRAN.R-project.org/package=sjPlot.

Mashburn, C. A., Tsukahara, J. S., & Engle, R. W. (2020). Individual differences in attention control: Implications for the relationship between working memory capacity and fluid intelligence. In R. Logie,

V. Camos, & N. Cowan (Eds.), *Working memory: state of the science* (pp. 175-211). Oxford University Press. https://doi.org/10.1093/oso/9780198842286.003.0007

Macnamara, B. N., & Conway, A. R. A. (2016). Working memory capacity as a predictor of simultaneous language interpreting performance. *Journal of Applied Research in Memory and Cognition, 5*(4), 434–444. https://doi.org/10.1016/j.jarmac.2015.12.001

Mielicki, M. K., Koppel, R. H., Valencia, G., & Wiley, J. (2018). Measuring working memory capacity with the letter–number sequencing task: Advantages of visual administration. *Applied Cognitive Psychology, 32*(6), 805–814. https://doi.org/10.1002/acp.3468

Miyake, A., & Shah, P. (Eds.) (1999a). *Models of working memory: mechanisms of active maintenance and executive control.* Cambridge University Press.

Miyake, A., & Shah, P. (1999b). Toward unified theories of working memory: emerging general consensus, unresolved theoretical issues, and future research directions. In A. Miyake & P. Shah (Eds.), *Models of working memory: mechanisms of active maintenance and executive control* (pp. 442–481). Cambridge University Press.

Oberauer, K. (2020). Towards a theory of working memory: from metaphors to mechanisms. In R. Logie, V. Camos, & N. Cowan (Eds.), *Working memory: state of the science* (pp. 116-149). Oxford University Press. https://doi.org/10.1093/oso/9780198842286.003.0005

Oberauer, K., Lewandowsky, S., Awh, E., Brown, G. D. A., Conway, A., Cowan, N., Donkin, C., Farrell, S., Hitch, G. J., Hurlstone, M. J., Ma, W. J., Morey, C. C., Nee, D. E., Schweppe, J., Vergauwe, E., & Ward, G. (2018). Benchmarks for models of short-term and working memory. *Psychological Bulletin, 144*(9), 885-958. https://doi.org/10.1037/bul0000153

R Core Team (2023). *R: A language and environment for statistical computing. R Foundation for Statistical Computing*. URL https://www.R-project.org/

Reips, U.-D. (2002). Standards for internet-based experimenting. *Experimental Psychology, 49*(4), 243-256. https://doi.org/10.1026//1618-3169.49.4.243

Revelle, W. (2021) *Psych: procedures for personality and psychological research*. Northwestern University, Evanston, Illinois, USA, https://CRAN.R-project.org/package=psych Version = 2.1.6

Rhodes, S., Doherty, J. M., Jaroslawska, A. J., Forsberg, A., Belletier, C., Naveh-Benjamin, M., Cowan, N., Barrouillet, P., Camos, V., & Logie, R. H. (2021). Exploring the influence of temporal factors on age differences in working memory dual task costs. *Psychology and Aging, 36*(2), 200-213. https://doi.org/10.1037/pag0000531

RStudio Team (2023). *RStudio: integrated development environment for R. RStudio*, URL http://www.rstudio.com/

Ruiz, S., Chen, X., Rebuschat, P., & Meurers, D. (2019) Measuring individual differences in cognitive abilities in the lab and on the web. *PLoS ONE 14*(12): e0226217. https://doi.org/10.1371/journal.pone.0226217

Shelton, J. T., Metzger, R. L., & Elliott, E. M. (2007). A group-administered lag task as a measure of working memory. *Behavior Research Methods, 39*(3), 482- 493. https://doi.org/10.3758/BF03193017

Steger, D., Schroeders, U., & Gnambs, T. (2020). A meta-analysis of test scores in proctored and unproctored ability assessments. *European Journal of Psychological Assessment, 36*(1), 174– 184. https://doi.org/10.1027/1015-5759/a000494

Stoet, G. (2010). Psytoolkit: A software package for programming psychological experiments using Linux. *Behavior Research Methods, 42*(4), 1096-1104. https://doi.org/10.3758/BRM.42.4.1096

Stoet, G. (2017). Psytoolkit: a novel web-based method for running online questionnaires and reactiontime experiments. *Teaching of Psychology, 44*(1), 24-31. https://doi.org/10.1177/0098628316677643

Webb, M. A., & Tangney, J. P. (2023). Too good to be true: bots and bad data from Mechanical Turk. *Perspectives on Psychological Science* (online first). https://doi.org/10.1177/17456916221120027

Wechsler, D. (2003). *WAIS III: test de inteligencia para adultos*. Paidós.

Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019). Welcome to the tidyverse. *Journal of Open Source Software, 4*(43), 1686. https://doi.org/10.21105/joss.01686

Appendix

Sensitivity Analyses

Sensitivity analyses are recommended because the studies involved reconsideration of secondary data. A sensitivity power analysis answers the question which effect size a study could detect given a pre-determined sample size, power, and alpha level (Lakens, 2022). One of the answers that a sensitivity analysis can provide is whether the study has sufficient power for effects considered "plausible and interesting", such as the smallest effect size of interest, or the effect size that is expected (Lakens, 2022). This is generally accomplished by simulations of which sample sizes would be necessary to find a desired effect, or range of effects, with appropriate statistical power.

Power calculations usually involve parameters unrelated to the hypotheses, such as several variances and when there are multiple factors, their correlations. A sensitivity analysis for the effects of modality and set size presented in this article must keep the mixed linear model approach employed in the studies, which allowed for initial variations within participants with a random intercept per subject, and reflect the underlying parameters' structure and covariance. Indeed, another way of posing the sensitivity analysis question is: if this study were to be replicated, which sample sizes would be necessary to find desired effects, or range of effects, with appropriate statistical power? Classical approaches to power analysis typically work with analytical formulas that cannot be applied to solve power estimations for linear mixed models. It would not be reasonable to simulate simpler structures such as paired comparisons, ANOVAS, or linear regressions.

The simulations need to be based on linear mixed models. The simr package v. 1.0.7 is a power analysis package for R designed to work with the lme4 package (Green & McLeod, 2016). A power analysis in simr starts with a model fitted in lme4 and allows for explorations of the power space given different fixed and random effects and sample sizes. Power is calculated by simulating new values for the response variable using the model provided, refitting the model to the simulated response, and applying a statistical test to the simulated fit. The power calculations are based on Monte Carlo simulations.

To specify a desired effect (Lakens, 2022) within generalized linear mixed models' simulations, the target measure is the estimated model coefficients (Green & McLeod, 2016). In our case, fixed effects coefficients for modality and set size, for Letter Number Sequencing and Running Span, with modality and set size as fixed factors and participants as random factor.

The simulations parameter that we must test for is the fixed effects coefficients. Their value estimates the change in performance with respect to the other category (for modality), or for an increase in set size, contingent on other parameters of the model. There is no universal rule to what constitutes a small, or big, or interesting, regression coefficient; it all depends on context (Miller, 2008). Therefore, we considered the estimated and standardized coefficients in the observed models to determine the minimum effect of interest, e.g. the values for the estimated modality and set size coefficients to adopt in the simulations. After this step, the sensitivity analyses involved running the simulations and obtaining power curves to explore trade-offs between sample size and power for fixed parameters under the model.

Results

Letter Number Sequencing

We run simulations for a linear mixed model (lmer) with modality and set size as fixed factors, a random intercept for participants, and performance as dependent variable. To explore power as a function of sample size for a small effect, we set the modality coefficient to 0.1, a mean difference of 0.1 points (over 3) in performance. Figure A1 shows the power curve for the simulated model at different sample sizes (500 simulations). The power simulations were estimated around 0 % with 75 participants or less, to around 80% with 100 participants and 100 % with 110 participants or more.

