Weiwei Zhang, Lawrence Jun Zhang* and Aaron J. Wilson

Strategic competence, task complexity, and foreign language learners' speaking performance: a hierarchical linear modelling approach

https://doi.org/10.1515/applirev-2022-0074
Received June 14, 2022; accepted September 16, 2022; published online October 24, 2022

Abstract: Understanding the intricate relationships among strategic competence, tasks and performance is an issue of perennial interest in the assessment of foreign/second languages especially in integrated speaking assessment, a field that is under-researched. Against this background, we investigated such complex relationships in the context of integrated speaking assessment of English as foreign language (EFL) learners, hoping to provide additional empirical evidence to address the problem. In the investigation, strategic competence was defined as metacognitive strategy use and was measured via an inventory administered on 120 Chinese university EFL students; task characteristics were conceptualised as task complexity and were measured on a self-rating scale by the students and five EFL teachers; and the students' speaking performance was indicated by their scores on four integrated speaking assessment tasks. Data analysis through a hierarchy linear modelling approach led to two primary findings: Monitoring, one form of strategic competence, moderated the effect of task complexity on performance; strategic competence had no substantial effects on performance which had an inverse relationship with task complexity. These findings will add validity evidence for the foreign language speaking assessment literature and provide implications for speaking instruction and test development.

Keywords: hierarchy linear modelling; integrated speaking assessment; strategic competence; task complexity

***Corresponding author: Lawrence Jun Zhang**, School of Curriculum and Pedagogy, Faculty of Education and Social Work, University of Auckland, Auckland, New Zealand, E-mail: lj.zhang@auckland.ac.nz. https://orcid.org/0000-0003-1025-1746
Weiwei Zhang, School of Foreign Languages, Quzhou University, Quzhou City, Zhejiang Province, China. https://orcid.org/0000-0003-0598-155X
Aaron J. Wilson, School of Curriculum and Pedagogy, Faculty of Education and Social Work, University of Auckland, Auckland, New Zealand. https://orcid.org/0000-0002-4593-2288

1 Introduction

Understanding the intricate relationships among strategic competence, the core component of language ability, test tasks and test performance is a significant endeavour in foreign/second language assessment, as Bachman (2007) reiterated almost two decades ago. Unfortunately, this still remains a great challenge (Hughes and Reed 2017), especially in integrated speaking assessment, a field that is under-researched (Frost et al. 2020; Huang et al. 2018). Against this background, we examined the complex relationships among English-as-a-foreign language (EFL) learners' strategic competence, complexity of test tasks and these learners' performance in the context of integrated speaking assessment, hoping that our research efforts can help to address the challenge.

To study the relationships between language ability, tasks and performance, researchers have adopted various approaches which can be categorised into the real-life approach, the interactionalist approach, and the interactional or interactive ability approach (Bachman 2007; Purpura 2016; Sun and Zhang 2022). However, each of these three approaches has its respective limitations: The real-life approach does not accommodate the interactional properties of foreign language assessment; the interactionalist approach lacks theoretical evidence associated with this research field; and the interactional/interactive ability approach focuses on test-takers and is not applicable in investigating the interactions between test-takers and test tasks in foreign language assessment (Bachman 2007; Luoma 2004; Purpura 2016). To address the limitations, researchers have come to a consensus that foreign language assessment is a specific language-use situation in which learners/test-takers should be regarded as language users (e.g., Bachman and Palmer 1996, 2010; Hidri 2018; Weir 2005). Bachman and Palmer (2010) proposed that researchers examine foreign language assessment within interactional language-use frameworks, including the non-reciprocal language-use framework which illustrates the interactions between learners and tasks where no inter-personal conversations occur and the reciprocal language-use framework with a focus on inter-individual communications.

As our study was conducted in the context of the Test of English as a Foreign Language (TOEFL) iBT integrated speaking test, a pioneering and world-widely recognised computer-assisted integrated assessment with high validity and reliability, where no reciprocal interactions between the tester and test-takers take place (Barkaoui et al. 2013), this assured compatibility between the non-reciprocal language-use framework and our research context. Because the compatibility and the recognition of the framework are influential in defining foreign language assessment (Hidri 2018), we framed our study in the non-reciprocal language-use framework. In this framework, strategic competence is conceived as metacognitive strategy use and test tasks are defined as a set of task characteristics corresponding to task complexity variables within Robinson's (2015) Triadic Componential Framework. Strategic competence and test tasks work independently and interactively to affect test performance which is indicated by scores measured via rubrics (Davis 2018; Sato and McNamara 2019).

The three variables mentioned above involve a hierarchical data structure composed of learners (strategic competence and performance) and tasks. In foreign language assessment research, although hierarchical data structure is very common (Barkaoui 2013), statistical techniques most widely applied to examine such a data structure are single-level techniques, including analysis of variance (ANOVA), multiple regression analysis, G-theory, and multi-faceted Rasch models. Consequently, statistical inaccuracy such as biased standard errors and confidence intervals may occur. To avoid this, some researchers advocate a hierarchical linear modelling approach to studying the hierarchical data structure in the assessment of language ability (e.g., Barkaoui 2013; In'nami and Barkaoui 2019). In line with this, we built a two-level hierarchical linear model (HLM), where variables at Level-1 are TOEFL iBT integrated speaking tasks and EFL learners' performance, and those at Level-2 concern the metacognitive strategies used by EFL learners (Raudenbush and Bryk 2002; Nezlek 2011).

In addition to mitigating statistical inaccuracy, the hierarchical linear modelling approach can also be applied to test specific theories and hypotheses (Barkaoui 2013; In'nami and Barkaoui 2019). This indicates that an investigation into the relationships between strategic competence, task characteristics and speaking performance within Bachman and Palmer's (2010) non-reciprocal language-use framework with task characteristics embedded in Robinson's (2015) Triadic Componential Framework in a two-level HLM, as was the case in our study, can not only help to address the perennial problem in foreign language assessment, as noted earlier, but also test the frameworks *per se*. In this sense, our study is expected to provide empirical evidence for the validation of the framework and accordingly contribute to the literature on foreign language assessment and task research. It is also expected to offer implications for foreign language speaking instruction aiming at fostering learners' strategic competence. With regard to foreign language assessment, our study is hoped to bring some insights into developing test tasks for measuring test-takers' strategic competence with high validity and reliability.

2 Review of the literature

2.1 Integrated speaking assessment

Integrated speaking assessment takes into consideration various language skills (e.g., listening, reading and speaking) in one single assessment task to duplicate authentic foreign language-use tasks in the real world. Speaking ability in such authentic contexts is closely related to learners' strategic competence and is valued highly as one of the critical factors affecting foreign language learners' academic success (Crossley and Kim 2019; Frost et al. 2020). Concomitantly, the authenticity of the assessment format empowers it to have positive backwash effects on learning. Many scholars therefore have posited that integrated language skills such as those elicited by integrated speaking assessment should be considered as a fundamental pedagogical component in foreign language classroom instruction (e.g., Alderson et al. 2017; Newton and Nation 2020). The authenticity also provides learners with textual and aural inputs as background knowledge and puts them on equal footing, which enhances test fairness (Crossley and Kim 2019).

Because of these characteristics, integrated speaking assessment tasks have gained increasing importance in both learning and teaching for improving foreign language learners' academic performance (Newton and Nation 2020) and in high-stakes tests such as the TOEFL iBT integrated speaking section for assessing learners' language ability (Frost et al. 2020). Nonetheless, there have been comparatively few studies on integrated speaking assessment, and in particular on the relationships between foreign language learners' strategic competence, task characteristics and performance in such an assessment context (Frost et al. 2020; Huang et al. 2018), which necessitates additional research efforts in this regard. We, therefore, nested our study in the context of integrated speaking assessment formulated by the TOEFL iBT integrated speaking section, a common practice in empirical studies on integrated speaking assessment.

2.2 Strategic competence

In Bachman and Palmer's (2010) non-reciprocal language-use framework, strategic competence is proposed to be the core component of language ability, and it is conceptualised as comprising three metacognitive strategies: Goal setting, appraising and planning. The definitions of the three strategies are shown in Table 1.

Table 1: Bachman and Palmer's (2010) strategic competence model.

