ISSN: 2754-6659

Journal of Artificial Intelligence & Cloud Computing

SCIENTIFIC Research and Community

Review Article

Open Access

Natural Language Processing for Script Writing Assistance

Kailash Alle

Sr. Software Engineer, Comscore, Inc

ABSTRACT

Despite the importance of monitoring in self-regulated learning (SRL) and second language (L2) speech production outside of testing situations, there's limited understanding of how this metacognitive skill operates during tests and its impact on learner performance. Given the interconnected relationship between L2 testing and learning, it's crucial to explore how monitoring works during speaking tests, particularly computer-delivered integrated speaking tests. These tests are increasingly used in L2 classrooms and are seen as the future of L2 assessment.

This study aims to investigate how 95 Chinese learners of English as a foreign language (EFL) used monitoring during computer-delivered integrated speaking tests. Participants completed a self-reported questionnaire after performing three speaking tasks. Descriptive analysis followed by Hierarchical Linear Modeling (HLM) showed that monitoring was frequently used by learners, but it didn't significantly affect their performance.

These findings have implications for SRL in pedagogy: educators should help learners develop self-regulation skills, including self-monitoring during tests, by simulating testing conditions in classroom activities. Additionally, the study provides insights into L2 testing by focusing on self-monitoring as a component of strategic competence, a crucial aspect of L2 assessment.

***Corresponding author**

Kailash Alle, Sr. Software Engineer, Comscore, Inc, USA.

Received: September 05, 2022; **Accepted:** September 12, 2022; **Published:** September 20, 2022

Keywords: Script Writing, AI, Natural Language, Speech Production

Introduction

"Language tests and the English language have significant impacts in today's global context, influencing policies and practices in English language education" (Shohamy, 2007, p. 1). Over 15 years ago, Elana Shohamy, a prominent researcher in language testing, highlighted the strong influence of L2 testing on language teaching. Shohamy's findings remain relevant, supported by studies showing how L2 testing evaluates teaching practices in language education (e.g., Huang et al., 2016). Building on this, this article examines self-monitoring, a key skill in self-regulated learning (SRL), within L2 testing contexts, specifically computer-delivered integrated speaking tests. The study aims to understand how well self-regulatory skills function during testing. The findings are expected to offer teaching insights: educators should replicate testing conditions in classrooms to help learners regulate themselves effectively during high-stakes tests, crucial for academic or career goals in L2 learning (Shohamy, 2007; Bachman and Palmer, 2010). Additionally, the study aims to shed light on L2 testing through self-monitoring, a component of strategic competence essential in L2 testing (Bachman and Palmer, 2010; Zhang et al., 2021a, 2022a,b). In the realm of L2 education, SRL is seen as a powerful tool for developing learners who can independently set learning goals and monitor their progress (Zimmerman, 1986; Schunk and Greene, 2018; Zhang and Zhang, 2019; Teng, 2022). While the systematic exploration of SRL's role began in the 1980s (Schunk and Greene, 2018), it is widely acknowledged that early studies (e.g., Zimmerman, 1989; Pintrich et al., 1993) paved the way for diverse theoretical perspectives: social-cognitive, cognitive/metacognitive (also known as information-processing), developmental, motivational, emotional, co-regulation, and socially shared regulation. These perspectives all emphasize the critical role of self-monitoring in SRL, where self-regulated learners actively manage their learning (e.g., Greene and Azevedo, 2007; Griffin et al., 2013; DiBenedetto, 2018; Schunk and Greene, 2018). Similarly, in L2 speaking research, self-monitoring significantly influences speech production covertly and overtly (Kormos, 2006, 2011; Bygate, 2011). Some scholars argue that SRL helps L2 learners overcome challenges in speaking, a daunting task for many (Uztosun, 2021; Gan et al., 2022). Despite the close connection between SRL and L2 speaking, there is limited literature contextualizing SRL specifically in L2 speaking, particularly from the perspective of self-monitoring (Uztosun, 2021; Gan et al., 2022), although extensive research has explored SRL in writing (e.g., Teng and Zhang, 2020), listening (e.g., Vandergrift and Goh, 2012), and reading (e.g., Cirino et al., 2017). Moreover, the focus on speaking's importance motivated this study: speaking directly facilitates communication in real-world settings (Luoma, 2004). Additionally, our emphasis on self-monitoring in computer-delivered L2 speaking tests reflects the increasing prevalence of this testing format due to technological advancements, accelerated by COVID-19's impact on traditional offline teaching, prompting virtual learning via computers (Zhang et al., 2021a). Given our

