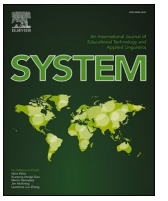
System 111 (2022) 102951

Contents lists available at ScienceDirect

System

journal homepage: www.elsevier.com/locate/system



Understanding assessment tasks: Learners' and teachers' perceptions of cognitive load of integrated speaking tasks for TBLT implementation

Weiwei Zhang [a], Lawrence Jun Zhang [b,*]

[a] School of Foreign Languages and International Education, Quzhou University, Quzhou City, Zhejiang Province, China
[b] School of Curriculum and Pedagogy, Faculty of Education and Social Work, University of Auckland, Auckland, New Zealand

ARTICLE INFO

Keywords:
TBLT
Learners' and teachers' perceptions of cognitive load
Task complexity
Robinson's triadic componential framework
Integrated speaking tasks

ABSTRACT

TBLTis an educational approach to L2 learning and teaching where tasks constitute the main focus of classroom instruction. In its implementation, one of the major challenges facing teachers is the lack of empirical evidence for task design and selection relating to task complexity and the cognitive load imposed by tasks perceived by learners. To address the issue, we investigated EFL learners' and teachers' perceptions of cognitive load of integrated speaking tasks, one type of advocated tasks for TBLT, and factors that affect such perceptions within Robinson's Triadic Componential Framework as proposed by some scholars in the task research field. In a mixed-methods design, we collected data using a self-rating scale, a self-rating questionnaire and semi-structured interviews in the context of integrated speaking tasks for helping shed light on task design and selection criteria. Our analysis revealed that task complexity factors contributed to participants' perceptions of cognitive load, and in the four factors under investigation, prior knowledge was perceived as a stronger determinant of cognitive load than planning time, steps involved and task type. These findings provide empirical evidence for task design and selection in TBLT implementation, especially for the use of integrated speaking tasks in TBLT aiming at EFL speaking instruction. They also provide implications for L2 assessment.

1. Introduction

Task-based language teaching (TBLT) is a pedagogical approach in foreign and/or second language (L2) learning and teaching, emphasising the inclusion of authentic tasks such as integrated skills tasks in a curriculum to create an ideal learner-centred learning condition (Ellis et al., 2019; Shehadeh, 2017). Many task-related issues still remain unclear, which imposes challenges to teachers in implementing the approach. One of them is task design and selection in a syllabus or a lesson plan to meet the requirements of learners with various L2 proficiency levels (Ariatina & Ellis, 2021). To address this issue, some scholars (e.g., Awwad, 2019; Révész & Gurzynski-Weiss; 2016; Tavakoli, 2009) proposed that research be conducted to gain insights into factors that affect task sequencing and grading through learners' and/or teachers' perceptions of the cognitive load of tasks or cognitive task demands denoted by task complexity (Sasayama, 2016). In classroom instruction, learners' perceptions of classroom activities, including the cognitive load involved in tasks, are not always consistent with those of their teachers (Liu et al., 2021; Tavakoli, 2009). Hence, it is pedagogically

* Corresponding author. Faculty of Education and Social Work, University of Auckland, Auckland, 1023, New Zealand.
E-mail address: lj.zhang@auckland.ac.nz (L.J. Zhang).

https://doi.org/10.1016/j.system.2022.102951
Received 7 February 2022; Received in revised form 31 October 2022; Accepted 9 November 2022
Available online 18 November 2022
0346-251X/

possible that a task judged by a teacher as suitable for their students may be perceived by the latter as either too difficult or too easy cognitively. When learners perceive a task as inappropriate, difficult or easy, their task motivation (indicating learners' confidence in performing tasks) and engagement are likely to be negatively influenced, which typically results in a discouraging instructional outcome (Dörnyei, 2019; Zhang et al., 2021a, 2022). As such, the perceptual mismatch pinpoints a problematic area in TBLT (Tavakoli, 2009), highlighting the significance of investigating sequencing and grading tasks through the lens of both learners' and teachers' perceptions of the cognitive load involved in the given tasks. However, limited attention has been devoted to the significance, and so studies along this research line are evidently few (Awwad, 2019; Révész & Gurzynski-Weiss; 2016).

We, therefore, investigated how English-as-a-foreign-language (EFL) learners and teachers perceived the cognitive load of the integrated speaking tasks, a typical type of task in TBLT as explained and discussed in the following section, and what factors affected such perceptions within Robinson's (2015) Triadic Componential Framework (hereinafter referred to as Robinson's Framework) built upon his Cognition Hypothesis (2011), one of the cognitive-interactionist models of TBLT. Our investigation is expected to explore and identify factors that contribute to task performers' perceptions of cognitive load, which will further help to "miminise the perceptual mismatch between task stake holders" (Awwad, 2019, p. 469) in language educators' (e.g., EFL teachers) designing and selecting tasks for effective TBLT implementation. Simultaneously, findings emerging from our investigation are hoped to provide empirical evidence for the use of integrated speaking tasks in TBLT and implications for L2 assessment.

2. Literature review

2.1. Tasks in TBLT: integrated speaking tasks

TBLT has been developing as an innovative and powerful language pedagogy approach since the 1980s (East, 2020; Ellis et al., 2019). Although the main focus of TBLT is tasks, how to define tasks has always been the subject of much debate in the research field of TBLT (Ellis et al., 2019). Scholars have come to an agreement that tasks for TBLT in classroom instruction should be meaning-based, such as communicative tasks that can meet L2 learners' learning needs by duplicating real-life language use tasks for communications, so that learners can engage in real-world language use in classrooms to improve their communicative language ability) (Ellis et al., 2019; Taguchi & Kim, 2018). In the present study, this agreement serves as the definition of tasks for TBLT. In real-life language use, language users commonly receive input in various forms (e.g., listening, reading, writing and speaking) before a communication event starts, and breaking language use into discrete language skills is almost impossible (Zhang et al., 2021b). Therefore, tasks that integrate multiple language skills, compared with independent skill tasks that involve only one language skill, better reflect real-life communication (Bryfonski, 2020; Wu & Alrabah, 2014). The closeness to the real world enables skills-integrated tasks to exemplify the features of tasks for TBLT as delineated above and to be recognised as an essential means of nurturing learners' overall communicative language ability in TBLT (Ellis et al., 2019; Wu & Alrabah, 2014).

Among skills-integrated tasks, integrated speaking tasks that require learners to read and/or listen before performing speaking illustrate Bachman and Palmer's (1990, 2010) Language Ability Model which emphasises L2 learners' knowledge of and strategic competence in tackling communicative tasks in real life (Brooks & Swain, 2014). Obviously, this type of task immediately meets the criterion for TBLT tasks that underscore fostering learners' communicative language ability (Cohen, 2014). In addition, in comparison with other language skills, speaking is "viewed as the most complex and difficult skill to master" (Tarone, 2005, p. 485), and speaking in a foreign language is considered even more complicated (Hughes & Reed, 2017; Gan et al., 2022; Luoma, 2004). EFL learners usually have limited access to practising spoken English outside classroom settings in EFL contexts (Albino, 2017; Sasayama, 2016), including China (Gan et al., 2022; Zhang, 2016) where this study was conducted. Such a situation calls for an exploration of EFL speaking, an under-studied field (Albino, 2017; Frost et al., 2020), for the purpose of TBLT implementation.

