Working memory tasks in interpreting studies - A meta-analysis

(Article begins on next page)

04 May 2024

Working memory tasks in interpreting studies

A meta-analysis

Serena Ghiselli
University of Bologna

Studies about working memory (WM) and interpreting have used a variety of methods and results are often conflicting. There is therefore the need to analyse the cognitive tasks which have been used so far to assess their effectiveness in detecting WM performance differences. This paper presents the findings of a meta-analysis that compares the results of interpreters and interpreting students (study group) to the results of non-interpreters (control group) in four cognitive tasks (reading span, n-back task, listening span and dual tasks). Interpreters show a significant WM advantage of medium size over non-interpreters in tasks based on verbal stimuli, but not in tasks based on non-verbal stimuli. In addition, differences are larger when there is a wider gap in interpreting expertise between the two groups.

Keywords: working memory, interpreting, cognitive tasks, methodology, literature review, meta-analysis

1. Introduction

Interpreting is a complex cognitive task, for which an efficient working memory (WM) is of paramount importance (Gile 2009; Liu et al. 2004; Moser-Mercer 2000). Many scholars investigated whether interpreting practice fosters WM efficiency (Babcock & Vallesi 2017; Morales et al. 2015; Nour et al. 2020a) and how WM changes in interpreting students during training (Babcock et al. 2017; Chmiel 2018; Dong et al. 2018). Various validated tests to assess WM have been used in the literature, but none of them was developed specifically for interpreters and results collected so far are mixed, which is probably due to the diverse research designs adopted (Dong & Cai 2015).

Systematic reviews are a way to find patterns in the data by looking at the big picture. A systematic review is defined as a review that uses explicit, systematic

methods to collate and synthesise findings of studies that address a clearly formulated question (Cochrane Handbook for Systematic Reviews of Interventions: Version 6.2 2021). Meta-analysis is a statistical technique used to synthesise results when study effect estimates and their variances are available, yielding a quantitative summary of results (McKenzie & Brennan 2019).

Research synthesis is a scientific process starting with the definition of the variables and relationships of interest of the research. Then evidence is collected and qualitatively evaluated through inclusion and exclusion criteria. Afterwards, the collected results are combined and analysed to detect differences and similarities, in order to interpret the cumulative evidence and highlight strengths and limitations (Cooper et al. 2009).

The study presented in this paper uses meta-analysis to address a very critical problem of studies on interpreting and WM, that is the reliability of the tasks which have been used so far to detect differences between study and control groups. Even if the published studies are relatively few, for a very specialised field such as cognitive interpreting studies evidence is enough to look at the results and reflect on best practices for the benefit of future research quality and effectiveness.

2.   Background and motivation

In cognitive interpreting studies literature two meta-analyses were published (Mellinger & Hanson 2019; Wen & Dong 2019) and they have the merit of having introduced this methodology into the field and of having highlighted interesting patterns for future developments. Mellinger and Hanson (2019) investigated the link between WM, short-term memory (STM) and simultaneous interpreting expertise. Data analysis showed that simultaneous interpreters performed better than non-interpreters both on tests of STM and on tests of WM. Differently from expectations, interpreters outperformed the various comparison groups on both auditory and visual WM tasks. The issue of causality remained, however, unresolved since research findings could not determine whether education, training and the practice of simultaneous interpreting produced stronger WM or whether individuals with stronger WM were those who were more likely to be successful in interpreting. Wen and Dong (2019) found the same correlation between the level of expertise and WM and determined that the level of interpreter expertise significantly moderated the interpreter advantage, which existed only for intermediate-level interpreters (with at least two year's training) and experienced interpreters (with a minimum of four years' professional experience) and not for beginners (with one year's training or less).

Both meta-analyses converge on the confirmation of the interpreter advantage, but many studies taken individually fail to find this advantage. The question arises whether the upstream problem is that the methodology used is not always appropriate to address the research questions on the cognitive interpreter advantage. Against this backdrop, the present study uses meta-analytical methods to focus on the WM tasks used so far in interpreting studies literature and to test to what extent each task is able to detect a performance difference between study and control groups. The goal is formulating evidence-based suggestions on the tasks that proved to be more useful to study WM in interpreters to inform future research.

The first hypothesis is that verbal tasks results should differ more between interpreters and non-interpreters than non-verbal ones since they are influenced by language skills. Interpreting scholars argued that tasks in L1 tend to be easier than tasks in L2 (Christoffels et al. 2006; Chincotta & Underwood 1998; Tzou et al. 2012; Cai et al. 2015) and the added difficulty of L2 is reduced in the interpreter group (Christoffels et al. 2006). As regards the task language, a second hypothesis is that tasks in L2, being more demanding, should be better at detecting differences between participants with various levels of expertise in interpreting than tasks in L1 and that difference in performance should increase as the difference in interpreting expertise increases. A third hypothesis is that tasks based on auditory stimuli, being more similar to the input that interpreters have to process on the job, should detect more differences between study and control groups than tasks based on visual stimuli.

3. Methods

The literature search to gather sources for the meta-analysis was carried out on the basis of the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) guidelines (Page et al. 2021). PRISMA is an evidence-based checklist of items to include in the report of systematic reviews and meta-analyses. The PRISMA statement provides a flow diagram template and guidelines for authors on how to prepare a systematic review which is valuable to users, meaning that it is complete, transparent, and based on clearly pre-defined inclusion and exclusion criteria for the identification and selection of studies. The tools and the keywords used for the search have to be specified as well as the methods used to assess the risk of bias in the included studies.

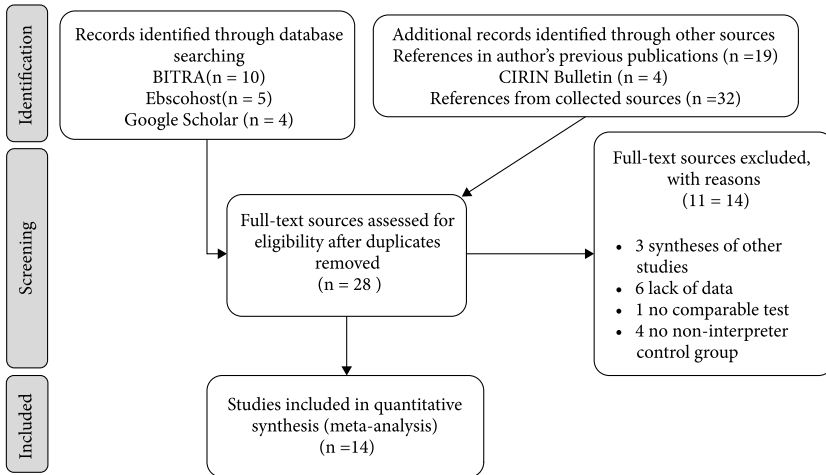## 3.1 Literature search and selection criteria

Figure 1. PRISMA flow diagram of studies including WM tasks comparing interpreters and interpreting students to non-interpreters

The literature search was carried out between September 1st 2020 and February 28th 2021 and it is schematised in Figure 1. The search stopped when results started to be repetitive and nothing new could be found. The keywords used for database searching were "psychometr* AND interpret*" and "working memory" AND interpret*". The keywords aimed at being as inclusive as possible, hence the use of the asterisk to retrieve "psychometrics, psychometric" and "interpreting, interpreter, interpreters". The sources which were taken into consideration were articles, books and PhD theses in English. An eligible source had to include at least a study group formed of interpreters or interpreting students and at least a control group of non-interpreters. The source needed to use a WM task comparing the study and the control groups and enough data to calculate the effect size Hedges' *g* (see § 3.5). The definition of WM task is not homogeneous in the articles collected, in the sense that some authors consider as WM tasks also tasks where only information maintenance (and not processing) is required (Yudes et al. 2011; Stavrakaki et al. 2012; Tzou et al. 2012; Dong et al. 2018). In this study a WM task is defined as a task where both maintenance and processing of information are required in the support of cognition and action (Logie et al. 2021). In order to be included in the meta-analysis, a task should also appear in the literature with an identical or similar procedure at least twice, to allow results comparison.

The database Bibliography of Interpreting and Translation (BITRA) (Franco Aixelá 2001) was used to search for the sources, without any time limit. Since it

is a very sector-specific resource, any result could be relevant. Other databases were also searched (see the list below). The time span for the search in the other databases was 2019–2021, in order to reduce the high number of irrelevant results which would otherwise be retrieved. The goal was to make sure to find recent publications which had not been indexed in BITRA yet. The databases searched were accessible through the online library service of the University of Bologna (AlmaStart). The following databases were used, in addition to BITRA, for data collection:

– GOOGLE SCHOLAR
– The research databases provider EBSCOHOST, inside which the following databases were selected:
  – APA Psycharticles,
  – APA PsycInfo
  – Catalogo Italiano dei periodi (ACNP)
  – eBook Academic Collection (Ebscohost)
  – eBook collection (Ebscohost)
  – Education source
  – Education Resources Information Center (ERIC)
  – Modern Language Association (MLA) Directory of Periodicals
  – Modern Language Association (MLA) International Bibliography with Full text
  – OpenDissertations
  – Psychology and Behavioural Sciences Collection

In addition to databases, the CIRIN BULLETIN (issues from January 2019 to January 2021) were screened (Gile 2021). CIRIN is an international information network on research into conference interpreting set up in Paris in 1990 with the purpose of facilitating the circulation of conference interpreting research. The goal was again to check for more recent publications which were not indexed in BITRA. Relevant references cited in the author's PhD thesis and previous publications on related topics were read again and assessed for eligibility. Finally, the references in the collected sources were screened to look for additional studies.

As can be seen from the PRISMA flow diagram (Figure 1) the results found using different methods were often the same. The 74 sources which were collected went down to 28 after duplicates were removed. These 28 sources were read full-text and assessed for eligibility. Half of them were excluded and 14 studies were included in the quantitative synthesis.

