IMPLEMENTING INTEGRATED FORMAT IN ASSESSING READING COMPREHENSION: A CASE OF VIETNAMESE EFL LEARNERS

Trinh Ngoc Thanh*

Ho Chi Minh City University of Technology and Education
01 Vo Van Ngan Street, Linh Chieu Ward, Thu Duc City, Ho Chi Minh City

Abstract: The present study evaluates the effect of test format on the performance of reading comprehension, which is the integrated format. Unlike the separation of text and test questions into two sections in the split format, the main modification of the integrated format is that the relevant text is integrated with test questions in each reading task. Through the comparison between learners' performance in the two test formats, this study tests the hypothesis that the overall test performance and task performance in the integrated format are higher than performance in the split format. Drawn on score data of 20 Vietnamese EFL learners, findings from the study showed no effect of test format on the overall test performance and marginally significant effect of test format on task performance. A further analysis on relevant aspects of test design is to be discussed.

Key words: cognitive load theory, split format, integrated format, reading comprehension

1. Introduction

The essence of comprehension is often observed in an indirect manner since the moment of comprehending a text takes place within a short time lapse (Pearson & Cervetti, 2017). In the notion of indirect observation, reading comprehension can be better assessed in the form of instruction as teachers can create teaching activities which can monitor comprehension or explicit teaching of reading skills and strategies (Palincsar & Brown, 1984; Afflerbach, Pearson & Paris, 2008). In the facilitation of reading skills and strategies, it was viewed reading comprehension itself is not a single entity involving reading ability but a combination of reader, text, and task factors (Kamhi & Catts, 2017) as well as the considerations of variables such as "content knowledge, motivation and interest, text organization, nature and content of the task, and characteristics of the setting in which reading occurs" (Lipson & Wixson, 1986; as cited in Kamhi & Catts, 2017, p. 1).

However, in the consideration of specific groups of learners, Carlson, Seipei, and McMaster (2014) indicated the necessity to analyze comprehension difficulties when learners move up to new requirements in their learning levels and the attention to individual needs in the mastery of reading skill. It could be explained that the assessment of learner's performance in different reading conditions is more important than the control of variables in the measurement of comprehension abilities (Lipson & Wixson, 1986; Wixson, 2017). Furthermore, some other listed factors causing comprehension difficulties could be the type of necessary knowledge and the role of inference making (Pearson & Cervetti, 2017), the process of text decoding from bottom-up direction (Kintsch, 1998), and the process of meaning-making in the mental model (Carlson et al., 2014).

One particular aspect concerning the reduction of comprehension difficulties is the extent to which the format of instruction plays a role in allocating the necessary amount of cognitive loads given for the learning tasks (Chandler & Sweller, 1991). In particular, it is the extent to which instructions are likely to produce higher amount of intrinsic cognitive load (i.e. "the intrinsic nature of the learning task") and reduce the amount of extraneous cognitive load (i.e. "the manner in which the task is presented") while learners approach the learning materials (Van Merrienboer & Sweller, 2005, p. 150). The borderline between the intrinsic and extraneous cognitive load

────────
* Tel.: 0837887889
Email: thanhtn@hcmute.edu.vn

however possibly results in split-attention effect, the result of which is distraction when there are two separate sources of information presented in reading materials (Yeung, Jin & Sweller, 1997).

To reduce split-attention effect in reading materials, there should be an effective initiation for mental integration through referents. For instance, Sweller et al. (1998) reported a replacement for "a single, integrated source of information" instead of "multiple sources of information" (as cited in Van Merrienboer & Sweller, 2005, p. 150). In this replacement, the physical integration of isolated materials should be reformatted to reduce "the need to search for relevant referents and mentally integrate them" (Chandler & Sweller, 1991, p. 295) and to facilitate automation in learning-mediated intellectual performance after the acquisition of schemas (Sweller, 1994). Therefore, an attention to selecting appropriate elements for reading materials (Chandler & Sweller, 1991) can be also considered the key to enhance the intrinsic cognitive load for reading comprehension.

All things considered, the present study places theoretical concerns of minimizing comprehension difficulties by implementing the integrated format into the purpose of assessing reading comprehension. This study particularly examines whether the physical integration of appropriate text and comprehension questions can improve the overall test performance (i.e. the total score of test) and task performance (i.e. the score of the designed reading tasks) in reading comprehension. Furthermore, it also analyses relevant aspects of test design as contributors to the appearance of possible extraneous cognitive load.