Metacognitive strategies	Definitions
Goal setting	Identifying the intended tasks Selecting tasks Deciding whether or not to complete the selected tasks
Appraising	Appraising task characteristics to determine the possibility of task completion and the relevant resources needed Examining the prior knowledge available Evaluating task performance
Planning	Selecting one's prior knowledge available for task completion Formulating plans to complete the tasks Selecting one particular plan for task completion

Bachman and Palmer (2010, p. 49) *Language Assessment in Practice: Developing Language Asessments and Justifying Their Use in the Real World*. Oxford University Press.

The three-metacognitive-strategy model is termed the strategic competence model which has extensive influence in foreign language assessment (Ellis et al. 2019). Despite this, the disagreement in the conceptualisation of strategic competence and the related lack of empirical evidence of the construct still exist (Xu et al. 2022b; Zhang et al. 2021). Therefore, researchers typically study strategic competence as test-takers' use of metacognitive strategies in an exploratory approach with reference to the literature on metacognition and language learning strategies (McNamara 1996; Purpura 2016; Seong 2014).

In the research literature on metacognition and language learning strategies, planning, monitoring and evaluating are the three most widely-acknowledged constituents of metacognitive strategies (e.g., Zhang and Zhang 2019). Problem-solving has also been proposed (Chamot and Harris 2019; Chamot et al. 1999) as one of the fundamental metacognitive strategies because of its usefulness and applicability in dealing with learning tasks as a key component in the available models that highlight the significance of metacognition in language learning strategies (e.g., Anderson 2002; Rubin 2001; Sato and Lam 2021). The salience of problem-solving as a metacognitive strategy in tandem with planning, monitoring, and evaluating and their definitions and taxonomies are illustrated by Chamot et al.'s (1999) Metacognitive Model of Strategic Learning, as summarised in Table 2.

Table 2 shows that planning, monitoring and evaluating in Chamot's et al. (1999) model correspond to goal setting, appraising and evaluating in Bachman and Palmer's (2010) strategic competence model illustrated in Table 1. In the process of foreign language speaking, the three metacognitive strategies work actively and cooperatively with problem-solving in the different stages to ensure

Table 2: Definitions and taxonomies of metacognitive strategies.

MS	Taxonomies	Definitions
Planning	Setting goals	Identify the purpose of the task
	Directed attention	Decide in advance to focus on particular tasks and ignore distractions
	Activate background information	Think about and use what you already know to help you do the task
	Prediction	Anticipate information to prepare and give direction for the task
	Organizational planning	Plan the task and content sequence
	Self-management	Arrange for conditions that help you learn
Problem-solving	Inference	Make guesses based on previous knowledge
	Substitute	Use a synonym or descriptive phrase for unknown words
Monitoring	Selective attention	Focus on key words, phrases, and ideas
	Deduction/induction	Consciously apply learned or self-developed rules
	Personalize/personal experience	Relate information to personal experiences
	Take notes	Write down important words and concepts
	Ask if it makes sense	Check understanding and production to keep track of progress and identify problems
	Self-talk	Talk to yourself to reduce anxiety by reminding self of progress, resources available, goals
Evaluating	Verify predictions and guesses	Check whether your predictions or guesses are correct
	Check goals	Decide whether goals are met
	Evaluating performance	Judge how well you do in the task

MS = metacognitive strategies. This table is adapted from Chamot, A. U., S. Barnhardt, P.B. El-Dinary & J. Robbins. 1999. *The Learning Strategies Handbook* (pp. 15–18). Longman.

smooth production of speech (Bygate 2011; Kormos 2011). In fact, the origin of Bachman and Palmer's (2010) strategic competence model that closely relates to the use of planning, monitoring and evaluating in solving problems, influenced by the Communicative Competence Model (Canale and Swain 1980), which regards strategic competence as problem-solving systems, further suggests the correspondence between planning, monitoring, evaluating to goal setting, appraising and planning and the indispensable part of problem-solving as a strategic competence component.

Taken together, we conceptualised strategic competence as EFL learners' use of metacognitive strategies: planning, problem-solving, monitoring and evaluating. To examine foreign language learners' use of the four metacognitive strategies, we employed Zhang et al.'s (2021) Strategic Competence Inventory for Computer-assisted Speaking Assessment (SCICASA). The rationales for the employment of the inventory were mainly based on the consideration that: (a) The research context of our study is a computer-assisted integrated speaking assessment; (b) an inventory is a commonly-used instrument in empirical studies to measure strategic competence (Craig et al. 2020); and (c) strategic competence assumed to be reported on the SCICASA is foreign language learners' perceived use of metacognitive strategies composed of planning, problem-solving, monitoring and evaluating.

2.3 Test task characteristics

In the non-reciprocal language-use framework, Bachman and Palmer (2010) proposed a Language Use Task Characteristics Model for examining task characteristics. The model accommodates multiple components, including the setting, rubrics, input, response, and the relationship between the input and the response, with each further parsed into various subcomponents. Some scholars (e.g., Fulcher and Reiter 2003; Gan 2012) argue that the model is difficult to apply widely because it is complicated and presented as an unordered checklist.

On the other hand, in task characteristics studies, including those on foreign language speaking assessment tasks, Robinson's (2015) Triadic Componential Framework has been extensively applied as the most detailed and operational framework to date (see e.g., Pallotti 2019; Tabari 2020). In Robinson's framework, componential dimensions of task characteristics are explicitly distinguished and how they affect performance is also explained (Lee 2019; Xu et al. 2022a). In the framework, task characteristics are task complexity, if we follow Robinson's conceptualisation of task complexity as a series of objective task characteristics or variables accounting for variances in task performance (Xu et al. 2022a). Robinson also made a clear distinction between task complexity and task difficulty with the latter referring to task-takers' perceptions of task demands which attribute to performance variability (see Robinson and Gilabert 2007, for elaboration; see also Xu et al. 2022b; Zhang et al. 2017).

Robinson's views on tasks are essentially consistent with Bachman and Palmer's (2010) definition of test tasks: A task is a multi-componential construct and it is a synthesis of diverse properties independent of task-takers. Also, his distinction between task complexity and task difficulty responds to Bachman (2002) call for studying test tasks without mixing them with test-takers. In addition, Robinson's framework demonstrates an information-processing approach (Ellis et al. 2019; Sasayama 2016), in which the cognitive demand/task complexity in foreign language speech production has close associations with task-takers' strategic competence illustrated in Bachman and Palmer's (2010) strategic competence model (Kormos 2011; Skehan 2018).

These features of Robinson's framework justify our establishing the correspondence between task complexity and Bachman and Palmer's (2010) Language Use Task Characteristics Model in examining foreign language speaking assessment tasks. This supports our conceptualisation of the task characteristics of the four integrated speaking assessment tasks as Robinson's task complexity within the non-reciprocal language use framework, which makes reasonable our statistical measurement of task characteristics of the four assessment tasks through the measurement of task complexity.

To measure task complexity, applied cognitive scientists have identified several validated methods, including, (a) self-rating scales/questionnaires, (b) subjective time-estimation, (c) dual-task methodology, (d) psycho-physiological techniques, and (e) expert judgement (Révész et al. 2016; Sasayama 2016). In our study, we administered the self-rating scale developed by Révész et al. (2016) on EFL learners and teachers (as expert judgement) to measure the complexity of the four integrated speaking tasks. The reasons are that the self-rating scale is used the most in task complexity studies and that two sources of data (learners and expert judgement) complement each other, contributing to a more accurate measurement of task complexity of the four speaking tasks (Révész et al. 2016; Sasayama 2016).

2.4 Available empirical studies

To our knowledge, only four existing studies have examined the complex relationships among EFL learners' strategic competence, task characteristics reflected in task complexity and learner performance in foreign language speaking assessment. Among them, Swain et al. (2009) explored 14 Chinese EFL learners' strategic behaviours in the TOEFL iBT speaking tests. Through analysing the think-aloud data, they found that integrated speaking tasks triggered a wide variety of metacognitive strategies, and there was no direct relationship between metacognitive strategy use and test scores. Later, Barkaoui et al. (2013) updated this study and arrived at similar conclusions.

