Influence of Working Memory on PTE Academic task performance

9 September 2024

Authors:

Ana Ulicheva

Sumita Ishaque

Rose Clesham

Contents

Abstract

In this study, we tested the relationship between individuals' performance on the PTE Academic (PTE A) research test and their working memory capacity. Forty-four university students completed both the English proficiency test and an additional task measuring their working memory span (backward digit span task). We hypothesised that individuals with a smaller working memory span might exhibit poorer performance on the PTE A test or its individual items – if it were true that PTE A places excessive demands on test-takers' working memory. However, we did not find such evidence: No significant correlations were observed between PTE A subscores, or individual item scores, and test-takers' working memory capacity. This piece of evidence indicates that PTE A does not suffer from this type of construct-irrelevant variance.

Introduction

Imagine your brain as a superhero with two special powers. Short-term memory is like a sticky note, holding onto information briefly, such as remembering a phone number just long enough to dial it (Baddeley, 2020). Working memory (WM), however, is like a super-powered clipboard that not only holds information but also helps you use it to solve problems and make decisions (Nee et al., 2013). This paper explores how this super-powered working memory impacts performance on English proficiency test tasks, investigating whether these tasks truly measure language skills or if they are influenced by our brain's ability to juggle information.

Working memory is of particular interest to us because it has been shown to underlie learning skills such as reading attainment (Sesma et al., 2009), written language (Alloway et al., 2005), oral language (McInnes et al., 2003), and following directions (Jaroslawska et al., 2016). Backward digit span is one of the most commonly used tests designed to measure working memory in clinical research and practice. Backward digit span involves visually or orally presented numerical sequences, and is traditionally interpreted as a measure of working memory because it requires respondents to recall digits in reverse serial order (e.g., 5-2-7 is correctly recalled as 7-2-5; as opposed to a forward digit span task that is more directly associated with short-term memory).

Working memory and L2 proficiency

There is a growing consensus in the scientific literature that WM plays an important role in bilingual language processing and moderates performance on a various measures of L2

processing (for a review see Linck et al., 2014). For example, speakers with larger WM capacity are better able to learn vocabulary (in both first and second languages), write more proficiently, and have better L1 reading and listening comprehension (Daneman & Hannon, 2007; Engle, 2001). WM capacity limitations seem to be particularly apparent in low-proficiency bilinguals as opposed to high-proficiency bilinguals (Hummel, 2009). The directionality of this effect is unknown. Some researchers argue that WM resources constrain performance on L2 tasks (Linck et al., 2014; see also Novick et al., 2014). Others suggest that prolonged exposure to multiple languages enhances executive functions in bilinguals (Bialystok, 2010).

More broadly, WM constrains other meta-linguistic cognitive processes, such as taking notes, following directions, or ignoring distractions (Engle, 2001). It should come as no surprise that language tasks that involve remembering information, make inferences, access information from long-term memory and integrating new information (Daneman & Hannon, 2007) may show some degree of WM involvement. This should also hold true in the context of high-stakes language proficiency testing.

However, the primary aim of high-stakes language proficiency tests is to measure language proficiency, and not working memory. In this context, it is important to keep in mind that more cognitively demanding tasks might give individuals with high WM capacities an inherent advantage over those with lower WM capacities, thus introducing construct-irrelevant variance and benefitting some test-takers more than others. Thus, test providers must ensure that performance on test tasks is not overly reliant on test-takers' memory capacity. For instance, nonverbal WM measures, on average, typically show a .18 correlation with L2 outcome measures (range .13-.22), which is lower than that for verbal WM measures (around .26; Linck et al., 2014). Higher correlations may be a cause for concern.

