Yang *Language Testing in Asia* (2023) 13:22
https://doi.org/10.1186/s40468-023-00234-4

Language Testing in Asia

RESEARCH Open Access

Using eye-tracking and retrospective verbal reports to explore the cognitive processes of banked gap-filling: a case study featuring methodological triangulation

Chengsong Yang[1*]

*Correspondence: ycs@mail.xjtu.edu.cn

[1] School of International Studies, Xi'an Jiaotong University, 28 Xianning West Avenue, Xi'an 710049, China

Abstract

This study made triangulated use of eye-tracking and retrospective verbal reports (RVRs) to compare the banked gap-filling processes of two same-scorers and those of a top scorer and a bottom scorer. The two same-scorers differed in their cognitive effort on global and local processing, fluency of choice making, and strategy use when completing the task and half of the mutual correct responses. Contrary to findings from previous studies, the top scorer exerted much greater effort on global and local processing than the bottom scorer, in her pursuit of perfection, and she reported much less use of syntax analysis as a strategy. The findings of this study increase our understanding of individual differences in the cognitive processes of English-as-a-foreign-language (EFL) reading and test-taking and prove the value of in-depth, multi-faceted process research. Featuring the use of heatmaps, eye-tracking metrics, choice-making graphs, gaze plots, and RVRs, this study also responds to an emergent concern in language assessment about how the enormous process data can be handled effectively. Implications for testing and learning EFL reading are further developed.

Keywords: Eye-tracking, Banked gap-fill tasks, EFL reading processes, Individual differences, Retrospection

Introduction

In language testing research, the importance of process explorations relative to score-based studies has been reiterated. Process information can be indispensable evidence for, or even a truer reflection of, reading proficiency or levels of comprehension, while scores may be a misleading indicator given that "a weak student may choose the appropriate alternative despite faulty reasoning" (Cohen, 1984, p. 71). Process studies also unveil what is measured by a test or an item, as against what is intended to be measured by it (e.g., Weir, 2005). Some process research focuses on individual differences (McCray & Brunfaut, 2018; Ranalli et al., 2018). Individual differences research "examines attributes on which learners vary and how such variations relate to language-learning success," to achieve the purpose of developing differential pedagogy (Skehan, 1991, p.

290). In language testing, such research may additionally offer opportunities for test validation (e.g., McCray & Brunfaut, 2018). In studies on how test-taking processes vary with proficiency, proficiency has been graded commonly based on scores (e.g., Anderson, 1991; McCray & Brunfaut, 2018). While such individual difference studies may be carried on with new approaches and methods, research that uncovers test-taking processes underlying the same scores is especially warranted. This alternative line of research intends to reveal the variety of processes leading to the same scores or correct answers and identify those processes that are unwanted, which may indicate differences in proficiency when it may be taken as equal, and provoke questions about test or item validity as well. The cognitive processes of banked gap-filling appear to be under-researched. The banked gap-fill task is a type of cloze which requires test-takers to restore words missing from the text by making choices from a box or a "bank" that collects all the candidate words for the gaps. It has been adopted in high-stakes English examinations or English proficiency tests, such as Pearson Test of English, Aptis developed by the British Council, China's Test for English Majors Band 4, and China's College English Test Band 4 and Band 6 (CET-4 and CET-6) for non-English majors.

Recently, eye-tracking has been utilized to explore the cognitive processes of test taking, in tests or assessments of reading (e.g., Bax, 2013; Brunfaut & McCray, 2015; Kong, 2019; McCray & Brunfaut, 2018), writing (e.g., Yu et al., 2017), speaking (e.g., Burton, 2022; Lee & Winke, 2018), and listening (e.g., Aryadoust, 2020). It has also been employed to study rating processes (e.g., Ma & Winke, 2022). This process-tracing technique records eye movements (e.g., eye fixations, saccades) in real time. Its application is based on assumptions about the very relationship between eye movements and cognitive processing or attention allocation, for example, "the immediacy assumption" and "the eye-mind assumption" put forward by Just and Carpenter (1980, p. 330–331), or "fixations are considerations", as Conklin and Pellicer-Sánchez recapitulated well (2016, p, 454). Eye-tracking is merited for its non-intrusiveness, millisecond-level fineness of data, and variety of built-in functions. Its value in revealing subtle individual differences, in diagnosis, and in validating tests has been endorsed (e.g., Bax, 2013). However, regarding its use, there have been emerging concerns about how the *enormous* data collected can be handled appropriately for assessment purposes (e.g., Chan, 2018). Its many affordances (e.g., gaze plots and logs of timestamped events, such as mouse clicks), in fact, can be further exploited, to achieve finer, multi-faceted analysis of test-taking processes. Besides, eye-tracking has limitations to overcome, one of which may be that eye movements do not directly show the contents of thoughts. Retrospective verbal reports (RVRs), including stimulated recall, can be used to complement this limitation (e.g., Latif, 2019; Lim, 2020), as they require informants to explicitly tell after completion of a task what they thought, when "there remain retrieval cues in short-term memory that allow effective retrieval of the sequence of thoughts" (Ericsson & Simon, 1993, p. xvi). With better exploitation of eye-tracking and the joint use of RVRs, methodological triangulation can be achieved, whereby findings can be confirmed and understandings may be increased with more comprehensive data. A case study approach has been adopted in eye-tracking studies (e.g., Ranalli et al., 2018). Yu et al. (2017) offer their support for this approach, suggesting, "It is the dynamics and the idiosyncratic nature of each participant's eye movements … that warrant further detailed qualitative analysis

for the purposes of theory building and test validation" (p. 80). A case study is likely to yield new findings regarding individual differences, and it offers space for fine and multi-faceted analyses.

This study exploits the benefits of eye-tracking and RVRs to delve into the banked gap-filling processes of two English-as-a-foreign-language (EFL) test-takers who obtained identical scores and another two who gained extreme scores. It contributes to the understanding of individual differences in the cognitive processes of taking reading tests and may promote triangulated use of eye-tracking and RVRs.

Literature review

Cognitive processes are "any of the mental functions assumed to be involved in the acquisition, storage, interpretation, manipulation, transformation, and use of knowledge", and "encompass such activities as attention, perception, learning, and problem solving" (American Psychological Association, n.d.). Reading involves lower-level processes (e.g., word recognition) and higher-level ones (e.g., constructing text representations), which draw upon the knowledge base while functioning, and are regulated by metacognitive activities (i.e., goal setting, monitoring, and remediation), based on Khalifa and Weir's (2009) socio-cognitive model of cognitive processing in reading. In goal setting, readers select the types of reading they need to engage in to complete a task. *Local reading* will be confined to the sentence and clause level, while more *global reading* will aim at understanding a wider range of text. McCray and Brunfaut (2018) adapted the two goals and made a distinction among *overall processing* (for completion of the whole task), *text processing* (including *global* and *local*), and *task processing* (i.e., bank processing), to better suit the context of eye-tracking research on banked gap-filling. Grabe and Yamashita (2022) also discuss at considerable length cognitive processes that characterize a "fluent" reader. Anderson's (1991) list of 47 reading processing strategies, which comprises supervising strategies, support strategies, paraphrase strategies, strategies for establishing coherence in text, and test-taking strategies, is also an important framework for understanding the series of cognitive processes learners consciously employ to achieve comprehension and task completion while reading and taking reading tests. Specific to cloze tests, Bachman (1985) categorized test-takers' information use for solving closure (e.g., within clause). The cognitive processes of reading may vary with languages (including L1 in the case of L2 reading), reading abilities, social contexts, purposes, and motivation (e.g., Grabe & Yamashita, 2022; Kuperman et al., 2022). The cognitive processes of taking reading tests have been found to differ across proficiency (or score) levels (e.g., Anderson, 1991; Bax, 2013), with cultural knowledge (e.g., Sasaki, 2000), and across tasks (e.g., Brunfaut & McCray, 2015).

