Studies in Second Language Acquisition (2024), **46**, 1475–1492
doi:10.1017/S0272263124000445

REPLICATION STUDY

Task design, L1 literacy, and second language oracy

A close replication of Tavakoli and Foster (2008)

Jonathon Ryan[1], Pauline Foster[2], Yi Wang[3], Anthea Fester[1] and Jia Rong Yap[4]

[1]Centre for Languages, Hamilton, Waikato Institute of Technology, Waikato Mail Centre, New Zealand; [2]Centre for Applied Linguistics, University College London, London, UK; [3]Centre for Languages, Waikato Institute of Technology, Hamilton, New Zealand and [4]Waikato Institute of Technology, Hamilton, New Zealand
Corresponding author: Pauline Foster; Email: pauline.foster@ucl.ac.uk

(Received 30 July 2023; Revised 19 June 2024; Accepted 03 July 2024)

Abstract

This paper reports a replication of part of Tavakoli and Foster's (2008) investigation into the influence of narrative task design on second language (L2) oral performance. The initial study found in part that narratives with both foreground and background information elicited significantly greater syntactic complexity than those with only foreground information. This close replication adds the variable of literacy, conducting the study with adult refugees to New Zealand with low first language (L1) literacy. Participants narrated two of the four cartoon strips in Tavakoli and Foster (2008). In contrast to the initial study, background information in the narrative tasks had no impact on the syntactic complexity, lexical diversity, or fluency of performances. However, given the tendency of participants to omit background events, this outcome is discussed in terms of visual literacy, and aptness to describe rather than connect the cartoon frames. The implications for the use of narrative tasks with such learners are explored.

Keywords: second language literacy; visual literacy; task-based learning; narrative task design

Introduction

Overwhelmingly, the theory and practice of second language teaching and learning have been founded on empirical evidence drawn from highly literate populations. Largely omitted from the investigation are the 781 million adults estimated by the United Nations (Education for All Global Monitoring Report Team, 2015) to have low literacy in their first language (LL1L). Their underrepresentation in the literature is likely due to their absence in the university system and its conduits from which second language acquisition (SLA) research typically recruits participants. A recent review of meta-

analyses in SLA research found that "a staggering 88% of all adult samples are university student samples" (Andringa & Godfroid, 2020, p. 138). Such convenience sampling means that LL1L adults are routinely overlooked during hypothesis generation and absent from participant recruitment, raising questions about the generalizability of SLA findings to populations with low L1 literacy (Andringa & Godfroid, 2020; Tarone & Bigelow, 2012).

Recent research has suggested that L1 alphabetic literacy substantially and consistently influences language processing and metalinguistic knowledge (e.g., Havron & Arnon, 2017; Tarone, Bigelow & Hansen, 2009; Vágvölgyi, Coldea, Dresler, Schrader & Nuerk, 2016). For instance, in lexical segmentation tasks, such as when participants first hear and then reverse the order of words in a phrase (e.g., "little boy" to "boy little"), a stronger level of L1 literacy has been associated with significantly more accurate second language (L2) performance (Havron & Arnon, 2017). In tasks designed to elicit word-for-word repetition, LL1Ls are significantly more inclined than literate adults to instead provide a synonym, suggesting greater attention to meaning than form (Kurvers, 2015). Such findings have been interpreted as indicating that the development of literacy skills, such as fluent encoding and decoding of sound and letter combinations, "changes the way in which the individual processes oral language" (Tarone, et al, 2009, p. 12). Literacy, in this view, entails skills for perceiving and inspecting utterances at a level of linguistic abstraction, independent of their immediate communicative function. Tarone et al. (2009) argued such abstraction is foundational to many of the established practices of second language teaching, including explicit teaching of forms and oral corrective feedback. If confirmed, such findings suggest a need to reexamine other assumptions in SLA, and how they are applied in language classrooms where explicit teaching and feedback are normal practice. One way to test the extent to which SLA research findings are generalizable to LL1L adults is through close replications, in which change is limited to a single major variable of interest (Porte & McManus, 2019), in this case through selecting participants with low literacy in their L1 (Marsden, Morgan-Short, Thompson & Abugaber, 2018). There have been some recent advances in this area, with *Language Learning* publishing a special edition on just such replications (Godfroid & Andringa, 2023).

One SLA topic ripe for exploration is the interaction between literacy and elements of language learning task design. Such work explores the optimal design and use of tasks in the classroom, thereby offering practical relevance to teachers. While previous studies include a range of first languages and contexts, they have almost invariably been conducted with highly literate populations, with implicit assumptions about the generalizability of their findings to LL1Ls. One exception is a prior study by the present authors, who recruited LL1Ls in a replication of a study into the impact of pretask planning on oral performance (Ryan et al., 2023). The initial study (Foster & Skehan, 1996) found that the provision of 10 minutes of planning time led to greater oral fluency, grammatical accuracy, and syntactic complexity, but the replication found no such effects for the LL1L learners. Aside from questions specific to planned performance, the study also raised more general questions about the generalisability to LL1Ls of other findings relating to task variables.

Initial study

Tavakoli and Foster's (2008) initial study is situated within a body of task-based learning and teaching research (see Skehan & Foster, 2012, for a comprehensive overview) which explores L2 proficiency and development through task performance measures such as accuracy, fluency, syntactic complexity, and lexical diversity.

https://doi.org/10.1017/S0272263124000445 Published online by Cambridge University Press

Table 1. Designs of initial and present studies

	Initial study		Replication study
Storyline complexity	Inherent narrative structure		Inherent narrative structure (not examined)
	Loose	Tight	
- background	Journey task	Football task	Football task
+ background	Walkman task	Picnic task	Picnic task

The approach is cognitive (Skehan, 1998), in that it assumes a learner's L2 performance is shaped by inevitably limited cognitive resources having to be assigned to the competing demands of speaking accurately, and fluently, with appropriately complex syntax, and appropriately diverse vocabulary.