Taken together, our definition of TBLT tasks, the relationship between integrated speaking tasks and communicative language ability and L2 learners' insufficient speaking practice beyond classroom instruction support our research focus on integrated speaking tasks for TBLT implementation. In research that involves integrated speaking tasks, tasks from TOEFL iBt (Test of English as a Foreign Language -\ Internet-based test), the widely-used format of integrated skills assessment with established high validity and reliability, are commonly used (e.g., Frost et al., 2020). Even though TOEFL iBt integrated speaking tasks are considered as assessment tasks, they are essentially activities that measure learners' communicative language ability by replicating real-world language use through integrating reading and/or listening with speaking (Brooks & Swain, 2014). So, the testing tasks are consistent with our definition of tasks in TBLT. The consistency and the above-mentioned common practice in researching integrated speaking tasks rationalise our employment of the TOEFL iBt integrated speaking tasks. Moreover, the central role of assessment tasks as determinant in L2 learning and the ever-increasing emphasis on the positive washback effect of L2 assessment on L2 learning (Wu et al., 2021a; see also Wu et al., 2021b) further support our employment of the test format in this study.

2.2. Cognitive task load: Perceptions and contributing factors

In studies on cognitive load of tasks for TBLT, researchers often refer to cognitive-interactionist models among which the two well-known are Skehan's (1998, 2018) Limited Attention Capacity Model and Robinson's (2015) Triadic Componential Framework (Sasayama, 2016; Zhang et al., 2017). With the two options available in framing our investigation into the cognitive load of the TOEFL iBt integrated speaking tasks for the purpose of TBLT, we reviewed the two models in order to make a choice. In general, Skehan's (1998, 2018) Limited Attention Capacity Model proposed that tasks are conceptualised in accordance with three sets of task characteristics: code complexity related to the linguistic demands of tasks on task-performers, cognitive complexity that captures the

cognitive processes elicited by tasks and communicative stress associated with performance conditions under which tasks are completed. On the other side, in Robinson's (2015) Framework, tasks are conceptualised from three aspects: task complexity referring to intrinsic cognitive task load or task demands, task difficulty concerning learners' perceptions of task demands and task condition regarding the conditions under which learners perform tasks (Rahimi & Zhang, 2018; Révész & Gurzynski-Weiss; 2016; Xu et al., 2022). The two models have their own strengths, in that they are both able to explain the cognitive processes of L2 learning despite their seemingly different foci. We decided to adopt Robinson's Framework due to its better alignment with the focus of our study, particularly from the perspectives of L2 speaking and L2 assessment.

In the research field of L2 speaking, Robinson's Framework is proposed to underpin the working mechanism of L2 speakers' attentional control, affecting their speech production (see Kormos, 2011, for the detailed mechanism). During L2 speech production, speakers need to process their attention to deal with various stages involved in speaking for their ultimate utterances. As one's attentional resources are limited, L2 speakers must control their attentional distribution appropriately to these stages for a smooth speech production. The control is presumed to be subject to the conceptual and procedural demands of L2 speaking tasks, and the demands are determined by task complexity factors along the two dimensions of resource-dispersing and resource-directing in Robinson's Framework (Bygate, 2011; Kormos, 2011), as reviewed later.

In L2 assessment studies, it is extensively recognised that a test task should be defined as accommodating various characteristics that are supposed to impact the cognitive task load or cognitive task demands on test-takers, causing a considerable effect on test performance. It is also recognised that test tasks should not be mixed up with test-takers regardless of the interactions between the two concepts in the actual test performance (Bachman, 1990; Fulcher & Reiter 2003; Gan, 2011; Wigglesworth & Frost, 2017). The definition of test tasks and the distinction between test tasks and test-takers empower Robinson's Framework to increasingly gain ground in research on L2 test tasks, as the framework is characterised by its definition of tasks as an incorporation of multiple features and its explicit distinction between tasks and learners (Lee, 2019; Pallotti, 2019). In Robinson's Framework, the distinction is demonstrated by the difference between task complexity and task difficulty. Task complexity refers to task characteristics that indicate the inherent cognitive load of tasks or cognitive task demands on learners, whilst task difficulty is proposed to denote learners' perceptions of the cognitive task load (Révész & Gurzynski-Weiss; 2016). In line with Robinson's Framework, we conceptualised learners' and teachers' perceptions of the cognitive load of the TOEFL iBt integrated speaking tasks as their perceptions of task difficulty and used the two concepts interchangeably in this article.

Apart from its role as underpinning the working mechanism of L2 speaking and its wide applicability in test task research, Robinson's Framework was also motivated by the need of syllabus design and task sequencing for pedagogical purposes. The motivation corroborates with our research purpose, further supporting our use of the framework. In Robinson's Framework, task complexity is "the result of the attention, memory, reasoning, and other information processing demands imposed by the structure of the task on the language learner" (Robinson, 2011, p. 29), and it can be manipulated along the two dimensions of resource-directing and resource-dispersing to generate tasks with varying degrees of complexity for duplicating the full demands of real-world tasks (Lee, 2019). The manipulation of task complexity variables along the resource-directing dimension is intended to change conceptual or linguistic demands of tasks on learners. These variables include: (a) reference to events occurring now in a shared context (Here--and-now) versus events happening in the past elsewhere (There-and-Then); (b) reference to few elements versus many elements (± few elements); (c) whether or not tasks require learners to reason about other people's intentions, beliefs and desires and the relationship between them (± intentional reasoning); (d) whether or not supportive information on spatial location is available (± spatial reasoning); (e) what learners need to do: processing simple information transmission or reasoning the intentions of other people (± causal reasoning); and (f) what perspectives learners have to take, such as the first person perspective or the third person perspective (± perspective-taking).

On the other hand, the manipulation of the variables along the resource-dispersing dimension is assumed to change informative or procedural demands of tasks on learners, and these variables are: (a) whether or not planning time is provided (± planning); (b) if background knowledge of or familiarity with task content is available (± prior knowledge); (c) if few steps or many steps are needed for task completion (± few steps); (d) steps that learners take are in sequence or independent from one another for completing tasks (± independency of tasks); (e) if there is one thing or multiple things to be done before task completion (± single task); and (f) whether or not the structure of a task is clear (± task structure).

According to Robinson and Gilabert (2007), Révész and Gurzynski-Weiss (2016), Awwad (2019), Wang and Zhang (2019), and Rahimi and Zhang (2019), the above manipulation of task complexity variables along the two dimensions is proposed to affect task performers' perceptions of task cognitive load. However, such a proposal has not been empirically and sufficiently evidenced by the existing literature. The reason is that prior studies into task performers' perceptions of task cognitive load and task complexity factors attributing to the perceptions within Robinson's Framework primarily gave weight to the respective relationships between the two constructs and task performance (Révész & Gurzynski-Weiss, 2016; Xu et al., 2022). Studies that examined the relationship between the two constructs *per se* are relatively scant, and these studies mostly investigated such a relationship independently from the lens of either learners or teachers (see Awwad, 2019, for a review).

2.3. Prior empirical studies and research questions

Earlier research warrants studies into task performers' perceptions of task cognitive load and factors accounting for the perceptions with Robinson's Framework. To date, only a few of such studies have been conducted, and Tavakoli's (2009) study is an example. In an exploratory approach through interview data, it found that six task complexity factors displayed by four narrative tasks influenced task difficulty perceived by participants of a small sample size ($n = 10$ for both learners and teachers): cognitive demand, linguistic demand,

clarity of prompts (pictures/stories), amount of information, task structure and affective factors. The study also found that learners' perceptions generally matched those of teachers. Another study by Révész (2011) used two versions (simple versus complex) of a task that differed in two variables within Robinson's Framework (± reasoning and ±few elements) and the self-rating scale to investigate how English as a second language (ESL) students ($n = 43$) and their teachers ($n = 6$) perceived the cognitive demand of the tasks. Results show that all the participants perceived the assumed complex task as more difficult and the assumed simple task as less difficult, which suggests the accountability of the two task complexity variables for the participants' perceptions of task cognitive load as intended and the consistency in the perceptions. A more recent mixed-methods study by Awwad (2019) employed two video-based narrative tasks characterised by the task complexity factor of intentional reasoning in a larger sample ($n = 68$ for learners, $n = 26$ for teachers). The study collected data on participants' perceptions of task cognitive load through open-ended self-rating questionnaires. Results of the study showed that learners' perceptions were consistent with those of teachers, and variance in task complexity resulted in the participants' different perceptions.

