

RESEARCH

Open Access



# The comparative impacts of portfolio-based assessment, self-assessment, and scaffolded peer assessment on reading comprehension, vocabulary learning, and grammatical accuracy: insights from working memory capacity

Anwar Hammad Al-Rashidi<sup>1\*</sup> , Balachandran Vadivel<sup>2</sup> , Nawroz Ramadan Khalil<sup>2</sup> and Nirvana Basim<sup>3</sup>

\*Correspondence:  
a.alrashidi@psau.edu.sa

<sup>1</sup> Department of Psychology,  
College of Education, Prince  
Sattam Bin Abdulaziz University,  
Al-Kharj, Saudi Arabia

<sup>2</sup> Department of English, Cihan  
University-Duhok, Kurdistan  
Region, Iraq

<sup>3</sup> Department of Literature  
and Linguistics, Kandahar  
University, Kandahar, Afghanistan

## Abstract

This research was carried out to comparatively study the impacts of portfolio-based assessment, self-assessment, and scaffolded peer assessment on reading comprehension, vocabulary learning, and grammatical accuracy of Afghan English as a foreign language learners. To accomplish this, 172 learners enrolled at a language institute, through an Oxford Quick Placement Test (OQPT), 120 lower-intermediate learners and 5 higher-intermediate learners were selected. These selected participants were assigned into four groups: portfolio group ( $N = 30$ ), self-assessment group ( $N = 30$ ), scaffolded peer assessment group ( $N = 35$ ), and control group ( $N = 30$ ). The five higher-intermediate learners were injected into the scaffolded peer assessment group to function as the mediators, hence more participants in the group. After selecting the participants, through a reading-span test developed by Shahnazari (2013), learners' working memory (WM) span was determined. It was discovered that 16 subjects in the portfolio condition, 14 self-assessment learners, 18 participants in the peer assessment group, and 13 participants in the control condition had high WM, while the rest of the participants had low WM. Thereafter, through validated instructor-made tests, subjects' reading comprehension, knowledge of targeted lexical items, and grammatical accuracy at baseline were determined. Then, a ten-session treatment began. After the treatment, a follow-up post-test was administered. The results of three two-way between-group MANOVA disclosed that all three experimental conditions outstripped the comparison group on the second occasion and that high WM learners outstripped low WM learners (with a large effect size on reading comprehension test (partial eta squared = .365), a moderate effect size on the same test among high vs. low WM learners (partial eta squared = .095), a large effect size on vocabulary post-test (partial eta squared = .465), a moderate effect size on the same test among high vs. low WM learners (partial eta squared = .083), a large effect size on grammar test (partial eta squared = .500), and a moderate effect size on the same test among high vs. low WM learners (partial eta squared = .072)). The results further revealed that subjects in the scaffolded peer assessment group outstripped subjects in other experimental

conditions, but the difference was non-significant. Additionally, the difference between the portfolio assessment and self-assessment group was not statistically significant. The implications of the study are reported.

**Keywords:** Alternative assessment, Grammatical accuracy, Portfolio assessment, Scaffolded peer assessment, Self-assessment, Reading comprehension, Vocabulary learning

## Introduction

There has always been a passion among L2 practitioners to improve their learners' language proficiency. To this end, they have always tried and tested different instructional techniques and instruments, so they might facilitate language learning among their learners. In ESL/EFL environments, alternative assessment techniques are frequently utilized to enhance learning. According to Hargreaves et al. (2002), alternative assessment is intended to create strong, productive learning for students themselves in contrast to standardized testing. As examples of alternate evaluation methods, they give conferences, observational checklists, self- or peer assessments, diaries, and learning logs in addition to portfolios. Portfolio evaluation, on the other hand, is undoubtedly the most well-known and significant example of an alternate assessment approach.

A topic that has gained some research attention in the ESL literature is portfolio assessment (Lam, 2017), a well-studied assessment-as-learning strategy (Alam, 2019; Lam, 2020). The portfolio is a planned pupil work collection that demonstrates the student's efforts, development, and achievements in one or more areas of the curriculum, according to Paulson et al. (1991). Portfolio evaluation, which is typically utilized in writing classes, has been shown to support writing improvement self-assessment, and peer assessment (Barrot, 2016; Lam, 2017). In a similar vein, it has a favorable impact on students' autonomy, motivation, and reflective thinking (Lee, 2017; Sultana et al., 2020).

Another example of alternative assessment is self-assessment, a self-monitoring procedure training language learners how to use metacognition (Esteve et al., 2012). This implies that when students evaluate themselves, they control metacognitive processes (Takarroucht, 2021). Metacognitive processes include self-regulation abilities, metacognitive knowledge, and metacognitive experiences (Iwai, 2011). The definition of metacognitive knowledge is the understanding of task demands and approaches. The capacity to identify performance problems through reflection and problem-solving is known as metacognition (Tarricone, 2011). Executing metacognitive methods, including planning, monitoring, and assessing, is a requirement for developing self-regulation skills (Iwai, 2011). A group of higher-order processes known as metacognitive strategies are in charge of identifying performance flaws and carrying out cognitive techniques (Tarricone, 2011). Metacognitive methods are a form of self-regulation.

Sorting through the literature reveals that peer assessment is another instantiation of alternative assessment. Students can discuss their personal performance and academic requirements with their peers using the communication strategy known as peer evaluation. Peer assessment is a type of collaborative learning and formative evaluation that can be utilized in EFL/ESL courses. Peer assessment can enhance students' production abilities by incorporating them into revisions (Zhao, 2010), make learners more interested in production (Shih, 2010), scaffold students' production process, and enhance

critical thinking (Hyland, 2000). Authors are allowed to exhaust their texts and get others' interpretations of them (Joordens et al., 2009). Moreover, peer evaluation might promote learner autonomy (Yang et al., 2006).

In addition to what went above, individual differences are thought to be crucial in language learning and processing (Kidd et al., 2018), and they can reduce or even modify the effects of instruction (Li, 2017). They have been demonstrated to have significant explanatory value when predicting learning outcomes in second or foreign language learning (Pawlak, 2017). One such individual difference is working memory (WM). It is an attentional mechanism with a finite capacity that facilitates sophisticated cognitive processing (Cowan, 2017). WM, according to Baddeley (2017), is a system made up of storage subsystems that are in charge of temporarily storing and processing both verbal and visual-spatial information (the phonological loop and the visuospatial sketchpad, respectively); a domain-general component that is in charge of controlling and regulating attention; and an episodic buffer that acts as a link between the storage subsystems and the episodic buffer. Attention management, analogical reasoning, explicit deduction, information retrieval, and decision-making are just a few of the cognitive processes that the WM is critical to for L2 learning (Tagarelli et al., 2016), as well as the storage of metalinguistic knowledge as L2 language learners comprehend and produce it.

It is impossible to understate the role that reading comprehension plays in academic success. People's lives are significantly impacted by learning to read (Alawajee & Almutairi, 2022). The secret to learning new things and succeeding at work is reading (Castles et al., 2018). According to Seymour (cited in Pallathadka et al., 2022), reading comprehension is the capacity to interpret information from texts. Reading comprehension is a cognitive process that involves deriving meaning from texts, according to Woolley (2011), and it strongly depends on the reader's ability to comprehend written texts accurately and fluently.

It is undeniable that vocabulary is crucial to learning a second language (Kargar Behbahani & Kooti, 2022). According to those that have studied vocabulary, Harmer (2001) considers it to be the language's main organ and its flesh. Furthermore, according to Mediha and Enisa (2014), vocabulary is essential to the communication of any message. Furthermore, Wilkins (1972) believes that a big vocabulary is more crucial than grammar while acquiring an L2. Consequently, learning new words is a crucial component of studying any second or foreign language.

Growing concerns about learners' language accuracy in recent years have led to a reassertion of the importance of grammar in syllabus design and class material, even to the point of paying explicit attention to grammatical forms and rules. It has been essential for English teachers to instruct students in grammar correctly. But as Ellis (1997) emphasized, there are several pedagogical approaches available to language practitioners; the question is how to teach grammar from among them. Grammar instruction always receives significantly greater attention from English teachers at high schools than other language instruction. This is primarily because of the school final exam, which focuses primarily on grammar and uses the pass percentage of the pupils as a measure of the effectiveness of the teachers (Torkabad & Fazilatfar 2014).

Being an English teacher working for Afghanistan's ministry of education, I frequently observe my students' less-than-satisfactory performance on language tests,

both teacher-made exams and high-stakes standardized ones. One contributing factor to this low performance on language tests is the fact that enough time is not dedicated to language instruction in the government-initiated curriculum. Therefore, there is certainly a need to look for alternative possibilities to make the most of the time at hand and ensure learners' language growth.

Despite the plethora of research on the above-mentioned alternative assessment examples and individual differences (i.e., WM capacity), studies investigating the interplay between individual variations and instructional circumstances or approaches are still somewhat rare (Benson & Dekeyser, 2019; Ruiz et al., 2018). Additionally, it is well-acknowledged that vocabulary serves as the foundation of language. Notwithstanding, according to Ritonga et al., (2022), almost no study has ever attempted to investigate the effect of alternative assessment on vocabulary learning. Furthermore, in a world wherein English is seen as the most significant lingua franca, Afghan EFL learners' general English proficiency and particularly their reading comprehension skill is extremely insufficient (Pallathadka et al., 2022). Grammar is given more emphasis in Afghan classrooms than other linguistic skills. This is because high-stake exams in Afghanistan are mostly dependent on grammar. Although grammar is highly valued in Afghan high schools, Afghan EFL students fail to acquire the grammatical features to which they are exposed, hence their grammatical knowledge is inadequate (Patra et al., 2022).

In addition to what has been explained above, numerous researchers have been concerned with the impacts of one of the examples of alternative assessment on language learning. After all, one question that remains is this: what alternative assessment is more facilitative of different language skills or language components? To accomplish it, this investigation seeks to vehemently fill this lacuna, add to the literature, and help language practitioners around the globe understand which alternative assessment is more helpful in developing learners' language abilities. Moreover, to gain a wider perspective into how these different alternative assessment procedures might help language learners sharpen their linguistic skills, the potential role of WM capacity is also investigated to see how this individual difference could mediate language learning through different examples of alternative assessment.

Based on the above explanations, this study has three major objectives. The first objective of this study is to investigate the comparative effect of portfolio assessment, self-assessment, and scaffolded peer assessment on vocabulary learning with WM as an intervening variable. Secondly, the study objectifies to see which of the aforementioned assessment types is more facilitative of reading comprehension with WM as a moderating variable. Finally, this experiment looks into the comparative effect of these kinds of alternative assessments on the grammatical accuracy of language learners across different WM capacities. Therefore, this study looks to find an appropriate answer to the below-mentioned research questions:

- Research question 1: Is there any significant difference between learners receiving portfolio assessment, those receiving self-assessment, and those receiving scaffolded peer assessment on reading comprehension across different WM capacities?

- Research question 2: Is there a noticeable difference in vocabulary learning across various WM capacities between students who receive portfolio evaluation, those who receive self-assessment, and those who receive scaffolded peer assessment?
- Research question 3: In terms of grammatical accuracy across various WM capacities, are there any notable differences between students who receive portfolio evaluation, those who receive self-assessment, and those who receive scaffolded peer assessment?

As mentioned above, language instructors have always been on the lookout to find a panacea for their learners' language growth. Despite the numerous studies which have independently explored the efficacy of the above-cited alternative assessment varieties, to the best of what the researcher knows, no research has ever attempted to compare different alternative assessment strategies to see which is more facilitative of language skills. Additionally, this study also examines WM contributing role to see if individuals with different WM capacities can develop their language skills similarly. The researcher hopes that the results of this study help language teachers understand which strategy is of more help in actual language classrooms. Furthermore, the researcher also hopes that the results of this endeavor pose several theoretical implications for researchers in instructed SLA domain. Besides, this study's findings might help course designers, materials developers, learners, and all stakeholders.

## Literature review

In this section, at first theoretical considerations in alternative assessment at stakes, portfolio assessment, self-assessment, and peer assessment are discussed. Then and only then the experimental studies regarding these instantiations of alternative assessment are dealt with.

### Theoretical underpinnings

#### *Portfolio assessment*

Electronic or printed dossiers containing student-written scripts are called portfolios. These scripts have been selected over time and are often supported by a reflective journal. In the field of education, portfolio assessment is frequently considered to be preferable to the more common, product-focused standardized tests (Kirkpatrick & Gyem, 2012). Numerous studies in second/foreign language (L2) have highlighted the benefits of portfolio assessment in terms of L2 teachers' positive experiences with various types of it (Lee, 2017); the contribution of the portfolio to L2 learners' autonomy, self-regulated learning, social awareness, and metacognitive awareness (Behbahani et al., 2011); and the mediation role of portfolio assessment in revising works-in-progress (Azizi & Namaziandost, 2023; Mphahlele, 2022). Because of the rigidity of L2 teachers (Xu & Brown, 2016), insufficient literacy in language assessment (Gan & Lam, 2020), and low student involvement (Lee & Coniam, 2013), its complex and comprehensive grading (Song & August, 2002), and the test-driven, dominant culture in most educational systems, portfolio assessment has remained highly contentious in actual classroom settings despite the claimed educational benefits (Lam, 2018). As a result, there have been

several difficulties with fully implementing portfolio assessment in L2 contexts, prompting Hyland and Hyland (2019) to ask for more extensive study on these issues.

The process-oriented peer assessment approach to L2 writing redefines it from a pedagogical standpoint as a recursive and metacognitive activity that involves L2 learners in routine reflection on their language development (Lam, 2019). According to Vygotsky's (1987) social constructivism model of learning, second-language learners learn best when they actively create their knowledge of the target language through social interactions rather than just receiving it and serves as the foundation for portfolio-based assessment. The L2 learners' "knowledge of writing as a socially situated practice in academic discourse groups," for instance, is strengthened by writing portfolios (Duff, 2010, p. 169). As a result, it can evaluate the development of L2 writers' higher-level writing abilities (such as textual and discursive writing) as well as their lower-level writing abilities (such as writing mechanics and punctuation) (Steen-Utheima & Hopfenbeck, 2018).

Successful learner engagement, according to Chappuis (2014), depends on how well L2 learners grasp the aims in writing portfolios, how quickly they can visualize the gap between their current situation and those aims, and how to attain the aims. In a similar vein, it is advised that L2 writing instructors foster self-reflection by scaffolding the students through tutorials to the entire portfolio assessment process (Kusuma et al., 2021; Rezai et al., 2023), using examples and prompts (Gregory et al., 2001), extending deadlines to further engage students (Lam, 2020), and disclosing the assessment rubrics to them (Panadero & Romero, 2014).

