Universal Journal of Educational Research 3(7): 451-459, 2015
DOI: 10.13189/ujer.2015.030704 http://www.hrpub.org

Factors Affecting Item Difficulty in English Listening Comprehension Tests

Pei-Ju Sung, Su-Wei Lin*, Pi-Hsia Hung

Department of Education, National University of Tainan, Taiwan

Abstract Task difficulty is a critical issue affecting test developers. Controlling or balancing the item difficulty of an assessment improves its validity and discrimination. Test developers construct tests from the cognitive perspective, by making the test constructing process more scientific and efficient; thus, the scores obtained more precisely represent the proficiency level. In this paper, a framework of cognitive factors related to English listening comprehension (LC) tests is proposed. Data used were derived from the After School Alternative Program (ASAP) English test item pool. A total of 150 items were analyzed, and item difficulty was concurrently calibrated according to three-parameter-logistic item response theory from the responses of 1,459 fifth- to seventh-grade participants. Components affecting item difficulty were proposed and discussed with regard to the cognitive perspective. The 53.5% of variance in item difficulty of LC can be explained by the cognitive predicting model. This result is expected to make the task constructing procedure more organized and to ensure the task is of the desired difficulty, thus enabling the intended ability to be effectively measured. In addition, the ASAP English test aims to assist low-achieving students in remedial instruction; the cognitive factors and study results provide a reference for developing teaching materials of English LC remedial instruction and for constructing LC test items.

Keywords Listening Comprehension Test, Cognitive Factors, Item Difficulty

1. Introduction

1.1. Introduction to the Cognitive Process of Listening Comprehension

Listening is a cognitive process and, in reality, its context is not visible to the listener. Listening comprehension (LC) occurs within the mind of the listener, and the context of interpretation is the cognitive environment of the listener [1]. Assessing the listening ability is one of the least understood, least developed, and yet one of the most crucial areas of language [2]. Research on language assessment is limited and concentrated on the specific constructs or abilities underlying LC, the procedure of listening assessment construction, and the validation and evaluation of listening tests.

This study explores factors affecting LC test difficulty by using participants at age from eleven to thirteen, investigates the LC process, and identifies the key components affecting LC.

Memory plays a critical role in LC. Examinees could not reaccess the text when constructing an answer in listening comprehension tests. The examinees must process information, perceive it, and accept an incoming discourse stream simultaneously. Information processing procedure causes a heavy cognitive load. Thus, both LC and the task difficulty are associated with the extent of cognitive load.

LC is an active dynamic and invisible process. Information is processed in listeners' working memory. Listeners first receive stimuli and then begin perceiving the stimuli and retrieving the prior knowledge from schema by performing a bottom-up or top-down process for managing the information. The bottom-up process uses basic language knowledge, such as phonological awareness, vocabulary, and grammar, for understanding the spoken text. Top-down process uses background knowledge and common sense for comprehending aural stimuli. This process involves "attention to a continuous stream of speech, which is not under the timing control of the listeners" [3]; this process imposes a heavy cognitive load on listeners.

The circumstances become more severe when responding to an LC assessment under a time limit and the exam anxiety of examinees increase. The examinees hardly have enough time for performing a top-down process. Paris and Brooks [4] investigated cognitive factors in children's listening and reading comprehension and concluded that children can act on incoming information by using different approaches similar to adults.

Interests in exploring the effect of cognitive factors that contributes to item difficulties are increasing [5-8]. Brown [9]

argued that cognitive load is one of the most crucial determinants of difficulty in listening. Cognitive factors analyzed and used in tests can be employed for predicting or adjusting item difficulty as well as providing the evidence of constructs validity. The finding of cognitive factors forms a reference for constructing test items in a certain domain.

Weir [10] places the processing of acoustic and visual inputs as one of several core processes that the present literature suggests to be essential for a theoretically-grounded and empirically-informed cognitive processing framework for understanding second language (L2) listening skills. Cognitive tasks (including test items) are considered to require multiple processing stages, strategies, and knowledge stored. Both tasks and people vary on the processing components. In other words, the primary sources of processing difficulty may vary between tasks, even when the tasks belong to the same item type. Embretson indicated that the cognitive test constructing system has the some properties, such as test content is prescribed using explicit principles; score meaning is linked to the underlying cognitive process; item parameter represents the sources of cognitive complexity in a test item; and abilities are linked to processes that underlie task difficulty [11].

