




## The Fluctuating Effect of Thinking on Language Performance: New Evidence for the Island Ridge Curve

Yuyang Cai & Huilin Chen


To cite this article: Yuyang Cai & Huilin Chen (2022): The Fluctuating Effect of Thinking on Language Performance: New Evidence for the Island Ridge Curve, Language Assessment Quarterly, DOI: [10.1080/15434303.2022.2080553](https://doi.org/10.1080/15434303.2022.2080553)

To link to this article: <https://doi.org/10.1080/15434303.2022.2080553>

 View supplementary material 

 Published online: 12 Jun 2022.



 Submit your article to this journal 

 View related articles 

 View Crossmark data 



# The Fluctuating Effect of Thinking on Language Performance: New Evidence for the Island Ridge Curve

Yuyang Cai <sup>a</sup> and Huilin Chen <sup>b</sup>

<sup>a</sup>Shanghai University of International Business and Economics, Songjiang District, Shanghai, China; <sup>b</sup>Shanghai International Studies University, Hongkou District, Shanghai, China



## ABSTRACT


Thinking skills play a critical role in determining language performance. Recent advancement in cognitive diagnostic modelling (CDM) provides a powerful tool for obtaining fine-grained information regarding these thinking skills during reading. Studies are scant, however, exploring the relations between thinking skills and language performance, not to mention studies examining the variation of this association with language proficiency. The current study explored this variation through the lens of the Island Ridge Curve (IRC). Drawing on an English reading test data by 2,285 students, we identified five thinking skills using CDM. Next, we followed guidelines of IRC and put students into four language proficiency groups to examine the relations of each skill identified through reading tasks to language performance across groups. Results of multi-group path analysis showed the effect of each skill identified through reading test fluctuated in the pattern of the IRC. The potential of IRC for examining the moderation of language proficiency on language factors is discussed.

思维技能对语言表现起着至关重要的作用。近年来认知诊断模型(CDM)的发展为通过阅读表现来精确测量这些思维技能提供了有效工具。然而,现有认知诊断研究却很少关注思维技能同语言表现之间的关系,更不用说探究这种关联在不同语言水平学习者中如何变化。本研究从“岛脊曲线理论”(IRC)视角对这种关联在不同语言水平学习者中的变化进行探索。基于2285名考生的英语阅读考试数据,我们用认知诊断的方法识别了五种思维技能。接着,我们参照IRC指导原则,依据学生的综合语言成绩,将学生从低到高分成四个能力组,并探索上述思维技能对综合语言能力贡献在四个水平组中的变化情况。通过多组路径分析方法,我们发现这些思维技能对综合语言表现的作用随着语言水平从低到高呈岛脊曲线状波动。本文讨论了将IRC理论和研究方法用于检验语言能力核心要素对语言能力贡献随语言水平不断提升而波动变化的重要意义和前景。

## Introduction

Thinking skills are critical to language performance (Perfetti et al., 2005; Bennett et al., 2016; Cai & Cheung, 2021; Grabe, 2009a; Perfetti et al., 2008). Fluent language performance requires resources such as linguistic base, prior world knowledge, working memory, and a set of thinking skills (Grabe, 2009b; Jeon & Yamashita, 2014; Luebke & Lorie,

**CONTACT** Huilin Chen  [chlwilliam@hotmail.com](mailto:chlwilliam@hotmail.com)  School of Education, Shanghai International Studies University, 550 Dalian Rd(W), Hongkou District, Shanghai, China, 200083

 Supplemental data for this article can be accessed online at <https://doi.org/10.1080/15434303.2022.2080553>

© 2022 Taylor & Francis

2013). A large volume of research has been devoted to the identification of these thinking skills through reading tests. Much of this strand of endeavours has gained insights from Bloom's taxonomy of Educational Objectives (Anderson et al., 2001; Bloom, 1974). The latest version of the taxonomy comprises six levels of thinking: remembering, understanding, applying, analyzing, evaluating, and creating. Drawing on this taxonomy, different lists of reading subskills have been developed. Adams-Smith (1981) put forward a menu consisting of seven subskills almost identical to Bloom's taxonomy: memory, translation, interpretation, application, analysis, synthesis, and evaluation; Alderson and Lukmani (1989) proposed a longer list with eight subskills: recognition of words, identification, discrimination, analysis, interpretation, inference, synthesis, and evaluation. More recently, Luebke and Lorié (2013) shortened the list into four categories: recognition, understanding and analyzing, inference, and application. An overview of these lists suggests that, while Bloom's taxonomy provides a useful scaffold for distinguishing the subtle skills underlying reading comprehension, the taxonomy is more of a guideline than a menu operationalizable for coding real reading tasks. Depending on the requirement of the language curriculum and the purpose of a particular reading task, practitioners and researchers need to develop their own working list, especially when doing post hoc coding with language performance data.

The advancement of psychometrics in cognitive diagnostic modeling (CDM; Leighton & Gierl, 2007) during the past decades has encouraged even more vibrant efforts in this strand of inquiries (H. Chen & Chen, 2016a, 2016b; Jang, 2009; Kim, 2015; Lee & Sawaki, 2009; Ravand & Robitzsch, 2018; Von Davier, 2008). Most of these studies used existing assessment data and have identified a large number of subskills across readers of different characteristics. Some of the most frequently addressed subskills include recognition, summarizing (obtaining the main ideas), interpretation, inferring, and evaluating (Kim, 2015; Lee & Sawaki, 2009; Pearson & Raphael, 1990), all of them involving thinking skills.