### ***Self-assessment***

Many language teachers and academics agree that self-assessment and other alternative modes of assessment have received much research and support. Numerous types of research in the area have shown that self-assessment is very important and effective in fostering different language learning techniques and skills as well as in increasing the awareness and motivation required for language acquisition (Birjandi & Hadidi Tamjid, 2010). Self-assessment in particular hence looks appropriate to be included in the language learning curriculum.

To provide a thorough image of what pupils know and need to learn, assessment describes the procedures used to gather, trade, and negotiate data from a variety of important sources (Ebadi & Rahimi, 2019). Bachman et al. (2010) speak of self-assessment when one assesses their work. The technique of self-assessment should therefore be promoted and taught to every learner. The core of self-assessment, according to Locke et al. (1996), is the basic evaluation of one's deservingness, effectiveness, and competence as a person. This idea is a wide, latent, higher-order attribute that includes neuroticism, self-efficacy, and self-esteem.

High levels of self-evaluation enable people to adapt to new circumstances and strive to fulfill their obligations to the best of their abilities (Al-Mamoory & Abathar Witwit, 2021; Jiang et al., 2022). Those with high levels of self-awareness can pause, reflect, and alter their emotional experiences (Putro et al., 2022). To enhance their learning, high-level self-awareness learners control their emotional experiences (Hu, 2022). Eysenck (1990) claimed that CSA can be used as a gauge of emotional stability in this regard. Additionally, self-evaluation promotes students' wellbeing (Jahara et al., 2022). Learners



should exercise their metacognitive skills, critical thinking, affective thinking, self-efficacy beliefs, and academic emotion (Wei, 2020; Zhang, 2022; Davoudi & Heydarnejad, 2020; Khajavy, 2021; Khajavy et al., 2020; Namaziandost & Cakmak, 2020) to implement self-assessment.

### ***Scaffolded peer assessment***

Feedback is the process through which students analyze critiques of their learning and apply them to themselves to become better students (Carless & Boud, 2018). For students to provide constructive critiques and comments on each other's work in an organized learning process, there are two options: peer assessment and peer review. Peer assessment procedures allow for building critical judgment in addition to improving the activities being evaluated (Lipnevich & Smith, 2022; Malecka et al., 2020; Nicol, 2020).

There are various ways that peer assessment can assist learners. First of all, learning through peer assessment can help assessors better their job. In particular, they can improve their knowledge of the project's specifications, evaluation criteria, and topic (Noroozi et al., 2016); produce additional ideas; learn from the work of their peers; and critically evaluate their work (Hsia et al., 2016). Students can gain insight into how to enhance their performance as assessees whose work is evaluated by peers (Hsia et al., 2016). The advantages of obtaining peer feedback are mostly dependent on how useful the feedback is and, more crucially, how effectively pupils apply it. Also, the utilization of feedback has a substantial impact on how well students' final projects turn out.

Unfortunately, pupils lack subject-matter expertise. Some comments might be false or deceptive. Assesseees may become confused when many assessors make conflicting comments (Mostert & Snowball, 2013). Students also doubt their peers' abilities to offer feedback and do not regard them as "knowledge authorities" (Gielen et al., 2010, p. 305). This cynicism can affect assesseees in both good and bad ways. In particular, the skepticism may lead to resistance to peer feedback or a reluctance to follow the recommendations of peer assessors. On the other side, a skeptic's mindset might inspire assesseees to come up with suggestions for improvement (Gielen et al., 2010; Jiangmei, 2023).

Peer learning, which is often referred to as collaborative learning, is based on social constructivism and holds that when learners socially interact with their peers outside of the classroom, learning occurs more actively (Roschelle & Teasley, 1995). Through exchanging personal tales, perceptions, and reflections, students positively rely on one another and aid one another's mental models (Johnson & Johnson, 1987). Members of the group attempt to individually contribute to progress learning and accomplish a group goal in a cooperative learning environment (Johnson et al., 2014). This approach supports students' cooperative knowledge-building (Naserpour & Zarei, 2021). While everyone in the group accepts responsibility for their learning, there is a strong interdependence among them (Bolukbas et al., 2011).

In Sawyer's (2006) work, the help provided during the educational process to meet students' needs when they are introduced to novel concepts and skills is referred to as scaffolding. This could lead to higher and more thorough levels of learning (Naserpour & Zarei, 2021). The zone of proximal development (ZPD), a main idea in socio-cultural theory, and folding have a close association. According to Vygotsky (1987), ZPD is the

difference between a child's actual and anticipated levels of development, which are determined by how well they can manage problems when given direction from adults or more proficient peers (Verenikina, 2008). Scaffolding is the temporary assistance of an expert given to a beginner to boost their independence. This help is gradually lessened or withdrawn as students demonstrate mastery, complete activities on their own, and develop their skills and capabilities (Diaz-Rico & Weed, 2002 cited in Homayouni, 2022).

### **Working memory**

The term “**working memory**” describes the capacity to retain and process data while performing continuous cognitive tasks (Li, 2023). The term “working memory” was first used to refer to a revised understanding of short-term memory as a cognitive resource for concurrent information storage and manipulation as opposed to only a passive storage device. WM is a subject of many studies in SLA because of its alleged impact on the procedure and results of language learning (SLA). Harrington and Sawyer (1992), who looked into the function of WM in text understanding, and Mackey et al. (2002), who looked into the relationships between WM and L2 interaction, are two pioneering research on WM in SLA. Since these landmark findings, interest in the mediating function of WM in numerous facets of L2 learning has steadily increased. Despite the increasing interest in WM in L2 research, there has been a lack of consensus regarding its conceptualization, measurement, and process. This has led to a variety of inconsistent, and occasionally contradictory, results from the research.

Several theories have been put up to explain the connections between the various WM components (Miyake & Shah, 1999; Namaziandost et al., 2022). Two models, the multi-componential model, and the unitary model serve as the fundamental representations of these theories. Baddeley (2017) promoted the multi-component model, which divides WM into four parts: the central executive, the phonological loop, the visual-spatial sketchpad, and the episodic buffer. According to Baddeley (2017), the central executive coordinates across various components focuses and shifts attention, allocates resources, and communicates with long-term memory. A passive storage system for keeping and practicing auditory information is the phonological loop. It is a tool for acquiring vocabulary and is crucial for learning new vocabulary, not just random correlations between well-known words. For storing and practicing knowledge in the form of pictures, shapes, colors, directions, places, and their arrangements, turn to the visual-spatial sketchpad. The episodic buffer serves as a temporary storage area for combining discrete information bits into larger units, connecting short-term and long-term memory, and connecting data from various sources and data in various formats.

The reading span test that Daneman and Carpenter (1980) devised, which simultaneously examines the storing and processing components, is where the unitary model's North American origins may be found. The storage and processing tasks in this architecture are interdependent and share the same resource pool. The storage and processing operations are trade-offs, thus providing more resources to one will result in fewer resources for the other. The unitary model states that executive control, despite playing a significant role, as well as storage alone, such as phonological loop and visuospatial sketch pad, cannot describe WM.



## Experimental underpinnings

### *Portfolio assessment*

As an example of alternative assessment, numerous researchers have investigated the efficacy of portfolio-based assessment of language growth. For example, Barrot (2021) looked into the impacts of e-portfolio on ESL learners' writing. Eighty-nine L2 English speakers from four English classrooms participated in the study. An e-portfolio was used by two classes in the treatment group ( $N=48$ ), whereas a traditional portfolio was utilized by the other two classes in the control group ( $N=41$ ). Findings showed that e-portfolio learners outstripped the traditional portfolio group. These outcomes were linked to the e-portfolio's flexible, accessible, interactive capabilities, and its capacity to expose learners to peer pressure.

Another research that has studied the potential of portfolio assessment in reading comprehension in an EFL setting is that of Amani and Salehi (2017). Their study objectified to evaluate the effects of the portfolio as a descriptive evaluation technique on the growth of Iranian EFL students' text understanding skills using Prospect 2 as the foundational text. To achieve this, 20 female EFL students from an Iranian guidance school were chosen. Members of the experimental group received the portfolio assessment, whereas the control group members received the traditional assessment. The students in both groups took two text understanding assessments as a pretest and post-test to gauge their level of reading comprehension before and after the intervention. Descriptive and inferential statistical techniques were used to conduct the statistical study. The results did not demonstrate that the portfolio was superior to the traditional scoring method in helping children develop their reading comprehension.

In another research, Nourdad and Banagozar (2022) examined the potential role of e-portfolio evaluation on vocabulary learning and retention. Ninety-two guidance schools were chosen as the study's subjects to achieve this goal. They were split into two experimental and control groups at random. The experimental group practiced e-portfolio evaluation while the control group adhered to the traditional in-class quizzes. The experimental group's members were instructed to make their e-portfolios and keep a log of the lessons they learnt both during and after the online sessions. Also, it was requested that they upload the reflection sheets to their e-portfolios. To collect information regarding the impact of portfolio assessment in each grade, three parallel tests were used: a pretest, an immediate post-test, and a delayed post-test (a total of nine tests). The treatment participants outstripped the control condition in terms of acquisition and retention of EFL vocabulary, according to the findings of a one-way ANCOVA.

Examining the efficacy of portfolio-based assessment in language growth is not restricted to the above-mentioned studies. In newly published research, Rezai et al., (2022a, 2022b) wondered whether e-portfolio assessment can cultivate EFL learners' vocabulary, motivation, and attitudes. After homogenizing 100 EFL male students for this project, 50 were randomly assigned to the experimental group and 50 were placed in the control group. Following that, they completed the pretest, interventions, and post-test procedures. Eighteen 1-h sessions were held twice a week, and the experimental group received their training using e-portfolios, whereas the control group received their training through more traditional means. Using the use of an independent-sample  $t$  test, mean calculations, and percent calculations, the acquired data were examined.

The post-test results showed that the experimental group fared better than the control group in terms of vocabulary knowledge improvements. The results also showed that in terms of motivation after the interventions, there was a significant difference between the two groups. The results also demonstrated that the participants' sentiments about the e-portfolios were quite favorable.

### ***Self-assessment***

Although numerous studies on the impacts of self-assessment on L2 learning have been undertaken over the past 10 years, none has looked into how self-assessment reports affect L2 learning. It was for this reason that Rezai et al., (2022a, 2022b) sought to examine Iranian teenagers' perceptions of the efficiency of self-assessment reports in developing writing skills as well as how self-assessment reports enhance their writing abilities. The researchers chose one whole grade 11 class for this study. A self-assessment report based on Nunan's (2004) template was created and distributed to the students to help them evaluate their writing each week during the 15 sessions of instruction, which were held twice a week. Six students were used in a focus group interview that followed. The students' writing abilities in terms of content, language, and organization showed considerable improvement, according to the findings. The focus group interview results also revealed four themes: improving students' understanding of evaluation standards, fostering greater self-control, giving students a say in their academic futures, and boosting students' writing drive.

The effects of self-evaluation, planning, goal-setting, and reflection on students' self-efficacy and writing performance before and after revision were examined by Chung et al. (2021). Their findings revealed that the treatment condition had significantly improved on the post-test in terms of writing performance. In addition, they discovered that participants' self-efficacy changed dramatically from before to after the revision.

One alternate method for gauging students' English-speaking prowess is self-evaluation. Students are allowed to learn about, practice, and improve their speaking skills through this evaluation. Nonetheless, it is unlikely that projects of this nature were typical throughout Indonesia. Alek et al. (2020) wanted to understand how pupils at Link and Match vocational high school felt about using self-assessment to evaluate their speaking performance. Five items about the use of self-assessment were included in the questionnaire used to collect the data for this study. The data in this qualitative study had undergone a descriptive analysis. Thirty students from vocational high schools who were majoring in multimedia were included in this study. The majority of students believed that self-evaluation was highly beneficial since it helped them understand their functional capabilities and how to improve them to meet course objectives, particularly the speaking course objective. Furthermore, some students believed that self-assessment was very helpful because the teacher did not frequently utilize this assignment and the students did not enjoy trying to evaluate themselves. These researchers concluded that to investigate and evaluate pupils' speaking abilities, self-assessment is highly helpful.

### ***Peer assessment***

Peer assessment has been more prevalent in classrooms and other learning environments in recent years. Despite the widespread belief that peer assessment improves

learning across empirical investigations, the outcomes are conflicting. Li et al. (2020) combined findings based on 134 impact sizes from 58 trials in a meta-analysis. The performance of peer assessment learners is improved by 0.291 standard deviation units as compared to those who do not. They also conducted a meta-regression study to look at the variables that may affect the peer assessment effect. The most important element is rating system training. Peer assessment effect size is significantly greater when students have received rater training than when they have not. Peer assessment that is computer-mediated rather than paper-based is also linked to larger learning gains. Other factors (including rating format, rating standards, and peer assessment frequency) also have observable effects but are not statistically significant. Finally, these L2 researchers suggested that researchers and educators can use the findings of the meta-analysis as a guide to decide how to use peer evaluation as a learning tool effectively.

In another study, Moghimi (2022) explored the comparative effects of peer assessment and self-assessment, and gender on Iranian EFL learners' accuracy in speech. Based on the Quick Oxford Placement Exam, 60 homogeneous were chosen. An OQPT, peer, and self-assessment questionnaires served as the study's tools. To calculate the results, SPSS version 20 was used. The means were similar, but the male students' mean score was slightly higher than the female students. Furthermore, assessment types had a substantial impact on speech accuracy performance and that peer assessment was superior to self-assessment in this area.

Another study that has dealt with the efficacy of peer assessment coupled with scaffolding on oral skills and lexical growth is that of Homayouni (2022). The researcher chose 5 intermediate English learners and 37 lower-intermediate English learners through cluster sampling to achieve this goal. Then, 5 more proficient students and 20 lower-intermediate participants were assigned at random to the experimental group. The intermediate learner was given the role of the mediator in groups of 5, and they were in charge of providing feedback to their peers. There was no mediator assigned to the control group, which included the remaining individuals. Throughout four training sessions, both the scaffolded peer assessment of speaking and vocabulary learning was conducted. A one-way repeated measures ANOVA and an independent sample *t* test were performed in this randomized pre-test-post-test-delayed post-test trial. The outcomes of the statistical analysis showed that scaffolded peer evaluation had a significant positive influence on learners' vocabulary growth and speaking ability. That is, both speaking abilities and vocabulary knowledge can be developed by using scaffolded peer assessment in a group-oriented setting. The study's pedagogical implications suggest that language instructors can use the sociocultural theory and social constructivism concepts put out by Vygotsky (1987) to widen and deepen students' ZPD.