The cognitive structure of a test is typically defined as a set of cognitive attributes (e.g., cognitive operations and processes) required for producing correct answers on the test items [12]. Knowledge about cognitive structures can help test developers, psychologists, and educators to construct test items with desirable measurement and cognitive characteristics, operate constructs, and satisfactorily understand the cognitive processes of thinking and performance. Validating cognitive structures is a major problem and involves integrating cognitive psychology and psychometric modeling [13, 14].

In this study, we investigated the cognitive factors affecting the task difficulty of the After School Alternative Program English LC test (ASAP-ENGLCT). Test constructers can use the factors obtained here for developing tests and manipulating item difficulty to ensure satisfactory construct validity in the test.

1.2. Study Aim

This study investigated the cognitive factors affecting the item difficulty of LC based on the ASAP-ENGLCT. The ASAP is a long-term project conducted by the Taiwan government. In addition, the ASAP, through an assessment, selects low-proficiency students with a low socioeconomic status and uses remedial instruction. Each year, several experts and teachers are involved in the project for developing a large-scale computer-based standardized assessment. By finding the major components predicting item difficulty, test developers can satisfactorily control and balance the test difficulty and simultaneously increase the test validity.

We analyzed the ASAP-ENGLCT construct and explored the LC process. In addition, we identified the components affecting the English LC process. The complexity of a comprehension process affects the performance and response of examinees, consequently reflecting item difficulty. The results are helpful for constructing an English LC test.

1.3. Research Questions

The research data are based on the ASAP-ENGLCT, and the researcher explores the following questions:

1. What are the major test characteristics most likely to affect the difficulty of the ASAP-ENGLCT?
2. What is the explained variance of the cognitive factors?
3. What are the effective predictors of the difficulty of the ASAP-ENGLCT?

2. Methods

2.1. Data Description

The samples of norm were stratified random samples from Northern, central, Southern, and eastern Taiwan. Four elementary and junior high schools, three elementary and junior high schools, three elementary and junior high schools, and two elementary and junior high schools were separately sampled from each area. Participants were students from grades 5 to 7. Of the 2,195 participants, 713 were from grade 5, 746 from grade 6, and 736 from grade 7. The test score of the participants was used for calibrating the item difficulties of 150 items from the ASAP-ENGLCT item pool of grades 5, 6, and 7. The item difficulties were coded and calibrated using multiple regressions to explore the correlation between item difficulties and cognitive factors.

2.2. Item Pool Description of the ASAP-ENGLCT

Table 1. Specification of ASAP-ENGLCT item type for Grade 5 to Grade 7

Item type	Item number	Total
Identification of onset sound and rime sound	24	150
Identification of vocabulary	22	
Picture comprehension	24	
Dialogue response	38	
Dialogue comprehension	42	

Note. From Hung, P. H., et al. [15]. *After School Alternative Program Assessment Report*, Technology-Based Educational Assessment center NUTN, Tainan, Taiwan.

The ASAP-ENGLCT is a test developed by the project of ASAP technology-based testing, sponsored by the Ministry of Education of Taiwan. The data for this study were from the After School Alternative Program Assessment Project, conducted by Hung et al [15]. The test is a national test with big data pool to support the reliability and quality of the study. The item pool of the ASAP-ENGLCT is shown in Table 1. All items were developed in a multiple choice

format. The parameters of item difficulty across grades 5 to 7 items were concurrently calibrated using item response theory (IRT) and, in total, 150 items were calibrated. The summary information of the item parameters is shown in Tables 1 and 2.

Table 2. The descriptive statistic IRT parameters of ASAP-ENGLCT Grade 5 to Grade 7 (N=150)

Parameter	Mean	SD	Basic level (52 items)	Proficient level (98 items)
a (Discrimination)	1.176	0.379	1.003	1.274
b (Difficulty)	-.235	.951	-1.282	0.182
c (Pseudo-chance)	0.221	0.057	0.222	0.221

Note. From Hung, P. H., et al. [15]. *After School Alternative Program Assessment Report*, Technology-Based Educational Assessment Center NUTN, Tainan, Taiwan.