Tavakoli and Foster's specific focus was how a narrative task might impose a greater or lesser cognitive load on the speaker, depending on aspects of its design. Their study involved four cartoon strips, characterized by combinations of two features: *storyline complexity* and *inherent narrative structure.* Storyline complexity was defined by the presence or absence of background events accompanying the foregrounded constituents of the narrative. A storyline lacking background events was deemed to be *simple*, whereas a storyline with background events was deemed to be *complex.* A narrative in which episodes led strictly from one to the next was deemed to have a *tight structure*, while one in which episodes could be rearranged without loss of coherence was deemed to have a *loose structure.* Tavakoli and Foster's research hypotheses followed partly from Skehan and Foster (1997) who found that a tight narrative structure supported greater fluency and accuracy, and partly from a post hoc finding of Tavakoli and Skehan (2005) that a complex storyline supported greater syntactic complexity. A further hypothesis predicted that storyline complexity would support greater lexical diversity. In a 2×2 factorial design, the participants each performed two of four tasks, setting up a between-participants comparison for narrative structure and a within-participants comparison for storyline complexity (Table 1). The participants were 100 intermediate English as Foreign Language learners, recruited in two locations: London and Teheran. The London cohort ($n = 40$, mostly female, age range 19–45) came from a wide variety of L1 backgrounds, whereas the Teheran cohort ($n = 60$, all female, age range 19–47) were L1 Farsi speakers. Having two contexts for data collection allowed the authors to explore a further, null, hypothesis that similar patterns of performance features would be obtained from learners based within the target language environment (London) and those based outside it (Teheran). Both of these contexts involved highly literate learners familiar with the conventions of communicative language teaching and its emphasis on task performance in class.

In summary, the results showed strong support for an association of greater syntactic complexity with narrative storyline complexity, good support for an association of greater accuracy with tight narrative structure, limited support for an association of fluency with tight narrative structure, ambiguous results for an association of lexical diversity with storyline complexity, no support for an association of learning environment with greater accuracy or fluency, and, finally, strong support for an association of the London learning environment with greater syntactic complexity and lexical diversity.

https://doi.org/10.1017/S0272263124000445 Published online by Cambridge University Press

The replication study design

As noted above, the prevalence in SLA research of sampling literate and educated populations arises in great part from convenience; academic researchers under time pressure to conduct and publish studies, and with limited budgets, tend to recruit from their adjacent student populations, which are invariably literate and highly educated. While potential LL1L participants are plentiful in the Global North, having fled violent conflict or extreme poverty in their native countries, they are not likely to be found in sufficient numbers in higher education. To recruit for this study required a sufficient and stable participant pool of learners of similar L2 proficiency and low L1 literacy, such as provided by Te Pūkenga in New Zealand, a national network of Institutes of Technology and Skills, delivering vocational and applied education. Even so, as found by Ryan et al. (2023), it was anticipated that adequate recruitment and retention might still prove difficult, and therefore the sample size required to replicate the 2 × 2 factorial design of the initial study would not be feasible. To mitigate this, the replication design focused on just one of the variables from the initial study. Since the strongest findings from the initial study related to storyline complexity in the prompt tasks, this was selected as the one independent variable for the replication, omitting the other independent variables of narrative structure and learning context. Table 1 lays out the design contrasts and Table 2 lists the differences between the initial and present studies.

The replacement of participants with advanced L1 literacy by those with low L1 literacy is the focus of our investigation and represents a major change. Our close replication also explored an additional hypothesis (that entailed no additional changes to the research method) concerning the effect of storyline complexity on fluency, which was motivated by indications of such an effect in Tavakoli and Foster's results (detailed below under H3 in the hypotheses section). Other minor differences were entailed by the practicalities of recruiting from a low L1 literacy participant pool, or by the simplification of the initial design. For example, in terms of English L2 proficiency, the initial study recruited intact lower–intermediate classes. Our replication, requiring the filter of low L1 literacy, could not recruit intact classes. To reach our minimum recruitment target, we drew on learners from various classes at the slightly broader proficiency level of lower– to mid-intermediate. The simplification of the design of the initial study meant storyline complexity was the single independent variable, and was investigated using two of the initial study's four tasks. This provided only a within-participants comparison for which a multivariate analysis of variance (MANOVA) test would not have been appropriate; paired samples t-tests were therefore employed.

Replication hypotheses

Hypotheses one (H1) and two (H2) are the same as in the initial study (where they are labeled H1 and H4) and explore how storyline complexity relates to syntactic complexity and lexical diversity in performance.

H1: Compared with a narrative with only foreground events, one with both foreground and background events will be associated with learners producing syntactically more complex language. H1 was explored in the initial study where it was strongly supported.

H2: Compared with a narrative with only foreground events, narratives with both foreground and background events will be associated with learners producing language with greater lexical diversity. H2 was explored in the initial study, with mixed results.

https://doi.org/10.1017/S0272263124000445 Published online by Cambridge University Press

Table 2. Differences between initial and present studies

	Tavakoli & Foster (2008)	Present	Magnitude of change
L1 literacy	Advanced	Low	Major
L2 Proficiency level	Low intermediate	Low-mid intermediate	Minor
Independent variables	Storyline complexity, Task structure, & Location	Storyline complexity	Minor
Tasks	*Picnic, Walkman, Journey* and *Football*, permitting comparisons of tight vs. loose narrative structure, and plus vs. minus storyline complexity	*Picnic* and *Football*, permitting comparison of plus vs. minus storyline complexity	Minor
Location of data gathering	London and Tehran	Hamilton	Minor
Comparisons	Between- and within-participants	Within-participants only	Minor
Participant recruitment	Recruited intact EFL classes	Recruited from a variety of classes.	Minor
Lexical diversity	Measured by VocD	Measured by VocD and MTLD	Minor
Hypotheses	Four hypotheses concerning effect of storyline complexity and narrative structure on task performance	Two hypotheses from the initial study, concerning effect of storyline complexity on task performance. (*Narrative structure not investigated)*	Minor
	One hypothesis on the effect of location on task performance	*(Effect of location on task performance not investigated*)	Minor
Analyses	Factor analysis MANOVA	Factor analysis Paired samples t-tests	Minor

H3: Compared with a narrative with background and foreground events, a narrative with only foreground events will be associated with learners producing more fluent language.

(H3 was explored in the initial study only with regards to tight and loose narrative structure.)

The motivation to include H3 in the replication arose from our observation of a trend in the initial study results. Means for all four repair fluency variables in the Teheran cohort and three of the four repair fluency variables in the London cohort were higher for *Picnic* than for *Football*, indicating the task with a simple storyline was performed more fluently (see Appendix A). This would be in line with Skehan's (2014) prediction that a lighter cognitive load (a story with only foreground events) would support greater oral fluency than a heavier cognitive load (a story with both background and foreground events).