Verbal report studies have revealed test-takers' cognitive processes of completing cloze or gap-fill tasks and how such processes may vary with reading abilities or proficiency. These studies show that test-takers employ a variety of strategies, including top-down, bottom-up, and test-wise processing strategies, for comprehension and closing blanks (e.g., Gao & Gu, 2008; Storey, 1997; Yamashita, 2003). They also made gap-filling decisions based on information derived from the clause level, the sentence level, the text level, and outside the text (e.g., Bachman, 1985; Sasaki, 2000; Storey, 1997; Yamashita,

2003). Yamashita, for example, found that EFL learners, regardless of reading skill, referred to text-level information most when completing a gap-filling test examining text-level understanding. Gao and Gu also reported EFL learners' general use of clause-level information and bottom-up processing strategies in CET banked gap-filling. As regards individual differences, learners with better performance are engaged in expected cognitive processes more often: they tend to rely on wider ranges of text, use the same level of information with more correctness, select answers based on understanding and resort less to test-wise strategies (e.g., Gao & Gu, 2008; Storey, 1997; Yamashita, 2003). In Gao and Gu's study on banked gap-filling, high-scoring sophomores made far more correct use of clause-level information and tapped it as supplementary information, while low-achieving learners depended solely on it for judgment. In strategy use, the high group surpassed the low group in the frequency of use of top-down strategies, while the low group more frequently resorted to several test-wise strategies. The group differences reported by these studies need to be verified. Especially, new findings regarding individual differences may arise with the introduction of new methods, such as eye-tracking.

Eye-tracking studies have supported those previous findings on gap-filling processes, but more importantly, developed new insights based on evidence from visual processing (Brunfaut & McCray, 2015; McCray & Brunfaut, 2018). McCray and Brunfaut (2018), for example, confirmed that test-takers at the higher end of performance tended to finish the banked gap-fill task more quickly, rely less on local text, and select answers from options more expeditiously because their study revealed significant negative correlations between scores and three eye-movement measures, namely, total fixation duration (TFD) on task, TFD on three words either side of a gap (Adjacency), and the count of saccades or visual visits to the bank area. They inferred that lower performers' scores might largely reflect lower-level reading abilities, although the tests also tested higher-level processes of reading (Khalifa & Weir, 2009). Again, general findings may not apply to each individual. Methodologically, there is a need for more exploitation of the affordances provided by eye-tracking and for the joint use of RVRs in exploring banked gap-filling.

In individual differences studies, empirical endeavors are rarely seen with a focal interest in uncovering how test-takers with the same scores may differ in cognitive processes, despite the common recognition that scores may be achieved via different ways (e.g., Cohen, 1984). Previous studies, which investigated process differences in relation to score differences, generally categorized learners as proficient or less proficient ones (e.g., Anderson, 1991; Bax, 2013; McCray & Brunfaut, 2018) and tended to treat learners of the same category as approximate or homogeneous in cognitive processes. In the process-based studies, only sporadic examples can be found that suggested how cognitive processes of test taking could differ with regard to the same correct answer. Gao and Gu (2008), for example, provided verbal reports illustrating how Item 9 of the banked gap-fill task (*when an El Niño will strike*) was completed correctly by two high-proficiency learners using sentence-level information (e.g., the testee finished reading the whole sentence and tended to think about it holistically before making the closure) and extra-textual information ("And according to my general knowledge, it is often said that the storm wind strikes somewhere"), respectively, while it was completed wrongly

by a low-proficiency learner, who appeared to make a guess or an intuitive choice ("When an EI Nino will, will come or happen. Yes, "stable". I choose "stable").

The present study

The above literature review reflects the need to explore individual differences in the cognitive processes of banked gap-filling further and to apply triangulation of methods that exploits more benefits of eye-tracking and the advantages of RVRs. This study responds to this need, by selecting two same-scorers and two extreme scorers (a top scorer and a bottom scorer) for a case study. The cognitive processes are compared in terms of cognitive effort, fluency of choice making, and use of information and test-taking strategies for closure. Multiple eye movement statistics are utilized to quantify the attentional resources deployed and assess the cognitive effort made to achieve *overall processing*, *global* and *local text processing*, and *global* and *local bank processing* (Khalifa & Weir, 2009; McCray & Brunfaut, 2018), and thereby to infer the extents of learner engagement with *lower-level* and *higher-level processes* defined in Khalifa and Weir's model. Heatmaps and gaze plots are used to aid judgment of the cognitive effort expended on local reading or processing. Timestamped mouse clicks for making choices are exploited to generate measures counting direct choices, skips over blanks (suspended choices), pauses, and answer changes, to examine the "fluency" of choice making. The concept of "fluency" is introduced, as an echo to Grabe and Yamashita's (2022) notion of "fluency" in reading, and because these measures resemble the fluency measures of speaking and writing performance, which include but are not limited to replacements, hesitations, and pauses for speaking (e.g., Foster & Skehan, 1996) and dysfluencies for writing (e.g., Ellis & Yuan, 2004). Anderson's (1991) list of test-taking strategies and Bachman's (1985) categories of information use for closure are adopted as the coding schemes for counting learners' test-taking strategies and information use for banked gap-filling. Adapted from Nevo's (1989) Multiple-Choice Strategy Checklist, Anderson's list included 18 test-taking strategies, which related primarily to why test-takers reached their answers (e.g., based on understanding, matching, guessing, or elimination) and how they approached test items (e.g., skipping a question and going back to it later). Bachman demarcated four different ranges of information test-takers used for closure: (1) within clause; (2) across clause, within sentence; (3) across sentences, within text; and (4) extra-textual. Test-takers' RVRs are examined against these schemes to decide on the types of strategies and the ranges of information used for closure.

The following research questions are formulated:

1) Do two same-scorers differ in cognitive effort on global and local processing, fluency of choice making, and use of test-taking strategies when completing a banked gap-fill task?
2) How do a top scorer and a bottom scorer differ in cognitive effort on global and local processing, fluency of choice making, and use of information and test-taking strategies for closure when completing a banked gap-fill task?

Methods

Participants

Seventeen university freshmen were recruited from four universities in China's Northwest for this case study, and four of them were finally selected for comparison. The advertisement had required that participants' scores on the National Matriculation English Test (NMET), which was held around two months before their entrance to university, should either surpass 140 or stay below 110 (the full scores being 150), to allow maximal differences to be shown. No participants had attended CET, nor been exposed to the two CET banked gap-fill tasks used (hereafter referred to as "Tower" and "Sarah"). Table 1 gives the descriptive statistics of the 17 participants' NMET scores and task scores.

Purposive sampling was applied. Two high-level learners, DX and FF, both of whom scored six out of ten for the CET-6 task "Sarah", were selected as the same-scorers, for answering Research Question 1. DX and FF were average scorers (group $M = 6.18$) and had four overlapping correct answers, which would be a manageable number for detailed item-by-item analyses. One high-level learner, AG, who achieved full scores for the CET-4 task "Tower", and one low-level learner, BR, who obtained two correct answers for "Tower", were selected as two extreme scorers for comparison, to answer Research Question 2. The contrasts between their cognitive processes of banked gap-filling were found to be most interesting. All these learners were selected also because of their good sampling quality. The bio-data and NMET scores of the four participants are given in Table 2.

Table 1 Participants' NMET and task scores

Participants	*n*	NMET		"Tower"			"Sarah"		
		M	*SD*	Range	*M*	*SD*	Range	*M*	*SD*
High-level	11	144.45	2.70	1–10	7.18	2.86	2–9	6.18	1.99
Low-level	6	101.33	8.71	2–4	2.67	0.82	1–3	2	0.63

Table 2 The biodata and NMET scores of the four participants selected

Participants		Age	Gender	University	Major	NMET
Same-scorers	DX	19	F	State-key	Literature	143
	FF	18	M	State-key	Engineering	144
Extreme scorers	AG	18	F	State-key	Electrics	147
	BR	20	F	Private (3rd-tier)	Education	100

Table 3 Time, scores, and sampling quality

Participants		Time spent			Scores		Sampling rate
		Practice	"Tower"	"Sarah"	"Tower"	"Sarah"	
Same-scorers	DX	1 min 35 s	11 min 56 s	13 min 51 s	4	6[a]	92%
	FF	1 min 52 s	6 min 40 s	8 min 37 s	9	6[a]	87%
Extreme scorers	AG	1 min 31 s	13 min 44 s	18 min 52 s	10[b]	9	94%
	BR	3 min 9 s	7 min 50 s	10 min 40 s	2[b]	3	90%

Note. AG took "Sarah" as the 1st task. [a] or [b] indicates pair comparison

Table 3 presents the time they spent on each step of the experiment, their scores, and the overall eye-tracking sampling quality.

