LANGUAGE LEARNING
A Journal of Research in Language Studies

Language Learning ISSN 0023-8333

REGISTERED REPORT

First Language Literacy and Second Language Oracy: A Partial Replication of Foster and Skehan (1996)

Jonathon Ryan [a], Pauline Foster,[b] Anthea Fester,[a] Yi Wang,[a] Jenny Field,[a] Celine Kearney,[a] and Jia Rong Yap[a]

[a]Wintec | Te Pūkenga [b]University College London

Abstract: This article responds to calls for greater inclusivity in second language acquisition research and, more specifically, to calls to explore further the impact of first language literacy on second language oracy (e.g., Tarone et al., 2009). We conducted a partial replication of Foster and Skehan's (1996) influential study of task complexity, planning time, and performance over measures of complexity, accuracy, and fluency. The initial study and others had provided robust evidence to suggest that planning time had a positive impact on task performance, particularly for more cognitively demanding tasks. We conducted our replication with adult second language learners with low first language literacy, most of whom were former refugees. Contrary to previous studies, the findings indicate little to no evidence that planning time led to improved linguistic performance. It is not immediately clear why this should be so, and our findings highlight the need for further research with this underrepresented group.

CRediT author statement – **Jonathon Ryan**: conceptualization; funding acquisition; formal analysis (supporting); project administration; investigation (equal); writing – original draft (equal); writing – review & editing (equal). **Pauline Foster**: methodology; formal analysis (lead); investigation (equal); resources; supervision; writing – original draft (equal); writing – review & editing (equal). **Anthea Fester**: investigation (equal); formal analysis (supporting); writing – review & editing (supporting). **Yi Wang**: data curation; investigation (equal). **Jenny Field**: investigation (equal). **Celine Kearney**: investigation (equal). **Jia Rong Yap**: investigation (equal).

A one-page Accessible Summary of this article in non-technical language is freely available in the Supporting Information online and at https://oasis-database.org

This study was supported by funding from Trust Waikato.

Correspondence concerning this article should be addressed to Jonathon Ryan, Wintec | Te Pūkenga, Private Bag 3036, Waikato Mail Centre, Hamilton 3240, New Zealand. Email: jonathon.ryan@wintec.ac.nz

The handling editor for this manuscript was Aline Godfroid.

DOI: 10.1111/lang.12557

Keywords first language; second language; literacy; planning time; task performance; syntactic variety; accuracy; fluency

Introduction

An estimated 781 million adults worldwide have limited first language (L1) literacy, including over half the populations of some countries such as Afghanistan and Somalia (Education for All Global Monitoring Report Team, 2015).[1] Substantial numbers have also been found in second language (L2) classes in major cities of the Global North, grappling not only with an unfamiliar language but learning for the first time a written script, different ways of thinking about language, and the fine motor skills and uses of eyesight required to manipulate pen, paper, and computers (e.g., Goody, 2000). Despite these populations remaining "almost completely unstudied" within second language acquisition (SLA; Tarone, 2014, p. 19), research has consistently suggested that literacy influences the processing of linguistic information (e.g., Havron & Arnon, 2017), and that this is associated with distinct ways of learning L2s (Tarone et al., 2009). There is, therefore, a pressing need to address this oversight in the literature, both from a theory-building SLA perspective and, more particularly, to ensure that the instructional needs of this group are better understood and addressed.

Background Literature

Effects of Literacy

Studies from a range of disciplinary views have pointed to individual and social changes brought about by literacy, from changes in the use of spoken language (Biber & Hared, 1992) to societal changes facilitated by the production of more complex texts resulting in greater creation and accumulation of knowledge (Goody, 2000). More pointedly for our purposes, it has further been argued that literacy functions as a cognitive tool that enables greater abstraction (Goody, 2000, p. 25). More tentatively, it has been suggested that literacy allows for greater metalinguistic awareness whereby linguistic elements may become the object of reflection (Ardila et al., 2010).

There have also been, however, compelling challenges to positions seen as venerating literacy. If there are cognitive advantages to literacy, they are difficult to disentangle from those of schooling and in some populations may also be confounded by experiences of urbanization, economic status, unsettled living conditions (e.g., refugee camps), traumatic experiences (e.g., civil war), and nowadays, digital technologies. In a major study that did seriously attempt

to disentangle these variables, Scribner and Cole (1981) found very little consistent or otherwise convincing evidence of literacy having major effects on cognition, with urbanization and schooling appearing more influential. In a further challenge, Finnegan (2002) argued that the thrust of the "great divide" argument on literacy is "parochial and west-centred" (p. 27), and that in emphasizing what is enabled by print, there has been a strong tendency to overlook the very rich performance and sonic qualities of speech, thereby downplaying its multisensory and highly complex nature (Finnegan, 2002, p. 67).

Despite this, Ardila et al.'s (2010) survey of subsequent psychological and neuropsychological literature concluded that "literacy is strongly reflected in the performance" (p. 707) of tasks, though the authors cautioned that this does not simply mean that literate individuals "have the same abilities… plus something else" (p. 707). Instead, they suggested that one should assume that those without literacy develop "different types of learning, more procedural, pragmatic and sensory oriented" (p. 707). Importantly, there is evidence that some differences in cognitive processing may persist when acquisition of literacy is delayed into adulthood (e.g., Castro-Caldas et al., 2009).

L1 Literacy and L2 Oracy

Though relatively few SLA studies have explored the influence of L1 literacy, the findings to date have suggested the potential for far-reaching implications. One apparently robust finding has been that those with alphabetic literacy appear better able to recognize phonological word boundaries and to more readily conceive of the notion of a word as a linguistic unit (Kurvers et al., 2007). In their seminal work with low-literacy learners, Tarone et al. (2009) suggested that such metalinguistic awareness has a range of effects both in oral language production and in responsiveness to certain teaching practices. For instance, they suggested that lower L1 literacy is associated with the use of less grammatically complex language. This, they speculated, may arise from orienting toward communicative success by way of a simplifying strategy whereby grammatically obligatory yet semantically redundant features of language (e.g., auxiliary verbs) are omitted. Of further interest are data showing participants having difficulty noticing differences between their own brief utterances containing an error and recasts immediately provided by an interlocutor. Overall, the lower learners' alphabetic literacy, the less they tended to notice and recall these output–input differences.

Such findings have important ramifications for both language teaching and theory. As Tarone and Bigelow (2012) pointed out, "current mainstream theories of SLA and formal classroom learning contexts privilege explicit L2

knowledge" (p. 11). This includes commonplace practices such as corrective feedback and rule-based teaching, both of which may be unsuitable for learners with insufficient metalinguistic resources. It may not be surprising that in typical English L2 classrooms, low L1 literacy is associated with "significantly lower gains in English proficiency" (Browder, 2015, p. 187) and that for many such learners overall progress in L2 literacy ranges from "slow to very slow" (Kurvers, 2015, p. 71). There is, therefore, a need to explore both the effectiveness of instructional approaches for these learners, such as through case studies of effective teaching (e.g., King et al., 2018), and to broaden the exploration of how these learners respond to pedagogical variables. A sound starting point would be replicating some of the foundational work in task performance as this literature has obvious implications for both theory and practice.

Tasks and Pretask Planning Effects

Language learning tasks are of central practical and theoretical importance to language teachers and SLA researchers. Skehan (2014a) traced initial interest in tasks to the rise of the communicative language teaching approach that emphasized employing meaningful tasks for L2 use rather than the teaching of discrete language forms. Over time, the nature of tasks became a major object of enquiry, with their value theorized by drawing on the concept of focus on form, whereby learners attend to linguistic form incidentally while in the process of communicating meaning (Long, 1991). The fundamental questions in task-based research relate to how best to design and use tasks in ways that promote language acquisition. From a practical perspective, since task design and use are within the control of teachers, research findings have often been readily applicable to the language classroom.

One of the main theoretical frameworks informing task-based research is the tradeoff hypothesis (Skehan, 2009, 2014b) that takes as a starting point the fact of limited "attentional capacity and working memory" (Skehan, 2009, p. 510). When engaging in a task, language learners are faced with the primary challenge of communicating meaning and need to manage various linguistic resources to do so. If the situation requires carefully attending to one or more elements (e.g., grammatical accuracy, environmental variables), Skehan argued, this will come at the cost of withdrawing a degree of focus from other elements. A crucial issue, therefore, is designing and implementing tasks in ways that allow cognitive resources to be directed toward improving task performance from which learning may arise.

A particularly fruitful area of inquiry has been the manipulation of pretask planning variables and exploring how these influence the performance

measures of accuracy, complexity, and fluency. In their seminal paper. Foster and Skehan (1996) examined performance across three tasks of differing levels of cognitive load: narrative, personal communication, and decision-making, under three pretask planning conditions (no planning, undetailed planning, and detailed planning) and across three performance dimensions (accuracy, complexity, and fluency). The data came from students' performing typical oral classroom communication tasks in dyads. The findings revealed an interaction of task type, planning type, and performance. In particular, planning led to increased fluency and syntactic complexity, and albeit with a statistically less strong effect, to greater accuracy. Follow-up studies produced similar findings, including Skehan and Foster's (2008) comparison of planning versus no planning conditions and a partial replication of their 1996 study (Skehan & Foster, 2005) that further suggested that the impact of pretask planning was restricted to the first 5 min of the task. Summarizing subsequent developments, Skehan (2014a, p. 10) noted strong overall support for the conclusion that planning leads to greater fluency and language complexity with large effect sizes and with smaller and less consistent increases in accuracy. This supported a central prediction of the tradeoff hypothesis that in easing the attentional load associated with planning, learners are able to allocate more of their attentional resources to other aspects of language use.

