財團法人語言訓練測驗中心
外語教學、測驗研究補助專案

結案報告

申請人	朱秀瑜副教授
工作學校及系所	明志科技大學通識中心英文組
研究計畫名稱	Examining the Conformity of the GEPT Test Takers' Output to the Noun Phrase Accessibility Hierarchy (NPAH)
計畫執行期間	2013年10月—2014年9月

摘要

語言測驗(LT)與第二語言習得(SLA)的研究一直都是相得益彰的。因此，本研究擬就全民英檢(GEPT)的資料來檢視第二語言習得研究中的一個假說－the Noun Phrase Accessibility Hierarchy (NPAH)。NPAH是一個經過觀察多種語言下所產生的推論，即在一種語言中，語言學習者習得該語言各類型關係子句是有先後順序(或難易程度)的。這樣的順序如下：關係代名詞為主詞(SU) >關係代名詞為直接受詞(DO) >關係代名詞為間接受詞(IO) >關係代名詞為介係詞之受詞(OBL) >關係代名詞為所有格(GEN) >關係代名詞為比較句型的受詞(OCOMP)。而語言訓練測驗中心(LTTC)提供給本研究的全民英檢測驗資料，正好符合檢視NPAH的條件。因為研究顯示，NPAH的推論只有在學習者的母語及外語的關係子句結構是不同時，檢驗的結果才算是顯著。而全民英檢正符合上述要求，因為絕大部分全民英檢考生的母語都是中文，而中文及英文在關係子句結構正好呈現所謂的「鏡像」(mirror images)，也就是完全的反置。本研究假設：如果全民英檢不同級數的考生在口說及寫作表現上，呈現關係子句結構的進階式成長，那麼我們就可以宣稱：從第二語言習得的觀點來看，全民英檢是能夠將語言學習者做一個適當的劃分。

本研究一開始先將170位全民英檢考生(共四級，包括初級50位、中級50位、中高級40位及高級30位)的口說及寫作答題資料進行聽繕、打字、關係子句擷取及分類，然後再做量化及質性的分析。量化研究結果顯示，全民英檢考生的語言輸出並未完全符合NPAH的預測，因為SU及DO出現的模式不像NPAH所示，但OBL的確是出現在比較高階考生的語料中，這點倒是與NPAH相符。而質性分析的結果顯示，全民英檢考生的關係子句結構和考生整體的語言程度，在不同級數上有不同的權重關係。也就是說，不論是在使用關係子句的次數，抑或關係子句的正確度，以及關係子句的類型，不同級數的考生在關係子句結構的習得的確呈現進步的狀況，即便這樣的進步模式並未完全符合NPAH的預測。然而，這種差異可能是來自考生母語的影響或是其他語言學上面的原因，倒不是因為全民英檢考試本身的關係。

總括來說，全民英檢是一個相當可靠的測量工具，其分級結果充分反映了考生在關係子句結構上不同的表現：通過初級者，其產出的關係子句極其有限；通過中級者，在關係子句的產出上可能還相當的不穩，但至少能夠在沒有使用關係子句的情況下將語意傳達；通過中高級者，大致都能夠產出關係子句，只是結構和內容上可能還有少許不足；至於通過高級者，在產出關係子句上幾已無結構上的問題，其語言的內容品質相對上更形重要。