The task complexity variables explored by the above studies only consume a small proportion of the task complexity factors proposed in Robinson's Framework (Awwad, 2019; Révész, 2011). This situation suggests that additional research is needed, in particular the research targeting the under-explored task complexity factors along the dimension of resource-dispersing that contribute to learners' perceptions of task cognitive load for TBLT implementation (Awwad, 2019; Révész & Gurzynski-Weiss, 2016). In response to the suggestion, our research focused on Chinese EFL learners and teachers' perceptions of cognitive load of four TOEFL iBt integrated speaking tasks formulated by the simultaneous variations in multiple task complexity factors along the resource-dispersing dimension and factors contributing to the perceptions. The research focus is reframed and represented through the following three specific research questions.

1) What are EFL learners' perceptions of cognitive load imposed by the integrated speaking tasks?
2) What are EFL teachers' perceptions of cognitive load imposed by the integrated speaking tasks?
3) What are the task complexity factors contributing to the EFL learners' and teachers' perceptions of cognitive load imposed by the integrated speaking tasks?

Compared with the above studies, our study utilized a larger sample size. It involved 626 students and six teachers, which will, to some extent, help to expand the existing literature with further empirical evidence.

3. Methodology

3.1. Design

As described in the following sections of data collection and data analysis, we adopted a convergent mixed-methods design. The design permitted us to investigate the quantitative data collected on the self-rating scale for EFL learners' and teachers' perceptions of the cognitive load of the four speaking tasks to answer the first two research questions. The qualitative data generated from the semi-structured interviews and the open-ended question explored the reasons for the learners' and the teachers' ratings respectively. In addition, the design also allowed us to integrate or "converge" the two sources of data to examine the factors accounting for the learners' and the teachers' perceptions of cognitive task load involved in the four speaking tasks for addressing the third research question. Simply put, the convergent mixed-methods design made it feasible for us to know not only "what" but also "why" in order to answer our research questions as suggested by Creswell and Creswell (2018).

3.2. Participants

A total of 626 Chinese EFL university students and six Chinese EFL teachers participated in our study through convenience sampling, with a subsample of eight students participating in the semi-structured interviews. In the sample, male ($n = 238$) and female ($n = 388$) students accounted for 37.66% and 62.34% respectively, and they were between 18 and 21 years old. As the TOEFL iBt integrated speaking has high requirements on test-takers' English language proficiency (Huang & Hung, 2018), to ensure that students are able to perform the given tasks and distinguish the varying degrees of task difficulty (Rahimi & Zhang, 2019), all the student participants had passed the College English Test—Band 4 (CET-4). CET-4 is an authoritative English language proficiency test used in China for university students with a good level of reliability and validity, and the participants' average score for the test was 465 points, indicating their intermediate level language proficiency (Ministry of Education, China, 2016).

Regarding the teachers, we anticipated a large sample size but were able to have only six participants. Apossible reason for this is probably due to the fact that our study was conducted near the end of an academic term when teachers were busy with their marking papers, which hindered their motivation to participate. As relevant literature has documented a similar (n = 6 in Révész, 2011) or even a smaller sample size (n = 2 in Révész et al., 2014), the number of the teacher participants is acceptable to ensure the validity and reliability of the research results concerning the teachers' judgement of cognitive task load. Among the six teachers, two of them were males and four were females, and they were aged between 30 and 41 years. All of them had more than 10 years of English teaching experience with a Master's degree in English Literature,oEnglish Linguistics or Applied Linguistics.

3.3. Instruments

In selecting TOEFL iBt integrated speaking tasks, we considered the students' language proficiency level so that they were capable of self-rating the task difficulty (Rahimi & Zhang, 2018). Because of the established high validity and reliability of the test (Zhang et al., 2021b), we did not make any modifications to the test tasks. Instead, we used the four tasks (Task 1, Task 2, Task 3, and Task 4) included in a regular TOEFL iBt integrated speaking section forenhancing the generalisability of our study (Creswell & Creswell, 2018).

According to ETS (2022), the four TOEFL iBt integrated speaking tasks differ from one another in four aspects: (1) a task is either about campus life or about academic lectures; (2) a task involves either reading, listening and speaking or listening and speaking; (3) a task provides a test-taker with either 30 s or 20 s for preparation; (4) a task requires a test-taker either to narrate or to make justification or to make decision before making justification. With reference to Robinson's Framework, the four aspects denote four task complexity factors along the resource-dispersing dimension: prior knowledge, steps needed, planning time, and reasoning demand or task type as defined by some researchers (e.g., Rahimi & Zhang, 2019; Xu et al., 2022; Youn & Bi, 2019). Table 1 displays the variance in the four task complexity factors, and Appendix A illustrates the detailed description of the four tasks.

In the literature on measuring individuals' perceptions of cognitive task load within Robinson's Framework, several instruments have been empirically tested as valid, including: (a) self-rating scales/questionnaires, (b) subjective time-estimation, (c) dual-task methodology, (d) psycho-physiological techniques such as eye tracking and (e) expert judgment (see Révész, 2014 and Sasayama, 2016 for a comprehensive review of the available instruments). By the same token, a semi-structured interview has been recommended to probe the cognitive task complexity factors by explaining learner's perceptions of cognitive load imposed by tasks (Awwad, 2019; Sasayama, 2016). In this study, we employed the self-rating scale/questionnaire and the semi-structured interview in accordance with the studies reviewed earlier. The employment of the two instruments was also based upon our consideration of the unsuitability of some instruments for test conditions (e.g., time estimation that may cause distraction to test-takers) and resource constraints (e.g., the unavailability of an eye tracking system). In fact, the self-rating scale/questionnaire has been documented as the most widely used instrument in relevant literature (Sasayama, 2016), which also rationalises our use of the instrument.

We used the self-rating scale employed in Révész et al. (2016) to examine EFL learners' overall perceptions of the cognitive load of the four integrated speaking tasks. The scale comprises one item that asks the students to assign a numeric value to their perceptions of cognitive load indicated by the variable of mental effort on a 9-point Likert scale. The value of 1 suggests that the task requires no mental effort at all, whereas 9 indicates extreme mental effort. The scale was translated from English into Chinese after we consulted regarding this translation two Chinese professors specialising in applied linguistics. Subsequently, we piloted the translated scale for its validity and reliability with 20 students. To understand the background information about the students (e.g., Sasayama, 2016; Teng, 2022), we also included six structured questions on the students' age, major and EFL learning experience (see Appendix B). Moreover, at the end of the scale, we included a closing question to recruit interviewees (Creswell & Creswell, 2018; Dörnyei & Taguchi, 2009) which asked those who had the interest in the interviews to provide their email address for further contact.

To explore task complexity factors that affected EFL learner's overall perceptions of cognitive load, we developed a semi-structured interview guide based on Tavakoli (2009). Five questions constitute the interview guide, and an example question is "You have done four speaking tasks, and in general which one do you think is the easiest and which one do you think is the most difficult? And why?" Considering that the students' native language is Chinese, we delivered the guide in Chinese translated from English to determine its comprehensibility (see Appendix C). The translation was made after our consultation with the two professors. The piloting work on the guide was administered to two students for its comprehensibility (Creswell & Creswell, 2018).

An open-ended self-rating questionnaire used in Révész et al. (2016) was distributed to six EFL teachers for expert judgement of the mental effort required by the four integrated speaking tasks. The questionnaire has three sections: one scale, an open-ended question and teacher participants' background information. The format and wording of the scale is the same with the self-rating scale for students, and the open-ended question is to examine teachers' perceptions of cognitive load. The questionnaire was presented in English after the teachers' language proficiency was taken into account (see Appendix D).