Of these three dimensions of language use, there has perhaps been most disagreement over the relationship between planning and accuracy. While Ellis (1987) reported increased accuracy in use of the past tense after planning, Crookes (1989) found no significant difference across a broader range of accuracy measures, and this has been mirrored in a number of subsequent studies, often not ruling out the null hypothesis of no planning effect and with effect sizes tending to be very small or negligible (e.g., Bui, 2014; Wang, 2014). However, these studies have presented a number of complications, including apparently differing impacts on accuracy across speech and writing (Ellis & Yuan, 2005; Kormos & Trebits, 2012; but see Vasylets et al., 2017), variations in criteria adopted for measuring accuracy (Foster & Wigglesworth, 2016), and more general variations in how planning was operationalized (for a useful summary of dimensions in planning, see Ellis & Shintani, 2014, p. 142). Also relevant is that at any particular moment in time, learners' present state of interlanguage will tightly constrain their ability to produce certain structures accurately. Thus, while task conditions may influence learners' oral fluency and utterance complexity, they will have virtually no effect on the accuracy of forms that learners have not yet acquired. Those forms most likely to be affected will be those on the cusp of becoming fully acquired, that is, ones that

learners can use correctly under ideal conditions but that tend to be error-prone when the cognitive load is increased. Taken together, then, overall accuracy effects should be expected to be more subtle than are fluency effects.

Following Foster and Skehan's (1996) study, subsequent studies have explored a further range of variables, including types of planning (e.g., rehearsal or strategy planning; Baleghizadeh & Shahri, 2013), planning pretask or within-task (Ellis & Yuan, 2004), amount of planning time (Mehnert, 1998), impacts on speaking versus writing (Vasylets et al., 2017), how students actually plan their work (Ortega, 2005), and teachers' influence through guiding planning (Foster & Skehan, 1999). To date, however, there has been little detailed exploration of variables related to the characteristics of learners themselves and an overwhelming reliance on samples drawn from western, educated, industrialized, rich, developed (WEIRD) populations in research. Among the potentially highly relevant variables on which researchers have not reported is the influence of L1 literacy.

The Present Study

Our study was a partial replication of Foster and Skehan (1996) with the important change of introducing the variable of limited L1 literacy. As we noted above, this variable has not been explored in previous studies. If literacy is shown to impact the effects of planning on the different dimensions of L2 performance, this would direct the design of follow-up studies that allow for a more fine-grained analysis of the specific effect(s) detected with low-literate learners. Conversely, if the findings mirror the directional effects reported elsewhere in the literature, this will more widely expand the generalizability of the tradeoff hypothesis and associated pedagogical applications for classroom-based task design for learners with limited L1 literacy.

For the hypotheses that we adopted in this study, we started from the assumption that, regardless of degree of literacy, all human beings have limited attentional capacity such that when they are faced with competing attentional demands in a speaking task, their performance is affected. The directional effects that we present below are in line with those established in the planning literature on L1-literate participants, except for Hypothesis 4 where we have proposed a different direction for the limited L1 literacy participants in our study.

- Hypothesis 1: As speaking tasks require a L2 user to attend to both language content and language form, pretask planning time allows greater attention to the conceptualization and organization of ideas and to the

L2 formulations required to express them. Accordingly, planned performance is characterized by greater fluency (i.e., more time spent speaking, less silence, and fewer hesitations, repetitions, replacements, and repairs).

- Hypothesis 2: As speaking tasks require a L2 user to attend to both language content and language form in performance, pretask planning time allows greater attention to the conceptualization and organization of ideas and to the L2 formulations required to express them. Accordingly, planned performance is characterized by greater complexity of expression as measured by number of clauses per analysis of speech (AS) unit.
- Hypothesis 3: As speaking tasks require a L2 user to attend to both language content and language form, pretask planning time allows greater attention to the L2 formulations required to express meaning. Accordingly, planned performance is characterized by greater variability in the tense, aspect, modality, and voice of verbs.
- Hypothesis 4: Pretask planning time is not associated with any increase in accuracy. In previous studies using WEIRD participants, the supportive effects of planning time on accuracy were discernible, but small and inconsistent. In this study, because the participants have limited literacy, they will have had little experience of written grammatical rules or the metalanguage that aids receptivity to, conscious reflection on, and memorization of L2 forms (Tarone et al., 2009). This means that the participants will have fewer resources to draw on to increase their L2 accuracy in performance.
- Hypothesis 5: Under the different cognitive demands of the tasks, the effects of pretask planning time is greatest on the task with the heaviest cognitive load (decision-making) and is least on the task with the lightest cognitive load (personal communication).

To allow for a manageable sample size, we did not explore in our study Foster and Skehan's (1996) additional hypothesis that planning time effects would be greatest under the detailed planning condition. Omitting Foster and Skehan's additional hypothesis meant a minor change in research design whereby we replaced the original three-way distinction of nonplanning, planning, and detailed planning by a simpler two-way distinction of planning and nonplanning. In all other respects, the methods of our study mirrored those of the initial study (see Appendix S3 in the Supporting Information online for an itemized table of changes). To ensure the fidelity of this replication, we drew upon the experience and expertise of the first author (Foster) of the initial study as a member of our research team alongside the description of methods

provided in the original publication and a methods-focused chapter reporting on the wider study (Foster, 1996).

Method

Participants

The initial study included 32 participants drawn from four preintermediate-level classes who were placed into four groups (16 participants across two control groups, eight participants in each of the planning groups). Nearly all were women, and Foster (1996, p. 128) described them as being typical of the majority of the part-time language students studying in Britain at the time. As we noted above, the motivated change in our study was replication with learners with limited L1 literacy. We drew our participants from Somali and Afghani former-refugee communities residing in Hamilton and Auckland, New Zealand.

We used G*Power (Faul et al., 2007) to calculate statistical power. Assuming two groups (planners and nonplanners), three dependent variable types (accuracy, complexity and fluency), an α error probability set at .05, and repeated measure ANOVAs with a medium effect size f of 0.25, a participant sample size of 44 generated an actual power of .96, which is very high. This sample size would have given us 22 in each group, a feasible recruitment for the limited L1 literacy population to whom we had access. We nevertheless aimed for 60 participants (30 in each group) to protect against the likelihood of participant attrition across the three data-gathering sessions in order to ensure that no fewer than 44 participants remained for our analyses. We were to use all participants with complete datasets, meaning a final participant number of between 44 and 60.

The participants came from and the data collection took place in the English classes of (a) a tertiary provider of vocational education and (b) a private language school. Through discussions with teachers at the CEFR B1–B2 levels, we identified a list of candidate participants. To avoid creating an obvious distinction within the class (which could have led to negative categorizations of other classmates), it was announced that we were recruiting participants for a study of spoken language performance across three tasks and that anyone was welcome to volunteer. We presented those interested in volunteering with an information form, a participant consent form, and a form for providing basic biographical information, including nationality, L1, and years of schooling in New Zealand. Later, during one-to-one tutorials, we invited learners whom we identified as having potentially low L1 literacy to complete a literacy screening instrument, the Native Language Literacy Screening Tool.

Inclusion and Exclusion Criteria: L1 Literacy

Following Tarone et al.'s (2009) procedure, we used the freely available Native Language Literacy Screening Tool (Florida Department of Education, 2021) to identify candidates who matched the inclusion criteria. The Native Language Literacy Screening Tool is a one-page form written in users' L1, with three sections in which users are prompted for a written response in their L1. Part 1 involved single-word prompts (e.g., name, date, and address); Part 2 provided four short-answer questions (e.g., Where were you born?) of increasing reading complexity; and Part 3 asked the test-takers to write a story about their family. The test administrator did not need to understand the language but monitored the test-takers' behavior, including reading speed, subvocalization while reading, back-tracking, and reading with a pen. We scored test-takers from 1–3, representing low, moderate, and high L1 literacy. For our purposes, we set limited literacy to include candidates with Native Language Literacy Screening Tool scores of 1 (low literacy) or 2 (moderate literacy); moderate literacy is defined according to descriptors such as "showing some difficulty in decoding" and "writes laboriously in native language." The Native Language Literacy Screening Tool is presently available in 27 languages, and we modified it slightly to fit our context (e.g., replacing "United States" with "New Zealand") and translated into further languages as required, for example, we commissioned a version for Dari that we have made available on the IRIS repository (https://www.iris-database.org).

Inclusion and Exclusion Criteria: English Language Level

A further consideration was to ensure that the English language level of the participants was broadly comparable to that of the initial study for whose participants the elicitation materials were designed. Participants in the initial study were studying part time toward the Cambridge First Certificate examination, which suggested that most of the initial study's participants were at a level equivalent to approximately B1 in the Common European Framework of Reference for Languages (Cambridge English, 2015).[2] For the replication, we selected participants on the basis of their oral language rather than on that of their overall language competence. This considered the association between low L1 literacy and substantially lower L2 literacy (e.g., Green & Reder, 1986) and the observation that it is not uncommon in our context for limited L1 literacy learners to have strikingly jagged language profiles[3] (e.g., see Tarone & Bigelow, 2007) that show that they are highly effective in oral communication while they struggle with reading and writing. Since the local practice was to place students in classes based

on their weakest skill, some limited L1 literacy learners who were highly proficient speakers and listeners of English were studying at a Common European Framework of Reference for Languages B1 level or even lower. We identified the initial pool of candidates using the assessment tools used by the education provider that are nationally recognized within the New Zealand Qualifications Authority framework, the New Zealand Certificates in English Language.