Working memory involvement in PTE A test tasks

While there has been no direct evidence regarding WM involvement in various PTE A tests tasks, such suggestions have been sporadically made in academic literature, as well as third-party preparation resources online. Wei & Zheng (2017), following Buck (2001), for example, speculate that the 'Repeat Sentence' item may assess "working memory (rather than cognitive listening skills)" (p. 11). 'Repeat Sentence' involves repeating a sentence verbatim (see Appendix for the description of the item). Length of sentences ranges from 9 to 15 words. These authors suggest that construct-irrelevant variance may be particularly problematic for longer sentences. Similarly, Pae et al. (2016) states that "the dictation and sentence repetition tasks measure[d] the capacity of working memory" without any direct measurement of working memory capacity administered in their study. In a similar vein, numerous third-party resources that focus on preparation for PTE A mention working

memory capacity as a prerequisite skill for succeeding on the 'Repeat Sentence' task (e.g., Vuong, 2020). The notion is also pervasive in test-taker perceptions of the test (Fan et al., 2021).

As this brief review demonstrates, the idea that performance on the PTE A 'Repeat Sentence' item is not just related to test-taker working memory, but is somehow determined by it, keeps circulating in academic and non-academic sources. As of now, we are not aware of any published evidence in favour or against this claim (although see Brunfaut & Revesz, 2015, who reported a significant .3 correlation between listening scores on PTE A and working memory measures). If it is indeed the case that a strong relationship exists between an individual's non-verbal working memory capacity and their performance on the 'Repeat Sentence' item, then this item may not be measuring the intended language skills (e.g., speech processing and production; Appendix) and a revision of this item might be warranted. If on the other hand, no such relationship can be proven or the relationship is of an expected nature, this evidence should be disseminated to help reduce the spread of misleading information.

The first type of evidence we sought to obtain on this matter was correlational: Is there a significant correlation between test-takers working memory span and their performance on the 'Repeat Sentence' item? We designed a study where the same group of participants took the PTE A test first and then completed a working memory assessment. We did not limit our focus to the 'Repeat Sentence' item and considered correlations between all PTE A test item scores and the working memory measure.

Methods

Participants

Participants were current university students in the UK (undergraduate or postgraduate). We recruited 44 students with any language background; 13 were native speakers of English. One-third came from University College London, 35% from Southampton University, and the rest came from London School of Economics, Royal Holloway University of London and other institutions in the country. Participants were, on average, 23 years old (ranging between 19 and 37); 34 identified as females and 10 identified as males. Half of the participants were born in China, and the rest came from Europe, Africa, Middle East, and Asia. Consequently, 59% indicated that their first language was Chinese, 32% English, and the remaining 9% spoke Arabic, Japanese or Russian as their first language.

Materials

PTE A research test. The PTE A exam evaluates candidates' English proficiency within an academic context to determine their preparedness for studying in English-speaking environments. Conducted entirely on computers at Pearson VUE test centres, the test is also automatically scored. Each test taker receives a comprehensive score report that includes an overall score, individual scores for listening, reading, speaking, and writing. PTE A tasks are integrated, meaning that they require a combination of listening, reading, writing and speaking abilities. The initial raw scores for items, generated by the machine, are transformed and combined into the final composite score(s). A research version of the PTE A test was administered. Using research versions allows us to explore particular areas of interest, such as field testing new or amended item types and rubrics. In this case, the research version was largely unchanged from the operational test in its composition, scoring, or administration. The differences concerned minor modifications to selected items (e.g., to have more word count), the addition of new items (summarise group discussion, respond to a situation), and changes to some item's scoring rubrics.

Remote tasks. Remote tasks were created and hosted on the Gorilla platform (https://gorilla.sc/; Anwyl-Irvine et al., 2019) – a service widely used among academics in cognitive psychology and life sciences to run behavioural experiments. Participants received a link and accessed the tasks via a web browser on their PC or laptop. No other devices, such as phones or tablets, were allowed. Aside from tasks that we report on in this paper, participants completed a range of other language tasks. The total duration of the whole suite of tasks was between 50 and 70 minutes. Short breaks in-between tasks were allowed. A quiet environment was recommended.

- Questionnaire. The questionnaire was always administered first. We collected data on participants' demographic and language background (these questions were adapted from the Language Experience and Proficiency Questionnaire; Kaushanskaya et al., 2020).
- Backward digit span task. This task was always administered last. Participants were presented with lists of digits on the computer screen, and had to repeat them backward by clicking on an answer pad. After two practice trials, five lists of two digits were presented. Participants had to succeed on at least three trials to move on to the next level, where three digits were presented. This procedure was repeated until subjects failed at a given level. The level reached by each participant was recorded reflecting participants' working memory span. The task was adopted from Massonnié et al. (2022).