Regardless of this promising trend, existing studies have been underlined by a common belief that, once identified, a particular subskill functions in the way and with the same magnitude with all students of different characteristics. These findings, however, contradict others in that there are some students who excel but do not seem to think much (Perfetti et al., 2008). More fine-grained studies are needed to explore the functions of these thinking skills by students of different characteristics, in different contexts, and perhaps during different stages of language learning (Bronfenbrenner, 1979).

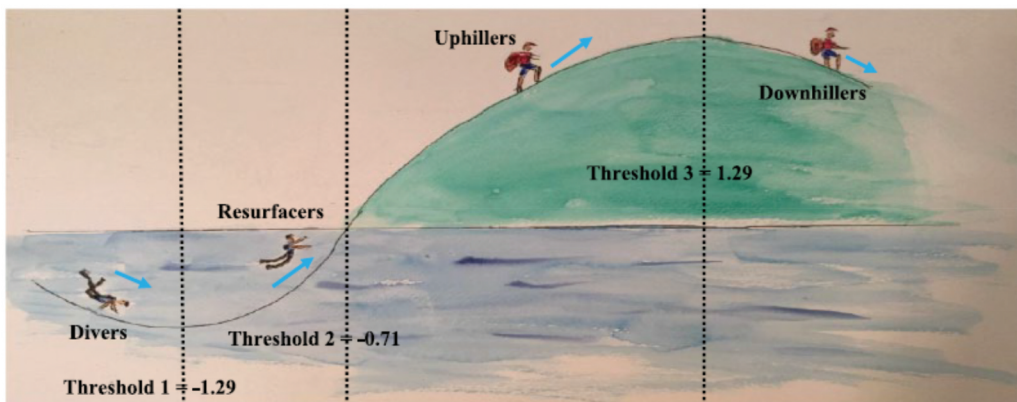
The focus of the current examination was to explore whether the effects of different thinking skills (embodied in reading subskills) on second language (L2) performance are moderated by students' L2 proficiency, or, whether the effects of different thinking skills on language performance vary across different levels of language proficiency. We framed our examination with the emerging theory of the Island Ridge Curve (IRC) in language testing research (Cai, 2020; Cai & Kunnan, 2019, 2020; Wang et al., 2021). The IRC has been developed based on the assumption of interaction in language use (Bachman & Palmer, 1996; Douglas, 2000) and on the threshold hypothesis in language research (Clapham, 1996). The former raised attention to the phenomenon that fluent language performance not only depends on a good mastery of key components of language competence (i.e., linguistic knowledge, strategy, and perhaps prior knowledge) but also on the interaction among them and with other contextual factors. Cai and Kunnan (2019) tested the

interaction between L2 proficiency and background knowledge, concluding that the effect of background knowledge on reading is largest with students of medium-L2 proficiency, but becomes smaller with students of low- or high-L2 proficiency.

Inspired by precedents, Cai (2020) proposed that the effects of non-linguistic components (i.e., strategic competence and background knowledge) on reading performance fluctuate as language proficiency increases. Drawing on a sample of 1491 nurse students' responses to a medical English reading test, a language knowledge test, and a strategic competence questionnaire, Cai and Kunnan (2020) conducted a Multi-Layered Moderation Analysis (MLMA) which allowed the variation of strategic competence effect on reading performance along the whole continuum of students' language proficiency. In the end, they found the effect of strategic competence on reading fluctuated in the 'down-up-down' pattern, which the authors metaphorically labelled as the Island Ridge Curve (IRC). To facilitate communication, the original diagram of the IRC was copied here (see Figure 1) with the authorization from the copyright holder:

The authors found that in the IRC there were three critical linguistic language thresholds (thetas =  $-1.29$ ,  $-0.71$ , and  $1.29$  standard units) that divided students into four groups: the divers (strategic competence effect was negative and gradually became larger as language proficiency increased), the resurfacers (strategic competence effect was negative but gradually became smaller), the uphillers (strategic competence was positive and continuously increased to its maximum), and the downhillers (strategic competence effect was positive and started to step down from its peak). The IRC has also found supporting evidence with motivation regulation strategies in academic English writing (Wang et al., 2021).

To explore the potential of the IRC for uncovering the mechanism in which thinking skills influence language performance, we stepped further and hypothesized that thinking skills affect language performance and the effect fluctuate as students' language proficiency increases. We expect that the (dis-)verification of this hypothesis can deepen our understanding of the role of thinking skills in determining language performance. The study was led by two questions:



**Figure 1.** A metaphoric illustration of the island ridge curve (IRC; Cai & Kunnan, 2020, p. 296). \*Authorization obtained from the copyright holder the Sage Publisher.

- (1) How does language proficiency moderate the prevalence of thinking skills during language performance?
- (2) How does language proficiency moderate the effect of thinking skills on language performance?

## Method

### *Data*

The current study used the Test for English Majors Band 4 (TEM-4) in China to seek answers to our questions. TEM4 is a compulsory exam for all English majors in China in the higher education system. The test is administered to sophomores to assess whether they have met the halfway objectives of the national curriculum for English majors. TEM-4 is roughly at the B2 level of the Common European Framework of Reference for Languages (CEFR; Liu & Wu, 2019; Yang & Liu, 2019). The whole dataset consisted of 236,586 test-takers from universities nationwide. The TEM Examination Board authorized us to a random sample of 2,285 test-takers (about 1% of the total) for research purposes only. For ethical concerns, the agency omitted all demographic information of the participants and only provided item-level response data for the reading subtest and the total raw scores combining the four TEM-4 subtests: vocabulary, grammar, reading, and listening.