2.3. Data Structure of Dependent Variables

Item difficulty (b value in Table 2) was used as a dependent variable, and the phonetic/nonphonetic discrimination item type, number of plausible distracters, necessity for inference, content familiarity, propositions, heterogeneity of sentence patterns in options, lexical overlap in the key, lexical overlap in the distractors, and redundant information were used as independent variables. The descriptive statistics of item difficulty are shown in Table 3. The average item difficulty was -.23.

Table 3. Descriptive statistics of ASAP-ENGLCT Grade5-Grade7 item difficulty (N=150)

	Mean	Std. Deviation	Minimum	Maximum
Item difficulty	-.233	.915	-2.861	2.312

2.4. Coding Rules of Cognitive Factors

Each item was coded with nine variables, the phonetic/nonphonetic discrimination item type, number of plausible distractors, necessity for inference, content familiarity, lexical overlap in the key, lexical overlap in the distractors, propositions, heterogeneity of sentence patterns in options, and redundant information. The coding rules are explained as follows:

Variable 1: Phonetic/nonphonetic discrimination item type

Phonetic discrimination, including onset or rime sound discrimination, and vocabulary identification were both coded as 0. Other item types classified as nonphonetic discrimination were coded as 1.

Variable 2: Number of plausible distractors

This variable was coded as the number of plausible response alternatives, except for the key.

Variable 3: Necessity for inference

Sometimes, stimuli include implicit information that examinees require several inference steps for accessing the correct answer. Therefore, this variable was coded as 0 when no inference was necessary for accessing the correct answer and was coded as 1 when inference was necessary for accessing the correct answer.

Variable 4: Content familiarity

The assumption of topic familiarity is whether the examinees are familiar with the context that could affect item difficulty. Stimuli involved an infrequent word, unfamiliar dialog, topic, and location that increase item difficulty. Thus, for this variable, stimuli involving a frequent task type, frequent word, or familiar content were coded as 0, whereas stimuli involving an unfamiliar task type, infrequent word, or unfamiliar topic were coded as 1.

Variable 5: Number of propositions

The number of propositions in a sentence is determined by the number of relations in a sentence. Only verbs, adjectives, and adverbs can constrain relations. Therefore, this variable was coded as the total number of verbs, adjectives, and adverbs in the text.

Variable 6: Lexical overlap in the key

If a lexical overlap existed between the text and the answer key, then this variable was coded as 1, otherwise it was coded as 0.

Variable 7: Lexical overlap in the distractors

If a lexical overlap existed between the text and the distractors, then this variable was coded as 1, otherwise the variable was coded as 0.

Variable 8: Heterogeneity of sentence patterns in options

Options comprising the same patterns and convergent constructs and that possibly reduced the cognitive load were coded as 0. By contrast, options comprising inconsistent patterns and discriminating constructs and that possibly increased the cognitive load were coded as 1.

Variable 9: Redundancy of necessary information

This variable was coded as 0 if the necessary information was represented repeatedly in a discourse. By contrast, this variable was coded as 1 if the necessary information was represented only once in a discourse.

Coding example:

Examinees listened to a dialog and responded accordingly.

Girl: Look! What is Tony doing in the park?
Boy: He is riding a bike.
Girl: What?
Boy: He is riding a bike.
Question: What is Tony doing?
(1) He is in the park.
(2) He is reading.
(3) ***He is riding a bike***. *
(4) He is jogging.

Table 4. Coding illustration of example

Component	Coding	Rule
1.Phonetic/non-phonetic discrimination item type	1	non-phonetic discrimination
2. Plausible distractor	1	He is "in the park."
3. Need for inference	0	No steps were necessary for inference
4. Content familiarity	0	The topic of the dialog was frequently presented in the text book.
5. Number of proposition	5	Look, is doing, is riding, is riding, is doing
6. Lexical overlap is in key	1	The lexical overlap is in the key
7. Lexical overlap in the distractors	1	Lexical overlap was present in the distractors
8. Heterogeneity of sentence patterns in options	0	Options display the same pattern
9. Redundancy of necessary information	0	"He is riding a bike."

