Studies in Second Language Acquisition (2023), **45**, 212–233
doi:10.1017/S0272263122000146

RESEARCH ARTICLE

The effects of proficiency level and dual-task condition on L2 self-monitoring behavior

Ghadah Albarqi[1] and Parvaneh Tavakoli[2]*

[1]Taif University, Taif, Saudi Arabia; [2]University of Reading, Reading, UK
*Corresponding author. E-mail: p.tavakoli@reading.ac.uk

(Received 10 September 2021; Revised 06 March 2022; Accepted 13 March 2022)

Abstract

The current study examined the effects of task condition (TC; single vs. dual) and proficiency level (PL) on self-monitoring of second language (L2) speakers. Data were collected from sixty-six female L2 learners of English performing two speaking tasks under two task conditions. While performance in the single-task condition involved only narrating a picture-based oral narrative, the dual-task condition involved performing the same oral narrative as well as a secondary task. Factor analysis, MANOVA, and two-way ANOVAs were used to examine the effects of PL and TC on a range of self-monitoring measures. The results indicated that the higher proficiency learners made significantly fewer filled pauses, repetitions, and hesitations, and a higher ratio of error correction and error-free clauses than the lower proficiency learners. These results suggest that with the development of proficiency L2 learners' performance becomes more fluent, and a more active and effective monitoring process seems to be at work. Compared to the single-task condition, performance in the dual-task condition led to significantly more repetitions implying the increased demand of TC triggers more dysfluency. These results are discussed in relation to the L1 monitoring models.

Understanding self-monitoring is central to understanding speech production, and therefore it is plausible to argue that "any theory of language production is incomplete without a theory of self-monitoring" (Hartsuiker, 2014, p. 417). Self-monitoring, referring to revising speech before and after articulation (Levelt, 1983), is observable through changes speakers make when inspecting and revising speech (e.g., hesitations and corrections). Self-monitoring has been investigated frequently in first language (L1) studies (Blackmer & Mitton, 1991; Oomen & Postma, 2001, 2002; Postma et al., 1990; Seyfeddinipur et al., 2008), with only a few second language (L2) studies examining L2 speakers' self-monitoring (e.g., Ahmadian & Tavakoli, 2014; Declerck & Kormos, 2012; Kormos, 2000b).

According to Levelt (1989), L1 speech production involves four components: conceptualization, formulation, articulation, and monitoring. The conceptualizer generates a preverbal message that is sent to the formulator where the required lemmas are activated and retrieved from the mental lexicon and put into syntactic structures

https://doi.org/10.1017/S0272263122000146 Published online by Cambridge University Press

through the process of grammatical encoding. The verbal message formulated will then move to the articulator that executes the phonetic plan of the speech. The production process is subject to monitoring during and after speech is produced, where the message and its linguistic form are revised. The L1 speech processes are hypothesized to be incremental, automatic, and parallel (Levelt, 1989), that is, L1 processing takes place speedily, effortlessly, and simultaneously. Unlike L1 speech, L2 speech processes are largely controlled especially at lower levels of proficiency (Kormos, 2006; Skehan, 2009). Therefore, lower proficiency level speakers are expected to engage in self-monitoring differently from those at higher proficiency levels (further discussion in "L2 Proficiency and Self-Monitoring" section). This hypothesis, although an under-researched area in self-monitoring studies, is of central interest to the current study.

L2 speech production is already a demanding process vulnerable to external cognitive demands, and as such performing a second task in parallel while engaged in L2 speech production, commonly known as dual-task condition (Declerck & Kormos, 2012; Oomen & Postma, 2002), would make the process even more demanding. The imposition of the dual-task condition means L2 speakers must divide their attentional resources between the speech production process and the secondary task; this is expected to affect L2 production processes. The use of dual-task condition and its impact on self-monitoring has rarely been investigated in L2 studies. This is the gap we aim to help fill by investigating the effects of dual-task condition and proficiency on L2 self-monitoring.

The effects of dual-task condition on self-monitoring (e.g., self-repair) have been of interest to both psycholinguistics and task-based language teaching and learning (TBLT) research. Psycholinguistic studies (e.g., Oomen & Postma, 2001, 2002; Postma et al., 1990), mainly interested in exploring the effects of cognitive load on psycholinguistic processes involved in language production, consider self-monitoring central to understanding speech production models. TBLT researchers (e.g., Robinson, 2001; Skehan, 2009), however, are primarily interested in self-monitoring in the light of L2 acquisitional processes. This body of research often evaluates the effects of task cognitive load on L2 performance within the complexity, accuracy, and fluency (CAF) framework, and claims that a careful analysis of L2 performance would help develop a better understanding of how tasks affect L2 acquisitional processes (Robinson, 2001; Skehan, 2009). Despite the differences, both disciplines are interested in how cognitive load affects monitoring behavior. The current study draws on a psycholinguistic approach to operationalizing cognitive load through the dual-task condition, but as will be discussed in the following text, it also draws on TBLT research in other aspects of the research design. In addition, aspects pertaining to self-monitoring (self-repair, accuracy, and disfluency) that have been studied in psycholinguistics and TBLT fields, will be examined in the current study (see the next section).

Literature review

Self-monitoring

Levelt's (1983) Perceptual Loop Theory (PLT) is adopted as the theoretical framework of this study. The PLT proposes that there is a single central monitor that is located within the conceptualizer, receiving feedback from three channels, known as loops (Levelt, 1989): the *perceptual loop*, the *inner loop*, and the *auditory loop*. PLT suggests that the monitor can only inspect the end products of the processing components using its loops, and that each loop inspects the outcome of a processing component. The

https://doi.org/10.1017/S0272263122000146 Published online by Cambridge University Press

perceptual loop checks the preverbal plan that is the end product of the conceptualizer; the inner loop inspects the phonetic plan that is the end product of the formulator; and the auditory loop scrutinizes the end product of the articulator (the overt speech). Although there are other theories of L1 self-monitoring,[1] Levelt's (1983) PLT is the most viable and empirically supported model in the field of psycholinguistics (Levelt, 1999; Levelt et al., 1999; Oomen & Postma, 2001, 2002; Postma et al., 1990; Seyfeddinipur et al., 2008).

The operations of the monitoring in the three loops result in two different types of repairs: *covert* and *overt*. The perceptual and inner loops' operations lead to *covert repair*, whereas the operations of the auditor loop result in *overt repair*. The key difference between the two types of repairs lies in whether they can be directly observed. According to PLT, covert repair is prearticulatory and reflected through disfluencies such as hesitations, repetitions, and filled pauses. Hesitation in this sense refers to repeating part(s) of a word without producing it in full. This is different from repetition that entails a complete reiteration of the same word or phrase. It is assumed that when a speaker predicts an error or faces a challenge in the production process (e.g., retrieving a word), she or he makes pauses or repetitions to buy time to correct the error or address the challenge before articulation. For example, in the utterance "go to a red, red node," repeating the same word is considered as a covert repair although no change is involved (Levelt, 1983, p. 45). Given its abstract nature, covert repair cannot be easily classified especially in L2 processing where disfluencies occur due to a range of purposes, including linguistic issues, online planning, and solving communication problems (De Jong et al., 2015; Derwing et al., 2009; Dörnyei & Kormos, 1998) (see definitions in Table 3). *Overt repair*, which can be directly observed and therefore is more reliably classified, includes different repair types: *D-repair* (different-information-repair), *A-repair* (appropriateness-repair), and *E-repair* (error-repair) (see Table 1). Repair production involves three phases that are error-to-cut-off, cut-off-to-repair, and repair execution. Further details of these phases are provided in the "Method" section. While the principles of Levelt's (1989) PLT model have been tested by a multitude of L1 studies, it is surprising that very few studies have examined this model in the L2 context. In the section that follows, we provide a summary of the key studies conducted in this area.