### *Measures*

#### *Measure of thinking skills*

Thinking skills were captured using the TEM-4 reading subtest tasks. The TEM-4 reading subtest contained four passages accompanied by 20 multiple-choice items. Each text addresses one of the following topics respectively: mobile phones and human behaviours, social mobility in Britain, computers as human companions, and an excerpt of the novel *Jane Eyre*. The average text length is around 400 words and the average Flesch Kincaid Grade Level Readability (Kincaid et al., 1975) is 8.8, a level suitable for Grade 8 to Grade 9 native English speakers. The reading subskills explicitly required by TEM-4 include grasping the general idea (summarizing), understanding the facts and details (recognizing), judging and reasoning (evaluating), and understanding the logic of the context. This demand on thinking skills is consistent with the Syllabus for English Majors (English Major Division of National Foreign Languages Advisory Board, 2000).

#### *Measure of language proficiency*

Language proficiency was represented by the total raw scores combining the four TEM-4 subtests: The Vocabulary Subtest, the Grammar Subtest, the Listening Subtest, and the Reading Subtest described above. The TEM4 Vocabulary Subtest contained 15 dichotomously scored multiple-choice items assessing the knowledge of distinguishing words from similar spellings, identifying the subtle differences in word meanings, and recognizing appropriate collocations in context. The Grammar Subtest comprised 15 dichotomously scored multiple-choice items covering subject-predicate agreement, tense, voice, mood, modal auxiliary, and complex sentence. The Listening Subtest included three sections: dialogue listening, lecture listening, and news listening, with 10 items for each section. All

items were dichotomously scored. Dialogue listening dealt with daily conversation full of repetitions, redundancies, interruptions, pauses, and simple and unfinished sentences. Lecture listening was in the form of a monologue based on written notes delivered for the purpose of describing academic topics. News listening was a prewritten edited monologue, delivered for the purpose of reporting social and political events.

### **Data analysis**

Our primary data analysis involved two stages: cognitive diagnostic modeling (CDM) to identify thinking skills underlying each reading question item, and multi-group path analysis to explore the variation of the association between these thinking subskills and language proficiency across students of different language proficiency levels.

#### **Cognitive diagnostic modeling (CDM)**

The whole procedure of CDM analysis consisted of four major steps: deciding a working list of reading subskills to capture thinking skills, constructing the Q-Matrix, fitting the CDM model, and estimating person parameters.

*Deciding on the reading skills.* We mainly referred to Bloom's Taxonomy of Educational Objectives (Anderson et al., 2001; Bloom, 1974) and kept open to subskills identified in previous studies. We invited five instructors teaching English majors (three full-time teachers with doctoral degrees in Applied Linguistics and two doctoral students studying English Linguistics) and familiarized them with the Bloom's Taxonomy of thinking skills and a list of reading subskills obtained from precedents (e.g., Adams-Smith, 1981; Lueke & Lorié, 2013). Next, we facilitated the coders to read carefully the TEM-4 test specifications, and TEM-4 reading subtest items, and then reached an agreement on a working list of reading subskills. As a result, five categories of reading subskills were decided by the panel, namely, recognizing, summarizing, interpreting, inferring, and evaluating. Table 1 presents detailed definitions for each category. These categories are intended to represent a hierarchy of reading subskills, with the succeeding one representing a higher cognitive level than its precedent. According to this scheme, recognizing and summarizing can be taken as lower-order skills, interpreting as middle-order, and inferring and evaluating as higher-order skills (Anderson et al., 2001).

*Constructing the Q-Matrix.* Q-Matrix is a table listing all reading subskills attempted by each of the reading test items. In doing so, we first invited the five judges to code the reading items independently according to the five subskills established earlier. When coding

**Table 1.** Thinking skills adopted for Q-matrix construction.

Thinking skills (Bloom's skills)	Definition
Recognizing (remembering)	Identifying details in a text by recalling the information explicitly stated in the reading items.
Summarizing (understanding)	Forming a global understanding of a paragraph and the whole text.
Interpreting (understanding)	Clarifying complex ideas or configurations and interpreting relationships by comparing, transposing, and giving descriptions.
Inferring (analyzing)	Making inference of implicit information according to context.
Evaluating (evaluating)	Compiling information together in a different way by combining elements in a new pattern or proposing alternative solutions.



disagreement occurred, we conducted an inter-rater agreement survey. A coding was decided as valid if more than half of the five judges reached an agreement. Based on that method, we established a coding matrix with an average agreement percentage of 85.8% (over 4 out of 5). We again sent the coding matrix to the coders to inquire about their opinions of the coding matrix and until all finally agreed on it. The final Q-Matrix is shown in Supplementary Table 1 and some other coding examples are shown in Supplementary Table 2.

*Fitting the CDM model.* Quite a few CDM models are available for undertaking diagnosing purposes (Ravand & Baghaei, 2020). Among them, the G-DINA model (De La Torre, 2011) was used for the current study for its excellent performance with reading assessment (Min & He, 2022; Ravand & Robitzsch, 2018). For a brief technical introduction to the G-DINA model please see Supplementary Material 2. The analysis was carried out on the R-based G-DINA package (Ma & de la Torre, 2020). Model-data fit was evaluated based on two statistics:  $z_r$  (the standardized residual between the observed and predicted Fisher transformed correlations between an item pair), and  $z_l$  (the standardized residual between the observed and predicted log-odds ratios of an item pair). A good model fit is suggested if the maximum  $z_r$  and  $z_l$  statistics are smaller than the Bonferroni adjusted critical z-score  $z_c$  at a certain significance level (Chen et al., 2013). For our study, the residuals are smaller than the Bonferroni adjusted critical z-score  $z_c$  at the cutoff significance level ( $p = .05$ ), suggesting the validity of using the Q-Matrix.