Participants

G*Power (Faul, Erdfelder, Lang & Buchner, 2007) was used to calculate statistical power. For matched pairs with an α error probability set at 0.05 and a medium effect size d_z (0.5) an actual power of 0.95 was generated with a participant sample size of 45. This was a feasible recruitment target within our context. Candidate participants were sought from lower-intermediate and intermediate English language classes at one Te Pūkenga site, a vocational and technical education provider in Hamilton, New Zealand. These classes were at Levels 2, 3 (general), and 3 (applied) of the New Zealand

https://doi.org/10.1017/S0272263124000445 Published online by Cambridge University Press

Certificate in English Language (NZCEL) programme, which the programme documents peg to Common European Framework of Reference (CEFR) levels from mid-B1 to low-B2 (NZQA, 2019).

Following Tarone et al. (2009), invitations to participate were based on results from the Native Language Literacy Screening Tool (NLLST) (Florida Department of Education, 2021). This provides written prompts in a learner's L1 (e.g., Name; Where were you born?), and assessment involves observing a candidate reading the prompt and writing a response. Indications of low literacy include lack of speed and fluency of responses, "facial and body language cues that indicate frustration and/or lack of understanding" and "struggling to write, hesitation, laboring over each letter" (Florida Department of Education, 2014, p. 2). The NLLST is available in 29 languages, to which were supplemented versions in Somali (courtesy of Tarone et al., 2009) and Dari (which we commissioned). Candidates who scored low to moderate literacy were invited to participate in the study. Ultimately, 54 participants were recruited, but upon inspection of their recordings, three had to be discarded: one did not attempt the second task, one could not be transcribed with confidence due to very poor articulation, and the third appeared to be of substantially lower proficiency than the others. The remaining 51 participants were former refugees from a range of nationalities and first language backgrounds who had arrived in New Zealand as older teens or adults. The largest representations were from Afghanistan (19), Somalia (eight), Pakistan (four) Colombia (four), and Congo (three) with one or two each from Cambodia, China, India, Indonesia, Myanmar, Philippines, Sri Lanka, Syria, and Thailand. These backgrounds of the participants contrast with those of the initial study, who were either L1 Farsi speakers from Iran or mainly European students from a wide variety of L1 backgrounds based in London. In addition, while in the initial study, nearly all participants were women (including all those recruited in Iran), men constituted more than a third of the replication sample (37%). The age ranges of participants appear broadly similar, with all but five participants in the replication falling within the 19–47 range of the initial study. Of the remaining five, two were 18, and three were older (55, 67 and 74). The mean age was 32 (figures are not available for the initial study).

Tasks

All four cartoon strips used in the initial study are provided in the supplemental materials. Our replication employed the two with a tight narrative structure: *Football* and *Picnic* (Heaton, 1966). *Football* depicts a group of boys losing and then ingeniously recovering their football from a deep hole. The storyline has only foreground events. *Picnic* depicts two children preparing for and setting off on a picnic, unaware that their small dog has hidden inside their basket and eaten the food by the time they unpack it. The storyline has foreground and background events. The instruction provided to participants was to look at the pictures and tell the story.

Data collection

Following the initial study, all recordings were elicited by the same researcher, who arranged individual meetings with participants in a quiet classroom. Participants were shown the first cartoon strip, and told their task was to tell the story in the pictures. They had 3 min to prepare, after which their narration was audio-recorded. The second strip was then presented and the same procedure followed. To control for a possible practice effect, the order of tasks was counterbalanced.

https://doi.org/10.1017/S0272263124000445 Published online by Cambridge University Press

Measures

Audacity software was used during transcription, with silences ≥ 0.5s coded as pauses. After transcription, the data were divided into AS-units and associated subclauses (Foster, Tonkyn, & Wigglesworth, 2000) and coded for the performance measures employed in the initial study. An additional measure of lexical diversity was added (MTLD) to triangulate with VocD, which had returned inconsistent results. All the AS-unit and clause boundaries were coded and checked by two researchers, and all disagreements were resolved. Following Foster et al. (2000), both finite and nonfinite clauses were included in the count of total clauses. For the performance measures of fluency, 10% of the coded transcripts were checked by another member of the team. We used Cohen's kappa coefficient to assess interrater reliability. This revealed very strong agreement in coding, $\kappa = .95$ (95% CI, .923 to .967), $p = < .001$. The consistency of VocD and MTLD scores is assumed through the respective algorithms (Weblingua Ltd, 2023).

Extensive work on complexity, accuracy, lexis, and fluency (CALF) measures has been undertaken since the initial study and has affirmed these as distinct constructs. (See, for two examples among many, Housen, Kuiken & Vedder, 2012, and Vercelotti, 2017.) Accordingly, while the initial study reported a factor analysis of CALF measures, such an evaluation was not strictly necessary for this replication. Nevertheless, a factor analysis was run, confirmed the four constructs as distinct, and is provided in the supplemental materials.

For the remaining analyses, a one-way analysis of variance (ANOVA) test was used to check for a practice effect and the three main hypotheses of the study were examined with *t*-tests. For each, the alpha level was set at 0.05. Effect sizes are reported using Cohen's *d* with 95% confidence intervals. Following recommendations for within-group contrasts in SLA research, *d* values of .60 were interpreted as indicating a small effect, 1.00 as a medium effect, and 1.40 as a large effect (Plonsky & Oswald, 2014). In relation to H1 and H2, replication of the initial study's findings would be assumed if the *Picnic* narratives contained greater syntactic complexity (H1) and greater lexical diversity (H2) than *Football* evidenced in values of $p \leq .05$ and $d = .60$. The dependent variables, their definitions, and associated measures are set out in Table 3.

Results

A one-way ANOVA for task order indicated no discernible practice effect with the possible exception of a tendency to make more reformulations in the second task, $F(1, 49) = 4.19$, $p = .046$, $r = .28$. Caution is required in interpreting this result due to the relatively low number of reformulations overall ($M = 6.6$ per task). A correlation matrix for *Football* revealed a variety of coefficients between .38 and .85. The Kaiser–Meyer–Olkin value of sampling adequacy was .55, and Bartlett's test of sphericity was significant at $p < .001$, confirming substantial correlation in the data. As in the initial study, a factor analysis confirmed syntactic complexity, lexical diversity, and fluency as distinct constructs (see Tables S1 and S2 in the supplemental materials).

Paired sample *t*-tests were then run to explore the three hypotheses. Assumptions for these tests are reported in the supplemental materials (Table S3). H1 predicted that a narrative with foreground and background events (i.e., *Picnic*) would be associated with learners producing syntactically more complex language compared with a narrative with only foreground events (i.e., *Football*).