Yi (2012) extended Swain's study by modelling it in two conditions: A testing condition and a classroom learning condition. She collected speech samples of six Korean EFL university students on TOEFL iBT speaking tests, and through analysing their stimulated recall verbalisations, she found that metacognitive strategies were used most frequently under both conditions. Also, she found positive correlations between metacognitive strategy use and task complexity in contrast to the weak relationship between metacognitive strategy use and speaking performance. Likewise, Huang (2013) probed the relationships, in testing and non-testing conditions with a sample of 40 Chinese EFL learners, among strategic competence, task variance, and performance in the IELTS (International English Language Testing System) Speaking test. Via stimulus recall, the study showed that metacognitive strategies did not significantly affect test scores across tasks and conditions.

We notice that in the four studies, the researchers collected data on a small sample, which placed the generalisability of the research findings into question (Seong 2014). Further, these researchers only assumed variance in the test tasks employed in their studies without validating the variance through independently measuring task complexity involved in the test tasks, which should have been done, as suggested by scholars (e.g., Révész et al. 2016; Sasayama 2016). Additionally, these studies examined merely the non-directional relationship between EFL learners' strategic competence and their oral performance (Huang et al. 2018); how EFL learners' strategic competence works in foreign language speaking assessment and what are its relationships with test tasks and test performance remains unknown.

3 Method

To fill the above research gaps, we designed a study in a one-way repeated measures design to investigate the relationships among EFL learners' metacognitive strategy use, task complexity in the four integrated speaking tasks and these learners' oral scores as speaking performance within Bachman and Palmer's (2010) non-reciprocal language-use framework. The investigation, in essence, manifested the research question this study addressed: What are the relationships between strategic competence, task complexity and speaking performance in the context of integrated foreign language assessment?

3.1 Participant

A total of 120 EFL students, five EFL teachers, and two trained EFL raters participated in our study voluntarily via convenience sampling, and the sample size met statistical requirements (Huang and Hung 2013; Raudenbush and Bryk 2002; Révész et al. 2016). The participants came from two comprehensive universities with one locally known for finance and economics and the other specialising in engineering in the east region of the People's Republic of China.

The students were aged from 18 to 21, and on average, they reported 10 years (M = 10.36, SD = 1.95) of formal English language learning experiences. Their average score on the College English Test – Band 4 (CET-4), a validated and nationally-recognised test in China with a high-level reliability and validity, is 460 points out of the full marks of 710 (a score above 425 indicating a pass of the examination). This shows the students' upper-intermediate level language proficiency (Zhang 2017), which validated their perceptions of task difficulty (Rahimi and Zhang 2019; Xu et al. 2022b). All the participants had no training experiences related to the TOEFL iBT integrated speaking section tasks. The five EFL teachers had more than 10 years of English teaching experience with a master's degree in English and the two trained EFL raters had the experience in rating the TOEFL iBT integrated speaking tests.

3.2 Instruments

3.2.1 Strategic competence inventory

As noted earlier, we used Zhang et al.'s (2021) Strategic Competence Inventory for Computer-assisted Speaking Assessment (SCICASA) to measure learners' strategic competence. The inventory includes 23 items under the four constructs of planning, problem-solving, monitoring and evaluating, and each item was rated on a 6-point Likert: 0 (never), 1 (rarely), 2 (sometimes), 3 (often), 4 (usually), and 5 (always). The value of the Cronbach's α for the inter-item consistency is 0.94, much higher than the thumb-up rule (≥0.70), indicating high reliability (Pallant 2016). The SCICASA has also five questions on the students' background information, and it has two versions: English version and Mandarin Chinese version. Considering that the native language of the students is Mandarin Chinese, we adopted the latter version to reduce possible misunderstandings for the validity and reliability of the students' responses (Creswell and Creswell 2018).

3.2.2 Self-rating scale

As stated previously, we adopted the self-rating scale developed by Révész et al. (2016) for measuring task complexity. The scale has two items: One has to do with the students' and the teachers' rating of their mental efforts in performing the assessment tasks; and the other regards how they rated the difficulty of the tasks. A 9 – point Likert scale was used for the ratings: 1 suggests that the task is the easiest and requires no mental efforts at all, and 9 reveals that the task is extremely difficult and requires a lot of mental efforts. The rating values from 1 to 9 indicate an increase in mental efforts and task difficulty perceived by the students and the teachers. The strong correlation between the two item variables suggests the indicator role of task difficulty as task complexity (Révész et al. 2016; Sasayama 2016).

As the original scale is presented in English and the student participants are Chinese, to avoid possible misunderstanding, we translated the scale from the original language of English to Mandarin Chinese and consulted two Chinese linguistics professors to confirm that our translated version expresses what is intended (Creswell and Creswell 2018). Given the high English language proficiency level of the teachers, we administered the original version of the self-rating scale on them (see Révész et al. 2016).

3.2.3 TOEFL iBT integrated speaking test tasks

We chose a set of TOEFL iBT integrated speaking section tasks (Task 1, Task 2, Task 3 and Task 4) from the practice online data of the test (TPO). TPO provides learners with official practice tests featuring real past test questions so that learners can experience the real TOEFL iBT test, which ensures task authenticity. Also, for instrument validity and reliability, we did not make any changes to the selected tasks. Within Robinson's framework, the four tasks varied in four task complexity variables: Prior knowledge (campus life vs. academic lectures), procedures involved (reading, listening and speaking vs. listening and speaking), preparation time (20 vs. 30 s), and reasoning demand (narrating, justification or decision-making).

3.2.4 TOEFL iBT integrated speaking rubric

As stated previously, speaking performance is often reflected by test scores measured via a specific scoring rubric in foreign language assessment (Davis 2018). Chinese EFL learners' speaking performance was therefore indicated by their oral scores in performing the TOEFL iBT integrated speaking test, measured with reference to the TOEFL iBT integrated speaking test rubric. The rubric consists of four criteria: Delivery (fluency, clarity of ideas, and pronunciation), language use (grammatical accuracy and use of vocabulary), topic development (cohesion and progression of ideas), and general description. The rubric was developed by the Educational Testing Service (Huang and Hung 2013).

3.3 Data collection

Data collection took each student approximately 40 min, during which they answered the SCICASA through a Chinese on-line survey system named

WenJuanXing https://www.wjx.cn on mobile phones for convenience each time they completed a task. They also responded to the self-rating scale after they finished all the four tasks. The students performed the tasks on computers installed with the TPO software package in multimedia laboratories. To counterbalance the carryover effect, we offered the students an interval of around 20 min between tasks, and a Latin square design was used to reduce the order effect (Corriero 2017).

The students' speaking performances were recorded automatically on the TPO and stored on computers as a single file which was named after the codes assigned to the student. All the recording files were backed up in case of data loss (Weir et al. 2006) before being ordered through a random list in Microsoft Excel and were given to the two raters for scoring. The scoring method and the rater training procedure for intra-rater and inter-rater reliability were consistent with what was reported in Huang and Hung (2013). In addressing ethics issues, we strictly followed the ethical guidelines by the Human Participants Ethics Committee of The University of Auckland, New Zealand, which approved our study (Reference No. 020972).

3.4 Data analysis

Descriptive analysis was run for the students' reported use of metacognitive strategies, and their oral scores. For scoring validity, inter-rater reliability was examined with reference to the Cronbach's α coefficient (≥0.70). Based on the students' ratings of task difficulty and mental efforts across tasks, Pearson product-moment correlation was used to examine the relationships between task difficulty and mental efforts, with $p \leq 0.05$ indicating a significant correlation between the two variables, which suggests the indicator role of task difficulty as task complexity (Révész et al. 2016). After this, one-way repeated measures ANOVA was conducted to inspect if there was significant variability in task complexity across tasks with the p-value ($p \leq 0.05$) for F-ratio, and the value of η^2 ($\eta^2 \geq 0.01$) for the effect size. The variability indicates various degrees of tasks complexity or different levels of task conditions (Pallant 2016), which is also a fundamental part of assumption testing for running a hierarchy linear model (HLM) to address the research question (Nezlek 2011; Raudenbush and Bryk 2002; Weng 2009).

With regard to the teachers' ratings, we used the average means as experts' judgement to complement the students' ratings so as to further examine if there was variability in task complexity across tasks, and simultaneously enhance the validity of the students' ratings (Révész et al. 2016). We then established a two-level HLM, and variables at the two levels are presented in Table 3.

Table 3: Variables at two levels in the study.