research focus and motivation, we adopted a multidisciplinary perspective integrating SRL, L2 speaking, computer-delivered testing, and integrated language skills to investigate how self-monitoring operates in computer-delivered integrated L2 speaking tests. To provide a comprehensive understanding of self-monitoring's roles across these disciplines, both theoretically and empirically, and to underscore its importance necessitating further research efforts like ours, we conducted an extensive literature review. This review focused on prior empirical studies exploring monitoring in L2 testing conditions, particularly computer-delivered integrated L2 testing, which informed the research questions for this study. Following the review, we presented our empirical study's findings, discussing their contributions to SRL in relation to L2 speaking and testing, and suggested avenues for future research. It's important to note that self-monitoring is the practical form of metacognitive monitoring, or simply monitoring, in both SRL and speaking (e.g., Levelt, 1983, 1989; Zimmerman, 2000; Kormos, 2006, 2011; Bygate, 2011; Schmitz and Perels, 2011; Nozari, 2020; Teng, 2022). Therefore, we used "self-monitoring," "metacognitive monitoring," and "monitoring" interchangeably throughout this article.

Monitoring in SRL

Since its beginning, self-regulated learning (SRL) has been studied from various angles (see Panadero, 2017, for an overview). Different SRL models, like Zimmerman's (1989) social-cognitive model, have been widely used in previous research (Panadero, 2017; DiBenedetto, 2018; Schunk and Greene, 2018). In our study, we focused on the role of self-monitoring in SRL, particularly from an information-processing perspective as described by Winne and Hadwin's (1998) SRL model. This choice was made to ensure consistency across the disciplines involved, as both L2 speaking and L2 speaking testing involve information processing (Hughes and Reed, 2017; Yahya, 2019). Winne and Hadwin's (1998) model is well-regarded and frequently cited in research where SRL is used in computer-supported learning (Panadero, 2017; DiBenedetto, 2018; Schunk and Greene, 2018).

In Winne and Hadwin's (1998) model, monitoring plays a central role. It involves two main activities: constructing and enacting metacognitive control, which are pivotal in SRL. The model divides SRL into four phases: task definition, goal setting and planning, tactics enactment, and metacognitive adaption. Each phase operates through interactions among five elements or task facets known as COPES (Panadero, 2017), illustrated in Figure 1. These facets include conditions, operations, products, evaluations, and standards.

Conditions refer to resources or constraints perceived by L2 learners that affect their learning, such as task demands and cognitive factors. Task conditions relate to how learners perceive task requirements, while cognitive conditions involve information retrieved from long-term memory, including prior knowledge and learning strategies. Motivational factors also influence cognitive conditions. Operations represent the actual information processing that occurs in each phase, encompassing tasks like searching, monitoring, assembling, rehearsing, and translating (SMART), which operate at both object and meta-levels. Products are the outcomes generated when SMART manipulates available information, serving as the outputs of each phase towards completing tasks effectively. Standards refer to the assumed qualities of products or the desired endpoints of ongoing phases, often set by teachers or derived from learners' expectations and prior experiences. Comparing products against standards through monitoring leads to cognitive evaluations, the fifth element in the SRL model.