2.5. Data Analysis

Our data analysis was based on the framework of cognitive factors, so the consistency between coders is very important. Two experienced English teachers were invited for coding 150 items from the item pools of grades 5, 6, and 7. Each item was coded with nine cognitive factors and assigned a value for each of the factor variables. Raters had discussions after coding for integrating discrepancies and revising coding. We performed a correlation analysis to provide the Pearson correlation coefficients as the measures of inter-coder agreement.

Multiple regression analysis was conducted for calibrating the correlation coefficients between independent variables and item difficulty to enable us to select the independent variables significantly correlated with item difficulty. Subsequently, the selected variables were utilized in multiple regression analysis for estimating the explained variance that accounted for item difficulty.

3. Results

Table 5. Coder consistency correlations

	R_1 vs. R_2	R_1 vs. R_3	R_2 vs. R_3
Phonetic/ non- phonetic discrimination item type	1.00	1.00	1.00
Plausible distractors	.882	.871	.851
Necessity for inference	.920	.902	.922
Content familiarity	.891	.890	.914
Number of propositions	.874	.900	.883
Overlap in the key	1.00	1.00	1.00
Overlap in the distractors	1.00	1.00	1.00
Heterogeneity of sentence patterns in options	.901	.893	.913
Redundancy of necessary information	.952	.953	.960

Note. R_1 is the researcher. R_2 and R_3 are two professional in-service teachers.

3.1. Consistency of Coders

The correlation of rater consistency is shown in Table 5. The correlation coefficients ranged from .85 to 1. High correlation coefficient values showed a general consensus among raters about the cognitive factors.

3.2. Predicting Item Difficulty by Using Cognitive Factors

A multiple regression model was used for investigating the correlation between item difficulty and the cognitive factors, Table 6 shows that all the nine cognitive factors were significantly correlated with item difficulty. Stepwise regression was used to sift the most effective components for predicting item difficulty. The model summary of stepwise regression is shown in Table 7.

The results of stepwise regression show that five predictors, the necessity for inference, number of plausible distractors, phonetic/nonphonetic discrimination item type, heterogeneity of sentence patterns in options, and lexical overlap in the key, were the most effective predictors. The variance explained by these five predictors is 53.5%. Model 5 in Table 8 shows the significant predictors of the predicting model.

In conclusion, the model obtained an overall R^2 of .535. In other words, the explained variance of using the cognitive factors for predicting the item difficulty of the ASAP-ENGLCT was 53.5%.

Furthermore, we obtained a standardized regression equation from the coefficients as follows:

$$Y = -.915 + .375X_1 + .537X_2 + .692X_3 + \\ +.574X_4 - .286X_5$$

In this equation, Y represents item difficulty, −.915 is a constant, X_1 represents "inference," X_2 represents "number of distractors," X_3 represents "phonetic/nonphonetic discrimination item type," X_4 represents "heterogeneity of sentence patterns in options," and X_5 represents "lexical overlap in key." The regression line shows that "item type," "heterogeneity of sentence patterns in options," and "number of distractors" contribute the most variance to item difficulty. A one unit change in any of these three components causes more than a .5 unit increase in item difficulty.

3.2.1. Test for Model Fit

ANOVA analysis results in Table 9 shows that $F_{(5,144)} = 35.94$ ($p < .01$), indicating that the model fits the data well.

3.2.2. Test for Independence

Durbin–Watson statistics is a test statistics used for detecting the presence of autocorrelation in the residual series from a regression analysis. This test statistics can be used for testing for correlation between the residuals. The value of this statistic was between 0 and 4; the most satisfactory value was approximately 2. The Durbin–Watson value in Table 10 shows the goodness of fit of the model.

3.2.3. Test for Collinearity

Collinearity is used for indicating that a predictor is an exact linear combination of other predictors. Collinearity analysis involves the relationship among the independent variables (predictors) and does not directly involve the response variable (dependent variable). The collinearity indices are variance inflation factor (VIF) and tolerance.

$$VIF = \frac{1}{1 - R_i^2}, i = 1,2,3...,k$$

R_i^2 is the squared multiple correlation based on regression. X_i is the remaining k-1 predictors. For k predictors, k such models exist.