*Person parameter estimates.* Upon the decision of the model quality, person parameters were computed. This was to calculate each individual's probability of mastering each subskill coded in the Q-Matrix. This person parameter represented the extent to which an individual mastered the specified subskill (prevalence of thinking skills). Supplementary Table 3 presents the estimates of the first 23 participants in our working data (1% of the total sample). These CDM scores represented the levels of the five thinking skills that students developed captured by the TEM-4 reading subtest.

### **Multiple-group path analysis**

To explore the variation of thinking skill effects on language performance across different language proficiency levels, we first put students into four groups using three cut-off points ( $-1.29$ ,  $-0.71$ , and  $1.29$  standard units of reading score) suggested in the original IRC (see, Cai & Kunnan, 2020). From low to high, the four groups were labelled as struggling learners, low-proficiency learners, medium-proficiency learners, and high-proficiency learners, each corresponding to the divers, resurfacers, uphillers, and downhillers in the original IRC model, respectively (Cai & Kunnan, 2020). We then conducted multi-group path analysis on *Mplus* 8.5 (Muthén & Muthén, 1998–2020) by regressing the overall language performance score on each of the five thinking skills.

## **Results**

### **Prevalence of thinking skills across language proficiency groups**

Our grouping treatment divided students into four unevenly distributed groups: struggling learners (mean of language proficiency  $M = 25.74$  out of 80 points, or 32% of the total), low-proficiency learners ( $M = 34.65$ , or 43%), medium-proficiency learners ( $M = 48.31$ , or 60%), and high-proficiency learners ( $M = 61.50$ , or 77%).

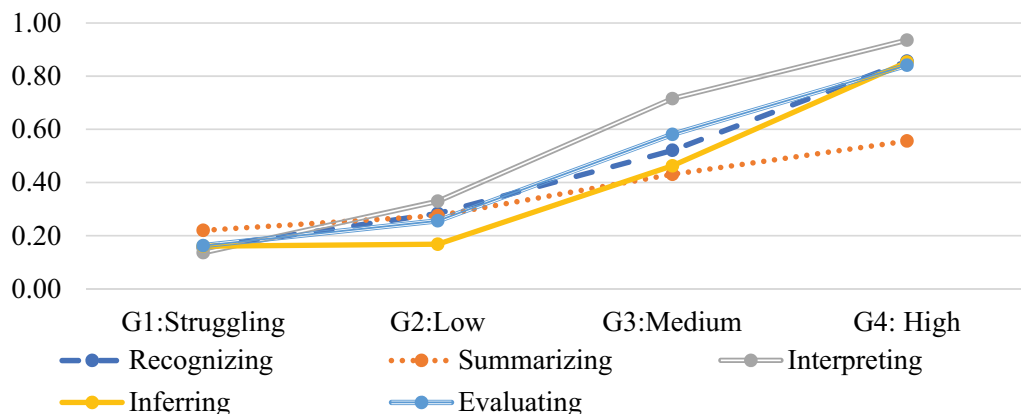
Table 2 shows the means of the CDM scores representing the five thinking skills by groups. Groups on the left end of language proficiency obtained the lowest scores in all skills. These scores then climbed up all the way along with high-proficiency learners. This variation suggested a linear relation in general between CDM scores and language proficiency.

Figure 2 illustrates the between-group change in each specific skill. All changes were small between the first two groups at the left end of language proficiency, which was especially so with the inferring skill. Starting from the second group (low-proficiency group), the changes in all skills accelerated across all other groups. Among these changes, the accelerations with inferring and summarizing were the most, and least salient, respectively.

Figure 2 also presents a dynamic illustration of the relative importance of each subskill across groups. First, the recognizing skill remained relatively steady across all groups, suggesting even prevalence of this skill and independency of language proficiency. Second, summarizing ranked the highest with struggling learners while its relative importance decreased as language proficiency moved up, indicating decreasing deployment of summarizing skill as learning proficiency continued to increase. The third skill interpreting, on the other hand, functioned in quite an opposite direction as did summarizing, indicating the increase in the deployment of this skill by students in higher language proficiency groups. A similar but relatively smaller reversing function was observed with evaluating. The fifth skill inferring consistently ranked low, until it surpassed summarizing in the last group.

**Table 2.** Means of the CDM scores representing thinking skills across groups (N = 2,285).

Group label (Mean total score)	Struggling Learners (M = 25.74)	Low-Proficiency Learners (M = 34.65)	Medium-Proficiency Learners (M = 48.31)	High-Proficiency Learners (M = 61.50)
Group size	295	246	1,559	185
Recognizing	0.15	0.28	0.52	0.86
Summarizing	0.22	0.28	0.43	0.56
Interpreting	0.14	0.33	0.72	0.94
Inferring	0.16	0.17	0.46	0.85
Evaluating	0.16	0.26	0.58	0.84