Table 4 shows the mean scores for syntactic complexity were higher for *Football* than *Picnic* (1.42 vs. 1.37), i.e., against H1. The pairwise *t*-test showed no statistical significance for this comparison, $t(50) = 1.36$, $p = .090$, $d = .19$, a small effect. The mean length

https://doi.org/10.1017/S0272263124000445 Published online by Cambridge University Press

Table 3. Dependent variables.

Variable	Definition	Measure
Syntactic complexity	The organization of what is said through progressively more elaborate language and greater variety of syntactic patterning (Foster & Skehan, 1996, pp. 303–304)	Total clauses divided by total AS units (Foster et al, 2000) Mean length of unit: expressed as mean number of words per AS-unit
Lexical diversity	The variety of active vocabulary deployed by a speaker (Malvern & Richards, 2009)	VocD (Malvern & Richards, 2009) MTLD (McCarthy & Jarvis, 2010)
Repair fluency	The need to engage in more frequent repairs in speech through: *Reformulations:* Phrases or clauses repeated with some modification to syntax, morphology, or word order *Replacements:* Lexical items that are immediately substituted *False starts:* Utterances abandoned before completion and that may or may not be followed by a reformulation *Repetitions:* Words, phrases, or clauses repeated without modification to syntax, morphology, or word order (Foster & Skehan, 1996, p. 310)	Reformulations, false starts, word replacements, repetitions; each expressed as mean per min on task
Breakdown fluency	Pauses and the total amount of silence that disrupts one's capacity to engage in continued performance (Foster & Skehan, 1996)	Mid- and end-clause pauses, expressed as mean number per min on task Total silence expressed as sum of pauses ≥ 0.5s per min on task

Table 4. Paired samples statistics for complexity.

	Mean	N	Std. deviation	Std. error mean
Syntactic complexity *Football*	1.42	51	.18	.03
Syntactic complexity *Picnic*	1.37	51	.19	.03
Mean Length Unit *Football*	7.84	51	1.40	.20
Mean Length Unit *Picnic*	7.99	51	1.26	.18

of unit score for *Football* is lower than for *Picnic* (7.84 vs. 7.99). This is predicted by H1, but the difference is very small, and the pairwise *t*-test was not statistically significant, $t(50) = -.84$, $p = .203$, $d = .19$, a small effect. H1 is not upheld.

H2 predicted that a narrative with foreground and background events (i.e., *Picnic*) would be associated with learners producing language with greater lexical diversity compared with a narrative with only foreground events (i.e., *Football*).

Table 5 shows the mean score for VocD was marginally higher for *Picnic* than for *Football* (38.07 vs. 37.21). This is in the direction of H2, but the pairwise *t*-test was not statistically significant, $t(50) = -.64$, $p = .262$, $d = .09$, a small effect. The MTLD mean for *Football* is marginally higher than for *Picnic* (35.33 vs. 33.88). This is against H2, but a pairwise *t*-test was not statistically significant, $t(50) = 1.17$, $p = .124$, $d = .16$, a small effect. H2 is not upheld.

https://doi.org/10.1017/S0272263124000445 Published online by Cambridge University Press

Table 5. Paired samples statistics for lexical diversity.

	Mean	N	Std. deviation	Std. Error Mean
VocD *Football*	37.21	51	10.74	1.50
VocD *Picnic*	38.07	51	9.15	1.29
MTLD *Football*	35.33	51	9.18	1.29
MTLD *Picnic*	33.88	51	8.02	1.12

Table 6. Paired samples statistics for repair fluency.

	Mean	N	Std. deviation	Std. error mean
Reformulation *Football*	2.74	51	1.40	.20
Reformulation *Picnic*	2.76	51	1.21	.17
Replacement *Football*	0.25	51	0.50	.07
Replacement *Picnic*	0.09	51	0.19	.03
False Start *Football*	2.46	51	1.20	.17
False Start *Picnic*	2.74	51	1.35	.19
Repetition *Football*	2.85	51	2.36	.33
Repetition *Picnic*	2.51	51	1.99	.28

H3 predicted that a narrative with only foreground events (i.e., *Football)* would be associated with learners producing more fluent language than a narrative with both background and foreground events (i.e., *Picnic*). Table 6 shows the results of the pairwise comparisons for the four repair fluency variables. For reformulations per min, the means were nearly identical (*Football* 2.74 vs. *Picnic* 2.76) with the pairwise *t*-test showing no significant difference, $t(50) = -.12$, $p = .453$, $d = .02$, a small effect. For replacements, the means were both very low (*Football* 0.25 vs. *Picnic* 0.09), and against H3. The pairwise *t*-test indicated a significant difference, $t(50) = 2.21$, $p = .016$, $d = .31$, a small effect[1]. The means for false starts were close (*Football* 2.46 vs. *Picnic* 2.74), in the direction of H3, but the pairwise *t*-test was statistically insignificant, $t(50) = -1.28$, $p = .104$, $d = .18$, a small effect. Finally, for repetitions, the means were again close (*Football* 2.85 vs. *Picnic* 2.51), against H3, with the pairwise *t*-test showing no significant difference, $t(50) = 1.49$, $p = .071$, $d = .06$, a small effect.

Notably, the means for repair fluency measures are all very low, indicating that the participants, in general, did not engage much in repair fluency behavior for either task. To focus this further, and to exclude individual preferences some learners might have for a particular repair gambit, a variable was computed that summed all the repair fluency scores together, and a further pairwise comparison was run. Table 7 shows the means are almost identical (*Football* 8.30 vs. *Picnic* 8.10) The paired *t*-test outcome was not significant, $t(50) = .49$, $p = .31$, $d = .07$, a small effect.

In Table 8 breakdown fluency variables are shown to pattern in the same way as the repair fluency variables. In terms of mid-clause pauses, the means were nearly identical (*Football* 10.28 vs. *Picnic* 10.08), with the paired samples *t*-test showing no significant difference, $t(50) = .43$, $p = .334$, $d = .06$, a small effect. For end-clause pauses the means were also very close (*Football* 6.15 vs. *Picnic* 6.61) with the paired samples *t*-test

[1]However, a frequency analysis showed that across the 102 narratives, replacements were very infrequent. The mode was zero: 36 of the 51 *Football* narratives and 40 of the 51 *Picnic* narratives had no replacements at all. The maximum frequency score for replacements was 2, recorded by one participant in his *Football* narrative.

https://doi.org/10.1017/S0272263124000445 Published online by Cambridge University Press

Table 7. Paired samples statistics for sum of repair fluency variables.