The reasons for excluding sources were the following:

- – 3 articles were excluded because they were not presenting new data but synthesising previous results. The articles are Mellinger & Hanson (2019), Nour et al. (2020b) and Wen & Dong (2019). These articles presented research syntheses on related topics and their references were useful to look for further studies.
- – 6 articles had to be excluded because the quantitative data they contained were not sufficient to calculate the effect size Hedges' *g*. The authors of all the sources that were potentially relevant but had insufficient data were contacted via email and asked for the missing information needed to include them in the meta-analysis. For these articles it was impossible to retrieve the data since the authors either did not answer or could no longer access the data of their study.
- – 1 PhD thesis (Jin 2010) had to be excluded because WM was assessed using a recall exercise, which has not been used in other research papers including a study and a control group, so it was not possible to compare results.
- – 4 sources had to be excluded because they did not include any control groups (Injoque-Ricle et al. 2015; Liu et al. 2004; Timarová et al. 2014, 2015).

At the end of the screening process, 14 articles were selected to carry out the meta-analysis. A table with the list of the data included is provided in Appendix A. Sources are ordered by task type because the focus of the present research is not to compare one study to another but one specific task in a study to similar tasks in other studies to see if interpreters and non-interpreters perform differently or not in that task. The table in Appendix A includes information about the task (type and channel of input and language if relevant), the participants (study and control groups), the publication (author and year), and whether an interpreter advantage was found. The presence of an interpreter advantage refers to the findings of the original study, not to the meta-analysis. The rationale adopted to categorise tasks is explained in the next section.

3.2 Task classification

The WM tasks which were selected at the end of the literature review process were categorised into four groups according to the type of activities they involved:

1. Reading span task – seven studies include reading span tasks. A reading span task is a task in which the participant has to read a set of sentences and keep in mind the last word of each sentence, in order to recall it at the end of the set. There is a certain degree of variation among the reading span tasks analysed, but the goal of the task remains unchanged and is measuring verbal WM. In Chmiel (2018) the task was remembering sequences of letters while

judging the sense of sentences (half of them made sense and half did not). At the end of the set participants had to recall the letters by selecting them from the computer screen in the same order they were presented. In some studies participants had to recall the sentence final words orally (Nour et al. 2020a; Signorelli et al. 2011; Ünlü & Şimşek 2018), whereas in others they had to write them down (Tzou et al. 2012; Yudes et al. 2011; Yudes et al. 2013). In some tasks the words had to be recalled in the same order (Tzou et al. 2012; Ünlü & Şimşek 2018; Yudes et al. 2011, 2013) and in others in any order (Nour et al. 2020a; Signorelli et al. 2011).

2. Listening span task – four studies include a listening span task. The listening span task is similar to the reading span task, but in this case sentences are presented orally. Participants had to listen to a set of sentences and remember the sentence's final word while either judging sentence sense (Chmiel 2018; Dong et al. 2018), repeating the sentences aloud (Köpke & Nespoulous, 2006), or judging whether sentences were true or false (Stavrakaki et al. 2012). The words had to be recalled either in the same order (Köpke & Nespoulous, 2006; Stavrakaki et al., 2012) or in any order (Chmiel 2018; Dong et al. 2018).

3. N-back task – four studies include n-back tasks. In the n-back task the participant has to keep in mind a series of items presented visually and react when the same item presented a certain number of steps before is shown again. In three studies the stimuli were images, either blue squares (Dong & Liu 2016; Dong et al. 2018) or black and white drawings of daily objects (Rosiers et al. 2019). One task was based on letter stimuli instead (Van der Linden et al. 2018).

4. Dual tasks – three studies include tasks that in this paper are labelled "Dual tasks" because they involve the performance of two activities in sequence. Dual tasks do not refer to tasks where a participant is required to perform two actions at the same time or to ignore a concurrent stimulus, which would measure different processes (split attention and inhibition respectively), but they refer to tasks where participants have to memorise items and then perform another task immediately afterwards. The four dual tasks identified are category and rhyme probe task (Köpke & Nespoulous 2006), automated operation span task (OSPAN) (Babcock et al. 2017), automated symmetry span task (Babcock et al. 2017) and phonological cued recall (Signorelli et al. 2011). In the category and rhyme probe task participants had to memorise lists of words followed by a probe. The probe could be either a word rhyming with one of the words from the list (phonological condition) or belonging to the same semantic category as one of the words from the list (semantic condition). In the OSPAN task participants were asked to recall sequences of consonants. In addition, prior to each item of the sequence, they had to perform

an arithmetic operation. The automated symmetry span task is the same as OSPAN but, instead of performing arithmetic operations, participants were asked to give a symmetry judgement on a drawing. Finally, the phonological cued recall is a task where participants saw a list of words presented one at a time and then they saw a cue word and had to recall the word that preceded the cue in the list.

In addition to categorising the tasks into four groups according to the type of activities they involved, each task was also assigned to a type and a channel of input. The type of input refers to the fact that stimuli are verbal or non-verbal, the channel of input can be either the visual or the auditory channel.

3.3 Groups: Interpreters, interpreting students and non-interpreters

In the studies included in the meta-analysis, the number of experienced interpreters per group ($M = 16.40$, $SD = 5.13$) is lower than the number of interpreting students ($M = 25.45$, $SD = 14.19$). This is not surprising, since it is easier to recruit students attending a Master's degree than experienced interpreters, who are often travelling and do not have a lot of time to take part in an experimental study. The mean age of experienced interpreters is 44.51 years ($SD = 8.40$) and the mean age of interpreting students is 22.27 ($SD = 2.06$). This age gap is again predictable, since time is needed to build interpreting experience, which is large in the sample of interpreters ($M = 15.97$ years, $SD = 6.54$). Interpreting students have on average of 14.94 months of interpreting training ($SD = 7.99$ months). In longitudinal studies, where interpreting students were tested before and after training, only results of tests taken after training were included in the meta-analysis.

Control groups are formed of non-interpreters. In the studies included in this meta-analysis there are six types of control groups: monolinguals (3), non-language students (7), multilinguals (6), English language students (5), foreign-language teachers (2) and translation students (5). Control groups have on average 22.54 participants ($SD = 9.69$), with a mean age of 25.46 ($SD = 11.39$). In the meta-analysis every study group was compared separately with every control group. Therefore, there are multiple effect sizes calculated from the same study and the same groups of participants. This situation creates a problem of interdependency of control the data that is addressed by means of three-level random effects models (see § 3.5).

3.4 Variables

The goal of this meta-analysis is to see which tasks are best suited to detect differences between the study group (interpreters or interpreting students) and the control group (non-interpreters).

A first set of analyses was carried out separately for every one of the four task categories (reading span task, listening span task, n-back task and dual tasks). In these analyses the independent variable was the group to which the participants belonged (either study or control) and the dependent variable was the effect size Hedges' g, that is the size of the difference in task performance between the study and the control group. There were three possible group comparisons: interpreters vs non-interpreters, interpreters vs interpreting students and interpreting students vs non-interpreters. Both interpreters and interpreting students were considered as a study group but their results were analysed separately and also compared one to the other (interpreters as study group and interpreting student as control group) to see whether differences in the level of expertise had an impact on task performance. The raw data to calculate every single Hedges' g effect belonged to the same scientific paper, so that the groups of participants had carried out exactly the same tasks. Not all group comparisons were possible for all the tasks because of lack of enough data to perform them.

When relevant and statistically preferable (see § 3.5), the following moderating dummy variables were added to the analysis: type of input, channel of input and language. The type of input refers to whether the content of the task is a verbal stimulus or a non-verbal stimulus. The channel is how stimuli are physically perceived, that is whether they are visual or auditory. The moderating variable of language is applicable only to verbal tasks and can be either the mother tongue (L1) or a foreign language (L2) for the participants. In the data collected non-verbal tasks were always based on visual stimuli, whereas verbal ones could have either visual or auditory stimuli.

A second set of analyses considered the results of all WM tests divided by study and control group. Under the study group label both interpreters and interpreting students were considered. The WM test results were compared adding type of input, channel of input and expertise difference as moderating variables in three separate models (see § 4.5).

The task measures on which Hedges' g was calculated were accuracy measures, i.e. the number of items recalled and, for the n-back task, also the response reaction time (RT).

3.5   Statistical methods

In order to compare the performance of study-control group pairs of different studies, a standardised measure is needed. In the present meta-analysis, the effect size Hedges' *g* was chosen. An effect size is defined as a quantitative reflection of the magnitude of some phenomenon that is used for the purpose of addressing a question of interest (Kelley & Preacher 2012, 140).

Hedges' g (Hedges 1981) was computed as a standardised measure of mean difference between the results of the two groups (study and control). It is the most widely used effect size in meta-analysis when you consider group differences (e.g., Mellinger & Hanson 2019; Wen & Dong 2019), it is similar to Cohen's d but it better controls the overestimation of the effect in studies with small sample sizes (< 20), which is the case for interpreting studies. An effect of 0.2 is considered to be small, 0.5 is a medium effect and from 0.8 upwards the effect is considered to be large (Cohen 1977). The effect sizes were not included in the primary studies and they were calculated using RStudio statistical software (RStudio 2021) and the contributed package metafor (Viechtbauer 2010), using group number, mean test score and standard deviation (or, alternatively, standard error). Metafor was also used to carry out all the statistical analyses and to create the plots.