**Figure 2.** Prevalence of thinking skills across groups.

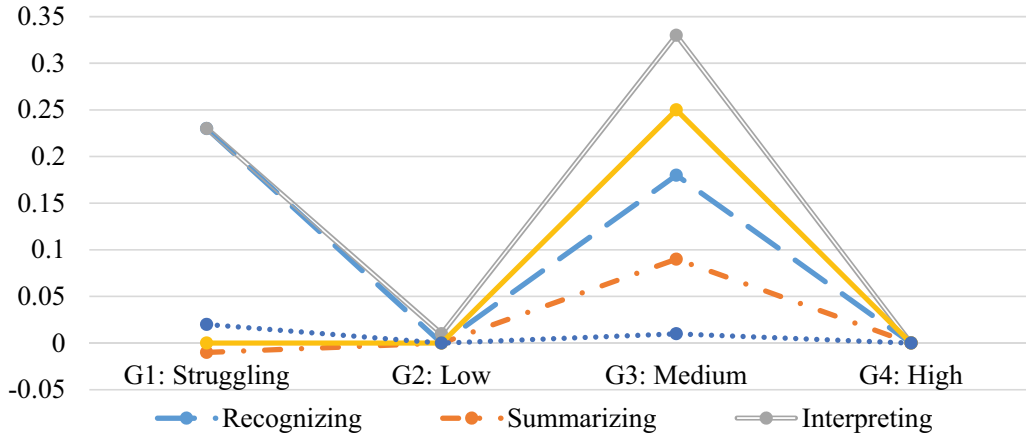
\* Inferring \*\*Evaluating



**Table 3.** Effects of thinking skills on language performance across groups.

Groups\Skills	Recognizing	Summarizing	Interpreting	Inferring	Evaluating
G1: Struggling (Divers)	0.23*	-.03 (p = .572)	0.23*	0.09(p = .354)	0.02 (p = .975)
G2: Low (Resurfacers)	0.04(p = .586)	0.03(p = .734)	0.15(p = .064)	-.06(p = .612)	0.02(p = .896)
G3: Medium (Uphillers)	0.18*	0.09*	0.33*	0.25*	0.01(p = .814)
G4: High (Downhillers)	-.01(p = .943)	-.09(p = .224)	0.01(p = .888)	0.26(p = .104)	0.04(p = .699)

\* All coefficient estimates were significant at  $p < .01$ , otherwise, the  $p$  values are provided in brackets. G1 to G4 = Groups 1 to 4, corresponding to the Divers, Resurfacers, Uphillers and Downhillers in Cai and Kunnan (2020).

**Figure 3.** Effects of thinking skills on language performance across groups.

### **Effects of thinking skills across language proficiency groups**

Results of multi-group path analyses are shown in Table 3 and plotted in Figure 3. The effects of evaluating on language performance across all groups were non-significant ( $\beta$ s =.01 to .04, with all  $p > .60$ ). All other four skills had significantly larger positive effects on language performance in the medium-proficiency group than in the two adjacent groups (i.e., low- or high-proficiency groups). Specifically, recognizing and interpreting had a positive effect on language performance in the struggling group. In all, the variation of these effects across the four language proficiency groups generally displayed a 'down-up-down' pattern (See Figure 3), a replication of the original IRC (Cai & Kunnan, 2020).

## **Discussion**

### **Question 1. How does language proficiency moderate the prevalence of thinking skills during language performance?**

Our study showed that the CDM scores for all thinking skills displayed an ascending trend across the four groups with continuous increase in language proficiency. This pattern suggested that, in general, there is a linear relationship between thinking skills and language proficiency. On the surface, this finding supports the widely-held belief that thinking is

beneficial to all students of all proficiency levels to similar extent. However, this result should not be over-interpreted as the higher a student scored in thinking skills, the higher he or she would benefit on their language performance. A more careful investigation needs to be undertaken to examine the actual effect that these thinking processes brings about on language performance across students of different language proficiency levels. We will return to this issue in later text.

The study also showed that the changes in the prevalence of thinking skills across groups were not even. For the two groups on the lower-proficiency end, the changes of prevalence in all skills were small, suggesting the function of language thresholds (Alderson, 1984; Clapham, 1996; Clarke, 1980; Shiotsu & Weir, 2007; Zhang, 2012). This explanation seemed to be more plausible when looking at the changes in the prevalence of thinking skills in the two groups on the higher-end of language proficiency. When students transitioned from low proficiency to medium proficiency, the prevalence of all thinking skills accelerated, suggesting some common language threshold existed with the low-proficiency students (i.e.,  $-0.71$  standard units in our case).

Moreover, the accelerations were uneven with different thinking skills: the effect was the largest with inferring and the smallest with summarizing. A most plausible interpretation for the uneven distribution of the moderation effect by language proficiency should relate to the levels of difficulty (or cognitive load) demanded by activating different types of thinking skills. According to arguments in Bloom's Taxonomy (Anderson et al., 2001; Bloom, 1974), thinking skills on the lower-order end such as recognizing and summarizing skills are relatively less demanding on cognitive resources, whereas skills such as inferring and evaluating are more demanding. Given this inequality, students with lower linguistic resources would need to flexibly deploy their processing resources by favouring thinking skills less demanding. This interpretation can be partly supported by the stable deployment of the less-demanding recognizing skill across all groups and the relatively later occurrence of the acceleration of the more-demanding inferring in the medium- and higher-proficiency groups.

More complex moderation by language proficiency also emerged from our study. In L1 research, Stanovich (1980) proposed the notion of compensatory processing, arguing that deficiencies in any knowledge source can be overcome by other knowledge sources. Bernhardt (2005) and McNeil (2012) extended this notion to L2 reading. Drawing on Bernhardt (2005), McNeil (2012) believed that L2 reading is influenced by the compensatory relations between L2 background knowledge at lower processing levels and L2 strategic knowledge at higher processing levels. In our study, three thinking skills at the higher-order end were not fully operationalized with the struggling learners possibly due to constraints by language thresholds, but these insufficient operations were compensated by summarizing and recognizing on the lower-order end of thinking levels.

With respect to the last two groups, as the constraints by language thresholds gradually loosened, the prevalence of summarizing skill gradually descended with replacements by two higher-order thinking: inferring and evaluating. The relatively stable position of recognizing suggested that the compensatory relation between recognizing and other skills was not salient. This is possibly because all students' language proficiency has far passed the threshold for the recognizing skill.