Instruments

The "Tower" task introduced the renovation of the CIS Tower in Manchester with solar panels, and the "Sarah" task related Sarah Josepha Hale's life story and discussed her controversial authorship regarding "Mary's Little Lamb" (see Appendices 1 and 2). The text of "Tower" had 244 words and that of "Sarah" had 268. Both had 15 options in the bank. Two banked gap-fill tasks were chosen to allow balanced observation of the participants' proficiency and because longer experiments would have caused tiredness. Both were authentic CET tests used in June 2018. Developed by The National College English Test Committee of China, the tasks were supposed to have high validity. They were also piloted with nine similar participants thinking aloud. As the scores in Tables 1 and 3 showed, the CET-6 task "Sarah" appeared appropriately difficult for the two same-scorers, as well as for the high-level learners; the CET-4 task "Tower" also could discriminate between the high-level and the low-level learners, especially the top scorer and the bottom scorer selected.

Tobii TX300 was the model of the eye-tracker used to collect data. It is a remote eye-tracker that allows a head movement of 37 cm × 17 cm at 65 cm. Each task was presented on one screen of the eye-tracker with triple line spacing set in HTML, using 18-point Verdana as the font. The choice was made by clicking on the blank, where a built-in window would allow letter symbols to be alternated.

Procedure

Data were collected in an eye-tracking lab, with the author working with one participant at a time. The participants first signed the informed consent form. Then, with their eyes properly tracked, they read the instructions, practiced with a simulated task, and then proceeded to the two tasks. The participants were instructed to achieve their best performance. There was no time constraint, given that in the actual test, only a limit of 40 min is set for the reading part, of which the banked gap-fill component is the first to be completed. To collect RVRs, they were asked to report whatever they had thought when making each response, with a screen recording of their answers replayed to them.

Data analysis

Analyses of cognitive processes based on the affordances of eye-tracking involved making use of heatmaps, eye movement statistics, choice-making graphs, and gaze plots. Hot spots in heatmaps, where fixations aggregated most, were pinpointed, which suggest the greatest cognitive effort made in local reading or processing. TFD, mean fixation duration (MFD, fixation duration per time), visits to the bank, and return visits were adopted as the eye-tracking measures. TFD and MFD in three global areas of interest (AOIs)—Task, Text, and Bank—measured attentional resources and assessed cognitive effort put in overall processing, global text processing, and global bank processing, respectively. TFD on three local AOIs—Sentence (having gaps), Clause (having gaps), and Adjacency—and the corresponding proportional measures (to TFD

on Text) assessed cognitive effort put in different levels of local text processing. Visits to Bank were counted to estimate the cognitive effort made on global bank processing or on local bank processing related to the completion of a specific blank. The gaze plot was generated to visualize the whole path of local processing related to a specific closure and was analyzed to calculate measures of the eye movements involved. The number of return visits was thereby counted. It is the number of visits to the bank area starting from (around) a blank and finally returning to that blank or its adjacency. This can be a meaningful indicator of the cognitive effort made to close a specific blank. Return visits may be associated with gap-filling trials or "matching the blank of an item with the options" (Gao & Gu, 2008, p. 14). Timestamped mouse clicks were marked in the choice-making line graph, where true or false choices and fluency measures that counted direct choices, skips over blanks, pauses during choice making, and answer changes were all shown. Pauses during mouse clicking that were longer than one second, an important threshold for pause measurement (e.g., Foster & Skehan, 1996), and during which participants looked at text outside the blank area (henceforth "while-clicking pauses"), were taken as suggesting hesitation in answering, and were used. Analyses of cognitive processes based on RVRs involved coding test-takers' information use and strategy use for closure.

With Tobii Studio 3.3.2, AOIs were depicted while heatmaps and gaze plots were repeatedly replayed to determine the inclusion and exclusion of fixation points. Additional mouse clicks in choice making with no fixating elsewhere had been removed manually before the calculation of the eye-tracking statistics. The choice-making graphs were depicted in reference to the gaze plots and screen recordings. While-clicking pauses were carefully identified and repeatedly examined. The four participants' RVRs were coded by the author and a lecturer with a PhD in applied linguistics, based on Bachman (1985) and Anderson (1991). The rate of code agreement (to code agreement plus disagreement) reached 96.97% and 87.27% respectively in the second round of coding.

Results

A comparison of the same-scorers DX and FF

DX's and FF's cognitive effort on global and local processing

Figure 1 presents DX's (upper) and FF's (lower) heatmaps for "Sarah". As is illustrated, D most prominently fixated around Blank 1, Blank 3, and Blank 8. In contrast, although FF also focused on Blank 1 and Blank 3, he processed the surrounding text area less intensely, and he displayed another focus on Blank 5. In the bank area, DX exerted strong cognitive effort to process every option, whereas FF showed focal concerns about the lower half of the options only.

Table 4 contrasts the two learners' eye movement statistics. As is shown, DX's TFD and MFD related to overall processing, global text processing, and global bank processing were greater than FF's, and her visits to Bank, which measured global bank processing as well, were also more frequent. She also spent longer TFD on and allocated larger proportions of TFD to the local processing of Sentence, Clause, and Adjacency. The contrast of these statistics suggests that DX expended more cognitive effort on both global and local processing than FF. DX's greater eye-tracking statistics concerning

Fig. 1 DX's (upper) and FF's (lower) heatmaps for "Sarah"

Table 4 DX's and FF's eye movement statistics

	Overall processing	Global text processing	Local text processing			Global bank processing	
	TFD & MFD	TFD & MFD	TFD & its proportions			TFD & MFD	Visits
			Sentence	Clause	Adjacency		
DX	618.21 0.234	395.96 0.221	332.03 83.85%	279.19 70.51%	197.36 49.84%	222.25 0.262	146
FF	361.04 0.201	230.67 0.186	192.16 83.31%	149.21 64.69%	107.48 46.59%	130.37 0.233	81

Bank confirm the highlights of fixation in the bank area of her heatmap. Quantification of eye movements about each blank also helps identify the focus of local processing. For example, DX's three longest TFD on Clause happened to Blank 1, Blank 3, and Blank 8 (68.94, 66.73, and 47.7 s, respectively), which points to the same focuses of local processing as can be observed from her heatmap.

DX's and FF's fluency of choice making

Figure 2 depicts DX's (upper) and FF's (lower) choice-making processes for "Sarah". Each circle or cross sign indicates a correct or wrong choice. In her first round of completion, DX chose Blank 5, Blank 6, Blank 7, and Blank 9 only. A similar range of skipping happened to FF, but he started to choose at Blank 2, and made two fewer skips. DX changed answers five times altogether, three of which happened to Blank 3, while FF made no answer changes. DX made 18 while-clicking pauses while FF made only