	Mean	N	Std. deviation	Std. error mean
Sum of repairs *Football*	8.30	51	3.97	.56
Sum of repairs *Picnic*	8.10	51	3.23	.45

Table 8. Paired samples statistics for breakdown fluency.

	Mean	N	Std. Deviation	Std. Error Mean
Pause mid-clause *Football*	10.28	51	3.90	.55
Pause mid-clause *Picnic*	10.08	51	3.50	.49
Pause end clause *Football*	6.15	51	2.58	.36
Pause end clause *Picnic*	6.61	51	2.28	.36
Silence (per min) *Football*	22.3	50	7.64	1.08
Silence (per min) *Picnic*	22.0	50	7.76	1.10

showing no significant difference, $t(50) = -1.22$, $p = .114$, $d = .17$, a small effect. One outlier was removed from the measure of total silence (see Table S3 in the supplemental material). The means for total silence per min were close (*Football* 22.3 vs. *Picnic* 22.0), and the paired samples *t*-test showed no significant difference, $t(49) = -.489$, $p = .313$, $d = .07$, a small effect.

Studies of L2 oral task performance suggest pause location is a surface indicator of the fluidity of the process underlying speech production. Mid-clause pausing interrupts a syntactic unit and is evidence of the speaker's local difficulty in encoding a pre-linguistic concept as language (Levelt, 1989; Skehan, Foster & Shum 2016). Computing mid-clause pauses as a percentage of total clauses, and then comparing the mean scores illuminates whether and to what degree speakers found the dual storyline of *Picnic* to be more syntactically demanding than the single storyline of *Football*. As *Picnic* requires the speaker to attend to both background and foreground events, and connect them syntactically, the proportion of mid- to end-clause pauses should be greater here than in the single storyline of *Football*. Accordingly, a variable was computed representing breakdown fluency as the percentage of pauses occurring mid-clause. Table 9 shows the means were very close (*Football* 61.7% vs. *Picnic* 59.7%), against the direction of H3. The paired samples *t*-test did not reach significance: $t(50) = 1.25$, $p = .110$, $d = .17$, a small effect. Hypothesis 3 is not upheld for any measure of repair or breakdown fluency.

A succinct summation of these results is that our participants' oral performance was not influenced by storyline complexity in terms of any measure of syntactic complexity, lexical diversity, or fluency.

DISCUSSION

For ease of reference, Table 10 lays out a comparison of the results for syntactic complexity and lexical diversity with those of the initial study. It is clear that telling a narrative with background events (*Picnic*) prompted the learners in London and Teheran to express themselves in AS-units that were significantly longer and with more complex syntax. Tavakoli and Foster (2008, p. 459) argue that "the presence of background events in a narrative necessitates, in English, using particular structures such as subordinated clauses to connect the background information to the events that are happening in the foreground." By contrast, the learners in Hamilton narrated both

https://doi.org/10.1017/S0272263124000445 Published online by Cambridge University Press

Table 9. Paired samples statistics for percentage of pauses occurring at mid-clause.

	Mean	N	Std. deviation	Std. mean error
% Pause mid-clause *Football*	61.7	51	15.1	.02
% Pause mid-clause *Picnic*	59.7	51	13.2	.02

Table 10. Comparisons of means for syntactic complexity and lexical diversity across initial and replication studies.

	Initial study		Replication
Measure	London cohort	Teheran cohort	Hamilton cohort
Syntactic complexity			
Football	1.41 (.13)	1.28 (0.16)	1.42 (.18)
Picnic	1.71 (.28)	1.59 (0.38)	1.37 (.19)
Comparison	Picnic > Football p = .001	Picnic > Football p = .001	Football > Picnic p = .091 ns
Mean length of utterance			
Football	8.27 (1.37)	7.55 (1.14)	7.84 (1.40)
Picnic	10.86 (2.50)	9.27 (1.68)	7.99 (1.26)
Comparison	Picnic > Football p = .001	Picnic > Football p = .001	Football > Picnic p = .230 ns
VocD			
Football	38.37 (11.18)	28.75 (11.26)	37.21 (10.74)
Picnic	36.59 (9.46)	27.76 (5.89)	38.07 (9.15)
Comparison	Football > Picnic p = .59 ns	Football > Picnic p = .67 ns	Picnic > Football p = .262 ns
MTLD			
Football	Not measured	Not measured	35.33 (9.19)
Picnic	Not measured	Not measured	33.88 (8.02)
Comparison			Football > Picnic p = .124 ns

Note: The *football* task includes only foreground events, whereas the *picnic* task includes both foreground and background events.

tasks in AS-units that were approximately the same length, and with a slightly (and insignificantly) more complex syntactic structure for the task *without* background events (*Football*). Crucially, inspection of the replication transcripts revealed that 34 of the 51 Hamilton participants did not take the *Picnic* background events into account in their narrations, and others gave them scant attention. For Tavakoli and Foster (2008) "[*Picnic*] carries through the idea of the dog hidden in the basket while the children leave the house, walk through the countryside, and choose their picnic spot, inviting the storyteller to remind us at several points that the children still do not know about the dog." However, the Hamilton participants do not interpret the pictures in this way. For most of them, the dog is an unseen or unimportant detail rather than a crucial one, effectively subverting the storyline's complexity. With both *Picnic* and *Football* largely narrated by the Hamilton participants as having only foreground events, it is not surprising that the cross-task complexity scores are so similar[2].

[2]A reviewer suggests that perhaps within-subjects variations in syntactic complexity are more closely related to proficiency than to the task prompt. However, Foster (2001) showed that task type and implementation conditions influence the syntactic complexity of L1 performance in the same way that it influences syntactic complexity in L2 performance (Foster & Skehan, 1996). Further, Foster and Tavakoli (2009) report a

https://doi.org/10.1017/S0272263124000445 Published online by Cambridge University Press