Name of variables	Outcome/predictor variables	Task level	Student level
Oral scores	Outcome variable	⁜	
Task difficulty	Predictor variable	⁜	
Planning	Predictor variable		⁜
Problem-solving	Predictor variable		⁜
Monitoring	Predictor variable		⁜
Evaluating	Predictor variable		⁜

⁜ indicates the level each variable is at in the HLM.

In building a two-level HLM, predictor variables can be entered into the model in a forward/backward elimination approach (entering/eliminating the predictor variables one by one sequentially) or in a block-entry approach (entering all predictor variables simultaneously). The second approach is applied in situations where a research focus is on the relationships among cross-level variables or direct and interactive cross-level effects. Considering that within Bachman and Palmer's (2010) non-reciprocal language-use framework, metacognitive strategies, and task characteristics are proposed to work independently and interactively to impact performance as reviewed previously, our research question relates to investigating a cross-level interaction involving EFL learners' use of metacognitive strategies at Level-2, and task difficulty and test scores at Level-1. Due to this, we employed the block-entry approach, inputting all the predictor variables simultaneously into the two-level model and building a full model or Model 2. The focus of the research question also determined the centring of the predictor variables: Task difficulty was entered into the model as group-mean centred and metacognitive strategies were treated as grand-mean centred (Anderson 2012; Weng 2009).

Before building the full model, we first built a null model or Model 1 without any predictor variables to inspect the Intra-class Coefficient (ICC) for assessing if the current dataset suits the hierarchy linear modelling approach. ICC reflects the proportion of the total variance in the students' oral scores that was accounted for by Level-2 individual differences, including metacognitive strategy use. The ICC ranges from 0 to 1 and if it is close to 1, it is necessary to use the hierarchy linear modelling approach in the current dataset. Also, the null model served as the benchmark value of the deviance for model comparison (Barkaoui 2013).

The estimation method for running the models is the Fully Maximum Likelihood and the models were examined with reference to two main indices: Deviance statistics for the comparison of the model fit (the decrease in the value of deviance indicates better model fit), and significance tests including t-tests for testing parameters' fixed effects ($p < 0.05$) and Chi-square tests to examine parameters' random effects ($p < 0.05$). To evaluate model fit, we also examined the reliability of Level-1 random coefficient, which was complemented by our visual inspecting of the normality of residuals of Levels-1 and Level-2 through Q–Q plots and scatter plots (Raudenbush and Bryk 2002; Weng 2009).

4 Results

4.1 EFL learners' metacognitive strategy use and oral scores across tasks

Descriptive analysis revealed that problem-solving was used most frequently, as reported by the students, followed by planning and evaluating (see Table 4). Monitoring was the least frequently used strategy. As for scoring, the inter-rater reliability in the current study was 0.91, above the rule of thumb-up requirement (>0.70) (Pallant 2016), indicating the statistical validity of the rated scores. Table 4 also shows that Task 1 elicited the highest oral scores, followed by Tasks 4 and 2, while the students' scores on Task 3 ranked the lowest.

Table 5 displays the descriptive statistics of the students' ratings of task difficulty and mental efforts. Pearson correlation analysis showed that mental efforts were significantly and positively correlated with task difficulty across tasks ($p \leq 0.05$). This suggests that task difficulty in this study can be used statistically to indicate task complexity. In the subsequent one-way repeated measures ANOVA, due to the violation of Sphericity assumption testing [$F(5) = 116.90$, $p = 0.000$], the value of Green-house-Geisser epsilon was used for correction (Pallant 2016). Results showed large effects of variance in task difficulty across the four tasks: [$F(2.65, 1{,}586.36) = 81.12$, $p < 0.001$; $\eta^2 = 0.12$]. This result suggests the substantial variability in task complexity across the four integrated foreign language speaking

Table 4: Metacognitive strategies and oral scores across tasks.

Metacognitive strategies	*Mean*	*SD*
Planning	3.45	0.62
Problem-solving	3.61	0.68
Monitoring	3.11	0.65
Evaluating	3.21	0.64
Oral scores	Mean	SD
Task 1	5.45	2.65
Task 2	4.40	3.15
Task 3	3.51	3.15
Task 4	4.86	2.99

Table 5: EFL learners' ratings of task difficulty and mental efforts across tasks.

Tasks	Task difficulty		Mental efforts	
	Mean	***SD***	***Mean***	***SD***
Task 1	5.13	1.85	5.25	1.88
Task 2	5.94	1.70	5.85	1.72
Task 3	6.00	1.82	6.00	1.80
Task 4	5.93	2.01	5.75	2.02

tasks (Pallant 2016; Révész et al. 2016), which validated the four levels of task conditions established by the four tasks.

4.2 EFL teachers' ratings as expert judgement

The teachers' ratings of task difficulty followed the same sequence with their ratings of mental efforts: Task 3 > Task 4 > Task 2 > Task 1, as shown in Table 6. The alignment of the two ratings demonstrates the positive correlation between task difficulty and mental efforts perceived by the EFL teachers. This result complemented the students' ratings and further validated task complexity variability across the four tasks.

4.3 Metacognitive strategy use, task complexity and oral scores

The results of HLM that indicate the relationships between the participants' use of metacognitive strategies, task complexity and their oral scores are manifested through the examination of Models 1 and 2 in Table 7.

Table 6: EFL teacher' ratings of task difficulty and mental efforts across tasks.

Teachers	ME				TD			
	T1	**T2**	**T3**	**T4**	**T1**	**T2**	**T3**	**T4**
Teacher A	2	4	6	7	3	5	6	7
Teacher B	1	2	4	1	2	2	5	2
Teacher C	4	5	6	7	5	6	7	7
Teacher D	5	4	6	4	6	5	8	5
Teacher E	4	5	5	4	4	5	5	4
Means	3.2	4	5.4	4.6	4	4.6	6.25	5

T = tasks, ME = mental efforts, TD = task difficulty.

Table 7: Results of Models 1 and 2.

	Model 1	Model 2
Fixed effects		
Level-1 coefficient (SE)		
Intercept (β_{00})	4.56[a] (0.26)	4.56[a] (0.26)
TD (β_{10})		−0.21[a] (0.08)
Level-2 MS coefficient (SE)		
P (β_{01})		−0.18 (0.53)
PS(β_{02})		0.32 (0.47)
M(β_{03})		0.10 (0.57)
E(β_{04})		0.46 (0.59)
Cross-level interaction coefficient (SE)		
P(β_{11})		0.12 (0.15)
PS (β_{12})		−0.24 (0.13)
M(β_{13})		0.29[b] (0.14)
E(β_{14})		−0.25 (0.15)
Random effect		
Between-students variance (r_0)	5.66[a]	5.81[a]
X^2 (*df*)	724.73 (94)	682.00 (77)
TD slope (r_1)		0.03
X^2 (*df*)		93.62 (77)
Within-student variance (*e*)	3.41	2.87
ICC	0.62	
Reliability		
Intercept (β_{00})	0.87	0.89
TD slope (β_{10})		0.05
Model fit		
Deviance (parameters)	1,737.86 (3)	1,673.00 (11)

SE = standard error; TD = task difficulty; P = planning; PS = problem-solving; M = monitoring; E = evaluating.
[a]$p < 0.01$; [b]$p < 0.05$.

From Table 7, it can be seen that the value of ICC was 0.62, meaning that 62% of the total variance in the students' oral scores was explained by their individual differences at Level-2, which indicated that task difficulty at Level-1 explained about 38% of the total variance in the students' scores. Such a result suggests the necessity and appropriateness of running a HLM for addressing our research question (Barkaoui 2013; Raudenbush and Bryk 2002; Weng 2009). Further, the coefficient of the fixed effects (β_{00}) in Model 1 or the null model was 4.56 ($p < 0.01$), denoting that the mean oral score across the four integrated speaking tasks and the students was 4.56 points, and substantial variance in the mean score existed. Moreover, the p-value of r_0 (5.66) was also less than 0.01, meaning that between-students variance at Level-2 affected significantly the overall mean of the students' oral test scores across the four tasks. Additionally, reliability estimate for the students' oral mean scores across the four integrated speaking tasks was around 0.87, indicating that almost 90% of the variation in the intercept for each student's oral scores across speaking tasks was potentially explicable by individual level or Level-2 predictors. The deviance of Model 1 was 1737.86, which was used in the following model comparisons for evaluating model improvement.