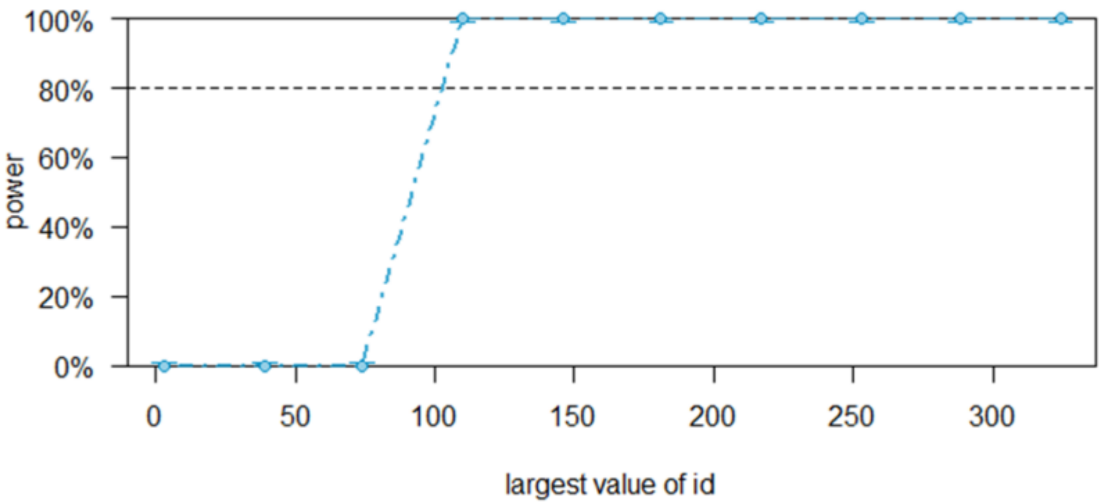

Figure A1. *Power (±95% CI) to Detect a Modality Coefficient = 0.1, Simulated over a Range of Sample Sizes, for Letter Number Sequencing.*

For the set size effects, observing the pattern of differences between set sizes, the last ones have approximately 0.4 points difference. This could be a minimum difference of interest, corresponding approximately to 0.3 standardized points. Therefore we set a pattern of fixed set size coefficients decreasing 0.4 points for each set size. Figure A2 shows the power curve for the simulated model at different sample sizes (500 simulations). The power was around 28%(CI 95 = 24.68 %, 32.78 %) with 350 participants.

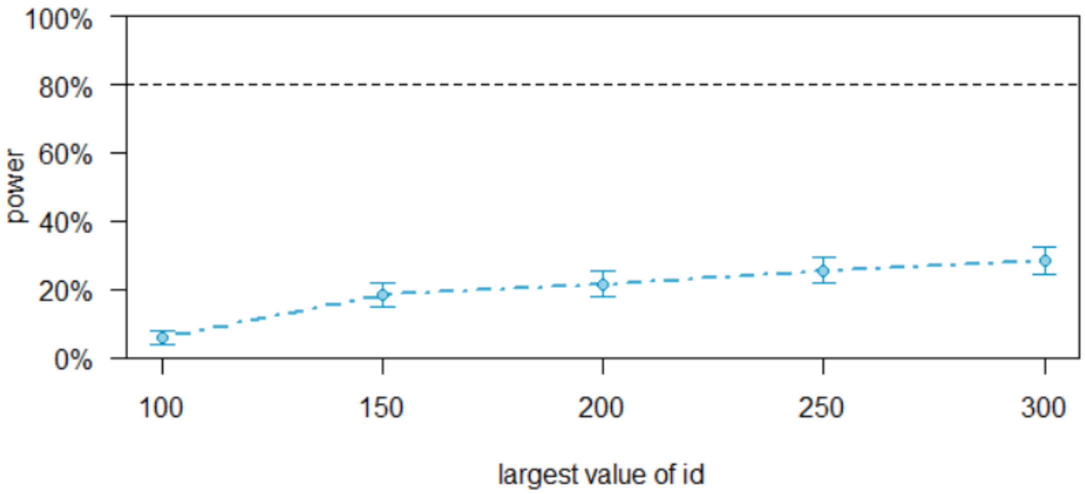

Figure A2. *Power (±95% CI) to Detect Set Size Coefficients Decreasing by 0.4 points and Modality Coefficient = 0.1, Simulated over a Range of Sample Sizes, for Letter Number Sequencing.*

However, this assumes a very small effect of modality. Power raises if the modality coefficient is assumed to be 0.2, with all other coefficients decreasing 0.4 points for each set size. Figure A3 shows the power curve for the simulated model at different sample sizes (500 simulations). The power was around 80% (CI 95 = 75.80 %, 83.05 %) with 300 participants.