### **Working memory**

As an individual difference trait, WM is claimed to mediate language learning. To verify this claim, Chow et al. (2021) investigated the roles of reading anxiety and WM in text understanding among Chinese EFL students. There were 105 Chinese ESL undergraduates altogether. The results revealed that verbal WM and reading anxiety, as reflected by reading traits and state anxiety, were the only two independent predictors of ESL reading comprehension. Moreover, there was no discernible connection between reading

anxiety and WM. The association between verbal WM and ESL reading comprehension was found to be somewhat mediated by reading anxiety, according to mediation analyses. These findings provide insight into the strategies for improving ESL learning and emphasize the significance of affective and cognitive components in determining ESL text grasping.

In another study, Teng and Zhang (2021) purported to investigate how WM functions in vocabulary learning with multimedia input. They focused on the potential connections between executive WM and phonological short-term memory (PSTM), as well as the effects of three different input conditions (definition + word information + video, definition + word information, and definition) on the acquisition of vocabulary in a second language (L2). Ninety-five students in all completed the three learning scenarios and passed the two WM tests: the reading span exam, which assesses complex executive WM, and the non-word span test, which evaluates PSTM. They tested both receptive and productive vocabulary knowledge both at the beginning and end of the 2 weeks. Based on repeated-measures analysis of covariance (ANCOVA), our results show that complex and phonological WM plays a significant role in vocabulary learning and retention under the three conditions. They also showed that the definition + word information + video condition has pronounced effects on vocabulary learning and retention.

In another study, Patra et al. (2022) looked at how learning English future tense was impacted by processing instruction (PI) and output-based activities, with WM serving as a mediating factor. To achieve this, 99 participants with pre-intermediate English proficiency as determined by the Oxford Placement Test were chosen for the study. They were split into three groups, each of which contained 33 learners: PI, output, and control. Utilizing a reading-span test, it was discovered that only 14 of the PI group's subjects, 15 of the output group's participants, and 13 of the comparison group's students had poor WM levels, while the other participants had strong WM levels. Then, a Bonferroni adjustment post hoc test and a two-way between-group analysis of variance were carried out. The analysis' findings demonstrated that the output and PI groups both outstripped the control group. The grammatical gain between the PI and output groups was also the same. Moreover, students with high WM did better than those with low WM. These L2 researchers concluded that output-based learning activities and PI can help teachers adopt powerful tactics to increase the knowledge and awareness of L2 learners.

All in all, the abovementioned studies point to the efficacy of portfolio assessment, self-assessment, and peer assessment. However, sorting through the literature reveals that there remains a paucity of research examining the comparative effects of these types of assessment on language development. Among the studies cited above, only Moghimi (2022) examined the comparative effects of peer assessment and self-assessment on learners' accuracy in speaking. One study is not enough in making sure whether peer assessment is superior to self-assessment. Additionally, to the best of what the researcher knows, no study has ever attempted to examine the mediating role of WM on the effects of different types of alternative assessment in language development. It is for these reasons that this study attempts to fill the gap and comparatively examine the effects of portfolio assessment, self-assessment, and scaffolded peer assessment on reading comprehension, vocabulary learning, and grammatical accuracy in an EFL setting. The researcher hopes that the results gleaned from this study will add to the literature,

fill a knowledge gap, help language teachers assist in language development in their learners, and help material designers how to design better textbooks.

## Method

In this section, the study's design, setting and subjects, instruments, data collection procedure, and method of data analysis are discussed in detail.

### Design

Since it was impossible for the researcher to randomly select the participants of the study, a quasi-experimental pretest-post-test control design (Ary et al., 2019) was employed in this current quantitative investigation. Four groups participated in this exploration: three treatment groups and a control group. The experimental groups included a portfolio group, a self-assessment group, and a peer assessment group. The variables of the study include an independent variable (i.e., type of treatment) with four levels discussed just above, three dependent variables (i.e., scores on tests of reading comprehension, vocabulary, and grammar), along with a moderating variable (WM capacity). It needs to be mentioned that learners' reading comprehension, vocabulary growth, and grammatical accuracy were checked on two occasions, once before the treatment (pretest), and once right after the treatment (post-test).

### Setting and participants

A hundred and twenty-five students studying English at a private language institute in Kandahar, Afghanistan, participated in this study. They were chosen for the study through convenient sampling. This sample was chosen out of 172 subjects. To be more specific, through an Oxford Quick Placement Test (OQPT), 120 subjects with lower-intermediate command of English, and five learners with intermediate level were chosen. The philosophy behind selecting the higher-intermediate learners was to assign more proficient learners in the peer assessment condition to serve as the mediator in the group. The participants were between the ages of 15 and 19. All participants in this study had Persian as their L1 with English serving as their target language. The subjects who had been selected were then assigned to four conditions: portfolio condition ( $N=30$ ), self-assessment condition ( $N=30$ ), peer assessment condition ( $N=35$ ), and control condition ( $N=30$ ), with 30 subjects in each. According to the results of the reading-span test (to be discussed in the following section), 16 subjects in the portfolio condition, 14 learners in the self-assessment group, 18 participants in the peer assessment group, and 13 participants in the control condition had high WM, while the rest of the participants had low WM. Additionally, a signed consent form was taken from all the participants before the research. For students below the legal age of 18, their parents were asked to sign the form.

### Instruments

At the beginning of the research, the researcher functioning as the teacher of the classrooms used an OQPT to determine the subjects' proficiency level. Thereafter, the researcher developed three instructor-made tests of reading comprehension, vocabulary knowledge, and grammar. Tests of vocabulary and text comprehension were based

on Focus on Vocabulary 1: Bridging vocabulary designed by Schmitt et al. (2011). Furthermore, the grammar test was based on Oxford Living Grammar (pre-intermediate level) designed by Harrison (2009). Furthermore, to check participants' WM capacity, a reading-span test developed and validated by Shahnazari (2013) was used. This measure of WM is a test in which testees need to read the sentences and make a judgment on whether the sentences are grammatically plausible. Additionally, they need to memorize the last word of each sentence. According to Shahnazari (2013), the number of words each examinee can recall constitutes their WM span. Because the researcher himself designed the items of the tests based on the aforementioned textbooks, these instructor-made tests were adopted by him, while the OQPT and the reading-span test were adapted for the study. To make sure of the validity and reliability of the adopted instruments certain procedures were undertaken. First of all, to construct and validate the instruments, the researcher used the known-group technique (Ary et al., 2019). In this group differential strategy, the researcher administered the adopted instruments to a group of English language teachers who knew the answers to the items. The difference between their performance and those of the participants at the pretest turned out to be statistically significant based on the independent sample *t* test results at  $p < 0.05$ , hence the validity of the instruments. Moreover, to check the reliability of the instruments, using SPSS software, alpha Cronbach's value was determined which turned out to be 0.76 verifying the reliability of the instruments. These adopted instruments had multiple-choice, fill-in-the-blank, and open-ended items. In addition to this, two versions of each instrument were adopted. A version was administered at the onset of the study (i.e., pretest), and another version with similar in form but with different items at the end of the treatment (i.e., post-test). It should not be forgotten that this study targeted the present continuous linguistic features. Furthermore, as far as the validity of the portfolio-assessment instrument is concerned, according to Lynch (2001), to have a valid portfolio instrument, we need fairness and consequential validity. Thus, learners were allowed to select the materials of their choice from among the submitted materials to raise the fairness of the instrument. Additionally, if it turns out that the participants in the portfolio assessment can gain the materials, the consequential validity of the instrument is automatically confirmed.

#### **Data collection procedure**

First of all, an OQPT was administered for the research to come up with a homogenized sample. For this study, based on the OQPT results only lower-intermediate learners of English were selected along with five higher-intermediate learners of English. These lower-intermediate learners were assigned to four conditions: a portfolio condition, a self-assessment condition, a scaffolded peer assessment condition, and a control group. In addition, the higher-intermediate learners were injected into the peer assessment group to function as the group's head and mediator. To be more specific, the participants in the scaffolded peer assessment condition were divided into five groups each with six learners, along with a higher-intermediate learner as the mediator. Then, the first version of the instructor made tests of reading comprehension, lexis, and grammar were given. After that, the researcher administered an adapted reading-span test discussed above to determine the participants' WM span. Thereafter, the treatment began. In a treatment



that lasted 10 sessions, the first two sessions were devoted to the administration of the OQPT, pretest, and reading-span test. The researcher decided to split the treatment into three halves. In the first phase, the researcher gave students enough guidance on how to choose, gather, and reflect on their activities in their portfolios as well as complete the self-assessment checklists, so they could become more independent and autonomous in their reading comprehension, lexical expansion, and grammatical accuracy.

In the first phase, the students in the portfolio and self-assessment conditions were given instructions during the first two instructional sessions. One assignment was due in the classroom, and the other was due outside the classroom, both on different subjects. To keep track of their tasks in chronological order, they created files. The researcher corrected the students' work using the checklists each session and addressed the substance of them in the class along with individual conferences because the researcher discovered that self-assessment using checklists requires comprehensive teaching. Students believed they could use the checklist to self-evaluate their papers after four weeks of teaching. Based on the qualitative observations, they improved in self-correction starting with the fifth instructional session.

Students improved in the second phase at using the checklist to self-evaluate their work. Except for some of the learners who required additional assistance, the teacher opted to reduce and eventually discontinue the teacher-student conferences. Nearly all of the students had the opportunity to self-evaluate their work throughout the second half of the treatment, complete the checklists, and add the papers to their portfolios for instructor random inspection. Following that, the researcher reviewed the pupils' portfolios every other session and noted the comments in the checklists for the portfolios. This allowed both the students and the teacher to reflect on all of the activities that were documented in the portfolio.

In the third phase and the scaffolded peer assessment condition, group participants were divided into different groups with a more proficient learner selected for each group to function as the head and mediator. Then, the instructional materials were given to the participants. In this cooperative scaffolded type of alternative assessment, attempts were made to develop learners' ZPD. That is, attempts were made to help learners do something under the guidance of a more proficient peer (i.e., mediator) that they could not do on their own. In this experimental condition, under the teacher's guidance, the mediators provided mediation to their peers. In other words, peers evaluated the comments produced by their buddies and advised on how those buddies can fix their inaccurate responses. This procedure was repeated in every session until the treatment finished.

At the last session of the treatment, the post-test was given, and learners' scores on both the pre and post-test were statistically compared using SPSS software which allowed the researcher to conduct statistical tests of significance.

### **Data analysis**

To perform tests of statistical significance, the researcher resorted to the SPSS software. At first, because the researcher needed to ensure the normal distribution of the data, a one-way Kolmogorov–Smirnov (K-S) test was conducted. Then, to check the effects of the treatment concerning the mediating role of WM, three two-way between-group

**Table 1** One-sample Kolmogorov–Smirnov test

		Condition	WM	Reading pretest score	Reading post-test score
<i>N</i>		125	125	125	125
Normal parameters	Mean	2.520	1.512	3.088	8.768
	Std. Deviation	1.104	.501	1.237	5.143
Most extreme differences	Absolute	.188	.347	.168	.121
	Positive	.161	.334	.168	.121
	Negative	–.188	–.347	–.129	–.111
Kolmogorov–Smirnov <i>Z</i>		2.103	3.875	1.882	1.350
Asymp. Sig. (2-tailed)		.130	.164	.096	.126

**Table 2** Levene's test of equality of error variances

	<i>F</i>	df1	df2	Sig
Reading pretest score	.311	7	117	.248
Reading post-test score	2.942	7	117	.307

MANOVAs were carried out. Post hoc tests will also be conducted to check the interaction effects.

## Results

The study's questions are attempted to be statistically analyzed in this section.

Research question 1: Is there any significant difference between learners receiving portfolio assessment, those receiving self-assessment, and those receiving scaffolded peer assessment on reading comprehension across different WM capacities?

In this research question, there is an independent variable (i.e., type of assessment) with three levels (i.e., portfolio assessment, self-assessment, and scaffolded peer assessment), a mediating variable (i.e., WM capacity) with two levels, and two interval-dependent variables (i.e., pre- and post-test scores on a reading comprehension test). In such a scenario, one needs to run two-way between-group MANOVA (Rezai, 2015). However, this test of statistical significance has some assumptions. Firstly, we need to make sure whether the data are normally distributed. Thus, a one-sample K-S test must be performed (Pallant, 2020).

Table 1 presents the results of a one-sample K-S test. As Table 1 shows, the Sig. (2-tailed) value in all four sub-parts of the table exceeds 0.05, so the normality assumption is confirmed. Now, we need to ensure the homogeneity assumption (Pallant, 2020). To ensure the homogeneity assumption, one needs to run Levene's test of equality of error variances (Rezai, 2015).

As Table 2 demonstrates, the *p* value regarding reading comprehension on both pre- and post-test exceeds 0.05; thus, the homogeneity assumption is confirmed. Now, we can safely carry out the MANOVA.

Table 3 presents the descriptive statistics regarding subjects' performance in all conditions on both pre- and post-test of reading comprehension. According to the table, in

**Table 3** Descriptive statistics

	Condition	WM	Mean	Std. deviation	N
Reading pretest score	Portfolio assessment	High	2.937	1.340	16
		Low	3.500	1.224	14
		Total	3.200	1.297	30
	Self-assessment	High	3.000	1.358	14
		Low	3.375	1.147	16
		Total	3.200	1.242	30
	Peer assessment	High	3.055	1.349	18
		Low	2.764	1.251	17
		Total	2.914	1.291	35
	Control	High	3.076	1.320	13
		Low	3.058	1.028	17
		Total	3.066	1.142	30
Reading post-test score	Portfolio assessment	High	3.016	1.310	61
		Low	3.156	1.171	64
		Total	3.088	1.237	125
	Self-assessment	High	11.125	4.209	16
		Low	8.928	3.604	14
		Total	10.100	4.028	30
	Peer assessment	High	10.000	5.023	14
		Low	8.687	2.868	16
		Total	9.300	4.001	30
	Control	High	14.888	4.701	18
		Low	7.705	5.108	17
		Total	11.400	6.049	35
	Total	High	3.692	1.548	13
		Low	3.941	1.519	17
		Total	3.833	1.510	30
	Total	High	10.393	5.710	61
		Low	7.218	4.001	64
		Total	8.768	5.143	125

the portfolio group, high WM spanners at 1.34 SD, had 2.93 as the mean, while learners with low WM had 3.5 at 1.22 SD. In the self-assessment group, learners with high WM scored 3.00 as the mean at 1.35 SD, whereas low WM learners scored 3.27 as the mean at 1.14 SD. In the peer assessment condition, high WM learners scored 3.05 as the mean with 1.34 SD, while low WM spanners scored 2.76 as the mean at 1.29 SD. High WM subjects in the control group had 3.07 as their mean with 1.32 SD, whereas low WM participants in the same condition had 3.05 as their mean with 1.02 SD. The table also summarizes the results of the post-test. Based on the table, in the portfolio condition, learners with high WM scored 11.12 as the mean with 4.20 SD, and low WM participants scored 8.92 as the mean with 3.60 SD. Besides, in the self-assessment condition, high WM learners, scored 10.00 as the mean with 5.02 SD, while learners with low WM scored 8.68 as the mean with 2.86 SD. High WM learners in the scaffolded peer assessment group had 14.88 as their mean with 4.70 SD, while their low WM counterparts had 7.70 as their mean with 5.10 SD. Furthermore, in the control group, high WM subjects, had 3.69 as their mean with 1.54 SD, and low WM learners had 3.94 means with 1.51 SD.