Cognitive resources and L2 self-monitoring

Self-monitoring as viewed in Levelt's PLT (1983, 1989, 1992) is a conscious process with limited resources (Postma, 2000), as its functioning relies on a human's limited working memory capacity (Levelt, 1989). In addition, self-monitoring is considered a demanding process because it requires checking both one's own speech and the speech of others to ensure comprehension and communication (Levelt, 1983, 1989). The literature presents a line of research that examines the association between self-monitoring and cognitive resources through manipulating dual-task demands. The rationale for employing this method draws on a principle of PLT that states that self-monitoring is sensitive to contextual effects (Levelt, 1983). Dual-task condition, that is, performing two tasks simultaneously, is regarded as an appropriate method to examine the effects

[1] Theories of L1 self-monitoring include the Node Structure Theory and the Production-Based Theory. There are some differences between these theories including the location of the monitor, its capacity, and its relationship with working memory.

https://doi.org/10.1017/S0272263122000146 Published online by Cambridge University Press

Table 1. Repair types

Repair Types	Definition	Examples
D-repair	Abandoning the message and replacing it with a different one.	|uh then the students wen uh 0.38 the teacher try to call the 911|
A-repair	Modifying the way in which an utterance is produced to become more appropriate or accurate in a particular context.	|when the stor uh the thunderstorm comes|
E-repair	Correcting lexical (e.g., phrases, idioms, preposition); grammatical (e.g., inflectional morphologies, auxiliaries); or phonological errors (e.g., intonation, stress, phoneme).	|all of them was uh 0.28 all of them were |

of cognitive resource depletion on self-monitoring (Broos et al., 2018; Oomen & Postma, 2002). To the best of our knowledge, there is only one L2 study (Declerck & Kormos, 2012) to date that has employed dual-task condition to examine L2 self-monitoring.

Declerck and Kormos (2012) investigated the effects of single and dual-task conditions on the efficiency of L2 monitoring on 20 Hungarian speakers belonging to lower and higher proficiency levels. They used a network description task to collect speech samples. This task requires learners to describe the movement path of a red dot moving differently on each network. A finger-tapping task was used as a secondary task. The results suggested that the dual-task condition had a negative effect on the accuracy of lexical selection and the efficiency of error-correction, but it did not affect fluency, speed of error-detection, or the overall repair frequency. The results also indicated that the ratio of error-correction decreased more significantly in the higher proficiency than the lower proficiency learners in the dual-task condition. This means that advanced speakers corrected less errors in the more demanding task condition. It has been assumed that self-monitoring was affected by conscious decisions taken by the L2 learners on whether or not to correct their errors (Declerck & Kormos, 2012; Mackay, 1992).

While this study provided valuable theoretical and methodological evidence about L2 self-monitoring under dual-task condition, it had some limitations that future research was called upon to address (Declerck & Kormos, 2012). First, the dual-task condition was not operationalized systematically. The similarity of the concurrent tasks employed in Declerck and Kormos (2012) was criticized as not being sufficiently demanding, which means that the two concurrent tasks might not have consumed the available cognitive resources (Duncan, 1980; Wickens, 2007). This is a limitation that the current study aims to address (see the “Method” section). Furthermore, the choice of tasks in their study, that is, the network description, might have led to some inadvertent consequences in terms of the monitoring foci. In the task, while task completion involved using language of direction and shape, it was not demanding in terms of conceptualization or generation of ideas. The task, however, is considered demanding in terms of lexical choices (Declerck & Kormos, 2012). Following Declerck and Kormos’s (2012) conclusion, we agree that their choice of task had an impact on how learners behave during L2 self-monitoring, that is, paying more attention to the correction of lexical errors. In addition, researchers recommend that dual-task studies need to employ different forms of secondary tasks “that are more likely to be encountered in real-life language use situations” (Révész et al., 2016, p. 735). To address these limitations, we are using two concurrent verbal tasks (see the “Method” section).

L2 Proficiency and self-monitoring

As discussed earlier, studies investigating effects of proficiency on L2 self-monitoring are often motivated by the question of whether the production process becomes more automatic with proficiency development. Following the literature in this area (DeKeyser, 2001; Segalowitz, 2003; Tavakoli, 2019), we assume that the automatization process is characterised by qualities such as ballistic, parallel, and attention-free processing, which predictably “draws on implicit-procedural knowledge” (Ortega, 2009, p. 85). The more automatic processing, in effect, enables L2 speakers to use the freed-up resources to deal with different aspects of performance (e.g., to check the appropriateness of their speech), and to be engaged in other tasks if needed. Research evidence suggests that certain subprocesses of the Formulator can reach automaticity (i.e., performing with reduced cognitive effort) such as lexical access (e.g., Hulstijn et al., 2009; Pellicer-Sánchez, 2015) and syntactic encoding (e.g., Robinson, 1997). Therefore, more attentional resources become available for other processes including L2 self-monitoring (Kormos, 2000b).

Given that development of proficiency is associated with automatization (DeKeyser, 2013; Tavakoli, 2019), it is expected that proficiency development affects different aspects of speech production including self-monitoring behavior. The impact of automatization on self-monitoring behavior can be observed through a range of different means, from measuring pauses to investigating repair behavior and examining the rate and success of self-correction. In this article, we are particularly interested in overt repair (self-repair) and covert repair (disfluency) as they are considered as distinctive features of L2 self-monitoring behavior (Levelt’s, 1983, 1989). We are also interested in accuracy as it is perceived as the main aim of L2 self-monitoring process (Gilabert, 2007; Kormos, 1999), and therefore, examining it as an end-product of self-monitoring is central to understanding the monitoring behavior. These key terms will be discussed in detail in what follows.

Accuracy in general refers to a decrease in the number of errors, indicating development in the underlying speech processes (DeKeyser, 2013), particularly at the formulator subprocesses (syntactic, lexical, and phonological encoding) where most errors occur (Kormos, 2006). Errors can be examined in different forms, but the two main types are accuracy process and accuracy product measures.

Disfluency features, according to the PLT, are produced as a corrective reaction to expected errors. Disfluency features, also of interest to TBLT researchers, are often examined in terms of pauses, hesitations, and repetitions. These features are reported to be among the best indicators of L2 proficiency development (De Jong, 2018; Révész et al., 2016). The existing research evidence (e.g., De Jong et al., 2015; Skehan, 2009; Tavakoli, 2019; Tavakoli et al., 2020) suggests that with the development of proficiency disfluency features decrease and speech becomes more fluent. The more automatic L2 processing and production at higher proficiency levels allows L2 speakers, for example, to have a faster lexical retrieval and less need for pausing to buy time when facing a challenge in the production process (Skehan, 2009; Suzuki, 2021; Tavakoli & Wright, 2020). The present study focuses on three disfluency features (i.e., filled pauses, hesitations, and repetitions) as indicators of self-monitoring, (see operationalization in Table 3).

Self-repair features, such as repair type and repair duration, are commonly examined in self-monitoring studies. While in L1 research employing Levelt’s (1989) self-repair taxonomy, discussed in detail in the text that follows, has been common, few L2

https://doi.org/10.1017/S0272263122000146 Published online by Cambridge University Press

studies have used this taxonomy to study L2 self-repair. Van Hest (1996) examined L1 and L2 self-repairs at three proficiency levels (beginning, intermediate, and advanced). The participants were 30 native speakers of Dutch learning English as an L2. The findings suggested that advanced L2 learners produced less error-repair (see the definition in Table 1), and more appropriateness-repair than the intermediate and lower proficiency learners. The findings of the study were limited as it did not examine the temporal phases of repair in relation to proficiency levels, or under different speaking tasks. Kormos (2000b) examined the effects of proficiency on repair types in terms of frequency of repairs and the differences between the temporal phases. Examining 30 L2 learners at advanced, upper-intermediate, and preintermediate levels, Kormos's (2000b) findings suggested that the higher proficiency learners produced more A-repair and less E-repair than lower proficiency learners. This was interpreted in the light of the fact that higher proficiency learners worked with more automatic processes.

Self-repair has also been examined in a number of TBLT studies (e.g., Lambert et al., 2017; Tavakoli & Skehan, 2005; Wang & Skehan, 2014). These studies, primarily investigating the effects of task design (e.g., its cognitive load) on L2 performance, used the CAF framework to analyze language in which self-repair is considered a subcategory of fluency. Such studies usually adopt Tavakoli and Skehan's (2005) taxonomy to analyze aspects of fluency in terms of speed, breakdown, and repair. Repair fluency in this taxonomy includes repetitions, replacements, reformulations, and false starts. Although this taxonomy has become central to understanding L2 fluency, it will not provide an effective and comprehensive framework for analyzing and understanding L2 self-monitoring processes. We argue that it is important for self-monitoring research to employ a taxonomy that allows for a careful analysis of the different types of repairs (e.g., A-repair and E-repair) and their duration. Adopting Levelt's (1989) classification will also allow us to compare our findings with those reported in previous research.