When discrepancies arise between products and standards, learners employ metacognitive control to refine products, adjust conditions and standards, or simultaneously revise both. This process ensures that learners' goals are met through effective learning strategies (Greene and Azevedo, 2007; Schunk and Greene, 2018; Winne, 2018). These five aspects interact throughout the SRL phases. For instance, in the task definition phase, learners assess conditions like available resources and constraints, personal interests, and relevant knowledge to define their tasks. They may use standards or prior knowledge to monitor if their task definitions are appropriate. Task definitions then set the stage for the goal-setting and planning phase, where learners establish specific learning objectives based on their task definitions. These goals serve as benchmarks that self-regulating learners continuously monitor during learning processes and final outcomes.

During tactics enactment, learners implement learning strategies while comparing their progress and outcomes against set goals and plans. Any mismatches identified prompt metacognitive adjustments in the optional adaptation phase. Winne and Hadwin's (1998) model suggests that these four phases are loosely and recursively sequenced, allowing learners to flexibly move across phases rather than in a strict linear sequence. Learners can return to previous phases or skip ahead based on their monitoring results, optimizing their learning efficiency (Winne and Hadwin, 1998; Greene and Azevedo, 2007; Schunk and Greene, 2018; Winne, 2018).

In summary, metacognitive monitoring plays a crucial role in enabling learners to become self-regulated. It facilitates continuous comparisons between products and standards, supporting efficient and effective learning across the recursive phases of SRL.

Monitoring in l2 Speech Production

Speaking, especially in a second language (L2), is a complex skill (Levelt, 1989; Yahya, 2019; Newton and Nation, 2020). It involves receiving and processing information before producing systematic speech to convey meaning in real-time situations (Tarone, 2005, p. 485). In both first language (L1) and L2 speaking research, speech production is seen as a process of handling information (Luoma, 2004; Bygate, 2011; Kormos, 2011; Zhang et al., 2022a,b), where monitoring plays a crucial role (Levelt, 1989; Levelt et al., 1999; Nozari and Novick, 2017; Broos et al., 2019). Speakers act as information processors who "monitor what they are saying and how they are saying it" (Levelt, 1989, p. 458). Through monitoring, they identify errors or problems and make adjustments to ensure their speech reflects their intentions and meets linguistic standards (Levelt, 1989; Levelt et al., 1999; Nozari and Novick, 2017; Broos et al., 2019).

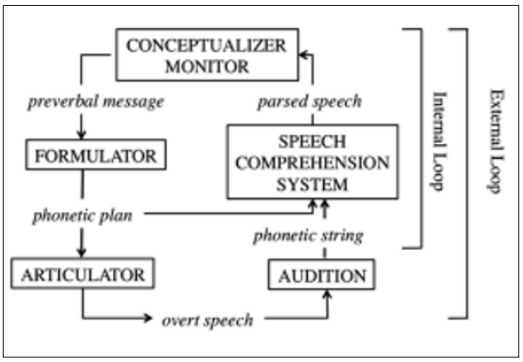
To replicate this monitoring mechanism in speech production, researchers have proposed four main approaches: comprehension-based monitoring, comprehension-perception based monitoring, production-based monitoring, and forward models of monitoring (see Nozari and Novick, 2017; Nozari, 2020, for details). These approaches encompass various models of self-monitoring in speech production, among which Levelt's (1983, 1989) perceptual loop model of self-monitoring (Figure 2) stands out as the most prominent and longstanding (Nooteboom and Quené, 2017; Gauvin and Hartsuiker, 2020; Nozari, 2020).

In Levelt's model, speech production involves several stages: conceptualization, formulation, articulation, and monitoring. Conceptualization begins with planning what to say based on long-term memory, discourse tracking, and understanding the listener's knowledge and expectations. Formulation then encodes this plan into linguistic terms, selecting words, arranging them grammatically, and preparing their pronunciation. Articulation is the physical process of speaking the formulated message. Throughout these stages, monitoring occurs both internally (covertly, before speech) and externally (auditorily, after speech), ensuring that the output matches the speaker's intentions and linguistic norms.