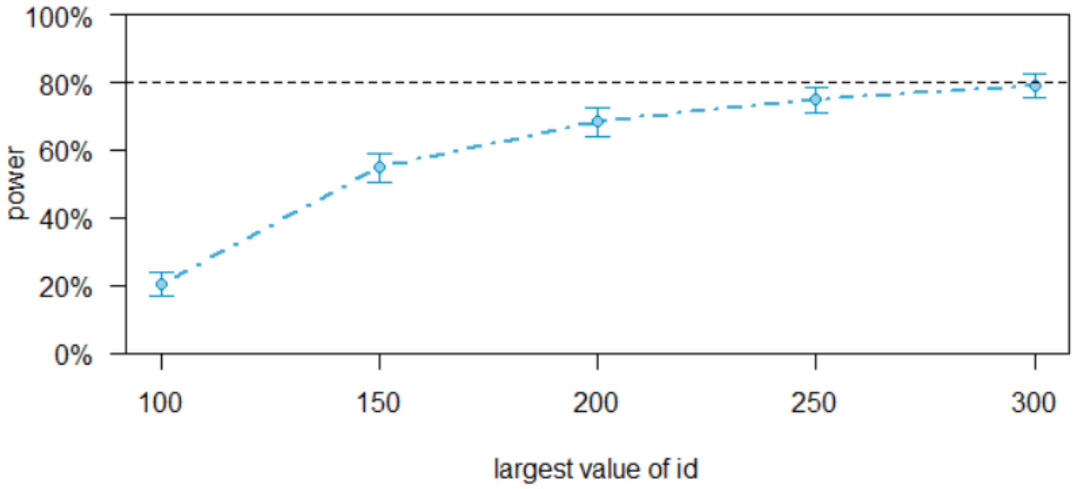

Figure A3. *Power (±95% CI) to Detect Set Size Coefficients Decreasing by 0.4 points and Modality Coefficient = 0.2, Simulated over a Range of Sample Sizes, for Letter Number Sequencing.*

Running span

Again, we run simulations with a small difference between in-person and online, setting the modality coefficient to 0.1 (which is smaller than the obtained estimated coefficient, 0.165). Figure A4 shows the power curve for the simulated model at different sample sizes (500 simulations). The power simulations were estimated around 0 % with 75 participants or less, to around 80% with 110 participants, and 100% power with 120 participants.

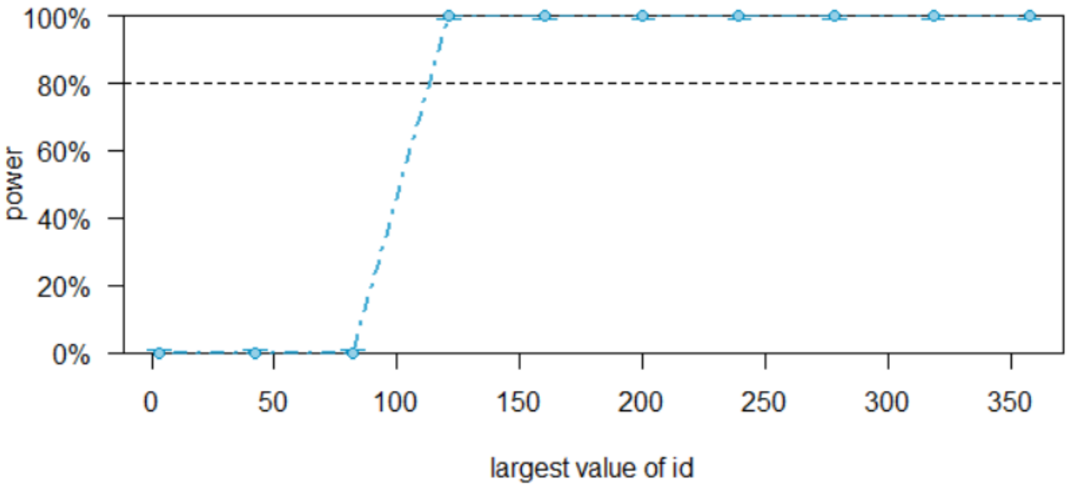

Figure A4. *Power (±95% CI) to Detect a Modality Coefficient = 0.1, Simulated over a Range of Sample Sizes, for Running Span.*

For set size, again, we set a 0.4 points difference in performance (out of 3 points) for each consecutive set size (which is smaller than the observed coefficients, with WM2 as baseline: setsizeWM3 = 0.75, setsizeWM4 = -1.96, setsizeWM5 = -2.45, and setsizeWM6 = -2.61). Figure A5 shows the power simulations across different sample sizes. Power was low (CI 95 = 31.79 %, 40.38 %) even with 350 participants.