Tasks and Materials

Data collection involved three tasks (see Appendix S2 in the Supporting Information online) and two planning conditions that very closely followed the design of Foster and Skehan's (1996) study. We have made the materials available on the IRIS repository (Ryan et al., 2022a). As Foster and Skehan (1996) discussed, they selected the three tasks because the tasks were reasonably familiar classroom/textbook tasks and because they presented "increasingly taxing cognitive load" (p. 306). The three tasks were (a) a personal information exchange task, (b) a narrative task, and (c) a decision-making task. The participants performed each task in dyads. The information exchange task represented familiar information in a familiar context; the narrative task was based on pictures that are loosely connected, allowing "scope for more complex language but also demanding greater cognitive effort" (p. 307); and the decision-making task was the most complex, involving the evaluation of new information and defense of an opinion.

For two of the three tasks from the initial study, the written task instructions for students were available in their complete and original form, having been published by Foster (1996) alongside the planning guidelines. Minor changes were warranted for both the personal and the decision-making tasks and are presented in Appendix S2 in the Supporting Information online. The original personal information task describes a scenario in which a student suddenly remembers that they had forgotten to turn off the oven; since they have an examination, they must now give their friend directions to their house to turn the oven off. Since this is intended to be the most familiar and least cognitively demanding task, it was important to consider whether providing directions to a home would still represent an ordinary activity in a world where GPS is commonplace. After some reflection, we felt that direction-giving remains reasonably common and suitable as a task, but we also felt that the task instructions required a pretext for not turning to GPS. For this, we included an explanation that the speaker's phone battery was flat and that the interlocutor had no mobile data.

The decision-making task also required relatively minor changes. These concerned the scenarios and their appropriacy for use with the expected participants, most of whom were former refugees who had fled a civil war. While we recognized the resilience of many in these communities, it was also the case that some members might still be suffering trauma and so certain topics would need to be avoided. We thus replaced two scenarios: We replaced a scenario involving a bombing with one involving industrial pollution and a scenario involving infidelity and domestic violence with one involving theft.

The third task involved a narrative in which "each member of the dyad had to construct a storyline from a set of five pictures that were loosely but not obviously connected and to relay their ideas to each other" (Foster & Skehan, 1996, p. 307). The set of pictures had been cut from a magazine article on Afghanistan and photocopied, one set per dyad. Unsurprisingly, given the technological limitations of the time, these pictures had existed only in hardcopies and so, unfortunately, had not been preserved. However, Foster, who had selected the initial five pictures, created a similar set for our use. We made considerable effort to ensure that the new set of pictures was a fair match to the original set for obvious thematic connections and obvious lexical demands. It is worth noting that the participants in the original study were likely to have had little direct experience with the Afghanistan context represented in the picture series. Thus, arguably, a replication involving Afghani students required a different set of pictures anyway to more accurately mirror the conditions of the initial study.

The pictures in both series were loosely but not obviously related, meaning that the participant dyads had to work a little to create an emerging story. The present set of pictures shared the following characteristics with those of the initial study:

- Both sets were from National Geographic (1975 and the late 1980s).
- Both contexts were wilderness landscapes and harsh living conditions (Afghanistan vs. Alaska).
- Both showed people who were trying to survive.
- Both showed people who might be the same from picture to picture but might also be different.
- Neither set had an obvious beginning or ending picture.
- The sets had nearly the same number of pictures (6 in the initial study vs. 5 in our study).

For two of the three tasks, then, we used exactly the same tasks and instructions as the initial study. The third task differed only in the content of the pictures, and we expected it to generate similar linguistic performance.

Task Procedures

All members of each class, whether selected as participants and interlocutors or not, completed the tasks under the direction of their usual teacher; data were collected through audio recording from the participating students. As in the original study, the class teacher was also a member of the research team in most cases. This had considerable advantages for ensuring that the tasks and the setup procedures for data collection were familiar to the participants. To ensure consistency, another researcher who had been introduced to the class on at least two occasions prior to data collection accompanied the teacher; this researcher was involved as an assistant at every data collection event.

We would ideally have assigned participants randomly to a control or experimental group, however, to manage a task that is simultaneously carried out in a planned and unplanned form in the same room presented the class teacher with considerable difficulties that might have proved insurmountable. Assignment of the participants to a control or experimental group thus followed class membership, using the same quasi-experimental design as in the original study and for the same logistical reasons. A week prior to each data collection session, we provided the class teachers with a practice activity that closely followed the activity used for the data collection in the procedures and communicative nature of the task. The teachers also received guidelines at least 48 hours before data collection (see Appendix S1 in the Supporting Information online).

In the composition of dyads, the participants and their interlocutors, that is, classmates who were not participants in the study, did not share the same L1 (or regional lingua franca). The participants and their interlocutors were all accustomed to working with one another. To mitigate the negative influence of other variables in task performance (e.g., personal histories, within-group social hierarchies), the students' usual teacher formed dyads in advance, matching the participants with an interlocutor with whom they regularly worked and appeared comfortable in doing so. Thus, the dyads reflected the students' own self-selection practices in forming pairs in the classroom.

The procedures followed the protocols from the initial study:

- There were to be two groups of equal size (22). One group functioned as a control group, performing the tasks with no planning time. The

other group was given 10 min of planning time before completing the task.
- Data collection took place within existing classes and during class time by a teacher/researcher who was known to the students.
- There were three tasks that were completed in class at 1-week intervals. The participants completed all the tasks under the same condition.
- The order of the tasks was: personal information task in Week 1, decision-making task in Week 2, and narrative task in Week 3.

Because of the strong likelihood of the participants' meeting outside of class times and sharing details of the tasks, which would have undermined the comparison of the no-planning condition and the 10-min planning condition, all the participants performed the tasks in the same order, and it was not possible for us to control for any effects for task order.

In the initial study, the participants received written instructions for the tasks and were allowed time to ask questions. We identified a minor change as being appropriate for participants with low L1 literacy; the teacher orally explained the tasks and discussed them as required in addition to providing the written instructions to the participants. We administered the two conditions relating to planning as follows:

- No planning: After the explanation of task, the students began with no preparation time.
- Planning: After explanation of the task, the students received 10 min of planning time with no directions given for how this time should be spent. The students had pens, and we provided them with paper. They were told that the teacher would collect their planning notes and that their notes would not be available during their performance of the task.

We audio recorded each interaction, and for each participant group, we collected task data over 3 weeks, with a 1-week interval between each task. We considered data usable only if the same dyad performed all three tasks.

Data Analysis

Following the procedures of the initial study, we transcribed recordings of the first 5 min of each performance on each task. We then coded the data according to the definitions provided in the initial study (Foster & Skehan, 1996, p. 310). Table 1 provides the categories and definitions used to analyze fluency.

The basic unit of analysis for measuring complexity is the AS-unit. This represented a slight departure from the initial study's use of the C-unit

Table 1 Fluency coding categories

Type	Definition	Measurement
Reformulations	Phrases or clauses repeated with some modification to syntax, morphology, or word order	Ratio of total number of reformulations to total number of clauses
Replacements	Immediate repetitions by way of synonym or other lexical substitute	Ratio of total number of replacements to total number of clauses
False starts	Utterances abandoned before their completion (with or without subsequent reformulation)	Ratio of total number of false starts to total number of clauses
Repetitions	Words, phrases, or clauses repeated without modification	Ratio of total number of repetitions to total number of clauses
Hesitations	Repetition of an initial phoneme or syllable	Ratio of total number number of hesitations to total number of clauses
Pauses	Breaks in the flow of speech; distinction between mid-clause and end-clause pauses (Skehan & Foster, 2005)	Ratio of total number of pauses of 1.0 s or longer to total number of clauses; ratio of end-clause pauses to total number of pauses
Silence total	The sum of timed pauses within the transcript of a minimum of 1.0 s and thereafter rounded to the nearest 0.1 s	Ratio of total elapsed seconds of pause to total length of speaking time

(communication unit) that had, in the previous literature, been rather vaguely defined and lacked exemplification (Foster et al., 2000). Foster and Skehan's (1996) definition and coding protocols aimed at a refinement that took into account the highly elipitcal nature of speech. Foster et al. (2000) later further refined this definition and relabelled it as the AS-unit; it has subsequently been widely adopted in research of this nature. The advantage of using the AS-unit in this replication was the transparency and consistency in analysis enabled by the detailed coding specifications available in Foster et al. (2000) while coding remained closely comparable to the C-unit used in the initial study. Briefly,

Table 2 Syntactic variety

Type	Definition	Measurement
Verb forms	Each recognizable verb form with a recognizable tense, aspect and voice combination, from present/past/future, simple/perfect/continuous, and active/passive	Number of different combinations used at least once
Modal verbs	Can, could, may, might, must, shall, should, will, would	Number of modal verbs appearing at least once
Nonfinite clauses	Clauses based on present and past participles, infinitives	Number of clause types used at least once

an AS-unit consists of "an independent clause or sub-clausal unity, together with any subordinate clause(s) associated with either" (Foster et al., 2000, p. 365). This definition was supported by numerous coding examples provided by Foster et al. We have reported complexity here as the ratio of the number of clauses to AS-units.

We measured syntactic variety as the number of different syntactic forms that are represented in Table 2 divided by the total number of clauses. For instance, a participant producing 50 clauses with five varieties of verb forms (tense–aspect–voice combinations; modal verbs; nonfinite clauses) would score 0.10.

We measured accuracy in two ways: by the number of error-free clauses and by the weighted clause ratio. An error-free clause is one in which there is no error in syntax, morphology, or word order. We counted errors in lexis when the word used was incontrovertibly wrong. In cases of fine decisions of appropriacy, we recorded no error. We scored the data as a ratio of error-free clauses to number of clauses produced. For the weighted clause ratio, following Foster and Wigglesworth's (2016) procedure, we coded clauses according to their degree of accuracy under the categories "entirely accurate" and the three levels of error severity shown in Table 3. We applied a weighting for each of these categories and then calculated totals (e.g., 12 clauses coded as Level 1 = 9.6) and then summed them to produce a raw total (e.g., $[4 \times 1.0] + [12 \times 0.8] + [7 \times 0.5] + [3 \times 0.1] = 17.4$). We then divided this raw total by the total number of clauses to produce the weighted clause ratio (in this example, $17.4 \div 26 = 0.67$).

