Journal of English for Academic Purposes 42 (2019) 100795

Contents lists available at ScienceDirect

Journal of English for Academic Purposes

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

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

Yuyang Cai [a, *], Antony John Kunnan [b]

[a] School of Languages, Shanghai University of International Business and Economics, 1900 Wenxiang Road, Songjiang, Shanghai, China
[b] Department of English, Faculty of Arts and Humanities, E21-4068, Department of English, Faculty of Arts and Humanities, University of Macau, Taipa, Macau, China

ARTICLE INFO

Article history:
Received 22 April 2019
Received in revised form 20 August 2019
Accepted 12 September 2019
Available online 13 September 2019

Keywords:
Background knowledge
Island ridge curve (IRC)
Language threshold
Multilayer Moderation Analysis (MLMA)

ABSTRACT

This study explored the nonlinear pattern by which language knowledge moderates the effect of background knowledge on Language for Specific Purposes (LSP) reading performance and the number of language thresholds that locate the turns of such an effect on background knowledge. We tested two hypothesized patterns in which students' language knowledge interferes with background knowledge: a linear pattern and a quadratic pattern (background knowledge effect moves up-then-down with the increase of language knowledge). A total of 1,491 nursing students from eight medical colleges in China participated in the study. Their background knowledge, language knowledge and LSP reading ability were measured using a nursing knowledge test, an English grammar knowledge test, and a nursing English reading test, respectively. Students' response data were first scored using multidimensional item response theory and then modeled using the innovative method of multi-layered moderation analysis (MLMA). The results supported the quadratic moderation pattern, which we labeled as island ridge curve (IRC). On the IRC, two language thresholds emerged: a resurfacing threshold (theta = -1.49; from where the positive effect of background knowledge emerged and started to increase) and a downhill threshold (theta= 1.00; from where the maximum effect of background knowledge emerged and started to decrease).

1. Introduction

A central issue in Language for Specific Purposes (LSP) assessment relates to the role of subject-matter background knowledge (hereafter, background knowledge) in determining LSP performance (Douglas, 2000). Numerous studies have shown that LSP readers with more background knowledge or more familiarity with topics in the reading tasks perform better than their peers who have less background knowledge (Alderson & Urquhart, 1985, 1988; Clapham, 1996; Krekeler, 2006; Lee & Schallert, 1997; Lin & Chern, 2014; Ridgway, 1997; Usó-Juan, 2006). Among these studies, quite a few have shown that the effect of background knowledge on LSP reading performance varies across readers of different language proficiency (Clapham, 1996; Krekeler, 2006; Ridgway, 1997). For example, Clapham (1996) found a significant effect of background

* Corresponding author.
E-mail addresses: sailor_cai@hotmail.com (Y. Cai), akunnan@umac.mo, akunnan@gmail.com (A. John Kunnan).

https://doi.org/10.1016/j.jeap.2019.100795
1475-1585/

knowledge with intermediate English proficiency students but not with low- or high-proficiency students. To interpret this interesting finding, Clapham posited the well-known two-threshold hypothesis: a lower-language threshold that constrains the beneficial effect of background knowledge and a higher-language threshold that renders background knowledge less useful. However, some scholars such as Ridgway (1997) only confirmed the lower-threshold, and others confirmed none (e.g., Krekeler, 2006; Lin & Chern, 2014). More empirical studies, therefore, are needed to uncover the mysterious function of background knowledge.

2. Review of literature

As early as the 1980s, LSP reading researchers have noted the effect of background knowledge on LSP reading performance. Alderson and Urquhart (1985) examined the effect of students' disciplinary background on their English Language Testing Service (ELTS) reading performance. Participants were students from three disciplines: Business and Economics, Science and Engineering, and Liberal Arts, each taking the Social Studies and the Technology Modules of the ELTS. Results of t-tests indicated significant, though inconsistent, background knowledge effect across the groups. The Science and Engineering group outperformed the other two groups on the Technology Module, but the Business and Economics group did not outperform the Science and Economics group on the Social Studies Module. This inconsistent effect of background knowledge led to the researchers' caution in interpreting the relationship between background knowledge and language proficiency (Alderson & Urquhart, 1988).

Clapham (1996) examined whether students from different disciplines (i.e. Business and Social Sciences, Life and Medical Science, and Physical Sciences and Technology) taking the IELTS (International English Language Testing System) reading module within their own disciplines could outperform students from other disciplines. Background knowledge was measured using self-reported topic familiarity and language proficiency was measured using a standardized language test (a grammar test). Based on the grammar test scores, students ($n = 787$) were divided into three groups: low level (below 60% of the summed score), intermediate level (between 60% and 80% of the summed score), and high level (above 80% of the summed score). Repeated measures analysis of variance of reading scores on the Business and Social Sciences and Life texts and on the Medical Science English texts showed that the effect of background knowledge was significant for students with intermediate language proficiency, but not for those with low or high language proficiency. This pattern, however, was not observed for the Physical Sciences and Technology students taking the Physical Sciences and Technology reading tasks. Drawing on findings from the Business and Social Sciences and Life and Medical Science groups, Clapham concluded with the two-threshold hypothesis to explain the interaction between background knowledge and language proficiency: a lower-threshold that releases the beneficial effect of background knowledge with low proficiency students, and a higher-threshold that clamps down the beneficial effect of background knowledge with high-proficiency students.

Clapham's study bore several merits. The first relates to the use of self-reported topic familiarity to represent background knowledge, whereas in previous studies background knowledge was only represented by students' self-reported disciplinary areas (e.g., Alderson & Urquhart, 1985; Ridgway, 1997). Such a measurement should allow for more accurate interpretation of results. After all, there is no way to tell whether students have actually acquired relevant topical knowledge or not simply by identifying their disciplinary areas (Usó-Juan, 2006). The second merit deals with how language proficiency (or grammar knowledge) was measured. The inclusion of a grammar test enabled the researcher to directly examine the interaction between language proficiency and background knowledge through the application of repeated measures analysis of variance and multi-group regression analysis. This procedure allowed her to go beyond a general examination of the interaction between background knowledge and language knowledge and to provide more detailed information about the interplay between the two factors. Regardless, Clapham's study still left room for a more accurate measurement of background knowledge (e.g., to use tests instead of self-reported topic familiarity). Further, her use of the arbitrary cut-off points of language proficiency to group students has made her findings less convincing. For instance, if we moved the lower cut-off point a little bit upwards (i.e., to 65% of the total score), the interaction between language proficiency and background knowledge might become significant due to the inclusion of students who were originally placed into the medium-level proficiency group. It is, therefore, unclear whether such an unjustified criterion has distorted the reality or not.

Following Clapham, Ridgway (1997) set out to examine the two-threshold hypothesis. The study had a small sample size of 69 students from the Faculty of Business and the Faculty of Environment. Building on an in-house university English proficiency test, students from each discipline were divided into a top group and a bottom group; hence a total of four groups. Each group was asked to perform tasks after reading texts from three disciplines: Business, Environment, and Sociology (on an academic topic that students from Business and Environment were equally familiar with). Non-parametric comparison of means showed some interesting patterns of the effect of background knowledge. On the Sociology text, all four groups performed with a high degree of uniformity. On the Business text, none of the two Business groups performed better than their Environment counterparts. On the Environment text, uniformity emerged again between the low proficiency Business group and the low proficiency Environment group. Things changed, however, to the two top-proficiency groups' performance on the Environment text: The Environment group outperformed the Business group ($p < .02$). Ridgway, therefore, concluded that a lower threshold might exist only for the Environment students. He interpreted the inconsistence between students' performance on the Business and on the Environment texts as the outcome of small sample size and different degrees of text specificity.