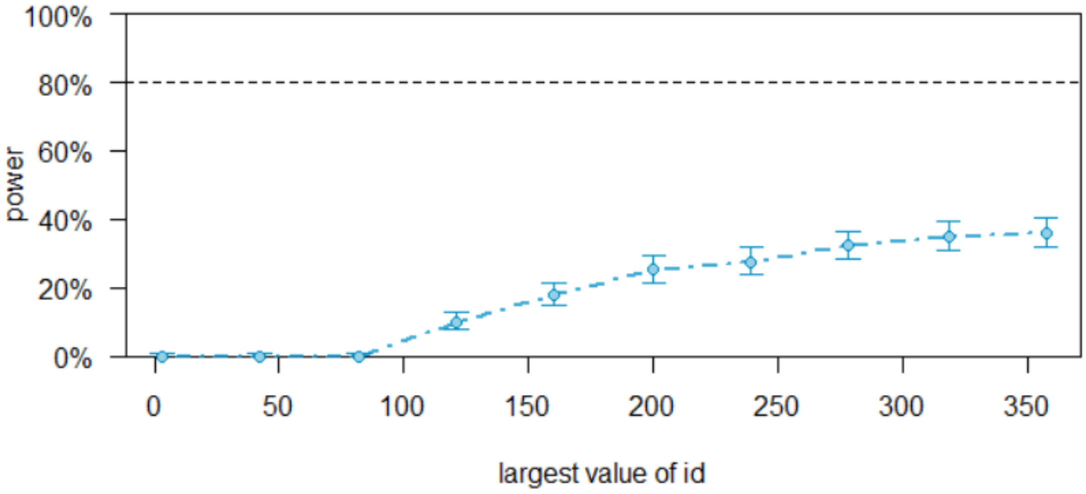

Figure A5. *Power (±95% CI) to Detect Set Size Coefficients Decreasing by 0.4 points and a Modality Coefficient = 0.1, Simulated over a Range of Sample Sizes, for Running Span.*

Again, we explored whether power would increase with a slightly bigger modality coefficient, 0.2. This raised power to 80% (CI 95 = 76.43 %, 83.61 %) with 250 participants, and 88% (CI 95 = 84.17 %, 90.18 %) with 350 participants, as shown in Figure A6.

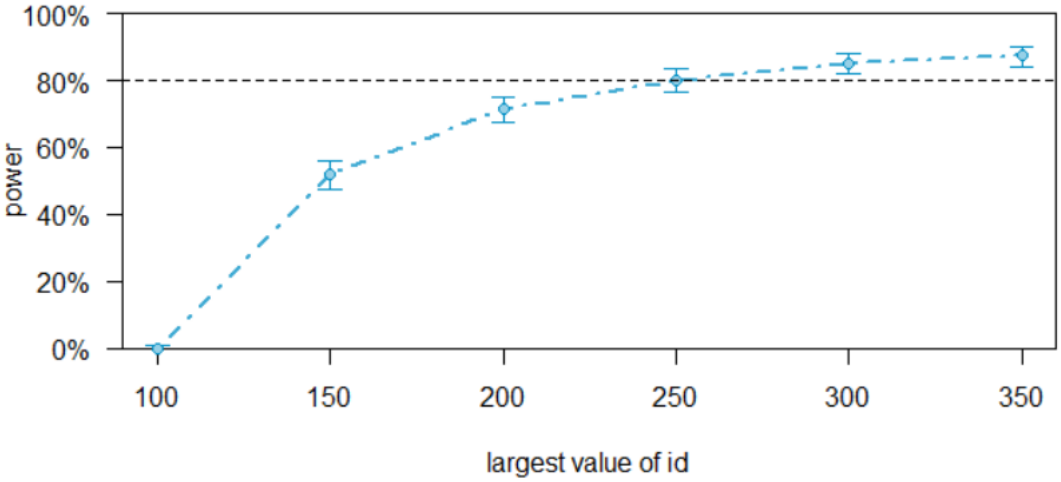

Figure A6. *Power (±95% CI) to Detect Set Size Coefficients Decreasing by 0.4 points and a Modality Coefficient = 0.2, Simulated over a Range of Sample Sizes, for Running Span*.

Also, the Running Span seems to have a steeper performance slope decline than the previous task, so we also run simulations increasing the set size coefficients' difference in 0.5 points. Figure A7 shows these simulations; they did not change much previous estimations. Power for a sample of 250 participants was estimated at 81 % (CI 95 = 77.71%, 84.72%) and for 350 participants, 88% (CI 95 = 85.26 %, 91.07%).