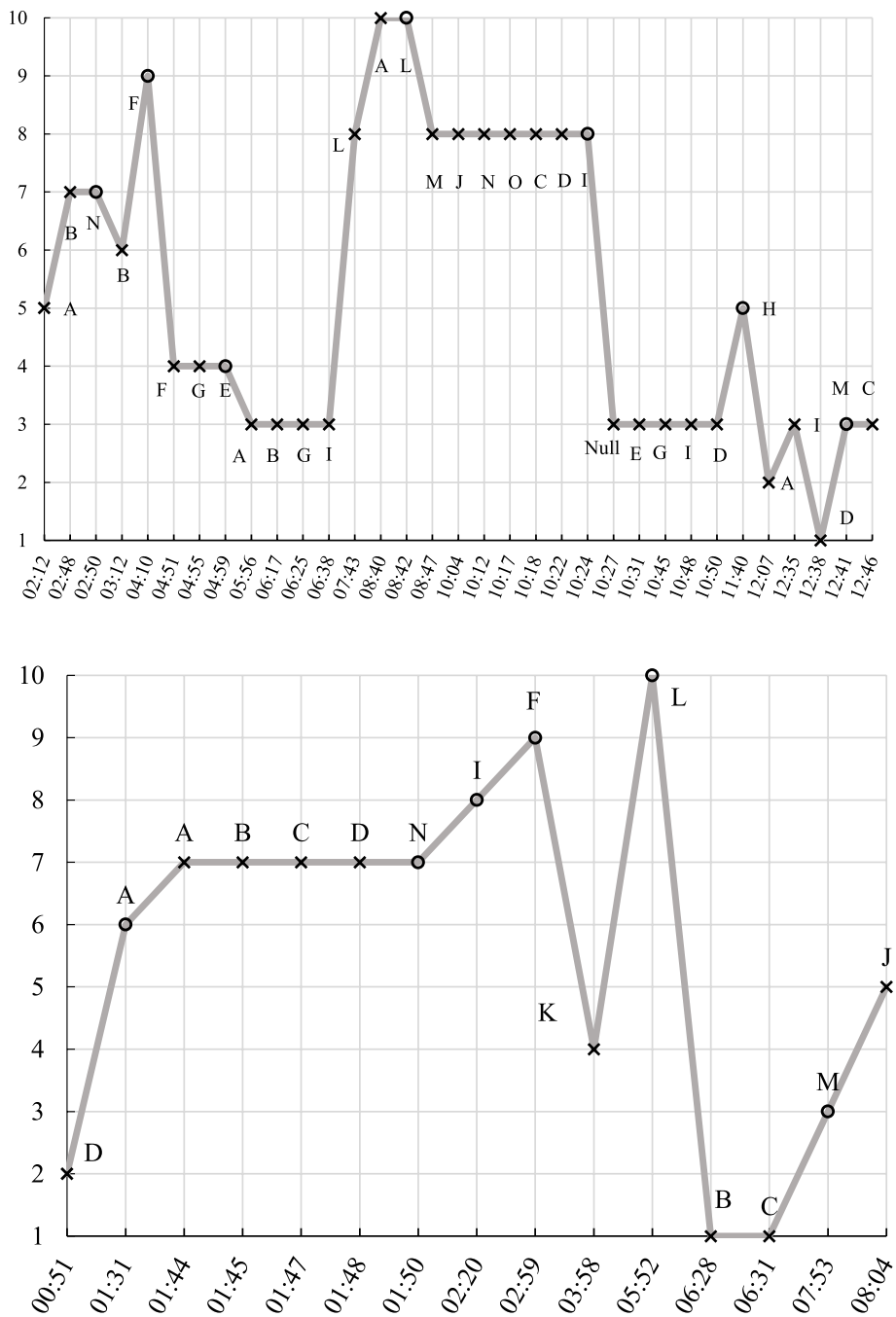

Fig. 2 DX's (upper) and FF's (lower) choice-making processes for "Sarah"

five. Among her six correct choices, DX made only one direct choice without answer changing and/or while-clicking pauses (Blank 9). In contrast, FF reached five of his six correct answers directly, with only Blank 7 causing four while-clicking pauses. The above counts of skips, answer changes, pauses, and direct choices show that DX may have met more difficulty or been more hesitant than FF while making the choices. This finding is supported by their eye-tracking statistics, which show that DX made more cognitive effort. It is also compatible with the observation from their heatmaps that DX processed the bank more intensely to decide on the answers. The choice-making graph also interestingly shows that, among DX's focuses of local text processing observed from her heatmap, Blank 1 and Blank 3 were the last completed, and Blank 8 caused the most pauses.

DX's and FF's cognitive effort and test-taking strategies for four mutual correct answers

DX and FF exerted a similar amount of cognitive effort on local processing, when completing Blank 7 and Blank 10. As an example, Fig. 3 presents their gaze plots for Blank 7, the easiest item (*while celebrating the* <u>*traditional*</u> *festivals*, all correct), which

Fig. 3 DX's (green) and FF's (cyan) gaze plots for Blank 7

visualize their comparable ease in both local text processing and local bank processing, including FF's wider but similarly swift search within the bank.

Table 5 juxtaposes DX's and FF's return visits, TFD on Sentence, and TFD on Bank concerning Blank 7, which were calculated based on their gaze plots. As Table 5 shows, they differed most noticeably in DX's greater attention to the distractor (O) *versatile* only.

DX spent more cognitive effort on local processing than FF when closing Blank 8 and Blank 9. Her struggle to complete Blank 8 (*issued a proclamation*), one of the three blanks attended to most, was especially obvious. Figure 4 gives DX's gaze plots (green) and FF's gaze plot (cyan) for Blank 8. It illustrates DX's three attempts in contrast to FF's single attempt and DX's much greater overall cognitive effort than FF's.

Table 6 further contrasts DX's and FF's eye-tracking metrics that quantified local text processing and local bank processing. In her three attempts, DX paid 20 more return visits altogether and allocated much greater TFD to sentence and bank processing than FF, her diversions to other blanks put aside.

DX and FF achieved three mutual correct answers using different strategies while both completed Blank 9 (*the poem was inspired by a real event*) based on proper understanding. Although DX and FF made a similar cognitive effort to close Blank 7 and Blank 10 (*for the rest of her life*), their strategies were different. DX turned to syntax analyses (i.e., "an adjective" and "fixed expression") that indicated the use of a narrower range of information. FF, however, appeared to rely on his (flawed) understanding of a longer text. He reported that *traditional festivals* would form

Table 5 Cognitive effort on local processing regarding Blank 7

	Return visits, Sentence TFD, and Bank TFD
DX	3 return visits (from N (753 ms), O (1867 ms), and N (480 ms); Sentence TFD (from *to* to *festivals*): 14.77 s; Bank TFD: 3.47 s
FF	2 return visits (from N (663 ms) and later N again (617 ms)); Sentence TFD: 14.35 s; Bank TFD 3.58 s

Note. The sentence TFD was calculated with the blank area excluded

Fig. 4 DX's three attempts and FF's single attempt for Blank 8

Table 6 Cognitive effort on local processing regarding Blank 8

	Return visits, Sentence TFD, and Bank TFD
DX	(With diversions to many other blanks and a transitional choice of L) 22 return visits (from H (314 ms), O (377 ms), and G (1013 ms) in the 1st attempt; from C (407 ms), L (706 ms), G (1790 ms), and L (477 ms, 1177 ms) before choosing it in the 2nd attempt; from H (747 ms), G (720 ms), D (1116 ms), C (721 ms), B (514 ms), C (260 ms, 1773 ms), D (2614 ms), G (1270 ms), I (2592 ms, 600 ms, 396 ms, 550 ms, and 2446 ms) in the 3rd attempt; Sentence TFD: 37.27 s; Bank TFD: 54.87 s
FF	2 return visits (H (640 ms), and I (2282 ms)); Sentence TFD: 6.39 s; Bank TFD: 9.63 s

a contrast with the new way of celebration Sarah advocated. He also mistook "*for the rest of her life*" as "for the purpose of her life in future being less disturbed", considering "the dispute about the figure" in the passage. For Blank 8, both admitted not knowing *proclamation*. With more effort paid and less fluency in making the choice, DX appeared to be able to figure out the (rough) meaning of *proclamation* at last ("… just felt, it should be, erm, should be *meeting*, or something"). In contrast, with relative ease and expeditiousness, FF in fact based his choice on part of speech and elimination ("… felt it like a noun and because I had filled in others").

A comparison of AG as a top scorer and BR as a bottom scorer

AG's and BR's cognitive effort on global and local processing

The heatmaps in Fig. 5 show that AG's local text processing centered around Blank 8 and Blank 1, and her local bank processing around the three *-ed* options (D) *competed*, (E) *constructed*, and (H) *discovered*. In contrast, when processing the text, BR appeared to focus on Blank 7 and on individual words adjacent to each blank (e.g., *renovation*, *cost-efficient*, *get*, and *race*). BR also showed much more concentrations of fixation than AG when processing the bank.