3.4. Data collection

Data on students' perceptions of cognitive load were collected through their ratings on the scale immediately after they completed the four integrated speaking tasks and through their interview responses concerning their ratings and task complexity factors that affected the ratings. To minimise the carry-over and the order effects, we deployed a Latin square design in administering the sequence of the four tasks (Révész, 2011). The self-rating process took each student approximately 10 min. The following interviews lasted around 30 min for each interviewee whose responses were audio-recorded for later transcription. Note-taking, audio-recording and

Table 1
Task complexity factors in the four integrated speaking tasks.

Tasks	Prior Knowledge	StepsNeeded	Planning Time (Seconds)	Task Type
Task 1	Campus-life	R-L-S	30	Narrative
Task 2	Lectures	R-L-S	30	Justification
Task 3	Campus-life	L-S	20	Decision-making and justification
Task 4	Lectures	L-S	20	Justification

Note: R = reading, L = listening, S = speaking.

researcher diary were used as a methodology of triangulation to increase the validity and reliability of the responses. For instance, during the face-to-face interviews, one female student told us that the planning time was not useful because the speaking tasks were too difficult. We noticed the interviewee's repeated facial expression of frowning when answering the question about how she viewed the planning time. We took note of the expression which, together with the female students' audio-recorded responses, indicated that planning time did not matter much for the interviewee. So, we inferred that planning time might not be useful for some interviewees in their tackling the speaking tasks if they perceived the tasks as being too difficult.

The collection of teachers' data was conducted around the end of an academic term. For their convenience, we invited the teachers to answer the open-ended questionnaires via emails for their ratings of the cognitive task load involved in the four speaking tasks and factors accounting for their ratings. In the end, we got valid responses from four teachers. During data collection, we addressed ethics issues according to the guidelines set by the Human Participants Ethics Committee of the University of Auckland, Auckland, New Zealand.

3.5. Data analysis

Through descriptive statistics analysis, we examined the variances in the means of the students' rating of the mental effort demanded by the four integrated speaking tasks. To examine whether the variances were significant or not, we further conducted a one-way repeated measures ANOVA during which the four tasks were the four task conditions working as an independent variable, and the means of the four tasks served as the dependent variables. After assumption testing, we checked the p-value for the F-ratio ($p \leq 0.05$) and the η^2 (ranging from 0.01 to 0.06) for effective size that indicate the substantial variance in cognitive load across the four asks (Pallant, 2016). As for the teachers' ratings of mental effort, we averaged the ratings because of the small datasets. The quantitative data analysis addressed the first two research questions.

To probe what task complexity factors contributed to the students' ratings, we used structure coding followed by content analysis in a deductive approach (Creswell & Creswell 2018) to code the interviewees' verbal responses which were transcribed by the first author. The coding scheme was formulated in accordance with the four task complexity factors under investigation (see Table 1) which worked as the subcategories. The presumed learners' attitudes (e.g., in support of longer versus shorter planning time) toward the four factors were employed as the codes (Saldaña, 2016). In total, the coding scheme included 10 codes (see Appendix E).

A research assistant and the first author were in charge of the data coding after a series of training in line with Saldaña (2016) and Richards (2015). Values for their intra-coder and inter-coder agreements were both above 0.9 (Barkaoui et al., 2013; Fernandez, 2018). In coding, the two coders independently coded the verbal responses of the interviewees on NVivo 12 (Windows version). They used vertical and horizontal analyses to code the data. The former targeted individual interviewee's responses separately with special attention to extreme cases, and the latter focused on the commonalities and differences in the individual interviewee's responses in a cross-case manner. Common themes, recurring patterns and uniqueness of any particular case were noted to see if they were within the coding scheme. Regarding those that were beyond the scheme, they were coded as new themes. Final data coding involves re-examination of the data to reassess the coding for refinement (Richards, 2015).

Due to the small datasets, the teachers' answers to the open-ended question were coded manually and independently by the authors. During coding, we deployed the descriptive coding method in a deductive approach with content analysis. Through the method, we used simple phrases to summarise each teacher's answers before annotating them in a particular category. A frequency count of all the annotations for the four tasks was obtained after the annotations were added up and put into a specific category (Révész et al., 2016; Zhang et al., 2021a). Finally, we integrated the qualitative coding results with the quantitative data analysis to investigate the task complexity factors accounting for the learners' and the teachers' perceptions of cognitive task load involved in the four integrated speaking tasks. The integration addressed the third research question.

4. Results

4.1. Learners' perceptions of cognitive load across tasks

The students' ratings of mental effort in Table 2 revealed that their perceptions of the cognitive load of the four tasks varied. They rated Task 3 as the one that involved the largest amount of cognitive load followed by Task 2, Task 4 and Task 1. These ratings helped to address the first research question.

In the assumption testing for the one-way repeated measures ANOVA, results of the Mauchly's test revealed that the assumption of Sphericity was violated [$F(5) = 298.81$, $p = 0.000$, the cut-off rule for p value is above 0.05]. In such situations, the value of

Table 2
Learners' Ratings of Mental effort across Tasks.

Tasks	Mental effort	
	Mean	SD
Task 1	5.61	1.94
Task 2	5.96	1.90
Task 3	6.42	1.91
Task 4	5.69	2.24

Greenhouse-Geisser epsilon was referred to for correction (Frey, 2018; Pallant, 2016). Results of the one-way repeated measures ANOVA with Greenhouse-Geisser correction showed large effects of variance in the within-participants factor of the four tasks on mental effort [F (2.28, 1422.01) = 44.40, $p < 0.05$; $\eta^2 = 0.07$], suggesting the significant variance in the mental effort across tasks perceived by the students (Frey, 2018; Pallant, 2016).

4.2. Task complexity factors affecting learners' perceptions

During the interviews on factors that attributed to task-taker's perceptions of cognitive load, the interviewees prioritised prior knowledge over the other task factors (planning time, steps involved and task type) in judging cognitive load or task difficulty, regarding it as a determinant. According to the interviewees, if prior knowledge of a given task was available, they were able to make effective use of the planning time to make a prediction or make a guess even when the task input was hard for them to comprehend. On the contrary, if relevant prior knowledge was not available, the interviewees tended to consider a given task too difficult to perform, and in such situations planning time made no sense and even was considered as a waste of time. The interviewees' view on prior knowledge is illustrated by one response from Interviewee 7:

My understanding of today's tasks is that if you have no prior knowledge of the topic, you will find the tasks are very difficult because you don't know how to develop your ideas. However, if you are very familiar with the topic, you will know how to express your ideas. Familiarity determines task difficulty. In addition to their agreement on prior knowledge, the coding results presented in Appendix F also indicate the following consensus among the student interviewees:

a) The longer planning time (30 s for Task 1 and Task 2 versus 20 s for Task 3 and Task 4) a task provided, the less difficult the task was.
b) The narrative task that required narrating the speakers' opinions (Task 1) was less difficult than the tasks on justification (Task 2 and Task 4); the task that involved decision-making and justification (Task 3) was rated as the most difficult.
c) Tasks (Task 1 and Task 2) involving three steps (reading, listening and speaking) were less difficult than tasks (Task 3 and Task 4) involving two steps (listening and speaking) because reading was deemed as a prior knowledge provider.

Furthermore, new themes that emerged from the coding results suggest that individual attributes, such as mental fatigue, peer pressure, anxiety and distraction from testing settings were also reported as contributors to the student interviewees' perceptions of task difficulty.

4.3. Teachers' perceptions of cognitive load across tasks

The teachers' ratings of mental effort in Table 3 showed that Task 3 was perceived to involve the largest amount of cognitive load in contrast with Task 1 which requested the smallest amount of cognitive load. Task 4 was sequenced in the second place followed by Task 2. These results answered the second research question.