The effect sizes Hedges' *g* were compared using three-level random-effects models. Random-effects models, differently from fixed-effects models, do not show Type I bias in significance tests for mean effect sizes and for moderator variables. In addition, fixed-effects models produce confidence intervals for mean effect sizes that tend to be narrower than their actual width, thereby overstating the degree of precision of the meta-analysis (Hunter and Schmidt 2000). Schmidt et al. (2009) showed that the precision of meta-analysis findings may have been systematically overstated in much of the research literature and recommended the use of random-effects models. The choice of multilevel models in particular was made to include sources of variability between studies. As mentioned in the introduction, the research designs adopted so far in interpreting studies have been very diversified. Moreover, in most studies there were multiple study and control groups who took the same tests. The groups of the same study were compared separately to calculate more reliable effects. On the one hand, distinguishing between the different levels of interpreting expertise (professionals vs students) was relevant for the research hypotheses (see § 2); on the other hand, research group participants came from a variety of backgrounds (see § 3.3), therefore calculating a separate effect for every group pair was considered to be a more precise representation of the data. The choice of calculating multiple effects from the same study creates, however, an issue of interdependency of the data (Cheung 2014). To take into account between-study differences and the interdependency of the

effects derived from the same groups of participants, the statistical procedure of three-level random-effects modeling was chosen, following the example of previous studies (Assink et al. 2015; Assink & Wibbelink 2016; Wen & Dong 2019). The models which were calculated include three levels: Level 1 corresponds to the sampling variance of the observed effect sizes, Level 2 to the variance between effect sizes from the same study and Level 3 to the variance between studies. Restricted maximum-likelihood was used in estimating the three-level random-effects models (Viechtbauer 2005).

For completeness, two separate log-likelihood-ratio tests were performed for every model to determine whether the within-study variance (Level 2) and between-study variance (Level 3) were significant. The reported value is the Bayesian information criterion (BIC) (Raftery 1995), that is the extent to which the hypothesized model is consistent with the data (Diamantopoulos & Siguaw 2000). An increase or decrease of more than 2 units in BIC is indicative of a non-trivial worsening or improving in model fit, respectively (Raftery 1995). However, the three-level model was reported independently from the log-likelihood-ratio test results because it represented the data-generating process better. According to Harrer et al. (2021), when data contain studies with multiple effect sizes, as is the case in this paper, effect sizes cannot be independent. Therefore, it is methodologically safer to include Level 2 and Level 3 in the models.

Heterogeneity was calculated using the formula of Cheung (2014: 215) and assessed according to the 75% rule as described by Hunter and Schmidt (1990). According to this rule, heterogeneity is considered to be substantial if less than 75% of the total amount of variance can be attributed to sampling variance (at Level 1). In case of substantial heterogeneity, and when potential moderating variables existed, a three-level model including moderators was calculated. In some models three moderating dummy variables could potentially be added: type of input and channel of input (for dual tasks), language (both for dual tasks and for the reading span task, see § 3.4) and interpreting expertise difference between study and control group (for the analysis on all WM tasks).

Model results are displayed visually using forest plots, which include the weight of every study in the model, the effect size and the confidence interval, with a bottom row providing an estimate of the overall effect size and its confidence interval. Forest plots also contain the effect size data that were used to perform the meta-analysis on the right of the plot. For each study, the effect size is represented with the point estimate on the *x*-axis supplemented by a line, which indicates the range of the confidence interval, and surrounded by a box (a larger box corresponds to a larger weight). At the bottom of the plot, a diamond shape represents the average effect. The length of the diamond symbolizes the confidence interval of the pooled result on the *x*-axis. The vertical reference line indicates the point on the *x*-axis equal to no effect (Harrer et al. 2021).

Finally, the file-drawer or publication bias problem was taken into account to check for the reliability of the model results. This problem states that a study with high effect sizes is more likely to be published than a study with low effect sizes (Rothstein et al. 2005) and this leads to publication bias, as the pooled effect estimated in a meta-analysis might be higher than the true effect size because there are missing studies with lower effects which were not considered due to the simple fact that they were never published. There are several methods to check for publication bias in a meta-analysis and more than one should be used for robustness, but most of them are not reliable if the number of effect sizes is small (Harrer et al. 2021; Pollet 2021), as is the case in this study. The two methods which were used to check for publication bias in the present research are failsafe N (Cooper et al. 2009) and a funnel-plot-based trim and fill method (Duval & Tweedie 2000), which have already been used in two other meta-analyses in interpreting studies (Mellinger & Hanson 2019; Wen & Dong 2019).

Failsafe N is the number of unpublished studies, with an average observed effect of zero, that would be needed to reduce the overall z-score to non-significance. If k=number of published studies, Rosenthal suggested that the failsafe N may be considered as being unlikely to exist if it is greater than a tolerance level of $5k + 10$.

In this paper, the trim and fill method is based on the random-effects model of the corresponding dataset. The statistics software performs three major steps: firstly, the funnel plot (a scatterplot of effect size data) helps with detecting potential publication bias. If there is no publication bias, the effect size data will have a symmetrical distribution; if there is, the distribution will be asymmetrical. Secondly, if publication bias is detected in the funnel plot, the trim and fill method will initiate an iterative process to trim studies (or effect size data) that contribute to the asymmetrical distribution until the distribution of the effect size data becomes symmetrical. Based on this symmetrical distribution, an adjusted mean effect will be yielded after rounds of iteration. Thus, thirdly, to correct the variance, another iterative procedure is initiated to reinstate the "trimmed" studies back into the analysis and then re-estimate the missing data based on the adjusted mean effect, aiming to compute a more accurate estimate of the mean effect.

The funnel plot consists of a funnel and two axes: the y-axis, showing the standard error and the x-axis showing the effect size of each study (Hedges' g). When there is no publication bias all studies would lie symmetrically around the pooled effect size (represented by a dotted line) within the shape of the funnel. When publication bias is present the funnel would look asymmetrical, because only studies with a large effect size were published, while studies without a significant, large effect would be missing.

4.   Results

The data collected are analysed by task and by study-control group pairs. For every comparison a three-level random-effects model was calculated and represented by means of a forest plot. An evaluation of the publication bias was also computed using two methods, failsafe N and the trim-and-fill method, and the latter was also visually represented with a funnel plot, included at the right of the corresponding forest plot in figures 2 to 10.

4.1   Reading span task

In the data collected there are 7 articles including a reading span task, from which 27 effect sizes were calculated. It was possible to compare interpreters to non-interpreters, interpreting students to non-interpreters and interpreters to interpreting students.

4.1.1   Reading span accuracy: Comparison between interpreters and non-interpreters

In four articles the reading span task results of interpreters were compared to those of non-interpreters, calculating a three-level random-effects model including 12 effect sizes (Figure 2). The overall effect is large ($g=1.11$, 95% CI [0.44, 1.79]) and the model is significant ($p=0.01$). The log-likelihood-ratio tests show that within-study variance is non-significant (BIC full model = 19.78 vs BIC reduced model = 17.39, $p=1.0$) and between-study variance is significant (BIC full model = 19.78 vs BIC reduced model = 22.39, $p=0.03$). 37.90% of the total variation can be attributed to sampling variance, 1.23% to within-study variance and 62.10% to between-study variance. The potential moderator of the model is language, since half of the tasks were in L1 and half in L2. The three-level model including the language moderator is not significant ($p=0.34$), with a larger effect of L1 ($g=1.22$, 95% CI [0.45, 1.99]) compared to L2 ($g=0.94$, 95% CI [0.10, 1.79]). Failsafe N indicates no publication bias (399, $p=<.0001$). The trim-and-fill funnel plot (Figure 2) estimates no missing effects and Hedges' g remains large and significant after the correction of publication bias ($g=1.07$, 95% CI [0.76, 1.37], $p=<0.0001$).

4.1.2   Reading span accuracy: Comparison between interpreting students and non-interpreters

In five articles the reading span task results of interpreting students were compared to those of non-interpreters. A three-level random-effects model was calculated including 11 effect sizes (Figure 3), 7 of which were based on a task in L1
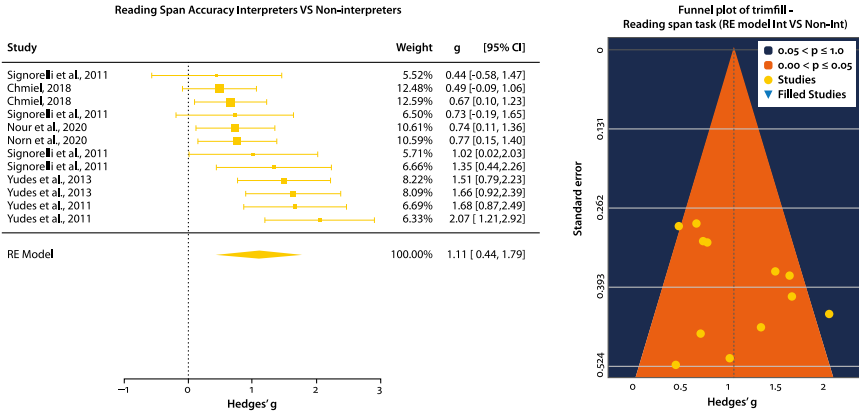
Figure 2.

and 4 on a task in L2. The overall effect is large ($g = 0.83$, 95% CI [0.01, 1.65]) and the model is significant ($p = 0.05$). The log-likelihood-ratio tests show that within-study variance is non-significant (BIC full model = 21.05 vs BIC reduced model = 18.74, $p = 1.0$) and between-study variance is significant (BIC full model = 21.05 vs BIC reduced model = 23.50, $p = 0.03$). 25.64% of the total variation can be attributed to sampling variance, 1.23% to within-study variance and 74.36% to between-study variance. The potential moderator of the model is language. The three-level model including the moderator is not significant ($p = 0.11$), with a larger effect of L1 ($g = 1.03$, 95% CI [0.15, 1.92]) compared to L2 ($g = y0.52$, 95% CI [−0.47, 1.52]). Fail-safe N indicates no publication bias (296, $p = <.0001$). The trim-and-fill funnel plot (Figure 3) estimates no missing effects and Hedges' g remains large and significant after the correction of publication bias ($g = 0.94$, 95% CI [0.57, 1.31], $p = <.0001$).