Overall, the table shows that on the pretest, high WM participants had a 3.01 mean with 1.31 SD, and low WM subjects had a 3.15 mean at 1.17 SD. On the post-test, these numbers rose dramatically such that high WM subjects had a 10.39 mean with 5.71 SD, and low WM learners had a 7.21 mean with 4.00 SD.

Table 4 presents tests of between-subject effects. According to this above-presented table, on the pretest of reading comprehension, at 3 degrees of freedom and with  $F=0.408$ , the difference between groups was not statistically significant ( $p=0.748$ ). The table further reveals that on the post-test, at 3 degrees of freedom with  $F=22.421$ , the difference between conditions was statistically significant at  $p<0.05$  with a large effect size (partial eta squared = 0.365). Concerning subjects' WM capacity on the pretest, at 1 degree of freedom with  $F=0.485$ , no statistical difference between subjects was found ( $p=0.523$ ). However, on the post-test, at 1 degree of freedom with  $F=14.116$ , there was a statistical difference between subjects with a moderate effect size ( $p<0.05$ , partial eta squared = 0.108). Concerning the interaction between condition and WM capacity on the pretest, at 3 degrees of freedom with  $F=0.573$ , there was not a statistical difference between conditions as the  $p$  value exceeds the threshold level 0.05; however, not the post-test, at 3 degrees of freedom with  $F=5.714$ , a statistical difference was observed at  $p<0.05$  with a moderate effect size (partial eta squared = 0.128).

Table 5 reveals pairwise comparisons between groups based on the Bonferroni adjustment test. According to the table, on the pretest, the difference between portfolio assessment and self-assessment was not statistically significant (mean difference = 0.031,  $p>0.05$ ). Additionally, the difference between portfolio assessment and scaffolded peer assessment did not turn out to be significant (mean difference = 0.309,  $p>0.05$ ). Furthermore, there was no statistical difference between the portfolio assessment group and the control condition (mean difference = 0.151,  $p>0.05$ ). The table further discloses that no group had a statistical difference with the control condition of the pretest ( $p>0.05$ ). However, on the post-test, post hoc analyses reveal that the

**Table 4** Tests of between-subject effects

Source	Dependent variable	Type III sum of squares	df	Mean square	F	Sig	Partial Eta squared
Corrected model	Reading pretest score	5.977	7	.854	.543	.800	.031
	Reading post-test score	1535.138	7	219.305	14.703	.000	.468
Intercept	Reading pretest score	1184.362	1	1184.362	752.874	.000	.865
	Reading post-test score	9183.254	1	9183.254	615.678	.000	.840
Condition	Reading pretest score	1.925	3	.642	.408	.748	.010
	Reading post-test score	1003.283	3	334.428	22.421	.000	.365
WM	Reading pretest score	.763	1	.763	.485	.488	.004
	Reading post-test score	210.543	1	210.543	14.116	.000	.108
Condition WM	Reading pretest score	3.552	3	1.184	.753	.523	.019
	Reading post-test score	255.674	3	85.225	5.714	.001	.128
Error	Reading pretest score	184.055	117	1.573			
	Reading post-test score	1745.134	117	14.916			
Total	Reading pretest score	1382.000	125				
	Reading post-test score	12,890.000	125				
Corrected total	Reading pretest score	190.032	124				
	Reading post-test score	3280.272	124				

**Table 5** Pairwise comparisons

Dependent variable	(I) Condition	(J) Condition	Mean difference (I-J)	Std. error	Sig.	95% Confidence interval for difference	
						Lower bound	Upper bound
Reading pretest score	Portfolio assessment	Self-assessment	.031	.325	1.000	-.840	.902
		Peer assessment	.309	.312	1.000	-.530	1.147
		Control	.151	.326	1.000	-.723	1.025
	Self-assessment	Portfolio assessment	-.031	.325	1.000	-.902	.840
		Peer assessment	.277	.312	1.000	-.561	1.116
		Control	.120	.326	1.000	-.754	.994
	Peer assessment	Portfolio assessment	-.309	.312	1.000	-1.147	.530
		Self-assessment	-.277	.312	1.000	-1.116	.561
		Control	-.158	.314	1.000	-1.000	.684
	Control	Portfolio assessment	-.151	.326	1.000	-1.025	.723
		Self-assessment	-.120	.326	1.000	-.994	.754
		Peer assessment	.158	.314	1.000	-.684	1.000
Reading post-test score	Portfolio assessment	Self-assessment	.683	.999	1.000	-1.999	3.365
		Peer assessment	-1.271	.962	1.000	-3.853	1.312
		Control	6.210	1.003	.000	3.519	8.901
	Self-assessment	Portfolio assessment	-.683	.999	1.000	-3.365	1.999
		Peer assessment	-1.954	.962	.268	-4.536	.629
		Control	5.527	1.003	.000	2.836	8.218
	Peer assessment	Portfolio assessment	1.271	.962	1.000	-1.312	3.853
		Self-assessment	1.954	.962	.268	-.629	4.536
		Control	7.481	.966	.000	4.889	10.073
	Control	Portfolio assessment	-6.210	1.003	.000	-8.901	-3.519
		Self-assessment	-5.527	1.003	.000	-8.218	-2.836
		Peer assessment	-7.481	.966	.000	-10.073	-4.889

mean difference between portfolio assessment and self-assessment is also non-significant (mean difference = 0.683,  $p > 0.05$ ), the difference between portfolio assessment and scaffolded peer assessment is also non-significant (mean difference = -1.271,  $p > 0.05$ ), and the difference between self-assessment and scaffolded peer assessment is also non-significant (mean difference = -1.954,  $p > 0.05$ ). A further inspection of the table shows that the difference between all three experimental conditions and the control group turns out to be statistically significant ( $p < 0.05$ ) (Table 6).

Further post hoc analyses based on the Bonferroni adjustment test reveal that on the post-test in the portfolio assessment condition, high WM learners had a higher mean than their low WM counterparts (mean difference = 2.196). In the self-assessment condition, high WM subjects had also a higher mean than their low WM peers (mean difference = 1.312). In the scaffolded peer assessment condition, learners with high WM had an amazingly higher mean than their low WM peers (mean difference = 7.183). However, in the control condition, low WM learners had a higher mean than their high WM peers (mean difference = -0.249) (Table 7).

**Table 6** Condition (WM)

Dependent variable	Condition	WM	Mean	Std. error	95% Confidence interval	
					Lower bound	Upper bound
Reading pretest score	Portfolio assessment	High	2.938	.314	2.317	3.558
		Low	3.500	.335	2.836	4.164
	Self-assessment	High	3.000	.335	2.336	3.664
		Low	3.375	.314	2.754	3.996
	Peer assessment	High	3.056	.296	2.470	3.641
		Low	2.765	.304	2.162	3.367
	Control	High	3.077	.348	2.388	3.766
		Low	3.059	.304	2.456	3.661
	Portfolio assessment	High	11.125	.966	9.213	13.037
		Low	8.929	1.032	6.884	10.973
Reading post-test score	Self-assessment	High	10.000	1.032	7.956	12.044
		Low	8.688	.966	6.775	10.600
	Peer assessment	High	14.889	.910	13.086	16.692
		Low	7.706	.937	5.851	9.561
	Control	High	3.692	1.071	1.571	5.814
		Low	3.941	.937	2.086	5.796

**Table 7** Pairwise comparisons

Dependent variable	(I) WM	(J) WM	Mean difference (I-J)	Std. error	Sig	95% Confidence interval for difference	
						Lower bound	Upper bound
Reading pretest Score	High	Low	-.157	.226	.488	-.604	.290
	Low	High	.157	.226	.488	-.290	.604
Reading post-test score	High	Low	2.611	.695	.000	1.235	3.987
	Low	High	-2.611	.695	.000	-3.987	-1.235

In addition to what went above, further pairwise comparisons concerning WM capacity reveal that the mean difference between high and low WM learners on the pretest was not statistically significant (mean difference = 0.157,  $p > 0.05$ ); however, on the post-test, the difference turned out to be significant (mean difference = 2.611,  $p < 0.05$ ). Additionally, calculations by hand revealed that the effect size was moderate (partial eta squared = 0.095).

Research question 2: Is there a noticeable difference in vocabulary learning across various WM capacities between students who receive portfolio evaluation, those who receive self-assessment, and those who receive scaffolded peer assessment?

In this scenario, similar independent and moderating variables as the first research question is at work. The only difference is that in this scenario, instead of scores on a reading comprehension test, the research deals with scores on a vocabulary test on two occasions (pre- and post-test scores) as the dependent variables. Thus, a further two-way between-group MANOVA needs to be conducted (Pallant, 2020). The two assumptions of normality and homogeneity were checked through a one-sample K-S test, and Levene's test of equality of variances, respectively. However, due to space limitations, their



respective tables are not represented here. The results showed that the sig. (2-tailed) for both tests exceeded the threshold level of 0.05, hence the conformation of normality and homogeneity assumption. Now, there is room to conduct the MANOVA.

Table 8 presents the descriptive statistics regarding subjects' performance in all conditions on both pre- and post-tests of vocabulary. According to the table, in the portfolio group, high WM spanners at 1.600 SD, had 3.187 as the mean, while learners with low WM had 3.785 at 2.044 SD. In the self-assessment group, learners with high WM scored 3.214 as the mean at 1.625 SD, whereas low WM learners scored 3.812 as the mean at 1.558 SD. In the peer assessment condition, high WM learners scored 3.277 as the mean with 1.447 SD, while low WM spanners scored 3.235 as the mean at 1.200 SD. High WM subjects in the control group had 3.538 as their mean with 1.391 SD, whereas low WM participants in the same condition had 3.588 as their mean with 1.175 SD. The table also summarizes the results of the post-test. Based on the table, in the portfolio condition, learners with high WM scored 11.375 as the mean with 3.896 SD, and low WM participants scored 9.214 as the mean with 3.533 SD. Besides, in the self-assessment condition,

**Table 8** Descriptive statistics

	Condition	WM	Mean	Std. deviation	N
Vocabulary pretest score	Portfolio assessment	High	3.187	1.600	16
		Low	3.785	2.044	14
		Total	3.466	1.814	30
	Self-assessment	High	3.214	1.625	14
		Low	3.812	1.558	16
		Total	3.533	1.591	30
	Peer assessment	High	3.277	1.447	18
		Low	3.235	1.200	17
		Total	3.257	1.313	35
	Control	High	3.538	1.391	13
		Low	3.588	1.175	17
		Total	3.566	1.250	30
	Total	High	3.295	1.487	61
		Low	3.593	1.487	64
		Total	3.448	1.488	125
Vocabulary post-test score	Portfolio assessment	High	11.375	3.896	16
		Low	9.214	3.533	14
		Total	10.366	3.828	30
	Self-assessment	High	10.571	4.847	14
		Low	9.312	2.242	16
		Total	9.900	3.679	30
	Peer assessment	High	15.333	3.613	18
		Low	8.352	4.581	17
		Total	11.942	5.379	35
	Control	High	3.769	1.535	13
		Low	3.705	1.649	17
		Total	3.733	1.574	30
	Total	High	10.737	5.479	61
		Low	7.546	3.919	64
		Total	9.104	4.990	125

high WM learners scored 10.571 as the mean with 4.847 SD, while learners with low WM scored 9.312 as the mean with 2.242 SD. High WM learners in the scaffolded peer assessment group had 15.333 as their mean with 3.613 SD, while their low WM counterparts had 8.352 as their mean with 4.581 SD. Furthermore, in the control group, high WM subjects, had 3.769 as their mean with 1.535 SD, and low WM learners had 3.705 means with 1.64 SD. Overall, the table shows that on a pretest, high WM participants had a 3.291 mean with 1.487 SD, and low WM subjects had a 3.593 mean with 1.487 SD. On the post-test, these numbers rose dramatically such that high WM subjects had a 10.737 mean with 5.479 SD, and low WM learners had a 7.546 mean with 3.919 SD.

Table 9 presents tests of between-subject effects. According to this above-presented table, on the pretest of vocabulary knowledge, at 3 degrees of freedom and with  $F=0.269$ , the difference between groups was not statistically significant ( $p=0.848$ ). The table further reveals that on the post-test, at 3 degrees of freedom with  $F=32.696$ , the difference between conditions was statistically significant at  $p<0.05$  with a large effect size (partial eta squared=0.456). Concerning subjects' WM capacity on the pretest, at 1 degree of freedom with  $F=1.223$ , no statistical difference between subjects was found ( $p=0.271$ ). However, on the post-test, at 1 degree of freedom with  $F=17.655$ , there was a statistical difference between subjects with a moderate effect size ( $p<0.05$ , partial eta squared=0.131). Concerning the interaction between condition and WM capacity on the pretest, at 3 degrees of freedom with  $F=0.410$ , there was no statistical difference between conditions as the  $p$  value exceeds the threshold level 0.05; however not the post-test, at 3 degrees of freedom with  $F=6.396$ , a statistical difference was observed at  $p<0.05$  with a large effect size (partial eta squared=0.140).