Krekeler (2006) also addressed the two thresholds hypothesis by examining 500 international students learning German as a Language for Specific Academic Purposes and with intention to study business or science/technology in Germany universities. Variables included participants' background knowledge (i.e., a mixture of self-reported topic familiarity, content familiarity, and intended areas of study), German language proficiency (i.e., language knowledge or, in her own term, organizational competence measured with C-test) and LSP reading. Visual examination of scatterplots and ANOVA analysis produced quite inconsistent results: on the one hand, the effect of background knowledge on reading performance appeared to be strong across both groups using all background knowledge indicators. On the other hand, no interaction between background knowledge and language proficiency was observed, except for that produced by analyzing reading performance on the business text (addressing inflation) using participants' self-reported topic familiarity as the indicator of background knowledge. For this proportion of data, the interaction effect was most evident for students with medium levels of language proficiency (scores on the C-TEST between 40% and 60% around the mean of 55.6%), a pattern consistent with the two-threshold hypothesis. Regardless, Krekeler concluded that the effect of background knowledge is significant, but the pattern was unpredictable.

Krekeler provided a relatively comprehensive view of the interaction between background knowledge and language proficiency using scatterplots along a whole range of language proficiency with more language proficiency points for testing interaction. Nevertheless, the cogency of his conclusion would attenuate if one looks closer into his study. Similar to previous studies, the key measures (e.g., background knowledge, language knowledge and LSP reading) were not derived from concrete theoretical models. Take for example the measurement of background knowledge. Although the researcher took great pains to use multiple indicators to measure background knowledge (i.e., topic familiarity before the reading, topic familiarity after the reading, and students' intended areas of study), none of the measures could be considered as real measures of background knowledge *per se*. For instance, the third indicator (i.e., intended areas of study) could tell us little about whether participants had acquired background knowledge for the moment or not. Another limitation, as Krekeler pointed out, was the small sample sizes of students with very low and very high language proficiency. The third limitation relates to the way that the researcher drew his conclusion: he first plotted the scores of background knowledge and language proficiency against those of LSP reading performance, and then eyeballed the regression trend. Compared with other methods such as ANOVA and multi-group regression that could tap into the interaction by grouping students according to their language proficiency, this approach was able to provide a more comprehensive sketch of the hypothesized interaction along the continuum of language proficiency. However, the inability of human eyes in capturing statistical subtlety make his conclusion less convincing.

A brief review of the literature suggests the mechanism of interaction between background knowledge and language proficiency. However, due to limitations in the ways background knowledge was measured and the analytical techniques applied to detect the interaction, two critical issues remained uncertain: the exact pattern of the interaction between language proficiency and background knowledge, and the corresponding number of language thresholds and their locations in affecting the effect of background knowledge.

The current study was designed to revisit these two critical issues. To achieve our goals, we developed a nursing knowledge test to measure students' background knowledge. Besides, we applied bifactor-multidimensional item response theory (bifactor-MIRT) to evaluate the measurement quality of each measure and to score students' responses to each measure (i.e., background knowledge, language knowledge and LSP reading performance). Compared with summed scores based on classic test theory, bifactor-MIRT based scores are more accurate in controlling for confounding factors from mismatch between person ability and item features (e.g., item difficulty and item discrimination) (Reckase, 2009). Most importantly, we developed the multi-layered moderation analysis (MLMA) to detect the nonlinear moderation of language knowledge on background knowledge. Previous studies usually relied on less powerful analytical methods such as *t*-test, ANOVA, or multiple regression to compare the effect of background knowledge across a limited number of groups. While such approaches can provide information regarding the variation of background knowledge effect across a limited number of language proficiency groups, this type of information is deemed crude (especially, for locating language thresholds), or likely to distort the reality if the grouping criteria is inappropriately determined. The MLMA, instead, allows us to explore the changing effect of background knowledge on LSP reading performance along the whole continuum of sampled students' language proficiency, and to detect the exact location(s) of the language threshold(s), if these language thresholds do exist.

3. The study

3.1. Research questions

This study was led by the following two research questions:

1. Does the effect of background knowledge on LSP reading performance change with the change of students' language proficiency (represented by grammatical knowledge)? If yes, to what extent is the effect influential, and when does the effect become more or less influential?
2. How many language thresholds affect the effect of background knowledge on LSP reading performance? What roles do these thresholds play in determining the effect of background knowledge?

4. Method

4.1. Participants

Participants involved 1491 second-year nurse students (1465 females and 26 males) from eight medical and healthcare colleges in China, with ages between 18 and 22. They were requested to respond to three tests: The Nursing Knowledge Test (NKT) measuring background knowledge, the English Grammar Test (GKT) measuring English grammar knowledge, and the Nursing English Reading Test (NERT) measuring LSP reading ability.

4.1.1. Instruments

The Nursing Knowledge Test (NKT). The NKT used retired items from the corresponding sections of a national nurse licensing exam for Chinese nurses: the National Qualification Examination for Nurse Practitioners (NQENP) (MOH, 2008). To correspond with the four topics addressed by the Nursing English Reading Test (NERT), the NKT was developed to contain four six-item subtests, each measuring knowledge in gynecology nursing, pediatrics nursing, basic nursing, and internal medicine nursing, respectively (See Table 1). The NKT was administered in Chinese mandarin. The internal consistency or the 24-items NKT was alpha = .72. A sample item measuring knowledge in internal medicine nursing was:

对头痛病人护理措施不正确的是 (Incorrect care solutions for patients with headache include):

[A] 鼓励病人用止痛药 (Encourage patients to use painkillers);

[B] 鼓励病人进行放松训练 (Encourage patients to learn to relax);

[C] 鼓励病人卧床休息 (Encourage patients to rest in bed);

[D] 鼓励病人进行理疗来缓解疼痛 (Encourage patients to undergo physical therapy to relieve pain).

The dimensionality of NKT was assessed using bifactor-multidimensional item response theory (bifactor-MIRT) and the results indicated that the NKT items could be represented by five uncorrelated factors: a domain-general factor representing common features underlying all NKT items and four domain-specific factors representing each of the four subjects: gynecology nursing, pediatrics nursing, basic nursing, and internal medicine nursing (see Cai, 2015).

The Grammar Knowledge Test (GKT). The GKT had 15 discrete sentences, each with a gap to be filled in by selecting the best answer from four alternatives. The GKT used retired items from the Public English Test System-Level Two (PETS-2) (NEEA, 2007). The language level of the PETS-2 corresponds to the 'intermediate medium' or 'intermediate high' level of the American Council on Teaching Foreign Languages Proficiency Guidelines (ACTFL, 2012) or to the level of A2 or B1 of the Common European Framework of Reference for Languages (Council of Europe, 2011). Purpura (2004) defines grammatical knowledge as knowledge of grammatical forms and knowledge of grammatical meanings. Following this definition, nine GKT items were coded as grammatical forms and six items coded as grammatical meanings (Please see Table 2).