Levelt's model suggests two main self-monitoring loops during speech production: an internal loop, which checks the speech plan before it's spoken, and an external loop, which evaluates the spoken speech based on auditory feedback. Some scholars have proposed a third loop, the conceptual loop, which evaluates appropriateness based on the contextual demands (Hartsuiker and Kolk, 2001; Oomen and Postma, 2001; Kormos, 2006, 2011). These loops operate within a single comprehension system, enabling speakers to monitor their own speech and understand others' speech using the same cognitive processes (Hartsuiker, 2007).

Recent research has expanded on Levelt's model, proposing new models of verbal monitoring and applying them to L2 speech production (Gauvin and Hartsuiker, 2020; Broos et al., 2016, 2019). Despite differences between L1 and L2 speech production, major models of L2 speech production (e.g., De Bot, 1992; Poulisse and Bongaerts, 1994; Kormos, 2006, 2011) are largely based on Levelt's framework, emphasizing the role of self-monitoring throughout the speaking process (Bygate, 2011; Lambert et al., 2021).

In summary, self-monitoring in L2 speech production, similar to L1 speech production, is essential for ensuring speech accuracy and appropriateness. It involves constant evaluation of speech output against internal intentions and external linguistic norms, supporting effective communication in both native and second languages.



Monitoring in Computer-Delivered L2 Integrated Speaking Testing

Despite the widely recognized importance of monitoring in self-regulated learning (SRL) and second language (L2) speech production, its role in L2 speaking tests remains unclear. The complexity of speaking, coupled with the intricacies of language testing, contributes to this lack of clarity. As Hughes and Reed (2017) noted, language testing is a challenging field even for experts, making it difficult to fully grasp.

In language testing research, monitoring is often considered part of strategic competence or metacognitive strategies crucial to L2 testing (Bachman and Palmer, 2010; Phakiti, 2016; Purpura, 2016; Zhang et al., 2021a, 2022a,b). However, because strategic competence lacks a definitive definition, researchers have explored how monitoring functions in L2 testing and its impact on test performance. Studies under strategic competence have shown mixed results: some studies found monitoring to enhance test scores minimally, while others found no significant correlation (Pan and In'nami, 2015; Phakiti, 2016).

In investigations specifically into L2 speaking testing, research is limited due to the complexity of speaking and the challenges posed by L2 testing formats. For example, Fernandez (2018) examined monitoring among L2 learners taking the IELTS speaking test, finding no clear positive relationship between monitoring and test performance. In contrast, Huang (2016) found that monitoring directly influenced test performance among L2 learners taking standardized English proficiency tests in Taiwan.

Overall, research on monitoring in L2 speaking tests is still evolving and warrants further exploration (Seong, 2014; Zhang et al., 2021a, 2022a,b). With advancements in technology and the shift to computer-assisted language learning and testing, there is growing interest in strategic competence in computer-delivered L2 integrated speaking tests (Swain et al., 2009; Barkaoui et al., 2013; Zhang et al., 2021a). However, despite some studies addressing monitoring as part of strategic competence, specific focus on monitoring in L2 speaking tests remains limited.

The existing studies, though insightful, often have limitations such as small sample sizes or single-method approaches, which can affect the validity and generalizability of findings (Creswell and Creswell, 2018; Creswell and Guetterman, 2019). Moreover, while some studies have explored interactions between test-takers and test tasks, they have not specifically focused on monitoring, highlighting a research gap that needs to be addressed.

Focus of this Study

Based on our review of monitoring in self-regulated learning (SRL) and L2 speech production, along with the varied findings on its role in L2 speaking tests, especially those delivered via computer, and recognizing the significant influence of L2 testing on teaching, we have formulated the following research questions:
Research Question (RQ) 1: Do L2 learners use self-monitoring during computer-delivered integrated L2 speaking test tasks?
If the answer to RQ1 is yes, then RQ2 and RQ3 are as follows:
RQ2: How does self-monitoring affect L2 learners' performance in handling computer-delivered integrated L2 speaking test tasks?
RQ3: What are the impacts of interactions between self-monitoring and tasks on L2 learners' speaking performance in computer-delivered integrated L2 speaking test tasks?