Table 10 reveals pairwise comparisons between groups based on the Bonferroni adjustment test. According to the table, on the pretest, the difference between portfolio assessment and self-assessment was not statistically significant (mean difference =  $-0.027$ ,

**Table 9** Tests of between-subject effects

Source	Dependent variable	Type III sum of squares	df	Mean square	F	Sig	Partial Eta squared
Corrected model	Vocabulary pretest score	7.304 <sup>a</sup>	7	1.043	.456	.864	.027
	Vocabulary post-test score	1686.955 <sup>b</sup>	7	240.994	20.130	.000	.546
Intercept	Vocabulary pretest score	1474.869	1	1474.869	644.823	.000	.846
	Vocabulary post-test score	9906.730	1	9906.730	827.510	.000	.876
Condition	Vocabulary pretest score	1.846	3	.615	.269	.848	.007
	Vocabulary post-test score	1174.275	3	391.425	32.696	.000	.456
WM	Vocabulary pretest score	2.797	1	2.797	1.223	.271	.010
	Vocabulary post-test score	211.363	1	211.363	17.655	.000	.131
Condition WM	Vocabulary pretest score	2.810	3	.937	.410	.746	.010
	Vocabulary post-test score	228.742	3	76.247	6.369	.000	.140
Error	Vocabulary pretest score	267.608	117	2.287			
	Vocabulary post-test score	1400.693	117	11.972			
Total	Vocabulary pretest score	1761.000	125				
	Vocabulary post-test score	13,448.000	125				
Corrected total	Vocabulary pretest score	274.912	124				
	Vocabulary post-test score	3087.648	124				

**Table 10** Pairwise comparisons

Dependent variable	(I) Condition	(J) Condition	Mean difference (I-J)	Std. error	Sig	95% Confidence interval for difference	
						Lower bound	Upper bound
Vocabulary pretest score	Portfolio assessment	Self-assessment	-.027	.391	1.000	-1.077	1.024
		Peer assessment	.230	.377	1.000	-.781	1.241
		Control	-.077	.393	1.000	-1.131	.977
	Self-assessment	Portfolio assessment	.027	.391	1.000	-1.024	1.077
		Peer assessment	.257	.377	1.000	-.754	1.268
		Control	-.050	.393	1.000	-1.104	1.004
	Peer assessment	Portfolio assessment	-.230	.377	1.000	-1.241	.781
		Self-assessment	-.257	.377	1.000	-1.268	.754
		Control	-.307	.378	1.000	-1.322	.708
	Control	Portfolio assessment	.077	.393	1.000	-.977	1.131
		Self-assessment	.050	.393	1.000	-1.004	1.104
		Peer assessment	.307	.378	1.000	-.708	1.322
Vocabulary post-test score	Portfolio assessment	Self-assessment	.353	.895	1.000	-2.050	2.756
		Peer assessment	-1.548	.862	.450	-3.862	.765
		Control	6.557	.898	.000	4.146	8.968
	Self-assessment	Portfolio assessment	-.353	.895	1.000	-2.756	2.050
		Peer assessment	-1.901	.862	.176	-4.215	.413
		Control	6.204	.898	.000	3.793	8.616
	Peer assessment	Portfolio assessment	1.548	.862	.450	-.765	3.862
		Self-assessment	1.901	.862	.176	-.413	4.215
		Control	8.106	.865	.000	5.783	10.428
	Control	Portfolio assessment	-6.557	.898	.000	-8.968	-4.146
		Self-assessment	-6.204	.898	.000	-8.616	-3.793
		Peer assessment	-8.106	.865	.000	-10.428	-5.783

$p > 0.05$ ). Additionally, the difference between portfolio assessment and scaffolded peer assessment did not turn out to be significant (mean difference = 0.230,  $p > 0.05$ ). Furthermore, there was no statistical difference between the portfolio assessment group and the control condition (mean difference = -0.077,  $p > 0.05$ ). The table further discloses that no group had a statistical difference with the control condition of the pretest ( $p > 0.05$ ). However on the post-test, post hoc analyses reveal that the mean difference between portfolio assessment and self-assessment is also non-significant (mean

**Table 11** Condition (WM)

Dependent variable	Condition	WM	Mean	Std. error	95% Confidence interval	
					Lower bound	Upper bound
Vocabulary pretest score	Portfolio assessment	High	3.188	.378	2.439	3.936
		Low	3.786	.404	2.985	4.586
	Self-assessment	High	3.214	.404	2.414	4.015
		Low	3.813	.378	3.064	4.561
	Peer assessment	High	3.278	.356	2.572	3.984
		Low	3.235	.367	2.509	3.962
	Control	High	3.538	.419	2.708	4.369
		Low	3.588	.367	2.862	4.315
Vocabulary post-test score	Portfolio assessment	High	11.375	.865	9.662	13.088
		Low	9.214	.925	7.383	11.046
	Self-assessment	High	10.571	.925	8.740	12.403
		Low	9.313	.865	7.599	11.026
	Peer assessment	High	15.333	.816	13.718	16.948
		Low	8.353	.839	6.691	10.015
	Control	High	3.769	.960	1.869	5.670
		Low	3.706	.839	2.044	5.368

**Table 12** Pairwise comparisons

Dependent variable	(I) WM	(J) WM	Mean difference (I-J)	Std. error	Sig.	95% Confidence interval for difference	
						Lower bound	Upper bound
Vocabulary pretest score	High	Low	-.301	.272	.271	-.840	.238
	Low	High	.301	.272	.271	-.238	.840
Vocabulary post-test score	High	Low	2.616	.623	.000	1.383	3.849
	Low	High	-2.616	.623	.000	-3.849	-1.383

difference = 0.353,  $p > 0.05$ ), the difference between portfolio assessment and scaffolded peer assessment is also non-significant (mean difference = -1.548,  $p > 0.05$ ), the difference between self-assessment and scaffolded peer assessment is also non-significant (mean difference = 0.862,  $p > 0.05$ ). A further inspection of the table shows that the difference between all three experimental conditions and the control group turns out to be statistically significant ( $p < 0.05$ ). (Table 11).

Further post hoc analyses based on the Bonferroni adjustment test reveal that on the post-test in the portfolio assessment condition, high WM learners had a higher mean than their low WM counterparts (mean difference = 2.161). In the self-assessment condition, high WM subjects had also a higher mean than their low WM peers (mean difference = 1.258). In the scaffolded peer assessment condition, learners with high WM had an amazingly higher mean than their low WM peers (mean difference = 6.980). In the control condition, high WM learners had a higher mean than their low WM peers (mean difference = 0.063). (Table 12).

In addition to what went above, further pairwise comparisons concerning WM capacity reveal that the mean difference between high and low WM learners on the pretest

**Table 13** Descriptive statistics

	Condition	WM	Mean	Std. deviation	N
Grammar pretest score	Portfolio assessment	High	3.687	1.537	16
		Low	4.285	1.728	14
		Total	3.966	1.629	30
	Self-assessment	High	3.857	1.747	14
		Low	3.437	1.711	16
		Total	3.633	1.711	30
	Peer assessment	High	3.500	1.886	18
		Low	3.705	1.263	17
		Total	3.600	1.594	35
	Control	High	3.384	1.445	13
		Low	3.882	1.317	17
		Total	3.666	1.372	30
	Total	High	3.606	1.645	61
		Low	3.812	1.500	64
		Total	3.712	1.569	125
Grammar post-test score	Portfolio assessment	High	11.562	3.723	16
		Low	9.357	3.650	14
		Total	10.533	3.794	30
	Self-assessment	High	10.928	4.322	14
		Low	9.562	1.931	16
		Total	10.200	3.284	30
	Peer assessment	High	15.666	3.217	18
		Low	8.588	4.302	17
		Total	12.228	5.173	35
	Control	High	3.923	1.320	13
		Low	3.823	1.590	17
		Total	3.866	1.455	30
	Total	High	11.000	5.316	61
		Low	7.734	3.838	64
		Total	9.328	4.883	125

was not statistically significant (mean difference = 0.301,  $p > 0.05$ ); however, on the post-test, the difference turned out to be significant (mean difference = 2.616,  $p < 0.05$ ). Additionally, calculations by hand revealed that the effect size was moderate (partial eta squared = 0.083).

Research question 3: In terms of grammatical accuracy across various WM capacities, are there any notable differences between students who receive portfolio evaluation, those who receive self-assessment, and those who receive scaffolded peer assessment?

In this scenario, similar independent and moderating variables as the first two research questions are at work. The only difference is that in this scenario, instead of scores on a reading comprehension test, and vocabulary test, the research deals with scores on a grammar test on two occasions (pre- and post-test scores) as the dependent variables. Thus, a further two-way between-group MANOVA needs to be conducted (Pallant, 2020). The two assumptions of normality and homogeneity were checked through a one-sample K-S test, and Levene's test of equality of variances, respectively. However, due to space limitations, their respective tables are not

**Table 14** Tests of between-subject effects

Source	Dependent variable	Type III sum of squares	df	Mean square	F	Sig.	Partial Eta squared
Corrected model	Grammar pretest score	8.815	7	1.259	.496	.836	.029
	Grammar post-test score	1744.023	7	249.146	24.021	.000	.590
Intercept	Grammar pretest score	1707.604	1	1707.604	673.106	.000	.852
	Grammar post-test score	10,404.504	1	10,404.504	1003.130	.000	.896
Condition	Grammar pretest score	2.975	3	.992	.391	.760	.010
	Grammar post-test score	1214.194	3	404.731	39.021	.000	.500
WM	Grammar pretest score	1.502	1	1.502	.592	.443	.005
	Grammar post-test score	223.076	1	223.076	21.507	.000	.155
Condition WM	Grammar pretest score	4.693	3	1.564	.617	.606	.016
	Grammar post-test score	231.571	3	77.190	7.442	.000	.160
Error	Grammar pretest score	296.817	117	2.537			
	Grammar post-test score	1213.529	117	10.372			
Total	Grammar pretest score	2028.000	125				
	Grammar post-test score	13,834.000	125				
Corrected total	Grammar pretest score	305.632	124				
	Grammar post-test score	2957.552	124				

represented here. The results showed that the sig. (2-tailed) for both tests exceeded the threshold level of 0.05, hence the conformation of normality and homogeneity assumption. Now, there is room to conduct the MANOVA.

Table 13 presents the descriptive statistics regarding subjects' performance in all conditions on both pre and post-test of grammar. According to Table 13, in the portfolio group, high WM spanners at 1.537 SD had 3.687 as the mean, while learners with low WM had 4.285 at 1.728 SD. In the self-assessment group, learners with high WM scored 3.857 as the mean at 1.747 SD, whereas low WM learners scored 3.437 as the mean at 1.711 SD. In the peer assessment condition, high WM learners scored 3.500 as the mean at 1.886 SD, while low WM spanners scored 3.705 as the mean at 1.263 SD. High WM subjects in the control group had 3.384 as their mean with 1.445 SD, whereas low WM participants in the same condition had 3.882 as their mean with 1.317 SD. The table also summarizes the results of the post-test. Based on the table, in the portfolio condition, learners with high WM scored 11.562 as the mean with 3.723 SD, and low WM participants scored 9.357 as the mean with 3.650 SD. Besides, in the self-assessment condition, high WM learners scored 10.928 as the mean with 4.322 SD, while learners with low WM scored 9.562 as the mean with 1.931 SD. High WM learners in the scaffolded peer assessment group had 15.666 as their mean with 3.3217 SD, while their low WM counterparts had 8.588 as their mean with 4.302 SD. Furthermore, in the control group, high WM subjects, had 3.923 as their mean with 1.320 SD, and low WM learners had 3.823 means with 1.590 SD. Overall, the table shows that on the pretest, high WM participants had a 3.606 mean with 1.645 SD, and low WM subjects had a 3.812 mean with 1.500 SD. On the post-test, these numbers rose dramatically such that high WM subjects had 11.000 means with 5.316 SD, and low WM learners had 7.734 means with 3.838 SD.

Table 14 presents tests of between-subject effects. According to Table 14, on the pretest of grammatical accuracy, at 3 degrees of freedom and with  $F=0.391$ , the



**Table 15** Pairwise comparisons

Dependent variable	(I) Condition	(J) Condition	Mean difference (I-J)	Std. error	Sig.	95% Confidence interval for difference	
						Lower bound	Upper bound
Grammar pretest score	Portfolio assessment	Self-assessment	.339	.412	1.000	-.767	1.445
		Peer assessment	.384	.397	1.000	-.681	1.449
		Control	.353	.414	1.000	-.757	1.463
	Self-assessment	Portfolio assessment	-.339	.412	1.000	-1.445	.767
		Peer assessment	.044	.397	1.000	-1.021	1.109
		Control	.014	.414	1.000	-1.096	1.124
	Peer assessment	Portfolio assessment	-.384	.397	1.000	-1.449	.681
		Self-assessment	-.044	.397	1.000	-1.109	1.021
		Control	-.031	.398	1.000	-1.100	1.038
	Control	Portfolio assessment	-.353	.414	1.000	-1.463	.757
		Self-assessment	-.014	.414	1.000	-1.124	1.096
		Peer assessment	.031	.398	1.000	-1.038	1.100
Grammar post-test score	Portfolio assessment	Self-assessment	.214	.833	1.000	-2.022	2.451
		Peer assessment	-1.668	.802	.239	-3.821	.486
		Control	6.587	.836	.000	4.342	8.831
	Self-assessment	Portfolio assessment	-.214	.833	1.000	-2.451	2.022
		Peer assessment	-1.882	.802	.124	-4.035	.272
		Control	6.372	.836	.000	4.128	8.617
	Peer assessment	Portfolio assessment	1.668	.802	.239	-.486	3.821
		Self-assessment	1.882	.802	.124	-.272	4.035
		Control	8.254	.805	.000	6.093	10.416
	Control	Portfolio assessment	-6.587	.836	.000	-8.831	-4.342
		Self-assessment	-6.372	.836	.000	-8.617	-4.128
		Peer assessment	-8.254	.805	.000	-10.416	-6.093

difference between groups was not statistically significant ( $p=0.760$ ). The table further reveals that on the post-test, at 3 degrees of freedom with  $F=39.021$ , the difference between conditions was statistically significant at  $p<0.05$  with a large effect size (partial eta squared = 0.500). Concerning subjects' WM capacity on the pretest, at 1 degree of freedom with  $F=0.592$ , no statistical difference between subjects was found ( $p=0.443$ ). However, on the post-test, at 1 degree of freedom with  $F=21.507$ , there was a statistical difference between subjects with a moderate effect size ( $p<0.05$ , partial eta

**Table 16** Condition (WM)

Dependent variable	Condition	WM	Mean	Std. error	95% Confidence interval	
					Lower bound	Upper bound
Grammar pretest score	Portfolio assessment	High	3.688	.398	2.899	4.476
		Low	4.286	.426	3.443	5.129
	Self-assessment	High	3.857	.426	3.014	4.700
		Low	3.438	.398	2.649	4.226
	Peer assessment	High	3.500	.375	2.757	4.243
		Low	3.706	.386	2.941	4.471
	Control	High	3.385	.442	2.510	4.259
		Low	3.882	.386	3.117	4.647
	Portfolio assessment	High	11.562	.805	9.968	13.157
		Low	9.357	.861	7.653	11.062
Grammar post-test score	Self-assessment	High	10.929	.861	9.224	12.633
		Low	9.563	.805	7.968	11.157
	Peer assessment	High	15.667	.759	14.163	17.170
		Low	8.588	.781	7.041	10.135
	Control	High	3.923	.893	2.154	5.692
		Low	3.824	.781	2.277	5.370

**Table 17** Pairwise comparisons

Dependent variable	(I) WM	(J) WM	Mean difference (I-J)	Std. error	Sig.	95% Confidence interval for difference	
						Lower bound	Upper bound
Grammar pretest score	High	Low	-.221	.287	.443	-.788	.347
	Low	High	.221	.287	.443	-.347	.788
Grammar post-test score	High	Low	2.687	.579	.000	1.540	3.835
	Low	High	-2.687	.579	.000	-3.835	-1.540

squared=0.155). Concerning the interaction between condition and WM capacity on the pretest, at 3 degrees of freedom with  $F=0.617$ , there was no statistical difference between conditions as the  $p$  value exceeds the threshold level 0.05; however, not the post-test, at 3 degrees of freedom with  $F=7.442$ , a statistical difference was observed at  $p < 0.05$  with a large effect size (partial eta squared=0.160).