A sample item measuring grammatical form:

She would rather stay at home than _____ with John.

**[A] go [B] went [C] going [D] to go*

A sample item measuring grammatical meaning:

2: My cousin sent me a wonderful gift from Africa.

Oh, is that so? _____?

*[A] Where is it from *[B] What's it like [C] How did it come [D] Who sent it to you*

The internal consistency alpha for the 15-items GKT was 0.71. The GKT was assessed using bifactor-multidimensional item response theory (bifactor-MIRT) and the form-meaning structure of the GKT was verified in Cai (2014).

The Nursing English Reading Test (NERT). The NERT used retired items from the reading section of the Medical English Test System Level Two (For Nurses) (METS, 2007). The METS is a four-level test battery developed to measure nurse students' language ability in English-speaking nursing workplace. The reading section of the METS measures nurse students' ability in understanding written materials related to healthcare (e.g., short stories or essays related to daily health issues) or documents used in healthcare environments (e.g., brochures, tables, or instructional manuals), which are again categorized into reading for implicit meanings and reading for explicit meanings. As an intermediate level for the METS program, the reading section of

Table 1
The nursing knowledge test (NKT).

Subtests	Items	Subtotal
Gynecology Nursing	NK1, NK2, NK3, NK4, NK5, NK6	6
Pediatrics Nursing	NK7, NK8, NK9, NK10, NK11, NK12	6
Emergency Nursing	NK13, NK14, NK15, NK16, NK17, NK18	6
Medical Nursing	NK19, NK20, NK21, NK22, NK23, NK24	6
Total		24

Table 2
The grammar knowledge test (GKT).

Components	Elements	Items	No. of Items	Subtotal
Grammatical form	Lexical form	GK3, GK12, GK14	3	9
	Morphosyntactic form	GK5, GK7, GK11, GK15	4	
	Cohesive form	GK8, GK13	2	
Grammatical Meaning	Lexical meaning	GK6, GK9, GK10	4	6
	Cohesive meaning	GK2	1	
	Interactional meaning	GK1, GK4	2	
Total				15

Table 3
The nursing English reading test (NERT).

Text	Items	Subtotal
Dystocia	NR1, NR2, NR3*, NR4, NR5	5
Baby Massage	NR6, NR7, NR8*, NR9, NR10	5
Emergency	NR11, NR12*, NR13*, NR14, NR15*	5
Migraine	NR16*, NR17, NR18, NR19, NR20*	5
Total		20

Note. NR1 to NR20 represent NERT Item 1 to Item 20, respectively; Items with the star symbol * represent reading for implicit meanings; items without the * symbol represent reading for explicit meanings.

the METS-2 (relabeled as NERT in this study) focused on skills of the second category. The NERT adapted for the study contained four passages addressing topics in four subject areas: gynecology nursing (Text 1), pediatrics nursing (Text 2), emergency nursing (Text 3), and internal medicine nursing (Text 4). Each passage had a length of about 190–300 words and was accompanied by five multiple-choice (MC) questions (see Table 3). The Flesch Reading Ease values are 54.17 (for Text 1), 70.67 (for Text 2), 66.84 (for Text 3), and 48.24 (for Text 4), with higher values representing easier tasks. A sample reading item for Text 4 (addressing internal medicine nursing) is:

Question: _____ are more vulnerable to Migraine.

[A] Americans; [B] Asians; [C] Africans; [D] Caucasians.

The internal consistency alpha for the 20-items NERT was 0.73. The measurement quality of the NERT was assessed using bifactor-MIRT and the results indicated a structure of five uncorrelated factors perfectly fit the data: a domain-general factor representing general reading ability and four domain-specific factors (i.e., testlet factors) representing background knowledge on each of the four subjects: gynecology nursing, pediatrics nursing, basic nursing, and internal medicine nursing (see Cai & Kunnan, 2018).

4.1.2. Data collection and analysis

Before data collection, ethics issues were reviewed by The University of Hong Kong. Field entry permits from administrators and student participant agreement signatures were obtained and participants were told of the purpose and background of the study before taking the tests. During the 90-min data collection, students took the tests in the sequence of GKT, NERT and NKT.

Our analyses involved four major steps. First, descriptive statistics and internal consistency were computed using SPSS Version 20.0. Second, bifactor-MIRT was applied to validate the tests and to compute bifactor-MIRT scores. Third, three sets of composite scores were derived from the bifactor-MIRT scores to represent background knowledge, language knowledge, and LSP reading ability, respectively. As these steps have been reported elsewhere,[1] this study only dealt with the next two steps.

As the fourth step, a baseline structural equation model was constructed, with LSP reading (i.e., nursing English reading) as the endogenous variable and language knowledge and background knowledge as exogenous variables. Finally, a series of multi-layered moderation analysis (MLMA) was conducted to test the exact pattern of interaction between language knowledge and background knowledge in affecting LSP reading performance. Briefly speaking, the MLMA tested whether language knowledge affects the relation of background knowledge to LSP reading performance linearly (i.e., the effect of background knowledge increases with the increase of language knowledge) or quadratically (i.e., the effect of background knowledge varies up-then-down as language knowledge increases). A detailed introduction to MLMA is provided in the Appendix.

The present study used *Mplus* 8.2 (L. K. Muthén & Muthén, 1998-2018) for the fourth and fifth steps. Four indices were used to evaluate structural equation models. These included the Comparative Fit Index (CFI; Bentler, 1990), Tucker–Lewis index

[1] See Cai (2015) for the Nursing Knowledge Test, Cai (2013) for the Grammar Knowledge Test, and Cai and Kunnan (2018) for the Nursing English Reading Test.

(TLI; Tucker & Lewis, 1973), Chi Square ($\chi 2$) statistic, the root mean square error of approximation (RMSEA; Steiger & Lind, 1980) and the standardized root mean square residual (SRMR). Following Hu and Bentler (1999) and Mueller and Hancock (2010), a model was considered to have good fit if CFI and TLI values were larger than 0.95 and if RMSEA and SRMR values were less than 0.05.

For MLMA model evaluation, the RMSEA, CFI, TLI, and SRMR were not appropriate as they are insensitive to nonlinear effects (L. K. Muthén & Muthén, 1998-2017). Following Muthén and Muthén, we consulted the chi-square significance of −2 times the loglikelihood difference between the simple model and the complex model (Δ-2LL) with the difference between the numbers of free parameters as the degrees of freedom. If the chi-square was significant, then the latent interaction would be justified.

5. Results

5.1. Model fit results

This section reports the model fit results for structural equation models without interaction terms (Step 4) and with MLMA interaction terms (Step 5). Prior to exploring the relationship between different key variables, a series of confirmatory factor analyses were performed on each individual scale, in the order of the NKT (Model 1), the NERT (Model 2), the combination of the NKT and NERT (Model 3), and finally, the combination of the NKT, NERT and GKT (Model 4). The fit indices are presented in the upper part of Table 4.