Abstract

Language Testing (LT) and Second Language Acquisition (SLA) have always benefited from each other in research. In that sense, the present study proposed to test the Noun Phrase Accessibility Hierarchy (NPAH) on the General English Proficiency Test (GEPT). The NPAH is a generalization found among human languages and discussed in many SLA studies, which predicts the ease of relativization as a function of the grammatical role of the head noun phrase (NP) modified by the relative clause (RC): subject (SU) > direct object (DO) > indirect object (IO) > oblique (OBL) > genitive (GEN) > object of comparison (OCOMP). The GEPT test data, sponsored by the Language Training and Testing Center (LTTC) for teachers and researchers in Taiwan, are perfect for testing such a theoretical assumption on the ground that, theoretically, it is believed the test of the NPAH would be salient only if the learners' native language and target language are different with regard to the RC construction. The GEPT meets those requirements in that its test data come mostly from native Chinese speakers learning English as a second language and, furthermore, Chinese and English happen to have so-called "mirror images" in terms of the RC construction. This study assumes if clear developmental sequences can be found among GEPT test-takers' RC production across levels as predicted by the NPAH, it could be claimed that the GEPT test tends to differentiate Chinese EFL learners based on the NPAH. In this present study, the speaking and writing test responses of a total of 170 GEPT test-takers across four levels (50 elementary, 50 intermediate, 40 high-intermediate and 30 advanced) were first transcribed and tagged and then analyzed quantitatively and qualitatively. The quantitative analysis indicates that the GEPT test-takers' language output did not entirely follow the NPAH predictions in that SU and DO relatives were not used in a way as predicted, but the use of OBL relatives seemed to have followed the predictions to appear at the later stages of development in English. On the other hand, the qualitative analysis suggests that the GEPT test-takers' RC production at different levels seems to weigh differently in relation to their language proficiency. In other words, the GEPT test-takers at the four levels did show progress in their RC production from level to level in terms of their RC attempts, accuracy and types, even though this progress did not entirely follow the developmental sequence as predicted by the NPAH. However, such a mismatch should probably be attributed to the test-takers' L1 influence or other linguistic features rather than the test itself. Generally speaking, the GEPT test is believed to work as a reliable assessment tool for anchoring test-takers' RC development over its four levels in the following ways: It is expected to see test-takers who passed only the elementary level produced very limited RCs. Those who passed the intermediate level might still be unstable (or unable) in producing RCs but they could still manage to get their meaning across without RCs. As for the high-intermediate level, test-takers who passed this level are believed to be able to produce RCs but not without some slight inadequacies. Finally, if test-takers managed to pass the advanced level, then supposedly they should have no problem producing RCs and their language proficiency is higher than simply knowing how to make correct RCs.

Keywords: NPAH, GEPT, SLA, relative clause, output

Introduction

Although Language Testing (LT) and Second Language Acquisition (SLA) focus on different aspects of applied linguistics, both fields have always benefited from each other in research. For example, SLA studies often use standardized language tests to determine second language learners' developmental stages; while many LT studies are based on SLA theories. In that sense, it is expected to see more research relating SLA proposals to standardized language tests or vice versa.

Based on one common claim in the SLA theories that more marked forms would be the last to be acquired or one could expect fewer errors in the less marked forms, Keenan and Comrie (1977) proposed the Noun Phrase Accessibility Hierarchy (NPAH), a generalization predicting the ease of relativization in the SLA research. The basic concept of the NPAH is that second language learners' production of relative clauses (RC) could be predicted according to the NPAH; that is, the following hierarchy reflects the ease of RC formation so as to constitute a developmental sequence on the part of learners:

SU > DO > IO > OBL > GEN > OCOMP[1]

Furthermore, based on the study of typological universals, the test of the NPAH would be salient only if the learners' native language and target language are different with regard to the specific universal in question; that is, relative clause formation. Otherwise, it could be claimed that the universal in question was only a matter of language transfer. In that regard, Chinese and English seem to be perfect candidates for testing the NPAH, since the two languages happen to have so-called "mirror images;" that is, structures that are reversed from one language to another.

Fortunately, since the year of 2013, the Language Training and Testing Center (LTTC) has started to sponsor researchers and teachers in Taiwan for conducting research on its locally-developed tests, such the General English Proficiency Test (GEPT)[2], College Student English Proficiency Test (CSEPT), etc. This present study was honored to get the sponsorship for the year of 2013-2014, including a research grant and access to the GEPT test data for research. The most important of all, the GEPT test data come mostly from native Chinese speakers learning English as a second language, which makes the NPAH good for testing the predicted acquisition order of Chinese speakers learning English.

Accordingly, this study analyzed the GEPT test-takers' responses in their speaking and writing tests across different levels. It is assumed that if clear developmental sequences can be found among different level test-takers' RC production as predicted by the NPAH, it could be claimed that the

[1] SU: Subject relative clause
DO: Direct object relative clause
IO: Indirect object relative clause
OBL: Object of preposition relative clause
GEN: Genitive relative clause
OCOMP: Object of comparative relative clause

[2] Developed by the LTTC since the year of 2000, the GEPT test is a criterion-referenced test on four language skills and administered at five levels (elementary, intermediate, high-intermediate, advanced and superior).

GEPT test tends to differentiate Chinese EFL learners based on the NPAH. One research question has thus been formulated for the present study:

Does the GEPT test-takers' RC production across levels conform to the predictions of the NPAH?