2. Rationale of the present study

2.1. The original study

The present study replicates Huynh (2015)'s investigation on how to reduce extraneous cognitive load for reading assessment in classroom context. 21 Vietnamese EFL students were randomly allocated into split-attention and integrated instruction formats in both the learning and testing phases.

In terms of research design, participants in Huynh (2015)'s study took part in the testing phase right after the learning phase. Choosing the same text "The early aborigines" as the reading material for the learning and testing phases, Huynh designed 10 questions for the learning phase and 12 questions for the testing phase. Concerning the design of the two test formats, it was noted that reading passage and the set of reading comprehension questions were physically separated as two sections in the split format of the reading questions. Meanwhile, smaller sets of reading comprehension questions are physically placed into relevant parts of the text in the integrated format. On the comparison of mean scores of assessment result in learning and testing phases, Huynh claimed the effectiveness of the integrated format in reducing extraneous cognitive load. In comparison with the split format, the integrated format proved its efficiency in reading performance in the learning phase and later on the testing phase.

Certain limitations were identified from the original study. The learning phase was immediately followed by the testing phase and even though there could be no significant interaction between the two phases, this condition could lead to cognitive effort in memorizing from the learning phase to answer the reading questions in a relatively short time span. Furthermore, the set of questions for reading comprehension in both the learning and testing phases required learners to provide written answers. The written form raised a concern of appropriate scoring due to a variety of responses and a lack of standard answers.

To remediate the above identified limitations, the research procedure in the present study would be modified as follows. Instead of administering learning and testing on the same day, the present study separated the two phases and designed more reading tasks for the test. The selection of various reading tasks possibly would reduce the flexibility of answers and thus enhance the standard of scoring.

2.2. Research questions

Following a similar research design in Huynh (2015)'s study, the present study evaluates the hypothesis whether the physical integration of relevant text with the reading questions in the integrated format could improve overall test performance (i.e. the total score of test) and task performance (i.e. the score of the designed reading tasks) in comparison with the separation between text and reading questions in the split format. For the purpose of hypothesis testing, two research questions for the present study are as follows:

RQ1: Is there a difference in the overall test performance between the integrated format and the split format?

RQ2: Is there a difference in task performance between the integrated format and the split format?

2.3. Data collection

This study employed the collection of secondary data from the homeroom teacher as the main provider of data. The main purpose was to reduce intrusive effects caused by the procedure of data collection. Furthermore, it would lead to a concern for conflict of interests if the researcher had direct contact with the participants.

The set of secondary data was collected from a group of 20 first-year English major students enrolling in an English reading course entitled Reading 1 at a local university in Ho Chi Minh City. Prior to the procedure of data collection, the participants enrolled in Reading 1 for six weeks of the first semester in 2015. In terms of reading ability, they should have mastered adequate training for reading skills and strategies for reading comprehension in order to complete all the procedures of data collection.

Discussion with the homeroom teacher was conducted to ensure that the designed instruments were appropriate to be used in the classroom or to be revised if necessary. A letter of consent was also sent to the homeroom teacher to seek for the agreement as the data provider and to indicate necessary actions for maintaining the code of research ethics.

2.4. Research instrument

An online free-access article entitled "Robin Hood: Fact or Fiction" from Linguapress publisher was chosen as the reading material for this study. This article was indicated for intermediate student-level and the text length was about 530 words. The present study incorporated IELTS reading test format into the design of reading tasks in the 20-minute reading mini-test at the classroom. Table 1 describes the task design of the test.

In terms of format, there were two versions to be used in the present study: the split format (test form A) and the integrated format (test form B). All three reading tasks were inserted after the reading text in the split format while relevant reading questions were physically integrated into relevant paragraphs of the reading text in the integrated format. For the purpose of scoring, each correct answer received 1 mark and 0 for each incorrect answer and the maximum score for each test version was 12.

2.5. Data analysis

The following steps were proposed to answer the two research questions. First, an Excel file was prepared to record the following information: name, student ID, diagnostic test result, test format (split-A or integrated-B), overall performance score (i.e. the total score out of 12), individual task performance (i.e. the score of task 1, task 2, and task 3-see Table 1), and records of wrong answers in each task. Second, the quantitative analysis for comparing the two test formats was conducted using one-way ANOVA Welch test because this test is appropriate for small sample size (less than 30). Finally, the discussion of overall performance and task performance was conducted on the basis of (1) calculation for the dependability of test items, (2) analysis of wrong answers, and (3) the writing of test items.