The fit indices for the NKT measurement model (Model 1) met the criteria for a good-fit model ($X^2(2) = 10.10$, $p < 0.01$, RMSEA = 0.05 (0.02, 0.09), SRMR = 0.00, CFI = 1.00, TLI = 1.00). Excellent fit indices were obtained for the NERT measurement model (Model 2): ($X^2(2) = 1.79$, $p < .41$, RMSEA = 0.00 (0.00, 0.01), SRMR = 0.00, CFI = 1.00, TLI = 1.00). These results indicated both the NKT and the NERT composite scores performed excellently in recovering the intended constructs of nursing knowledge and nursing English reading ability. However, when combining the two measurement models together (Model 3a), the model fit indices dropped substantially to a level of poor fit ($X^2(19) = 1296.27$, $p < 0.00$, RMSEA = 0.21 (0.20, 0.22), SRMR = 0.03, CFI = 0.90, TLI = 0.85). The modification indices suggested the connection between the uniquenesses of the NERT measurement model to their corresponding uniquenesses of the NKT measurement model. Substantively, this means that the variance of the NERT uniquenesses could be attributed to the domain-specific knowledge represented by their corresponding NKT uniquenesses. A modified model (Model 3b) by making the four connections produced good fit indices again ($X^2(4) = 268.20$, $p < 0.00$, RMSEA = 0.04 (0.03, 0.05), SRMR = 0.02, CFI = 1.00, TLI = 1.00). The third individual measurement model for GKT was then added to make a full measurement model (Model 4a). This model showed a set of fit indices mostly above the good fit criteria ($X^2(28) = 167.40$, $p < 0.00$, RMSEA = 0.06 (0.05, 0.07), SRMR = 0.03, CFI = 0.99, TLI = 0.98).

To examine the reason for the relatively low RMSEA statistics, the modification indices were consulted and potential covariance between the GKT composite (representing grammatical form) in the NKT and the TXT1 composite (the text addressing the topic of gynecology) in the NERT were suggested. Substantively, this could mean that the gynecology text was more demanding in test taker's knowledge of grammatical form than other three texts. This modification was made accordingly and produced a set of fit indices all above the good fit criteria ($X^2(27) = 123.62$, $p < 0.00$, RMSEA = 0.05 (0.04, 0.06), SRMR = 0.02, CFI = 0.99, TLI = 0.99).

The fit indices and change of fit for the MLMA models are shown in the lower part of Table 4. The −2 times the loglikelihood (with the number of estimated parameters) for the baseline model (Model 4b) was 26,671.90 (38). The ΔG^2s (and associated changes in degree of freedom) due to successively adding the linear moderation (Model 5) and the quadratic moderation (Model 6) were 3.32 (1) and 6.36 (1), respectively, both significant at $p < 0.01$. The results indicate that quadratic

Table 4
Fit indices for CFA and SEM models.

	Model	X^2	df	X^2/df	p	RMSEA (90% C·I.)	SRMR	CFI	TLI
Part 1	Model 1	10.10	2	5.05	0.05	0.05 (0.02, 0.09)	0.00	1.00	1.00
	Model 2	1.79	2	0.90	0.41	0.00 (0.00, 0.05)	0.00	1.00	1.00
	Model 3a	1296.27	19	68.23	0.00	0.21 (0.20, 0.22)	0.03	0.90	0.85
	Model 3b	268.20	4	2.85	0.00	0.04 (0.03, 0.05)	0.02	1.00	1.00
	Model 4a	167.40	28	5.98	0.00	0.06 (0.05, 0.07)	0.03	0.99	0.98
	Model 4b	123.62	27	4.58	0.00	0.05 (0.04, 0.06)	0.02	0.99	0.99
	Model	AIC	BIC	-2LL	df	-2LL Change	df change	p	
Part 2	Model 4b	26747.90	26949.58	26671.90	38	NA	NA	NA	
	Model 5	26746.58	26953.56	26668.58	39	3.32	1	0.01	
	Model 6	26742.22	26954.51	26662.22	40	6.36	1	0.01	

Note. Model 1: The NKT measurement model (hypothesized); Model 2: The NERT measurement model (hypothesized); Model 3a: the measurement model combining NKT and NERT (hypothesized); Model 3b: The measurement model combining NKT and NERT (modified); Model 4a: The full measurement model combining NKT, NERT, and GKT (hypothesized); Model 4b: the full measurement model combining NKT, NERT and GKT (modified); also the baseline model for testing Models 5 and 6 (specified by the equation: b0 + b1LK + b2 BK + e); Model 5: The baseline model plus the product of BK and LK (specified by the equation: NERA = b0 + b1LK + b2BK + b3BK × LK + e); Model 6: Model 5 plus the product of BK and LK squared (specified by the equation: NERA = b0 + b1*LK + b2BK + b3BK × LK + b4BK × LK2 + e). Note that the final equation is equivalent to Equation (3.1) detailed in the Appendix).

moderation was the optimal representation of the interaction between language knowledge and background knowledge in affecting LSP reading performance.

5.2. Results of the effect of background knowledge

Fig. 1 shows the diagram of the final MLMA model (Model 6). The largest direct effect on LSP reading performance was from grammatical knowledge ($\beta = .47$, $p < .01$), followed by background knowledge ($\beta = 0.26$, $p < .01$), the linear moderation ($\beta = 0.10$, $p < .01$), and lastly by the quadratic moderation ($\beta = -0.05$, $p < .01$).

The curve in Fig. 2 illustrates the effect of background knowledge on LSP reading performance as captured by the quadratic moderation pattern. For the sake of communication, the curve was thereafter metaphorically labeled as the island ridge curve (IRC). The IRC is 'sliced' by two language thresholds (i.e., theta 1 = −1.49 and theta 2 = 1) and consists of three ridges (zones): the resurfacing ridge (Zone 1), the uphilling ridge (Zone 2), and the downhilling ridge (Zone 3). Within Zone 1, the effect of background knowledge started from an extreme negative value beta = −0.05 and then resurfaced towards the sea level beta = 0 at Threshold 1; within Zone 2, the effect of background knowledge continued to increase from beta = 0 toward the peak of beta = 0.31 at Threshold 2; and within Zone 3 the effect of background knowledge stepped down towards a smaller but positive value of beta = 0.29.

6. Discussion

Question 1. Does the effect of background knowledge on LSP reading performance change with the change of students' language proficiency (represented by grammatical knowledge)? If yes, to what extent is the effect influential, and when does the effect become more or less influential?