The larger the VIF value, the more troublesome is X_i. Any value of VIF_i greater than 10.0 should be a cause of concern.

$$Tolerance_i = \frac{1}{VIF_i}$$

The larger the value of tolerance, under 1.0, the more favorable it is.

The VIF value in Table 11 is small and the tolerance value is large and under 1.0, indicating no collinearity problem in this model.

In conclusion, the most satisfactory predictors for the ASAP-ENGLCT were the phonetic/nonphonetic discrimination item type, number of distractors, necessity for inference, lexical overlap in the key, and heterogeneity of sentence patterns in options. The overall explained variance was 53.5%, and the test for the goodness-of-fit showed the importance of the F-test and the correlation coefficient. The results for independence and collinearity also demonstrate the accuracy with a high explained variance of the predicting model.

Table 6. Significance of correlation coefficient

	Item_difficulty	Var1	Var2	Var3	Var4	Var5	Var6	Var7	Var8
Var1	.437**								
Var2	.443**	-.041							
Var3	.452**	.381**	.162*						
Var4	.344**	.168*	.116	.439**					
Var5	.280**	.513**	.179*	.321**	.224**				
Var6	-.406**	-.401**	-.044	-.424**	-.207*	-.139			
Var7	.334**	.574**	.171*	.132	-.028	.359**	.002		
Var8	.445**	.116	.274**	.178*	.265**	.139	-.226**	.175*	
Var9	.289**	.342**	.026	.349**	-.005	.099	-.180*	.324**	.089

var1: Phonetic/ non- phonetic discrimination item_type, var2: N. of distractors, var3: Inference, var4: Content familiarity
var5: N. of Proposition, var6: Overlap_in_key, var7: Overlap_in_distractor, var8: Heterogeneity of sentence patterns in options
var9: Redundancy of NI
*p < .05; **p < .01

Table 7. Model summary of stepwise regression

Model	R	R^2	Adjusted R^2	Std. Error of the Estimate
1	.452[a]	.204	.199	.819
2	.587[b]	.345	.336	.746
3	.674[c]	.454	.443	.683
4	.722[d]	.521	.508	.642
5	.732[e]	.535	.519	.634

a. Predictors: (Constant), Inference
b. Predictors: (Constant), Inference, N_distractors
c. Predictors: (Constant), Inference, N_distractors, phonetic/ non- phonetic discrimination item_type
d. Predictors: (Constant), Inference, N_distractors, phonetic/ non- phonetic discrimination item_type, Heterogeneity of sentence patterns in options
e. Predictors: (Constant), Inference, N_distractors, phonetic/ non- phonetic discrimination item_type, Heterogeneity of sentence patterns in options, lexical overlap_in_key

Table 8. Coefficients of stepwise regression

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-.470	.077		-6.106	.000
	Inference	.956	.155	.452	6.162	.000
2	(Constant)	-.732	.084		-8.692	.000
	Inference	.826	.143	.390	5.769	.000
	N_distractors	.580	.103	.380	5.612	.000
3	(Constant)	-1.181	.113		-10.431	.000
	Inference	.523	.143	.247	3.667	.000
	N_distractors	.638	.095	.418	6.699	.000
	Phonetic/ non- phonetic discrimination item_type	.713	.132	.360	5.415	.000
4	(Constant)	-1.203	.106		-11.295	.000
	Inference	.464	.135	.219	3.444	.001
	N_distractors	.529	.093	.347	5.712	.000
	Phonetic/ non- phonetic discrimination item_type	.665	.124	.337	5.362	.000
	Heterogeneity of sentence patterns in options	.622	.138	.273	4.503	.000
5	(Constant)	-.915	.174		-5.274	.000
	Inference	.375	.140	.177	2.687	.008
	N_distractors	.537	.092	.352	5.855	.000
	Phonetic/ non- phonetic discrimination item_type	.592	.128	.299	4.640	.000
	Heterogeneity of sentence patterns in options	.574	.138	.252	4.147	.000
	Lexical overlap_in_key	-.286	.137	-.138	-2.086	.039