Rationale, aim, and research questions of the study

As discussed earlier, there are few studies investigating the effects of dual-task condition on L2 self-monitoring behavior. Among those studies, only one has examined the effects of dual-task condition on L2 self-monitoring (Declerck & Kormos, 2012). Our primary aim is to examine how resource limitations manipulated along dual-task condition can influence L2 self-monitoring. The study is also interested in finding out whether such effects, if any, are different at different levels of proficiency. We also aim to address the limitations found in previous studies (e.g., ibid.) in terms of choice of primary and secondary tasks. To develop a more in-depth understanding of how L2 self-monitoring functions, we will examine a wide range of monitoring measures which will be discussed in detail in the "Method "section. The research questions of the study are:

1. How does dual-task demand affect L2 self-monitoring behavior in terms of disfluency, repair types, duration of repair, and accuracy?
2. How does proficiency affect self-monitoring behavior in terms of disfluency, repair types, duration of repair, and accuracy?
3. Is there an interaction between dual-task condition and proficiency on L2 self-monitoring in terms of disfluency, repair types, duration of repair, and accuracy?

https://doi.org/10.1017/S0272263122000146 Published online by Cambridge University Press

Method

Design

The study had a between-participant factorial design in which task condition (single or dual task) and proficiency level are between-participants independent variables. Each participant performed two different picture prompts under either single or dual-task condition (see Appendix A). Using two picture stories allowed us to investigate a more diverse set of linguistic forms and a richer performance from each participant. A range of measures of self-monitoring, discussed in the following text, were the dependent variables of the study.

Participants

Data were collected from 66 Arabic L1 speaking female undergraduates, aged between 18 and 23, who volunteered to take part in the study.[2] They were all majoring in English at a University in Saudi Arabia and took L2 English courses in linguistics and literature. Forty of the participants were in first year and 26 in second year of their bachelor's degree. For this reason, the participants are regarded as a special group of learners due to their knowledge of English. Prior to their participation in the study and based on the results of an institutionally developed grammar placement test, they had been placed at levels corresponding to A2 and B1 of the Common European Framework of Reference for languages (CEFR). To examine their oral proficiency for the purpose of the study, however, we used an Elicited Imitation Test (see the following text). The participants had similar L2 learning background in that they had received 8–9 years of formal English instruction at school and university and had not lived in an English-speaking country before. All the participants volunteering to take part in the study gave formal consent for their participation.

Instruments

Language proficiency test

To examine the participants' proficiency, we used Wu and Ortega's (2013) Elicited Imitation Test (EIT). The rationale for using an EIT in this study was based on previous research calling for a valid and reliable measurement of L2 spoken ability (e.g., Tremblay, 2011) when examining the speech production process. EIT, validated in several previous studies (e.g., Ellis, 2005; Erlam, 2006; Gaillard & Tremblay, 2016; Wu & Ortega, 2013), allows researchers to examine not only the speakers' mastery of the L2 ability but their procedural knowledge in their L2 speaking. A recent meta-analysis of studies investigating the use of EIT as a measure of proficiency confirms that EITs are "a fairly dependable measure of L2 proficiency" (Kostromitina & Plonsky, 2021, p. 18). Other researchers argue that since completing EIT relies on fast language processing and producing speech in real time, using EIT is suited to measuring procedural oral language ability and degree of automaticity in their speech (Suzuki & DeKeyser, 2015). Finally, we chose EIT as previous research has suggested that speech samples elicited by EIT are comparable to spontaneous language production (Baten & Cornillie, 2019; Erlam, 2006), and therefore suitable for examining L2 processing.

[2] The choice of female participants in this study was due to practical reasons in the context of data collection.

https://doi.org/10.1017/S0272263122000146 Published online by Cambridge University Press

Wu and Ortega's (2013) EIT comprises 19 sentences with an increasing number of syllables (from 7 to 19) spoken by a native speaker of English. The participants were asked to repeat as much of the sentence as they could after being given only one chance to listen to and repeat the sentence. Sentences were given scores ranging from 0 to 4 points. Each participant was given a maximum of four points for a perfect repetition (repeat the whole sentence correctly), three for accurate content repetition, two for changes that affected meaning (in content or form), one for repetition of half of the sentence or less, and zero for a single-word repetition or failure to repeat anything.

Picture prompts

Oral narrative picture prompts were used to elicit L2 learners' oral performance in the current study. Oral narrative tasks are frequently used in L2 classrooms and considered an ecologically valid task in L2 studies (Prefontaine & Kormos, 2015; Tavakoli & Foster, 2011). Oral narratives have also been frequently used in TBLT research as it is effective in collecting samples of speech in a semicontrolled manner (see Suzuki, 2021, for a full discussion). Several dimensions of task design, recommended by the literature (De Jong & Vercellotti, 2016; Faez & Tavakoli, 2019; Tavakoli & Foster, 2011), including the number of elements, task structure, and storyline complexity, were considered when developing the tasks. It has been argued that a single prompt for each task condition could result in a confounding effect, as the prompts might not elicit similar speech samples (De Jong & Vercellotti, 2016). Some researchers argue that seemingly similar tasks differed in the language they elicited (De Jong & Vercellotti, 2016). As such, following De Jong and Vercellotti's (2016) guidelines, two comparable oral narrative picture prompts were designed. The two tasks had very similar linguistic demands in terms of vocabulary and structures required for task completion. To ensure they elicited vocabulary of the same complexity level, the written scripts of the stories were submitted to VocabProfilers.[3] The analysis suggested they elicited comparable vocabulary in terms of their frequency. The similarity in terms of number of elements and prompts, task structure, and storyline complexity helped ensure they had similar cognitive demands. Although the communicative nature of oral narratives is believed to encourage attention to both meaning and linguistic form (Skehan, 2009), we are aware that learners may vary in what they attend to, and some may prioritize one over the other. The picture stories were counterbalanced in the two task conditions to reduce any potential task effects (see Appendix A).

Task condition

Task condition included primary and secondary tasks. The primary task was narrating a picture story (see preceding text), and the secondary task involved bubbles appearing on a computer screen simultaneously as the L2 learners were narrating the picture story. A bubble appeared every five seconds on the screen, stayed for only five seconds and disappeared if no response was made. Each bubble contained the name of either an animate or inanimate noun (e.g., cat, dog, car). The names were written in English (the target language). The participants were asked to press the Z button on the computer's keyboard if the word was an animate name object, and the M button if it was inanimate. The two keyboard buttons (Z and M) were marked with Arabic translations of "animate" and "inanimate" (حي and جماد), respectively, to make it easy for the

[3] VocabProfiler classifies words according to their frequency levels (e.g., 1K, 2K, offlist) (Cobb, 2017).

participants to focus on the experiment (Albarqi, 2018). E-Prime Psychological Software (3.0)[4] was used to design and run the dual-task experiment. As discussed earlier, our choice of the secondary task was aimed at addressing the limitations of previous research (e.g., Declerck & Kormos, 2012), which was done by operationalizing the dual-task condition more systematically and designing a more demanding secondary task.

To ensure that the dual-task condition was systematically operationalized in the study, we followed the guidelines provided by Wickens (2007) in his limited-capacity multiple-resources model. Assuming that performing two tasks simultaneously is more difficult if the two tasks draw on the same resource pool, Wickens (2007) argues that the degree of similarity between tasks should be assessed in terms of which resource pools they depend on. Wickens (2007) proposes three dimensions to define which resource pools the tasks draw on: *perceptual modality* (the processing of visual or auditory modes of language), *processing code* (verbal and nonverbal or spatial processing demands), and *processing stages* (the stages of processing in which the task is involved). Wickens (2007) maintains that performing two tasks simultaneously is easier if (1) the input is received across different modalities rather than within the same modality (e.g., it is easier to read and listen than to read two texts at the same time), (2) the tasks require different processing codes rather than the same code (e.g., listening and driving is easier than listening and reading), and (3) the tasks are going through different stages of processing (e.g., perceptual, cognition and responding) (ibid.).