Literature Review

Ever since Keenan and Comrie (1977) proposed the Noun Phrase Accessibility Hierarchy (NPAH), many SLA studies have been conducted based on the proposal. Most of the early NPAH studies focused on the European languages and have reached quite convergent conclusions in supporting the NPAH as a linguistic universal (Croteau, 1995; Doughty, 1991; Dasinger & Toupin, 1994). However, later NPAH studies, especially those on non-European languages, have found disagreements and concluded language typological diversities should also be taken into consideration, such as studies on Japanese (Shirai & Kurono, 1998; Ozeki & Shirai, 2005), Korean (Jeon & Kim, 2007) and Cantonese (Matthews & Yip, 2002).

Among the NPAH studies on typologically diverse languages, researchers have attributed the divergence of these languages from the NPAH predictions to a number of different factors. Yip and Matthews (2007) concluded from their research on Cantonese-English bilingual children that these children's object relatives emerged before subject relatives in English, which is contrary to the NPAH predictions. They appealed to transfer effects (from Cantonese prenominal relatives to English postnominal relatives) to explain the divergence as they saw the effects interacting with the NPAH. Diessel (2004), viewing RCs constituting a network of interrelated constructions that children pick up bit by bit, proposed word order might be a factor that influences the acquisition sequence of RCs. Simply put, languages that have different word order are supposed to demonstrate different RC acquisition patterns, such as English (SVO) vs. Korean (SOV) and English (with postnominal RCs) vs. Chinese (with prenominal RCs). Ozeki and Shirai (2007) in a study about Japanese RCs found that semantic features of a language, such as animacy of head nouns, could also affect L2 learners' development of the RC structure in Japanese.

What is worth noticing is that, after 30 years, Comrie (2007) reinterpreted his 1977 original proposal of the NPAH as he would rather take the original NPAH as a reflection of more fundamental "psycholinguistic principles" other than "clear-cut differentiation in grammaticality judgment." He also argued that "any linguistic principle must be seen in its interaction with other linguistic principles…there is no claim that the grammatical positions (which are for the most part also grammatical relations) are the primitives that drive accessibility (p.304)." His reinterpretation of the NPAH seems to suggest that for the NPAH to work as a linguistic universal, more cross-linguistic evidence and generalizations are still needed.

Methodology

To answer the research question formulated in the introduction section, this present study was conducted on the GEPT test data provided by the LTTC. First in this section, the data inclusion criteria, data transformation procedures and all sorts of tagging criteria and approaches will be made clear for subsequent analysis.

Data inclusion criteria

The GEPT speaking and writing tests at lower levels include test questions that elicit responses not exactly reflecting test-takers' production, such as reading aloud, paraphrasing, etc., so not all the test response data in the speaking and writing tests fit into the present research framework. In other words, only responses fully involved with test-takers' production were used in the present study. Table 1 summarizes the parts of the speaking and writing tests that were included for analysis in the present study at each of the four levels.

The intended data that the LTTC provided for this present study came from a total of 170 GEPT test-takers' speaking test responses (audio files) and writing test responses (scanned files), along with their listening, reading, speaking and writing sub-scores on the GEPT test. The 170 test-takers belonged to four different level groups: 50 at the elementary, 50 at the intermediate, 40 at the high-intermediate and 30 at the advanced level. All of them passed both stages of the GEPT test at that level and were differentiated accordingly in terms of their English proficiency.

For the GEPT elementary, intermediate and high-intermediate level, the perfect score for the listening or reading test is 120; while a combined score of 160 on the listening and reading tests (with neither sub-score lower than 72) is required for passing the first stage of the level. The perfect score and passing score for the speaking or writing test are 100 and 80 respectively (except that the elementary writing test's passing score is 70). For the GEPT advanced level, the perfect score for the listening or reading test is 120; while a combined score of 150 on the listening and reading tests (with neither sub-score lower than 64) is required for passing the first stage of the level. A point of 3.0 on a 5-point scale indicates a passing score for either the speaking or writing test. Table 1 also shows the scoring details for the speaking and writing tests at each level.

Table 1. The parts of the speaking and writing tests included in the present study at each of the four levels and the scoring details for each part

Levels	Speaking Test	Scoring	Data used	Writing Test	Scoring	Data used
Elementary	**Part I** **Repeating**	Holistic* (0-5 points, converted to the 100-point scale later)	No	**Part I** **Sentence Writing** -Paraphrasing -Sentence combination -Rearrangement	50%	No
	Part II **Reading Aloud**		No			
	Part III **Answering Questions**		**Yes**	**Part II** **Paragraph Writing**	50%	**Yes**