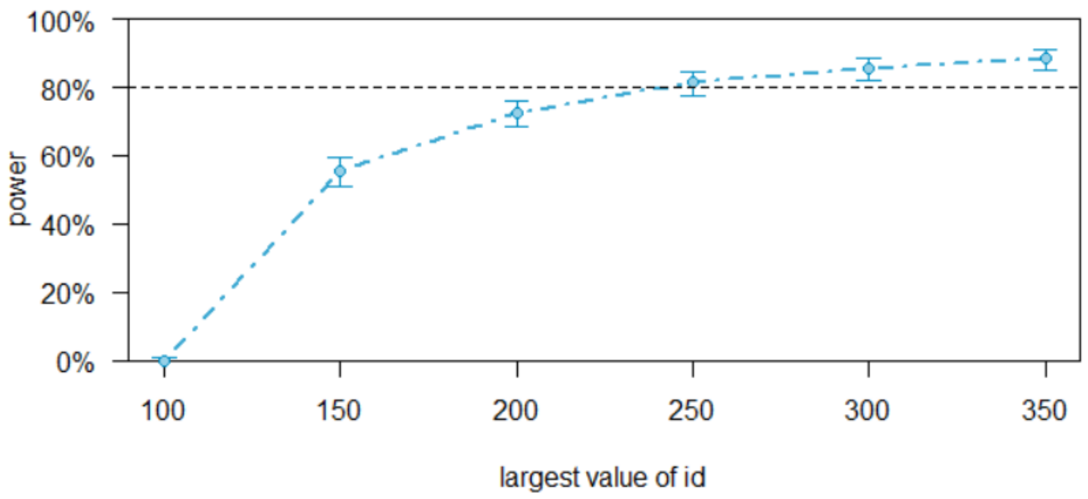

Figure A7. *Power (±95% CI) to Detect Set Size Coefficients Decreasing by 0.4 points and a Modality Coefficient = 0.2, Simulated over a Range of Sample Sizes, for Running Span*.

In synthesis, these simulations explored power curves for sample sizes around the smaller effects of interest, assuming a linear mixed model with modality and set size as fixed factors, a random intercept for participants, and performance as dependent factor. The underlying parameters' structure and values was estimated with the observed data and analyses.

Letter Number Sequencing task. With the modality coefficient set as 0.1 at least 100 participants are needed for 80% power, and with 110 participants power is around 100%. For the set size effect, setting a pattern of fixed set size coefficients decreasing 0.4 points for each set size, power was low (CI 95 = 24.68 %, 32.78 %) with 350 participants. Changing the modality coefficient in the same model with a slight increase, 0.2, and the aforementioned pattern of set size coefficients, raised power to 80% (CI 95 = 75.80 %, 83.05 %) with 300 participants.

Running Span task. With the modality coefficient set as 0.1 at least 110 participants are needed for 80% power, and with 120 participants power is around 100%. For the set size effect, setting a pattern of fixed set size coefficients decreasing 0.4 points for each set size, power was low (CI 95 = 31.79 %, 40.38 %) with 350 participants. Changing the modality coefficient in the same model with a slight increase, 0.2, and the same previous pattern of set size coefficients, raised power to 80% (CI 95 = 76.43 %, 83.61 %) with 250 participants, and 88% (CI 95 = 84.17 %, 90.18 %) with 350 participants.

Overall, these sensitivity analyses show that the samples in the present studies, Letter Number Sequencing N = 353 and Running Span N = 395, were adequate for finding the smaller effects of interest within the linear mixed model analytic approach.

In addition, these analyses show the need to perform power simulations considering the full model to be explored, and not just paired comparisons or simple effects

Appendix references

Lakens, D. (2022). *Improving Your Statistical Inferences*. https://lakens.github.io/statistical_inferences/. https://doi.org/10.5281/zenodo.6409077

Green, P. & MacLeod, C.J. (2016), SIMR: an R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution, 7*, 493- 498. https://doi.org/10.1111/2041-210X.12504

Miller, J.E. (2008). Interpreting the substantive significance of multivariable regression coefficients. *Proceedings of the American Statistical Association, Statistical Education Section*. http://www.statlit.org/pdf/2008MillerASA.pdf