Note. Dependent Variable: Item difficulty

Table 9. ANOVA table of multiple regression

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	66.745	5	13.349	33.163	.000a
	Residual	57.964	144	.403		
	Total	124.710	149			

a. Predictors: (Constant), options, item_type, N_distractors, Inference, lexical overlap_in_key
b. Dependent Variable: Item_difficulty

Table 10. Result of Durbin- Watson statistics

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	.732[a]	.535	.519	.63445	.535	33.163	5	144	.000	1.994

Note a. Predictors: (Constant), options, phonetic/ non- phonetic discrimination item_type, N_distractors, Inference, lexical overlap_in_key
b. Dependent Variable: Item_difficulty

Table 11. Results of collinearity statistics

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
(Constant)	-.915	.174		-5.274	.000		
Phonetic/ non-phonetic discrimination item_type	.592	.128	.299	4.640	.000	.775	1.290
N_distractors	.537	.092	.352	5.855	.000	.895	1.117
Inference	.375	.140	.177	2.687	.008	.741	1.350
Lexical overlap in key	-.286	.137	-.138	-2.086	.039	.732	1.366
Heterogeneity of sentence patterns in options	.574	.138	.252	4.147	.000	.876	1.142

Note. Dependent Variable: Item difficulty

The results demonstrated that five predictors, namely inference, the number of plausible distractors, the phonetic/nonphonetic discrimination item type, the heterogeneity of sentence patterns in options, and lexical overlap in the key, are the most effective predictors in the model. The variance explained by these five predictors was 53.5%. In conclusion, the model obtained an overall R^2 of .535. In other words, the explained variance of using the cognitive factors for predicting the item difficulty of the ASAP-ENGLCT was 53.5%. Compared to other studies, this value is relatively high. This high value emphasizes the importance of this predicting model.

4. Conclusions

The effect of using cognitive factors for predicting the item difficulty of EFL (English as foreign language) LC shows that the explained variance of the predicting framework is approximately 53%, which responds to research question 2. Moreover, in response to research question 1, nine cognitive factors proposed in this study significantly correlated with item difficulty. The coding of the nine cognitive factors also revealed a high consistency among raters. The five cognitive factors, namely the phonetic/nonphonetic discrimination item type, number of plausible distractors, necessity for inference, lexical overlap in the key, and heterogeneity of sentence patterns in options, that significantly contributed to item difficulty can be considered as effective predictors, and they also responded to research question 3. The results provided reference for test constructers when developing a large-scale EFL English LC test. Because the database was derived from the ASAP-ENGLCT, the effective predictors represented the characteristics of basic-level LC items and could be generalized to the item construction.

The main purpose of an ASAP assessment is to select students below grade 9 with both low academic achievement and low socioeconomic status. Remedial instruction is implemented after the target students are selected. The components significantly correlated with item difficulty can be used to develop teaching strategies.

4.1. Significant Components

This study proposed nine components in the predicting model, and the results showed that five components were significantly correlated with item difficulty. Among these five components, Rubin [16] and Brindley and Slatyer [17] obtained the item type, whereas Nissan, DeVincenzi, and Tang [18] obtained inference. The present study indicated that the necessity of making an inference for answering an item considerably affects the difficulty of dialog items in TOEFL LC. Both Kostin [19] and Hu [20] obtained lexical overlap in the key. These findings show that our results were consistent with those of the aforementioned studies.

4.2. Insignificant Components

Multiple regression results showed that content familiarity does not significantly affect the item difficulty of the ASAP-ENGLCT. The results demonstrate that the test is unaffected by construct-irrelevant factors, suggesting that examinees' performance on the test is primarily associated with the abilities being measured.

In this study, the number of propositions did not considerably affect item difficulty and this finding was consistent with that obtained by Buck [1]. According to the author, examinees process listening information with units of idea but not with individual words, phrases, or sentences. In other words, under a time limit, examinees must efficiently process listening information by using idea chunks and incoming information. Examinees cannot or do not have to recognize every single word. This result is not consistent with reading comprehension. Kintsch [21] suggested that the reading rates of expository texts are directly associated with the number of propositions in the text. The author's suggestion indicates the different characteristics of listening and reading. In addition, the number of propositions must be associated with the item type, that is, the more sentences involved in an item, the more propositions. However, this does not necessarily increase item difficulty. For example,

almost every dialog response item included only one sentence, whereas almost every dialog comprehension item included more than four sentences. The number of propositions increases with the number of sentences; however, item difficulty does not necessarily increase with the number of sentences. Thus, it is a contrasting fact that more sentences provide more adequate information for a correct answer to reduce item difficulty. Furthermore, when constructing items, test developers cannot consider the number of propositions for every item.