Method

In our study, we used a method where we repeatedly measured how Chinese students learning English as a foreign language performed on TOEFL iBT integrated speaking tasks. We collected and analyzed data on their use of monitoring through questionnaires. We chose TOEFL iBT integrated speaking tasks because it's a pioneer in computer-delivered L2 testing and combines various language skills like reading, listening, and speaking. It's well-known for its reliability and validity. Also, TOEFL iBT aligns with China's Standards of English Language Ability (CSE), which helped us select participants with suitable language proficiency levels, as they are Chinese EFL learners.

Participants

We analyzed data from 95 Chinese university students studying English as a foreign language. They scored between 425 and 500 points on the College English Test—Band 4 (CET-4), which is a widely used English proficiency test in Chinese universities. According to China's Standards of English Language Ability (CSE), these students had an intermediate level of English proficiency suitable for TOEFL iBT integrated speaking tasks. The students were between 18 and 21 years old, with 38% male and 62% female participants. Two experienced Chinese EFL teachers rated the students' performance in TOEFL iBT integrated speaking tasks. These teachers were from the universities where we recruited the students. Participation in our study was voluntary, and we used convenience sampling to select participants. The sample size of students met the requirements for the statistical testing procedure (Hierarchical Linear Modeling - HLM) used in our study.

Measures

We assessed the Chinese EFL learners' use of self-monitoring during the TOEFL iBT integrated speaking tasks using the Chinese version of the Strategic Competence Inventory for Computer-Assisted Speaking Assessment (SCICASA) by Zhang et al. (2021a). The Chinese version of the inventory is in Appendix A, and the English version is in Appendix B for international readers. This inventory was chosen because it is in the learners' native language and measures strategic competence in computer-delivered speaking tests with high reliability (α=0.87). We only used the self-monitoring section, which has 7 items rated on a 6-point Likert scale: 0 (never), 1 (rarely), 2 (sometimes), 3 (often), 4 (usually), and 5 (always). An example item is, "I knew what to do if my intended plan did not work efficiently during the task." The SCICASA also collects participants' background information such as English proficiency, age, and gender.

For the TOEFL iBT integrated speaking tasks, we used a complete set of new tasks designed for intermediate-level learners, including three tasks: Task 1, Task 2, and Task 3 (ETS, 2022a). To maintain the test's validity and reliability, we used the original tasks without any modifications (Creswell and Creswell, 2018). Our study did not focus on the specifics of the test tasks, so we did not provide detailed descriptions here. For more information on the tasks, you can refer to Zhang and Zhang (2022) for the old version or the ETS website (2022a) for the new version.

The students' performance was rated by two experienced EFL teachers using the TOEFL iBT integrated speaking rubric (ETS, 2022b), which evaluates oral performance based on four criteria: Delivery (fluency, clarity, and pronunciation), Language Use (grammatical accuracy and vocabulary), Topic Development (cohesion and progression of ideas), and General Description. Each criterion is scored from 0 to 4 points.

Data Collection

We collected data in multimedia lecture rooms where the Chinese EFL learners completed the three test tasks on computers using the TOEFL iBT integrated speaking practice software. After each task, learners filled out the SCICASA questionnaire on their mobile phones via the Chinese online survey platform Wenjuanxing (2021), rating the frequency of their self-monitoring behaviors. For instance, if they often used monitoring, they would select the number 3 for the item "I knew what to do if my intended plan did not work efficiently during the task." This method was chosen for convenience and efficiency (Dörnyei and Taguchi, 2009).