Table 1

Criterion-Referenced Design of the Classroom Reading Mini-Test

Task	Test items	Theoretical grounds	Description
1	Multiple choice questions (MCQs)	*MCQs* are common in the assessment for group settings and easy to administer with one correct choice and alternative distracters (Carlson et al., 2014). While answering MCQs, the retrieval of relevant information from the text is the major cognitive activity involved in the determination of the correct choice (Ferrer et al., 2017).	This task involves the selection of the correct answer among four choices (A, B, C, or D).
2	Validating (True-False-Not Given)	*Validating questions* are grounded on validation in text comprehension. Validation refers to a mechanism where readers are involved in the main cognitive activity of judging the plausibility of the given information (Richter, 2015) and balancing the controversies from inconsistent information in the mental representation (Richter & Maier, 2017).	This task involves the main activity of validating the accuracy of the given statements. The value True is applicable if the statement agrees with the information and False if the statement contradicts the information. Meanwhile, the value Not Given is defined when there is no information derived from the text on this statement.
3	Cloze-test (Fill in the blank with no more than two words)	*Cloze-test questions* include the gap-filling of the appropriate word and are applicable to assess reading comprehension for both lower and higher level of learners (Mizumoto, Ikeda & Takeuchi, 2016).Considering text readability, test designers may modify the original text across levels of learners (Crossley et al., 2017).	This task requires test-takers to fill in the gap with the appropriate word or group of words. The answer is limited within the length of two words and the gap-filled words should be appropriate and grammatical.

3. Findings

3.1. The integrated format and overall test performance

Due to the method of scoring (1 for correct and 0 for incorrect answers), the range of overall performance was supposed not to be considerably varied among individual performances. Therefore, with reference to results from the calculation of means and standard deviations (SDs), it was decided that the range of cut-score would be between the lower and upper levels of 1.6 SD for both the integrated and split formats. As a result, the range of the cut-score of split group was $3 <$ score split < 8 (N=11, M=5.27, SD=1.76);

meanwhile, the cut-score of integrated group was $5 <$ score integrated < 9 (N=9, M=6.78, SD=1.55).

Out of 20 participants, there were in total 3 outliers to be eliminated before the computation of ANOVA analysis. Two participants from the split and integrated groups achieved the overall score below the range of cut scores: their overall scores were 2 (split) and 4 (integrated) respectively. The other was eliminated because the performance of task 1 was uniquely recorded as zero. After eliminating outliers of overall test performance from both groups, 17 participants from both groups (N_{split}=9 & $N_{integrated}$=8) were left for

one-way ANOVA Welch analysis.

One-way Welch ANOVA test was computed with regards to small and unequal sample size for each group. The significant level of *p* value is set at 0.05. Although integrated group (M=7.12, SD=1.36) achieved a higher mean of overall test performance than split group (M=5.9, SD=1.36), there were no significant differences between group means as determined by one-way Welch ANOVA ($F(1,15)=3.496$, $p>0.05$). It can be concluded that there is no difference in the overall test performance between the two test formats.

Table 2

One-way Welch ANOVA for Overall Test Performance

	Sum of squares	df	Mean square	F	Sig.
Between groups	6.471	1	6.471	3.496	.081
Within groups	27.764	15	1.851		
Total	34.235	16			

3.2. The integrated format and task performance

The evaluation of the effect of integrated format on task performance was structuralized at individual (i.e. one single task) and collective (i.e. a pair of tasks) levels. Table 3 presents the report of mean score of task performance in split and integrated formats. The mean scores suggest a higher performance of task performance for the integrated format at individual and collective levels.

Although integrated group achieved a higher mean of task performance than split group across individual and collective levels, results from Table 4 overall indicated that the differences between group means for individual task 2 (validating questions) (F(1,15)=4.642, p=0.048<0.05) and for the pair of task 2 (validating questions) & task 3 (cloze-test questions) (F(1,15)=4.644, p=0.048<0.05) were marginally significant.