Table 7 shows AG's and BR's eye movement statistics for overall processing, global text processing, local text processing, and global bank processing. The top scorer AG spent longer TFD on the six AOIs and paid more visits to Bank, indicating that she made more cognitive effort to complete the task, process the text both locally and globally, and process the bank. In contrast, BR cast longer MFD on Task, Text, and Bank, which suggests her greater effort made for word processing. BR's longer average fixation within the bank (0.287), together with her large proportion of TFD on Bank to TFD on Task (35.45%, compared with AG's 33.65%), is consistent with the widespread hot pots apparent in the bank area of her heatmap.

Fluency of task completion: choice-making graphs

Figure 6 depicts AG's (upper) and BR's (lower) choice-making graphs. As is shown, the top scorer AG made seven choices in one go and correctly. She changed answers to Blank 1, Blank 4, and Blank 8 only, due to the conflicts of choice. Noticeably, after she closed all the items at 6:16, she engaged in an over-7-min-long checking process, until 13:44, as can be seen via a replay of her screen recording. During this checking, she first restored (E) *constructed* for Blank 1, then changed her original (E) option for Blank 8 (*much less pollution than that caused by energy ____ through fossil*

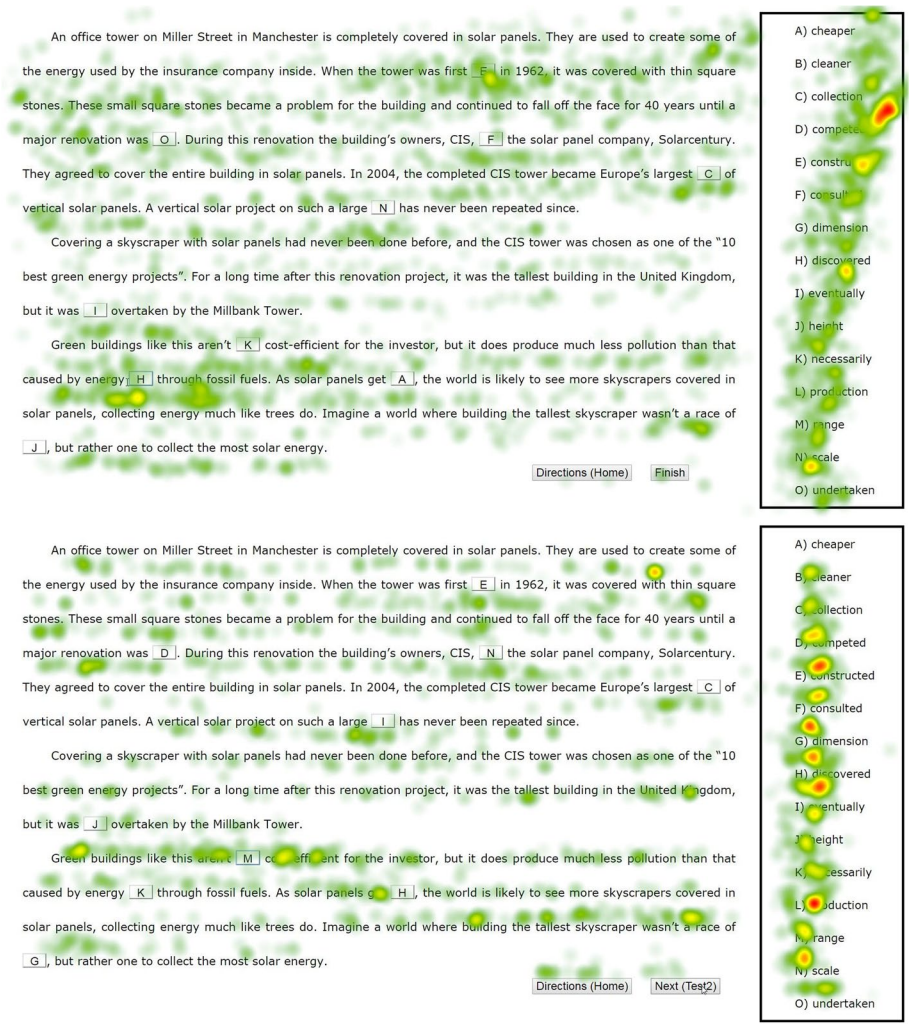

Fig. 5 AG's (upper) and BR's (lower) heatmaps for "Tower"

Table 7 AG's and BR's eye movement indexes

	Overall processing	Global text processing	Local text processing			Global bank processing	
	TFD & MFD	TFD & MFD	TFD & its proportions			TFD & MFD	Visits
			Sentence	Clause	Adjacency		
AG	589.43 0.238	391.1 0.222	343.12 87.73%	276.62 70.73%	164.79 42.14%	198.33 0.274	138
BR	344.39 0.252	222.29 0.236	183.2 82.41%	138.74 62.41%	85.40 38.42%	122.1 0.287	57

fuels) to (H) *discovered*, and, at last, after a 2-min interval, successfully rechanged (H) to (L) *production*, a noun rather than an *-ed* form, for Blank 8. Within the 2 min prior to her click at Finish, she initiated three more rounds of checking. Her much longer checking, during which she solved Blank 8 at last, explained why she showed more cognitive effort in terms of eye-tracking measures. Besides, her choice

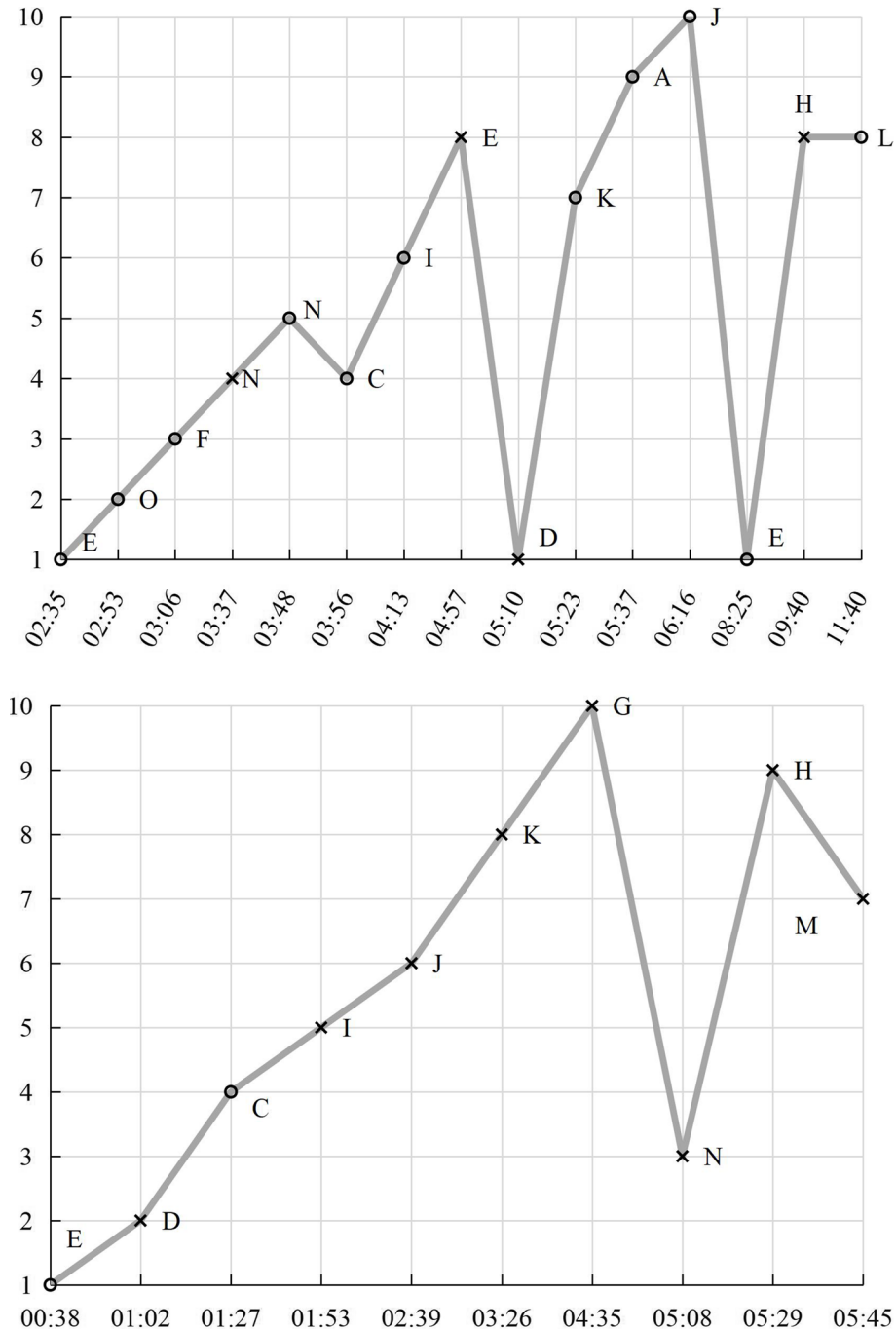

Fig. 6 AG's (upper) and BR's (lower) choice-making graphs for "Tower"