## ***Question 2. How does language proficiency moderate the effects of thinking skills on language performance?***

A first glance over the results would show that each of the four thinking skills (i.e., recognizing, summarizing, interpreting, and inferring) except for evaluating positively predicted language proficiency across all groups. This is consistent with the general belief in the literature that, to achieve fluent language performance, learners not only need to have a good mastery of linguistic knowledge and world knowledge (Grabe, 2009a; Kintsch & Mangalath, 2011; Perfetti & Adlof, 2012), they also need to develop a list of thinking skills that can help them to build up the mental representation of comprehension (Perfetti et al., 2005; Bennett et al., 2016; Cai & Cheung, 2021; Grabe, 2009b; Kintsch & Mangalath, 2011; Luebke & Lorié, 2013).

A closer look at the results would produce more subtle information. In general, the effects of all thinking skills except for evaluating transited in a pattern of ‘down-up-down’ motion from the divers (students with lowest language proficiency) to the resurfacers (students with second-lowest language proficiency), then to the uphillers (students with medium language proficiency), and finally to the downhillers (students with high language proficiency). This result produced a good replication of the pattern in the original IRC found with strategic competence in medical English reading with undergraduate students (Cai & Kunnan, 2020).

According to the explanation provided by Cai and Kunnan (2020), when students’ language proficiency was too low (e.g., for divers and resurfacers), they lack sufficient linguistic resources to build up a mental representation even at the textbase level (Kintsch, 1998). The fragmented coding due to linguistic insufficiency leads the students to a moment of floundering, during which other non-linguistic factors such as prior knowledge, strategies, and thinking skills might not be able to work efficiently; or even worse, the harder the students struggle with skill activation, the worse the results may turn out to be (Cai & Kunnan, 2020).

The floundering of thinking with the divers and the resurfacers could also come from students’ low processing capacity in thinking skills. As the development of language proficiency involves a continuous process of the integration of linguistic and non-linguistic factors (Bachman & Palmer, 1996, 2010; Cai, 2020; Cai & Cheung, 2021), low language proficiency is generally accompanied by low thinking capacity during language use. This explanation can be partly supported by results regarding the effect sizes of thinking skills within each proficiency group. As plotted in Figure 3, among all five thinking skills, evaluating as a demanding thinking skill ranked the lowest across all groups. Meanwhile, recognizing and interpreting as two less-demanding skills ranked the highest two.

Regarding the uphillers, as their language proficiency moved beyond a certain threshold (e.g., above  $-0.71$  standard units), the accuracy rate of decoding might increase significantly to such an extent that the floundering moment gradually turned away. Released beneficial effects of thinking, thus, gradually increased and reached a peak (Cai & Kunnan, 2020).

Another interesting feature of the replicated IRC related to the decreasing effect of thinking skills with the downhillers. A widely-held view is that thinking skills as a type of human mental resources is something that should always bring about good results. Put another way, the more one uses these thinking skills, the better outcome it should produce. This is the reason why researchers and practitioners in the education sector

have invested so much effort and energy to foster such kinds of human resources in students. Nonetheless, thinking skills in our study with the downhillers did not seem to completely follow this linearity. This phenomenon might find interpretation from cognitive load theory (Shehab & Nussbaum, 2015). According to researchers in cognitive load theory, when students' language proficiency reaches a high threshold (e.g., 1.29 standard units), their thinking becomes an automatic processing mechanism such that students need to spare no effort to activate these load-heavy mental resources (Shehab & Nussbaum, 2015).

It is worth noting that the effects of thinking skills were the largest with the uphillers among all proficiency groups. Closer observation would show that the uphillers were roughly located in the middle of the continuum of language proficiency. This phenomenon of 'golden centrality' somehow echoed Aristotelian philosophers' concept of the 'golden mean', which posits that human virtuous disposition usually lies in a middle position of two ends (Bartlett & Collins, 2011). Put another way, the maximum contribution of a beneficial factor to an outcome is largest when the value of the contributing factor is near the middle, but not near the two ends (too low or too high). In our case, the largest effect of critical thinking is not with students with the lowest value in critical thinking (i.e., divers), nor with students with the largest value in critical thinking (i.e., the downhillers).

### **Limitations and implications**

This study has several strengths including its cross-national data, relatively large sample size, fine-grained analysis of thinking skills, and the appropriate framing under the heuristic model of the IRC (Cai & Kunnan, 2020). However, some key limitations need to be noted. First, the current study used TEM-4 real test data for analysis. Although this is a convenient way of obtaining quality and authentic data for research, there were some key demographic variables (e.g., gender, age, and universities) not provided due to the testing agency's concern for test information security. It is possible that the deployment of thinking skills was also affected by these individual variables apart from language proficiency. Future studies may consider alternative ways to collect data so that these variables can be accounted for.

Second, the cross-sectional design prevented us from making objective conclusions regarding the dynamics of thinking skills that function concurrently with the development of language proficiency. Future studies may consider adopting a longitudinal design to determine whether the fluctuation of these thinking skills effects exists within individuals across time.

Third, although our studies conceptualized the reading subskills as thinking skills, the data produced by test-takers are indeed a combined outcome of thinking, language, and prior knowledge, among other things. Future studies may design tools more independent of language performance to measure thinking skills.

Regardless of these limitations, the study has important implications. Theoretically, it provided additional evidence for the appropriateness of using the IRC to frame research to examine the role of thinking skills in determining language performance. Previously, the IRC was discovered with cognitive variables such as metacognitive and cognitive strategies (Cai & Kunnan, 2020), subject-matter background knowledge (Cai & Kunnan, 2019), and extended to other intelligence-based factors such as motivation regulation strategies (Wang

et al., 2021). Our study demonstrated the capacity of the IRC for explaining the mechanism of another ‘hard intelligence’ variable (i.e., thinking). It is worthwhile for future scholars to explore whether the IRC also applies to other ‘hard intelligence’ variables such as creative thinking, design thinking, system thinking, and equally importantly, other ‘soft intelligence’ variables related to cognition such as self-concept, self-efficacy, and fluid intelligence (or growth mindset), among others.