In our study the primary task, oral narrative picture prompts, was of visual modality and verbal processing code, and involved the processing stages of perception, cognition, and verbal responding. The secondary task was similarly of visual modality and verbal processing. Given the similarity of the dimensions of the secondary task to the primary task, we considered the secondary task would, to a great extent, increase the demands of performing the primary task. The secondary task comprised 20 trials and 4 practice trials to familiarize participants with the experiment.

Procedures

After explaining the general aim of the study and gaining informed consent from the participants, the EIT was administered to each participant individually. The participants were then randomly assigned to the single- or dual-task condition. The participants were then asked to narrate the picture stories, under either the single- or dual-task condition. In the single-task condition, the participants looked at the picture prompts shown on a Microsoft PowerPoint and narrated the story. Under the dual-task condition, the participants were asked to perform the secondary task simultaneously as they were narrating the picture stories. They were asked to pay equal attention to both tasks. Oral performances were recorded on a digital voice-recording machine and dual-task performances were recorded on E-Prime software. All the instructions during the data collection were given in students' L1, namely Arabic.

Measures

A total of 132 speech samples (66 × 2 performances) were collected from participants. Following previous studies (e.g., Duran Karaoz & Tavakoli, 2020; Tavakoli, 2011;

[4] E-Prime Psychological Software is suitable for computerized experimental design because it handles milliseconds precision timing efficiently (Schneider et al., 2002).

https://doi.org/10.1017/S0272263122000146 Published online by Cambridge University Press

Tavakoli et al., 2016), 1 minute of performance per person per task was used for the purpose of the analysis. The 1-minute performance was chosen from the beginning of their performance. The total of spoken data collected from a participant is two minutes, as two picture stories were described in either task condition.

Once the data were transcribed, the transcriptions were coded for a range of measures of self-monitoring. Fourteen measures were employed to assess L2 self-monitoring behavior including disfluencies, repair types, temporal phases of repair, and accuracy. Pauses and temporal phases of repair were calculated using PRAAT software (Boersma & Weenink, 2008). Following Kormos and Dénes (2004), disfluency measures (filled pauses, repetitions, hesitations) were divided by the total speech time (60), multiplied by 60 (Table 3). Silent pauses were only included as a measurement of the interruption length (see Table 2). Self-repair types and temporal phases of repair are two aspects of self-repair measured in the present study. Self-repairs included the main repair types classified by Levelt (1983) and adopted by Kormos (1999), see Table 1 where examples were taken from the current data. The figures reported for each measure of repair types are frequencies of the measures per 60 seconds.

Repair temporal phases entail three phases of repair (error-to-cut-off, cut-off-to-repair, and repair) (Figure 1). Coding these phases of repair is time consuming, thus for practical reasons, we only include the first two temporal phases: The first phase (error-to-cut-off) and the second phase (cut-off-to-repair) as presented in Table 2.

We employed two measures of accuracy, self-correction ratio and percentage of error-free clauses, to show two different aspects of accuracy during self-monitoring. While self-correction shows accuracy-as-a-process as it directly reflects the monitoring process, the percentage of error-free clauses indicates accuracy-as-a-product. The ratio of error-correction is calculated by dividing the number of repaired errors by the total number of errors in the speech sample (Kormos, 2006; Oomen & Postma, 2001). The percentage of error-free clauses, a global measure of accuracy is calculated by the number of error-free clauses divided by the total number of clauses in the speech sample multiplied by 100. Some researchers (Foster & Wigglesworth, 2016) have criticized this measure as it fails to show the gravity of the error, arguing an alternative global measure (e.g., Weighted Clause Ratio) that considers errors' weighting is needed. Despite such criticism, percentage of error-free clauses is still a reliable measure of accuracy widely used in TBLT studies (Skehan, 2009; Tavakoli, 2019).

The first author coded all of the repair measures in the data, while the second author second rated 10% of randomly selected speech samples to check the coding reliability. In the case of disagreement between the two raters, a third rater was consulted to ensure the reliability of the coding process. The two raters agreed on 83.43% of repair type classification. This percentage is high, comparable to the 73% of Levelt (1983) and the 75% of Declerck and Kormos (2012). Concerning accuracy measures, 10% of the data were second rated by a native speaker of English with linguistic expertise. The

Table 2. Temporal phases of repair

Repair temporal phases	Definition	Measurement
Phase 1 (Error-to-cut-off)	It involves erroneous or inappropriate word(s) (Levelt, 1983).	It is calculated in seconds from the onset of erroneous word(s) to the moment of interruption.
Phase 2 (Cut-off-to-repair)	It entails producing silent and/or filled pauses before executing the repair (Levelt, 1983).	It is calculated in seconds from the moment where speech stops to the moment of resumption.

https://doi.org/10.1017/S0272263122000146 Published online by Cambridge University Press

Table 3. Disfluency features

Dimension	Measure	Definition
Disfluency	Hesitations	The total number of hesitation (i.e., repeating part(s) of a word) was divided by the total time of speech in seconds and multiplied by 60.
	Repetitions	The total number of repetitions (i.e., words, phrases) was divided by the total time of speech in seconds and multiplied by 60.
	Filled pauses	The total number of filled pauses (uh, umm, err) divided by the total time of speech in seconds and multiplied by 60.

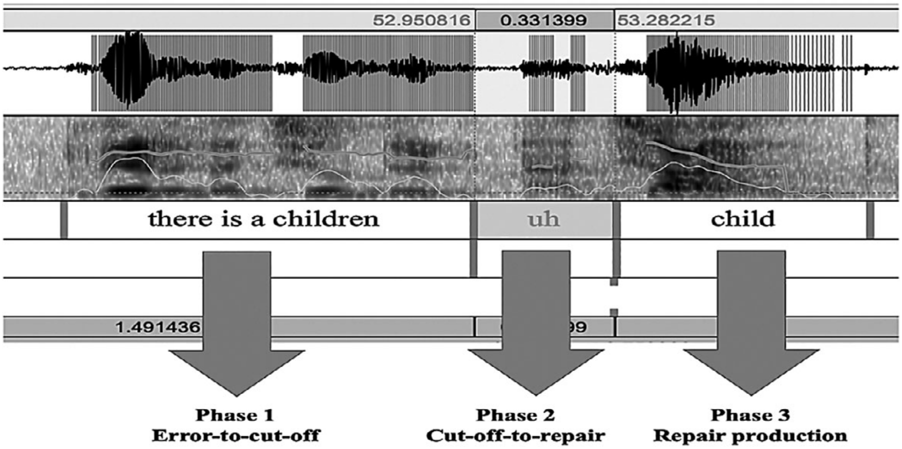

Figure 1. Calculating repair temporal phases.

convergence between the two raters was 83%. The high interrater reliability achieved confirmed the consistency of coding procedure. Before coding the data, we segmented the transcripts to AS units, using Foster et al.'s (2000) guidelines.

Results

Our data analysis includes factor analysis that is important for selecting representative measures to be submitted to the MANOVA, while the two-way ANOVAs contained all measures. This section presents the purpose and details of each analysis. Descriptive statistics for PL and TC are provided in Tables 4 and 5.