4.1.3   Reading span accuracy: Comparison between interpreters and interpreting students

In three articles the reading span task results of interpreters were compared to those of interpreting students by means of a three-level random-effects model including 4 effect sizes (Figure 4). The overall effect is small ($g = 0.23$, 95% CI [−0.95, 1.41]) and the model is not significant ($p = 0.48$). The log-likelihood-ratio tests show that within-study variance is non-significant (BIC full model = 6.30 vs BIC reduced model = 5.20, $p = 1.0$) and between-study variance is non-significant (BIC full model = 6.30 vs BIC reduced model = 5.61, $p = 0.52$). 42.41% of the total variation can be attributed to sampling variance, 2.01% to within-study variance and 57.59% to between-study variance. The potential moderator of the model is language, since half of the tasks were in L1 and half in L2. The three-level model
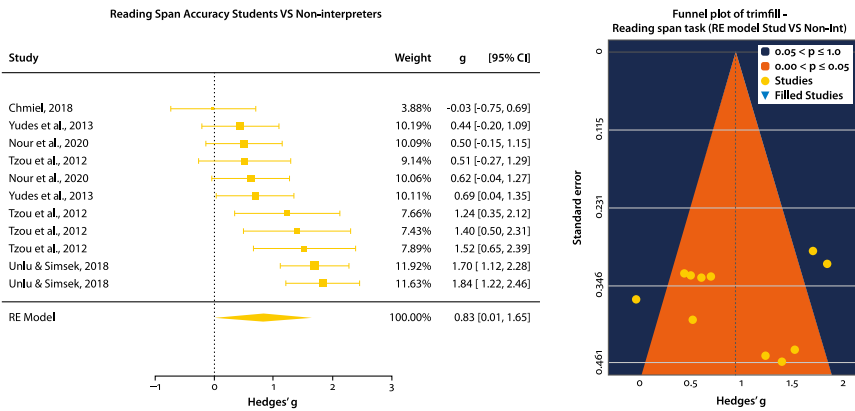
Figure 3.

including the language moderator is not significant ($p = 0.40$), with a larger effect of L1 ($g = 0.44$, 95% CI [−3.38, 4.27]) compared to L2 ($g = 0.01$, 95% CI [−3.70, 3.71]). Fail-safe N indicates a potential publication bias (0, $p = 0.08$). The trim-and-fill funnel plot (Figure 4) does not estimate missing studies and the model results remain non-significant after the correction of publication bias ($g = 0.22$, 95% CI [−0.20, 0.64], $p = 0.30$).
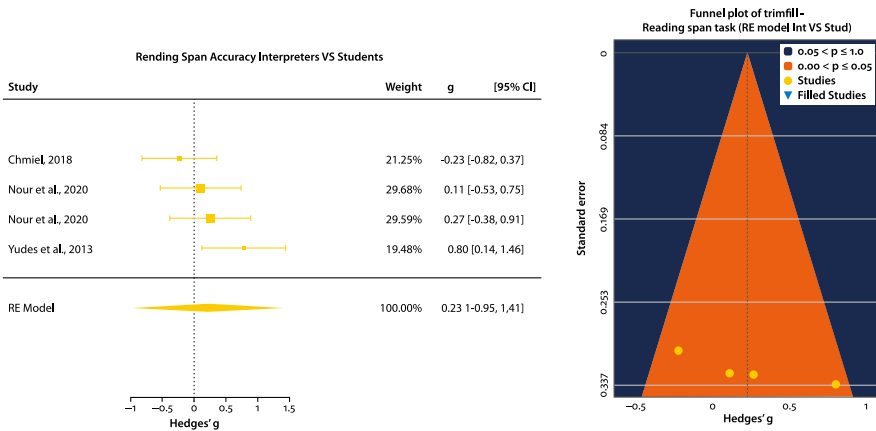

Figure 4.

4.2   N-back task

The dataset of the n-back task is rather small, it includes three studies comparing interpreting students to non-interpreters on accuracy and RT (five effect sizes each).

4.2.1 N-back task accuracy: Comparison between interpreting students and non-interpreters

A three-level random-effects model including five effect sizes of n-back accuracy was calculated (Figure 5). The overall effect is negligeable and negative ($g = -0.04$, 95% CI [-0.52, 0.43]) and the model is not significant ($p = 0.73$). The log-likelihood-ratio tests show that within-study variance is non-significant (BIC full model = 1.51 vs BIC reduced model = 0.12, $p = 1.0$) and between-study variance is non-significant (BIC full model = 1.51 vs BIC reduced model = 0.12, $p = 1.0$). 100% of the total variation can be attributed to sampling variance, 3.95% to within-study variance and 3.75% to between-study variance. Fail-safe N indicates a potential publication bias (0, $p = 0.35$). The trim-and-fill funnel plot (Figure 5) does not estimate missing studies and the model results remain non-significant after the correction of publication bias ($g = -0.04$, 95% CI [-0.26, 0.17], $p = 0.69$).
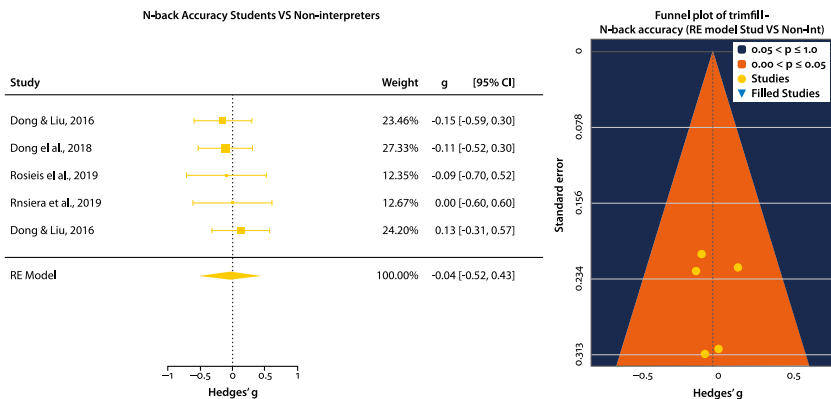

Figure 5.

4.2.2 N-back task RT: Comparison between interpreting students and non-interpreters

A three-level random-effects model including five effect sizes of n-back RT was calculated (Figure 6). The overall effect is negligeable ($g = 0.11$, 95% CI [-0.98, 1.20]) and the model is not significant ($p = 0.71$). The log-likelihood-ratio tests show that within-study variance is non-significant (BIC full model = 9.11 vs BIC reduced model = 8.19, $p = 0.50$) and between-study variance is non-significant (BIC full model = 9.11 vs BIC reduced model = 8.42, $p = 0.40$). 33.77% of the total variation can be attributed to sampling variance, 19.91% to within-study variance and 46.32% to between-study variance. There are no potential moderators. Fail-safe N indicates a potential publication bias (0, $p = 0.25$). The trim-and-fill funnel plot (Figure 6) estimates two missing studies on the left side and, after the correc-

tion of publication bias, the model results give a negative effect size and remain non-significant ($g = -0.15$, 95% CI [-0.61, 0.31], $p = 0.53$).
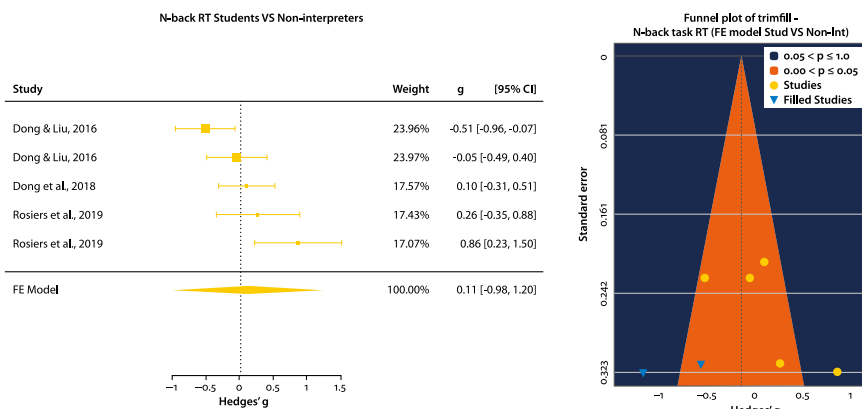

Figure 6.

4.3 Listening span task

In the data collected there are four articles including a listening span task, from which nine effect sizes were calculated. All the tasks except one are in the L1 of the informants. It was possible to compare interpreters to non-interpreters and interpreting students to non-interpreters.

4.3.1 Listening span accuracy: Comparison between interpreters and non-interpreters

In three articles the listening span task results of interpreters were compared to those of non-interpreters and a three-level random-effects model including five effect sizes was calculated (Figure 7). The overall effect is medium ($g = 0.61$, 95% CI [-0.01, 1.22]) and the model is significant ($p = 0.05$). The log-likelihood-ratio tests show that both within-study variance and between-study variance are non-significant (BIC full model = 3.94 vs BIC reduced model = 2.56, $p = 1.0$). 100% of the total variation can be attributed to sampling variance, 0% to within-study variance and 4.42% to between-study variance. Fail-safe N indicates a potential publication bias (28, $p = <.0001$). The trim-and-fill funnel plot (Figure 7) estimates a missing study on the right side, but Hedges' g remains medium and significant after the correction of publication bias ($g = 0.65$, 95% CI [0.39, 0.90], $p = <.0001$).
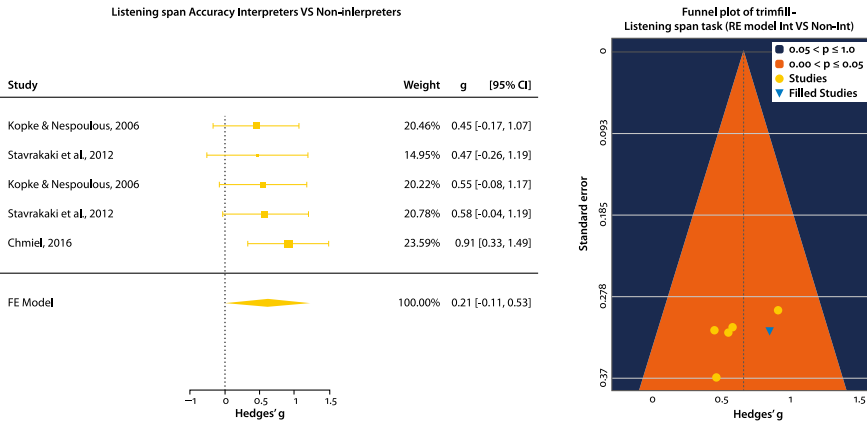
Figure 7.