Practically, our findings provided useful information that can help teachers and students become more cognizant of the mechanism of thinking skills, and the interaction between these thinking skills and language proficiency. This information can be useful for teachers and students in designing thinking skills interventions that attend to remedy or strengthen language proficiency by identifying the weak spots of thinking skills and by considering the different levels of language proficiency of the trainees.

## Conclusion

Beyond just focusing on the identification of thinking skills involved during language performance, researchers may benefit from the more fine-grained studies on thinking skills by taking the perspective of the IRC. What matters is not just what thinking skills are used and to what extent they are used, but how the use of these subskills and the effects of the use vary across different levels of language proficiency. The findings that the effects of thinking skills fluctuate as language proficiency increases underscore our caution that an understanding of the effects of thinking skills on language performance would be incomplete without properly attending to the moderation by language proficiency.

## Acknowledgments

This work was partly supported by The Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning given to the first author (Code: TP2018068) and by the project “Cognitive studies on the English language ability structure of English major college students in China” sponsored by National Social Science Foundation in China to the second and correspondence author (Code: 17BYY101).

Acknowledgements should also be extended to the TEM Examination Board which provided us with the test data.

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## Funding

This work was supported by the The Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning [TP2018068]; National Social Science Foundation in China [17BYY101].

**ORCID**Yuyang Cai  <http://orcid.org/0000-0002-0320-4602>Huilin Chen  <http://orcid.org/0000-0001-8040-3472>**References**

- Adams-Smith, D. (1981). Levels of questioning: Teaching creative thinking through ESP. *English Teaching Forum*, 19(1), 15–21.
- Alderson, J. C. (1984). Reading in a foreign language: A reading problem or a language problem? In J. C. Alderson & A. H. Urquhart (Eds.), *Reading in a foreign language* (pp. 122–135). Longman.
- Alderson, J. C., & Lukmani, Y. (1989). Cognition and reading: Cognitive levels as embodies in test questions. *Reading in a Foreign Language*, 5(2), 253–270.
- Anderson, L. W., Krathwohl, D. R., & Bloom, B. S. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Longman.
- Bachman, L. F., & Palmer, A. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford University Press.
- Bachman, L. F., Palmer, A. S., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford University Press.
- Bartlett, R. C., & Collins, S. D. (2011). *Aristotle's Nicomachean ethics*. University of Chicago Press.
- Bennett, R. E., Deane, P., & van Rijn, W. 2016. From cognitive-domain theory to assessment practice. *Educational psychologist*. 51(1), 82–107, <https://doi.org/10.1080/00461520.2016.1141683>
- Bernhardt, E. (2005). Progress and procrastination in second language reading. *Annual Review of Applied Linguistics*, 25, 133–150. <https://doi.org/10.1017/S0267190505000073>
- Bloom, B. S. (1974). *Taxonomy of educational objectives: The classification of educational goals*. Longman.
- Bronfenbrenner, U. (1979). *The ecology of human development*. Harvard University Press.
- Cai, Y., & Kunnan, A. J. (2019). Detecting the language thresholds of the effect of background knowledge on a language for specific purposes reading performance: A case of the island ridge curve. *Journal of English for Academic Purposes*, 42, 1–13. <https://doi.org/10.1016/j.jeap.2019.100795>
- Cai, Y. (2020). *Examining the interaction among components of English for specific purposes ability in reading: The triple-decker model*. Peter Lang.
- Cai, Y., & Kunnan, A. J. (2020). Mapping the fluctuating effect of strategy use ability on English reading performance for nursing students: A multi-layered moderation analysis approach. *Language Testing*, 37(2), 280–304. <https://doi.org/10.1177/0265532219893384>
- Cai, Y., & Cheung, H. (2021). A dynamic language ability system framework for diagnosing EMI students' readiness of English language ability. In L. I. Su, W. H. Cheung, & J. R. Wu (Eds.), *Rethinking EMI: Multidisciplinary perspectives from Chinese-speaking regions* (pp. 141–160). Taylor & Francis.
- Chen, J., de la Torre, J., & Zhang, Z. (2013). Relative and absolute fit evaluation in cognitive diagnosis modeling. *Journal of Educational Measurement*, 50(2), 123–140. <https://doi.org/10.1111/j.1745-3984.2012.00185.x>
- Chen, H., & Chen, J. (2016a). Exploring reading comprehension skill relationships through the G-DINA model. *Educational Psychology*, 36(6), 1049–1064. doi:<https://doi.org/10.1080/01443410.2015.1076764>
- Chen, H., & Chen, J. (2016b). Retrofitting non-cognitive-diagnostic reading assessment under the generalized DINA model framework. *Language Assessment Quarterly*, 13(3), 218–230. <https://doi.org/10.1080/15434303.2016.1210610>
- Clapham, C. (1996). *The development of IELTS: A study of the effect of background on reading comprehension*. Cambridge University Press.