Procedure

After students expressed their interest, all received instructions via email on how to prepare for and sit the PTE A test. The test was administered in one of the UK centres in person. The cost of PTE A was waived. Participants were compensated for their travel to and from their closest test centre. They were offered additional remuneration for completing all requirements for the study at an hourly rate exceeding minimum wage in the UK. Data collection took place between the 8th of January and the 8th of March 2024. Participants were also offered a free PTE A voucher. There were not given any feedback on their performance in the PTE A test.

After participants sat PTE A, they received instructions on how to complete remote tasks. They were urged to complete the tasks as soon as possible. All participants received detailed information on Pearson's privacy policy, data handing and use and had to give consent to participate in the study.

Analysis

We considered the following measures from the PTE A test: overall score, skill subscores (listening, speaking, reading, writing), as well as raw scores on all individual items (machine-generated). The working memory span measure was that obtained from the backward digit span task. Our main analysis involved computing pairwise Pearson's correlations between each PTE A measure and the working memory span measure. Significant correlations were examined in depth. This included a Bayes Factor analysis to gauge the strength of evidence we obtained (Heck et al., 2022). As described in the Introduction, we were particularly interested in the relationship between 'Repeat Sentence' and WM scores. Therefore, this correlation was examined in depth irrespective of its significance status.

Results

In our sample, the values for WM span measure ranged from 2 to 10 (mean was 4.7). Overall PTE scores ranged from 47 to 90 (mean was 69) indicating a range of proficiency levels. Tables 1 and 2 include results of the main analysis. None of the correlations that were examined came out as significant.

Next, we performed a further test (Bayes Factor) in order to gauge strength of evidence we obtained regarding one item of a priori interest (16-LS-REPT 'Repeat Sentence'; $r(44) = .05$, $p = .74$).

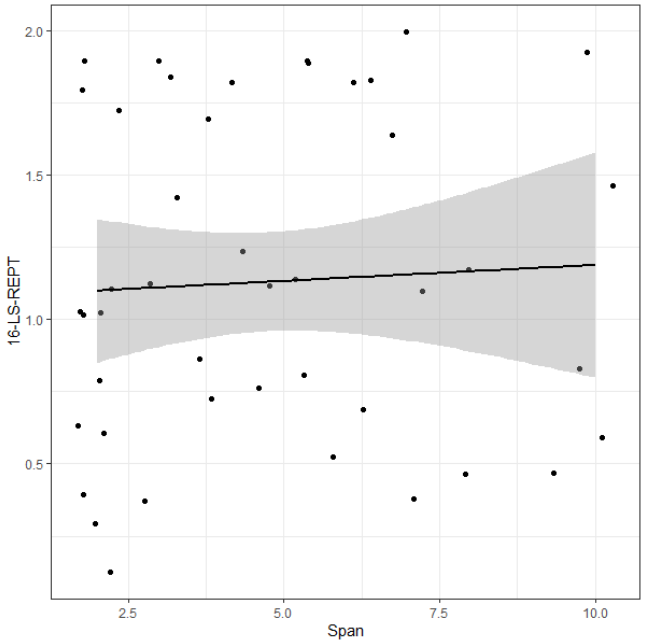
Table 1. Pearson's correlations between participants' scores on the whole test, subskill scores, and their WM span.

PTE SECTION	CORRELATION WITH WM
overall	0.09
listening	0.15
reading	0.02
speaking	0.12
writing	0.12

Table 2. Pearson's correlations between participants' scores on individual items and their WM span.