4.3.2 Listening span accuracy: Comparison between interpreting students and non-interpreters

In two articles the listening span task results of interpreting students were compared to those of non-interpreters and a three-level random-effects model including three effect sizes was computed (Figure 8). The overall effect is medium (g = 0.51, 95% CI [−6.21, 7.22]) and the model is not significant (p 0.51). The log-likelihood-ratio tests show that within-study variance is non-significant (BIC full model = 4.38 vs BIC reduced model = 3.69, p = 1.0) and between-study variance is non-significant (BIC full model = 4.38 vs BIC reduced model = 5.22, p = 0.22). 14.78% of the total variation can be attributed to sampling variance, 3.87% to within-study variance and 85.22% to between-study variance. Fail-safe N indicates a potential publication bias (11, p = 0.0003). The trim-and-fill funnel plot (Figure 8) estimates no missing studies and Hedges' g is medium and non-significant after the correction of publication bias (g = 0.64, 95% CI [−0.09, 1.38], p = 0.08).

4.4 Dual tasks

In the data collected there were three articles including dual tasks, from which 13 effect sizes were calculated. It was possible to compare interpreters to non-interpreters and interpreting students to non-interpreters.
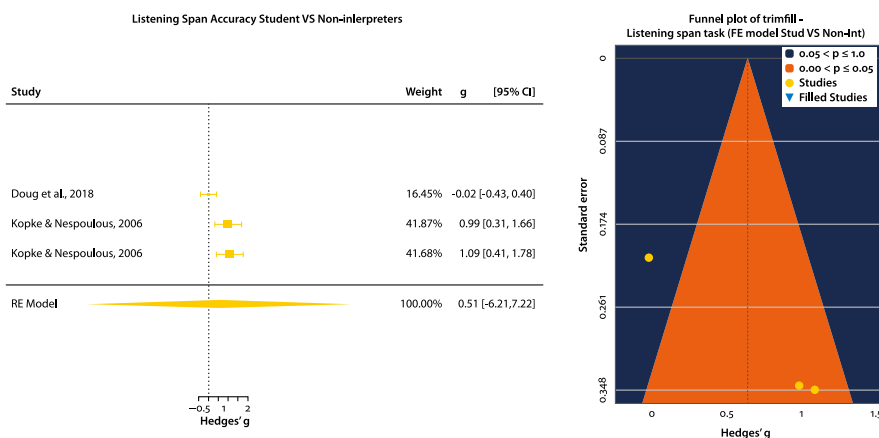
Figure 8.

4.4.1 Dual tasks accuracy: Comparison between interpreters and non-interpreters

In two articles dual tasks results of interpreters were compared to those of non-interpreters and a three-level random-effects model including six effect sizes was computed (Figure 9). Two effects were calculated from a dual task in L1 based on auditory stimuli (category and rhyme probe task) and four from a dual task in L2 based on visual stimuli (phonological cued recall). The overall effect is small ($g = 0.26$, 95% CI [−3.69, 4.21]) and the model is not significant ($p = 0.56$). The log-likelihood-ratio tests show that within-study variance is non-significant (BIC full model = 13.27 vs BIC reduced model = 12.70, $p = 0.31$) and between-study variance is non-significant (BIC full model = 13.27 vs BIC reduced model = 11.89, $p = 0.63$). 39.24% of the total variation can be attributed to sampling variance, 34.05% to within-study variance and 26.71% to between-study variance. The potential moderator of the model is language/channel of input. The three-level model including the moderator is not significant ($p = 0.42$), with a larger effect of L1/auditory ($g = 0.60$, 95% CI [−4.08, 5.28]) compared to L2/visual ($g = −0.02$, 95% CI [−3.96, 3.91]). Fail-safe N indicates a potential publication bias (0, $p = 0.08$). The trim-and-fill funnel plot (Figure 9) estimates no missing studies, but Hedges' g is small and the model non-significant after the correction of publication bias ($g = 0.22$, 95% CI [−0.24, 0.68], $p = 0.35$).
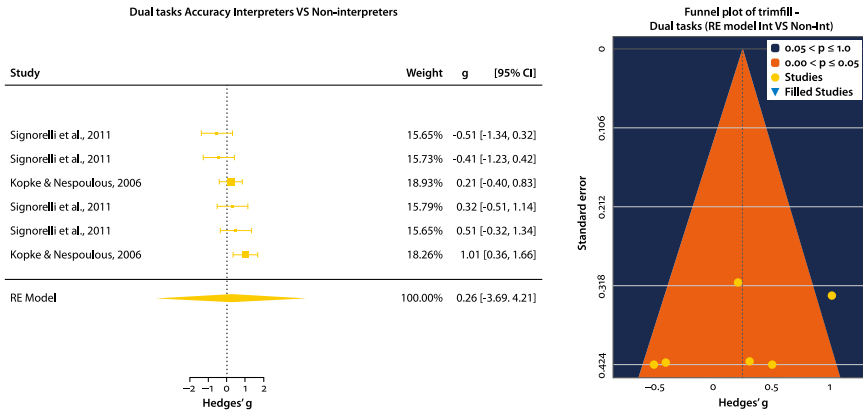
Figure 9.

4.4.2 Dual tasks accuracy: Comparison between interpreting students and non-interpreters

In two articles dual tasks results of interpreting students were compared to those of non-interpreters and a three-level random-effects model including six effect sizes was calculated (Figure 10). Two effects were computed from a dual task with verbal and auditory stimuli (category and rhyme probe task) and four from two dual tasks based on non-verbal and visual stimuli (OSPAN and symmetry span task). The overall effect is medium ($g = 0.61$, 95% CI [−6.17, 7.40]) and the model is not significant ($p = 0.46$). The log-likelihood-ratio tests show that within-study variance is non-significant (BIC full model = 11.98 vs BIC reduced model = 11.05, $p = 0.41$) and between-study variance is non-significant (BIC full model = 11.98 vs BIC reduced model = 13.07, $p = 0.10$). 13.30% of the total variation can be attributed to sampling variance, 8.91% to within-study variance and 77.80% to between-study variance. The potential moderator of the model is type/channel of input. The three-level model including the moderator is not significant ($p = 0.42$), with a larger effect of verbal/auditory ($g = 1.18$, 95% CI [−2.92, 5.27]) compared to non-verbal/visual ($g = 0.11$, 95% CI [−2.61, 2.82]). Fail-safe N indicates a potential publication bias (20, $p = 0.0003$). The trim-and-fill funnel plot (Figure 10) estimates no missing effects and Hedges' g is medium and non-significant after the correction of publication bias ($g = 0.46$, 95% CI [−0.05, 0.97], $p = 0.08$).
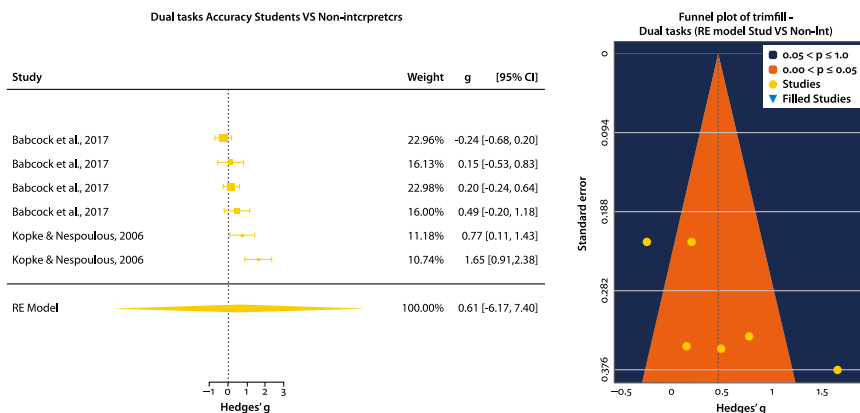
Figure 10.

4.5 Analyses on all WM tasks

In addition to the analyses by task, an analysis on the whole dataset was carried out through a three-level random-effects model. For this analysis, the results of both types of study groups (interpreters and interpreting students) were collected under a single variable and an additional variable was added to categorise the effects according to the three possible group pairs (interpreters vs non-interpreters, interpreters vs interpreting students and interpreting students vs non-interpreters). There are three variables that can potentially moderate task results: channel of input (auditory or visual), type of input (verbal or non-verbal), interpreting expertise difference between study and control group (interpreters vs non-interpreters, interpreting students vs non-interpreters, interpreters vs interpreting students).

The overall effect is medium (g=0.57, 95% CI [0.23, 0.90]) and the model is significant (p=0.0029). The log-likelihood-ratio tests show that within-study variance is significant (BIC full model=112.25 vs BIC reduced model=120.38, p=0.0005) and between-study variance is significant (BIC full model=112.25 vs BIC reduced model=116.20, p= 0.0048). 21.64% of the total variation can be attributed to sampling variance, 24.76% to within-study variance and 53.60% to between-study variance.

The three-level model including the channel of input as moderator is not significant (p=0.78), with a medium effect both for auditory (g=0.49, 95% CI [−0.22, 1.20]) and for visual stimuli (g=0.58, 95% CI [0.21, 0.96]). The three-level model including the type of input as moderator is significant (p=0.0063), with a medium effect for verbal stimuli (g=0.76, 95% CI [0.46, 1.06]) and a negligeable effect for non-verbal stimuli (g=0.07, 95% CI [−0.38, 0.52]). The three-level model

including the interpreting expertise difference between study and control group as moderator is significant ($p = 0.0002$), with a medium effect both for the difference between interpreters and non-interpreters ($g = 0.70$, 95% CI [0.29, 1.11]) and for the difference between interpreting students and non-interpreters ($g = 0.56$, 95% CI [0.17, 0.95]). The effect of the comparison between interpreters and interpreting students was, instead, negligeable and negative ($g = -0.15$, 95% CI [−0.66, 0.36]).