Table 15 reveals pairwise comparisons between groups based on the Bonferroni adjustment test. According to the table, on the pretest, the difference between portfolio assessment and self-assessment was not statistically significant (mean difference = 0.339,  $p > 0.05$ ). Additionally, the difference between portfolio assessment and scaffolded peer assessment did not turn out to be significant (mean difference = 0.384,  $p > 0.05$ ). Furthermore, there was no statistical difference between the portfolio assessment group and the control condition (mean difference = 0.353,  $p > 0.05$ ). The table further discloses that no group had a statistical difference with the control condition of the pretest ( $p > 0.05$ ). However, on the post-test, post hoc analyses reveal that the mean difference between portfolio assessment and self-assessment is also non-significant (mean difference = 0.214,  $p > 0.05$ ), the difference between portfolio assessment and scaffolded peer

assessment is also non-significant (mean difference =  $-1.668$ ,  $p > 0.05$ ), and the difference between self-assessment and scaffolded peer assessment is also non-significant (mean difference =  $-0.214$ ,  $p > 0.05$ ). A further inspection of the table shows that the difference between all three experimental conditions and the control group turns out to be statistically significant ( $p < 0.05$ ). (Table 16).

Further post hoc analyses based on the Bonferroni adjustment test reveal that on the post-test in the portfolio assessment condition, high WM learners had a higher mean than their low WM counterparts (mean difference =  $2.205$ ). In the self-assessment condition, high WM subjects had also a higher mean than their low WM peers (mean difference =  $1.366$ ). In the scaffolded peer assessment condition, learners with high WM had an amazingly higher mean than their low WM peers (mean difference =  $7.079$ ). In the control condition, high WM learners had a higher mean than their low WM peers (mean difference =  $0.099$ ). (Table 17).

In addition to what went above, further pairwise comparisons concerning WM capacity reveal that the mean difference between high and low WM learners on the pretest was not statistically significant (mean difference =  $-0.221$ ,  $p > 0.05$ ); however, on the post-test, the difference turned out to be significant (mean difference =  $2.687$ ,  $p < 0.05$ ). Additionally, calculations by hand revealed that the effect size was moderate (partial eta squared =  $0.072$ ).

## Discussion

In this section of the study, the impacts of portfolio assessment, self-assessment, and scaffolded peer assessment on reading comprehension, lexical growth, and grammatical accuracy each concerning WM capacity are discussed. Concerning each research question a two-way between-group MANOVA was performed. The results disclosed that all three experimental conditions outstripped the comparison condition on all dependent variables of text understanding, lexical gain, and grammatical accuracy. The results further revealed that regarding WM capacity, in all experimental conditions, high WM participants outperformed their low WM counterparts. In addition, based on the obtained results, no statistical difference was found between all three experimental conditions. To be more specific, subjects in scaffolded peer assessment conditions obtained more reading comprehension skills, vocabulary knowledge, and grammatical accuracy, but their difference from those of other participants in other experimental settings was negligible ( $p > 0.05$ ).

In terms of the promising effects of portfolio assessment established based on the results, findings are in sharp contrast with that of Amani and Salehi (2017). These L2 researchers had shown that portfolio assessment cannot facilitate reading comprehension any more than traditional methods can. Therefore, they were skeptical about the enhancing role of this type of alternative assessment. However, based on this study's findings, portfolio assessment can result in text comprehension improvement as well as vocabulary growth, and grammatical accuracy. The findings are also in line with those of Barrot (2021), Nourdad and Banagozar (2020), and Rezai et al., (2022a, 2022b). Barrot (2021) found that portfolio assessment can improve learners' writing performance. Although this study did not assess EFL learners' writing skills, the results imply that

portfolio assessment can result in overall language development. In this way, the results are consistent with that of Barrot. Additionally, Nourdad and Banagozar (2022) investigated the effect of portfolio assessment on vocabulary gain and retention. The results of their study pointed to the efficacy of this type of alternative assessment on both immediate post-test and delayed post-test. Although this current study did not measure the long-term effects of portfolio assessment on vocabulary development, the findings are in line with the abovementioned researchers. In another study, Rezai et al., (2022a, 2022b) found that portfolio assessment can improve vocabulary knowledge which is in line with this study's results.

This study also found support for the facilitative role of self-assessment as a teaching technique in reading comprehension, lexical growth, and grammatical accuracy in an EFL context. The results are consistent with Rezai et al., (2022a, 2022b), Chung et al. (2021), and Alek et al. (2020). Rezai and his associates were concerned about the contribution of the self-assessment procedure to writing development. Their study found support for the procedure. Although this current study did not directly measure EFL learners' writing ability, it is safe to say that Rezai et al.'s findings are to some extent relevant to this paper's results as the researcher also found support for the enhancing role of self-assessment on reading comprehension, vocabulary knowledge, and grammatical expansion. Chung et al. (2021) also came to an understanding that portfolio assessment can result in writing improvement. Additionally, Alek et al. (2020) conducting a mixed-methods investigation found that self-assessment can improve learners' speaking skills.

Our results also indicated that scaffolded peer assessment can improve learners' text understanding, knowledge of lexical items, and structural understanding. Thus, the results are consistent with Li et al. (2020) and Homayouni (2022). In a meta-analysis, Li and his colleagues (2020) found that peer assessment can result in language learning gain which is consistent with the findings of this current exploration. Additionally, Homayouni (2022) found that peer assessment coupled with scaffolding and group work can improve both vocabulary knowledge on both immediate post-test and delayed post-test as well as learners' oral skills. Homayouni's (2022) findings are consistent with our findings on the basis that we also found support for the efficacy of scaffolded peer assessment in lexical growth. However, our study's results are in contrast with that of Moghimi (2022). Moghimi (2022) compared and contrasted the effects of peer assessment with self-assessment on learners' accuracy in speech. This researcher came to an understanding that peer assessment is statistically superior to self-assessment in terms of its effect on accuracy in speech. This finding is somehow in contrast with our findings. Although we found that scaffolded peer assessment can result in more learning gain than self-assessment does, this difference was not significant. That is, both types of assessment can improve learners' text comprehension, vocabulary knowledge, and structural accuracy.

The results of this exploration also corroborated that WM as an individual difference can facilitate language learning. This study found that high WM can learners learn more than their low WM peers. This finding supports the earlier claim made by Chow et al. (2021). Chow et al. (2021) found that verbal WM and reading anxiety were two independent predictors of ESL reading comprehension. Additionally, Teng and Zhang (2021) found that complex and phonological WM plays a decisive role in vocabulary

learning and retention. Thus, Teng and Zhang's (2021) results are consistent with this study's findings. In addition to these studies, Patra et al. (2022) also found that learners with high WM can gain more grammatical knowledge than learners with low WM. This finding is completely in line with our finding as this study also found that learners with high WM who are exposed to portfolio assessment, self-assessment, and scaffolded peer assessment can not only fare better on a test of reading comprehension, but also on tests of vocabulary knowledge and grammatical accuracy.

This study tried to add a cognitive individual difference moderating variable (i.e., WM capacity) to the contribution of different types of assessment, namely portfolio assessment, self-assessment, and scaffolded peer assessment to text understanding, lexical gain, and structural accuracy. The novelty of the study lies in the addition of the moderating role of WM to the gain as a result of the abovementioned types of assessment. The results showed that all experimental groups outperformed the control group on the post-test; however, there was no statistical difference between subjects in different experimental conditions. To be more specific, learners in scaffolded peer assessment condition gained more in terms of reading comprehension, vocabulary knowledge, and grammar, but the difference with those of other subjects in other treatment groups was not significant in a statistical sense. The findings further elucidated that learners with high WM outperformed learners with low WM. This does not imply that low WM learners cannot learn text comprehension, vocabulary, and grammar as the result of the different levels of the independent variable of the study (i.e., portfolio assessment, self-assessment, and scaffolded peer assessment), but it implies that learners with high WM have the advantage to learn more of text understanding, lexical items, and grammatical structures than learners with poor WM.

On the whole, the results were more supportive of peer assessment as an instantiation of alternative assessment to the improvement of EFL learners' text understanding, vocabulary growth, and grammatical expansion. Recently, researchers have made a growing case for the use of evaluation to encourage learning in academic practice (William, 2018). Peer assessment is a crucial part of formative assessment theories since it is believed to provide teachers or students with new information about the learning process, improving subsequent performance. The results of this study support the notion that peer evaluation, at least in language programs that focus on reading comprehension, vocabulary, and grammar development, might be a useful instructional strategy for increasing student progress. The results suggest that peer evaluation, which is more successful than other types of assessment, can play a key formative role in classrooms. According to the findings, creating classroom activities that incorporate peer assessment can be a helpful way to promote learning and make the most of instructional resources by allowing the teacher to focus on assisting students with harder and more involved tasks. This demonstrates, practically speaking, that teachers can implement peer assessment in several ways and tailor the design to the particular features and constraints of their contexts (Double et al., 2020).

Although the benefits of peer evaluation on language productive skills have been extensively studied, very few to no studies have examined its effect on vocabulary learning (Ritonga et al., 2022). In light of the potential effects of this type of assessment on a novel-dependent variable and the extent to which the modified independent variable

(i.e., scaffolded peer judgment) can explain variance in the dependent variable, this research can be viewed as innovative in the strictest sense.

Students can participate in cooperative learning where they are enthusiastic to assist and evaluate their classmates and take responsibility for their language learning accomplishments by using peer assessment. This may result in enhanced social abilities, greater evaluation, and more precise feedback (Homayouni, 2022). Because they care about the group members and want to achieve the same objective, learners learn better when they assist one another. This is according to the social interdependence theory (Slavin, 2011). The results of this study thus support the social constructivism provided by Vygotsky (1987) given that learners with a similar age cohort jointly cooperate and widen each other's ZPD (Webb, 2008).

### **Implications of the study**

This study poses several theoretical and pedagogical implications. The first theoretical implication of the study is that if learners are given their voice and choice in assessing their learning and assessing those of their peers, and decision-making, their learning will be improved. Another theoretical implication of this study is that through collaborative learning with the help of a more proficient peer (i.e., a mediator), learners can do the tasks they cannot on their own, so learners' ZPD is broadened in this way (Vygotsky, 1987). Another theoretical implication of the study is that when a teacher decides to use the portfolio as a classroom-based assessment tool, they should plan and prepare well in advance (Mathur & Mahapatra, 2022). Making the implementation effective can be greatly aided by identifying specific skill areas (subskills), task types, materials, and a progress-check system. A further theoretical implication of the study is that, through the researcher's own experience as an English instructor, it has been revealed that English teachers are not inclined to allow learners to self-assess their progress in the language. Thus, this study can shed new light on self-evaluation, which could be applied as an English teaching strategy. A final theoretical implication of this study is that self-assessment can make learners more independent (Masruria & Anam, 2021).

In addition to the abovementioned theoretical implications, this study poses several pedagogical implications as well. One way teachers might promote cooperative learning is through exercises that incorporate peer review. If they want to better their language learning, students learning English as a second language may find it helpful to get familiar with various methods of assessment in general and peer assessment in particular. Also, students can identify exactly where they require help and support so that they can ask for it from their teachers by using peer and self-assessment tasks.

Peer evaluation and scaffolded learning can offer an engaging, thrilling, nearly stress-free atmosphere for learning the bolts and nuts of language, according to a second pedagogical implication of this study. Students can increase their language proficiency, reading comprehension, vocabulary size, and grammatical accuracy by using cooperative learning strategies through peer assessment and scaffolded learning. Additionally, a stress-free learning environment will be created where cooperation is encouraged rather than unhealthy competition that stunts mental development.



Another pedagogical implication of the study is that peer assessment and portfolios should go hand in hand with routine activities and self-evaluation. Peer assessment can lessen students' fear of receiving negative feedback in addition to giving them feedback that can help them improve (Cepik & Yastibas, 2013). The more feedback the students receive from their peers, the more accustomed they become to managing their anxieties and emotions. The majority of pupils lack social interaction while learning a language, as Guo et al. (2018) noted. The students' assumption that their peers will judge their performance badly because they do not have close relationships with their peers may be a result of this lack of social engagement.

This study also has implications for the administration. The heads of language institutes and the administrators of public schools might urge their academic staff to use the study's findings in the classroom to foster autonomy and improve students' reading comprehension, vocabulary acquisition, and grammatical accuracy.

This study has implications for those who create educational materials. The results of this analysis can also be thoroughly examined by curriculum developers and/or syllabus designers, who can then include the findings presented in this study in their upcoming materials. It is recommended that those who create syllabi provide a range of assessment types in their materials. The results of this study could help task and activity designers produce a range of tasks and activities that are specifically tailored to EFL students' reading comprehension, vocabulary growth, and grammatical improvement.

## Conclusion

This investigation was carried out to fill a knowledge gap and provide language instructors with pedagogical implications over the comparative effects of portfolio assessment, self-assessment, and scaffolded peer assessment with the mediating role of WM on reading comprehension, vocabulary learning, and grammatical accuracy. The results of this study showed that all three types of assessment are facilitative of reading comprehension, lexical gain, and structural accuracy. The results further showed that learners with high WM can gain more in terms of reading comprehension, vocabulary development, and grammatical structures. Although the findings of the study revealed the supremacy of scaffolded peer assessment groups over other experimental conditions, the difference was non-significant, pointing to the efficacy of all three types of assessment targeted in this study posing several theoretical and pedagogical implications for the study discussed above in full detail.