With regard to the full model or Model 2, Table 7 reveals that the coefficient for task difficulty ($\beta_{10} = -0.21$) was significant ($p < 0.01$), suggesting substantially negative effect of task difficulty at Level-1 on students' oral scores. This result cross-validated the variance in the student's oral scores across tasks as presented in Table 4. Chi-square test shows that the variance component for task difficulty (r_1) was not significant ($p > 0.05$), indicating that the relationship between task difficulty and test scores did not statistically vary across the students significantly. On the other hand, the coefficients for planning (β_{01}), problem-solving (β_{02}), monitoring (β_{03}) and evaluating (β_{04}) that referred to the respective fixed effects of the four metacognitive strategies reported by the students on the average mean of their oral scores across tasks were all not significant ($p > 0.05$), which suggested that variances at Level-2 in the students' use of the four metacognitive strategies had no direct and substantial effects on their oral scores across tasks. As for the cross-level interactions, the p values of β_{11}, β_{12}, β_{14}, the three coefficients denoting the respective effects of planning, problem-solving and evaluating on the relationship between task difficulty and the students' oral scores, were all greater than 0.05, revealing that the students' reported use of the three metacognitive strategies did not have statistically significant effects on the relationship between task difficulty and their oral scores. By contrast, the coefficient for the effect of the students' reported use of monitoring on the relationship between task difficulty and their oral scores (β_{13}) was significant ($p < 0.05$), which pointed to a substantial effect of monitoring on the relationship between task difficulty and the students' oral scores. Alternatively stated, monitoring moderated the negative effect of task difficulty on the EFL learners' test scores across the four integrated speaking tasks: The more frequently EFL learners used monitoring, the weaker was the negative effect of task difficulty on their test scores. Figure 1 shows the moderating effect of monitoring (Barkaoui 2013; Raudenbush and Bryk 2002; Weng 2009).

In terms of random effects, the p value of the between-student variance ($r_{0i} = 5.81$) was less than 0.01, implying significant difference in students' mean orals scores across tasks. As for model fit, the examination of the ordinary standard errors and the robust standard errors showed that there was no significant

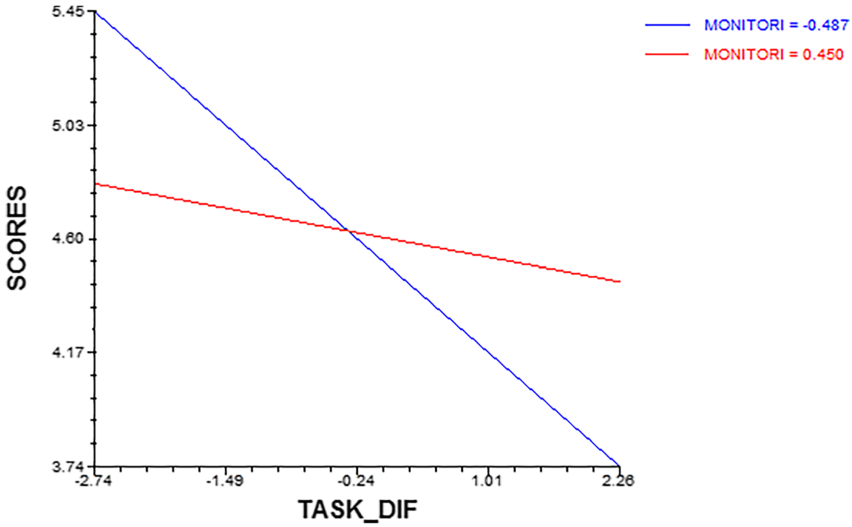

Figure 1: Model graph on the moderating effect of monitoring. SCORES = EFL learners' oral scores; TASK_ DIF = task difficulty; Monitori = monitoring. Blue line = student cohort with more use of monitoring; Red line = student cohort with less use of monitoring.

variance, and hence, model specification was acceptable (Barkaoui 2013; Raudenbush and Bryk 2002; Weng 2009). In addition, the decrease in the values of deviance from 1,737.86 in the null model to 1,673.00 in the full model demonstrated an improvement of model fit. Finally, the investigation of Level-1 random coefficient reliability (β_{00} = 0.89, large than 0.05, the thumb-up rule) and of visual inspecting the Q–Q plots and scatter plots of the residuals for the two levels revealed that the full model fitted well the current dataset (Barkaoui 2013; Raudenbush and Bryk 2002; Weng 2009).

5 Discussion

In our study, we conceptualised strategic competence as foreign language learners' metacognitive strategy use in the forms of planning, problem-solving, monitoring, and evaluating. Statistical analyses revealed that these metacognitive strategies were used to varying degrees by Chinese EFL learners in their completing the four integrated speaking assessment tasks.

To be specific, problem-solving, against the other three metacognitive strategies, unexpectedly demonstrated the highest frequency across the four tasks, although it is not incorporated in Bachman and Palmer's (2010) strategic competence model. The result may relate to the way in which the students performed the integrated speaking tasks. According to O'Malley and Chamot (1990), EFL learners commonly use strategies in a problem-solving manner; so, it is possible that the students considered their use of various strategies as their applications of problem-solving and reported them on the SCICASA. Such a result is also consistent with Oxford's (2017) view that EFL learners tend to use a specific strategy in a particular language skill area and they prefer to use problems-solving in performing speaking tasks.

It is surprising that monitoring was the least frequently used strategy reported by the Chinese EFL students, despite the fact it is widely recognised as indispensable in foreign language speech production (Bygate 2011; Kormos 2011). We tend to think that a possible explanation of the result has to do with the complexity of L2 speech production. During the production, as proposed by Kormos (2011), monitoring is one of the fundamental stages, operating in both covert and overt forms. Since the students reported that they had no prior knowledge regarding how to use metacognitive strategies in responding to our initial survey used for recruiting purposes, it is likely that they were not aware of their actual use of monitoring when the strategy functioned covertly in their speech production. Therefore, when reporting their strategic competence on the SCICASA, they may not truly retrospect their use of the monitoring strategy (Fazilatfar 2010). Furthermore, Barkaoui et al. (2013) have postulated that the immediate and online characteristics of speaking performance impose higher demands on speakers, in comparison with other language skills. These demands may provide few chances for the students to monitor their speaking process where they were challenged simultaneously by the huge cognitive load from the integrated speaking assessment tasks and the time pressure from the assessment (Kormos 2011).

Regardless of its lowest frequency among the four metacognitive strategies, monitoring was unexpected to significantly moderate the negative effect of task complexity on performance. This result may be explained by the critical role of monitoring, as documented in the literature on metacognition (e.g., Efklides 2008; Sun and Zhang 2022; Sun et al. 2021; Zhang and Zhang 2019). It is generally accepted that individuals' metacognition functions at their meta-level, and it associates with the objective world though monitoring and control. Such a view elucidates why Flavell (1979) labelled his metacognition model as the model of cognitive monitoring and why in all the metacognitive strategies, monitoring has been reported as a strong predictor of individuals' academic performance (Shih and Huang 2020). By the same token, in the foreign language learning domain, monitoring operates in an omnipresent form, and is acknowledged as a key factor in an individual's learning process, as the strategy is essential in assisting learners to develop and understand complicated information in the learning process (Zhang and Zhang 2019). Moreover, in foreign language speaking, as reviewed previously, monitoring works in the whole process of speech production as a core error inspector. It is very possible that because of such importance of monitoring across disciplines, this metacognitive strategy, not the other three strategies which had higher frequency under investigation, moderated the effect of task complexity on test scores.

In fact, the fundamental reason for the unexpected results concerning problem-solving and monitoring may relate to the characteristics of Bachman and Palmer's (2010) strategic competence model *per se*. First, the model is proposed as a macro model applicable in the general context of foreign language assessment. In actual studies, due to variability in language skills, the macro model may not be applicable at the micro level featured by a particular language skill. For instance, as reviewed earlier, problem-solving fulfils an essential role in the specific foreign language skill of speaking, but it is excluded in the macro model. Likewise, monitoring is assumed to play an indispensable part in foreign language speech production, but how it works in foreign language speech production under testing/assessment conditions is not clear yet despite its inclusion in the model mainly in the form of appraising (refer to Table 1). Hence, the identification of problem-solving as a salient metacognitive strategy and the seemingly conflicting roles of monitoring in our study not only confirms the importance of the two metacognitive strategies in foreign language speech production (Bygate 2011; Kormos 2011), but also indicates the necessity of contextualising Bachman and Palmer's (2010) macro strategic competence model in line with the language skill/skills under investigation at the micro level in a specific study.