Fail-safe N indicates no publication bias (2815, $p = {<}.0001$). The trim-and-fill method estimates no missing studies, with a significant model after the correction of publication bias ($p = {<}.0001$) having a medium Hedges' g ($g = 0.55$, 95% CI [0.38, 0.72]).

5.   Discussion

The hypotheses which were formulated have been partially confirmed. The first hypothesis stated that verbal tasks results should find greater differences between study and control groups than non-verbal ones. In the reading span task a significant large effect was found with no publication bias, with the exception of the comparison between interpreters and interpreting students, where the effect was small and not significant. The overall effect was larger when comparing interpreters to non-interpreters ($g = 1.11$) than when comparing interpreting students to non-interpreters ($g = 0.83$). The listening span performance difference is significant and of medium size when comparing interpreters to non-interpreters, non-significant when comparing interpreting students to non-interpreters, but the dataset is very limited, so results need to be viewed with caution. The Hedges' g calculated from verbal dual tasks is small when comparing interpreters to non-interpreters and medium when comparing interpreting students to non-interpreters, but neither of the models is significant. This dataset is again very small and diversified. Moreover, the comparison between interpreters and non-interpreters includes the study of Signorelli et al. (2011), where groups differ in age. Since WM capacity declines with age (Park et al. 2002) and, since the interpreters' mean age was higher compared to the other study groups ($M = 56.75$, $SD = 7.85$, § 3.2.1), it can be argued that the age factor influenced task results. Considering all these findings, the first hypothesis has been partially confirmed because the models based on the larger datasets for the reading and listening span tasks proved to be adequate to find performance differences. This result was also confirmed by the three-level random-effects model of the whole dataset including the type of input as a moderating variable.

The second hypothesis about tasks in a foreign language being better at detecting differences between participants with various levels of expertise in interpreting was not confirmed. L1 tasks produced larger effects than L2 tasks, even if the models including the language moderator are never significant. This result is in contrast with interpreting studies literature, but in line with meta-analytical findings about the impact of bilingualism on WM capacity (Grundy & Timmer 2017), where a significant small to medium population effect size of 0.20 was discovered in favour of greater WM capacity for bilinguals, with larger effects in L1 than in L2 tasks. A possible explanation of the larger effects of L1 tasks in this study is that the number of participants is significantly higher in L1 tasks than in L2 tasks. The Shapiro-Wilk normality test highlighted that the number of informants per study was not normally distributed ($W = 0.87$, $p = 0.0001536$), so the difference in sample size between tasks in L1 and tasks in L2 was calculated using the non-parametric test Wilcoxon Rank Sum, which found a significant difference of moderate size in the total number of informants ($W = 333.5$, $p = 0.0242$, $r = 0.34$).

The second hypothesis also stated that the difference in performance should increase as the difference in interpreting expertise increases. This assumption is confirmed by the data, because the observed Hedges' g becomes increasingly larger as the expertise gap between study and control group widens (§ 4.5). This finding is in line with the results of Wen and Dong (2019), who found that the level of interpreter expertise significantly moderated the interpreter advantage.

Finally, the third hypothesis was that tasks based on auditory stimuli should detect more differences between study and control groups than tasks based on visual stimuli. This hypothesis was not confirmed, because the channel of input as a moderating variable is not significantly influencing task results, which produce in both cases a medium effect size. On the contrary, the effect of the type of input (verbal or non-verbal) has a significant impact on task results, in favour of verbal tasks. A visual and non-verbal task widely used as the N-back task did not prove to be a good option, at least according to the limited dataset collected for this meta-analysis. This result is in line with Mellinger and Hanson (2019), who found that interpreters outperformed control groups on both auditory and visual WM tasks. It can also be argued that, even if interpreting is an activity based on the processing of auditory input, in some situations interpreters are required to process both auditory and visual input, for example during sight translation or simultaneous interpreting with text. If, on the one hand, this means adding another input to an already multitasking activity, on the other hand empirical findings show an overall facilitating role of seeing a written text congruent with the speech during simultaneous interpreting (Desmet et al. 2018; Lambert 2004).

6.   Conclusion

This study found that interpreters and interpreting students have a significant WM advantage of medium size over non-interpreters in tasks based on verbal stimuli, but not in tasks based on non-verbal stimuli. The effect of the channel of input of the task stimuli is medium but not significant, implying that both auditory and visual tasks are sensitive to performance differences between study and control groups, provided that they include verbal stimuli. Finally, differences are larger when the two groups differ more in interpreting expertise.

Some limitations need to be considered. Meta-analysis is a quantitative method and the amount of data processed is very relevant to yield reliable results. WM tasks in interpreting studies literature is a very specific research topic, so the available data are limited. In addition, participants' characteristics vary a lot among studies, which is a problem for the reliability of data comparison. The measures of interpreting expertise are very diversified in the literature, from years of experience to hours of practice, often without considering the quality of the interpreted text (García 2014). More data are needed to confirm or reverse the observed trends, together with harmonising research practices. Researchers can benefit from meta-analytical efforts to make informed methodological decisions and there is a need for recognizing the value of reporting null results, as well as for study replication (Olalla-Soler 2020). Pre-registration and data repositories are also valuable ways to improve the quality of future meta-analyses (Mellinger and Hanson 2020).

References

Assink, Mark, Claudia E. van der Put, Machteld Hoeve, Sanne L. A. de Vries, Geert Jan J. M. Stams, and Frans J. Oort. 2015. 'Risk Factors for Persistent Delinquent Behavior among Juveniles: A Meta-Analytic Review'. *Clinical Psychology Review 42* (December): 47–61. https://doi.org/10.1016/j.cpr.2015.08.002

Assink, Mark, and Carlijn J. M. Wibbelink. 2016. 'Fitting Three-Level Meta-Analytic Models in R: A Step-by-Step Tutorial'. *The Quantitative Methods for Psychology 12* (3): 154–74. https://doi.org/10.20982/tqmp.12.3.p154

Babcock, Laura, Mariagrazia Capizzi, Sandra Arbula, and Antonino Vallesi. 2017. 'Short-Term Memory Improvement After Simultaneous Interpretation Training'. *Journal of Cognitive Enhancement 1* (3): 254–67. https://doi.org/10.1007/s41465-017-0011-x

Babcock, Laura, and Antonino Vallesi. 2017. 'Are Simultaneous Interpreters Expert Bilinguals, Unique Bilinguals, or Both?' *Bilingualism: Language and Cognition 20* (2): 403–17. https://doi.org/10.1017/S1366728915000735

Cai, Rendong, Yanping Dong, Nan Zhao, and Jiexuan Lin. 2015. 'Factors Contributing to Individual Differences in the Development of Consecutive Interpreting Competence for Beginner Student Interpreters'. *The Interpreter and Translator Trainer 9* (1): 104–20. https://doi.org/10.1080/1750399X.2015.1016279

Cheung, Mike W.-L. 2014. 'Modeling Dependent Effect Sizes with Three-Level Meta-Analyses: A Structural Equation Modeling Approach'. *Psychological Methods 19* (2): 211–29. https://doi.org/10.1037/a0032968

Chincotta, Dino, and Geoffrey Underwood. 1998. 'Simultaneous Interpreters and the Effect of Concurrent Articulation on Immediate Memory: A Bilingual Digit Span Study'. *Interpreting 3* (1): 1–20. https://doi.org/10.1075/intp.3.1.01chi

Chmiel, Agnieszka. 2018. 'In Search of the Working Memory Advantage in Conference Interpreting – Training, Experience and Task Effects'. *The International Journal of Bilingualism; London 22* (3): 371–84. https://doi.org/10.1177/1367006916681082

Christoffels, Ingrid K., Annette M. B. de Groot, and Judith F. Kroll. 2006. 'Memory and Language Skills in Simultaneous Interpreters: The Role of Expertise and Language Proficiency'. *Journal of Memory and Language 54* (3): 324–45. https://doi.org/10.1016/j.jml.2005.12.004

Cochrane Handbook for Systematic Reviews of Interventions: Version 6.2. 2021. 6.2. https://training.cochrane.org/handbook/current

Cohen, Jacob. 1977. *Statistical power analysis for the behavioral sciences.*

Cooper, Harris, Larry V. Hedges, and Jeffrey C. Valentine. 2009. *The Handbook of Research Synthesis and Meta-Analysis.* Russell Sage Foundation.

Desmet, Bart, Mieke Vandierendonck, and Bart Defrancq. 2018. 'Simultaneous Interpretation of Numbers and the Impact of Technological Support'. In *Interpreting and Technology*, edited by Claudio Fantinuoli, 13–27. Berlin: Language Science Press.