The results of MLMA supported the quadratic moderation, or metaphorically, the pattern of the island ridge curve (IRC). The IRC consisted of three connected ridges (zones): the *resurfacing ridge (Zone 1)*, the *uphilling ridge (Zone 2)*, and the *downhilling ridge (Zone 3)*. The first ridge is located to the left most of the curve. Within the range students had language knowledge between 1.67 minus standardized units (Threshold 1) and 1.49 minus standardized units (Threshold 2). Correspondingly, the magnitudes of the effect of background knowledge on LSP reading performance ranged from 0.05 minus to 0.00. The trail of the effect of background knowledge within this ridge showed a resurfacing trend, thereby the resurfacing ridge. The whole resurfacing ridge was below the x-axis, suggesting that for students with extremely low language proficiency the activation of background knowledge did no good but might have harmed their LSP reading comprehension. This is

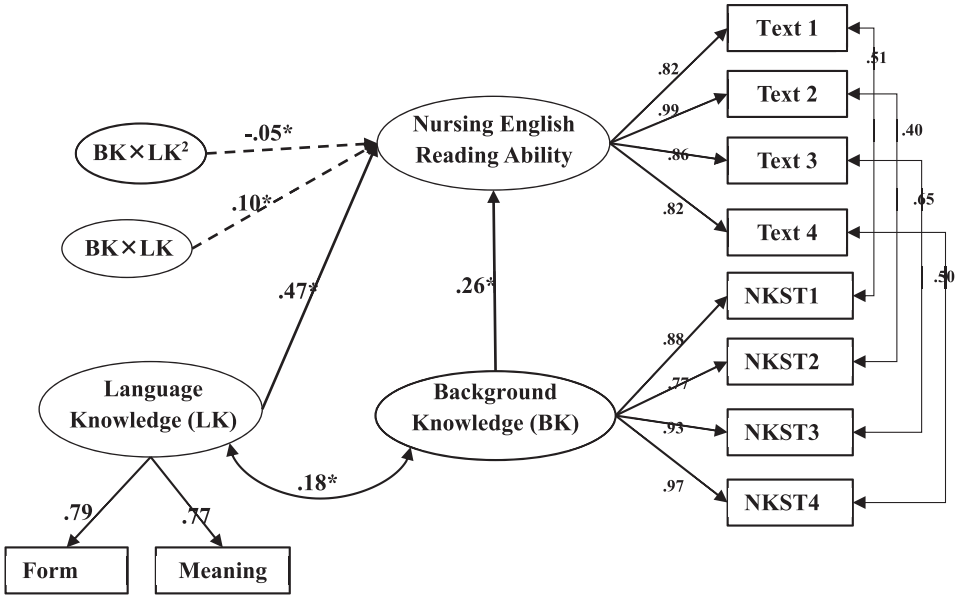
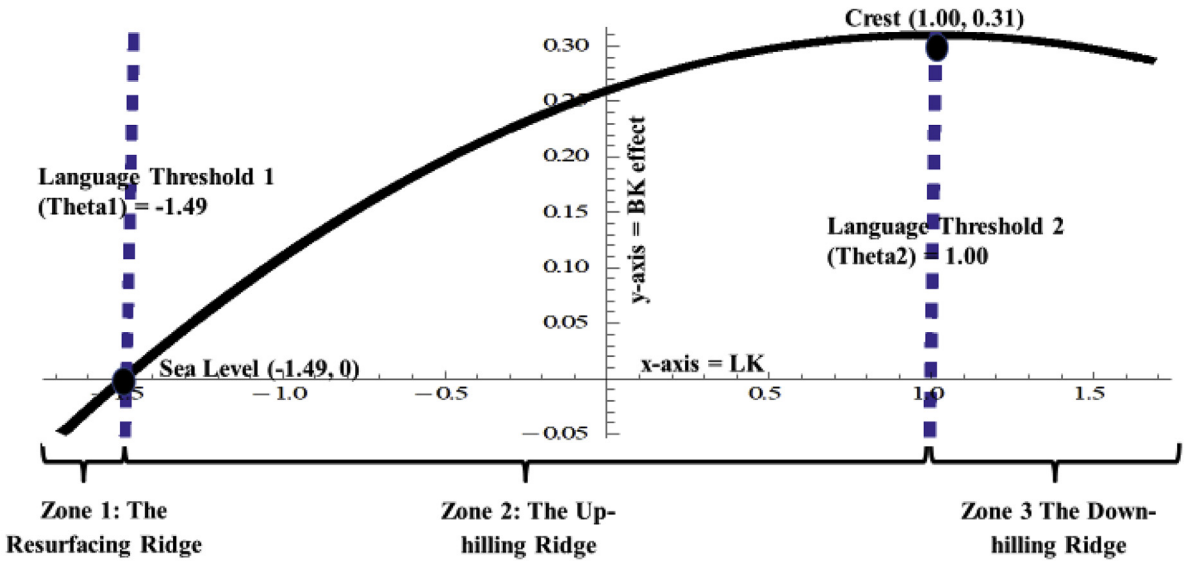


Fig. 1. Diagram for the MLMA (Model 6 with standardized estimates). **Note**. *$p < .05$; $BK \times LK$ = product term of background knowledge with language knowledge (first layer moderation); $BK \times LK^2$ = product term of background knowledge with language knowledge squared (second layer moderation). Text 1 (addressing gynecology nursing) corresponded with NKTST1 (measuring gynecology nursing knowledge); Text 2 (addressing pediatrics nursing) corresponded with NKTS2 (measuring pediatrics nursing knowledge); Text 3 (addressing emergency nursing) corresponded with NKTST3 (measuring emergency nursing); and Text 4 (addressing internal medicine nursing) corresponded with NKST4 (measuring internal medicine nursing knowledge).



Note. *BK = background knowledge, LK = language knowledge*

Fig. 2. The island ridge curve (IRC) illustrating the moderation of language knowledge on background knowledge effect in affecting LSP reading performance.

possible, as readers might have decoded the text wrongly such that the miscoding again misled students to activate irrelevant background knowledge (Perfetti & Hart, 2001); or that even if they were able to activate certain relevant background knowledge, the activation was insufficient to some extent; or that they were able to activate relevant background knowledge but ended up using them wrongly when inferring meanings in the text. These interpretations are consistent with findings with ESL learners that partial or false background knowledge interfere with comprehension (Carrell & Eisterhold, 1983).

The uphilling ridge is a smooth ascending slope. It was located from the resurfacing moment (language knowledge = −1.49) to the peak (language knowledge = 1.00). This result suggested that, starting from the resurfacing moment, the potential beneficial effect of background knowledge was gradually released from the constraint of language knowledge and provided readers with foundational power to use background knowledge in a beneficial way (Perfetti & Hart, 2001). As language knowledge kept increasing, more and more relevant background knowledge could be rightly activated and deployed to right use, pushing the beneficial function of the effect of background knowledge further and faster until the effect reached its peak.

The downhilling ridge is perhaps most interesting. It started as soon as background knowledge effect reached the peak (at Threshold 2). Within this ridge, the effect of background knowledge gradually stepped down from the peak with the increase of language knowledge. A plausible interpretation is that, when students' language knowledge reached certain point (language knowledge = 1.00), background knowledge acquired its full freedom and had its full beneficial capacity released. At this point, language knowledge and background knowledge collaborated in such a harmonious way that any additional activation of background knowledge would make no further contribution to comprehension but adding extra online workload (Madrid & Cañas, 2009). As language knowledge continued to increase, the need to activate background knowledge decreased, thereby leading to reduced effect of background knowledge that is executed. In a certain sense, this interpretation is consistent with Clapham's (1996) explanation of the mechanism of the higher-language threshold. However, as we will discuss below, there is a caveat to this interpretation.

> *Question 2. How many language thresholds affect the effect of background knowledge on LSP reading performance? What roles do these thresholds play in determining the effect of background knowledge?*

Fig. 1 shows that two language thresholds determined the pattern of the moderation of language knowledge on the relation of background knowledge to LSP reading performance. The first threshold (theta = −1.49) separated the resurfacing ridge from the uphilling ridge and the second threshold (theta = 1.00) separated the uphilling ridge from the downhilling ridge. The first threshold is where negative effect of background knowledge switched to be positive, whereas the second threshold is where the full potential of background knowledge was reached and started to fade out.