making and viewing to and fro between Blank 1 and Blank 8 confirm the heatmap highlights surrounding the two blanks and the *-ed* options, namely, (D) *competed*, (E) *constructed*, and (H) *discovered*.

BR demonstrated a seemingly "fluent" process. All her choices were first and final. She made a quicker start of choice, at 00:38. In the following 4 min or so, she made another six choices. Then, within 1 min and 10 s, she responded to Blank 3, Blank 9,

and Blank 7, which she had skipped over. BR spent 2 min and 5 s, about a quarter of her task time, checking, but only checked the first three and the last of her choices. Her choice-making "fluency" is consistent with the less cognitive effort she made to complete the task, as her eye movement statistics have shown.

Information and strategy use

Table 8 presents the range of information and strategies that AG and BR relied on to close "Tower", which were coded from their RVRs. AG relied primarily on proper comprehension of the textual information and was able to obtain the meanings of the chosen words in the context. For AG, part-of-speech judgment played an auxiliary role, and she was able to avoid its interference. Initially, AG persisted in searching for an *-ed* verb for Blank 8, "so that it constituted an attributive with what was behind it to modify *energy*". She said that she then looked for all the verbs with *-ed* and found none seem to have the right meaning. She finally turned to the noun *production*. "That's, these pollution, is probably caused by, uh, *energy production*, power generation, through fossil fuels". On the contrary, BR resorted all to grammar (part-of-speech) analyses and demonstrated little understanding of the clausal and sentential information. This explains her less cognitive effort and "fluency".

Discussion

The first research question, which asks if two learners with identical scores differ in cognitive processes, is answered affirmatively. The distributions of hot spots in DX's and FF's heatmaps show that these two same-scorers had different focuses of local text processing and bank processing, with DX appearing to exert strong effort to process every option. Their eye movement statistics further reveal that DX exerted much more cognitive effort on overall processing, local and global text processing, and global bank processing than FF. It is inferred, following McCray and Brunfaut's (2018) interpretations of the eye-tracking metrics and types of processing, that DX may have engaged in more lower-level processes than FF, so that she spent more (proportions of) time processing

Table 8 AG's and BR's information use and strategies for closure

Cognitive processes based on RVRs		AG	BR
Information use	Within clause	10	10
	Across clause, within sentence	6	5
	Across sentences, within text	1	0
	Extra-textual	1	0
Strategies	Select a response based on understanding the material read	10	9(8[a])
	State the meaning of the word chosen	10	4
	Analyze grammar (mostly parts of speech)	3	10(6[b])
	Eliminate other options	1	1
	Match options (with the same part of speech) one by one with the blank	1	2
	Benefit from a previous encounter	1	0
	Collocation/fixed expression	1	0
	Refer to a word with the same form in the following text	0	1
	State failure to know the word chosen	0	1

Note. For Blank 3, Blank 4, and Blank 5, Clause = Sentence; [a] indicates piecemeal understanding or misunderstanding; [b] indicates misjudgments

the local texts, read the whole text part more (words in Sentence constituted 73.36% of those in Text), visited the bank more frequently, engaged in more reading of the options in the bank, and finally spent more time completing the task. DX's longer MFD on both Text and Bank, as well as her heatmap hot spots on Bank, also suggests her more engagement with lower-level processes such as word recognition. Then, the choice-making graphs illustrate DX's less fluency in completing the task. As regards the four mutual correct responses, the gaze plot analyses find that DX exerted greater cognitive effort on local processing than FF to complete two of the four responses, and their RVRs indicate that they employed different strategies to achieve three of these mutual answers. The above findings add empirical evidence to the common recognition that scores may be achieved via different ways, including faulty reasoning and test-wise strategies (e.g., Anderson, 1991; Cohen, 1984; Gao & Gu, 2008). They also point to the limitation of judging proficiency simply by scores. One most likely explanation for the two same-scorers' different cognitive processes, apart from other personal factors, may be that they differed in actual reading proficiency, considering the close relationship between reading abilities and cognitive processes discussed and confirmed in the literature (e.g., Anderson, 1991; Bax, 2013; Grabe & Yamashita, 2022; McCray & Brunfaut, 2018). That FF was more proficient can be inferred from McCray and Brunfaut's (2018) findings of the negative correlations between reading proficiency (scores in general) and TFD on Task, TFD on Adjacency, and visits to Bank, as DX had greater values of these three. DX's lack of fluency in choice making is also indicative of her less proficiency. Still another evidence is that FF performed much better than DX in the other task, "Tower", as is shown in Table 3. Their different cognitive processes (and the same scores) may also be attributed to task or item characteristics. Certain items might allow different cognitive processes, or tolerate unwanted cognitive processes, for their correct closure, causing concerns about their validity (e.g., DX's and FF's different strategies for Blank 7, Blank 8, and Blank 10).

The second research question asks how a top scorer and a bottom scorer may differ. Most noticeably, the eye-tracking measures show that the top scorer AG made more cognitive effort on overall processing, local and global text processing, and global bank processing than the bottom scorer BR, although both appeared similarly "fluent" in choice making. This is contrary to McCray and Brunfaut's (2018) findings and to Bax's (2013) conclusion that successful test-takers would read more expeditiously than unsuccessful ones. AG's much stronger processing effort obviously should not be attributed to her having any difficulty in lower-level processes in general, such as word recognition, as is suggested by her shorter MFD. The major reason, instead, was AG's particular effort to solve Blank 8, in relation to Blank 1, and her longer checking. Her RVRs, heatmaps, and choice-making graphs have revealed the processes of her working hard to tackle Blank 8. Further analysis of AG's and BR's choice-making processes has shown that AG spent 5 min and 23 s more on checking than BR. Such processes might be characteristic of her as a top student (her NMET score was very close to full marks), who is usually more ready to solve problems. In contrast, as BR's RVRs indicate, when encountering a task beyond her ability, BR resorted overwhelmingly to part-of-speech analyses. Such analyses, which represented presumably shallower, narrower-ranging processing of the task text than was involved in AG's use of meaning-based strategies,

should entail less effort. More probably, BR might have refrained from exerting longer effort, because she realized her limits. Her case then may suggest that a more expeditious process may not necessarily relate to more proficiency. Test-takers' psychology, including motivation, should be studied when their cognitive processes are considered (e.g., Grabe & Yamashita, 2022; Ranalli et al., 2018).

The RVRs show that the two extreme performers also differed in information and strategy use, in that the top scorer AG used a wider range of information, achieved a much better understanding, and employed fewer test-wise strategies, than the bottom scorer BR. These differences confirm previous findings on how proficiency is related to or affects cognitive processes (e.g., Gao & Gu, 2008; Yamashita, 2003). For example, AG relied outstandingly on a correct understanding of the texts read and the words chosen, while BR resorted overwhelmingly to grammar (parts of speech) analyses and to flawed understanding. These contrasts are in line with the findings from Gao and Gu's (2008) study, which reported that the learners of higher proficiency utilized information more correctly, decided on answers based more on understanding, and resorted less to test-wise strategies. Noticeably, however, one major difference between this study and Gao and Gu's (2008) is that in this study, the top scorer employed much less grammar analysis than the bottom scorer (3 to 10), whereas in their study, the high-proficiency learners utilized this strategy twice as often as the low-proficiency learners did (47 to 23). The primary reason for this inconsistency may be that the top scorer and the bottom scorer in this study were extreme cases, compared with the six high-proficiency and six low-proficiency learners from an army university. The particularly high achiever AG relied primarily on comprehension and may have developed automatized or quite expeditious syntactic parsing, which fell out of the range of verbal reports (Ericsson & Simon, 1993; Yang, 2019).