Table 3 Weighted clause ratio

Type	Definition	Weighting
Accurate	Entirely accurate	1.0
Level 1	Minor errors with little or no effect on meaning	0.8
Level 2	Serious errors (e.g., word order, tense); meaning is discernible	0.5
Level 3	Very serious errors; meaning is only partly discernible	0.1

Coding

At least two researchers coded each variable. To ensure coding reliability, the two coders and a third member of our research team worked together to code the first two transcripts. Through this process, we established a consistent approach to coding, with any potentially tricky cases recorded in notes and resolved by consensus. The two coders then worked independently to separately code a quarter of the transcripts, meeting together and with other members of the research team to discuss unclear or problematic cases. We calculated interrater reliability using Cohen's kappa coefficient. We recoded any coding categories that fell below 80% agreement until we reached agreement of higher than 85%, making changes to the coding protocol as required. We then coded the remaining transcripts.

Statistical Analyses

The design included planning as a between-subjects variable and task complexity as a within-subjects variable. We have provided the following descriptive statistics for all measures: means, 95% confidence intervals around the means, and standard deviations. We assessed the assumption of normal distribution within groups with the Kolmogorov-Smirnov test and the assumption of homogeneity of variance with Levene's test. If these assumptions were met (p values greater than .05), for each dependent variable, we performed mixed within–between ANOVAs using SPSS. We employed partial eta squared including 95% confidence intervals to report effect sizes for multivariate ANOVAs, interpreting benchmarks of .14 as a large effect size, .06 as medium, and .01 as small (Cohen, 1988). Cohen's d was employed to report effect sizes for pairwise comparisons, with 0.80 interpreted as a large effect size, 0.50 as medium, and 0.20 as small (Cohen, 1988). If the assumption of normal distributions was not met, then we used the Kruskal-Wallis test. If the assumption of homogeneity of variance was not met, we used Welch's adjusted F ratio. The main thrust of this replication was to explore how planning time might

affect limited L1 literacy participants differently from the way it affected the WEIRD participants who featured in Foster and Skehan's (1996) study. We largely followed Foster and Skehan's statistical analysis procedures rather than adding others such as linear mixed-effects modelling. If the findings suggested that L1 literacy had an unexpected effect, this would be explored in a future study involving a larger participant pool with distinguishable levels of L1 literacy.

Deviations From the Stage 1 Registered Manuscript

The preregistration for this study is available via the OSF (https://osf.io/kqwmu). We made three deviations from the registered manuscript, each of which was approved by the editors of *Language Learning*. Following the procedures of the initial study, we had proposed to measure silences rounded to the nearest 0.5 s but, since current technology allows for more precision, we decided to report to the nearest 0.1 s. This allowed for greater accuracy and detail in the reported values, potentially revealing more subtle effects.

Our recruitment target was a minimum of 44 participants, with 22 in each group. Although we had initially recruited 49 participants (24 planners, 25 nonplanners), two from the planning group were absent for Task 2. Analysis later revealed that two further participants had not adequately followed the instructions for the decision-making task (one participant merely described the scenario), thereby compromising the key distinction between task types and resultant cognitive load. We therefore proceeded to analyze full datasets for 45 participants, with 21 in the planning group (one fewer than our minimum target) and 24 in the nonplanning group. Due to the 2021 COVID-19 situation and prolonged closure of schools to in-class teaching, recruitment of further participants was not practical, and we considered the numbers acceptable given the consistent findings and normal distributions within the data.

Similarly, we specified in Stage 1 that participants would be aged 18–45 years on the basis of assumptions of the likely student body. Enrollment data subsequently revealed that some participants fell outside this age range—one was one aged 17 years, four were aged 60+ years, and we decided to retain them due to the difficulty in reaching this population. We considered this justifiable for three reasons. First, the mean ages in the two groups were practically identical: planning group ($M = 35.57$ years, $SD = 15.65$) and nonplanning group ($M = 35.79$ years, $SD = 13.27$). Second, as our purpose was to explore the effectiveness of ordinary classroom tasks in ecologically

valid settings, there was a case for including all students who met the L1 literacy inclusion criteria. Third, a substantial body of literature has indicated that negative correlations between L2 grammaticality judgment scores and age of L2 onset are significant between the ages of approximately 15 and 40 years but not significant between 40 and at least 60 years (see especially DeKeyser et al., 2010). We reasoned, therefore, that learners who were in their late 40s and older would not perform tasks differently from learners who were in their mid-40s.

Results

We have made the data available on Harvard Dataverse (Ryan et al., 2022b). We conducted extensive training on how to code the data. Because the AS-unit was the fundamental unit of the study, we peer-checked every coding decision for AS-units and subclausal boundaries at least once and in most cases twice, with a third coder reviewing all discrepancies, which we discussed and resolved.[4] For other variables, posttraining agreement between the first and second coder was very high, especially so for syntactic variety, $\kappa = .972$, $p = < .001$, and fluency, $\kappa = .955$, $p = < .001$, although there were also additional errors of coding omission (3.5% and 5.9%, respectively). The infrequently used fluency category "False start" had the lowest agreement rate (75%), mostly in opposition to "Reformulation" (11/15). Although coding for accuracy (especially weighted clause ratio) involved a more subjective judgment than did coding the other variables, the kappa value was also high, $\kappa = .849$, $p = <.001$, and there were very few omitted codes (0.6%). Subsequently, a third coder rechecked each of the accuracy codes, and we discussed and resolved discrepancies.

We submitted the data for accuracy, complexity, and fluency measures to the Kolmogorov-Smirnov test for normality of distribution, to Mauchly's test of sphericity, and to Levene's test for homogeneity of variance. Unless specifically indicated below, we found no p values less than .05 in these tests (see Appendix S4 in the Supporting Information online for the full details of the results of these statistical tests). In the sections below, we report descriptive statistics and results of the two-way mixed between–within ANOVAs for each dimension of task performance.

Accuracy

We employed two measures of accuracy: error-free clause ratio and weighted clause ratio. As Table 4 shows, for both planned and unplanned conditions the highest error-free clause ratio was in Task 1 (personal information). An advantage held for the planned condition in Task 2 (decision making) and to a lesser

Table 4 Descriptive statistics for grammatical accuracy ratios

	Planned			Unplanned		
Variable	*M*	*SD*	95% CI	*M*	*SD*	95% CI
Error-free clause ratio						
Task 1	.484	.152	[.418, .549]	.485	.147	[.423, .546]
Task 2	.403	.105	[.344, .463]	.348	.157	[.292, .403]
Task 3	.396	.112	[.330, .463]	.393	.180	[.331, .456]
Weighted clause ratio						
Task 1	.826	.081	[.788, .864]	.798	.089	[.763, .833]
Task 2	.744	.064	[.703, .786]	.704	.113	[.665, .742]
Task 3	.781	.059	[.746, .816]	.755	.094	[.722, .788]

Note. Task 1 = Personal information. Task 2 = Decision making. Task 3 = Narrative.

extent in Task 3 (narrative). The two-way mixed ANOVA showed a significant within-participants main effect of task on error-free clause ratio, $F(2, 86) = 17.81$, $p < .001$, $\eta_p^2 = .293$, but no significant interaction of the task and planning condition, $F(2, 86) = 1.32$, $p = .274$, $\eta_p^2 = .030$. Between-participants results showed no significant main effect of planning on accuracy scores, $F(1, 43) = 0.27$, $p = .609$, $\eta_p^2 = .006$. Bonferroni-adjusted pairwise comparisons showed that for the two groups as a whole, Task 1 differed significantly from Task 2, $M_{diff} = .11$, 95% CI [0.06, 0.16], $p < .001$, $d = 0.78$, and from Task 3, $M_{diff} = .09$, 95% CI [0.04, 0.14], $p < .001$, $d = 0.60$, but Task 2 did not differ significantly from Task 3, $M_{diff} = -.02$, 95% CI [−0.06, 0.02], $p = .761$, $d = -0.15$. As displayed in Table 4, planning time was only associated with greater accuracy in Task 2 although the effect did not reach significance. The significant influence on error-free clause ratio was task-type, with the less cognitively demanding Task 1 enabling the participants to produce more error-free clauses than in the more cognitively demanding Tasks 2 and 3.