The IRC (island ridge curve) representing the change effect of background knowledge neither corroborated the null-threshold hypothesis (Krekeler, 2006) nor the lower-threshold hypothesis (Ridgway, 1997). Instead, from one point of view, the IRC was consistent with the two-threshold hypothesis (Clapham, 1996). A closer comparison between the IRC and the two-threshold hypothesis, however, would lead to a somewhat different conclusion.

Before delving into this discussion, we would like to visit the connotation of 'threshold'. In the Merriam-Webster dictionary, 'threshold' is defined as 'the point or level at which something begins or changes'. In the *short-circuit theory*, a language threshold refers to the minimum level of L2 language proficiency (in our case, language knowledge) required for releasing the beneficial effect of L1 skills (Clarke, 1980). Under this theory, the assumption is that L1 skills play no role in affecting L2 comprehension (neither positive nor negative) until L2 proficiency reaches a certain threshold. The term "threshold" therefore means the critical point along a language proficiency continuum where L1 effect switches from zero to a positive value. This dichotomous view of change has found its way in the two-language threshold hypothesis, in which the lower-threshold is posited to be where background knowledge effect 'clicks' and becomes positively significant, and the higher-threshold is where the effect of background knowledge 'clicks' and switches to nothing (Clapham, 1996; Ridgway, 1997).

While Clapham's interpretation of background knowledge effect with students of middle language proficiency is consistent with the information contained in the uphilling ridge, her interpretations of the two-extreme groups are not. For this reason, her interpretations of the two language thresholds are not consistent with the two language thresholds in the IRC. As the resurfacing ridge shows, the effect of background knowledge is not always helpful to all students but can be detrimental to some students with extremely low language proficiency. According to the IRC, the first language threshold is where the negative effect of background knowledge is exhausted and where the beneficial potential begins to be released. Similarly, the second language threshold is the point where the maximum potential of the effect of background knowledge starts to fade out, rather than the point where background knowledge effect disappears suddenly, as posited in Clapham's two-threshold hypothesis. A more sensible interpretation is that the change is transitional rather than a 'click' and 'switch-off', as suggested by the IRC.

7. Conclusion

This study explored the interaction between language knowledge and background knowledge in affecting LSP reading performance. The study has generated evidence as to the variation of the effect of background knowledge on LSP reading as language knowledge continued to increase. The up-then-down pattern of the effect of background knowledge on LSP reading was illustrated metaphorically in an island ridge curve (IRC): (1) when language knowledge was extremely low (below Threshold 1 = −1.49), background knowledge effect was negative but showed a trend to move up the sea level as language knowledge continuously increases; (2) when language knowledge was medium (from Threshold 1 = −1.49 to Threshold 2 = 1.00), background knowledge effect surfaced out of the sea level and moved upwards towards the peak (where background knowledge effect was 0.31); and (3) when language knowledge became high enough (beyond Threshold 2 = 1.00), background knowledge stepped down from its full potential and switched its way downward. These findings advance our understanding of the effect of background knowledge and its interaction with language knowledge (and the locations of language thresholds) in affecting LSP reading performance.

Our investigation is unique for at least two reasons. First, the measurement accuracy of each key variable was assessed using multidimensional item response theory, a sophisticated approach for assessing measurement validity. Using this statistical technique, potential measurement errors due to mismatch between person features (i.e., ability in language or background knowledge) and item features (e.g., item difficulty and item discrimination) were clamped down to the minimum (van der Linen & Hambleton, 1997). More importantly, the MLMA allowed the researchers to detect quadratic interaction between language knowledge and background knowledge in determining LSP reading performance. The MLMA is superior in that it allowed full information of the effect of background knowledge along the continuum of language knowledge. This would be impossible through the multi-group approach that divides students into language knowledge groups based on some arbitrary cut-off points that were either inconsistent across studies nor justified in any way. We believe that the efforts we invested into this study have allowed us to present a relatively more comprehensive and convincing picture of the effect of background knowledge in the backdrop of language knowledge.

The findings of our study should contribute to the teaching and learning of LSP reading. It calls to attention that background knowledge may have different effects on LSP reading comprehension, both in terms of quality (harmful or beneficial) and quantity (more harmful or more beneficial). For students with extremely low language knowledge, although they would still rely on their background knowledge to facilitate their comprehension, it is suggested that they should not expect background knowledge can help them in their text comprehension. Instead, more attention could be paid to improving their language knowledge. For students who have very high level of language knowledge, their language knowledge would have been internalized to such an extent that their language knowledge and background knowledge can automatically collaborate with each other in a subconscious way. That is to say, they could let go their awareness of using background knowledge at the conscious level for a natural flow of comprehension. This idea of subconscious collaboration of language knowledge and background knowledge is similar to the idea of flow theory in cognitive psychology (Csikszentmihalyi, 1990).

To conclude, the authors are aware of several limitations in the study. One relates to the measurement of key constructs such as language knowledge and background knowledge. To fit in with mainstream research, we kept the label of language proficiency but did not include other key components embedded in this construct such as textual knowledge, pragmatic language knowledge and strategic competence. Similarly, we used background knowledge to refer to medical and nursing knowledge required for nurses in their workplaces, but the whole range of this knowledge should extend to knowledge of other subjects. In addition, our study only involved nursing students reading English texts in their own domain. It is not sure

whether this language and background knowledge interaction would hold for other disciplines. Regardless, the IRC that emerged from our data could still work as a heuristic model for future researchers to test the interaction between background knowledge and language knowledge in determining LSP reading, listening, speaking, writing, or translating performance in other domains (e.g., law, business, and so forth). To achieve these goals, the MLMA technique should be able to provide a promising tool together with other qualitative methods such stimulated recall or think aloud.

Data availability statement

The data used were part of the first author's doctoral thesis project. The data will be available at request.

Declarations of interest

None.

Acknowledgement

The work was partly supported by three grants offered to the first author: 1) The Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning (Code: TP2018068), 2) the TOEFL Small Grants for Doctoral Research in Second or Foreign Language Assessment, Educational Testing Service, USA, and 3) by the Grants for Graduate Students in Psychological and Educational Measurement Programs, Assessment Systems Corporation, USA. The authors would also like to extend their sincere thanks to scholars that have generously provided invaluable consultations on the project: Professor Bengt Muthen and Professor Zhonglin Wen on latent interaction, Professor Mark Reckase and Professor Li Cai on multidimensional item response theory, and Professor Jim Purpura on the coding of the grammar knowledge test. Regardless, all possible errors remained on the authors.

Appendix

Technical report: Multi-layered moderation analysis (MLMA)

This technical report assumes that readers have basic knowledge in structural equation modeling (SEM). For readers who might not have this knowledge, there are good resources such as textbooks describing basic SEM concepts (Hoyle, 1995) and guide books for using SEM computer programs such as AMOS (Byrne, 2013), LISREL (Joreskog, Olsson, & Wallentin, 2016), Mplus (Geiser, 2012), and so forth.