Tavakoli and Foster (2008) predicted that lexical diversity, measured by VocD, would increase with storyline complexity. In both their London and Teheran cohorts, the means were, marginally, in the opposite direction to that predicted (*Picnic* < *Football*) but neither difference reached statistical significance. For the replication study, the VocD means were, marginally, in the predicted direction (*Picnic* > *Football*), but also did not reach statistical significance. Tavakoli and Foster (2008) reported mixed findings for VocD in their two loosely structured narrative tasks: *Walkman* (with background events) and *Journey* (without background events). In line with the hypothesis that background events would result in greater lexical diversity, *Walkman* in the Teheran cohort had significantly higher VocD scores than *Journey* (p =.001), and in the London cohort, the difference came close to significance (p = .06 ns), indicating that background details can have some impact on lexical diversity. To explore if a triangulating measure could bring greater clarity, the replication study employed MTLD (McCarthy et al., 2010). The means for MTLD gave a slight advantage to *Football*, i.e., against the predicted direction, but this did not reach statistical significance. Thus, the relationship between task design and the performance dimension of lexical diversity remains slippery. With many learners in the replication study letting the background events in *Picnic* pass unremarked, it is unsurprising that no significant effect for lexical diversity was found. However, the London and Teheran learners in the initial study did weave the dog into their *Picnic* narratives and still did not produce any greater lexical diversity than in their *Football* narratives, so the explanation for a lack of effect does not lie in L1 literacy. Nor can we say that the learners in Hamilton had poorer L2 vocabularies to draw on. The VocD means in Table 10 show that for both tasks they employed considerably more diversity than the learners in Teheran, and equal diversity to the learners in London. Their lack of L1 literacy had not affected their acquisition of vocabulary, which was presumably supported by their living inside rather than outside the target language environment. (See Foster and Tavakoli, 2009, for a full analysis of this aspect of their data.) In their discussion, Tavakoli and Foster (2008) acknowledge that cross-task comparisons for lexical diversity can be explained by different picture stimuli entailing different vocabulary items, and are thus problematic. Nevertheless, the replication did show that the Hamilton learners, responding to the same prompts as learners in the initial study, had lexical resources to draw on that were as diverse as the London cohort's and more diverse than the Teheran cohort's, regardless of storyline complexity.

The influence of storyline complexity on fluency was added to the replication following the observation that mean scores for all but one repair fluency variables in the Teheran and London cohorts were higher for *Picnic* than for *Football*, suggesting that the oral performance of a task with only foreground events was associated with less repair behavior (Tavakoli & Foster, 2008, pp. 454–455, Tables 10–13). For breakdown fluency, Tavakoli and Foster (2008) report only on mid-clause pausing, where the mean scores for *Picnic* were lower than for *Football*, suggesting that the task with both background and foreground events was associated with fewer breakdowns. In the replication, the results for the comparison of breakdown variables in *Picnic* and *Football* reveal that the Hamilton participants maintained their wonted fluency across both tasks. Tables 6, 7, 8, and 9 all show remarkable consistency between the mean scores. Thus, with the *Picnic* background events either unseen or ignored, many of the

significant effect for storyline complexity on syntactic complexity in L1 task performance, mirroring that found for the low-intermediate L2 participants of Tavakoli and Foster (2008).

https://doi.org/10.1017/S0272263124000445 Published online by Cambridge University Press

Hamilton participants made *Picnic* into a task with a single storyline and described it with the same levels of repair and breakdown fluency they exhibited in *Football.*

Inspection of the transcripts revealed that we were incorrect in our assumption that LL1L participants would respond to the two narrative tasks in the way the initial study participants had. The replication participants tended to describe the cartoons frame by frame, not so much to make a story, but rather to list things in each picture, i.e., more inventory than narrative. For example, in *Picnic* both the London and Teheran cohorts had taken account of both the foreground and background details in the six frames and woven them into a narrative line. The dog appears in the background of frame 1, while in the foreground the children are making their sandwiches. In the foreground of frame 2, the dog is climbing into the picnic basket, while in the background the children and their mother are consulting a map. The dog does not appear in frames 3 and 4, though the basket does, being carried by the children as they set off for their picnic. The dog reappears in the foreground of frame 5, jumping out of the basket, and then running around in the background of frame 6 as the children realize their sandwiches have been eaten. In the replication, some participants did not mention the dog at all, some mentioned it as present in frame 1 but did not describe it getting into the basket in frame 2, and some noticed (with bemusement) that, suddenly, it was back in frames 5 and 6. In general, the replication participants focused on what the humans were doing, no matter if they were in the background or the foreground.

In considering why this should be so, accounts based on oral language processing appear to offer little purchase. The replication participants did not require literacy-supported cognitive tools (e.g., awareness of word boundaries) to understand how events follow in a sequence. That ability is either innate or a habit of mind developed in childhood, independent of literacy. The cartoon tasks asked them to put into words the events depicted across the sequence of frames. By ignoring the *Picnic* dog after it climbed into the picnic basket, as 34 of the 51 participants did, there were no background events to weave into the human storyline, and no need to employ more complex syntax than they used for the single storyline of *Football.* That is the simplest explanation for H1 not receiving support.

Setting literacy aside, other explanations for the LL1L participants responding to picture narratives in this way can be sought in their levels of visual literacy, schooling experience, or a combination of both. For example, the cartoon strip nature of the elicitation materials may have introduced a complication for these participants that is distinct from the variable of storyline complexity. Although narrative aspects of pictorial representation are likely universal, cartoon strips are a relatively recent cultural artifact arising from print literacy, and even when they are free of speech or other representations of language, they still involve "reading" in terms of a left-to-right orientation and making sense of semiotic resources that are ultimately learned through experience (Huang & Archer, 2012). This includes deriving meaning from matters of layout, visual focus, and in particular cultural conventions of visual representation.

In the process of learning to read, children in literate societies amass vast experience of decoding visual storylines in picture books, both at school and in the home. By contrast, low–literate L2 English learners cannot draw on such resources of childhood experience and may therefore encounter difficulties in decoding symbolic visuals such as arrows showing location or movement, clocks indicating the passage of time, or cloud "bubbles" depicting thought (Brod, 1999; Bruski, 2012; Hvitfeldt, 1985). While it is true that *Picnic* and *Football* comprise only iconic visuals, rather than symbolic ones, we cannot be sure that the cartoon people and places in *Picnic* and *Football* were drawn

in such a way to make obvious to someone with limited visual literacy what the artist expected to be noticed[3].

To expand on the point about visual and print literacy being dependent on experiential factors, it can be argued that they are inseparable from other factors related to early-years education, such as familiarity with picture stories as pedagogic devices, and imaginative play as a route to discovery and learning. These are typical elements of primary education in many developed countries and are recognized as legitimate learning activities if encountered later in adult education, such as in communicative or task-based foreign language classrooms. Our replication participants may not have understood education as "playful." It is typical in many of their countries of origin for education to entail strict discipline, teacher-led oral instruction, and student-rote learning. (In relation to education in Afghanistan, for example, see Orfan, Noori, Hashemi & Akramy, 2021, and Sakena Fund, 2023). So, either by their limited first-hand experience of primary education, or their general expectation of education as a serious rather than "fun" undertaking, they may not have been inclined to engage seriously with cartoons showing children enjoying themselves.