Second, although problem-solving is not explicitly included in the strategic competence model, Bachman and Palmer (1996, 2010) proposed the model with reference to the human intelligence theory (Sternberg 1985) and Communicative Competence Model (Canale and Swain 1980), both of which treat problem-solving as one influential internal working component. Thus, the inexplicit but functioning mode of problem-solving underpinning the strategic competence model is likely to explain why problem-solving, in spite of its absence in the model, was reported to be used by the Chinese EFL learners the most frequently of the four metacognitive strategies. In essence, the confrontation between the absence and the highest frequency demonstrated by problem-solving revealed in our study is supposed to provide empirical evidence for facilitating the comprehensive validation of the strategic competence model, which warrants the inclusion of metacognitive strategies validated by empirical studies as advocated by some scholars (e.g., Phakiti 2016; Seong 2014).

In addressing the research questions, the four metacognitive strategies reported by the students had no significant effect on their test scores. This may be accounted for by task complexity. As displayed in Table 5, the means of the students' ratings of the four tasks ranged from 5.13 to 6. Since the number of 9 on the scale denotes extreme difficult tasks. These values suggest that the students perceived the four integrated speaking test tasks as difficult. When the students found the four tasks difficult, they might turn to whatever resources available to deal with them. Under such conditions, it is understandable that the students used more strategies on all the four difficult tasks, and consequently, their test scores could not manifest considerable changes across the four tasks as reported by Barkaoui et al. (2013) and Swain et al. (2009).

Of the current literature on foreign language assessment, the non-significant effect of metacognitive strategy use on test performance has been reported in a good number of studies. For example, Fernandez's (2018) study showed no positive correlation between metacognitive strategy use and participants' test performance reflected by their test response quality in the IELTS speaking test tasks. Pan and in'nami (2015) also reported a weak relationship between the two variables: Metacognitive strategy use only accounted for 7% of the variance in listening test scores. However, the non-significant effect of metacognitive strategy use on test performance is not consistent with Bachman and Palmer's (2010) strategic competence model, where metacognitive strategy use is proposed to have direct and substantial effects on test performance. Indeed, extensive empirical studies have yielded inconclusive results on the relationship between metacognitive strategy use and test performance (e.g., Pan and In'nami 2015; Purpura 1999). In this sense, our study lends some support to these studies which reflect many researchers' views on the strategic competence model: additional empirical evidence is needed for the model' validation (Seong 2014).

Regarding the considerable effect of task complexity on the students' oral scores discovered through the HLM, theoretically, this result aligns well with Bachman and Palmer's (2010) framework in which test task characteristics are proposed to affect test performance. The result also lends validation support to Robinson' (2015) Triadic Componential Framework where task complexity is assumed to impose impact on task performance. Empirically, the result has been confirmed by an impressive body of literature on foreign language assessment. An example is Zhang et al. (2014), who reported variance in Chinese EFL learners' test performance when they performed four reading tasks characterised by different complexity. In task-based research, the influence of task complexity on speaking performance has also been investigated widely with almost a universal result: Negative effect of task complexity on task performance exists.

6 Conclusions

This study investigated the complex relationships among EFL learner's strategic competence, task characteristics and learner performance in the context of foreign language integrated speaking assessment. The investigation was conducted within Bachman and Palmer's (2010) non-reciprocal language-use framework in a hierarchy linear modelling approach, and it was primarily expected to provide additional empirical evidence for the framework, and hence enrich our understanding of foreign language assessment. Equally, it is hoped to provide pedagogical implications for foreign language speaking instruction and for foreign language test task development.

Pedagogically, despite variations in conducting metacognitive instruction, we can see that the participants' reported use of problem-solving and monitoring in our study suggests that foreign language teachers should pay special attention to the two strategies in designing their syllabuses for metacognitive instruction in speaking so as to foster learners' strategic competence. Such a suggested pedagogical practice is supported by Oxford (2017) and Plosky (2019), both of whom argued for the effectiveness of narrowing down target metacognitive strategies in classroom instruction in accordance with foreign language learners' retrospect of their experiences.

On the other hand, variances in task complexity across tasks and the substantially negative effect of task complexity on performance indicate that in designing tasks for pedagogic purposes, foreign language teachers need to consider whether the complexity of a given task meets the requirements of learners with various levels of language proficiency. If tasks are too easy or too complex, learners' motivation and engagement are likely to be adversely affected, and hence may not perform as expected, which may result in failures in the teachers' classroom instruction (Lynch et al. 2019). The negative effect of task complexity on test performance will provide similar implications for test task development. Test developers should also consider the appropriate level of cognitive complexity that test tasks impose on test-takers, as a test task that is too easy or too complex may generate a test score that cannot truly reflect a test-taker's language ability. This will challenge the validity and reliability of the test and its usefulness (Bachman and Palmer 2010; Hughes and Reed 2017).

Regardless of these promising findings, we need to point out the limitations of our study. In examining task complexity of the integrated foreign language speaking assessment tasks, we treated it as a holistic construct for statistical measurement in order to address our research question. However, as proposed by Robinson (2015), task complexity is a series of task characteristics (e.g., planning time and prior knowledge), and therefore a qualitative in-depth probe into these characteristics, the inter-relationships within them, and the effects of such relationships on performance will bring about a comprehensive understanding of task characteristics involved in foreign language assessment and their influence on learner/test-taker performance. Thus, we suggest that in future studies, if conditions permit, researchers should adopt a mixed-methods research design in examining task complexity as an indicator of task characteristics in integrated foreign language speaking assessment.

Funding: This work was not financially supported by any agency.
Ethics statement: The studies involving human participants were reviewed and approved by The University of Auckland Ethics Committee. The patients/participants provided their written informed consent to participate in this study. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

References

Alderson, J. Charles, Tineke Brunfaut & Luke Harding. 2017. Bridging assessment and learning: A view from second and foreign language assessment. *Assessment in Education: Principles, Policy and Practice* 24(3). 379–387.

Anderson, Daniel. 2012. *Hierarchical linear modeling (HLM): An introduction to key concepts within cross-sectional and growth modeling frameworks*. Eugene: Behavioral Research and Teaching.

Anderson, Neil J. 2002. The role of metacognition in second language teaching and learning. ERIC ED463659. Available at: https://eric.ed.gov/?id=ED463659.

Bachman, Lyle F. 2002. Some reflections on task-based language performance assessment. *Language Testing* 19(4). 453–476.

Bachman, Lyle F. 2007. What is the construct? The dialectic of abilities and contexts in defining constructs in language assessment. In Janna Fox, Mari Wesche, Doreen Bayliss, Liying Cheng, Carolyn E. Turner & Christine Doe (eds.), *Language testing reconsidered*, 41–71. Ottawa, Canada: University of Ottawa Press.

Bachman, Lyle F. & Adrian S. Palmer. 1996. *Language testing in practice: Designing and developing useful language tests*. Oxford, England: Oxford University Press.

Bachman, Lyle F. & Adrian S. Palmer. 2010. *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford, UK: Oxford University Press.

Barkaoui, Khaled. 2013. Using multilevel modeling in language assessment research: A conceptual introduction. *Language Assessment Quarterly* 10(3). 241–273.

Barkaoui, Khaled, Lindsay Brooks, Merrill Swain & Sharon Lapkin. 2013. Test-takers' strategic behaviors in independent and integrated speaking tasks. *Applied Linguistics* 34(3). 304–324.