The multi-layered moderation analysis (MLMA) is a byproduct of the first author's doctoral thesis (Cai, 2013). It is an innovative application of latent moderation modeling (Klein & Moosbrugger, 2000; Muthén, 2012) embedded in SEM. The general idea of latent moderation is that, in a SEM model that involves two or more predictor variables, the variance of the outcome variable is not only explained by the individual variance of each predictor variable, but also by the variance of their multiplicative product term (Marsh, Hau, Wen, Nagengast, & Morin, 2013; Marsh, Wen, & Hau, 2004). The significance of the multiplicative product term in the SEM suggests that, as the value of another predictor variable (e.g., language proficiency) increases, the effect of one predictor variable (e.g., background knowledge) on the outcome variable would linearly increases (when the coefficient of the multiplicative product term is positive) or linearly decrease (when the coefficient of the multiplicative product term is negative). However, in real life the interaction between two predictor variables may display more complex pattern than linearity, such as the fluctuation of background knowledge on LSP reading performance as suggested by the two-threshold hypothesis. More sophisticated interaction analysis technique needs to be developed to help solve this substantive problem. The MLMA was developed for this right reason.

The MLMA involves two phases, one for constructing and testing multi-layered moderation terms to scientifically justify the existence of a specific interaction pattern, the other for plotting and interpreting the analytical results to help interpret the substantial meaning of a justified interaction pattern. The following text covers the major procedures directly involved in the current study for detecting two types of interaction (i.e., linear and quadratic moderation) between language knowledge and background knowledge during nursing English reading performance.

Phase 1. The construction of the MLMA model

The MLMA approach developed for the current study involved four steps. First, a conventional latent moderation model was created (Klein & Moosbrugger, 2000). Specifically, the exogenous (outcome) variable of nursing English reading performance was expressed as a function of a constant variable (represented by b_0), the random effect of background knowledge (represented by b_1 BK), the random effect of language knowledge (represented by b_2 LK), and the effect of unknown factors (represented by *e).* This resulted into the following equation:

$$NERA = b_0 + b_1\,BK + b_2\,LK + e \quad (1.0)$$

where NERA represents nursing English reading ability.

A limitation of Equation (1.0) is it does not capture the additional interaction effect between background knowledge and language knowledge on nursing English reading performance. It is therefore straightforward to move to the second step to incorporate the latent moderation algorithm proposed by Klein and Moosbrugger (2000). This is shown with the following equation:

$$NERA = b_0 + b_1BK + b_2LK + b_3BK \times LK + e \quad (2.1)$$

Thus, the additional moderation effect of language knowledge on background knowledge is captured by the term of $b_3BK \times LK$, where b_3 represent the random effect of the moderation term. According to Muthén (2012), to depict the relationship between the moderator (i.e., language knowledge) and the moderated variable (i.e., background knowledge), one can transform Equation (2.1) and obtains its equivalence:

$$NERA = b_0 + b_2\,LK + (b_1 + b_3\,LK) \times BK + e \quad (2.2)$$

In this equation, $(b_1 + b_3\,LK) \times BK$ specifies the moderation of language knowledge on the relation of background knowledge to nursing English reading performance. Hence, the actual magnitude of background knowledge effect on nursing English reading performance (M_{L1}) depends on language knowledge in such a way that:

$$M_{L1} = (b_1 + b_3\,LK) \times BK \quad (2.3)$$

The moderation pattern specified in Equation (2.3) is, however, insufficient when the moderation becomes more complex, for example, when the moderation functions in a U-shaped or a reversed U-shaped pattern. A reversed U-shaped pattern indeed corresponds to the two-thresholds hypothesis, which posits that background knowledge use effect is strongest for students with medium language proficiency while weaker (or nonsignificant) for students falling below or above this range. Mathematically, the nonlinear interference of language knowledge with the relation of background knowledge to nursing English reading can be expressed using the product between language knowledge squared and background knowledge: $LK^2 \times BK$. Hence, as a third step, the following equation can be constructed:

$$NERA = b_0 + b_1\,LK + b_2\,BK + b_3BK \times LK + b_4\,BK \times LK^2 + e \quad (3.1)$$

Transformed, Equation (3.1) becomes

$$NERA = b_0 + b_2\,LK + (\underline{b_1 + b_3\,LK + b_4\,LK^2}) \times BK + e \quad (3.2)$$

The actual magnitude of background knowledge on nursing English reading ability and its relation to the change of language knowledge is now expressed by the underscored part of the equation as

$$M_{L2} = (b_1 + b_3\,LK + b_4\,LK^2) \times BK \quad (3.3)$$

A significant feature of the MLMA approach is that, the final shape of the modeling is determined by the significance of the higher-layered moderation term. In the current study, the two-layered moderation term fit the data most idealistically, which is captured by Equation (3.4). Equation (3.4) with unstandardized estimates now becomes

$$NERA = 0.52 \times LK + 0.25 \times NK + 0.08 \times NK \times LK - 0.10 \times NK \times LK^2 + 0.43 \quad (3.4)$$

To facilitate interpretation, the estimates can be standardized by multiplying the standardized deviation of the related variable(s) and then divided by the standardized deviation of the endogenous variable (i.e., nursing English reading) (for detailed description of this approach please refer to Muthén, 2012). This resulted into the following equivalent form of Equation (3.4):

$$_{NERA} = 0.47 \times LK + (\underline{0.26 + 0.10 \times LK - 0.05 \times LK^2}) \times NK + 0.55 \quad (3.5)$$

Building on our earlier discussion, the moderation relationship between background knowledge effect and its moderator language knowledge is expressed as

$$M_{L2} = 0.26 + 0.10 \times LK - 0.05 \times LK^2 \quad (3.6)$$

The next part plots this relationship and provides relevant interpretation.

Phase 2. Plotting and interpretation

A particular feature of the MLMA approach is its ability to detect the continuous projection of the concurrent change of modeled variables (i.e., the changing effect of background knowledge with the change of language knowledge). This strength not only provides the researcher a general view of the dynamic change of the projection (e.g., the switch between ups and

downs), but also produces more detailed information such as where the effect of background knowledge trends up or down, and where the effect is negative, zero or positive. Mathematically, this information can be obtained by factoring Equation (3.6):

$$M_{L2} = -0.05\,(-2.51 + LK)\,(0.71 + LK)\,(1.79 + LK) \quad (3.7)$$

Drawing on Equation (3.7), one can further obtain the locations of two sets of interesting points: one set consisting of moments where the effect of background knowledge was zero: (−1.79, 0), (−0.71, 0), and (2.51, 0). Substantively, these points represented the particular language knowledge levels that the effect of background knowledge became null. Note that the actual range of language knowledge of the participants was from −1.52 to 1.52. Hence, the points at the extremes only conceptually existed. The other set of points were (−1.29, −0.06) and (1.29, 0.38), indicating that the effect of background knowledge on nursing English reading had the minimum magnitude of −0.06 when language knowledge was at the level of −1.29 and reached its maximum when language knowledge increased to 1.29.

Plotted, the relationship specified in Equation (3.4) (with the language range from −1.52 to 1.52) shows the pattern in Fig. 1. Note that the x-axis represents the continuum of language knowledge and the y-axis represents the magnitude of background knowledge effect on nursing English reading ability.