A related point is made by Norton (2007, p. 9) in the distinction between "real reading" and "fun reading"; the former is seen as teacher-prescribed, educational, and not designed to be amusing, while the latter, such as reading comic books, is dismissed as "garbage" and "a waste of time." (See further in Norton, 2003; Norton & Vanderheyden, 2004.) Although Norton's (2003) study was based in Canada, it is not unreasonable to suppose that LLIL adult refugees also do not expect learning materials to involve playfulness.

By contrast, the participants in the initial study were all very familiar with communicative language classrooms, in which opportunities to perform light-hearted or imaginative speaking tasks are routinely offered to help develop fluency, vocabulary, pronunciation, and grammar. The *Picnic* and *Football* narratives were chosen as research instruments for the initial study because they were typical of such tasks and likely to engage interest. In this regard, they did not disappoint. The participants described the storylines as expected, taking into account both background and foreground details. As the initial study explored how task design influences syntax, it was observed that describing parallel events in English entails joining clauses with *as, while, during, at the same time*, *but*, and *however*. The replication has not shown that LL1L participants were incapable of using such syntactic devices, only that, in this context, they did not use them.

Future replications

Given the widespread use of communicative language teaching in the Global North and the increasing numbers of LL1L language learners, it is important to study how the latter are able to engage with the former. Task-based learning and teaching (TBLT) tends to assume learners are experienced with decoding visual prompts and are acquainted with picture stories as pedagogic devices. Our results suggest this might not hold for LL1L learners; the way they performed *Football* and *Picnic* could be explained by a lack of familiarity with such classroom activities, coupled with an inclination not to take them seriously. Accordingly, it would be useful to closely

[3]After data collection many participants observed informally that they were not familiar with comic strip narratives, and had never been asked to describe one.

https://doi.org/10.1017/S0272263124000445 Published online by Cambridge University Press

replicate other TBLT studies involving visual prompts, with LL1L participants grouped according to an assessment of their visual literacy. A suitable starting point may be a close replication of the other major variable investigated by Tavakoli and Foster (2008), tight vs. loose narrative structure. Conceptual and approximate replications (Porte & McManus, 2019) could also substitute an initial study's prompt tasks with something that better aligns with LL1L learners' expectations of language learning activities. As adults needing urgently to establish themselves in a new country, LL1L learners are often highly motivated to acquire practical life skills in their target language, so research tasks with more obvious real-life associations might engage them better.

Typically, TBLT research is cross-sectional, and it would be useful to set LL1L approximate replications inside a more longitudinal design. An initial study using picture prompts could be replicated with LL1L participants, and then posttask questionnaire or interview data could illuminate the extent to which LL1L learners took the research task seriously and/or viewed it as intrinsically valuable. The learners could then engage in a series of orientation sessions in which the pedagogic usefulness of pictures as L2 language prompts would be discussed. After this orientation, a task of a similar type to the initial task could be administered, and the questionnaire or interview repeated. The qualitative data arising from such a conceptual replication would give insights into whether LL1L learners' view of picture prompts changed over time. The quantitative data would show whether the LL1L learners' task performance changed over time to be more in line with that of the initial participants, or did not significantly change. This would, in turn, suggest ways of designing tasks that better align with LL1L learners' expectations of language classrooms or ways to socialize LL1L learners towards the norms of "Western" education. Another longitudinal tack would be to see how LL1L learners' visual literacies might be improved by practice in decoding picture sequences, and encouragement to supply links between the frames. This last suggestion in particular could form a further conceptual replication of Tavakoli and Foster (2008), investigating whether training in cartoon narratives leads to more integrated accounts of the pictures, and greater syntactic complexity at points where storylines intersect.

Supplementary material. The supplementary material for this article can be found at http://doi.org/10.1017/S0272263124000445.

Data availability statement. The experiment in this article earned Open Data and Materials badges for transparent practices. The data and materials are available at https://www.iris-database.org/details/iNTZjaRIBb.

Acknowledgements. The authors wish to acknowledge the very helpful comments of the reviewers on earlier drafts of the paper. We especially acknowledge the support and guidance of the Special Issue Editor, Kevin McManus. The project was assisted by a grant from Trust Waikato. Our thanks are also due to the participants who cheerfully gave their time to the research.

Competing interest. The author(s) declare none.

References

Andringa, S., & Godfroid, A. (2020). Sampling bias and the problem of generalizability in applied linguistics. *Annual Review of Applied Linguistics*, *40*, 134–142. https://doi.org/10.1017/S0267190520000033

Brod, S. (1999). *What non-readers or beginning readers need to know: Performance-based ESL adult literacy*. Spring Institute for International Studies. https://eric.ed.gov/?id=ED433730

Bruski, D. (2012). Graphic device interpretation by low-literate adult ELLs: Do they get the picture? *MinneTESOL/WITESOL Journal*, *29*, 7–29. https://hdl.handle.net/11299/162757

Education for All Global Monitoring Report Team. (2015). *Education for all 2000-2015: Achievements and challenges*. UNESCO.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007, 2007/05/01). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Florida Department of Education. (2014). *Native language literacy screening manual.* https://www.fldoe.org/core/fileparse.php/7522/urlt/0061376-native.pdf

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

Foster, P. (2001). Rules and routines: A consideration of their role in the task-based language production of native and non-native speakers. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Language tasks: Teaching, learning and testing* (pp. 75–93). Longman.

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., & Tavakoli, P. (2009). Native speakers and task performance: Comparing effects on complexity, fluency, and lexical diversity. *Language Learning*, *59*(4), 866–896. https://doi.org/10.1111/j.1467-9922.2009.00528.x

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

Godfroid, A., & Andringa, S. (2023). Uncovering sampling biases, advancing inclusivity, and rethinking theoretical accounts in second language acquisition: Introduction to the special issue SLA for All? *Language Learning*, *73*(4), 981–1002. https://doi.org/10.1111/lang.12620

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

Heaton, J. B. (1966). *Composition through pictures*. Longmans.

Housen, A., Kuiken, V. & Vedder, I. (2012). Complexity, accuracy and fluency: Definitions, measurement and research. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency—Investigating complexity, accuracy and fluency in SLA* (pp. 1–20). John Benjamins.

Huang, C.-W., & Archer, A. (2012). Uncovering the multimodal literacy practices in reading manga and the implications for pedagogy. In A. A. Zenger & B. Williams (Eds.), *New media literacies and participatory popular culture across borders* (pp. 44–60). Routledge.