Bygate, Martin. 2011. Teaching and testing speaking. In Michael H. Long & Catherine J. Doughty (eds.), *The handbook of language teaching*, 412–440. West Sussex, England: Wiley.
Canale, Michael & Merrill Swain. 1980. Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics* 1(1). 1–47.
Chamot, Anna Uhl, Sarah Barnhardt, Pamela Beard El-Dinary & Jill Robbins. 1999. *The learning strategy handbook*. New York, NY: Longman.
Chamot, Anna Uhl & Vee Harris. 2019. Language learning strategy instruction for critical cultural awareness. In Anna Uhl Chamot & Vee Harris (eds.), *Learning strategy instruction in the language classroom: Issues and implementation*, 123–139. Bristol, England: Multilingual Matters.
Corriero, Elena F. 2017. Counterbalancing. In Mike Allen (ed.), *The sage encyclopaedia of communication research methods*, 278–281. Los Angeles, CA: Sage.
Craig, Kym, Daniel Hale, Catherine Grainger & Mary E. Stewart. 2020. Evaluating metacognitive self-reports: Systematic reviews of the value of self-report in metacognitive research. *Metacognition and Learning* 15(2). 155–213.
Creswell, John W. & J. David Creswell. 2018. *Research design: Qualitative, quantitative, and mixed methods approaches*, 5th edn. Los Angeles, CA: Sage.
Crossley, Scott A. & You Jin Kim. 2019. Text integration and speaking proficiency: Linguistic, individual differences, and strategy use considerations. *Language Assessment Quarterly* 16(2). 217–235.
Davis, Larry. 2018. Analytic, holistic, and primary trait marking scales. In John I. Liontas & Ali Shehadeh (eds.), *The TESOL encyclopaedia of English language teaching, vol. II: Approaches and methods in English for speakers of other languages*, 1–6. Hoboken, NJ: Wiley-Blackwell.
Efklides, Anastasia. 2008. Metacognition: Defining its facets and levels of functioning in relation to self-regulation and coregulation. *European Psychologist* 13(4). 277–287. https://psycnet.apa.org/doi/10.1027/1016-9040.13.4.277.
Ellis, Rod, Peter Skehan, Shaofeng Li, Natsuko Shintani & Craig Lambert. 2019. *Task-based language teaching: Theory and practice*. Cambridge, England: Cambridge University Press.
Fazilatfar, Ali Mohammad. 2010. A study of reading strategies using task-based strategy assessment. *Journal of English Language Teaching and Learning* 53(217). 20–44.
Fernandez, Christina Judy. 2018. Behind a spoken performance: Test-takers' strategic reactions in a simulated part 3 of the IELTS speaking test. *Language Testing in Asia* 8(1). 1–20.
Flavell, John H. 1979. Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist* 34(10). 906–911. https://psycnet.apa.org/doi/10.1037/0003-066X.34.10.906.
Frost, Kellie, Josh Clothier, Annemiek Huisman & Gillian Wigglesworth. 2020. Responding to a TOEFL iBT integrated speaking task: Mapping task demands and test takers' use of stimulus content. *Language Testing* 37(1). 133–155.
Fulcher, Glenn & Rosina Marquez Reiter. 2003. Task difficulty in speaking tests. *Language Testing* 20(3). 321–344.
Gan, Zhengdong. 2012. Complexity measures, task type, and analytic evaluations of speaking proficiency in a school-based assessment context. *Language Assessment Quarterly* 9(2). 133–151.

Hidri, Sahbi. 2018. Introduction: State of the art of assessing second language abilities. In Sahbi Hidri (ed.), *Revisiting the assessment of second language abilities: From theory to practice*, 1–19. Cham, Switzerland: Springer.
Huang, Heng-Tsung Danny & Shao Ting Alan Hung. 2013. Comparing the effects of test anxiety on independent and integrated speaking test performance. *TESOL Quarterly* 47(2). 244–269.
Huang, Heng-Tsung Danny, Shao Ting Alan Hung & Lia Plakans. 2018. Topical knowledge in L2 speaking assessment: Comparing independent and integrated speaking test tasks. *Language Testing* 35(1). 27–49.
Huang, Li-Shih. 2013. *Cognitive processes involved in performing the IELTS speaking test: Respondents' strategic behaviours in simulated testing and non-testing contexts. IELTS Research Reports Series, No. 1*. IDP and IELTS Australia.
Hughes, Rebecca & Beatrice Szczepek Reed. 2017. *Teaching and researching speaking*, 3rd edn. New York, NY: Routledge.
In'nami, Yo & Khaled Barkaoui. 2019. Multilevel modeling to examine sources of variability in second language test scores. In Vahid Aryadoust & Michell Raquel (eds.), *Quantitative data analysis for language assessment, vol. 2: Advanced methods*, 150–170. New York, NY: Routledge.
Kormos, Judit. 2011. Speech production and the cognition hypothesis. In Peter Robinson (ed.), *Second language task complexity: Researching the cognition hypothesis of language learning and performance*, 39–60. Amsterdam, The Netherland: Benjamins.
Lee, Jiyong. 2019. Task complexity, cognitive load, and L1 speech. *Applied Linguistics* 40(3). 506–539.
Luoma, Sari. 2004. *Assessing speaking*. Cambridge, England: Cambridge University Press.
Lynch, Raymond, Adrian Hurley, Olivia Cumiskey, Brian Nolan & Bridgeen McGlynn. 2019. Exploring the relationship between homework task difficulty, student engagement and performance. *Irish Educational Studies* 38(1). 89–103.
McNamara, Timothy Francis. 1996. *Measuring second language performance*. London, England: Longman.
Newton, Jonathan M. & I. S. P. Nation. 2020. *Teaching ESL/EFL listening and speaking*, 2nd edn. New York, NY: Routledge.
Nezlek, John Bruce. 2011. *Multilevel modelling for social and personality psychology*. London, England: Sage.
O'Malley, J. Michael & Anna Uhl Chamot. 1990. *Learning strategies in second language acquisition*. Cambridge, England: Cambridge University Press.
Oxford, Rebecca L. 2017. *Teaching and researching language learning strategies*. New York, NY: Routledge.
Pallant, Julie. 2016. *SPSS survival manual: A step by step guide to data analysis using IBM SPSS*, 6th edn. Sydney, NSW: Allen & Unwin.
Pallotti, Gabriele. 2019. An approach to assessing the linguistic difficulty of tasks. *Journal of the European Second Language Association* 3(1). 58–70.
Pan, Yi-Ching & Yo In'nami. 2015. *Re*lationships between strategy use, listening proficiency level, task type, and scores in an L2 listening test. *Canadian Journal of Applied Linguistics* 18(2). 45–77.
Phakiti, Aek. 2016. Test-takers' performance appraisals, appraisal calibration, and cognitive and metacognitive strategy use. *Language Assessment Quarterly* 13(2). 75–108. https://doi-org.ezproxy.auckland.ac.nz/10.1080/15434303.2016.1154555.

Plosky, Luke. 2019. Language learning strategy instructing: Recent research and future directions. In Anna Uhl Chamot & Vee Harris (eds.), *Learning strategy instruction in the language classroom: Issues and implementation*, 3–21. Bristol, England: Multilingual Matters.

Purpura, James E. 1999. *Learner strategy use and performance on language tests: A structural equation modeling approach*. Cambridge: Cambridge University Press.

Purpura, James E. 2016. Second and foreign language assessment. *Modern Language Journal* 100. 190–208. https://www.jstor.org/stable/4413500.

Rahimi, Muhammad & Lawrence Jun Zhang. 2019. Writing task complexity, students' motivational beliefs, anxiety and their writing production in English as a second language. *Reading and Writing* 32(3). 761–786.

Raudenbush, Stephen W. & Anthony S. Bryk. 2002. *Hierarchical linear models: Applications and data analysis methods*, 2nd edn. Thousand Oaks, CA: Sage.

Révész, Andrea, Marije Michel & Roger Gilabert. 2016. Measuring cognitive task demands using dual-task methodology, subjective self-ratings, and expert judgments: A validation study. *Studies in Second Language Acquisition* 38(4). 703–737.

Robinson, Peter. 2015. The cognition hypothesis, second language task demands, and the SSARC model of pedagogic task sequencing. In Martin Bygate (ed.), *Domains and directions in the development of TBLT: A decade of plenaries from the international conference*, 123–155. Amsterdam, the Netherlands: Benjamins.

Robinson, Peter & Roger Gilabert. 2007. Task complexity, the cognition hypothesis and second language learning and performance. *International Review of Applied Linguistics in Language Teaching* 45(3). 161–176.

Rubin, Joan. 2001. Language learner self-management. *Journal of Asian Pacific Communication* 11(1). 25–37.