To minimize order effects, we used a Latin-square design to sequence the three tasks and gave learners a 10-minute break between tasks (Weir et al., 2006; Verma, 2015; Corriero, 2017). The TOEFL iBT software recorded each learner's responses, saved under anonymous codes to protect their privacy. These files were randomly assigned to two raters for scoring.

The two raters, experienced Chinese EFL teachers, underwent training to ensure consistency in their ratings, achieving reliability scores above 0.70 (Frey, 2018; Teng, 2022). They scored each of the four segments of the participants' oral performances independently on a scale of 0 to 4 points. The scores were then combined into a composite score, which was averaged to provide a holistic score for each learner's overall performance (Huang and Hung, 2013).

We strictly followed ethical guidelines from the relevant university departments. We obtained official permission from the universities, and participants were fully informed about the study's purpose and their voluntary participation. Consent forms were signed, and participants could request their data be destroyed at any time. Each participant received a small gift worth about 100 CNY and a thank-you letter as a token of appreciation.

Data analysis

We used descriptive analysis to find the average use of self-monitoring by Chinese EFL learners during the three test tasks, addressing RQ1. We also calculated the average speaking performance based on their test scores for further analysis to answer RQ2 and RQ3 (Barkaoui et al., 2013; Ellis et al., 2019). We checked assumptions like data normality by examining standard deviation, skewness, and kurtosis (Pallant, 2016).

To answer RQ2 and RQ3, we created two hierarchical linear models (HLM), each with two levels. This method is recommended for analyzing performance data collected in testing conditions through one-way repeated measures (Barkaoui, 2013, 2015; In'nami and Barkaoui, 2019).

In the full model, Level-1 included the three TOEFL iBT integrated speaking tasks with the test scores as the outcome variables and the tasks as predictor variables. Level-2 included the Chinese EFL learners with their use of self-monitoring as the predictor variable. We examined the effects of interactions between self-monitoring (Level-2) and test tasks (Level-1) on test scores, focusing on cross-level interactions. Data on self-monitoring were centered on the grand mean before entry, while data on test tasks were entered as dummy variables using the k-1 formula. Task 1 served as the baseline, and we created two dummy variables: Task 2 and Task 3. When a learner did Task 2, it was marked as 1 and Task 3 as 0, and vice versa.

The null model had no predictor variables and was used to calculate the Intra-class Coefficient (ICC) to determine the suitability of using HLM on our data. An ICC value close to 1 indicated a good model fit. We evaluated model fit using deviance statistics (smaller values indicate better fit) and significance tests: t-tests for fixed effects and Chi-square tests for random effects (both $p < 0.05$). We also checked the reliability of Level-1 random coefficients and visually inspected the normality of residuals for both levels using Q-Q plots and scatter plots. We used the Fully Maximum Likelihood estimation method (see Barkaoui, 2013, 2015; Zhang et al., 2022b for detailed application of HLM in L2 testing).

Results

Descriptive analysis showed that the data on self-monitoring and speaking performance followed a normal distribution, with skewness and kurtosis values within acceptable ranges (−3≤skewness ≤3; −8≤kurtosis ≤8; Pallant, 2016; Frey, 2018). Based on these results, we examined the averages of the Chinese EFL learners' self-monitoring and speaking performance across the three integrated L2 speaking tasks, as shown in Table 1. The self-monitoring averages ranged from 3 to 3.30, indicating that learners often used self-monitoring while performing the tasks, as 3 means "often" and 4 means "usually" on the SCICASA scale. This answered RQ1 about whether L2 learners use self-monitoring during computer-delivered integrated L2 speaking tests.

The averages of the test scores were used to build two hierarchical linear models (HLM) to address RQ2 and RQ3. Table 2 shows the results of the null model and the full model based on our evaluation of model fit.

In the table, γ01 represents the fixed effects of self-monitoring on the average oral scores across the three tasks, while μ0 indicates the random effects due to learners' individual differences, including their use of self-monitoring, that couldn't be explained by the models. γ11 and γ21 represent the cross-level effects of the interactions between self-monitoring and Task 2 and Task 3, respectively, on the learners' oral scores. These indices were the main focus of our research questions.