Hvitfeldt, C. (1985). Picture perception and interpretation among preliterate adults. *Passage: A Journal of Refugee Education*, *1*(1), 27–30. https://archive.org/details/ERIC_ED254099

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

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

Malvern, D., & Richards, B. (2009). A new method of measuring rare word diversity: The example of l2 learners of French. In B. Richards, M. H. Daller, D. D. Malvern, P. Meara, J. Milton, & J. Treffers-Daller (Eds.), *Vocabulary studies in first and second language acquisition: The interface between theory and application* (pp. 164–178). Palgrave MacMillan.

Marsden, E., Morgan-Short, K., Thompson, S., & Abugaber, D. (2018). Replication in second language research: Narrative and systematic reviews and recommendations for the field. *Language Learning*, *68*(2), 321–391. https://doi.org/10.1111/lang.12286

McCarthy, P. M., & Jarvis, S. (2010, 2010/05/01). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, *42*(2), 381–392. https://doi.org/10.3758/BRM.42.2.381

New Zealand Qualifications Authority. (2019). *New Zealand certificates in English language (NZCEL) guiding document.*https://www.nzqa.govt.nz/assets/qualifications-and-standards/qualifications/EL-quals/NZCEL-Guiding-Document.pdf

Norton, B. (2003). The motivating power of comic books: Insights from Archie comic readers. *The Reading Teacher*, *57*(2), 140–147. http://www.jstor.org/stable/20205333

Norton, B. (2007). Critical literacy and international development. *Critical Literacy: Theories and Practices*, *1*(1), 6–15. http://www.criticalliteracyjournal.org/cljournalissue1volume1.pdf

Norton, B., & Vanderheyden, K. (2004). Comic book culture and second language learners. In B. Norton & K. Toohey (Eds.), *Critical Pedagogies and Language Learning* (pp. 201–221). Cambridge University Press. https://doi.org/10.1017/CBO9781139524834.011

Orfan, S. N., Noori, A. Q., Hashemi, A., & Akramy, S. A. (2021). Afghan EFL instructors' use of teaching methods. *International Journal of English Language Studies*, *3*(5), 31–38. https://doi.org/10.32996/ijels.2021.3.5.5

Plonsky, L., & Oswald, F. L. (2014). How big is 'big'? Interpreting effect sizes in L2 research. *Language Learning*, *64*(4), 878–912. https://doi.org/10.1111/lang.12079

Porte, G., & McManus, K. (2019). *Doing replication research in applied linguistics*. Routledge.

Ryan, J., Foster, P., Fester, A., Wang, Y., Field, J., Kearney, C., & Yap, J. R. (2023). L1 literacy and L2 oracy: A partial replication of Foster & Skehan (1996). *Language Learning*, *73*(4), 1345–1368. https://doi.org/10.1111/lang.12557

Sakena Fund. (2023). *Education: Afghan Institute of Learning*. https://www.sakena.org/afghan-institute-of-learning/education.php

Skehan, P. (1998) *A cognitive approach to language learning*. Oxford: Oxford University Press.

Skehan, P. (2014). Limited attentional capacity, second language performance, and task-based pedagogy. In P. Skehan (Ed.), *Processing perspectives on task performance* (pp. 211–260). John Benjamins. https://doi.org/10.1075/tblt.5.08ske

Skehan, P., & Foster, P. (1997). Task type and task processing conditions as influences on foreign language performance. *Language Teaching Research*, *1*(3), 93–120. https://doi.org/10.1177/136216889700100302

Skehan, P. & Foster, P. (2012). Complexity, accuracy, fluency and lexis in task-based performance: A synthesis of the Ealing research. In Housen, A., Kuiken, F. & Vedder, I. (Eds.) *Dimensions of L2 Performance and Proficiency: Complexity, Accuracy and Fluency in SLA* (pp. 199–220). John Benjamins.

Skehan, P., Foster, P. & Shum, S. (2016). Ladders and snakes in second language fluency. *International Review of Applied Linguistics in Language Teaching*, *54*(2), 97–101. https://doi.org/10.1515/iral-2016-9992

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.

Tavakoli, P., & Foster, P. (2008). Task design and second language performance: The effect of narrative type on learner output. *Language Learning*, *58*(2), 439–473. https://doi.org/10.1111/j.1467-9922.2008.00446.x

Tavakoli, P., & Skehan, P. (2005). Strategic planning, task structure and performance testing. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 239–277). Benjamins John.

Vágvölgyi, R., Coldea, A., Dresler, T., Schrader, J., & Nuerk, H.-C. (2016). A Review about functional illiteracy: Definition, cognitive, linguistic, and numerical aspects. *Frontiers in Psychology*, *7*, Article 1617. https://doi.org/10.3389/fpsyg.2016.01617

Vercellotti, M. (2017). The development of complexity, accuracy, and fluency in second language performance: A longitudinal study. *Applied Linguistics*, *38*(1), 90–111. https://doi.org/10.1093/applin/amv002

Weblingua Ltd. (2023). *Text Inspector*. https://textinspector.com/

Appendix A

Means and SDs for repair fluency measures for Picnic *and* Football *(Tavakoli & Foster, 2008)*

Cohort	Measure	Picnic	Football	Fluency comparison
Teheran	reformulations	0.83 *(0.79)*	0.63 *(0.61)*	picnic < football
	false starts	3.63 *(2.73)*	2.66 *(2.05)*	picnic < football
	replacements	1.90 *(1.66)*	1.26 *(1.28)*	picnic < football
	repetitions	4.64 *(4.16)*	2.70 *(3.08)*	picnic < football
London	reformulations	2.35 *(1.16)*	1.40 *(1.31)*	picnic < football
	false starts	5.40 *(3.51)*	3.60 *(2.83)*	picnic < football
	replacements	1.85 *(1.49)*	1.30 *(1.03)*	picnic < football
	repetitions	4.00 *(3.50)*	5.50 *(4.18)*	football < picnic

The tables show that *Football* was consistently performed more fluently than *Picnic* by the Teheran cohort across all five measures of repair fluency, and by the London cohort across four of the five measures.

Cite this article: Ryan, J., Foster, P., Wang, Y., Fester, A., & Yap, J. R. (2024). Task design, L1 literacy, and second language oracy: A close replication of Tavakoli and Foster (2008). *Studies in Second Language Acquisition,* 46: 1475–1492. https://doi.org/10.1017/S0272263124000445