4.4. Task complexity factors affecting teachers' perceptions

The coding results of the teachers' answers to the open-ended question revealed three common task complexity factors accounting for their perceptions of task difficulty (see Appendix G). These factors are: (a) prior knowledge; (b) task type; (c) steps involved. Among the factors, prior knowledge was reported the most frequently in the teachers' explanations of task complexity factors that contributed to their perceptions. Additionally, the speed at which the listening materials were delivered and task structure, coded as new themes, were also perceived by the teachers as factors influencing task difficulty. Moreover, the coding results showed that the teachers primarily based their perceptions on their teaching experiences. Of note, for an easy and fast locating of the new themes in Appendix F and Appendix G, we highlighted those themes in bold letters in the two appendixes.

5. Discussion

5.1. Learners' perceptions versus teachers' perceptions

The variance in mental effort perceived by the students across tasks in Table 2 suggests the variability in the synergetic effects of the three task complexity factors (planning time, steps involved and prior knowledge) along the resource-dispersing dimension on the

Table 3
Teacher' Ratings of Mental effort across Tasks.

MentalEfforts	Task 1	Task 2	Task 3	Task 4
Teacher 1	2	4	6	7
Teacher 2	1	2	4	1
Teacher 3	4	5	6	7
Teacher 4	5	4	6	4
Means	3	3.75	5.5	4.75

students' perceptions. This result borrows support from Robinson's Framework in which when variables along the two dimensions of task complexity are manipulated simultaneously, the complexity of a specific task is assumed to vary. The variation indicates the variability in the cognitive demands of the task, which further suggests the variance in learners' perceptions of the cognitive task load. Similarly, the variability in task difficulty across task types perceived by the students is supported by Foster and Skehan (1996) who posited that different task types impose various amounts of cognitive load on task performers, generating varying degrees of task difficulty. The students' perceptions of task type are also supported by Gan (2012a,b) who postulated that variance in task types leads to variance in task difficulty.

Table 2 also showed that the students rated Task 3 as the one that required the most amount of mental effort followed by Task 2, Task 4 and Task 1. The rating was not in agreement with that of the teachers reflected in Table 3, and the disagreement did not confirm the three studies reviewed earlier that bear the closest resemblance to our study (Awwad, 2019; Révész, 2011; Tavakoli, 2009), where teacher' perceptions of the cognitive load of a specific task demonstrated agreement with those of learners. It is noticeable that the disagreement in the students' and the teachers' perceptions of the cognitive load relates to Task 4 and Task 2. Despite the small sample size of the teachers and the variance in their ratings, given the fact that the vast majority of Chinese EFL teachers have the educational background of English literature or English Linguistics (Zhang & Gao, 2019), the disagreement may have to do with the sample of the student participants. The background information in the self-rating scale showed that a large percentage of the students were majoring in finance and international trade. Their prior knowledge of finance possibly explained why the students perceived Task 4 as imposing less cognitive load in comparison with Task 2. Task 4 explains an academic concept of money related to finance of which most of the students had prior knowledge that was not available to the Chinese EFL teachers due to the latter's educational background. In contrast, Task 2 regards a psychological term that was not related to the students' majors, denoting an unfamiliar topic to them. The disagreement suggests that the availability of prior knowledge was highly likely to reduce task difficulty perceived by the participants, which borrows much support from current literature on the relationship between prior knowledge and task difficulty (e.g., Ellis et al., 2019; Robinson, 2015; Skehan, 2018).

Regardless of the disagreement between the students' and the teachers' perceptions of cognitive load involved in Task 2 and Task 4, the sequence of the four tasks is generally consistent with Robinson's Framework. When a task (e.g. Task 3) provides less planning time and involves various task types, it will impose more cognitive load on task performers. Likewise, a task (e.g. Task 1) that offers prior knowledge and more planning time and involves fewer task types will lead to a less difficult task perceived by task performers.

5.2. *Task complexity factors affecting leaners' and teachers' perceptions*

The above disagreement on task sequence in terms of task difficulty among the participants also corroborate their perceptions of prior knowledge as a stronger determinant of cognitive load involved in a specific task than planning time, steps involved, and task type. Such perceptions echo Oxford et al. (2004), who suggested that familiarity with tasks makes a big difference for a learner in considering whether a certain task is to be difficult or not. The determinant role of prior knowledge has also borrows some support from Newton and Nation (2020), who advocated for the importance of prior knowledge in determining task difficulty as well as from Wang and Yu (2018), who found that Chinese EFL students tended to refer to their prior knowledge of given tasks in performing those tasks, and whether or not they had the knowledge affected their evaluation of the task difficulty and their task performance.

When assessing task difficulty established by task type, the participants gave their lowest rating to Task 1 which required them to narrate opinions followed by Task 4 and Task 2, the two tasks that asked them to justify an academic concept with examples in a lecture. On the contrary, the participants gave the highest rating to Task 3 that involved decision-making and justification. These results show that the narrative task was perceived by the participants as requiring less cognitive load compared with the justification task and the decision-making task. The perceptions partially vindicate the study conducted by Foster and Skehan (1996), who investigated task difficulty in three types of tasks (personal information exchange, narrative and decision-making) and its impacts on learners' oral performance. The study disclosed that narrative tasks required less cognitive load than decision-making tasks. Further, the results on task type of our study are also evidenced by Rezazadeh et al.'s (2011) study on writing in which descriptive and narrative tasks were discovered to be less cognitively demanding than argumentation/justification tasks.

Regarding the effect of the steps involved in the four integrated speaking tasks on the participants' perceptions of task difficulty, Tables 2 and 3 display that the two tasks (Task 3 and Task 1) that require more steps (reading, listening and speaking) were measured by the participants as the most difficult (Task 3) and the least difficult (Task 1) respectively. In accordance with the task complexity factors operating in the four integrated speaking tasks in Table 1, both Task 3 and Task 1 were about campus life, which indicates the availability of prior knowledge to the participants, and hence the difference between the two tasks lay in task type and planning time. With Task 1 providing longer planning time (30 s versus 20 s) and simpler task type (narrative versus decision-making plus justification) than Task 3, it is explainable that Task 3 was perceived as more difficult than Task 1, in that longer planning time is intended to generate less difficult tasks in Robinson's Framework. However, the participants rated Task 1 as easier than Task 2 and Task 4 that involved fewer steps (two steps), which seemed to conflict with Robinson's Framework where more steps are predicted to impose more cognitive demands and accordingly result in more difficult tasks. Apart from the intricateness of the simultaneous manipulation of multiple task complexity factors, explanation of the conflict may also have to do with the participants' perceptions of the additional step (reading) in Task 1 as a prior knowledge provider instead of additional cognitive demands as intended. The perceptions indeed additionally exemplify the role of prior knowledge as the most important determinant of task difficulty in the four task factors under examination as reported previously by the participants.

Collectively, the participants' perceptions of planning time, prior knowledge and steps involved in the four integrated speaking tasks partially agreed with Robinson's Framework on task complexity: the independent presence of planning time and prior knowledge

was perceived to lead to less difficult tasks, and more steps in comparison with less steps were reported to impose almost similar cognitive load on the students. Further, the participants' perceptions of task type not only echo the above-cited Foster and Skehan (1996) and Rezazadeh et al. (2011) but are substantially documented elsewhere (e.g., Rahimi & Zhang, 2019; Youn & Bi, 2019).

As regards the new themes reported by the participants as factors leading to task difficulty, they involve individual attributes (e.g., mental fatigue, peer pressure and anxiety), task complexity (e.g., task structure) and context of task performance (e.g., distraction caused by testing settings; speed at which listening materials were delivered), the three core components of Robinson's Framework. Specifically, individual attributes associate with learner factors accounting for inter-person variation in individuals' perceptions of task difficulty and task performance. Task structure is one of the task complexity variables and context of task performance indicates task conditions related to variables that are assumed to influence one's perceptions of task difficulty (Robinson, 2011, 2015). The relationships between the new themes and individuals' perceptions of task difficulty may explain why these themes were introspected by the participants as factors accounting for their perceptions of task difficulty. Such relationships have been reported in Tavaloli's (2009) study, in which task structure and individual attributes (affective factors) were identified as the contributors to both EFL learners' and EFL/ESL teachers' perceptions of task difficulty.