ITEM	ITEM DESCRIPTION	CORRELATION WITH WM
01-RR-SAMC	Multiple Choice, Single Answer	0.07
02-RR-MAMC	Multiple Choice, Multiple Answers	-0.06
04-RR-DRDR	Re-order Paragraphs	0.16
05-RR-GAPS	Fill in the Blanks	-0.11
06-LR-HILI	Highlight Correct Summary	0
07-SR-READ	Read Aloud	-0.02
08-RW-SUMM	Summarize Written Text	0.04
09-LL-SAMC	Multiple Choice, Single Answer	0.19
10-LL-MAMC	Multiple Choice, Multiple Answers	0.2
11-LL-GAPS	Select Missing Word	-0.13
12-LR-HOTS	Highlight Incorrect Words	0.14
13-LW-GAPS	Fill in the Blanks	0.1
14-LW-DICT	Write from Dictation	0.19
15-LW-SUMM	Summarize Spoken Text	0.17
16-LS-REPT	Repeat Sentence	0.05
17-WW-ESSA	Write Essay	-0.05
18-RW-GAPS	Fill in the Blanks	-0.09
19-SS-DESC	Describe Image	0.16
20-LS-PRES	Retell Lecture	0.17
21-LS-SAQS	Answer Short Question	-0.08
22-LS-SGD	Summarise Group Discussion	0
23-SS-SITU	Respond to a Situation	-0.04

Figure 1. Scatterplots illustrating relationships between participants' scores on the item 16-LS-REPT 'Repeat Sentence' and their WM spans (plotted along x-axes). Note that the data points were jittered for readability purposes to prevent visual overlap.



Bayes Factor analysis was performed using the correlationBF() function from the BayesFactor library in R (Morey & Rouder, 2014) with default parameters. Bayes Factor value was r = 0.35 for item 16-LS-REPT. This value can be interpreted as anecdotal to moderate evidence in favour of the null hypothesis (Lee & Wagenmakers, 2014). In other words, we can be somewhat confident that there is no relationship between the scores on the 'Repeat Sentence' item and WM span, but more data collection would be warranted.

Discussion

It is well-known that complex, demanding tasks engage higher-level cognitive processes including working memory (Engle, 2001). This is also true for various tasks measuring of L2 processing (Linck et al., 2014). In this study we tested whether scores on a high-stakes language proficiency test, PTE, are associated with test-takers' working memory capacity to a greater extent than we would expect. Strong associations may suggest that the test suffers from construct-irrelevant variance and puts some test-takers at a disadvantage.

The study was designed to be exploratory in nature. One a priori hypothesis that we had formulated based on previous research and anecdotal evidence was that one item in particular, 'Repeat Sentence', may show a strong positive relationship with the working memory span because it requires memorising and reproducing verbal material verbatim. However, we also tested all other test items for completeness. The correlation between scores on the 'Repeat Sentence' item and working memory spans was negligible ($r(44) = .05$, $p = .74$), which was confirmed through the Bayes Factor analysis. These findings suggest that 'Repeat Sentence' does not tax working memory excessively, with a reasonable level of confidence (see also Klem et al., 2015, for similar results). However, further data collection

would be warranted. Furthermore, the overall test, subskill, and almost all other item-specific performance, does not appear to be dependent on individuals' working memory capacity. This is a reassuring result that raises our confidence in PTE A test's fairness and validity.

We should note here some limitations of this study. Firstly, our sample was made up of young adults who are already immersed in the English-speaking environment through their academic study in the UK. It would be warranted to repeat this study in a more heterogeneous sample of individuals to ensure that those with a lower English proficiency and working memory span do not show a different pattern of performance. Secondly, a larger sample may be required for replicating these findings and further increasing our level of confidence in them (e.g., 195 participants; Brysbaert, 2019); otherwise, alternative designs should be considered (for example, performing tasks in question under concurrent memory load).

To conclude, we found no evidence that performance on PTE A test items is constrained by test-takers' working memory capacity. This is also true for the 'Repeat Sentence' item where test-takers are required to memorise and repeat sentences up to 15-words long. While performance on this item has been suggested in academic and non-academic literature to be reliant on working memory capacity, we have shown this not to be the case in our sample of participants.