Table 3

Mean Score of Task Performance in Split and Integrated Formats

Test format		Task 1	Task 2	Task 3	Task 1 & 2	Task 1 & 3	Task 2 & 3
Split	Mean	2.00	1.33	2.56	3.33	4.56	3.89
	N	9	9	9	9	9	9
	SD	.866	.500	.726	1.000	1.333	.782
Integrated	Mean	2.12	2.25	2.75	4.38	4.88	5.00
	N	8	8	8	8	8	8
	SD	.641	1.165	1.165	1.188	1.356	1.309
Total	Mean	2.06	1.76	2.65	3.82	4.71	4.41
	N	17	17	17	17	17	17
	SD	.748	.970	.931	1.185	1.312	1.176

Table 4

Task Performance at Individual and Collective Measurements

		Sum of squares	Df	Mean square	F	Sig.
Task 1	Between groups	.066	1	.066	.112	.743
	Within groups	8.875	15	.592		
	Total	8.941	16			
Task 2	Between groups	3.559	1	3.559	4.642	.048
	Within groups	11.500	15	.767		
	Total	15.059	16			
Task 3	Between groups	.160	1	.160	.175	.682

		13.722	15	.915		
	Within groups	13.722	15	.915		
	Total	13.882	16			
	Between groups	4.596	1	4.596	3.856	.068
Task 1 & 2	Within groups	17.875	15	1.192		
	Total	22.471	16			
	Between groups	.432	1	.432	.239	.632
Task 1 & 3	Within groups	27.097	15	1.806		
	Total	27.529	16			
	Between groups	5.229	1	5.229	4.644	.048
Task 2 & 3	Within groups	16.889	15	1.126		
	Total	22.118	16			

4. Discussion

4.1. Calculation of dependability index

In the previous one-way Welch ANOVA analysis of overall test performance, no significant difference was found in the group means of the score between the integrated format and the split format. It is assumed that there should be an effect of the dependability of test administration on overall test performance. Phi-lambda ($\varphi\lambda$) index was computed with reference to the following statistics from the original data of overall test performance from 20 participants left for the determination of cut-score: the number of items on the test (K=12), mean of the proportion scores as measured by the average of proportion of correct answers (Xp=0.5), standard deviation of the proportion scores (Sp=0.15), and cut-point expressed as a proportion (λ=0.6) (see Fulcher, 2010 for more details on the calculation of these statistics).

Applying the formula for calculating Phi Lambda index, the dependability index remained at 0.37 at the cut-score of 6. The dependability level at 0.37 could be interpreted that the test administration with split and integrated format reports some agreement value of dependability, and therefore results in no significant difference in the group means of the score for overall test performance.

4.2. Analysis of test items

The second finding concerning the effect of integrated format on task performance revealed a remarkable borderline to significant level of the individual task 2 (validating questions) and the pair of task 2 and task 3 (cloze-test questions). This finding assumed that the integrated format for the same sets of questions lowered the cognitive load for task performance. To clarify this claim, the frequency of wrong answers from 17 participants who were left after the determination of cut-score for the computation of ANOVA analysis was recorded in Table 5.

Statistics in Table 5 demonstrated a constant reduction of the number of wrong answers for task 2 in the integrated format. Except for item 3, the reduced number of wrong answers is consistent for other items in task 3.

In the evaluation of pairs of tasks, the constant reduction of wrong answers was found in items 1, 2, and 4 of both task 2 and 3 in the integrated format. These statistics also suggest that item 4 of task 2 is the most challenging question in the split format because 9 out 9 participants failed to answer this question. Also, item 3 in task 3 in the integrated format may attain problems with the identification of the correct answer for the participants.

Table 5

Statistics of Wrong Answers for Split and Integrated Format

Test Format	Test Items	Task 2 (x_1)	Task 3 (x_2)	∑ Task 2 3 ($\sum x_1 \ x_2$)	Summary	Items	Task 2 $x_{1 \ split} -$ $x_{1 \ integrated}$	Task 3 $x_{2 \ split} -$ $x_{2 \ integrated}$	∑ Task 2 3
split (n=9)	1	8	6	14	Reduced wrong answers (raw data)	1	4	2	6
	2	2	5	7		2	2	2	4
	3	5	0	5		3	1	-3	-2
	4	9	4	13		4	3	4	7
integrated (n=8)	1	4	4	8	Reduced wrong answers (%)	1	50%	33%	43%
	2	0	3	3		2	100%	40%	60%
	3	4	3	7		3	20%	N/A	-40%
	4	6	0	6		4	33%	100%	55%

Table 6 showed further analysis of validating questions in task 2. Higher frequency of task performance score at 1 and 2 was found a popular norm in the split format.