Data reduction

Given that this is one of the few studies in this area, we used a wide range of measures to examine self-monitoring. To control for any potential overlap between these measures,

Table 4. Descriptive statistics of PL

PL	*N*	Median	Mean	*SD*	Minimum	Maximum	95% *CI*
Elicited imitation test	66	30.63	32.44	13.05	12	66	1.26–1.49

Table 5. Descriptive statistics of TC

	Median		Mean		*SD*		Minimum		Maximum		
Measures	Single	Dual	Single	Dual	Single	Dual	Single	Dual	Single	Dual	95% *CI*
Filled pauses	17.00	18.00	17.11	17.97	8.31	6.54	2.00	2.00	45.00	34.00	17.09–20.71
Hesitation	1.00	1.00	1.41	1.48	1.50	1.42	.00	.00	5.00	6.00	1.18–1.91
Repetition	1.00	2.54	1.86	2.55	2.22	1.94	.00	.00	10.00	10.00	2.0–3.08
A-repair	.000	.000	.47	.67	.90	.87	.00	.00	4.00	3.00	.16–.47
D-repair	.00	.00	.35	.55	.59	.64	.00	.00	2.00	3.00	.19–.44
E-repair	2.00	1.00	1.74	1.59	1.27	1.44	.00	.00	5.00	6.00	1.41–2.10
A-repair Phase (1)	.00	.00	.24	.48	.49	.68	.00	.00	2.42	2.45	.09–.31
A-repair Phase (2)	.00	.00	.11	.24	.24	.41	.00	.00	1.03	2.54	.04–.14
D-repair Phase (1)	.00	.00	.25	.50	.52	.88	.00	.00	2.52	5.72	.12–.50
D-repair Phase (2)	.00	.00	.22	.44	.47	.98	.00	.00	2.20	6.59	.10–.37
E-repair Phase (1)	1.02	1.06	1.94	1.42	4.27	1.53	.00	.00	4.27	6.23	1.02–3.03
E-repair Phase (2)	.76	.46	1.04	1.11	1.24	1.72	.00	.00	6.17	10.01	.84–1.51
Error-free clauses	36.93	40.46	37.26	41.72	19.19	19.34	.00	.00	75.00	93.74	30.20–39.25
Ratio of error-correction	.19	.19	.21	.25	.13	.13	.00	.00	.50	.67	.17–.22

all 14 measures were submitted to a principal component analysis with varimax rotation. The principal component analysis produced five factors, each containing a number of the measures. The factorability of the dataset was checked against Bartlett's test of sphericity ($\chi^2 = 603.23$, $p = .000$) and the Kaiser–Meyer–Olkin measure of sampling adequacy (.59). Based on eigenvalues above 1 (Pallant, 2016) and the visual inspection of the scree plot, five factors were identified in the data set, accounting for 69.43% of the variance in the self-monitoring measures. Following Pallant (2016), we reported the Pattern Matrix that displayed the highest loading items on each component. This helps in identifying and labelling the components. Factor 1 included A-repair and its temporal phases. Factor 2 included E-repair, its temporal phases and the ratio of error-correction. Factor 3 represented D-repair and its temporal phases. Factor 4 contained disfluency measures (e.g., hesitation and repetition). Factor 5 included measures of accuracy and filled pauses. The only negative loading of the factors, that is, filled pauses on Factor 5, suggests that a decrease in frequency of filled pauses is associated with an increase in accuracy. Given the small sample size of the study, we suggest the results of the factor analysis are considered cautiously. Table 6 shows all the loadings for the underlying factors.

Analysis of variance

To explore the overall impact of proficiency level (PL) and task condition (TC) on L2 self-monitoring, the five factors obtained from the factor analysis were entered into the MANOVA as representatives of L2 self-monitoring: A-repair (first phase); E-repair frequency; D-repair (second phase); hesitations; and error-free clauses (the highest loading items on each component). The use of MANOVA is controlled by a number of assumptions that need to be checked prior to proceeding with the analysis. All assumptions of normality, equality of variance, linearity, and multicollinearity were met in the current study. Partial eta squared were calculated to assess the magnitude of the effects obtained in the analysis. Cohen's (1988) guidelines suggest partial eta squared values of .2 should be regarded as small, .5 as medium, and .8 as large. More recently, however, Norouzian and Plonsky (2018) argue that in multiway designs, partial eta squared figures should be interpreted more carefully as "ηp2 values are

Table 6. Factor analysis of L2 self-monitoring measures

	Factors					
Measures	1	2	3	4	5	Dimensions
A-repair (First phase)	.876					A-repair
A-repair Frequency	.864					
A-repair (Second phase)	.843					
E-repair Frequency		.882				E-repair
E-repair (Second phase)		.800				
E-repair (First phase)		.644				
Ratio of error-correction	.370	.587			.420	
D-repair (Second phase)			.823			D-repair
D-repair (First phase)			.801			
D-repair Frequency			.791			
Hesitations				.842		Disfluencies
Repetitions				.837		
Filled pauses				.461	–.348	
Error-free clauses					.950	Accuracy

https://doi.org/10.1017/S0272263122000146 Published online by Cambridge University Press

Table 7. Results of multivariate analysis of variance

Effect	Wilks' Lambda Value	*F*	Sig.	Partial Eta Squared
Proficiency	.025	1.86	.000*	.521
Task Condition	.912	1.428	.22	–
Proficiency × Task Condition	.647	.840	.72	–

*$p < .025$.

invariably larger—often much larger—than their η2 counterparts" (Norouzian & Plonsky, 2018, p. 261). Following these guidelines, we suggest our results are interpreted cautiously.

The results of the MANOVA indicate that PL had a significant effect on L2 self-monitoring, F (220, 373) = 1.86, $p = .000$; Wilks Lambda = .025; partial eta squared (ηp^2) = .521 (Table 7). The results suggest that L2 self-monitoring was significantly influenced by differences in proficiency levels (based on the EIT scores). The analysis does not show any significant effect of TC on L2 self-monitoring. This means that there may not be great differences in L2 self-monitoring behavior in the two task conditions. Likewise, there was no interaction effect between the two variables which means that L2 performances were not mediated by TC. To understand how individual aspects of L2 self-monitoring were influenced by proficiency level and task demands, the 14 measures were submitted to a series of two-way ANOVAs (Table 8). The purpose of the analyses was to have a fine-grained examination of the effects of PL and TC on different aspects of L2 self-monitoring, and potential interaction effects. A Bonferroni correction was considered to correct the alpha level (0.05/14) for the ANOVAs ($alpha < 0.004$).

Table 8. Two-way between-group analyses of variance

	Proficiency Level			Task Condition			PL*TC		
Measures	*F*	*Sig.*	η^2	*F*	*Sig.*	η^2	*F*	*Sig.*	η^2
Frequency of filled pauses	4.95	.000*	.736	.049	.826	.001	2.09	.046	.177
Frequency of hesitation	2.54	.000*	.589	.519	.025	.062	.785	.617	.074
Frequency of repetition	3.22	.000*	.645	22.82	.000*	.226	2.20	.036	.184
Frequency of A-repair	1.63	.030	.479	.848	.36	.011	1.05	.409	.097
Frequency of D-repair	1.82	.011	.506	3.10	.082	.038	1.32	.245	.119
Frequency of E-repair	1.33	.138	.428	.475	.493	.006	1.06	.399	.098
Duration of A-repair Phase1	1.38	.106	.438	1.97	.165	.025	.743	.653	.071
Duration of A-repair Phase 2	1.09	.360	.381	1.28	.261	.016	.528	.832	.051
Duration of D-repair Phase 1	.99	.509	.358	5.19	.026	.062	1.87	.076	.161
Duration of D-repair Phase 2	1.61	.033	.476	1.09	.300	.014	.422	.904	.042
Duration of E-repair Phase 1	.298	1.00	.144	.808	.371	.010	.483	.865	.047
Duration of E-repair Phase 2	.927	.601	.343	.007	.934	.000	.927	.499	.087
Error-free clauses	3.13	.000*	.639	.004	.949	.000	1.20	.309	.110
Ratio of error-correction	2.80	.000*	.612	.286	.594	.004	.267	.012	.215

*$p = .004$.
PL df (44, 78), *TC df* (1, 78), *PL*TC df* (8, 78)

However, we would like to remind our readers that given the strict nature of a Bonferroni correction, many of the potentially significant differences in the ANOVAs might not reach the corrected alpha level.

Effects of proficiency on L2 self-monitoring

The results of the two-way ANOVAs show that while disfluency and accuracy were significantly affected by proficiency ($p < 0.004$), self-repair measures were not. For disfluency measures, the lower proficiency learners made significantly more filled pauses, $F(44, 78) = 4.95$, $p < .000$, $\eta^2 = .736$, more repetitions, $F(44, 78) = 3.22$, $p < .000$, $\eta^2 = .645$, and more hesitations, $F(44, 78) = 2.54$, $p < .000$, $\eta^2 = .589$, compared to the higher proficiency learners. The effect sizes of these comparisons, all above .5, imply that PL accounts for a considerable amount of the variance in measures of filled pauses, repetitions, and hesitations.