For the weighted clause ratio (see Table 4), mean scores in the planning condition were higher than those in the nonplanning condition across all tasks, with the difference being smallest in Task 1 and largest in Task 2. For both planners and nonplanners, the highest weighted clause ratio mean score was in Task 1 and the lowest in Task 2. The two-way mixed ANOVA showed a significant within-participants main effect of task on weighted clause ratio scores, $F(2, 86) = 25.00$, $p < .001$, $\eta_p^2 = .368$, but no significant interaction of the task and planning conditions, $F(2, 86) = 0.20$, $p = .817$, $\eta_p^2 = .005$. Between-participants tests showed no significant main effect for planning on weighted

Table 5 Descriptive statistics for syntactic complexity ratios

	Planned			Unplanned		
Variable	*M*	*SD*	95% CI	*M*	*SD*	95% CI
Syntactic complexity						
Task 1	1.279	0.133	[1.219, 1.338]	1.270	0.137	[1.215, 1.326]
Task 2	1.396	0.141	[1.321, 1.472]	1.487	0.194	[1.416, 1.558]
Task 3	1.540	0.291	[1.422, 1.658]	1.525	0.249	[1.414, 1.636]
Syntactic variety						
Task 1	0.100	0.050	[0.081, 0.118]	0.109	0.033	[0.091, 0.125]
Task 2	0.127	0.053	[0.102, 0.153]	0.162	0.062	[0.138, 0.186]
Task 3	0.153	0.076	[0.122, 0.185]	0.178	0.068	[0.149, 0.207]

Note. Task 1 = Personal information. Task 2 = Decision Making. Task 3 = Narrative.

clause ratio scores, $F(1, 43) = 2.13$, $p = .151$, $\eta_p^2 = .047$. Bonferroni-adjusted pairwise comparisons showed that for the two groups as a whole, Task 1 differed significantly from both Task 2, $M_{diff} = 0.09$, 95% CI [0.06, 0.12], $p < .001$, $d = 0.98$, and Task 3, $M_{diff} = 0.04$, 95% CI [0.01, 0.08], $p = .008$, $d = 0.53$, while Task 2 differed significantly from Task 3, $M_{diff} = -0.40$, 95% CI [−0.07, −0.02], $p = < .001$, $d = -0.51$. These results suggested an association of planning time with higher accuracy across tasks, but this effect did not reach significance. The statistically significant effect on weighted clause ratio was task type for the groups as a whole, with the least cognitively demanding Task 1 associated with the most accurate performance, and the most cognitively demanding Task 2 associated with the least accurate performance.

In relation to planning effects, neither accuracy measure produced scores that reached statistical significance although weighted clause ratio perhaps suggests a trend toward greater accuracy in the planning condition. Both measures showed that the significant influence on accuracy in task performance was level of cognitive demand imposed by the task design; the lighter the cognitive burden, the higher the accuracy in performance.

Complexity

We explored the complexity dimension of task performance in two measures: syntactic complexity and syntactic variety. We operationalized syntactic complexity as the number of clauses per AS-unit and syntactic variety as the number of syntactic structures per clause. As Table 5 shows, for clauses per

AS-unit, performance in Task 1 was almost identical for the participants in the two conditions, with a very small advantage in the planned condition for Task 3 but a considerably higher advantage for the unplanned condition in Task 2. A two-way mixed ANOVA showed that there was a significant within-participants main effect of task, $F(2, 86) = 21.73$, $p < .001$, $\eta_p^2 = .336$, but no significant interaction of the task and planning conditions, $F(2, 86) = 1.11$, $p = .336$, $\eta_p^2 = .025$. Between-subject tests showed no significant main effect of planning on complexity scores, $F(1, 43) = 0.34$, $p = .564$, $\eta_p^2 = .008$. Bonferroni-adjusted pairwise comparisons showed that for the two groups as a whole, Task 1 differed significantly both from Task 2, $M_{diff} = -0.17$, 95% CI [−0.25, −0.09], $p < .001$, $d = -1.09$, and from Task 3, $M_{diff} = -0.26$, 95% CI [−0.37, −0.15], $p = .008$, $d = -1.22$, while Task 2 did not differ significantly from Task 3, $M_{diff} = -0.09$, 95% CI [−0.19, 0.01], $p = .094$, $d = -0.39$. The results indicated no significant effect for planning on syntactic complexity, rather it was task design that supported syntactic complexity, with the cognitively more demanding tasks leading to the participants formulating language with more clauses per AS-unit.

For syntactic variety, across all tasks the unplanned condition was associated with a greater number of syntactic structures per clause. As Table 5 shows, this difference was smallest in Task 1 and largest in Task 2. A two-way mixed ANOVA showed a significant within-participants main effect of task on syntactic variety scores, $F(2, 86) = 14.19$, $p < .001$, $\eta_p^2 = .248$, but no significant interaction of the task and planning conditions, $F(2, 86) = 0.61$, $p = .545$, $\eta_p^2 = .014$. Between-subject tests revealed a significant main effect for planning on syntactic variety scores, $F(1, 43) = 4.39$, $p = .042$, $\eta_p^2 = .093$. Bonferroni-adjusted pairwise comparisons showed that for the two groups as a whole, Task 1 differed significantly both from Task 2, $M_{diff} = -0.41$, 95% CI [−0.07, −0.01], $p = .002$, $d = -0.81$, and from Task 3, $M_{diff} = -0.06$, 95% CI [−0.09, −0.03], $p < .001$, $d = -1.07$, while Tasks 2 and 3 did not differ significantly from each other, $M_{diff} = -0.02$, 95% CI [−0.05, 0.01], $p = .244$, $d = -0.31$.

These results showed that the planned condition produced less varied language than did the unplanned condition, especially in the more cognitively demanding Tasks 2 and 3. This result did rise to the level of statistical significance for the three tasks as a whole ($p = .042$), with a medium to large effect size ($\eta_p^2 = .093$). In this regard, while increasing variety in performance was associated with cognitively more demanding tasks, the pretask planning condition reduced the number of syntactic options speakers chose in formulating speech.

Table 6 Descriptive statistics for fluency scores

	Planned			Unplanned		
Variable	*M*	*SD*	95% CI	*M*	*SD*	95% CI
Repair fluency						
Task 1	0.551	0.249	[0.448, 0.655]	0.476	0.222	[0.379, 0.572]
Task 2	0.552	0.287	[0.427, 0.677]	0.591	0.283	[0.474, 0.708]
Task 3	0.517	0.242	[0.406, 0.629]	0.512	0.262	[0.408, 0.616]
Breakdown fluency						
Task 1	0.786	0.602	[0.550, 1.023]	0.723	0.474	[0.501, 0.944]
Task 2	1.147	0.719	[0.835, 1.460]	1.078	0.703	[0.785, 1.370]
Task 3	1.069	0.600	[0.822, 1.316]	0.988	0.526	[0.757, 1.219]
Total silence						
Task 1	0.136	0.107	[0.096, 0.176]	0.086	0.075	[0.049, 0.123]
Task 2	0.205	0.103	[0.150, 0.261]	0.188	0.144	[0.135, 0.240]
Task 3	0.247	0.132	[0.193, 0.300]	0.183	0.110	[0.133, 0.233]
Pause positioning						
Task 1	0.577	0.223	[0.479, 0.676]	0.500	0.218	[0.406, 0.595]
Task 2	0.502	0.131	[0.439, 0.564]	0.514	0.151	[0.454, 0.574]
Task 3	0.610	0.156	[0.541, 0.680]	0.619	0.156	[0.561, 0.694]

Note. Task 1 = Personal information. Task 2 = Decision making. Task 3 = Narrative.

Fluency

We employed a range of variables to measure repair fluency (repetitions, reformulations, replacements, and false-starts) and to measure breakdown fluency (pausing and hesitations). We calculated ratio of silence to speech for each task performance as well as position of pausing relative to clause boundaries. We expressed repair fluency as the number of repairs per clause. Table 6 shows that the participants' mean scores for repair fluency were similar across all tasks in both the planned and unplanned conditions. The difference was greatest in Task 1, smaller in Task 2, and almost nonexistent for Task 3. There was no discernible pattern in either the planned or the unplanned condition. A two-way mixed ANOVA showed no within-participants main effect for task, $F(2, 86) = 1.50$, $p = .228$, $\eta_p^2 = .034$, and no significant interaction of the task condition with the planning condition, $F(2, 86) = 1.15$, $p = .322$, $\eta_p^2 = .026$. Between-participants tests showed no significant main effect for planning, $F(1, 43) = 0.05$, $p = .828$, $\eta_p^2 = .001$. These results showed that neither the cognitive demands of the task nor the planning condition was related to the incidence of repairs in the data.

We expressed breakdown fluency as number of pauses and hesitations per clause. Mean scores across all tasks (see Table 6) showed that the planned condition was associated with a higher incidence of breakdowns than was the unplanned condition. The difference was small in Task 1, larger for Task 2, and largest for Task 3. A two-way mixed ANOVA showed a significant within-participants main effect for task, $F(2, 86) = 13.77$, $p < .001$, $\eta_p^2 = .242$, but no significant interaction of the task and planning conditions, $F(2, 86) = 0.01$, $p = .992$, $\eta_p^2 < .001$. Between-participants tests showed no main effect for planning, $F(1, 43) = 0.20$, $p = .661$, $\eta_p^2 = .005$. Bonferroni-adjusted pairwise comparisons showed that for the two groups as a whole for breakdowns, Task 1 differed significantly from Task 2, $M_{diff} = -0.36$, 95% CI [−0.56, −0.16], $p < .001$, $d = -0.57$, and from Task 3, $M_{diff} = 0.27$, 95% CI [−0.41, −0.13], $p < .001$, $d = -0.50$, but Task 2 did not differ significantly from Task 3, $M_{diff} = 0.08$, 95% CI [−0.10, 0.27], $p = .803$, $d = 0.13$. These results showed that breakdown fluency was not significantly affected by the planning condition, but in contrast to repair fluency, there was more frequent pausing and hesitation in the cognitively more demanding Tasks 2 and 3.