6. Conclusions and implications

In a convergent mixed-methods design, we studied EFL learners' and teachers' perceptions of cognitive load of integrated speaking tasks and task complexity factors that explained such perceptions. In general, three major findings of our study corroborate Robinson's Framework, and therefore are expected to provide implications for TBLT implementation.

First, task complexity variability led to EFL learners' and teachers' perceptions of varying degrees of cognitive load involved in integrated speaking tasks. This finding suggests a causal relationship between learners' perceptions of cognitive load of the integrated speaking tasks and external manipulation of task complexity factors within Robinson's Framework. The relationship renders empirical evidence in support of EFL teachers' task design and selection in implementing TBLT through simultaneous variations in multiple task complexity factors for integrated speaking tasks with different difficulty levels to meet the requirements of their students with various levels of language proficiency (Ariatina & Ellis, 2021).

Second, prior knowledge was perceived by both EFL learners and teachers as a stronger determinant of cognitive task load in comparison with planning time, steps involved, and task type. The result suggests that teachers' purposeful manipulation of task complexity factors for designing tasks characterised by intended task difficulty should be done with cognisance of setting a special focus on prior knowledge, in that a task without providing an appropriate amount of prior knowledge might be perceived by EFL learners as difficult. Such perceptions may adversely affect learners' task motivation and engagement, resulting in ultimate failures of a TBLT lesson despite the teachers' painstaking work on task preparations (Dörnyei, 2019; Zhang et al., 2021a).

Third, the general match between EFL learners' and the teachers' perceptions of task difficulty tapped by the self-rating scale

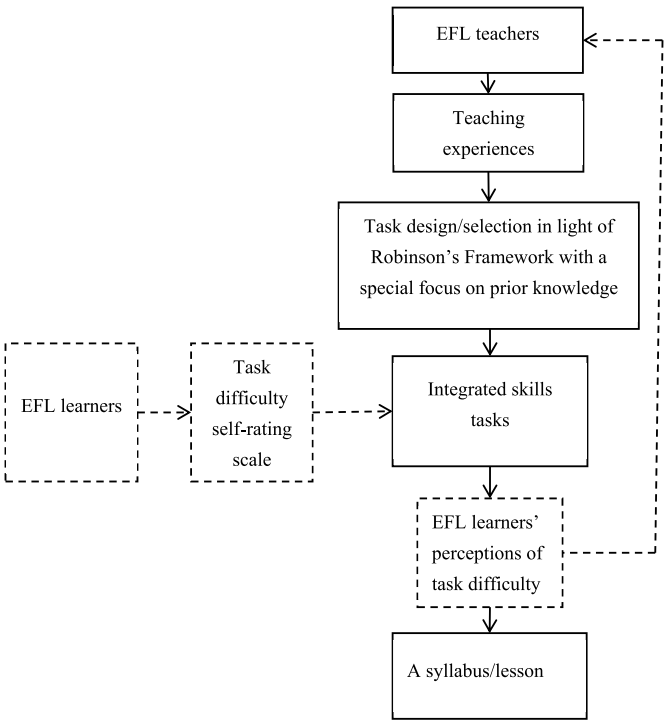


Fig. 1. Procedures of formulating a syllabus/lesson plan for TBLT implementation.

implies that EFL teachers can use the scale to ascertain how students perceive task difficulty in their task design and selection for a TBLT syllabus or lesson plan. Such a practice has been endorsed by scholars, including Robinson (2007) and Sasayama (2016). Moreover, considering the fact that the teachers' ratings of task difficulty were reported to be mainly based on their teaching experience, the match also indicates another pedagogical practice advocated for TBLT by task research experts, such as Ellis et al. (2019). In other words, teachers can refer to their teaching experience to predict their students' evaluation of task difficulty in preparing tasks. In fact, as practitioners actively engaging in training EFL teachers for years, we often implement the two practices in our TBLT classroom instruction to help teachers, especially pre-service and novice teachers to understand their students' perceptions of task difficulty with reference to the students' feedback on tasks during their task design and selection or TBLT practice.

Integrating our research findings and borrowing some support from our own training practice, we propose two procedures that EFL teachers, in particular the pre-service and novice EFL teachers can follow in formulating a syllabus or a lesson plan that meets their students' learning needs for successful TBLT implementation: a procedure with learners' engagement and a procedure without. In the former, EFL teachers design and select integrated skills tasks characterised by simultaneous variability in multiple task complexity factors in Robinson's Framework with a special focus on prior knowledge. These tasks are then rated by students on a task-difficulty-self-rating scale. If the students rate the tasks as appropriately matching their language proficiency, the teachers can include these tasks in their syllabi or lesson plans. Otherwise, the teachers have to re-design or re-select tasks until the students assess the tasks as appropriate. In the second procedure, EFL teachers can design and select tasks based only on their teaching experience without students' engagement.

Clearly, for pre-service and novice teachers, a shift between the two procedures will be, to a great extent, determined by their teaching experience regarding a correct judgement of their students' perceptions of task difficulty which they can also acquire from following the first procedure in addition to getting ideas from other teaching activities. Fig. 1 demonstrates the two procedures with the sections in dotted lines illustrating the double roles of learners' engagement. When these sections are included, the figure demonstrates the first procedure with learners' engagement; when they are removed, the second procedure without learners' engagement is presented.

Given the fact that the four tasks came directly from the authentic TOEFL iBt integrated speaking test, our findings are likely to have some implications for L2 assessment. In developing test tasks, test developers might need to consider task complexity factors that affect test-takers' perceptions of task difficulty so that test tasks can match test-takers' language proficiency levels to elicit expected responses for examining test-takers' language ability and accordingly for ensuring test validity (Bachman & Palmer, 2010; Tavakoli, 2009). Besides, the employment of the test tasks in our study suggests that in preparing tasks for TBLT classroom instruction if teachers can take into careful consideration the appropriate cognitive load of test tasks, then test tasks can also function as well as tasks that are specifically designed for pedagogical purposes. Last but not least, our findings on individual attributes and context of task performance entail the necessity of taking into account not only task factors residing in task complexity but also factors embedded in task performers and task conditions in task preparations for TBLT implementation. The necessity, while complying with Robinson's Framework, points to the complexity of studying tasks for the purpose of TBLT implementation *per se*, which suggests additional work along this strand of research enquiry.

7. Recommendations for further research and limitations

Further research work is necessary due to the limitations of our study. In spite of a large sample size of student participants (n = 626), the number of EFL teachers is comparatively small (n = 4 for the valid sample). Regardless of the rationales for the small sample size as presented earlier, we strongly suggest that a larger sample size of teacher participants be included in future studies of relevance because a small sample size may hinder the generalisability of any study (Creswell & Creswell, 2018). In light of the suggestion, it will be worthwhile to investigate if the findings reported here regarding teacher participants can be transferred to the research context where a large sample size of teachers is available.

The task complexity factors primarily examined in our study were along the resource-dispersing dimension within Robinson's Framework. In the research area of examining task performers' perceptions of cognitive task load and factors leading to the perceptions, both dimensions (resource-dispersing plus resource-directing) of the framework have been acknowledged to be under researched (Awwad, 2019). Therefore, further research following the two directions is recommended.

Author statement

The studies involving human participants were reviewed and approved by The University of Auckland Human 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. Both authors contributed equally to the writing up of the manuscript, with W. Zhang being the person who conceptualised the study with L. J. Zhang's supervision.

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.

Funding

This work was not financially supported by any agency.

Author contributions

The authors have contributed equally to the manuscript, with Professor L. J. Zhang serving as the corresponding author.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.system.2022.102951.