References

Alloway, T.P., Gathercole, S.E., Willis, C., & Adams, A.-M. (2004). A structural analysis of working memory and related cognitive skills in young children. Journal of Experimental Child Psychology, 87 (2), 85–106. doi:10.1016/j.jecp.2003.10. 002

Alloway, T. P., Gathercole, S. E., Adams, A. M., Willis, C., Eaglen, R., & Lamont, E. (2005). Working memory and phonological awareness as predictors of progress towards early learning goals at school entry. British Journal of Developmental Psychology, 23(3), 417-426.

Anwyl-Irvine, A.L., Massonié, J., Flitton, A., Kirkham, N.Z., Evershed, J.K. (2019). Gorilla in our midst: an online behavioural experiment builder Behavior Research Methods. Doi: https://doi.org/10.3758/s13428-019-01237-x

Baddeley, A. (2020). Working memory. In Memory (pp. 71-111). Routledge.

Buck, G. (2001). Assessing Listening. Cambridge: Cambridge University Press.

Bialystok, E. (2010). Bilingualism. Wiley interdisciplinary reviews: Cognitive science, 1(4), 559-572.

Brunfaut, T., & Revesz, A. (2015). The role of task and listener characteristics in second language listening. Tesol Quarterly, 49(1), 141-168.

Brysbaert M. (2019). How Many Participants Do We Have to Include in Properly Powered Experiments? A Tutorial of Power Analysis with Reference Tables. Journal of cognition, 2(1), 16. https://doi.org/10.5334/joc.72

Cornelius, N., & Brown, J. L. (2020). The interaction of repetition and difficulty for working memory in melodic dictation tasks. Research Studies in Music Education, 42(3), 368-382. https://doi.org/10.1177/1321103X18821194

Daneman, M., & Hannon, B. (2007). What do working memory span tasks like reading span really measure. The cognitive neuroscience of working memory, 21-42.

Engle, R. W. (2001). What is working memory capacity?

Fan, J., Jin, Y., Kong, X., & Zhang, X. How do test takers prepare for computerized speaking tests? A comparative study of the PTE Academic and the CET-SET.

Heck, D. W., Boehm, U., Böing-Messing, F., Bürkner, P. C., Derks, K., Dienes, Z., ... & Hoijtink, H. (2022). A review of applications of the Bayes factor in psychological research. Psychological Methods.

Hummel, K. M. (2009). Aptitude, phonological memory, and second language proficiency in nonnovice adult learners. Applied Psycholinguistics, 30(2), 225-249.

Jaroslawska, A. J., Gathercole, S. E., Allen, R. J., & Holmes, J. (2016). Following instructions from working memory: Why does action at encoding and recall help?. Memory & Cognition, 44, 1183-1191.

Kaushanskaya, M., Blumenfeld, H. K., & Marian, V. (2020). The language experience and proficiency questionnaire (LEAP-Q): Ten years later. Bilingualism: Language and Cognition, 23(5), 945-950.

Klem, M., Melby-Lervåg, M., Hagtvet, B., Lyster, S. A. H., Gustafsson, J. E., & Hulme, C. (2015). Sentence repetition is a measure of children's language skills rather than working memory limitations. Developmental science, 18(1), 146-154.

Lee, M. D., & Wagenmakers, E. J. (2014). Bayesian cognitive modeling: A practical course. Cambridge university press.

Linck, J. A., Osthus, P., Koeth, J. T., & Bunting, M. F. (2014). Working memory and second language comprehension and production: A meta-analysis. Psychonomic bulletin & review, 21, 861-883.

Massonnié, J., Mareschal, D. & Kirkham, N. (2022). Individual differences in dealing with classroom noise disturbance. Mind, Brain, Education. doi:10.1111/mbe.12322

McInnes, A., Humphries, T., Hogg-Johnson, S., & Tannock, R. (2003). Listening comprehension and working memory are impaired in attention-deficit hyperactivity disorder irrespective of language impairment. Journal of abnormal child psychology, 31, 427-443.

Morey, R. D., & Rouder, J. N. (2014). BayesFactor version 0.9.9: An R package for computing Bayes factor for a variety of psychological research designs. Software http://bayesfactorpcl.r-forge.r-project.org/

Nee, D. E., Brown, J. W., Askren, M. K., Berman, M. G., Demiralp, E., Krawitz, A., & Jonides, J. (2013). A meta-analysis of executive components of working memory. Cerebral cortex, 23(2), 264-282.