- Clarke, M. A. (1980). The short circuit hypothesis of ESL reading-or when language competence interferes with reading performance. *Modern Language Journal*, 64(2), 203–209. <https://doi.org/10.2307/325304>
- De La Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, 76(2), 179–199. <https://doi.org/10.1007/s11336-011-9207-7>
- Douglas, D. (2000). *Assessing language for specific purposes*. Cambridge University Press.
- English Major Division of National Foreign Languages Advisory Board. (2000). *English teaching syllabus for English majors [Chinese]*. Foreign Language Teaching and Research Press.
- Grabe, W. (2009a). *Reading in a second language: Moving from theory to practice*. Cambridge University Press.
- Grabe, W. (2009b). Teaching and testing reading. In M. H. Long & C. J. Doughty (Eds.), *The handbook of language teaching* (pp. 441–462). Wiley-blackwell.
- Jang, E. E. (2009). Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for fusion model application to LanguEdge assessment. *Language Testing*, 26(1), 31–73. <https://doi.org/10.1177/0265532208097336>
- Jeon, E. H., & Yamashita, J. (2014). L2 reading comprehension and its correlates: A meta-analysis. *Language Learning*, 64(1), 160–212. <https://doi.org/10.1111/lang.12034>
- Kim, A.-Y. (2015). Exploring ways to provide diagnostic feedback with an ESL placement test: Cognitive diagnostic assessment of L2 reading ability. *Language Testing*, 32(2), 227–258. <https://doi.org/10.1177/0265532214558457>
- Kincaid, J. P., Fishburne, R. P., Rogers, R. L., & Chissom, B. S. (1975). *Derivation of new readability formulas (automated readability index, fog count, and Flesch reading ease formula) for Navy enlisted personnel*. Naval Air Station.
- Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge University Press.
- Kintsch, W., & Mangalath, P. (2011). The construction of meaning. *Topics in Cognitive Science*, 3(2), 346–370. <https://doi.org/10.1111/j.1756-8765.2010.01107.x>
- Lee, Y. W., & Sawaki, Y. (2009). Application of three cognitive diagnosis models to ESL reading and listening assessments. *Language Assessment Quarterly*, 6(3), 239–263. <https://doi.org/10.1080/15434300903079562>
- Leighton, J. P., & Gierl, M. (2007). *Cognitive diagnostic assessment for education: Theory and applications*. Cambridge University Press.
- Liu, J., & Wu, S. (2019). *Research on inventory of China standards of English [Chinese]*. Higher Education Press.
- Luebke, S., & Lorie, J. (2013). Use of Bloom's Taxonomy in developing reading comprehension specifications. *Journal of Applied Testing Technology*, 1(1), 1–27. <http://jattjournal.com/index.php/atp/article/view/45250>
- Ma, W., & de la Torre, J. (2020). GDINA: An R package for cognitive diagnosis modeling. *Journal of Statistical Software*, 93(14), 1–26. <https://doi.org/10.18637/jss.v093.i14>
- McNeil, L. (2012). Extending the compensatory model of second language reading. *System*, 40(1), 64–76. <https://doi.org/10.1016/j.system.2012.01.011>
- Min, S., & He, L. (2022). Developing individualized feedback for listening assessment: Combining standard setting and cognitive diagnostic assessment approaches. *Language Testing*, 39(1), 90–116. <https://doi.org/10.1177/0265532221995475>
- Muthén, L. K., & Muthén, B. Q. (1998–2020). Mplus 8.5 [Computer software]. Muthén & Muthén.
- Pearson, P. D., & Raphael, T. E. (1990). Reading comprehension as a dimension of thinking. In D. Person & T. Raphael (Eds.), *Dimensions of thinking and cognitive instruction* (Vol. 1, pp. 209–240). NCREL.
- Perfetti, C. A., Landi, N., & Oakhill, J. (2005). The Acquisition of reading comprehension skill. In M. J. Snowling & C. Hulme (Eds.), *The science of reading: A handbook* (pp. 227–247). Blackwell Publishing.
- Perfetti, C., Yang, C. L., & Schmalhofer, F. (2008). Comprehension skill and word-to-text integration processes. *Applied Cognitive Psychology*, 22(3), 303–318. <https://doi.org/10.1002/acp.1419>

- Perfetti, C., & Adlof, S. M. (2012). Reading comprehension: A conceptual framework from word meaning to text meaning. In J. Sabatini, E. Albro, & T. O'Reilly (Eds.), *Measuring up: Advances in how we assess reading ability* (pp. 3–20). R&L Education.
- Ravand, H., & Robitzsch, A. (2018). Cognitive diagnostic model of best choice: A study of reading comprehension. *Educational Psychology, 38*(10), 1255–1277. <https://doi.org/10.1080/01443410.2018.1489524>
- Ravand, H., & Baghaei, P. (2020). Diagnostic classification models: Recent developments, practical issues, and prospects. *International Journal of Testing, 20*(1), 24–56. <https://doi.org/10.1080/15305058.2019.1588278>
- Shehab, H. M., & Nussbaum, E. M. (2015). Cognitive load of critical thinking strategies. *Learning and Instruction, 35*, 51–61. <https://doi.org/10.1016/j.learninstruc.2014.09.004>
- Shiotsu, T., & Weir, C. (2007). The relative significance of syntactic knowledge and vocabulary breadth in the prediction of reading comprehension test performance. *Language Testing, 24*(1), 99–128. <https://doi.org/10.1177/0265532207071513>
- Stanovich, K. E. (1980). Toward an interactive-compensatory model of individual differences in the development of reading fluency. *Reading Research Quarterly, 16*(1), 32–71. <https://doi.org/10.2307/747348>
- Von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology, 61*(2), 287–307. <https://doi.org/10.1348/000711007X193957>
- Wang, C., Cai, Y., Zhao, M., & You, X. (2021). Disentangling the relation between motivation regulation strategy and writing performance: A perspective of the Island Ridge Curve [Chinese]. *Foreign Languages World, 204*(3), 46–54+72.
- Yang, M., & Liu, J. (2019). China's standards of English language ability and business English testing and assessment [Chinese]. *Foreign Languages in China, 16*(3), 13–20.
- Zhang, D. (2012). Vocabulary and grammar knowledge in second language reading comprehension: A structural equation modeling study. *The Modern Language Journal, 96*(4), 558–575. <https://doi.org/10.1111/j.1540-4781.2012.01398.x>