In the case of accuracy measures, PL had significant impact on error-free-clauses, $F(44, 78) = 2.80$, $p < .000$, $\eta^2 = .612$, and ratio of error-correction, $F(44, 78) = 3.22$, $p < .000$, $\eta^2 = .645$, with moderate effect sizes. These results suggest that the higher proficiency learners made more attempts at correcting their utterances (i.e., a higher ratio of error correction), and they produced more accurate clauses (i.e., more error-free-clauses).

In terms of repair types, as demonstrated in Table 8, PL slightly affects A-repair and D-repair with the higher proficiency learners making slightly more A-repair and D-repair than the lower proficiency learners, but these differences come short of reaching the Bonferroni adjusted *p* level. The analysis does not show significant main effects of PL on temporal phases of repair that suggest the proficiency development may not affect the duration of producing repair. To conclude, PL seems to have significant effects on disfluency and accuracy measures.

Effects of task condition on L2 self-monitoring

To provide an overall picture of the participants' behavior during performance under the dual-task condition, details of performance on the secondary task is illustrated in Table 9.

Table 9 summarizes the accuracy rate of keyboard responses and the reaction times during the secondary task, that is, the time that participants spent when responding to stimuli in the secondary task. The data demonstrate that the average of accuracy of keyboard responses was (72%) which likely means that the majority of participants were engaged with the secondary task while they were describing the oral narrative picture prompts. Reaction times data show that the average time of responding to stimuli was about 1.81 seconds (1895.1 ms) out of 5 seconds, which suggests that participants responded to stimuli in a relatively speedy manner.

Table 9. Average performance on the secondary task

Secondary task data	Mean	Min.	Max.	Std. Dev	Std. Err
Accuracy of responses Max = 1	0.72	.45	1.00	.136	.017
Reaction times Max = 5,000 ms	1895.1	634.5	2888.5	461.6	58.15

https://doi.org/10.1017/S0272263122000146 Published online by Cambridge University Press

Although Table 8 does not show a significant effect of TC for most self-monitoring measures, the results indicate that repetition was affected by TC, *F(8, 78)* = 2.20, $p <$ .000, η^2 = .226. Descriptive statistics in Table 5 indicates that more repetitions were made in the dual-task condition ($M = 2.55$, $SD = 1.94$) compared to performance in the single-task condition ($M = 1.86$, $SD = 2.22$). The increase in repetitions in the dual-task condition suggests that performance in this condition is less fluent than that in the single-task condition.

Interaction effects of proficiency and task condition on self-monitoring

The data in Table 8 shows no interaction effect between PL and TC on any of the fourteen measures according to the adjusted alpha level. This suggests that TC did not interact with PL in their impact on the oral performance of L2 learners. These results will be discussed in the next section.

Discussion

To examine the effects of PL and TC on L2 self-monitoring, we subjected our data to a range of different statistical analyses. Firstly, we used factor analysis to control for any overlap among the measures to be submitted to MANOVA. The results of the analyses suggested that PL had a statistically meaningful impact on L2 self-monitoring in terms of disfluency and accuracy of oral performance. The results of the analysis examining the effects of TC on performance suggested TC only influenced repetitions. In what follows, we discuss the findings of the study in relation to our research questions and in the light of the literature discussed previously.

The effects of proficiency on self-monitoring behavior

The results of our study indicate that higher proficiency speakers produced significantly fewer filled pauses, repetitions, and hesitations than the lower proficiency learners. In general, this is in line with previous research in this area (e.g., De Jong et al., 2015; Skehan, 2009; Tavakoli, 2019; Tavakoli et al., 2020) suggesting that filled pauses, hesitations, and repetitions are characteristics of performance at lower proficiency levels; these features are often perceived as opportunities for L2 learners to buy time to deal with the demands of L2 processing, particularly at conceptualization and formulation stages of speech production (Skehan, 2009; Tavakoli & Wright, 2020).

The results of our study also indicate that the higher proficiency learners, compared to lower proficiency ones, produced considerably more error-free clauses. L2 learners are typically expected to improve their accuracy when they develop their proficiency, and as such this finding seems rather anticipated. The finding is in line with Nakatsuhara et al. (2019) who reported that the development of proficiency was clearly observed in an increase in percentage of error-free clauses, whereas development in other aspects of proficiency (e.g., syntactic complexity) was not always consistently observed between different levels. Our analysis also suggested that the ratio of error-correction was higher for the higher proficiency learners. This is an interesting finding that implies activation of monitoring processes is more likely to occur at higher levels of proficiency. The finding is in line with Declerck and Kormos's (2012) study where the ratio of error-correction increased in the advanced rather than the intermediate learner

https://doi.org/10.1017/S0272263122000146 Published online by Cambridge University Press

group. The authors argued that monitoring processes were functioning more efficiently in the advanced group (ibid.). Further research is certainly needed in this regard.

We have referred to the two measures of ratio of error-correction and error-free clauses as accuracy-process and accuracy-product measures of L2 self-monitoring respectively. The results, in effect, suggest that the lower proficiency learners were less successful at both accuracy process and accuracy product measures. Our finding implies that the lower proficiency learners may not have been able to identify their errors and may not have been able to correct the errors. Our study design, however, does not allow us to examine whether the former caused the latter. Neither does our study indicate whether the accuracy process and product measures were affected by linguistic knowledge restrictions or processing capacity limitations. Further research is needed to examine these hypotheses. The combined results of disfluency and accuracy measures in our study are in line with research investigating performance across proficiency levels (Nakatsuhara et al., 2019; Tavakoli et al., 2016) suggesting accuracy and fluency are closely linked to L2 learners' proficiency. However, these results cannot confirm Levelt's (1983) assumption that disfluencies (i.e., covert repair) are made as corrective actions to anticipated errors. Our results show that the higher frequency of dysfluencies in our lower proficiency learners was not related to anticipating corrective actions. These learners produced a high number of disfluencies, but they were not successful in anticipating or identifying many errors. This finding highlights the potential differences between L1 and L2 monitoring processes and draws our attention to the need for developing an appropriate L2 model of speech production. It is also worth noting that disfluencies in L2 speech might not necessarily reflect self-monitoring; they may represent other processes or personal traits (see De Jong et al., 2015; Derwing et al., 2009; Dörnyei & Kormos, 1998; Skehan & Foster, 2005). Duran-Karaoz and Tavakoli (2020), for example, provided evidence that L2 disfluencies, to a great extent, reflect L1 speaking style. Therefore, future studies are needed in which L1 styles are controlled for when investigating L2 monitoring processes. Retrospective interviews are also needed to examine the purpose of producing disfluencies.

Regarding repair types, our results do not confirm the findings of previous research in which an increase was reported in A-repair in the speech of the higher proficiency learners (e.g., Gilabert, 2007; Kormos, 2000a, 2006; Van Hest, 1996). One possible interpretation of the discrepancy in these studies is that L2 learners in previous studies were at advanced levels of proficiency where speech production has become more automatic, particularly at the Formulator subprocesses where lexical retrieval and syntactic processing are needed. The availability of cognitive resources emerging from the automatization of the speech production processes has been claimed to account for the increase of A-repair among proficient learners (Gilabert, 2007; Kormos, 2000a, 2006; Van Hest, 1996). In the current study, L2 learners belonged to elementary and intermediate levels of proficiency where some speech processes may not have been automatized yet.

The effects of task condition on self-monitoring behavior

Our analyses indicate that TC did not have a statistically significant effect on most L2 self-monitoring measures. The only measure influenced by TC was repetitions where L2 learners produced significantly more repetitions in the dual-task condition compared to single-task condition. This finding is important as previous studies employing dual-task condition did not report any significant influence of TC on disfluencies either in L1 (Oomen & Postma, 2002) or L2 (Declerck & Kormos, 2012) contexts. This may suggest that the dual-task condition as operationalized in the current study has likely increased the task demand with an impact on the number of repetitions. It is possible to explain the higher number of repetitions in the dual-task condition in the light of the need the learners may have felt to buy time during a cognitively demanding task. This is in line with previous research that considers repetitions as a strategy to cope with the increased demand of task condition (see De Jong et al., 2015; Derwing et al., 2009; Dörnyei & Kormos, 1998; Skehan & Foster, 2005).