We expressed the amount of silence in a task performance as the sum of pauses in proportion to time spent on the task. Table 6 shows that for all tasks, the planning condition was associated with a greater proportion of silence in a task. This difference was smallest in Task 2, larger in Task 1, and largest in Task 3. Levene's test showed a p value of .043 for mean ratio of silence in Task 1; a subsequent Kruskal-Wallis test was not significant, $H(1) = 2.69$, $p = .101$. A two-way mixed ANOVA showed a significant within-participants main effect for task, $F(2, 86) = 22.81$, $p < .001$, $\eta_p^2 = .347$, but no significant interaction of the task condition with the planning condition, $F(2, 86) = 1.03$, $p = .361$, $\eta_p^2 = .023$. Between-participants tests showed no significant main effect for planning, $F(1, 43) = 2.41$, $p = .128$, $\eta_p^2 = .053$. Bonferroni-adjusted pairwise comparisons indicated that for the two groups as a whole, Task 1 differed significantly both from Task 2, $M_{diff} = -0.09$, 95% CI [−0.13, −0.04], $p < .001$, $d = -0.78$, and from Task 3, $M_{diff} = -0.10$, 95% CI [−0.14, −0.07], $p < .001$, $d = -0.94$, but Task 2 did not differ significantly from Task 3, $M_{diff} = -0.02$, 95% CI [−0.06, 0.03], $p = .927$, $d = -0.13$. The results indicated that the participants in the planned condition tended to produce proportionally more silence in their task performance than did those in the unplanned condition, though this did not reach statistical significance. Again, the significant effect on performance was task type, with the cognitively more demanding Tasks 2 and 3 associated with greater proportion of silence for both groups.

We conducted one final fluency analysis targeting the position of pausing, either at midclause or at clause boundaries. We calculated this by expressing clause-boundary pauses as a proportion of total pauses. A pause at a clause boundary indicated more fluent language processing than did a pause which interrupted a clause. Mean scores not rising or falling much above 0.50 would have suggested an even distribution of pausing midclause and pausing at boundaries. As Table 6 shows, we found this outcome in Task 2 for both the planned and unplanned condition. In Task 1, the distribution was relatively even in the unplanned condition, but there were more pauses at clause boundaries in the planned condition. Task 3 showed no advantage for planning time, with pauses more likely to occur at clause boundaries in both conditions. Mauchly's test indicated that the assumption of sphericity had been violated, $\chi^2 = 10.72$, $p = .005$, therefore in the subsequent two-way mixed ANOVA, we corrected within-participants degrees of freedom using Greenhouse-Geisser estimates. The results showed a significant within-participants main effect for task on proportion of pausing at clause boundaries, $F(1.63, 84) = 5.49$, $p = .009$, $\eta_p^2 = .117$, but no significant interaction of the task condition with the planning condition, $F(1.63, 84) = 0.53$, $p = .554$, $\eta_p^2 = .010$. Between-participants tests showed no main effect for planning, $F(1, 42) = 0.264$, $p = .610$, $\eta_p^2 = .006$. Bonferroni-corrected pairwise comparisons showed that for the two groups as a whole, Task 1 did not differ significantly from Task 2, $M_{diff} = 0.03$, 95% CI [−0.04, 0.11], $p = .909$, $d = 0.15$, or from Task 3, $M_{diff} = -0.08$, 95% CI [−0.19, 0.03], $p = .191$, $d = -0.43$, but Task 2 differed significantly from Task 3, $M_{diff} = -0.11$, 95% CI [0.04, 0.18], $p = .001$, $d = -0.75$. These results indicated that the planning condition did not have an effect on the position of pausing. For Tasks 2 and 3 the planning and nonplanning condition produced almost identical results, and although Task 1 results showed a trend for the planning condition to favor pausing at clause boundaries, this was not statistically significant. By contrast, task type was a significant variable, with pausing patterns in Task 3 significantly different from the other two tasks. Compared to their speech production in the other tasks, in Task 3, both planners and nonplanners were significantly less likely to interrupt a syntactic structure with a pause.

Impact of Different Cognitive Demands

The tasks were designed to present the participants with different levels of cognitive demands, such that Task 1 < Task 3 < Task 2 (where < indicates increasing task demands), with the prediction that effects of planning would be least discernible for Task 1 and most for Task 2. The two-way mixed

between–within ANOVAs reported above consistently showed no statistically significant interaction of the task and planning conditions. For mean scores, the divergence between the planned and unplanned performances was widest for Task 2 for five of the eight measures: accuracy (error-free clause ratio and weighted clause ratio), syntactic complexity, syntactic variety, and breakdown fluency. This pattern of results suggested that, for some dimensions of performance, the greatest scope for a planning effect was on the task with the heaviest cognitive burden, but this was without reaching statistical significance and not necessarily in the direction of improved complexity, accuracy, or fluency.

Discussion

Effects of Planning Time and Task

We can now relate the various analyses of performance measures to the study's hypotheses (see Table 7 for a summary). In contrast to the results obtained by Foster and Skehan (1996), pretask planning time in this study had very little statistically significant effect on measures of oral task performance. There was only one significant outcome—for syntactic variety—but this was in the opposite direction to Hypothesis 3; limited L1 literacy learners in the planned condition drew on a narrower range of syntactic structures than did the learners in the unplanned condition when performing the tasks. It was as if the planning condition had led them to be more conservative in how they expressed themselves. In the opposite direction to Hypothesis 1, though not reaching significance, limited L1 literacy learners in the planned condition tended to task performances that were less fluent than those of learners in the unplanned condition, that is, with more breakdowns and a higher proportion of silence, suggesting that planning time inhibited rather than supported performance. In terms of syntax, the planning condition did not lead to any statistically significant advantage in complexity, but there was a sizeable difference in the mean scores for Task 2, in the opposite direction to Hypothesis 2; limited L1 literacy learners in the planned condition formulated fewer complex AS-units than their counterparts in the unplanned condition. Taken together, these outcomes show the effects of planning time to be largely without statistical significance but tending toward making task performance more hesitant and less ambitious than the effects of not planning.

In the same vein, the analyses of accuracy showed no statistically significant effect for planning. This was in line with the prediction of Hypothesis 4, which argued that as limited L1 literacy learners have little experience of metalanguage or of conscious reflection on grammar, they would not be able to draw on these during planning time. Yet, the means of weighted

Table 7 Support for the study hypotheses

Hypothesis	Upheld	Detail
H1 Repair fluency	No	Across all tasks, there was no difference in repair fluency between the planning and nonplanning condition.
H1 Breakdown fluency	No	Means scores for breakdown fluency were in the opposite direction to H1, but this did not reach significance.
H1 Proportion of silence	No	Mean scores for proportion of silence were in the opposite direction to H1, but this did not reach significance.
H2 Complexity of expression (syntactic complexity)	No	There was no statistically significant effect for planning on syntactic complexity.
H3 Variability of expression (syntactic variation)	No	Mean scores for syntactic variation were in the opposite direction to H3. This was statistically significant, $p = .045$, $\eta_p^2 = .093$.
H4 Accuracy	Yes	There was no significant effect on planning on either accuracy measure. Weighted clause ratio means across all tasks were in the opposite direction to H4, but this did not reach significance.
H5 Planning effects and cognitive complexity of task	No	There was no significant interaction between planning and task type for any of the performance variables investigated by H1–H4. There was a highly significant effect for task type, with performances in the cognitively least demanding (Task 1) tending to be the most accurate, least complex, least syntactically varied and most fluent.

clause ratio across the three tasks were all higher in the planned condition. This outcome might indicate that limited L1 literacy learners do have L2 grammatical knowledge that, owing little or nothing to metalanguage or the conscious rehearsal of rules, is recruited if planning time provides extra attentional space and the speaker opts for less complex, less varied, and less fluent performance.

For task effects, Hypothesis 5 predicted that pretask planning time would impact most on Task 2, the task with the heaviest cognitive load and least on

Task 1, the task with the lightest cognitive load. There was no statistical support for an interaction of the task and planning conditions, although as we noted above, the difference between the planned and unplanned scores for five of the eight performance measures was widest in the most cognitively demanding Task 2. It is striking, however, that pairwise comparisons showed that performance on Task 1 was significantly different from performance on Tasks 2 and 3 for error-free clauses, weighted clause ratio, syntactic complexity, syntactic variety, breakdown fluency, proportion of silence (all p values $< .001$). There was no significant difference between Tasks 2 and 3 for any of these measures apart from weighted clause ratio ($p = .008$) and clause boundary pausing ($p < .001$). For repair fluency alone, there was no discernible difference between the task performances. This pattern of results indicated that for limited L1 literacy participants, regardless of whether they had time or not to plan, their oral task performance was influenced by the familiarity of the information about which they were asked to talk and the amount of cognitive work that they had to do when processing information. For Task 1, with the most familiar information and the lightest processing demands, the participants produced language that was generally more accurate, less varied, and less complex than the language that they produced for the unfamiliar information and heaviest processing demands of Tasks 2 and 3.

In Foster and Skehan's (1996) study, the impact of task type on performance was viewed as the result of information-processing pressure, with speakers needing to allocate their limited attention between the competing demands of L2 form and L2 content. This would be most onerous for the most cognitively taxing tasks and most alleviated by pretask planning time. Indeed, the pattern of results obtained by Foster and Skehan demonstrated that their participants in the planning condition were able to combine accuracy with complexity in the most cognitively demanding decision task. In the absence of significant planning effects in our study, the highly significant task effects show that task design by itself is an influence on oral performance, with the lightest cognitive burden of Task 1 suggesting a possible tradeoff between accuracy on the one hand and fluency and complexity on the other. One final observation about syntactic complexity is that all our limited L1 literacy participants were able to formulate syntactically more complex language in response to the more complex demands of Tasks 2 and 3. Both planners and nonplanners increased their mean number of clauses per AS-unit when prompted by the demands of the task to make a narrative or explain a judgement.

Role of L1 Literacy

The key independent variable for the participants in this study was limited or nonexistent literacy in their L1, contrasting with the Foster and Skehan's (1996) participants who, having had the benefit of a complete primary and secondary education, were literate in both their L1(s) and English L2. The question of whether we can ascribe the absence of planning effects on our limited L1 literacy participants' oral L2 performance to their lack of experience of learning through reading and writing is not however straightforward. The profiles of the two groups of participants inevitably differed along many dimensions, it being very difficult (if not impossible) to separate the educational experience of our limited L1 literacy language learners from a host of other variables that might have influenced their individual language learning goals, motivation, and strategies.