References

Albino, G. (2017). Improving speaking fluency in a task-based language teaching approach: The case of EFL learners at PUNIV-cazenga. *Sage Open, 7*(2), 1–11. https://doi.org/10.1177/2158244017691077

Ariatina, & Ellis, R. (2021). A case study of one teacher's introduction to task-based language teaching Insights for the development of a professional development programme. *Language Teaching for Young learners, 3*(1), 137–158. https://doi.org/10.1075/ltyl.00020.ell

Awwad, A. (2019). L2 learners' and professionals' perceptions of cognitive task complexity: Towards developing an index of task difficulty. *The Journal of Asia TEFL, 16*(2), 468–484. https://doi.org/10.18823/asiatefl.2019.16.2.2.468.

Bachman, L. F. (1990). *Fundamental considerations in language testing.* Oxford University Press.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world.* Oxford University Press.

Barkaoui, K., Brooks, L., Swain, M., & Lapkin, S. (2013). Test-takers' strategic behaviours in independent and integrated speaking tasks. *Applied Linguistics, 34*(3), 304–324. https://doi.org/10.1093/applin/ams046

Brooks, L., & Swain, M. (2014). Contextualizing performances: Comparing performances during TOEFL iBT™ and real-life academic speaking activities. *Language Assessment Quarterly, 11*(4), 353–373. https://doi.org/10.1080/15434303.2014.947532

Bryfonski, L. (2020). Current trends and new developments in task-based language teaching. *ELT Journal, 74*(4), 492–511. https://doi.org/10.1093/elt/ccaa043

Bygate, M. (2011). Teaching and testing speaking. In M. H. Long, & C. J. Doughty (Eds.), *The handbook of language teaching* (pp. 412–440). Wiley.

Cohen, A. D. (2014). *Strategies in learning and using a second language.* Routledge.

Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed- methods approaches* (5th ed.). Sage.

Dörnyei, Z. (2019). From integrative motivation to directed motivational currents: The evolution of the understanding of L2 motivation over three decades. In M. Lamb, K. Csizér, A. Henry, & S. Ryan (Eds.), *The Palgrave handbook of motivation for language learning* (pp. 39–69) (Palgrave).

Dörnyei, Z., & Taguchi, T. (2009). *Questionnaires in second language research: Construction, administration, and processing.* Routledge.

East, M. (2020). Task-based language teaching as a tool for the revitalisation of Te Reo Māori: One beginning teacher's perspective. *Language Learning Journal, 48*(3), 272–284. https://doi.org/10.1080/09571736.2020.1719433

Ellis, R., Skehan, P., Li, S., Shintani, N., & Lambert, C. (2019). *Task-based language teaching: Theory and practice.* Cambridge University Press.

ETS. (2022). TOEFL iBT® integrated speaking tasks. Retrieved from https://www.ets.org/toefl/test-takers/ibt/about/content/speaking/.

Fernandez, C. (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. https://doi.org/10.1186/s40468-018-0073-4

Foster, P., & Skehan, P. (1996). The influence of planning and task type on second language performance. *Studies in Second Language Acquisition, 18*(3), 299–323. https://doi.org/10.1017/S0272263100015047

Frey, B. B. (2018). *The sage encyclopaedia of educational research, measurement, and evaluation.* Sage.

Frost, K., Clothier, J., Huisman, A., & Wigglesworth, G. (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. https://doi.org/10.1177/0265532219860750

Fulcher, G., & Reiter, R. M. (2003). Task difficulty in speaking tests. *Language Testing, 20*(3), 321–344. https://doi.org/10.1191/0265532203lt259oa

Gan, Z. (2012a). Second language task difficulty: Reflections on the current psycholinguistic models. *Theory and Practice in Language Studies, 1*(8), 921–927. https://doi.org/10.4304/tpls.1.8.921-927.

Gan, Z. (2012b). Complexity Measures, Task Type, and Analytic Evaluations of Speaking Proficiency in a School-Based Assessment Context. *Language Assessment Quarterly, 9*(2), 133–151. https://doi.org/10.1080/15434303.2010.516041

Gan, Z. et al. (2022). Development and validation of an EFL speaking self-efficacy scale in the self-regulated learning context. *Journal of Asia TEFL, 19*(1), 35–49.

Huang, H. D., Hung, S. A., & Plakans, L. (2018). Topical knowledge in L2 speaking assessment: Comparing independent and integrated speaking test tasks. *Language Testing, 35*(1), 27–49. https://doi.org/10.1177/0265532216677106

Hughes, R., & Reed, B. S. (2017). *Teaching and researching speaking* (3rd). Routledge.

Kormos, J. (2011). Speech production and the cognition Hypothesis. In P. Robinson (Ed.), *Second language task complexity: Researching the Cognition Hypothesis of language learning and performance.* John Benjamins.

Lee, J. (2019). Task complexity, cognitive load, and L1 speech. *Applied Linguistics, 40*(3), 506–539. https://doi.org/10.1093/applin/amx054

Liu, Y., Mishan, F., & Chambers, A. (2021). Investigating EFL teachers' perceptions of taskbased language teaching in higher education in China. *Language Learning Journal, 49*(2), 131–146. https://doi.org/10.1080/09571736.2018.1465110

Luoma, S. (2004). *Assessing speaking.* Cambridge University Press.

Ministry of education, China. https://cet.neea.edu.cn/html1/folder/19081/5124-1.htm, (2016).

Newton, J. M., & Nation, I. S. P. (2020). *Teaching ESL/EFL listening and speaking* (2nd. ed.). Routledge.

Oxford, R., Cho, Y., Leung, S., & Hae-Jin, K. (2004). Effect of the presence and difficulty of task on strategy use: An exploratory study. *International Review of Applied Linguistics in Language Teaching, 42*(1), 1–47. https://doi.org/10.1515/iral.2004.001

Pallant, J. (2016). *SPSS survival manual: A step by step guide to data analysis using IBM SPSS* (6th ed.). Allen & Unwin.

Pallotti, G. (2019). An approach to assessing the linguistic difficulty of tasks. *Journal of the European Second Language Association, 3*(1), 58–70. https://doi.org/10.1080/09571736.2018.1465110

Rahimi, M., & Zhang, L. J. (2018). Effects of task complexity and planning conditions on L2 argumentative writing production. *Discourse Processes, 55*(8), 726–742. https://doi.org/10.1080/0163853X.2017.1336042

Rahimi, M., & Zhang, L. J. (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. https://doi.org/10.1007/s11145-018-9887-9

Révész, A. (2011). Task complexity focus on L2 constructions, and individual differences: A classroom based study. *The Modern Language Journal, 95*(4), 162–181. https://doi.org/10.1111/j.1540-4781.2011.01241.x