From, the ICC value in the null model is 0.63, meaning that 63% of the total variance in the learners' oral scores was due to individual differences (including self-monitoring) at Level-2, while 37% was explained by the tasks at Level-1. This showed that using HLM was necessary and appropriate for our data set to address RQ2 and RQ3 (Raudenbush and Bryk, 2002; Weng, 2009). Additionally, the reliability estimate for the learners' mean oral scores across the three tasks was 0.84, suggesting that about 84% of the variation in each learner's scores could be explained by Level-2 predictors. The deviance value of the null model was 1298.60, which was used in subsequent model comparisons for evaluating model fit (Raudenbush and Bryk, 2002; Weng, 2009).

Tasks	Self-monitoring		Test scores	
	Means	SD	Means	SD
Task 1	3.16	0.88	5.45	2.65
Task 2	3.21	0.87	4.40	2.95
Task 3	3.30	0.89	4.40	2.95

	Null model	Full model
Fixed effects		
Level1coefficient(r)	3.05	
Intercept(γ_{00})(sig.)	4.88(0.00)	5.20(0.00)
Task 2(γ_{10})(sig.)		−0.96(0.00)
Task 3(γ_{20})(sig.)		−0.50(0.00)
Level 2 coefficient (sig.)		
Self - monitoring(γ_{01})		0.10(0.51)
Cross – level interaction coefficient (sig.)		
Self - monitoring in Task 2(γ_{11})		0.21(0.48)
Self - monitoring in Task 3(γ_{21})		−0.17(0.57)
Random effect		
Between – students variance(μ_0) (sig.)	5.32(0.00)	5.34(0.00)
X^2(df)	592.03(94)	640.00(93)
ICC	0.63	
Reliability	0.84	0.85
Model fit		
Deviance (parameters)	1298.60(3)	1280.94(8)

For the full model, Table 2 shows that the coefficient for self-monitoring (γ01) was 0.10, with a p-value of 0.48, which is much larger than the cut-off value of 0.05. This suggests that differences in self-monitoring among Chinese EFL learners had no direct or significant effect on their oral scores across the three tasks. Similarly, the p-values for γ11 (0.48) and γ21 (0.57), which represent the effects of the interactions between self-monitoring and Task 2 and between self-monitoring and Task 3 on the learners' oral scores, were both greater than 0.05. This indicates that these interactions also did not have statistically significant effects on their oral scores. Therefore, self-monitoring did not affect oral scores for Task 1 either, as it was the baseline task.

The reliability value in the full model was 0.85, suggesting that individual differences at Level-2 accounted for 85% of the variance in the learners' oral scores at Level-1. Although the effects of tasks on oral scores were not the main focus of this study, we reported them in Table 2 for a complete interpretation of the results. The table shows that the p-values for γ00 (0.00), representing the average means of the oral scores across the three tasks in both the null and full models, were below 0.05. This indicates significant variance in the mean scores across tasks and individuals. Likewise, the p-values for γ10 (0.00) and γ20 (0.00), which represent the effects of Task 2 and Task 3 on oral scores, were also smaller than 0.05, meaning these tasks had substantial effects on the learners' oral scores. Accordingly, Task 1 also significantly impacted oral scores.

When evaluating model fit, we compared the ordinary standard errors and the robust standard errors and found no significant variance, indicating that the model specification was acceptable. The decrease in deviance values from 1298.60 in the null model to 1280.94 in the full model showed an improvement in model fit. Finally, the high reliability of the Level-1 random coefficient ($\gamma 00=0.85$) and the visual inspection of Q-Q plots and scatter plots of residuals for both Level-1 and Level-2 confirmed that the full model fit well with the current data set (Raudenbush and Bryk, 2002; Weng, 2009; Barkaoui, 2013, 2015) [1-11].