Table 6

Frequency of Wrong Answers in Task 2 From Both Formats

Task 2	Test Format Split	Test Format Integrated	Frequency
Item 1	6	3	9
Item 2	3	1	4
Item 3	0	3	3
Item 4	0	1	1
Total	9	8	17

4.3. The writing of test items

The statistics of item performance in Table 5 and Table 6 indicated the challenging level of the following test items: item 1 and item 4 in task 2 of the split format and item 3 in task 3 of the integrated format. Further analyses of the design for these test items are as follows:

4.3.1. Item 1 of task 2

Extracted text: Other stories claim that Robin was not an Anglo Saxon nobleman, but a common fugitive; they say that his real name was "Robert Hood", and that he only fought against his personal enemies, in particular the Sheriff of Nottingham, not against the Normans.

Item 1: The Sheriff of Nottingham were not the only enemies for Robin Hood. _____

Data on frequency of wrong answers from 6 participants achieving the score of 1 in the split group showed that these participants provided wrong answers for item 1 of task 2.

Answer: False

The writing of item 1 may fail to consider the identification of "other stories" for the determination of false value for the above written statement. Furthermore, the presentation of facts in the previous paragraph (paragraph 8) in line with this paragraph may puzzle participants whether there should be two enemies (the Sheriff of Nottingham, the Normans) for Robin Hood rather than one (the Sheriff of Nottingham). The addition of the phrase "In other stories" should have been included in the above statement for further clarity. It is also noted that the item writing in negative condition may also have contributed to the puzzlement for participants to determine the correct answer for item 1 of task 2.

4.3.2. Item 4 of task 2

Extracted text: Many old stories said that Robin lived in Yorkshire. However, later stories had him living in Sherwood Forest, near Nottingham; and today, Robin's name is definitely attached to the city of Nottingham, and to Sherwood Forest.

Structure: Sentence 1 [past simple]. <Compound sentence> However, sentence 2 [past simple PP1 PP2]; and today, sentence 3 [(source of answer) past simple PP3 PP4]

Item 4: There is a conclusion about where Robin Hood lived._____

Answer: True

Item 4 demands inference skill to answer the question because it is necessary to synthesize relevant details from the extracted test. The key word "a conclusion" in the item is initially designed to match with the source of answer from the phrase "is definitely attached" from the line "… and today, Robin's name is definitely attached to the city of Nottingham, and to Sherwood Forest."

There are three possible causes for the challenging level of item 4. First, it could be explained that the mixture of verb tense in item 4 is probably the potential factor which imposed cognitive demands for participants in the split group. The past tense "lived" in the noun clause of the prepositional phrase "…about where Robin Hood lived" of the test item may direct the attention to refer to sentence 1 [...that Robin lived in Yorkshire…]. Another explanation is due to the assumption that participants may have difficulty in processing a lengthy compound sentence where the clue to answer the question resided. Furthermore, the repetition of the proper nouns "Sherwood Forest" in PP1 and PP4 and "Nottingham" in PP2 and PP3 is also predicted to distract the attention to the relevant detail.

Concerning the clause "Robin's name is definitely attached", the revision from "a conclusion" to "a definite conclusion" could have reduced the frequency of wrong answers for item 4. Consideration for transforming words carrying the same root will provide more specific clues for participants to answer this item.

4.3.3 Item 3 of task 3

Extracted text: In Nottingham, Robin is now a very popular character. Visitors to the city can learn all about him at the "Tales of Robin Hood" exhibition, where Robin and his adventures are brought to life; and in Sherwood Forest, "the Major Oak", a massive old tree, is said to be Robin Hood's tree.

Structure: Sentence 1 [PP1 Main clause 1]. <Compound sentence> Sentence 2 [Main clause 2, Relative clause]; and PP2, Sentence 3[Subject (Appositive) + Passive Voice]

Item 3:_____is said to be the location of Robin Hood's tree.