Révész, A. (2014). Towards a fuller assessment of cognitive models of task-based learning: Investigating task-generated cognitive demands and processes. *Applied Linguistics, 35*(1), 87–92. https://doi.org/10.1093/applin/amt039
Révész, A., & Gurzynski-Weiss, L. (2016). Teachers' perspectives on second language task difficulty: Insights from think-aloud and eye tracking. *Annual Review of Applied Linguistics, 36*, 182–204. https://doi.org/10.1017/S0267190515000124
Révész, A., Michel, M., & Gilabert, R. (2016). Measuring cognitive task demands using dual task methodology, subjective self-ratings, and expert judgments: A validation study. *Studies in Second Language Acquisition, 38*(4), 703–737. https://doi.org/10.1017/S0272263115000339
Révész, A., Sachs, R., & Hama, M. (2014). The effects of task complexity and input frequency on the acquisition of the past counterfactual construction through recasts. *Language Learning, 64*, 615–650. https://doi.org/10.1111/lang.12061
Rezazadeh, M., Tavakoli, M., & Rasekh, A. E. (2011). The role of task type in foreign language written production: Focusing on fluency, complexity, and accuracy. *International Education Studies, 4*(2), 169–176. https://doi.org/10.5539/ies.v4n2p169
Richards, L. (2015). *Handling qualitative data: A practical guide* (3rd ed.). Sage.
Robinson, P. (2007). Task complexity, theory of mind, and intentional reasoning: Effects on L2 speech production, interaction, uptake and perceptions of task difficulty. *International Review of Applied Linguistics in Language Teaching, 45*(3), 193–213. https://doi.org/10.1515/iral.2007.009
Robinson, P. (2011). Task based language learning: A review of issues. *Language Learning, 61*(1), 1–36. https://doi.org/10.1111/j.1467-9922.2011.00641.x
Robinson, P. (2015). The Cognition Hypothesis, second language task demands, and the SSARC model of pedagogic task sequencing. In M. Bygate (Ed.), *Domains and directions in the development of TBLT: A decade of plenaries from the international conference* (pp. 123–155) (Benjamins).
Robinson, P., & Gilabert, R. (2007). Task complexity, the Cognition Hypothesis and second language learning and performance. *International Review of Applied Linguistics in Language Teaching, 45*(3), 161–176. https://doi.org/10.1515/iral.2007.007
Saldaña, J. (2016). *The coding manual for qualitative researchers* (3rd). Sage.
Sasayama, S. (2016). Is a 'complex' task really complex? Validating the assumption of cognitive task complexity. *The Modern Language Journal, 100*(1), 231–254. http://www.jstor.org/stable/44134982.
Shehadeh, A. (2017). Foreword: New frontiers in task-based language teaching research. In M. Ahmadian, & M. Mayo (Eds.), *Recent perspectives on task-based language learning and teaching* (pp. vii–xxii). De Gruyter.
Skehan, P. (1998). *A cognitive approach to language learning*. Oxford University Press.
Skehan, P. (2018). *Second language task-based performance: Theory, research, assessment*. New Routledge.
Taguchi, N., & Kim, Y. J. (2018). Task-based approaches to teaching and assessing pragmatics. In N. Taguchi, & Y. J. Kim (Eds.), *Task-based approaches to teaching and assessing pragmatics* (pp. 1–26). John Benjamins.
Tarone, E. (2005). Speaking in a second language. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 485–502). Lawrence Erlbaum.
Tavakoli, P. (2009). Investigating task difficulty: Learners' and teachers' perceptions. *International Journal of Applied Linguistics, 19*(1), 1–25. https://doi.org/10.1111/j.1473-4192.2009.00216.x
Teng, S. L. (2022), *26. Self-regulated Learning and Second Language Writing: Fostering strategic language learners*. Springer Nature.
Wang, H., & Yu, G. (2018). Test-takers' cognitive processes during a listen-to-summarize cloze task. *International Journal of Listening, 35*(1), 1–28. https://doi.org/10.1080/10904018.2018.1473040
Wang, L., & Zhang, L. J. (2019). Peter Skehan's influence in research on task difficulty: A bibliometric analysis using CiteSpace. In Z. E. Wen, & M. J. Ahmadian (Eds.), *Researching L2 task performance and pedagogy: In honour of Peter Skehan* (pp. 183–196). https://doi.org/10.1075/tblt.13.09wan. Benjamins.
Wigglesworth, G., & Frost, K. (2017). Task and performance-based assessment. In E. Shohamy (Ed.), *Encyclopaedia of language and education* (2nd ed., pp. 121–133). Springer.
Wu, S., & Alrabah, S. (2014). Tapping the potential of skill integration as a conduit for communicative language teaching. *English Language Teaching, 7*(11), 119–129. https://doi.org/10.5539/elt.v7n11p119
Wu, X. M., Zhang, L. J., & Dixon, H. R. (2021a). EFL teachers' understanding of Assessment for Learning (AfL) and the potential challenges for its implementation in Chinese university EFL classes. *System, 101*, 1–14. https://doi.org/10.1016/j.system.2021.102589, 102589.
Wu, X. M., Zhang, L. J., & Liu, Q. (2021b). Using Assessment for Learning (AfL): Multi-case studies of three Chinese university English-as-a-foreign-language (EFL) teachers engaging students in learning and assessment. *Frontiers in Psychology, 12*, 1–16. https://doi.org/10.3389/fpsyg.2021.725132, 725132.
Xu, T. S., Zhang, L. J., & Gaffney, J. S. (2022). Examining the relative effectiveness of task complexity and cognitive demands on students' writing in a second language. *Studies in Second Language Acquisition, 44*(2), 483–506. https://doi.org/10.1017/S0272263121000310
Youn, S. J., & Bi, N. Z. (2019). Investigating test-takers' strategy use in task-based L2 pragmatic speaking assessment. *Intercultural Pragmatics, 16*(2), 185–218. https://doi.org/10.1515/ip-2019-0009
Zhang, L. J. (2016). Reflections on the pedagogical imports of western practices for professionalizing ESL/EFL writing and writing-teacher education. *Australian Review of Applied Linguistics, 39*(3), 203–232. https://doi.org/10.1075/aral.39.3.01zha
Zhang, L., & Gao, S. (2019). 40 years of research on foreign language teacher education and development in China: Review and prospect. *Foreign Language Education in China, 2*(1), 3–12.
Zhang, L. J., Wang, L., & Wu, H. (2017). A meta-analysis of linguistic complexity research (1990-2015) conducted within cognitive linguistics frameworks. *Fudan Forum on Foreign Languages and Literature, 10*(1), 50–62. https://mall.cnki.net/magazine/Article/PFDW201701009.htm.
Zhang, W., Zhang, L. J., & Wilson, A. J. (2021b). Supporting learner success: Revisiting strategic competence through developing an inventory for computer-assisted speaking assessment. *Frontiers in Psychology, 12*, Article 689581. https://doi.org/10.3389/fpsyg.2021.689581
Zhang, W., Zhang, L. J., & Wilson, A. J. (2022). Strategic competence, task complexity, and foreign language learners' speaking performance: A hierarchical linear modelling approach. *Applied Linguistics Review*. https://doi.org/10.1515/applirev-2022-0074. online First).
Zhang, W., Zhang, D., & Zhang, L. J. (2021a). Metacognitive instruction for sustainable learning: Learners' perceptions of task difficulty and use of metacognitive strategies in completing integrated speaking tasks. *Sustainability, 13*(11), 6275. https://doi.org/10.3390/su13116275

Weiwei Zhang has earned her PhD in Education (Applied Linguistics/TESOL) from The University of Auckland, New Zealand. She is currently a fulltime lecturer at Quzhou University, 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, Applied Linguistics Review, Sustainability, Frontiers in Psychology,* and other Chinese journals of foreign languages and foreign language education. Email: weiwei.zhang@qzc.edu.cn

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 various aspects of the psychology of language learning and teaching in *Applied Linguistics; Applied Linguistics Review; Modern Language Journal; British Journal of Educational Psychology*; *Current Psychology; Discourse Processes; English Today; Reading & Writing; Reading & Writing Quarterly; Journal of Psycholinguistic Research; Instructional Science*; *System*; *TESOL Quarterly; Language Awareness; Language, Culture and Curriculum; Language and Education; Language Teaching Research; Journal of Second Language Writing; Journal of Multilingual and Multicultural Development; International Journal of Bilingualism; International Journal of Bilingual Education and Bilingualism; International Review of Applied Linguistics in Language Teaching; RELC Journal*, among others. 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 a good number of journals, including *Applied Linguistics*, *Language Learning, Modern Language Journal, Reading & Writing, Reading & Writing Quarterly*, *Language Teaching, Language Teaching Research, Computer-Assisted Language Learning,* and *ELT Journal*, among others. In the recent Stanford University Rankings of scholars' academic impact, he has been listed in the top 2% Scientists in the World in the disciplinary areas of Linguistics/Language/Applied Linguistics. Email: lj.zhang@auckland.ac.nz