Our nonsignificant results from the effects of TC on other measures is different from Declerck and Kormos's (2012) findings in which they observed significant effects on the ratio of the error-correction and lexical errors. We interpret the difference between the two studies in the light of task designs used in the two studies. As discussed earlier, Declerck and Kormos's (2012) task involved a tightly controlled network description that required the participants to produce a set of utterances requiring a good degree of precision involving colours and directions. In this task, it is highly important to be correct about the choice of lexical items (e.g., colors), directions and movements (e.g., verb structures). Our task, in contrast, allowed the participants to express their meaning in any lexical and syntactic units of their own choice as long as the main events of the story were narrated. We postulate that the controlled nature of the network description task in Declerck and Kormos's (2012) may have encouraged a focus on accuracy, with an effect on the learners' L2- self-monitoring in terms of the accuracy measures.

There are two possible explanations for the lack of influence of dual-task condition on other L2 monitoring measures in this study. First, it has been argued that even in the single-task condition L2 speech processes require substantial cognitive resources, and therefore, performing in the dual-task condition might not lead to noticeable effects on speech processes (ibid.). That is to say, in the case of L2 speech production where cognitive resources are already consumed, the increased demand of task condition would have little impact on L2 self-monitoring. Second, it is plausible to argue that with the increased cognitive demand of the task condition, the monitor becomes robust, so that no noticeable differences are observed between the two task conditions. That is to say, the monitor was able to correct the same number of errors, make the same amount of repair, maintaining the rate of accuracy and fluency even with the increased demand of the task condition. This assumption is in line with the data of Levelt et al. (1999), which reported that the monitor becomes intense in the more demanding task condition. In other words, the auditory loop of the monitor may operate actively with the increased cognitive demand of task condition so that it detects the same number of errors and maintains accuracy and fluency (see "Self-Monitoring" section). However, this is not conclusive and further research is still needed in this regard.

Conclusion

In response to the calls for L2 researchers to test L1 self-monitoring theories in the L2 context (e.g., Kormos, 2000a; Van Hest, 1996), the current study set out to examine the effects of PL and TC on L2 speakers' performance in single and dual-task conditions. The study was also rightly placed to inspect the principles of PLT in L2 speech production. One of the main premises of the PLT is that self-monitoring draws largely on cognitive resources and how attentional resources are consumed during speech production (Levelt, 1983, 1989, 1992, 1999; Levelt et al., 1999). The findings of the current study indicate that with proficiency development, considerably fewer filled pauses, repetitions, and hesitations are observed in L2 learners' performance. Similarly,

https://doi.org/10.1017/S0272263122000146 Published online by Cambridge University Press

a greater ratio of errors was corrected, and a higher percentage of error-free clauses was produced by higher proficiency learners implying a more active and effective monitoring process is at play. These findings are important for the development of L2 speech production models, as it highlights which feature of self-monitoring is more relevant to L2 speaking processes.

Another principal premise of the PLT is that self-monitoring is sensitive to contextual effects (Levelt, 1983). To examine the impact of such contextual factors in terms of resource limitation on L2 self-monitoring, the dual-task condition was used in the current study. The results showed that making the L2 speaking process more demanding by adding a secondary task had a considerable impact on repetition of L2 utterances. The increased demand of TC has likely led to more repetitions suggesting L2 learners may use repetition as a strategy to cope with task demand or an opportunity to buy time to process their speech before articulation.

Finally, further research will need to address the limitations of the current study. First, we suggest that future research should include more heterogeneous samples (male and female), and a wider range of proficiency levels to examine monitoring in relation to different stages of development. This would allow researchers to see if certain features of monitoring progress with the development of proficiency. Researchers should also investigate learners' L1 performance (as well as their L2 performance) to determine which monitoring features are triggered by L2 processing and which are related to personal styles. While this study focused solely on self-repair types, their temporal phases and disfluencies, future studies should examine different types of errors (lexical, grammatical, phonological) in relation to self-repair in different proficiency levels. This would allow us to understand the sensitivity of the monitor toward different types of errors in different levels of proficiency. Last but not the least, future research should investigate the distribution of the disfluencies relative to the content of speech and to the timing and execution of the secondary task. Such careful examinations would provide important information about self-monitoring and the nature of disfluencies.

Supplementary Materials. To view supplementary material for this article, please visit http://doi.org/10.1017/S0272263122000146.

Acknowledgments. We would like to thank Taif University students who participated in the study for their time and commitment. We also thank the reviewers for their thoughtful comments and efforts toward improving our manuscript.

References

Ahmadian, M. J., & Tavakoli, M. (2014). The effects of simultaneous use of careful online planning and task repetition on accuracy, complexity, and fluency in EFL learners' oral production. *Language Teaching Research*, *15*, 35–59.

Albarqi, G. (2018). What can dual task paradigm tell us about second language self-monitoring behaviour? *Language Studies Working Papers*, *9*, 3–13.

Baten, K., & Cornillie, F. (2019). Elicited imitation as a window into developmental stages. *Journal of the European Second Language Association*, *3*, 23–34.

Blackmer, E. R., & Mitton, J. (1991). Theories of monitoring and the timing of repairs in spontaneous speech. *Cognition*, *39*, 173–194.

Boersma, P., & Weenink, D. (2008). Praat, a system for doing phonetics by computer. Computer software. http://www.fon.hum.uva.nl/praat/

Broos, W. P., Duyck, W., & Hartsuiker, R. J. (2018). Monitoring speech production and comprehension: Where is the second-language delay? *Quarterly Journal of Experimental Psychology*, 1–19.

Cobb, T. (2017). Web Vocabprofile. An adaptation of Heatley, Nation & Coxhead's (2002) Range. http://www.lextutor.ca/vp/

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Lawrence Erlbaum Associates.

De Jong, N. H. (2018). Fluency in second language testing: Insights from different disciplines. *Language Assessment Quarterly*, *15*, 237–254.

De Jong, N. H., Groenhout, R., Schoonen, R., & Hulstijn, J. H. (2015). Second language fluency: Speaking style or proficiency? Correcting measures of second language fluency for first language behavior. *Applied Psycholinguistics*, *36*, 223–243.

De Jong, N., & Vercellotti, M. L. (2016). Similar prompts may not be similar in the performance they elicit: Examining fluency, complexity, accuracy, and lexis in narratives from five picture prompts. *Language Teaching Research*, *20*, 387–404.

Declerck, M., & Kormos, J. (2012). The effect of dual task demands and proficiency on second language speech production. *Bilingualism: Language and Cognition*, *15*, 782–796.

DeKeyser, R. (2001). Automaticity and automatization. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 125–151). Cambridge University Press.

DeKeyser, R. M. (2013). Age effects in second language learning: Stepping stones toward better understanding. *Language learning*, *63*, 52–67.

Derwing, T. M., Munro, M. J., Thomson, R. I., & Rossiter, M. J. (2009). The relationship between L1 fluency and L2 fluency development. *Studies in Second Language Acquisition*, *31*, 533–557.

Dörnyei, Z., & Kormos, J. (1998). Problem-solving mechanisms in L2 communication: A psycholinguistic perspective. *Studies in Second Language Acquisition*, *20*, 349–385.

Duncan, J. (1980). The locus of interference in the perception of simultaneous stimuli. *Psychological Review*, *87*, 272.

Duran-Karaoz, Z., & Tavakoli, P. (2020). Predicting L2 fluency from L1 fluency behaviour: The case of L1 Turkish and L2 English speakers. *Studies in Second Language Acquisition*, 1–25.

Ellis, R. (2005). Measuring implicit and explicit knowledge of a second language: A psychometric study. *Studies in Second Language Acquisition*, *27*, 141–172.

Erlam, R. (2006). Elicited imitation as a measure of L2 implicit knowledge: An empirical validation study. *Applied Linguistics*, *27*, 464–491.

Faez, F., & Tavakoli, P. (2019). *Task Based Language Teaching. English Language Teaching Development Series*. TESOL International Association.

Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics*, *21*, 354–375.