Answer: Sherwood Forest

Item 3 requires participants to locate the proper noun "Sherwood Forest" from PP2 in order to fill in the gap with no more than two words. Although there was a physical integration of the relevant text boundary, the prepositional phrase "In Nottingham" in PP1 may distract participants to provide a correct answer for item 3. Another possibility could be addressed to the complexity of the compound sentence structure where the details with distractors come from the relative clause "…where Robin and his adventures are brought to life…", the appositive "…a massive old tree…", and the passive voice of sentence 3. Therefore, revising or simplifying sentences with complex structure is suggested for the purpose of eliminating unnecessary cognitive loads which potentially prohibit the process of reading comprehension.

5. Conclusion

The present study investigates whether the administration of physical integration of questions into its relevant part would improve reading comprehension performance. Replicating similar administration with modification in research design in Huynh's study (2015), findings from the present study further suggests relevant contributors to extraneous cognitive loads for the processing of the reading test materials, namely the design of reading tasks and the writing of test items.

Findings from the present study lead to the consideration on how task performance could be influenced by the writing of test items. Drawn from theory-based validity in language testing (Weir, 2005), these factors suggest more collection for prior evidence before assessing reading comprehension. Furthermore, this study also raises the concern to increase the dependability index in reading assessment and the increase of Phi-Lambda could contribute to the validation of test items in reading assessment (Ross & Hua, 1994).

Certain limitations are identified in the present study. First, the small size number of participants and the scoring method may limit the scatter plot of performance score, considerably affecting the calculation of mean and standard deviation. Moreover, findings in this study are subject to task design and the

recruited group of participants Therefore, generalizability to other contexts should take these factors into consideration.

References

Afflerbach, P., Pearson, P. D., & Paris, S. (2008). Clarifying differences between reading skills and reading strategies. *The Reading Teacher*, *61*(5), 364–373. https://doi.org/10.1598/RT.61.5.1

Carlson, S. E., Seipel, B., & McMaster, K. (2014). Development of a new reading comprehension assessment: Identifying comprehension differences among readers. *Learning and Individual Differences*, *32*, 40-53. https://doi.org/10.1016/j.lindif.2014.03.003

Chandler, P., & Sweller, J. (1991). Cognitive load theory and the format of instruction. *Cognition and instruction*, *8*(4), 293-332. https://doi.org/10.1207/s1532690xci0804_2

Crossley, S. A., Skalicky, S., Dascalu, M., McNamara, D. S., & Kyle, K. (2017). Predicting text comprehension, processing, and familiarity in adult readers: New approaches to readability formulas. *Discourse Processes*, *54*(5-6), 340-359. https://doi.org/10.1080/0163853X.2017.1296264

Ferrer, A., Vidal-Abarca, E., Serrano, M. A., & Gilabert, R. (2017). Impact of text availability and question format on reading comprehension processes. *Contemporary Educational Psychology*, *51*, 404-415. https://doi.org/10.1016/j.cedpsych.2017.10.002

Fulcher, G. (2010). *Practical Language Testing*. Hodder Education.

Huynh, C. M. H. (2015). Split-attention in reading comprehension: A case of English as a foreign/second language. In *6th International Conference on TESOL* (pp. 1-12).

Kamhi, A. G., & Catts, H. W. (2017). Epilogue: Reading comprehension is not a single ability - Implications for assessment and instruction. *Language, Speech, and Hearing Services in Schools*, *48*(2), 104-107. https://doi.org/10.1044/2017_LSHSS-16-0049

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge University Press.

Lipson, M. Y., & Wixson, K. K. (1986). Reading disability research: An interactionist perspective. *Review of Educational Research*, *56*(1), 111-136. https://doi.org/10.3102/00346543056001111

Mizumoto, A., Ikeda, M., & Takeuchi, O. (2016). A comparison of cognitive processing during cloze and multiple-choice reading tests using brain activation. *ARELE: Annual Review of English Language Education in Japan*, *27*, 65-80. https://doi.org/10.20581/arele.27.0_65

Palincsar, A. M., & Brown, A. L. (1984). Reciprocal teaching of comprehension - fostering and comprehension - monitoring activities. *Cognition and Instruction*, *1*(2), 117–175. https://doi.org/10.1207/s1532690xci0102_1

Pearson, P. D., & Cervetti, G. N. (2017). The roots of reading comprehension instruction. In S. E. Israel (Ed.), *Handbook on reading comprehension* (2nd ed., p. 12-56). The Guilford Press.