The participants in the original study were aged 18–30 years, all temporary visitors to the UK whose common goal was to leave with a globally recognized English language qualification. To that end, they enrolled in English as a foreign language classes directed at the Cambridge First Certificate exam (Cambridge English, 2015). All needed to use English in their daily lives, in low-stakes interactions such as shopping and casual conversations rather than in high-stakes interactions with UK government agencies or local authority education and social services. Some worked part time as family au pairs, but most did not have or need paid employment. By contrast, the limited L1 literacy participants in our study were older ($M = 35.69$ years, $SD = 14.26$) and more culturally heterogenous. Settled in New Zealand, they had no prospect or plans to return permanently to their countries of origin, having fled from war or situations of violent conflict. Many had experienced great trauma and extended periods in refugee camps. Facing the challenges of integrating into New Zealand society, they enrolled in classes of English as another language because they needed to operate not just in socializing or managing everyday conversations but in high-stakes interactions around housing, health, employment, children's education, immigration status, tax, and other social obligations. With English the essential tool for their present lives and future prosperity, the limited L1 literacy learners needed English for essential communication rather than for a certificated qualification.

Faced with the personal urgency of building a new life in a new language, a reasonable learning strategy would be not to attend to grammatical elements, such as auxiliary verbs, plural morphemes, and third person *–s* that are often semantically redundant and/or phonologically indistinct. This would not necessarily imply an inability to attend to such features arising from a nonliterate L1

background and associated lack of metalinguistic knowledge. It could imply rather a choice not to expend precious time and effort on elements that deliver little in terms of added communicative effectiveness. This could account for why, in the unplanned condition, limited L1 literacy participants appeared to prioritize fluency over accuracy. This learner predilection is not uncommon; it has been manifested by early fossilizers such as Wes in Schmidt's (1983) case study. Further, anecdotal evidence from those in our research team who oversaw the task performances described the learners in the planning condition as visibly frustrated by the 10 min that they had to devote to planning the task; they were eager to get going on performing the task.

Another possibility is suggested in a line of research indicating that literacy facilitates a process of isolating linguistic components from the stream of speech, making phonemes, words, and morphemes more readily available as objects for inspection, and thereby supporting the development of metalinguistic awareness (e.g., Kurvers et al., 2007; Tarone & Bigelow, 2012; Tarone et al., 2009). During planning time, it could be that a more developed metalinguistic awareness promotes attending to certain aspects of the message such as applying explicit knowledge of grammatical rules to enhance accuracy and linguistic variety.

The nonsignificant impact of planning could also have been related to differences in the planning strategies used by the limited L1 literacy participants compared to those in the initial study. Relevant here may be that Foster and Skehan's (1996) participants had extensive schooling; this almost certainly involved training in literacy-related planning strategies that may be transferable to planned oral production. Specifically, prewriting strategies such as identifying and organizing content are deeply instilled during primary school and beyond (Torrance et al., 2007), and related strategies are reported by literate students in successful oral task planning (Pang & Skehan, 2014). Such strategies are not necessarily instinctive, with substantial evidence indicating that students of all ages benefit from additional training (e.g., Torrance et al., 2007) and that their planning behavior varies considerably in quality and quantity (Flower & Hayes, 1980). It could be, therefore, that the highly L1 literate participants in Foster and Skehan's (1996) study successfully transferred deeply ingrained composition planning strategies to oral task performance in their English as a foreign language classes. In contrast, as we noted above, many of the limited L1 literacy participants appeared to do little planning; they simply wanted to "get on with it."

Limitations and Future Research Directions

The data in this replication showed little to no significant effect for pretask planning on a host of performance measures. We have discerned, however, certain trends in the data for the planned condition to produce higher accuracy, less variety, and more breakdown fluency than the unplanned condition that merit further investigation. Future studies should adopt a stricter operationalization of the nonplanning condition. In Foster and Skehan's (1996) study, the participants in the nonplanning condition started the task immediately after very limited oral instructions, with no details of the content of the task until they turned over the worksheet. In the replication, to avoid the limited L1 literacy participants' struggling with written instructions, we provided an oral introduction to the task, including their seeing the narrative pictures and learning details of the criminal cases. Thus, the nonplanning condition contained an element of planning absent in the original study. Mehnert (1998) showed that even 1 min of planning time produced performances that were more fluent and more accurate compared to no planning time. While it is difficult to imagine a totally unplanned condition for limited literacy learners, as tasks have to be understood before being performed, any future study should maximize the contrast between the planned and unplanned conditions by limiting the amount of pretask content information. In addition, while recruitment may prove challenging, future studies would ideally aim for more uniform cohorts to avoid any potential confound with variables such as L1 influence, language aptitude, and multilingualism. Future studies should also investigate the learning motivations of limited L1 literacy learners in order to explore any interaction of motivation with planned performance and also the planning strategies used by limited L1 literacy learners, perhaps partially replicating Pang and Skehan's (2014) study.

Following the lead of Tarone et al.'s (2009) study, future studies could also include screening for L2 literacy; in some cases, the L2 literacy acquired while living in New Zealand may have influenced task performance. Future studies could also tease apart the effects of literacy and strategy adoption through comparing the performance of learners with low literacy in L1 and L2 with those with midlevel literacy in at least one language. To confirm the link between planned performance and literacy, it will also be important to control for other ways in which limited L1 literacy participants tend to diverge from WEIRD populations, the most relevant of which is the entanglement of literacy with schooling. Finally, a future study could investigate how limited L1 literacy language learners perform tasks in their L1, with or without planning time. Even

with little to no acquaintance with their L1 in written form, it is not expected this would inhibit their ability to conceive and articulate complex, varied, or fluent language, but it would give a baseline comparison against which to set various dimensions of their L2 performance.

Conclusion

In this partial replication of Foster and Skehan (1996), we worked with adult L2 learners who had not developed literacy in childhood, exploring how the variable of planning time and its interaction with task complexity influenced oral performance across measures of accuracy, complexity, and fluency. Our hypotheses reflected the assumption that the results for these learners would mirror those of the highly literate participants of the initial study, but this proved not to be the case: The provision of 10 minutes planning time was not associated with improved oral performance, and for some measures the results trended in the opposite direction. Why this should be so remains unclear, though possibilities for further investigation include the elicitation procedures, metalinguistic knowledge, social factors, and experience of schooling. The motivation for this study was the desire for a more inclusive SLA, in which theoretical models will be more broadly reflective of the diversity of human populations and learning environments. As the other articles in this special issue demonstrate, replicating pivotal SLA studies can be a highly effective way of exploring the special characteristics of understudied populations and confirming or disconfirming the generalizability of existing findings and theoretical models. We echo the call of Tarone et al. (2009) for a great deal more investigation of the language production, pragmatic strategies, and processing skills of adult L2 learners with low L1 literacy. Such work is not only of substantial theoretical importance within SLA but may also lead to the development of pedagogical approaches that minimize the early fossilization reported for many such learners (Field & Ryan, in press).

Final revised version accepted 5 November 2022

Open Research Badges

This article has earned Open Data and Open Materials badges for making publicly available the digitally-shareable data and the components of the research methods needed to reproduce the reported procedure and results. The article has also earned a Preregistered Research Designs badge for having a

preregistered research design. All data, materials and the design that the authors have used and have the right to share are available at https://doi.org/10.7910/DVN/TAFRCO, http://www.iris-database.org and https://osf.io/kqwmu. All proprietary materials have been precisely identified in the manuscript.

Notes

1 Due to limitations and inconsistencies in data collection and reporting, this report indicated that the true figure was likely to be much higher.
2 Common European Framework of Reference level B1 equates to an intermediate level. Speakers are able to understand the L2 when it concerns familiar topics, can deal with most common situations that arise when visiting places where the L2 is spoken, can produce simple text on familiar or personal topics, and can describe experiences and events and briefly give explanations for opinions and plans.
3 It could be argued that jagged profiles might also have occurred in the initial study, and certainly L2 users typically do have varied strengths and weaknesses. However, it has also been generally assumed to be the case that a (WEIRD) learner's level of proficiency in one skill area (e.g., writing) will not typically diverge dramatically from another. Indeed, the identification of jagged profiles in language proficiency examinations has been taken to be worthy of further investigation as it may be an indication of a possible clerical mistake (e.g., IELTS, 2021).
4 Establishing unit boundaries does not lend itself to Cohen's kappa due to the vast number of coding possibilities.

References

Ardila, A., Bertolucci, P. H., Braga, L. W., Castro-Caldas, A., Judd, T., Kosmidis, M. H., Matute, E., Nitrini, R., Ostrosky-Solis, F., & Rosselli, M. (2010). Illiteracy: The neuropsychology of cognition without reading. *Archives of Clinical Neuropsychology*, *25*(8), 689–712. https://doi.org/10.1093/arclin/acq079

Baleghizadeh, S., & Shahri, M. N. N. (2013). The effect of online planning, strategic planning and rehearsal across two proficiency levels. *The Language Learning Journal*, *45*(2), 171–184. https://doi.org/10.1080/09571736.2013.808258

Biber, D., & Hared, M. (1992). Literacy in Somali: Linguistic consequences. *Annual Review of Applied Linguistics*, *12*, 260–282. https://doi.org/10.1017/S0267190500002269

Browder, C. T. (2015). The educational outcomes of U.S. high school English-learner students with limited or interrupted formal education. In M. G. Santos & A. Whiteside (Eds.), *Low educated second language and literacy acquisition: Proceedings of the Ninth Symposium* (pp. 172–198). Lulu Publishing.