This study pointed to the efficacy of different types of alternative assessments in learning different bits of a foreign language in a foreign context. Based on the results of the study, it is possible to claim that we should move away from teacher-fronted classrooms and traditional testing, to new approaches and possibilities that have emerged in recent times. By applying the principles of alternative assessment, not the least among its types of peer assessment which is based on Vygostkian thinking, learners' ZPD could be broadened, and learning a new language is facilitated.

Although this study appears to be an innovation, it is not flawless. First of all, the study did not use randomization which is a necessary condition of experimental designs

(Mackey & Gass, 2022). It is suggested that prospective researchers randomly select the participants to improve the internal validity of their studies (Ary et al., 2019). Furthermore, this study was conducted in just a particular setting. Future studies should target several geographical areas to see if the results will be the same. Another limitation of this study is that the research did not take into account the long-term effects of these types of assessments. Accordingly, it is not clear whether the effects can be long-lasting or not. For this reason, we suggest that future studies add a delayed post-test to their instruments to check whether the effects can remain or not.

#### Abbreviations

EFL	English as a foreign language
OQPT	Oxford Quick Placement Test
WM	Working memory
MANOVA	Multivariate analysis of variance
ZPD	Zone of proximal development
SLA	Second language acquisition
ANCOVA	Analysis of covariance

#### Acknowledgements

Not applicable.

#### Authors' contributions

All authors had adequate and equal contributions. The authors read and approved the final manuscript.

#### Authors' information

Anwar Hammad Al-Rashidi got her PhD in Counseling Psychology from Imam Mohamed bin Saud Islamic University. She is an associate professor in the Psychology Department in Education College, Prince Sattam Bin Abdulaziz University, Al-Kharj, Saudi Arabia, and she published several studies in international journals.

Dr. Balachandran Vadivel completed his Bachelor of Arts degree in English Literature from Bharathidasan University, India, in 2003 and went on to secure his Master of Arts degree in English Literature from the same university in 2006. He also holds a Bachelor of Education degree from the University of Madras, which he completed in 2004. In 2008, he obtained his Master of Philosophy degree from Bharathidasan University and started his career as a lecturer at H.H. The Rajah's College, India, in the same year. In 2010, he was appointed as an Assistant Professor of English at Mount Zion College of Engineering and Technology, which is affiliated with Anna University, India. In 2018, he earned his Ph.D. in English Language Teaching (ELT) from Bharathidasan University. His research interests lie in the areas of ELT, Second Language Acquisition, and Applied Linguistics, and he has published several research papers in these fields. Currently, he is serving as an Assistant Professor at Cihan University-Duhok, located in the Kurdistan Region of Iraq.

Nawroz Ramadan Khalil earned her B.A. in Language and Literature from Aleppo University, Syria in 1997. She went on to complete her M.A. in Literature, Drama, and Poetry from the same university in 2002, with a thesis titled "The Influence of Western Literature on Eastern Literature." In 2006, she obtained her Ph.D. in Literature, Drama, and Poetry from Aleppo University with a dissertation titled "The Influence of Literature on University Courses." From 1997 to 2010, Nawroz worked at Aleppo University in various positions, including heading international research missions, serving as a member of the university committee, leading the English department, and supervising student research papers. She then worked at Duhok University's College of Basic Education/Department of English from 2010 to 2014. Currently, she is the head of the English Department at Cihan University. Nawroz has published numerous research articles and has participated in various workshops and conferences related to language and literature in Iraq (Kurdistan region), Syria, and Lebanon. She has also worked for international organizations such as ACTED as a supervisor for awareness campaigns aimed at refugees who have been traumatized by war.

Nirvana Basim graduated from the Department of Literature and Linguistics, Kandahar University, Kandahar, Afghanistan.

#### Funding

This study is supported via funding from Prince Sattam Bin Abdulaziz University Project Number (PSAU 2023 /R/1444).

#### Availability of data and materials

The data that support the findings of this study are available from the corresponding author upon reasonable request.

#### Declarations

##### Competing interests

The authors declare that they have no competing interests.

Received: 3 April 2023 Accepted: 19 April 2023

Published online: 05 May 2023

## References

- Alam, M. (2019). Assessment challenges & impact of formative portfolio assessment (FPA) on EFL learners' writing performance: A case study on the preparatory English language course. *English Language Teaching*, 12(7), 161–172.
- Alawajee, O. A., & Almutairi, H. A. (2022). Level of readiness for in-class teaching among teachers of students with special educational needs: Post-COVID-19. *Eurasian Journal of Educational Research*, 98(98), 1–20.
- Alek, A., Marzuki, A. G., Farkhan, M., & Deni, R. (2020). Self-assessment in exploring EFL students' speaking skill. *Al-Ta Lim Journal*, 27(2), 208–214.
- Al-Mamoory, S., & Abathar Witwit, M. (2021). Critical discourse analysis of oppression in "To Kill a Mockingbird." *Journal of Social Science and Humanities Research*, 9(2), 11–24.
- Amani, F., & Salehi, H. (2017). Impacts of portfolio assessment on Iranian EFL students' reading comprehension ability based on junior high school English textbook (PROSPECT 2). *Journal of English Language and Literature-JOELL*, 4(4), 69–84.
- Ary, D., Jacobs, L. C., Sorensen, C. K., & Walker, D. (2019). *Introduction to research in education* (10th ed.). Cengage Learning.
- Azizi, Z., & Namaziandost, E. (2023). Implementing peer-dynamic assessment to cultivate Iranian EFL learners' inter-language pragmatic competence: A mixed-methods approach. *International Journal of Language Testing*, 13(1), 18–43. <https://doi.org/10.22034/ijlt.2022.345372.1171>.
- Bachman, L. F., Palmer, A. S., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford University Press.
- Baddeley, A. D. (2017). Modularity, working memory and language acquisition. *Second Language Research*, 33, 299–311.
- Barrot, J. S. (2016). Using Facebook-based e-portfolio in ESL writing classrooms: Impact and challenges. *Language, Culture and Curriculum*, 29(3), 286–301.
- Barrot, J. S. (2021). Effects of Facebook-based e-portfolio on ESL learners' writing performance. *Language, Culture and Curriculum*, 34(1), 95–111. <https://doi.org/10.1080/07908318.2020.1745822>
- Behbahani, S. M. K., Pourdana, N., Maleki, M., & Javanbakht, Z. (2011). EFL task-induced involvement and incidental vocabulary learning: succeeded or surrounded. In International Conference on Languages Literature and Linguistics. *IPEDR Proceedings*, 26, 323–325.
- Benson, S., & DeKeyser, R. (2019). Effects of written corrective feedback and language aptitude on verb tense accuracy. *Language Teaching Research*, 23(6), 702–726.
- Birjandi, P., & Hadidi Tamjid, N. (2012). The role of self-, peer and teacher assessment in promoting Iranian EFL learners' writing performance. *Assessment & Evaluation in Higher Education*, 37(5), 513–533. <https://doi.org/10.1080/02602938.2010.549204>
- Bolukbas, F., Keskin, F., & Polat, M. (2011). The effectiveness of cooperative learning on the reading comprehension skills in Turkish as a foreign language. *Turkish Online Journal of Educational Technology-TOJET*, 10(4), 330–335.
- Carless, D., & Boud, D. (2018). The development of student feedback literacy: Enabling uptake of feedback. *Assessment and Evaluation in Higher Education*, 43(8), 1315–1325.
- Castles, A., Rastle, K., & Nation, K. (2018). Ending the Reading Wars: Reading Acquisition From Novice to Expert. *Psychological Science in the Public Interest*, 19(1), 5–51. <https://doi.org/10.1177/1529100618772271>
- Cepik, S., & Yastibas, A. E. (2013). The use of e-portfolio to improve English speaking skill of Turkish EFL learners. *Anthropologist*, 16(1–2), 307–317.
- Chappuis, J. (2014). *Seven strategies of assessment for learning* (2nd ed.). Pearson.
- Chow, B. W. Y., Mo, J., & Dong, Y. (2021). Roles of reading anxiety and working memory in reading comprehension in English as a second language. *Learning and Individual Differences*, 92, 102092. <https://doi.org/10.1016/j.lindif.2021.102092>.
- Chung, H. Q., Chen, V., & Olson, C. B. (2021). The impact of self-assessment, planning and goal setting, and reflection before and after revision on student self-efficacy and writing performance. *Reading and Writing*, 34, 1885–1913.
- Cowan, N. (2017). The many faces of working memory and short-term storage. *Psychonomic Bulletin & Review*, 24, 1158–1170.
- Daneman, M., & Carpenter, P. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, 19, 450–466.
- Davoudi, M., & Heydarnejad, T. (2020). The interplay between reflective thinking and language achievement: A case of Iranian EFL learners. *Language Teaching Research Quarterly*, 18, 70–82.
- Double, K. S., McGrane, J. A., & Hopfenbeck, T. N. (2020). The impact of peer assessment on academic performance: A meta-analysis of control group studies. *Educational Psychology Review*, 32(2), 481–509.
- Diaz-Rico, L., & Weed, K. (2002). *The Crosscultural, Language, and Academic Development Handbook* (3rd ed.). Boston: Pearson.
- Duff, P. (2010). Language socialization into academic discourse communities. *Annual Review of Applied Linguistics*, 30, 169–192.
- Ebadi, S., & Rahimi, M. (2019). Mediating EFL learners' academic writing skills in online dynamic assessment using Google Docs. *Computer Assisted Language Learning*, 32(5–6), 527–555.
- Ellis, R. (1997). *Second Language Acquisition*. Oxford University Press.
- Esteve, O., Trenchs, J. T., Pujola, J., Arumi, M., & Birello, M. (2012). The ELP as a mediating tool for the development of self-regulation in foreign language learning university contexts: An ethnographic study. In B. Kuhn & M. Perez Cavana (Eds.), *Perspectives from the European Language Portfolio: Learner autonomy and self-assessment* (pp. 73–99). Routledge.
- Eysenck, H. J. (1990). Biological dimensions of personality. In L. A. Pervin (Ed.), *Handbook of personality: Theory and research* (pp. 244–276). Guilford Press.
- Gan, L., & Lam, R. (2020). Understanding university English instructors' assessment training needs in the Chinese context. *Language Testing in Asia*, 10(11). <https://doi.org/10.1186/s40468-020-00109-y>.
- Gielen, S., Peeters, E., Dochy, F., Onghena, P., & Struyven, K. (2010). Improving the effectiveness of peer feedback for learning. *Learning and Instruction*, 20(4), 304–315.
- Gregory, K., Cameron, C., & Davies, A. (2001). *Knowing what counts: Conferencing and reporting*. Connections Publishing.

- Guo, Y., Xu, J., & Liu, X. (2018). English language learners' use of self-regulatory strategies for foreign language anxiety in China. *System*, 76, 49–61. <https://doi.org/10.1016/j.system.2018.05.001>.
- Hargreaves, A., Earl, L., & Schmidt, M. (2002). Perspectives on alternative assessment reform. *American Educational Research Journal*, 30(1), 69–95.
- Harmer, J. (2001). The practice of English language teaching. Pearson. <http://www.scribd.com/Jeremy-Harmer-The-Practice-of-English-Language-Teaching-New-Edition1/d/15602107>.
- Harrington, M., & Sawyer, M. (1992). L2 working memory capacity and L2 reading skill. *Studies in Second Language Acquisition*, 14, 25–38.
- Harrison, M. (2009). *Oxford-living grammar: Pre-intermediate student's book pack*. Oxford University Press.
- Homayouni, M. (2022). Peer assessment in group-oriented classroom contexts: On the effectiveness of peer assessment coupled with scaffolding and group work on speaking skills and vocabulary learning. *Language Testing in Asia*, 12(1), 61. <https://doi.org/10.1186/s40468-022-00211-3>.
- Hsia, L.-H., Huang, I., & Hwang, G.-J. (2016). Effects of different online peer-feedback approaches on students' performance skills, motivation, and self-efficacy in a dance course. *Computers & Education*, 96, 55–71.
- Hu, N. (2022). Investigating Chinese EFL learners' writing strategies and emotional aspects. *LEARN Journal: Language Education and Acquisition Research Network*, 15(1), 440–468.
- Hyland, F. (2000). ESL writers and feedback: Giving more autonomy to students. *Language Teaching Research*, 4(1), 33–54.
- Hyland, K., & Hyland, F. (Eds.). (2019). *Feedback in second language writing: Contexts and issues* (pp. 1–22). Cambridge University Press.
- Iwai, Y. (2011). The effects of metacognitive reading strategies: Pedagogical implications for EFL/ESL teachers. *The Reading Matrix*, 11(2), 150–157.
- Jahara, S. F., Hussain, M., Kumar, T., Goodarzi, A., & Assefa, Y. (2022). The core of self-assessment and academic stress among EFL learners: The mediating role of coping styles. *Language Testing in Asia*, 12, 21. <https://doi.org/10.1186/s40468-022-00170-9>.
- Jiang, P., Namaziandost, E., Azizi, Z., & Razmi, M. H. (2022). Exploring the effects of online learning on EFL learners' motivation, anxiety, and attitudes during the COVID-19 pandemic: a focus on Iran. *Current Psychology*, 42. <https://doi.org/10.1007/s12144-022-04013-x>.
- Jiangmei, Y. U. A. N. (2023). Guidelines for preparing for, designing, and implementing peer assessment in online courses. *TOJET: The Turkish Online Journal of Educational Technology*, 22(1). 22111.pdf (tojet.net).
- Johnson, D. W., Johnson, R. T., & Smith, K. A. (2014). Cooperative learning: Improving university instruction by basing practice on validated theory. *Journal on Excellence in University Teaching*, 25(4), 1–26.
- Johnson, D. W., & Johnson, R. T. (1987). *Learning together and alone: cooperative, competitive, and individualistic learning*. Prentice-Hall, Inc.
- Joordens, S., Pare, D. E., & Pruesse, K. (2009). PeerScholar: an evidence-based online peer assessment tool supporting critical thinking and clear communication. *Proceedings of the 2009 International Conference on e-Learning* (pp. 236–240).
- Kargar Behbahani, H., & Kooti, M. S. (2022). Long-term Effects of Pictorial Cues, Spaced Retrieval, and Output-based Activities on Vocabulary Learning: The Case of Iranian Learners. *Glob Acad J Linguist Lit*, 4(3), 49–55.
- Khajavy, G. H. (2021). Modeling the relations between foreign language engagement, emotions, grit and reading achievement. Student Engagement in the Language Classroom. In P. Hiver, A. Al-Hoorie, & S. Mercer (Eds.), *Multilingual Matters* (pp. 241–259).
- Khajavy, G. H., MacIntyre, P. D., & Hariri, J. (2020). A closer look at grit and language mindset as predictors of foreign language achievement. *Studies in Second Language Acquisition*, 43(2), 379–402.
- Kidd, E., Donnelly, S., & Christiansen, M. H. (2018). Individual differences in language acquisition and processing. *Trends in Cognitive Sciences*, 22, 154–161.
- Kirkpatrick, R., & Gyem, K. (2012). Washback effects of the new English assessment system on secondary schools in Bhutan. *Language Testing in Asia*, 29(5). <https://doi.org/10.1186/2229-0443-2-4-5>.
- Kusuma, I., Mahayanti, N. W. S., Adnyani, L. D. S., & Budiarta, L. G. R. (2021). Incorporating E-portfolio with flipped classrooms: An in-depth analysis of students' speaking performance and learning engagement. *JALT CALL Journal*, 17(2), 93–111.
- Lam, R. (2017). Taking stock of portfolio assessment scholarship: From research to practice. *Assessing Writing*, 31, 84–97.
- Lam, R. (2018). Promoting self-reflection in writing: A showcase portfolio approach. In A. Burns & J. Siegel (Eds.), *International perspectives on teaching skills in ELT* (pp. 219–231). Palgrave MacMillan.
- Lam, R. (2019). Writing portfolio assessment in practice: individual, institutional, and systemic issues. *Pedagogies: An International Journal*, 15(3), 169–182.
- Lam, R. (2020). Writing portfolio assessment in practice: individual, institutional, and systemic issues. *Pedagogies: An International Journal*, 15(3), 169–182.
- Lee, I. (2017). Portfolios in classroom L2 writing assessment. In I. Lee (Ed.), *Classroom writing assessment and feedback in L2 school contexts* (pp. 105–122). Springer.
- Lee, I., & Coniam, D. (2013). Introducing assessment for learning for EFL writing in an assessment of learning examination-driven system in Hong Kong. *Journal of Second Language Writing*, 22(1), 34–50.
- Li, S. (2017). Cognitive differences and ISLA. In S. Loewen & M. Sato (Eds.), *The Routledge handbook of instructed second language acquisition* (pp. 396–417). Routledge.
- Li, H., Xiong, Y., Zang, X., Kornhaber, M. L., Lyu, Y., Chung, K. S., & Suen, H. K. (2016). Peer assessment in the digital age: A meta-analysis comparing peer and teacher ratings. *Assessment & Evaluation in Higher Education*, 41(2), 245–264.
- Li, H., Xiong, Y., Hunter, C. V., Guo, X., & Tywoniw, R. (2020). Does peer assessment promote student learning? A meta-analysis. *Assessment & Evaluation in Higher Education*, 45(2), 193–211.
- Li, S. (2023). Working memory and second language learning: a critical and synthetic review. *The Routledge handbook of second language acquisition and psycholinguistics* (pp. 348–360). <https://doi.org/10.4324/9781003018872-32>.
- Lipnevich, A. A., & Smith, J. K. (2022). Student – feedback interaction model: revised. *Studies in Educational Evaluation*, 75, 101208. <https://doi.org/10.1016/j.stueduc.2022.101208>.