Inter-mediate	**Part I** **Reading Aloud**	Holistic (0-5 points, nverted to the 100-point scale later)	No	**Part I** **Chinese-English Translation**	40%	No
	Part II **Answering Questions**		**Yes**			
	Part III **Picture Description**		**Yes**	**Part II** **Guided Writing**	60%	**Yes**
High-intermediate	**Part I** **Answering Questions**	Holistic (0-5 points, converted to the 100-point scale later)	**Yes**	**Part I** **Chinese-English Translation**	40%	No
	Part II **Picture Description**		**Yes**	**Part II** **Guided Writing**	60%	**Yes**
	Part III **Discussion**		**Yes**			
Advanced	**Part I** **Warm-up Interviews**	Holistic (1-5 points)	**Yes**	**Task I**	Holistic (1-5 points)	**Yes**
	Part II **Information Exchange**		**Yes**	**Task II**		**Yes**
	Part III **Discussion**		**Yes**			

****For the elementary level speaking test, the total score consists of two sub-scores: (1) pronunciation, intonation and fluency and (2) grammatical accuracy and vocabulary, with the former based on the performance in Part I, II and III and the latter on Part III only. Each sub-score accounts for 50% of the total score.***

Data transformation

To prepare the above-mentioned data for further analysis, the researcher alone first transcribed the 170 test-takers' recorded speech samples and then retyped their writing samples to appear in Word format. During the long and tedious process of transcribing and retyping, the researcher roughly tagged the data at the same time by following the minimum criteria for determining a qualifying RC (as shown in Table 2). The purpose of rough tagging at this stage was to include not only qualifying RCs without or with minor grammatical errors, but also attempted but disqualifying RCs for final tuning of the data to be analyzed.

Table 2. Minimum criteria for determining a qualifying relative clause

Basic requirements		**Examples**
1.	Presence of both a subject and a verb in a relative clause	(O) *We are designed to digest any food sources* ***that are*** *available to us.* (X) *There are lots of people want to get some cheap things* (X) *I would buy a cell phone which* ***easier using.***
2.	Presence of an antecedent prior to a relative clause	(O) *I can do* ***something*** *I want.* (X) *We might feed the animals* ***that the food*** *is unhealthy to them.*
3.	Presence of an accurate relative pronoun or adverb in a relative clause unless omitted (as in DO / IO / OBL / SC / OTH types)	(O)*There is one thing* ***(that)*** *they agreed on.* (O) *The reason* ***(why)*** *I am writing this letter is because I felt this topic should draw more attention from the public.* (X) *There are four boys* ***which*** *is playing soccer.*
4.	Presence of a preposition in a relative clause if required (as in IO / OBL types)	(O) *These are the problems we have to deal* ***with****.* (O) *It's a good place that parents should take their children* ***to*** *on holidays.* (X) *There are many cell phones nowadays I can use the internet* ___.
5.	No redundant occurrence of an object in a DO / IO / OBL relative clause other than the relative object pronoun	(O) *It is something I completely dislike.* (X) *She decided to buy the broken vase which she dropped* ***it*** *on the floor.*

Formal tagging of the data

At the next stage, the transformed and roughly tagged data were ready for formal tagging. It was now important to define an 'accurate' RC and an 'inaccurate' one among those roughly tagged RC attempts, as well as to define different RC types.

(1) Determining an 'accurate' RC: Basically, this present study did not apply a very strict standard in determining an accurate RC, since the primary goal of the study was to examine whether the GEPT test-takers across levels demonstrate developmental progress in RC formation. In that regard, RCs were classified as 'accurate' as long as they met the minimum criteria for forming a qualifying RC, as listed in Table 2, even though they might not be 100% grammatically correct. Among the pre-tagged qualifying RCs with minor grammatical errors, the tolerated grammatical error types were generalized and summarized in Table 3. Thus, qualifying RCs without or with the tolerated grammatical errors were classified as 'accurate'. On the other hand, attempted but disqualifying RCs were classified as 'inaccurate' RCs.

Table 3. Tolerated grammatical error types in an 'accurate' relative clause

Tolerated grammatical error types		Examples
1.	Inappropriate usage of the tense, aspect or voice of the verb in the relative clause	*-Last time I* ***get*** *together with my relatives* ***is*** *three weeks ago.* *-There are some foods that* ***are feed*** *for animals to eat.*
2.	Wrong usage of the singular or plural form of nouns in the relative clause	*-I'll buy a cell phone that has not so* ***much button****.* *-The food we feed the* ***sheeps*** *are sold by the farm.* *- Parents should put more attention on whom your* ***childrens*** *are making friends with.*
3.	Improper word usage in the relative clause	*-Are there a lot of guests you have to* ***service****?* *-I should like one that is with a big screen and not too* ***weight****.*