Sasayama, Shoko. 2016. Is a "complex" task really complex? Validating the assumption of cognitive task complexity. *Modern Language Journal* 100(1). 231–254.

Sato, Masatoshi & Claudia Dussuel Lam. 2021. Metacognitive instruction with young learners: A case of willingness to communicate, L2 use, and metacognition of oral communication. *Language Teaching Research* 25(6). 899–921.

Sato, Takanori & Timothy Francis McNamara. 2019. What counts in second language oral communication ability? The perspective of linguistic laypersons. *Applied Linguistics* 40(6). 894–916.

Seong, Yuna P. 2014. Strategic competence and L2 speaking assessment. *Teachers College, Columbia University Working Papers in TESOL and Applied Linguistics* 14(1). 13–24.

Shih, Hui-Chia Judy & Sheng-Hui Cindy Huang. 2020. EFL learners' metacognitive development in flipped learning: A comparative study. *Interactive Learning Environments*. 1–13. https://doi-org.ezproxy.auckland.ac.nz/10.1080/10494820.2020.1728343.

Skehan, Peter. 2018. *Second language task-based performance: Theory, research, assessment*. New York, NY: Routledge.

Sternberg, Robert J. 1985. *Beyond IQ: A triarchic theory of human intelligence*. Cambridge, England: Cambridge University Press.

Sun, Peijian Paul & Lawrence Jun Zhang. 2022. Speech competence and speech performance: A dual approach to understanding Chinese-as-a-second language learners' speech production ability. *Journal of Second Language Studies* 1–27. https://doi.org/10.1075/jsls.22002.jun.

Sun, Qiyu & Lawrence Jun Zhang. 2022. Examining the effects of English as a foreign language student-writers' metacognitive experiences on their writing performance. *Current Psychology*. https://doi.org/10.1007/s12144-022-03416-0.

Sun, Qiyu, Lawrence Jun Zhang & Susan Carter. 2021. Investigating students' metacognitive experiences: Insights from the English as a foreign language learners' writing metacognitive experiences questionnaire (EFLLWMEQ). *Frontiers in Psychology* 12(744842). 1–15.

Swain, Merrill, Li-Shih Huang, Khaled Barkaoui, Lindsay Brooks & Sharon Lapkin. 2009. *The speaking section of the TOEFL ÏBT (SSTiBT): Test-takers' reported strategic behaviors (TOEFL iBT Research Report No. iBT-10)*. Princeton, NJ: Educational Testing Service.

Tabari, Mahmoud Abdi. 2020. Differential effects of strategic planning and task structure on L2 writing outcomes. *Reading and Writing Quarterly* 36(4). 320–338. https://doi-org.ezproxy.auckland.ac.nz/10.1080/10573569.2019.1637310.

Weir, Cyril J. 2005. *Language testing and validation: An evidence-based approach*. Hampshire, England: Palgrave.

Weir, Cyril J., Barry O'Sullivan & Tomoko Horai. 2006. *Exploring difficulty in speaking tasks: An intra-task perspective. IELTS Research Report No. 6*. British Council and IDP Australia.

Weng, Fuxing. 2009. *The theory and applications of multilevel modelling*. Beijing, China: China Light Industry Press.

Xu, Ting Sophia, Lawrence Jun Zhang & Janet S. Gaffney. 2022a. Examining the relative effectiveness of task complexity and cognitive demands on students' writing in a second language. *Studies in Second Language Acquisition* 4(2). 483–506.

Xu, Zhiqin, Lawrence Jun Zhang & Judy Parr. 2022b. Incorporating peer feedback in writing instruction: Examining its effects on English-as-a-foreign-language (EFL) learners' writing performance. *International Review of Applied Linguistics in Language Teaching*. 1–25. https://doi.org/10.1515/iral-2021-0078.

Yi, Jong-Il. 2012. *Comparing strategic processes in the iBT speaking test and in the academic classroom*. Leicester, UK: University of Leicester Doctoral dissertation.

Zhang, Donglan & Lawrence Jun Zhang. 2019. Metacognition and self-regulated learning (SRL) in second/foreign language teaching. In Xuesong Gao (ed.), *Second handbook of English language teaching*, 883–897. Cham, Switzerland: Springer.

Zhang, Limei, Vahid Aryadoust & Jun Zhang Lawrence. 2014. Development and validation of the test takers' metacognitive awareness reading questionnaire (TMARQ. *The Asia-Pacific Education Researcher* 23(1). 37–51.

Zhang, Lawrence Jun, Liping Wang & Hongyun Wu. 2017. A meta-analysis of linguistic complexity research (1990–2015) conducted within cognitive linguistics frameworks. *Fudan Forum on Foreign Languages and Literature* 10(1). 58–69.

Zhang, Limei. 2017. *Metacognitive and cognitive strategy use in reading comprehension: A structural equation modelling approach*. Singapore: Springer.

Zhang, Weiwei, Lawerence Jun Zhang & Aaron John Wilson. 2021. Supporting learner success: Revisiting strategic competence through developing an inventory for computer-assisted speaking assessment. *Frontiers in Psychology* 12. 689581. 1–14.

Supplementary Material: The online version of this article offers supplementary material (https://doi.org/10.1515/applirev-2022-0074).

Bionotes

Weiwei Zhang
School of Foreign Languages, Quzhou University, Quzhou City, Zhejiang Province, China
https://orcid.org/0000-0003-0598-155X

Weiwei Zhang has earned her PhD in Education (Applied Linguistics & TESOL) from The University of Auckland, New Zealand. She is currently a fulltime lecturer in the Faculty of Foreign Languages and Interactional Education at Quzhou University, Zhejiang Province, China, having had rich experience in managing language programmes and teaching English as a foreign/second language in China and New Zealand. Her publications have appeared in SSCI journals such as *System*, *Sustainability*, *Frontiers in Psychology*, and other Chinese journals of foreign languages and foreign language education.

Lawrence Jun Zhang
School of Curriculum and Pedagogy, Faculty of Education and Social Work, University of Auckland, Auckland, New Zealand
lj.zhang@auckland.ac.nz
https://orcid.org/0000-0003-1025-1746

Lawrence Jun Zhang, PhD, is Professor of Linguistics-in-Education and Associate Dean, Faculty of Education and Social Work, University of Auckland, New Zealand. A past Post-Doctoral Fellow at University of Oxford, he has published extensively on the psychology of language learning and writing in *Applied Linguistics*; *Modern Language Journal*; *Applied Linguistics Review*; *British Journal of Educational Psychology*; *Discourse Processes*; *Reading & Writing*; *Reading & Writing Quarterly*; *Journal of Psycholinguistic Research*; *Instructional Science*; *System*; *TESOL Quarterly*; *Studies in Second Language Acquisition*; *Language, Culture and Curriculum*; *Language and Education*; *Language Teaching Research*; and *Journal of Second Language Writing*. His current interests lie in reading and writing development and language teacher education. He was the sole recipient of the "TESOL Award for Distinguished Research" in 2011 for his article in *TESOL Quarterly*. He is a co-editor for *System*, serving on the editorial boards of *Journal of Second Language Writing*, *Applied Linguistics Review*, *Metacognition & Learning*, *Australian Review of Applied Linguistics*, *Chinese Journal of Applied Linguistics*, and *RELC Journal*. Additionally, he reviews manuscripts for *Applied Linguistics*, *Language Learning*, *MLJ*, *Reading & Writing*, *Reading & Writing Quarterly*, *Language Teaching*, *Language Teaching Research*, *Computer-Assisted Language Learning*, and *ELT Journal*, among other journals.

Aaron J. Wilson
School of Curriculum and Pedagogy, Faculty of Education and Social Work, University of Auckland, Auckland, New Zealand
https://orcid.org/0000-0002-4593-2288

Aaron J. Wilson, PhD, is an Associate Professor in Literacies Education and Associate Dean (Research) at the Faculty of Education and Social Work, The University of Auckland, Auckland, New

Zealand. He researches and writes mainly about literacy, particularly disciplinary and adolescent literacy, as well as about teacher professional learning and development. He has won major competitive research grants with a total of almost one million New Zealand dollars and published extensively research on literacies education in journals such as *Australian Journal of Language and Literacy*; *Reading Research Quarterly*; *Review of Educational Research*; Teachers College Record; The Curriculum Journal, Literacy, among others.