References

ACTFL. (2012). *The American Council on the teaching of Foreign languages (ACTFL) proficiency Guidelines 2012*. Retrieved from http://www.actfl.org/sites/default/files/pdfs/public/ACTFLProficiencyGuidelines2012_FINAL.pdf.

Alderson, J. C., & Urquhart, A. H. (1985). The effect of students' academic discipline on their performance on ESP reading tests. *Language Testing, 2*(2), 192–204.

Alderson, J. C., & Urquhart, A. H. (1988). This test is unfair: I'm not an economist. In P. L. Carrell, J. Devine, & D. E. Eskey (Eds.), *Interactive approaches to second language reading* (pp. 168–182). Cambridge: Cambridge University Press.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*, 238–246.

Byrne, B. M. (2013). *Structural equation modeling with AMOS: Basic concepts, applications, and programming*. New York: Routledge.

Cai, Y. (2013). *Modeling ESP ability in reading: A focus on interaction among grammatical knowledge, background knowledge and strategic competence* (Unpublished doctoral thesis). Hong Kong, China: The University of Hong Kong.

Cai, Y. (2014). Comparing two theories of grammatical knowledge assessment: A bifactor-MIRT analysis. *Language Learning in Higher Education, 4*(1), 59–76.

Cai, Y. (2015). The value of using test responses data for content validity: An application of the bifactor-MIRT to a nursing knowledge test. *Nurse Education Today, 35*(2), 1181–1185. https://doi.org/10.1016/j.nedt.2015.05.014.

Cai, Y., & Kunnan, A. J. (2018). Examining the inseparability of content knowledge from LSP reading ability: An approach combining bifactor-multidimensional item response theory and structural equation modeling. *Language Assessment Quarterly, 15*(2), 109–129.

Carrell, P., & Eisterhold, J. C. (1983). Schema theory and ESL reading pedagogy. *Tesol Quarterly, 17*(4), 553–573.

Clapham, C. (1996). *The development of IELTS: A study of the effect of background on reading comprehension*. Cambridge: Cambridge University Press.

Clarke, M. A. (1980). The short circuit hypothesis of ESL reading-or when language competence interferes with reading performance. *The Modern Language Journal, 64*(2), 203–209.

Council of Europe. (2011). *Common european Framework of reference for languages: Learning, teaching, assessment*. Retrieved from http://www.coe.int/t/dg4/linguistic/source/framework_en.pdf.

Csikszentmihalyi, M. (1990). *Flow: The psychology of optimal experience*. New York: Harper and Row.

Douglas, D. (2000). *Assessing language for specific purposes*. Cambridge: Cambridge University Press.

Geiser, C. (2012). *Data analysis with Mplus*. New York and London: Guilford Press.

Hoyle, R. H. (1995). *Structural equation modeling: Concepts, issues, and applications*. New York: SAGE publications, Inc.

Hu, L.t., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1–55.

Joreskog, K. G., Olsson, U. H., & Wallentin, F. Y. (2016). *Multivariate analysis with LISREL*. New York: Springer.

Klein, A., & Moosbrugger, H. (2000). Maximum likelihood estimation of latent interaction effects with the LMS method. *Psychometrika, 65*(4), 457–474.

Krekeler, C. (2006). Language for special academic purposes (LSAP) testing: The effect of background knowledge revisited. *Language Testing, 23*(1), 99–130 (Retrieved from).

Lee, J. W., & Schallert, D. L. (1997). The relative contribution of L2 language proficiency and L1 reading ability to L2 reading performance: A test of the threshold hypothesis in an EFL context. *Tesol Quarterly, 31*(4), 713–739.

Lin, Y.-H., & Chern, C.-L. (2014). The effects of background knowledge and L2 reading proficiency on Taiwanese university students' summarization performance. *Contemporary Educational Research Quarterly, 22*(4), 149–186.

van der Linen, W. J., & Hambleton, R. K. (Eds.). (1997). *Handbook of modern item response theory*. New York: Springer.

Madrid, R., & Cañas, J. J. (2009). The effect of reading strategies and prior knowledge on cognitive load and learning with hypertext. *The Ergonomics Open Journal, 2*, 124–132.

Marsh, H. W., Hau, K.-T., Wen, Z., Nagengast, B., & Morin, A. J. S. (2013). Moderation. In T. D. Little (Ed.), *The oxford handbook of quantitative methods in psychology* (Vol. 2, pp. 361–387). Oxford and New York: Oxford University Press. Statistical Analysis (Vol. 2.

Marsh, H. W., Wen, Z., & Hau, K. T. (2004). Structural equation models of latent interactions: Evaluation of alternative estimation strategies and indicator construction. *Psychological Methods, 9*(3), 275–300.

METS. (2007). *Medical English test System level two (for nurses)*. Shanghai: Higher Education Press.

MOH. (2008). Test syllabus for the national qualification examination for nurse Practitioners (NQENP). In NQENPC (Ed.), *Test preparation guide for the national qualification examination for nurse Practitioners* (pp. 1–45). Beijing: People's Medical Publishing House.

Mueller, R. O., & Hancock, G. R. (2010). Structural equation modeling. In G. R. Hancock, & R. O. Mueller (Eds.), *The reviewer's guide to quantitative methods in the social sciences* (pp. 373–383). New York: Routledge.

Muthén, B. Q. (2012). *Latent variable interactions*. Retrieved from http://statmodel2.com/download/Latent%20variable%20interactions.pdf.

Muthén, L. K., & Muthén, B. Q. (1998-2017). *Mplus user's guide (8th)* (8th ed.). Los Angeles, CA: Muthén & Muthén.

Muthén, L. K., & Muthén, B. Q. (1998-2018). *Mplus 8.2 [computer software]*. LA: Muthén & Muthén.

NEEA. (2007). *Publich English test system-level two*. Shanghai: Higher Education Press.

Perfetti, C. A., & Hart, L. (2001). The lexical basis of comprehension skill. In D. S. Gorfien (Ed.), *On the consequences of meaning selection: Perspectives on resolving lexical ambiguity* (pp. 67–86). Washington, DC: American Psychological Association.

Purpura, J. E. (2004). *Assessing grammar*. Cambridge: Cambridge University Press.

Reckase, M. D. (2009). *Multidimensional item response theory*. London, New York: Springer Verlag.

Ridgway, T. (1997). Thresholds of the background knowledge effect in foreign language reading. *Reading in a Foreign Language, 11*(1), 151–168.

Steiger, J. H., & Lind, J. C. (1980). Statistically based tests for the number of common factors. In *Paper presented at the the Psychometric Society annual meeting* (Iowa City, IA).
Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38*(1), 1–10.
Usó-Juan, E. (2006). The compensatory nature of discipline-related knowledge and English-language proficiency in reading English for Academic purposes. *The Modern Language Journal, 90*(2), 210–227.

Yuyang CAI is Professor, School of Languages, Shanghai University of International Business and Economics. He had a PhD in language testing and assessment. His areas of research interest include educational psychology and quantitative methods. He had rich experience developing English for Specific/Academic Purposes assessments in mainland China and Hong Kong.

Antony KUNNAN is Professor, Department of English, The University of Macau. His research interests include language assessment, ethics, language policy and statistics. He was the founding editor of language Assessment Quarterly and the founding president of Asian Association of Language Assessment.