Diamantopoulos, Adamantios, and Judy A. Siguaw. 2000. 'Assessment of Model Fit'. In *Introducing LISREL: A Guide for the Uninitiated*, 82–100. London, UNITED KINGDOM: SAGE Publications. http://ebookcentral.proquest.com/lib/unibo/detail.action?docID =1191061. https://doi.org/10.4135/9781849209359.n7

Dong, Yanping, and Rendong Cai. 2015. 'Working Memory in Interpreting: A Commentary on Theoretical Models'. In *Working Memory in Second Language Acquisition and Processing*, edited by Zhisheng (Edward) Wen, Borges Mailce Mota, and Arthur McNeill, 63–84. Bristol: Multilingual Matters. https://doi.org/10.21832/9781783093595-008

Dong, Yanping, and Yuhua Liu. 2016. 'Classes in Translating and Interpreting Produce Differential Gains in Switching and Updating'. *Frontiers in Psychology 7* (August). https://doi.org/10.3389/fpsyg.2016.01297

Dong, Yanping, Yuhua Liu, and Rendong Cai. 2018. 'How Does Consecutive Interpreting Training Influence Working Memory: A Longitudinal Study of Potential Links Between the Two'. *Frontiers in Psychology 9*: 875. https://doi.org/10.3389/fpsyg.2018.00875

Duval, Sue, and Richard Tweedie. 2000. 'Trim and Fill: A Simple Funnel-Plot-Based Method of Testing and Adjusting for Publication Bias in Meta-Analysis'. *Biometrics* 56 (2): 455–63. https://doi.org/10.1111/j.0006-341X.2000.00455.x

Franco Aixelá, Javier. 2001. 'BITRA (Bibliography of Interpreting and Translation). Open-Access Database'. http://dti.ua.es/en/bitra/introduction.html

García, Adolfo M. 2014. 'The Interpreter Advantage Hypothesis: Preliminary Data Patterns and Empirically Motivated Questions'. *Translation and Interpreting Studies. The Journal of the American Translation and Interpreting Studies Association 9* (2): 219–38. https://doi.org/10.1075/tis.9.2.04gar

Gile, Daniel. 2009. *Basic Concepts and Models for Interpreter and Translator Training*. John Benjamins Publishing Company. https://doi.org/10.1075/btl.8

Gile, Daniel. 2021. 'CIRIN Bulletin'. 2021. https://cirin-gile.fr/

Grundy, John G., and Kalinka Timmer. 2017. 'Bilingualism and Working Memory Capacity: A Comprehensive Meta-Analysis'. *Second Language Research 33* (3): 325–40. https://doi.org/10.1177/0267658316678286

Harrer, Mathias, Pim Cuijpers, Toshi A. Furukawa, and David D. Ebert. 2021. *Doing Meta-Analysis with R: A Hands-On Guide*. CRC Press. https://doi.org/10.1201/9781003107347

Hedges, Larry V. 1981. 'Distribution Theory for Glass's Estimator of Effect Size and Related Estimators'. *Journal of Educational Statistics 6* (2): 107–28. https://doi.org/10.3102/10769986006002107

Hunter, John E., and Frank L. Schmidt. 1990. *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings*. Newbury Park, CA: Sage.

Hunter, John E., and Frank L. Schmidt. 2000. 'Fixed Effects vs. Random Effects Meta-Analysis Models: Implications for Cumulative Research Knowledge'. *International Journal of Selection and Assessment 8* (4): 275–92. https://doi.org/10.1111/1468-2389.00156

Injoque-Ricle, Irene, Juan Pablo Barreyro, Jesica Formoso, and Virginia I. Jaichenco. 2015. 'Expertise, Working Memory and Articulatory Suppression Effect: Their Relation with Simultaneous Interpreting Performance'. *Advances in Cognitive Psychology 11* (2): 56–63. https://doi.org/10.5709/acp-0171-1

Jin, Ya-shyuan. 2010. 'Is Working Memory Working in Consecutive Interpreting?' The University of Edinburgh. https://era.ed.ac.uk/handle/1842/4451

Kelley, Ken, and Kristopher J. Preacher. 2012. 'On Effect Size'. *Psychological Methods 17* (2): 137–152. https://doi.org/10.1037/a0028086

Köpke, Barbara, and Jean-Luc Nespoulous. 2006. 'Working Memory Performance in Expert and Novice Interpreters'. *Interpreting 8* (1): 1–23. https://doi.org/10.1075/intp.8.1.02kop

Lambert, Sylvie. 2004. 'Shared Attention during Sight Translation, Sight Interpretation and Simultaneous Interpretation'. *Meta : Journal Des Traducteurs / Meta: Translators' Journal 49* (2): 294–306. https://doi.org/10.7202/009352ar

Liu, Minhua, Diane L. Schallert, and Patrick J. Carroll. 2004. 'Working Memory and Expertise in Simultaneous Interpreting'. *Interpreting 6* (1): 19–42. https://doi.org/10.1075/intp.6.1.04liu

Logie, Robert, Valérie Camos, and Nelson Cowan, eds. 2021. *Working Memory – State of the Science*. Oxford: Oxford University Press.

McKenzie, Joanne E., and Sue E. Brennan. 2019. 'Synthesizing and Presenting Findings Using Other Methods'. In *Cochrane Handbook for Systematic Reviews of Interventions*, 321–47. John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119536604.ch12

Mellinger, Christopher D., and Thomas A. Hanson. 2019. 'Meta-Analyses of Simultaneous Interpreting and Working Memory'. *Interpreting* 21 (2): 165–95. https://doi.org/10.1075/intp.00026.mel

Mellinger, Christopher D., and Thomas A. Hanson. 2020. 'Meta-Analysis and Replication in Interpreting Studies'. *Interpreting* 22 (1): 140–49. https://doi.org/10.1075/intp.00037.mel

Morales, Julia, Francisca Padilla, Carlos J. Gómez-Ariza, and M. Teresa Bajo. 2015. 'Simultaneous Interpretation Selectively Influences Working Memory and Attentional Networks'. *Acta Psychologica* 155 (February): 82–91. https://doi.org/10.1016/j.actpsy.2014.12.004

Moser-Mercer, Barbara. 2000. 'Simultaneous Interpreting: Cognitive Potential and Limitations'. *Interpreting* 5 (2): 83–94. https://doi.org/10.1075/intp.5.2.03mos

Nour, Soudabeh, Esli Struys, and Helene Stengers. 2020a. 'Adaptive Control in Interpreters: Assessing the Impact of Training and Experience on Working Memory'. *Bilingualism: Language and Cognition* 23 (4): 772–79. https://doi.org/10.1017/S1366728920000127

Nour, Soudabeh, Esli Struys, Evy Woumans, Ily Hollebeke, and Hélène Stengers. 2020b. 'An Interpreter Advantage in Executive Functions?: A Systematic Review'. *Interpreting* 22 (2): 163–86. https://doi.org/10.1075/intp.00045.nou

Olalla-Soler, Christian. 2020. 'Practices and Attitudes toward Replication in Empirical Translation and Interpreting Studies'. *Target: International Journal of Translation Studies* 32 (1): 3–36. https://doi.org/10.1075/target.18159.ola

Page, Matthew J., Joanne E. McKenzie, Patrick M. Bossuyt, Isabelle Boutron, Tammy C. Hoffmann, Cynthia D. Mulrow, Larissa Shamseer, et al. 2021. 'The PRISMA 2020 Statement: An Updated Guideline for Reporting Systematic Reviews'. *BMJ* 372 (March): n71. https://doi.org/10.1136/bmj.n71

Park, Denise C., Gary Lautenschlager, Trey Hedden, Natalie S. Davidson, Anderson D. Smith, and Pamela K. Smith. 2002. 'Models of Visuospatial and Verbal Memory across the Adult Life Span'. *Psychology and Aging* 17 (2): 299–320. https://doi.org/10.1037/0882-7974.17.2.299

Pollet, Thomas. 2021. 'Meta-Analysis Course (in R)'. 2021. http://tvpollet.github.io/Meta-analysis_course

Raftery, Adrian E. 1995. 'Bayesian Model Selection in Social Research'. *Sociological Methodology* 25 (January): 111–63. https://doi.org/10.2307/271063

Rosiers, Alexandra, Evy Woumans, Wouter Duyck, and June Eyckmans. 2019. 'Investigating the Presumed Cognitive Advantage of Aspiring Interpreters'. *Interpreting* 21 (1). https://doi.org/10.1075/intp.00022.ros

Rothstein, Hannah R., Alexander J. Sutton, and Michael Borenstein. 2005. *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments*. john Wiley & Sons. http://ezproxy.unibo.it/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=cat05251a&AN=at.UBO2140726&lang=it&site=eds-live&scope=site. https://doi.org/10.1002/0470870168

RStudio, Team. 2021. *RStudio: Integrated Development Environment for R*. Boston: PBC. https://rstudio.com/.

Schmidt, Frank L., In-Sue Oh, and Theodore L. Hayes. 2009. 'Fixed- versus Random-Effects Models in Meta-Analysis: Model Properties and an Empirical Comparison of Differences in Results'. *British Journal of Mathematical & Statistical Psychology* 62 (1): 97–128. https://doi.org/10.1348/000711007X255327

Signorelli, Teresa M., Henk J. Haarmann, and Loraine K. Obler. 2011. 'Working Memory in Simultaneous Interpreters: Effects of Task and Age'. *International Journal of Bilingualism* 16 (2): 198–212. https://doi.org/10.1177/1367006911403200

Stavrakaki, Stavroula, Kalliopi Megari, Mary H. Kosmidis, Maria Apostolidou, and Eleni Takou. 2012. 'Working Memory and Verbal Fluency in Simultaneous Interpreters'. *Journal of Clinical and Experimental Neuropsychology* 34 (6): 624–33. https://doi.org/10.1080/13803395.2012.667068

Timarová, Šárka, Ivana Čeňková, Reine Meylaerts, Erik Hertog, Arnaud Szmalec, and Wouter Duyck. 2014. 'Simultaneous Interpreting and Working Memory Executive Control'. *Interpreting* 16 (2): 139–68. https://doi.org/10.1075/intp.16.2.01tim

Timarová, Šárka, Ivana Čeňková, Reine Meylaerts, Erik Hertog, Arnaud Szmalec, and Wouter Duyck. 2015. 'Simultaneous Interpreting and Working Memory Capacity'. In *Psycholinguistic and Cognitive Inquiries into Translation and Interpreting*, edited by Aline Ferreira and John W. Schwieter, 115:101–26. Amsterdam: John Benjamins Publishing Company. https://doi.org/10.1075/btl.115.05tim

Tzou, Yeh-Zu, Zohreh R. Eslami, Hsin-Chin Chen, and Jyotsna Vaid. 2012. 'Effect of Language Proficiency and Degree of Formal Training in Simultaneous Interpreting on Working Memory and Interpreting Performance: Evidence from Mandarin–English Speakers'. *International Journal of Bilingualism* 16 (2): 213–27. https://doi.org/10.1177/1367006911403197

Ünlü, Elena Antonova, and Çiğdem Sağın Şimşek. 2018. 'Testing the Impact of Formal Interpreting Training on Working Memory Capacity: Evidence from Turkish–English Students–Interpreters'. *Lingua* 209 (July): 78–88. https://doi.org/10.1016/j.lingua.2018.04.003