Bui, H. Y. G. (2014). Task readiness: Theoretical framework and empirical evidence from topic familiarity, strategic planning, and proficiency levels. In P. Skehan (Ed.), *Processing perspectives on task performance* (pp. 63–93). John Benjamins.

Cambridge English (2015). *The Cambridge English scale explained* [Brochure]. Retrieved from https://www.cambridgeenglish.org/images/177867-the-methodology-behind-the-cambridge-english-scale.pdf

Castro-Caldas, A., Nunes, M. V., Maestu, F., Ortiz, T., Simoes, R., Fernandes, R., de la Guia, E., Garcia, E., & Goncalves, M. (2009). Learning orthography in adulthood: A magnetoencephalographic study. *Journal of Neuropsychology*, *3*(1), 17–30. https://doi.org/10.1348/174866408x289953

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum.

Crookes, G. (1989). Planning and interlanguage variation. *Studies in Second Language Acquisition*, *11*(4), 367–383. https://doi.org/10.1017/S0272263100008391

DeKeyser, R., Alfi-Shabtay, I., & Ravid, D. (2010). Cross-linguistic evidence for the nature of age effects in second language acquisition. *Applied Psycholinguistics*, *31*(3), 413–438. https://doi.org/10.1017/S0142716410000056

Ellis, R. (1987). Interlanguage variability in narrative discourse: Style shifting in the use of the past tense. *Studies in Second Language Acquisition*, *9*(1), 1–19. https://doi.org/10.1017/S0272263100006483

Ellis, R., & Shintani, N. (2014). *Exploring language pedagogy through second language acquisition research*. Routledge.

Ellis, R., & Yuan, F. (2004). The effects of planning on fluency, complexity, and accuracy in second language narrative writing. *Studies in Second Language Acquisition*, *26*(1), 59–84. https://doi.org/10.1017/S0272263104026130

Ellis, R., & Yuan, F. (2005). The effects of careful within-task planning on oral and written task performance. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 167–192). John Benjamins.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175–191. https://doi.org/10.3758/BF03193146

Field, J., & Ryan, J. (in press). The influence of prior schooling on second language learning: A longitudinal study with former refugees. *New Zealand Studies in Applied Linguistics*.

Finnegan, R. (2002). *Communicating: The multiple modes of human interconnection*. Routledge.

Florida Department of Education (2021). *Native Language Screening Tool in 28 languages*. http://www.fldoe.org/academics/career-adult-edu/adult-edu/native-language-literacy-screening-too.stml

Flower, L., & Hayes, J. R. (1980). The cognition of discovery: Defining a rhetorical problem. *College Composition and Communication*, *31*(1), 21–32. https://doi.org/10.2307/356630

Foster, P. (1996). Doing the task better: How planning time influences students' performance. In J. Willis & D. Wills (Eds.), *Challenge and change in language teaching* (pp. 126–135). Heinemann.

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

Foster, P., & Skehan, P. (1999). The influence of source of planning and focus of planning on task-based performance. *Language Teaching Research*, *3*(3), 215–247. https://doi.org/10.1177/136216889900300303

Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics*, *21*(3), 354–375. https://doi.org/10.1093/applin/21.3.354

Foster, P., & Wigglesworth, G. (2016). Capturing accuracy in second language performance: The case for a weighted clause ratio. *Annual Review of Applied Linguistics*, *36*, 98–116. https://doi.org/10.1017/S0267190515000082

Goody, J. (2000). *The power of the written tradition*. Smithsonian Institution Press.

Green, K. R., & Reder, S. (1986). Factors in individual acquistion of English: A longitudinal study of Hmong adults. In G. L. Hendricks, B. T. Downing, & A. S. Deinard (Eds.), *The Hmong in transition* (pp. 299–328). The Centre for Migration Studies.

Havron, N., & Arnon, I. (2017). Reading between the words: The effect of literacy on second language lexical segmentation. *Applied Psycholinguistics*, *38*(1), 127–153. https://doi.org/10.1017/S0142716416000138

IELTS (2021). *Test performance 2021*. https://www.ielts.org/for-researchers/test-statistics/test-performance

King, K. A., Bigelow, M., & Hirsi, A. (2018). New to school and new to print: Everyday peer interaction among adolescent high school newcomers. *International Multilingual Research Journal*, *11*(3), 137–151. https://doi.org/10.1080/19313152.2017.1328958

Kormos, J., & Trebits, A. (2012). The role of task complexity, modality, and aptitude in narrative task performance. *Language Learning*, *62*(2), 439–472. https://doi.org/10.1111/j.1467-9922.2012.00695.x

Kurvers, J. (2015). Emerging literacy in adult second-language learners: A synthesis of research findings in the Netherlands. *Writing Systems Research*, *7*(1), 58–78. https://doi.org/10.1080/17586801.2014.943149

Kurvers, J., van Hout, R., & Vallen, T. (2007). Literacy and word boundaries. In N. Faux (Ed.), *Low-educated second language and literacy acquisition: Research, policy, and practice* (pp. 45–64). *The Literacy Institute*.

Long, M. H. (1991). Focus on form: A design feature in language teaching methodology. In K. de Bot, R. Ginsberg, & C. Kramsch (Eds.), *Foreign language research in cross-cultural perspective* (pp. 39–52). John Benjamins.

Mehnert, U. (1998). The effects of different lengths of time for planning on second language performance. *Studies in Second Language Acquisition*, *20*(1), 83–108. https://doi.org/10.1017/S0272263198001041

Ortega, L. (2005). What do learners plan? Learner-driven attention to form during pre-task planning. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 77–109). John Benjamins.

Pang, F., & Skehan, P. (2014). Self-reported planning behaviour and second language performance in narrative retelling. In P. Skehan (Ed.), *Processing perspectives on task performance* (pp. 95–127). John Benjamins.

Ryan, J., Foster, P., Fester, A., Wang, Y., Field, J., Kearney, C., & Yap, J. (2022a). *Elicitation tasks. Materials from "First language literacy and second language oracy: A partial replication of Foster and Skehan (1996)"* [Language test]. IRIS Database, University of York, UK. https://doi.org/10.48316/e7zp-ez27

Ryan, J., Foster, P., Fester, A., Wang, Y., Field, J., Kearney, C., & Yap, J. R. (2022b). *Data from "First language literacy and second language oracy: A partial replication of Foster and Skehan (1996)*" (Version 1) [Dataset]. Harvard Dataverse. https://doi.org/10.7910/DVN/TAFRCO

Schmidt, R. W. (1983). Interaction, acculturation, and the acquisition of communicative competence: A case study of an adult. In N. Wolfson & E. Judd (Eds.), *Sociolinguistics and language acquisition* (pp. 137–174). Newbury House.

Scribner, S., & Cole, M. (1981). *The psychology of literacy*. Harvard University Press.

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics*, *30*(4), 510–532. https://doi.org/10.1093/applin/amp047

Skehan, P. (2014a). The context for researching a processing perspective on task performance. In P. Skehan (Ed.), *Processing perspectives on task performance* (pp. 1–26). John Benjamins.

Skehan, P. (2014b). Limited attentional capacity, second language: Performance, and task-based pedagogy. In P. Skehan (Ed.), *Processing perspectives on task performance* (pp. 211–260). John Benjamins.

Skehan, P., & Foster, P. (2005). Strategic and on-line planning: The influence of surprise information and task time on second language performance. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 193–218). John Benjamins.

Skehan, P., & Foster, P. (2008). Complexity, accuracy, fluency and lexis in task-based performance: A meta-analysis of the Ealing research. In S. V. Daele, A. Housen, F. Kuiken, M. Pierrard, & I. Vedder (Eds.), *Complexity, accuracy, and fluency in second language use, learning, and teaching* (pp. 207–226). University of Brussels Press.

Tarone, E. (2014). Enduring questions from the interlanguage hypothesis. In Z. Han & E. Tarone (Eds.), *Interlanguage: Forty years later* (pp. 7–26). John Benjamins.

Tarone, E., & Bigelow, M. (2007). Alphabetic print literacy and oral language processing in SLA. In A. Mackey (Ed.), *Conversational interaction in second language acquisition: A series of empirical studies* (pp. 101–121). Oxford Univesity Press.

Tarone, E., & Bigelow, M. (2012). A research agenda for second language acquisition of pre-literate and low-literate adult and adolescent learners. In P. Vinogradov & M. Bigelow (Eds.), *Low educated second language and literacy acquisition: Proceedings of the 7th Symposium* (pp. 5–26). University of Minnesota.

Tarone, E., Bigelow, M., & Hansen, K. (2009). *Literacy and second language oracy*. Oxford University Press.

Torrance, M., Fidalgo, R., & García, J.-N. (2007). The teachability and effectiveness of cognitive self-regulation in sixth-grade writers. *Learning and Instruction*, *17*(3), 265–285. https://doi.org/10.1016/j.learninstruc.2007.02.003

Education for All Global Monitoring Report Team (2015). *Education for all 2000–2015: Achievements and challenges*. UNESCO. https://en.unesco.org/gem-report/report/2015/education-all-2000-2015-achievements-and-challenges

Vasylets, O., Gilabert, R., & Manchon, R. M. (2017). The effects of mode and task complexity on second language production. *Language Learning*, *67*(2), 394–430. https://doi.org/10.1111/lang.12228

Wang, Z. (2014). On-line time pressure manipulations: L2 speaking performance under five types of planning and repetition conditions. In P. Skehan (Ed.), *Processing perspectives on task performance* (pp. 27–61). John Benjamins.

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Accessible Summary (Chinese)
Accessible Summary (English)
Appendix S1. Guidelines for Data Collection.
Appendix S2. Tasks.
Appendix S3. Table of Itemized Changes to the Original Study.
Appendix S4. Tests of Mixed ANOVA Assumptions and Descriptive Statistics.