The findings in this study are useful for examining construct validity and improving task design. The National College English Test Committee of China (2016) stipulates that the banked gap-fill task "examines students' ability to comprehend and use vocabulary in a text context" (p. 6). The two tasks appeared to have achieved this purpose largely, as the scorers demonstrated different aspects of knowledge of vocabulary (regarding its meaning and use, e.g., part of speech, collocation, meaning inferred from context) while they referred to different ranges of text content. Several concerns, however, may be raised regarding the design of the banked gap-fill tasks. First, it may be worth considering including more items that elicit use of information from broader contexts, as the present items appeared to examine within-sentence text comprehension mainly (e.g., the top scorer's primary use of information at the clause and sentence levels to complete "Tower"). Then, items should be designed with proper degrees of difficulty. The basic requirement for closing Blank 7 in "Sarah" at the Band 6 level, and therefore, its intended purpose of testing, was knowing *traditional* and *festivals* as a "pair", but such formulaic knowledge may have been well formed before university. The economy of DX's and FF's cognitive effort shown in Table 5 and Fig. 3 supports this easiness. Finally, items should be best determined with validation measures to check if desired cognitive processes are elicited. For example, although DX had chosen I) *proclamation* for Blank 3 twice (Fig. 2), she finally kept it for Blank 8 and tried to speculate about its meaning. This indicates that Blank 8 left the opportunity for inference of word meaning

to be tested, for those who did not know the word. FF might as well have achieved a rough idea of what *proclamation* meant, judging from his much more expeditious and earlier choice. However, had he achieved the answer merely by part-of-speech judgment and elimination, as he reported, the item (*issued a proclamation*) would be a weak one. Chances were that the possible answers could be narrowed down with preliminary syntactic knowledge (i.e., a singular noun follows the indefinite article *a*), *proclamation* could be identified with suffix knowledge (i.e., *-tion* indicative of a noun), and other options might be eliminated for not going with *issued*, provided that the verb and the others were known. All these would mean a transfer of what was tested. One advice may be that distractors be expanded (e.g., *conference* for *proclamation*) (see also Gao & Gu, 2008) or words functioning as multiple parts of speech offered.

The four learners' banked gap-filling processes may be exploited for diagnosis and treatment purposes to benefit EFL learners, whom they may be exemplary for or representative of. DX's, FF's, and especially BR's cases indicate the fundamental need to develop abilities to conduct lower-level processes. Word recognition and lexical access, for example, would have facilitated identifying choices (e.g., *proclamation*) and skirting interference options (e.g., *versatile*, *rectified*) to achieve understanding-based closure. Depth of vocabulary knowledge should also be attended to, considering, for example, FF's inaccurate interpretation concerning *for the rest* (relaxation*) *of*. Syntactically, in light of the interference AG as a top student encountered when solving Blank 8, a teaching design on noun-noun phrases should help more learners. Grabe and Yamashita (2022) believe that second language learners need "continual practice" at word recognition, and for this, suggest that reading be expanded and recommend a series of learning activities (p. 52). They also maintain that basic grammar knowledge is very necessary for second language reading development and suggest that advanced grammar be taught or learned only when it appears frequently or hinders understanding.

Conclusion

By exploiting the benefits of eye-tracking and RVRs, this study compares the cognitive processes of two EFL learners who obtained the same scores and those of a top scorer and a bottom scorer when they took banked gap-fill tasks. It finds that the same scores and the same correct answers may be achieved with different cognitive efforts exerted on global and local processing, fluency of choice making, and strategies. This raises concerns about test-takers' real proficiency and about item design. This study also finds that, contrary to findings from previous studies, a high-achieving learner may exert more cognitive effort while a low-achieving learner may spare effort and appear similarly "fluent". This finding calls for attention paid to test-takers' psychology, especially motivation, when their test-taking processes are studied. The study also reveals individual differences in patterns of strategy use that are related to proficiency but are different from previous conclusions and relates such differences to individual peculiarities. Overall, the above findings contribute to the individual differences literature in language testing and assessment and support the long-lasting arguments for process research against the limitations of scores. Methodologically, this study demonstrates how the affordances of eye-tracking, including heatmaps, eye movement statistics, logs of timestamped events, and gaze plots, can be made more use of, and how

eye-tracking may be used jointly with RVRs, to achieve methodological triangulation. These methods and techniques may inform future research. Implications for test design and EFL learning of reading have also been given. Admittedly, to develop sound validity and pedagogical claims, large-scale studies are warranted. Test designers and teachers should also be interviewed, if possible, to understand their concerns and constraints.

Appendix 1

The "Tower" task

An office tower on Miller Street in Manchester is completely covered in solar panels. They are used to create some of the energy used by the insurance company inside. When the tower was first (1)____ in 1962, it was covered with thin square stones. These small square stones became a problem for the building and continued to fall off the face for 40 years until a major renovation was (2)____. During this renovation the building's owners, CIS, (3)____ the solar panel company, Solarcentury. They agreed to cover the entire building in solar panels. In 2004, the completed CIS tower became Europe's largest (4)____ of vertical solar panels. A vertical solar project on such a large (5)____ has never been repeated since.

Covering a skyscraper with solar panels had never been done before, and the CIS tower was chosen as one of the "10 best green energy projects". For a long time after this renovation project, it was the tallest building in the United Kingdom, but it was (6)____ overtaken by the Millbank Tower.

Green buildings like this aren't (7)____ cost-efficient for the investor, but it does produce much less pollution than that caused by energy (8)____ through fossil fuels. As solar panels get (9)____, the world is likely to see more skyscrapers covered in solar panels, collecting energy much like trees do. Imagine a world where building the tallest skyscraper wasn't a race of (10)____, but rather one to collect the most solar energy.

A) Cheaper	I) Eventually
B) Cleaner	J) Height
C) Collection	K) Necessarily
D) Competed	L) Production
E) Constructed	M) Range
F) Consulted	N) Scale
G) Dimension	O) Undertaken
H) Discovered	

Appendix 2

The "Sarah" task

Did Sarah Josepha Hale write "Mary's Little Lamb," the eternal *nursery rhyme* (儿歌) about a girl named Mary with a stubborn lamb? This is still disputed, but it's clear that the woman (1)____ for writing it was one of America's most fascinating (2)____. In honor of the poem's publication on May 24, 1830, here's more about the (3)____ author's life.

Hale wasn't just a writer, she was also a (4)____ social advocate, and she was particularly (5)____ with an ideal New England, which she associated with abundant Thanksgiving

meals that she claimed had "a deep moral influence." She began a nationwide (6)____ to have a national holiday declared that would bring families together while celebrating the (7)____ festivals. In 1863, after 17 years of advocacy including letters to five presidents, Hale got it. President Abraham Lincoln, during the Civil War, issued a (8)____ setting aside the last Thursday in November for the holiday.

The true authorship of "Mary's Little Lamb" is disputed. According to the New England Historical Society, Hale wrote only part of the poem, but claimed authorship. Regardless of the author, it seems that the poem was (9)____ by a real event. When young Mary Sawyer was followed to school by a lamb in 1816, it caused some problems. A bystander named John Roulstone wrote a poem about the event, then, at some point, Hale herself seems to have helped write it. However, if a 1916 piece by her great-niece is to be trusted, Hale claimed for the (10)___ of her life that "some other people pretended that someone else wrote the poem".