Foster, P., & Wigglesworth, G. (2016). Capturing accuracy in second language performance: The case for a weighted clause ratio. *Annual Review of Applied Linguistics*, *36*, 98–116.

Gaillard, S., & Tremblay, A. (2016). Linguistic proficiency assessment in second language acquisition research: The elicited imitation task. *Language Learning*, *66*, 419–447.

Gilabert, R. (2007). Effects of manipulating task complexity on self-repairs during L2 oral production. *International Review of Applied Linguistics in Language Teaching*, *45*, 215–240.

Hartsuiker, R. (2014). Monitoring and control of the production system. In M. Goldrick, V. S. Ferreira, & M. Miozzo (Eds.), *The Oxford handbook of language production* (pp. 417–436). Oxford University Press.

Hulstijn, J. H., Van Gelderen, A., & Schoonen, R. (2009). Automatization in second language acquisition: What does the coefficient of variation tell us? *Applied Psycholinguistics*, *30*, 555–582.

Kormos, J. (1999). Monitoring and self-repair in L2. *Language Learning*, *49*, 303–342.

Kormos, J. (2000a). The role of attention in monitoring second language speech production. *Language Learning*, *50*, 343–384.

Kormos, J. (2000b). The timing of self-repairs in second language speech production. *Studies in Second Language Acquisition*, *22*, 145–167.

Kormos, J., & Dénes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System*, *32*, 145–164.

Kormos, J. (2006). *Speech production and second language acquisition*. Routledge.

Kostromitina, M., & Plonsky, L. (2021). Elicited imitation tasks as a measure of l2 proficiency: A meta-analysis. *Studies in Second Language Acquisition*, 1–26.

Lambert, C., Kormos, J., & Minn, D. (2017). Task repetition and second language speech processing. *Studies in Second Language Acquisition*, *39*, 167–196.

https://doi.org/10.1017/S0272263122000146 Published online by Cambridge University Press

Levelt, W. (1983). Monitoring and self-repair in speech. *Cognition*, *14*, 41–104.

Levelt, W. J. (1989). *Speaking: From intention to articulation* (Vol. *1*). MIT Press.

Levelt, W. J. (1992). The perceptual loop theory not disconfirmed: A reply to MacKay. *Consciousness and Cognition*, *1*, 226–230.

Levelt, W. J. (1999). Language production: A blueprint of the speaker. *Neurocognition of Language*, 83–122.

Levelt, W. J., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral Brain Sciences*, *22*, 1–38.

Mackay, D. C. (1992). Awareness and error detection: New theories and research paradigms. *Consciousness and Cognition*, *1*, 199–225.

Nakatsuhara, F., Tavakoli, P. & Awwad, A. (2019). Towards a model of multi-dimensional performance of C1 level speakers assessed in the Aptis speaking test. *Technical Report.* British Council. ISSN 2398-7979.

Norouzian, R., & Plonsky, L. (2018). Eta- and partial eta-squared in L2 research: A cautionary review and guide to more appropriate usage. *Second Language Research*, *34*, 257–271.

Oomen, C., & Postma, A. (2001). Effects of time pressure on mechanisms of speech production and self-monitoring. *Journal of Psycholinguistic Research*, *30*, 163–184.

Oomen, C., & Postma, A. (2002). Limitations in processing resources and speech monitoring. *Language Cognitive Processes*, *17*, 163–184.

Ortega, L. (2009). *Understanding second language acquisition*. Hodder.

Pallant, J. (2016). *SPSS survival manual.* McGraw-Hill Education.

Pellicer-Sánchez, A. (2015). Developing automaticity and speed of lexical access: The effects of incidental and explicit teaching approaches. *Journal of Spanish Language Teaching*, *2*, 126–139.

Postma, A. (2000). Detection of errors during speech production: A review of speech monitoring models. *Cognition*, *77*, 97–132.

Postma, A., Kolk, H., & Povel, D.-J. (1990). On the relation among speech errors, disfluencies, and self-repairs. *Language and Speech*, *33*, 19–29.

Préfontaine, Y., & Kormos, J. (2015). The relationship between task difficulty and second language fluency in French: A mixed methods approach. *The Modern Language Journal*, *99*, 96–112.

Révész, A., Michel, M., & Gilabert, R. (2016). Measuring cognitive task demands using dual-task methodology, subjective self-ratings, and expert judgments: A validation study. *Studies in Second Language Acquisition*, *38*, 703–737.

Robinson, P. (1997). Generalizability and automaticity of second language learning under implicit, incidental, enhanced, and instructed conditions. *Studies in Second Language Acquisition*, *19*, 223–247.

Robinson, P. (2001). Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. *Applied Linguistics*, *22*, 27–57.

Schneider, W., Eschman, A., & Zuccolotto, A. (2002). E-Prime (Version 3.0). Computer Software and Manual. Psychology Software Tools Inc.

Segalowitz, N. (2003). Automaticity and second languages. In C. J. Doughty (Ed.), *The handbook of second language acquisition* (pp. 383–408). Blackwell.

Seyfeddinipur, M., Kita, S., & Indefrey, P. (2008). How speakers interrupt themselves in managing problems in speaking: Evidence from self-repairs. *Cognition*, *108*, 837–842.

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics*, *30*, 510–532.

Skehan, P., & Foster, P. (2005). Strategic and on-line planning: The influence of surprise information and task time on second language performance. In R. Ellis (Ed.), *Planning and task performance in a second language* (Vol. *11*, pp. 193–216). John Benjamins.

Suzuki, S. (2021). Multidimensionality of second language oral fluency: The interface between cognitive, utterance, and perceived fluency. Doctoral dissertation. Lancaster University, UK.

Suzuki, Y., & DeKeyser, R. (2015). Comparing elicited imitation and word monitoring as measures of implicit knowledge. *Language Learning*, *65*, 860–895.

Tavakoli, P. (2011). Pausing patterns: Differences between L2 learners and native speakers. *ELT Journal*, *65*, 71–79.

Tavakoli, P. (2019). Automaticity, fluency and second language task performance. In Z. Wen & M. J. Ahmadian. (Eds.), *Researching L2 task performance and pedagogy: In honour of Peter Skehan*. John Benjamins.

https://doi.org/10.1017/S0272263122000146 Published online by Cambridge University Press

Tavakoli, P., Campbell, C., & McCormack, J. (2016) Development of speech fluency over a short period of time: effects of pedagogic intervention. *TESOL Quarterly*, *50*, 447–471.

Tavakoli, P., & Foster, P. (2011). Task design and second language performance: The effect of narrative type on learner output. *Language Learning*, *61*, 37–72.

Tavakoli, P., Nakatsuhara, F., & Hunter, A.-M. (2020). Aspects of fluency across assessed levels of speaking proficiency. *Modern Language Journal*, *104*, 169–191.

Tavakoli, P., & Skehan, P. (2005). Strategic planning, task structure, and performance testing. In R. Ellis (Ed.), *Planning and task performance in a second language* (Vol. *11*, pp. 239–273). John Benjamins.

Tavakoli, P., & Wright, C. (2020). *Second language speech fluency: From research to practice*. Cambridge University Press.

Tremblay, A. (2011). Proficiency assessment standards in second language acquisition research: "Clozing" the gap. *Studies in Second Language Acquisition*, *33*, 339–372.

Van Hest, E. (1996). *Self-repair in L1 and L2 production*. Tilburg University Press.

Wang, Z., & Skehan, P. (2014). Structure, lexis, and time perspective. In P. Skehan (Ed.), *Processing perspectives on task performance* (Vol. *5*, pp. 155–185). John Benjamins.

Wickens, C. D. (2007). Attention to the second language. *IRAL-International Review of Applied Linguistics in Language Teaching*, *45*, 177–191.

Wu, S. L., & Ortega, L. (2013). Measuring global oral proficiency in SLA research: A new elicited imitation test of L2 Chinese. *Foreign Language Annals*, *46*, 680–704.

Cite this article: Albarqi, G. and Tavakoli, P. (2023). The effects of proficiency level and dual-task condition on L2 self-monitoring behavior. *Studies in Second Language Acquisition,* **45,** 212–233. https://doi.org/10.1017/S0272263122000146