The presence of negative items in text has been affirmed to be positively associated with item difficulty [18, 22], but in the item pool of this research, fewer than 10 of the 150 items represented negative items in the text. Therefore, the influence of negative items may be trivial; hence, the components were omitted after the exploratory study. Negative items may be an effective predictor if the number of negative items were sufficient.

5. Implications

5.1. Implication of Test Construction

The study results suggest that five cognitive factors, item type, the number of plausible distractors, the necessity of making inference for answering an item, lexical overlap between words in the listening text and the words in the answer key, and the heterogeneity of sentence patterns in options, are appropriate predictors of item difficulty in a basic-level EFL LC test.

The other four cognitive factors, content familiarity, the number of propositions, lexical overlap between words in the listening text and the words in distractors, and the redundancy of necessary information, were all significantly correlated to item difficulty. In addition, the skipped factor, that is, the presence of negative items in the text, affected item difficulty. Test developers are recommended to consider these components for test construction.

5.2. Application for Instructions

As EFL language learners, students often have problems with LC, and the present study results could help LC instruction in class. Because numerical factors, such as the number of words, sentences, and propositions, are not considerably correlated to item difficulty, teachers should train students to pay attention to the main ideas in the listening text instead of to every single word. When examinees encounter an unknown word, they should not be baffled and skip the incoming information; they should attempt to find more information from the whole text, and even from the answer options.

If a student fails in letter sound or vocabulary identification, teachers should teach the student the basic knowledge of language, such as the alphabet and phonics, to reconstruct the student's basic language ability. Subsequently, the student can acquire additional LC skills.

Our results are also associated with the strategy for using LC. When students simultaneously employ top-down and bottom-up cognitive process, both processes have their advantages and disadvantages. The bottom-up process helps students to catch details in the text, whereas the top-down process helps students to make inferences for more complex information.

When students are allowed to preview the item stems or options, the students should be instructed methods to examine the stems and options to identify the main message. Stems may provide a hint for students to form predictions, whereas options preview may help students to identify overlaps between the text and the answer key or to ignore distractors.

Applying the study results to LC instruction can help students to effectively improve their listening performance.

6. Limitations and Suggestions

This study is based on a curriculum-based assessment. Therefore, the construction of the assessment must be limited to the content area of the subject matter; the extent of factors that can be manipulated should also be limited. For instance, the presence of infrequent words, unfamiliar topics, and idioms affects item difficulty [15-16]. However, these factors were not suitable in the present study. The reason may be that the examinees of the ASAP were students who were elementary school and junior high school students below grade 9. The Ministry of Education recommends that students below grade 9 should be able to use 1,200 basic words. In this study, the content and test were constructed according to the aforementioned suggestion. The manipulation of words and topics is less frequent in curriculum-based assessment than in other types of language assessment.

In addition, phonological factors, such as speech rate and accent, and text factors, such as discourse utterance and interrogation and answer format, were not investigated. Our results suggested that options with a typical pattern in a dialog item increase item difficulty. Because sentence patterns in textbooks always provide a typical method for answering an interrogative sentence, for example, we usually answer questions beginning with auxiliary verbs, such as "do," "does," "can," and "may," with a yes or no reply; if the options do not provide this type of response alternative, examinees may need more time to process information, thus increasing item difficulty. Because of the insufficiency of items with negation, the factor is not able to be discussed in the research. All the aforementioned components are suggested to be investigated in other language assessments.

Related future studies should construct a valid, scientific assessment tool from the cognitive perspective. The tool should also measure latent traits. The present methodology could be applied to diverse assessment tools and provide the same benefits for organizing the procedure of assessment

construction and increase test validity, which is essential in assessment.