A) Campaign	I) Proclamation
B) Career	J) Rectified
C) Characters	K) Reputed
D) Features	L) Rest
E) Fierce	M) Supposed
F) Inspired	N) Traditional
G) Latter	O) Versatile
H) Obsessed	

Abbreviations

RVRs	Retrospective verbal reports
EFL	English as a foreign language
CET-4/6	College English Test Band 4/6
TFD	Total fixation duration
NMET	National Matriculation English Test
MFD	Mean fixation duration
AOIs	Areas of interest

Authors' contributions

The author is the sole contributor to this article and the work involved in it. The author read and approved the final manuscript.

Funding

This research has received grants from The Ministry of Education, China, as part of its Planned Humanities and Social Sciences Project 19YJA740070.

Availability of data and materials

All data and materials have been presented in or as appendixes to the article.

Declarations

Ethics approval and consent to participate

The author declares that this study has protected the interest of the human participants by complying with the country, provincial, and local regulations and with the relevant codes established by the professional groups; risks to the human participants have been minimized, and the risks are reasonable given the expected benefits. The author/researcher has informed participants of the key elements of the study protocol. The author declares that the privacy of participants and the confidentiality of data have been maintained.

Consent for publication

This study has gained the informed consent of all participants.

Competing interests

The author declares no competing interests.

Received: 22 November 2022 Accepted: 8 April 2023
Published online: 28 April 2023

References

American Psychological Association. (n.d.). Cognitive process. In *APA dictionary of psychology*. Retrieved February 25, 2023, from https://dictionary.apa.org/cognitive-process

Anderson, N. J. (1991). Individual differences in strategy use in second language reading and testing. *The Modern Language Journal, 75*(4), 460–472. https://doi.org/10.1111/j.1540-4781.1991.tb05384.x

Aryadoust, V. (2020). Dynamics of item reading and answer changing in two hearings in a computerized while-listening performance test: An eye-tracking study. *Computer Assisted Language Learning, 33*(5–6), 510–537. https://doi.org/10.1080/09588221.2019.1574267

Bachman, L. F. (1985). Performance on cloze tests with fixed-ratio and rational deletions. *TESOL Quarterly, 19*(3), 535–556. https://doi.org/10.2307/3586277

Bax, S. (2013). The cognitive processing of candidates during reading tests: Evidence from eye-tracking. *Language Testing, 30*(4), 441–465. https://doi.org/10.1177/0265532212473244

Brunfaut, T., & McCray, G. (2015). *Looking into test-takers' cognitive processes while completing reading tasks: A mixed-method eye-tracking and stimulated recall study* (ARAGs Research Reports Online AR-G/2015/001). British Council. https://www.britishcouncil.org/sites/default/files/brunfaut_and_mccray_report_final.pdf

Burton, J. D. (2022). Gazing into cognition: Eye behavior in online L2 speaking tests. *Language Assessment Quarterly*. Advance online publication. https://doi.org/10.1080/15434303.2022.2143680

Chan, S. (2018). Paper-based vs computer-based writing assessment: Divergent, equivalent or complementary? *Assessing Writing, 36*, 1–2. https://doi.org/10.1016/j.asw.2018.04.001

Cohen, A. D. (1984). On taking language tests: What the students report. *Language Testing, 1*(1), 70–81. https://doi.org/10.1177/026553228400100106

Conklin, K., & Pellicer-Sánchez, A. (2016). Using eye-tracking in applied linguistics and second language research. *Second Language Research, 32*(3), 453–467. https://doi.org/10.1177/0267658316637401

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

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (Rev. ed.). The MIT Press. https://doi.org/10.7551/mitpress/5657.001.0001

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

Gao, X. Y., & Gu, X. D. (2008). An introspective study on test-taking process for banked cloze. *Chinese Journal of Applied Linguistics, 31*(4), 3–16.

Grabe, W., & Yamashita, J. (2022). *Reading in a second language: Moving from theory to practice*. Cambridge University Press. https://doi.org/10.1017/9781108878944

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review, 87*(4), 329–354. https://doi.org/10.1037/0033-295X.87.4.329

Khalifa, H., & Weir, C. J. (2009). *Examining reading: Research and practice in assessing second language reading*. Cambridge University Press.

Kong, J. F. (2019). *Investigating the role of test methods in testing reading comprehension: A process-focused perspective*. Springer. https://doi.org/10.1007/978-981-13-7021-2

Kuperman, V., Siegelman, N., Schroeder, S., Acartürk, C., Alexeeva, S., Amenta, S., Bertram, R., Bonandrini, R., Brysbaert, M., Chernova, D., Da Fonseca, S. M., Dirix, N., Duyck, W., Fella, A., Frost, R., Gattei, C. A., Kalaitzi, A., Lõo, K., Marelli, M., ... Usal, K. A. (2022). Text reading in English as a second language: Evidence from the multilingual eye-movements corpus. *Studies in Second Language Acquisition*. Advance online publication. https://doi.org/10.1017/S0272263121000954

Latif, M. M. A. (2019). Eye-tracking in recent L2 learner process research: A review of areas, issues, and methodological approaches. *System, 83*, 25–35. https://doi.org/10.1016/j.system.2019.02.008

Lee, S., & Winke, P. (2018). Young learners' response processes when taking computerized tasks for speaking assessment. *Language Testing, 35*(2), 239–269. https://doi.org/10.1177/0265532217704009

Lim, H. (2020). Exploring the validity evidence of a high-stake, second language reading test: An eye-tracking study. *Language Testing in Asia, 10*, Article 14. https://doi.org/10.1186/s40468-020-00107-0

Ma, W. Y., & Winke, P. (2022). An investigation of the impact of jagged profile on L2 speaking test ratings: Evidence from rating and eye-tracking data. *Language Assessment Quarterly, 19*(4), 394–421. https://doi.org/10.1080/15434303.2022.2078720

McCray, G., & Brunfaut, T. (2018). Investigating the construct measured by banked gap-fill items: Evidence from eye-tracking. *Language Testing, 35*(1), 51–73. https://doi.org/10.1177/0265532216677105

Nevo, N. (1989). Test-taking strategies on a multiple-choice test of reading comprehension. *Language Testing, 6*(2), 199–215. https://doi.org/10.1177/026553228900600206

Ranalli, J., Feng, H.-H., & Chukharev-Hudilainen, E. (2018). Exploring the potential of process-tracing technologies to support assessment for learning of L2 writing. *Assessing Writing, 36*, 77–89. https://doi.org/10.1016/j.asw.2018.03.007

Sasaki, M. (2000). Effects of cultural schemata on students' test-taking processes for cloze tests: A multiple data source approach. *Language Testing, 17*, 85–114. https://doi.org/10.1177/026553220001700104

Skehan, P. (1991). Individual differences in second language learning. *Studies in Second Language Acquisition, 13*, 275–298. https://doi.org/10.1017/S0272263100009979
Storey, P. (1997). Examining the test-taking process: A cognitive perspective on the discourse cloze test. *Language Testing, 14*(2), 214–231. https://doi.org/10.1177/026553229701400205
The National College English Test Committee of China. (2016). *The National College English Test Band 4 and Band 6 examination outline* [Brochure]. http://cet.neea.edu.cn/html1/folder/16113/1588-1.htm
Weir, C. J. (2005). *Language testing and validation*. Palgrave Macmillan. https://doi.org/10.1057/9780230514577
Yamashita, J. (2003). Processes of taking a gap-filling test: Comparison of skilled and less skilled EFL readers. *Language Testing, 20*(3), 267–293. https://doi.org/10.1191/0265532203lt257oa
Yang, C. S. (2019). The veridicality of think-aloud protocols and the complementary roles of retrospective verbal reports: A study from EFL writing. *The Asia-Pacific Education Researcher, 28*(6), 531–541. https://doi.org/10.1007/s40299-019-00453-5
Yu, G. X., He, L. Z., & Isaacs, T. (2017). *The cognitive processes of taking IELTS academic writing task 1: An eye-tracking study* (IELTS Research Reports Online Series 2017/2). British Council, IDP: IELTS Australia, & Cambridge English Language Assessment. https://www.ielts.org/-/media/research-reports/ielts_online_rr_2017-2.ashx

Publisher's Note