- Lynch, M. M. (2001). Effective Student Preparation for Online Learning. <http://ts.mivu.org/default.asp?show=article&id=901>.
- Locke, E. A., McClellan, K., & Knight, D. (1996). Self-esteem and work. *International Review of Industrial/organizational Psychology*, 11, 1–32.
- Mackey, A., & Gass, S. (Eds.). (2022). *Second language research: Methodology and design* (3rd ed.). Routledge.
- Mackey, A., Philp, J., Fujii, A., Egi, T., & Tatsumi, T. (2002). Individual differences in working memory, noticing of interactional feedback and L2 development. In P. Robinson (Ed.), *Individual differences and instructed language learning* (pp. 181–208). Benjamins.
- Malecka, B., Boud, D., & Carless, D. (2020). Eliciting, processing and enacting feedback: mechanisms for embedding student feedback literacy within the curriculum. *Teaching in Higher Education*, 1–15. <https://doi.org/10.1080/13562517.2020.1754784>
- Masruria, W. W., & Anam, S. (2021). Exploring self-assessment of speaking skill by EFL high school students. *Linguistic, English Education and Art (LEEA) Journal*, 4(2), 387–400.
- Mathur, M., & Mahapatra, S. (2022). Impact of ePortfolio assessment as an instructional strategy on students' academic speaking skills: An experimental study. *Computer Assisted Language Learning*, 23(3), 1–23.
- Mediha, N., & Enisa, M. (2014). A comparative study on the effectiveness of using traditional and contextualized methods for enhancing learners' vocabulary knowledge in an EFL classroom. 5th World Conference on Educational Sciences, (116), 3443–3448. <https://doi.org/10.1016/j.sbspro.2014.01.780>.
- Miyake, A., & Shah, P. (Eds.). (1999). *Models of working memory*. Cambridge University Press.
- Moghim, A. (2022). On the comparative impact of self-assessment and peer assessment on Iranian male and female EFL learners' accuracy in speech. *Contemporary Educational Research Journal*, 12(4), 204–213.
- Mostert, M., & Snowball, J. D. (2013). Where angels fear to tread: Online peer-assessment in a large first-year class. *Assessment & Evaluation in Higher Education*, 38(6), 674–686.
- Mphahlele, R. S. (2022). Digital assessment literacy in online courses (formative/summative): Rethinking assessment strategies in the open distance and e-learning institutions. *Handbook of research on managing and designing online courses in synchronous and asynchronous environments* (pp. 404–417). IGI Global.
- Namaziandost, E., & Çakmak, F. (2020). An account of EFL learners' self-efficacy and gender in the Flipped Classroom Model. *Education and Information Technologies*, 25, 4041–4055.
- Namaziandost, E., Heydarnejad, T., & Rezai, A. (2022). Iranian EFL teachers' reflective teaching, emotion regulation, and immunity: examining possible relationships. *Current Psychology*, 42. <https://doi.org/10.1007/s12144-022-03786-5>.
- Naserpour, A., & Zarei, A. A. (2021). Visually-mediated instruction of lexical collocations: The role of involvement load and task orientation. *Iranian Journal of Learning & Memory*, 3(12), 39–50.
- Nicol, D. (2020). The power of internal feedback: Exploiting natural comparison processes. *Assessment and Evaluation in Higher Education*, 46(5), 756–778.
- Norozi, O., Biemans, H., & Mulder, M. (2016). Relations between scripted online peer feedback processes and quality of written argumentative essay. *The Internet and Higher Education*, 31, 20–31.
- Nourdad, N., & Banagozar, M. A. (2022). The effect of e-portfolio assessment on EFL vocabulary learning and retention. *Indonesian Journal of Applied Linguistics*, 12(2), 466–475.
- Nunan, D. (2004). *Task-based language teaching*. Cambridge University Press.
- Pallant, J. (2020). *SPSS survival manual: a step by step guide to data analysis using IBM SPSS*. McGraw-Hill Education (UK).
- Pallathadka, H., Xie, S., Alikulov, S., Al-Qubanchi, H. S., Alshahrani, S. H., Yunting, Z., & Behbahani, H. K. (2022). Word recognition and fluency activities' effects on reading comprehension: an Iranian EFL learners' experience. *Education Research International*, 2022. <https://doi.org/10.1155/2022/4870251>.
- Panadero, E., & Romero, M. (2014). To rubric or not to rubric? The effects of self-assessment on self-regulation, performance, and self-efficacy. *Assessment in Education Principles Policy and Practice*, 21(2), 133–148.
- Patra, I., Suwondo, T., Mohammed, A., Alghazali, T., Mohameed, D. A. A. H., Hula, I. R. N., & Behbahani, H. K. (2022). The effects of processing instruction and output-based activities on grammar learning: the mediating role of working memory. *Education Research International*, 2022. <https://doi.org/10.1155/2022/3704876>.
- Paulson, L. F., Paulson, P. R., & Meyer, C. A. (1991). What makes a portfolio a portfolio? *Educational Leadership*, 48(5), 60–63.
- Pawlak, M. (2017). Overview of learner individual differences and their mediating effects on the process and outcome of L2 interaction. In L. Gurzynski-Weiss (Ed.), *Expanding individual difference research in the interaction approach* (pp. 19–40). John Benjamins.
- Putro, H. P. N., Hadi, S., Rajiani, I., Abbas, E. W., & Mutiani, (2022). Adoption of e-learning in Indonesian higher education: Innovation or irritation? *Educational Sciences: Theory and Practice*, 22(1), 36–45.
- Rezai, A., Namaziandost, E., & Rahimi, S. (2022a). Developmental potential of self-assessment reports for high school students' writing skills: A qualitative study. *Teaching English as a Second Language Quarterly (Formerly Journal of Teaching Language Skills)*, 41(2), 163–203.
- Rezai, A., Namaziandost, E., & Amraei, A. (2023). Exploring the effects of dynamic assessment on improving Iranian Quran learners' recitation performance. *Critical Literary Studies*, 5(1), 159–176. <https://doi.org/10.34785/J014.2023.010>.
- Rezai, A., Rahul, D. R., Asif, M., Omar, A., & Reshad Jamalyar, A. (2022b). Contributions of E-portfolios assessment to developing EFL learners' vocabulary, motivation, and attitudes. *Education Research International*. <https://doi.org/10.1155/2022/5713278>.
- Rezai, M. J. (2015). *ABC of SPSS for students of applied linguistics*. Yazd University Press.
- Ritonga, M., Tazik, K., Omar, A., & Saberi Dehkordi, E. (2022). Assessment and language improvement: The effect of peer assessment (PA) on reading comprehension, reading motivation, and vocabulary learning among EFL learners. *Language Testing in Asia*, 12(1), 1–17.
- Roschelle, J., & Teasley, S. D. (1995). The construction of shared knowledge in collaborative problem solving. *Computer supported collaborative learning* (pp. 69–97). Springer.
- Ruiz, S., Tagarelli, K. M., & Rebuschat, P. (2018). Simultaneous acquisition of words and syntax: Effects of exposure condition and declarative memory. *Frontiers in Psychology*, 9, 1168.
- Sawyer, R. K. (2006). Educating for innovation. *Thinking Skills and Creativity*, 1(1), 41–48.

- Schmitt, D., Schmitt, N., & Mann, D. (2011). *Focus on vocabulary 1: Bridging vocabulary*. Longman.
- Shahnazari, M. (2013). The development of a Persian reading span test for the measure of L1 Persian EFL learners' working memory capacity. *Applied Research on English Language*, 2(2), 107–116. <https://doi.org/10.22108/are.2013.15473>
- Shih, C. M. (2010). The washback of the general English proficiency test on university policies: A Taiwan case study. *Language Assessment Quarterly*, 7(3), 234–254.
- Slavin, R. E. (2011). *Student team learning: a practical guide to cooperative learning*, (3rd ed.,). National Education Association.
- Song, B., & August, B. (2002). Using portfolios to assess the writing of ESL students: A powerful alternative? *Journal of Second Language Writing*, 11(1), 49–72.
- Steen-Utheim, A., & Hopfenbeck, T. N. (2018). To do or not to do with feedback: A study of undergraduate students' engagement and use of feedback within a portfolio assessment design. *Assessment & Evaluation in Higher Education*, 44(1), 80–96.
- Sultana, F., Lim, C. P., & Liang, M. (2020). E-portfolios and the development of students' reflective thinking at a Hong Kong University. *Journal of Computers in Education*, 7, 277–294.
- Tagarelli, K. M., Ruiz, S., Moreno-Vega, J. L., & Rebuschat, P. (2016). Variability in second language learning: The roles of individual differences, learning conditions, and linguistic complexity. *Studies in Second Language Acquisition*, 38, 293–316.
- Takaroucht, K. (2021). The effect of self-assessment on the development of EFL reading comprehension skills. *Journal of English Education and Teaching*, 5(2), 231–247.
- Tarricone, P. (2011). *The taxonomy of metacognition*. Psychology Press.
- Teng, M. F., & Zhang, D. (2021). The associations between working memory and the effects of multimedia input on L2 vocabulary learning. *International Review of Applied Linguistics in Language Teaching*. <https://doi.org/10.1515/iral-2021-0130>.
- Torkabad, M. G., & Fazilatfar, A. M. (2014). Textual Enhancement and Input Processing Effects on the Intake of Present and Past Simple Tenses. *Procedia - Social and Behavioral Sciences*, 98, 562–571. <https://doi.org/10.1016/j.sbspro.2014.03.452>
- Verenikina, I. (2008). *Scaffolding and learning: its role in nurturing new learners*. <https://ro.uow.edu.au/edupapers/43>.
- Vygotsky, L. S. (1987). Thinking and speech. In R. W. Rieber & A. S. Carton (Eds.), *The collected works of L. S. Vygotsky: Vol. 1: Problems of general psychology* (pp. 39–285). Plenum.
- Webb, N. M. (2008). Learning in small groups. In T. L. Good (Ed.), *21st century education: A reference handbook* (pp. 203–211). Sage.
- Wei, X. (2020). Assessing the metacognitive awareness relevant to L1-to-L2 rhetorical transfer in L2 writing: the cases of Chinese EFL writers across proficiency levels. *Assessing Writing*, 44, 100452. <https://doi.org/10.1016/j.asw.2020.100452>.
- William, D. (2018). How can assessment support learning? A response to Wilson and Shepard, Penuel, and Pellegrino. *Educational Measurement: Issues and Practice*, 37(1), 42–44.
- Wilkins, D. A. (1972). *Linguistics in Language Teaching*. Cambridge: MIT Press.
- Woolley, G. (2011). *Reading comprehension. Assisting children with learning difficulties*. Queensland: Springer.
- Xu, Y., & Brown, G. T. L. (2016). Teacher assessment literacy in practice: A reconceptualization. *Teaching and Teacher Education*, 58, 149–162.
- Yang, M., Badger, R., & Yu, Z. (2006). A comparative study of peer and teacher feedback in a Chinese EFL writing class. *Journal of Second Language Writing*, 15(3), 179–200.
- Zhang, Y. M. (2022). The research on critical thinking teaching strategies in college English classroom. *Creative Education*, 13, 1469–1485.
- Zhao, H. (2010). Investigating learners' use and understanding of peer and teacher feedback on writing: A comparative study in a Chinese English writing classroom. *Assessing Writing*, 15(1), 3–17.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Submit your manuscript to a SpringerOpen<sup>®</sup> journal and benefit from:**

- Convenient online submission
- Rigorous peer review
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

---

Submit your next manuscript at ► [springeropen.com](https://www.springeropen.com)

---