REFERENCES

[1] Buck, G. (2006). Assessing listening. New York: Cambridge University Press.

[2] Alderson, J. C., & Bachman, L. F. (2001). Series editors' preface. In G. Buck, Assessing listening (p. x). Cambridge: Cambridge University Press.

[3] McDonough, S.H. (1995). Strategies and skill in learning a foreign language. London: Edward Arnald.

[4] Paris, S. G. & Brooks, P. (1977). Cognitive factors in children's listening and reading comprehension: Assessment and facilitation. (Final report). Retrieved from http://www.eric.org.

[5] Hsiao, C. W., & Hung, P. H. (2008). The relationship between Cognitive loading and Item difficulty for the numerical operation items. Poster session presented in The Joint Meeting of the 32nd Conference of the IGPME (PME 32) and the XXXth Annual Meeting of the North American Chapter of the IGPME (PME-NA XXX). Mexico, Morelia.

[6] Hung, P. H., Hsiao, C. W., & Lin, S.W. (2009). The implication of cognitive component analysis of PISA mathematics literacy to remedy instruction. Curriculum and instruction quarterly, 13, (1), 47-66.

[7] Hung, P. H., Lin, S. W., & Lin, C. J. (2006). The framework of cognitive complexity analysis for the 6th graders online TASA-MAT. Journal of educational research and development, 4, 69-86.

[8] Robinson, P. (2001).Task complexity, Task difficulty, and task production: Exploring interactions in a componential framework. Applied Linguistic, 22, (1), 27-57.

[9] Brown, A. (1995). The effect of rater variables in the development of an occupation-specific language performance test. Language Testing, 12(1), 1-15.

[10] Weir, C. J. (2005). Language testing and validation: An evidence-based approach. England: Palgrave Macmillan

[11] Embretson, S. E. (1994). Applications of cognitive design systems to test development. In C. R. Reynolds (Ed.), Cognitive assessment: A multidisciplinary perspective (pp. 107-136). NY: Plnum Press.

[12] Taylor, L. & Geranpayeh, A. (2011). Assessing listening for academic purposes: Defining and operationalising the test construct. Journal of English for Academic Purposes,10,(2), 89-101.

[13] Embretson, S. E. (1996). Cognitive design principles and the successful performer: A study on spatial ability. Journal of education measurement, 33, (1), 29-39

[14] Mislevy, R. J. (1993). Test theory for a new generation of tests. Hillsdale, NJ: Erlbaum.

[15] Hung, P. H., Shyu, C. Y, Lin, S.W, Lin, P.H, Hung, Y.T., & Wang, Y. W. (2010). 2010-2012 After School Alternative Program Report 1.Tainan: Technology- Based Educational Assessment Center in Tainan, NUTN. In Chinese.

[16] Rubin, J. (1994). A review of second language listening comprehension research. Modern language journal,78,(2),199-220.

[17] Brindley, G., & Slatyer, H. (2002). Exploring task difficulty in ESL listening assessment. Language Testing, 19, 369-394.

[18] Nissan, S., DeVincenzi, F., & Tang, K. L. (1996). An analysis of factors affecting the difficulty of dialogue items in TOEFL listening comprehension. (TOEFL Research Report RR-51). Retrieved from NJ: Educational Testing Service. Website: http://www.ets.org/toefl/research/archives/research_report.

[19] Kostin, I. (2004). Exploring item characteristics that are related to the difficulty of TOEFL dialogue items (TOEFL Research Report RR-04-11). Retrieved from NJ: Educational Testing Service. Website: http://www.ets.org/toefl/research/archives/research_report

[20] Hu, Y. H. (2006). An Investigation into the task features affecting EFL Listening comprehension test performance. The Asian EFL Journal, 8, (2), 33-54

[21] Kintsch, W. (1998) Comprehension: A paradigm for cognition. New York: Cambridge University Press.

[22] Freedle, R., & Kostin, I. **(**1996): The prediction of TOEFL listening comprehension item difficulty for minitalk passages: implications for construct validity. (TOEFL Research Report 56). Retrieved from NJ: Educational Testing Service. Website: http://www.ets.org/toefl/research/archives/research_report.