Van der Linden, Lize, Eowyn Van de Putte, Evy Woumans, Wouter Duyck, and Arnaud Szmalec. 2018. 'Does Extreme Language Control Training Improve Cognitive Control? A Comparison of Professional Interpreters, L2 Teachers and Monolinguals'. *Frontiers in Psychology* 9. https://doi.org/10.3389/fpsyg.2018.01998

Viechtbauer, Wolfgang. 2005. 'Bias and Efficiency of Meta-Analytic Variance Estimators in the Random-Effects Model'. *Journal of Educational and Behavioral Statistics* 30 (3): 261–93. https://doi.org/10.3102/10769986030003261

Viechtbauer, Wolfgang. 2010. 'Conducting Meta-Analyses in R with the Metafor Package'. *Journal of Statistical Software* 36 (August): 1–48. https://doi.org/10.18637/jss.v036.i03

Wen, Hao, and Yanping Dong. 2019. 'How Does Interpreting Experience Enhance Working Memory and Short-Term Memory: A Meta-Analysis'. *Journal of Cognitive Psychology* 31 (8): 769–84. https://doi.org/10.1080/20445911.2019.1674857

Yudes, Carolina, Pedro Macizo, and Teresa Bajo. 2011. 'The Influence of Expertise in Simultaneous Interpreting on Non-Verbal Executive Processes'. *Frontiers in Psychology* 2. https://doi.org/10.3389/fpsyg.2011.00309

Yudes, Carolina, Pedro Macizo, Luis Morales, and M. Teresa Bajo. 2013. 'Comprehension and Error Monitoring in Simultaneous Interpreters'. *Applied Psycholinguistics* 34 (5): 1039–57. https://doi.org/10.1017/S0142716412000112

Appendix A. Tasks included in the meta-analysis

Task	Type of input	Channel of input	L	Study group	Control group	Study	Interpreter advantage (in the original results)
dual task (category probe task)	verbal	auditory	L1	professional interpreters (21)	multilinguals (20)	Köpke & Nespoulous (2006)	NO
dual task (category probe task)	verbal	auditory	L1	professional interpreters (21)	non-language students (20)	Köpke & Nespoulous (2006)	YES
dual task (category probe task)	verbal	auditory	L1	interpreting students (18)	multilinguals (20)	Köpke & Nespoulous (2006)	NO
dual task (category probe task)	verbal	auditory	L1	interpreting students (18)	non-language students (20)	Köpke & Nespoulous (2006)	YES
dual task (automated symmetry span task)	non-verbal	visual	/	interpreting students (47)	translation students (10)	Babcock et al. (2017)	NO
dual task (automated symmetry span task)	non-verbal	visual	/	interpreting students (47)	non-language students (35)	Babcock et al. (2017)	NO
dual task (OSPAN)	non-verbal	visual	/	interpreting students (47)	translation students (10); non-language students (35)	Babcock et al. (2017)	NO
dual task (OSPAN)	non-verbal	visual	/	interpreting students (47)	non-language students (35)	Babcock et al. (2017)	NO
dual task (phonological cued recall)	verbal	visual	L2	younger interpreters (12)	younger multilinguals (11)	Signorelli et al. (2011)	YES
dual task (phonological cued recall)	verbal	visual	L2	younger interpreters (12)	older multilinguals (11)	Signorelli et al. (2011)	YES
dual task (phonological cued recall)	verbal	visual	L2	older interpreters (12)	younger multilinguals (11)	Signorelli et al. (2011)	NO

Task	Type of input	Channel of input	L	Study group	Control group	Study	Interpreter advantage (in the original results)
dual task (phonological cued recall)	verbal	visual	L2	older interpreters (12)	older multilinguals (11)	Signorelli et al. (2011)	NO
listening span	verbal	auditory	L1	professional conference interpreters (24)	English language students (27)	Chmiel (2018)	YES
listening span	verbal	auditory	L2	consecutive interpreting students (48)	English language students (43)	Dong et al. (2018)	NO
listening span	verbal	auditory	L1	professional interpreters (21)	multilinguals (20)	Köpke & Nespoulous (2006)	NO
listening span	verbal	auditory	L1	professional interpreters (21)	non-language students (20)	Köpke & Nespoulous (2006)	NO
listening span	verbal	auditory	L1	interpreting students (18)	multilinguals (20)	Köpke & Nespoulous (2006)	YES
listening span	verbal	auditory	L1	interpreting students (18)	non-language students (20)	Köpke & Nespoulous (2006)	YES
listening span	verbal	auditory	L1	simultaneous interpreters (15)	foreign-language teachers (15)	Stavrakaki et al. (2012)	NO
listening span	verbal	auditory	L1	simultaneous interpreters (15)	monolinguals (35)	Stavrakaki et al. (2012)	NO
n-back (accuracy)	non-verbal	visual	/	interpreting students (44)	translation students (35)	Dong & Liu (2016)	NO
n-back (accuracy)	non-verbal	visual	/	interpreting students (44)	English language students (37)	Dong & Liu (2016)	NO
n-back (RT)	non-verbal	visual	/	interpreting students (44)	translation students (35)	Dong & Liu (2016)	YES
n-back (RT)	non-verbal	visual	/	interpreting students (44)	English language students (37)	Dong & Liu (2016)	YES

Task	Type of input	Channel of input	L	Study group	Control group	Study	Interpreter advantage (in the original results)
n-back (accuracy)	non-verbal	visual	/	consecutive interpreting students (48)	English language students (43)	Dong et al. (2018)	NO
n-back (RT)	non-verbal	visual	/	consecutive interpreting students (48)	English language students (43)	Dong et al. (2018)	YES
n-back (accuracy)	non-verbal	visual	/	interpreting students (21)	translation students (21)	Rosiers et al. (2019)	NO
n-back (RT)	non-verbal	visual	/	interpreting students (21)	multilingual communication students (21)	Rosiers et al. (2019)	NO
n-back (accuracy)	non-verbal	visual	/	professional interpreters (17)	monolinguals (18)	Van der Linden et al. (2018)	NO
n-back (accuracy)	non-verbal	visual	/	professional interpreters (17)	foreign-language teachers (19)	Van der Linden et al. (2018)	NO
n-back (RT)	non-verbal	visual	/	professional interpreters (17)	monolinguals (18)	Van der Linden et al. (2018)	NO
n-back (RT)	non-verbal	visual	/	professional interpreters (17)	foreign-language teachers (19)	Van der Linden et al. (2018)	NO
reading span	verbal	visual	L2	professional conference interpreters (24)	multilinguals (24)	Chmiel (2018)	YES
reading span	verbal	visual	L2	conference interpreting trainees (20)	multilinguals (24)	Chmiel (2018)	YES

Task	Type of input	Channel of input	L	Study group	Control group	Study	Interpreter advantage (in the original results)
reading span	verbal	visual	L1	professional conference interpreters (24)	English language students (27)	Chmiel (2018)	YES
reading span	verbal	visual	L1	interpreting students (17)	translation students (21)	Nour et al. (2020a)	NO
reading span	verbal	visual	L1	professional interpreters (21)	translation students (21)	Nour et al. (2020a)	NO
reading span	verbal	visual	L2	interpreting students (17)	translation students (21)	Nour et al. (2020a)	NO
reading span	verbal	visual	L2	professional interpreters (21)	translation students (21)	Nour et al. (2020a)	NO
reading span	verbal	visual	L2	younger interpreters (12)	younger multilinguals (8)	Signorelli et al. (2011)	YES
reading span	verbal	visual	L2	younger interpreters (12);	older multilinguals (11)	Signorelli et al. (2011)	YES
reading span	verbal	visual	L2	older interpreters (7)	younger multilinguals (8)	Signorelli et al. (2011)	YES
reading span	verbal	visual	L2	older interpreters (7)	older multilinguals (11)	Signorelli et al. (2011)	YES
reading span	verbal	visual	L1	1st year interpreting students (11)	non-language students (16)	Tzou et al. (2012)	YES
reading span	verbal	visual	L1	2nd year students (9)	non-language students (16)	Tzou et al. (2012)	YES
reading span	verbal	visual	L2	1st year interpreting students (11)	non-language students (16)	Tzou et al. (2012)	YES
reading span	verbal	visual	L2	2nd year students (9)	non-language students (16)	Tzou et al. (2012)	YES

Task	Type of input	Channel of input	L	Study group	Control group	Study	Interpreter advantage (in the original results)
reading span	verbal	visual	L1	last-year student interpreters (26)	last-year English language students (32)	Ünlü & Şimşek (2018)	YES
reading span	verbal	visual	L1	last-year student interpreters (26)	first-year English language students (38)	Ünlü & Şimşek (2018)	YES
reading span	verbal	visual	L1	professional interpreters (16)	multilinguals (16)	Yudes et al. (2011)	YES
reading span	verbal	visual	L1	professional interpreters (16)	Spanish monolinguals (16)	Yudes et al. (2011)	YES
reading span	verbal	visual	L1	professional interpreters (19)	English monolinguals (19)	Yudes et al. (2013)	YES
reading span	verbal	visual	L1	professional interpreters (19)	multilinguals (19)	Yudes et al. (2013)	YES
reading span	verbal	visual	L1	interpreting students (19)	English monolinguals (19)	Yudes et al. (2013)	YES
reading span	verbal	visual	L1	interpreting students (19)	multilinguals (19)	Yudes et al. (2013)	YES

serena.ghiselli3@unibo.it

Biographical notes

Serena Ghiselli holds an MA in Interpreting and a PhD from the Department of Interpreting and Translation (DIT) of the University of Bologna at Forlì campus. Her PhD thesis is entitled "Working Memory and Selective Attention in Interpreting: Cognitive Development and Improvement Strategies". She is a post-doctoral researcher at the same Department, where she is working on a psychometric framework to test working memory, focused attention and split attention in interpreting-specific tasks.

iD https://orcid.org/0000-0002-1252-4018