Richter, T. (2015). Validation and comprehension of text information: Two sides of the same coin. *Discourse Processes*, *52*(5-6), 337-355. https://doi.org/10.1080/0163853X.2015.1025665

Richter, T., & Maier, J. (2017). Comprehension of multiple documents with conflicting information: A two-step model of validation. *Educational Psychologist*, *52*(3), 148-166. https://doi.org/10.1080/00461520.2017.1322968

Ross, S., & Hua, T. F. (1994). An approach to gain score dependability and validity for criterion-referenced language tests. In *The 6th Annual Language Testing Research Colloquium* (pp. 1-28). Educational resources information center.

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction*, *4*(4), 295-312. https://doi.org/10.1016/0959-4752(94)90003-5

Sweller, J., Van Merrienboer, J. J., & Paas, F. G. (1998). Cognitive architecture and instructional design. *Educational Psychology Review*, *10*(3), 251-296. https://doi.org/10.1023/A:1022193728205

Van Merrienboer, J. J., & Sweller, J. (2005). Cognitive load theory and complex learning: Recent developments and future directions. *Educational Psychology Review*, *17*(2), 147-177. https://doi.org/10.1007/s10648-005-3951-0

Weir, C. J. (2005). *Language Testing and Validation*. Macmillan.

Wixson, K. K. (2017). An interactive view of reading comprehension: Implications for assessment. *Language, Speech, and Hearing Services in schools*, *48*(2), 77-83. https://doi.org/10.1044/2017_LSHSS-16-0030

Yeung, A. S., Jin, P., & Sweller, J. (1998). Cognitive load and learner expertise: Split-attention and redundancy effects in reading with explanatory notes. *Contemporary educational psychology*, *23*(1), 1-21.

Reading material

Robin Hood: Fact or fiction. (2002). Retrieved from http://linguapress.com/intermediate/robin-hood.htm

Appendix

Formula of calculation

1. $Reduced\ wrong\ answers\ (raw) = x_{split} - x_{integrated}$

2. $Reduced\ wrong\ answers\ (\%) = \dfrac{x_{split} - x_{integrated}}{x_{split}} \times 100$

Notes:

x_{split} $\sum$wrong answers of task performance in split-format test

$x_{integrated}$ $\sum$wrong answers of task performance in integrated-format test

Dependability index

$$\Phi\lambda = 1 - \frac{1}{K-1}\left(\frac{\bar{X}p\,(1-\bar{X}p) - Sp^2}{(\bar{X}p - \lambda)^2 + Sp^2}\right)$$

Notes:

K *number of items on the test*

$\bar{X}p$ *mean of proportion score*

Sp *standard deviation of the proportion score*

λ *cut − point expressed as a proportion*

ỨNG DỤNG DẠNG BÀI TÍCH HỢP VÀO ĐÁNH GIÁ ĐỌC HIỂU TIẾNG ANH: TRƯỜNG HỢP NGƯỜI HỌC TẠI VIỆT NAM

Trịnh Ngọc Thành

Đại học Sư phạm Kỹ thuật Thành phố Hồ Chí Minh
01 Võ Văn Ngân, Phường Linh Chiểu, TP. Thủ Đức, TP. Hồ Chí Minh

Tóm tắt: Nghiên cứu này đánh giá tác động của dạng bài tích hợp lên kết quả kiểm tra đọc hiểu tiếng Anh. Khác với dạng bài phân vùng ở phần thiết kế tách biệt văn bản đọc hiểu với các câu hỏi đọc hiểu thành hai phần riêng biệt, dạng bài tích hợp được thiết kế bằng việc gắn kết phần trích đoạn của văn bản bài đọc với câu hỏi kiểm tra đọc hiểu tương ứng. Việc so sánh kết quả kiểm tra đọc hiểu giữa hai dạng đề nhằm kiểm định lại giả thuyết kết quả kiểm tra đọc hiểu của dạng bài tích hợp cao hơn so với dạng bài phân vùng. Kết quả nghiên cứu từ 20 người học Việt Nam cho thấy không có tác động của dạng bài tích hợp lên kết quả kiểm tra đọc hiểu. Ngoài ra, dạng bài tích hợp có tác động không đáng kể lên kết quả thực hiện nhiệm vụ đọc hiểu. Nghiên cứu cũng thực hiện các phân tích bổ sung về những khía cạnh liên quan trong việc thiết kế đề kiểm tra đọc hiểu.

Từ khóa: thuyết tải nhận thức, dạng bài phân vùng, dạng bài tích hợp, đọc hiểu