Novick, J. M., Hussey, E., Teubner-Rhodes, S., Harbison, J. I., & Bunting, M. F. (2014). Clearing the garden-path: Improving sentence processing through cognitive control training. Language, Cognition and Neuroscience, 29(2), 186-217.

Pae, H. K., Sevcik, R. A., Greenberg, D., & Kim, S. A. (2016). Relationships among metacognitive skills, listening, and academic reading in English as a foreign language. Linguistic Research, 33.

Re, A. M., Mirandola, C., Esposito, S. S., & Capodieci, A. (2014). Spelling errors among children with ADHD symptoms: The role of working memory. Research in developmental disabilities, 35(9), 2199-2204.

Sesma, H. W., Mahone, E. M., Levine, T., Eason, S. H., & Cutting, L. E. (2009). The contribution of executive skills to reading comprehension. Child neuropsychology, 15(3), 232-246.

Vuong, M. (2020). How to improve memory for PTE Repeat Sentence. https://ptemagic.com.au/how-to-improve-memory-for-pte-repeat-sentence/

Wei, W., & Zheng, Y. (2017). An investigation of integrative and independent listening test tasks in a computerised academic English test. Computer Assisted Language Learning, 30(8), 864-883. https://eprints.soton.ac.uk/413819/1/Wei_Zheng_2017_Author_check_version.pdf

Appendix

Repeat Sentence (16-LS-REPT)

In this task, test-takers hear a sentence of normal length (9-15 words) once and are asked to immediately repeat it back verbatim. Repeat Sentence items assess the integrated skills of Listening and Speaking, measuring and scoring features of accuracy, pronunciation, and fluency. Such tasks have been used extensively in both first- and second-language speech studies since the 1960s (Vinther, 2002) and have been recognized as a useful measure of second language listening and speaking proficiency in many languages (Tracy-Ventura,

McManus, Norris, & Ortega, 2013). Sentence repetition tasks are known to require the following four processes: speech perception and processing (hearing the input and processing it into meanings), representation (structuring the various meanings into a larger meaning or representation of a story, action, description, etc.), memory (maintaining this representation as well as a representation of the actual language used), and speech production (re-producing the prompt as one's internal representations allow) (Bley-Vroman & Chaudron, 1994). While sentence repetition is a brief and constrained task, it requires a series of complex underlying linguistic processing skills.

The format of Repeat Sentence items assesses the automaticity of a test taker's linguistic processing in real time. Test takers must understand spoken utterances as well as recognize and process linguistic units in order to repeat a sentence, just as they need to do in real-time conversations or other listening and speaking situations in the real world. By requiring test takers to process language in real time, just as people do in everyday conversation, the task taps into underlying or implicit L2 competence (Erlam, 2006; Hulstijn, 2015; Sarandi, 2015). The focus on automaticity is key here – while some tasks on PTE give planning time, others require a quick response with no time to plan or think back, just as in real conversation.
Studies in the field of language assessment suggest that the performance on this elicited imitation task is a good predictor of the performance on more open-ended, communicative tasks (Van Moere, Xu, & Klungtvedt, 2012). Also, Kostromitina and Plonsky (2021) showed in their meta-analysis that the scores on this task demonstrate excellent internal consistency as well as substantial correlations with various criterion measures of L2 proficiency including self-assessment and standardized proficiency assessment. In addition, providing that higher proficiency speakers can repeat longer, complex sentences as long as the sentence is meaningful and the syntax is familiar to them (Radloff & Hallberg, 1991), this task enables efficient, reliable discrimination across different levels of test-takers by systematically changing the length and complexity of input sentences.

Finally, although this task might appear low in face validity and far from a communicative task at first glance, speakers regularly adopt their interlocutor's words and grammar into their own speech in everyday conversation (Levinson, 1983; Brown & Yule, 1983). From a sociolinguistic perspective, by repeating phrases or sentences of other speakers, listeners can show acceptance of others' utterances and give evidence of their own participation so that speakers feel a sense of endorsement (Tannen, 1989).