Contributions and Limitations

This study shows that even though Chinese EFL learners reported actively using self-monitoring, it didn't significantly impact their performance in computer-based tests. This finding suggests that self-monitoring might not be effective in L2 speaking test contexts. Therefore, it's recommended that L2 teachers not only teach self-monitoring in regular learning settings but also create specific learning environments that mimic testing conditions. This approach would allow students to practice self-monitoring in more realistic test situations. Additionally, integrating self-monitoring practice into the syllabus, especially in speaking instructions, would give learners more opportunities to refine this skill in authentic tests. This way, learners can effectively use self-monitoring both in everyday learning and high-stakes testing situations, achieving true learner autonomy.

If self-monitoring is only practiced in familiar learning conditions, students might struggle to perform well in various contexts, including tests. This reflects the idea that having good tools is important, but knowing how to use them correctly is crucial. The study also indicates that L2 teachers can use test formats like the TOEFL iBT integrated speaking tasks in their teaching to help students practice self-monitoring in computer-delivered speaking tasks. This practice is especially relevant during COVID-19, which has increased the demand for online learning and testing. Using integrated speaking tasks can help students become familiar with real-world language use, benefiting their self-regulated learning beyond the classroom.

Besides pedagogical implications, the study's findings provide evidence for the self-monitoring process in L1 and bilingual speech production models. The results also offer insights into the definitions and classifications of strategic competence in L2 testing, an area needing further exploration.

However, the study has limitations. The convenience sampling used resulted in a homogenous group of Chinese EFL university students with similar language proficiency and learning experiences, limiting the generalizability of the results to other contexts. Additionally, the study didn't explore why self-monitoring worked as it did in the testing context. Future research should investigate not only "if" and "how" self-monitoring works but also "why" by using methods like interviews, think-aloud protocols, and self-reflections for a comprehensive understanding.

References

1. Bachman LF (2007) "What is the construct? The dialectic of abilities and contexts in defining constructs in language assessment," in Language Testing Reconsidered. eds. https://books.openedition.org/uop/1563?lang=en.
2. J Fox, M Wesche, D Bayliss, L Cheng, CE Turner, et al. (2010) Language Assessment in Practice: Developing Language Assessments and Justifying their Use in the Real World. Oxford: Oxford University Press https://elt.oup.com/teachers/bachmanpalmer/?cc=global&selLanguage=en.
3. Barkaoui K (2013) Using multilevel modeling in language assessment research: a conceptual introduction. Lang. Assess. Q 10: 241-273.
4. Barkaoui K (2015) Test-takers' Writing Activities during the TOEFL iBT® Writing Tasks: A Stimulated Recall Study (Research Report No. RR-15-01), Princeton, NJ: Educational Testing Service.
5. Barkaoui K, Brooks L, Swain M, Lapkin S (2013) Test-takers' strategic behaviours in independent and integrated speaking tasks. Appl. Linguis 34: 304-324.
6. Boksem MA, Tops M, Wester AE, Meijman TF, Lorist MM (2006) Errorrelated ERP components and individual differences in punishment and reward sensitivity. Brain Res 1101: 92-101.
7. Broos WP, Duyck W, Hartsuiker RJ (2016) Verbal self-monitoring in the second language. Lang. Learn 66: 132-154.
8. Broos WP, Duyck W, Hartsuiker RJ (2019) Monitoring speech production and comprehension: where is the second-language delay? Q. J. Exp. Psychol 72: 1601-1619.
9. Bygate M (2011) "Teaching and testing speaking in M." in The Handbook of Language Teaching. eds.
10. H Long, CJ Doughty, Bygate M (2018) Learning Language through Task Repetition. Amsterdam: John Benjamins https://benjamins.com/catalog/tblt.11.
11. Cirino PT, Miciak J, Gerst E, Barnes MA, Vaughn S, et al. (2017) Executive function, self-regulated learning, and reading comprehension: a training study. J. Learn. Disabil 50: 450-467.