(2) Defining RC types: As mentioned in the earlier section, Keenan and Comrie's NPAH consists of five types of relative clause based on the position of the antecedent of the relative clause to be relativized, namely, SU, DO, IO, OBL, GEN and OCOMP. However, the present data show that there was also quite heavy use of relative clauses not belonging to any of the five types in the GEPT test-takers' language output, such as the use of *where*, *when* and *why* in the relative clause (relative adverbials), *which* referring to the antecedent as a whole proposition, the antecedent relativized as the subject complement and so forth. In order to put these non-basic relatives into analysis as well, the formal tagging was executed using type labels of the relatives on the NPAH (excluding OCOMP for no such occurrences in the present data) and the extra categories that appeared in the present data, as illustrated in Table 4.

Table 4. RC types tagged for the present study

NPAH RC types	**Definitions**	**Examples**
SU	Subject to be relativized	-The man **who / that** lives next door is very friendly. -The machine **which / that** broke down has now been repaired.
DO	Direct object to be relativized	-The woman **(who / that)** I wanted to see was away. -Have you found the keys **(which / that)** you have lost?
IO	Indirect object to be relativized	-He knows the girl **(who / that)** I wrote a letter to.
OBL	Relative pronoun as the object of a preposition	-This woman **(who / that)** he fell in love with left him. -Are these the keys **(which / that)** you were looking for? *The hotel **(which / that)** we stayed at wasn't very clean.
GEN	The antecedent having a possessive role in the relative clause	-This school is only for children **whose** first language is not English.
Non-basic RC types	**Definitions**	**Examples**
SC	Subject complement to be relativized	-They help us to become the person **(who / that)** we want to be.
OTH (others)	The antecedent as a whole proposition	-Tom passed his driving test, **which** surprised everybody.
	Adverbial clauses functioning as relative clauses	*The hotel **where** we stayed wasn't very clean.
		-The last time **(when / that)** I saw her, she looked fine.
		-The reason **(why / that)** I'm calling you is to invite you to a party.
		-We hope the government can change the way **(how / that)** elementary schools educate their students.

These two sentences, though having the exact same meaning, were tagged under two different labels.

To implement tagging, the emergence of test-takers' accurate and inaccurate use of each type of relative clause in their oral and written production were judged and marked in different colors manually. Manual tagging was used in this project because some of the omissions of relative pronouns as direct or indirect objects have caused difficulties for automatic retrieval of such types of relative clauses from the texts. In addition, some of the attempted but inaccurate RCs could not be easily parsed without human judgment either.

After the formal tagging was executed manually, the data were then partly double-checked using Microsoft Word's 'Find' function. That is, partial automatic checkup of the manually tagged texts was done by searching key words such as *which, who, that, whom, whose, where* and *why* with Microsoft Word's 'Find' function for any missed or incorrect tagging in the manual process. At last, final proofreading was performed to make sure (1) the RCs were categorized into the types as defined and (2) the 'accurate' and 'inaccurate' RCs in each type were distinguished as defined. The formal tagging, partial automatic checkup and final proofreading were all completed by the researcher alone as well.

Results and Discussion

In this section, the results of the data analysis will be presented and discussed to examine whether the GEPT test-takers' language output conforms to the predictions of the NPAH. More discussions will follow to finalize a few minor details.

Data summary

As shown in Table 5, the transcribed data of the intended speaking test responses contain a total of 4935, 22115, 26940 and 30501 tokens for the elementary, intermediate, high-intermediate and advanced levels, leading to an average of 98.7, 442.3, 673.5 and 1016.7 tokens per person for each of the four levels respectively. For the writing test responses, the total numbers of tokens for each level are 3417, 9068, 9407 and 27923 and the average tokens per person for each level are 68.3, 181.4, 235.18 and 930.8 respectively.

After the completion of the tagging process, a total of 4, 106, 148 and 251 attempted RCs (including accurate and inaccurate ones) were extracted from the speaking test data for each of the four levels (from elementary to advanced) and the numbers turn to 20, 43, 88 and 402 from the writing test data, as shown in Table 6. Obviously, there seems to be a surprisingly big gap between the numbers of attempted spoken RCs and written RCs at the intermediate and high-intermediate levels (106 vs. 43 and 148 vs. 88). However, if the total numbers of tokens in both tests were taken into consideration, the situation would not seem so bizarre. As shown in Table 5, the total number of tokens produced in the speaking test is far greater than the number in the writing test at either the intermediate or high-intermediate level (22115 vs. 9068 and 26940 vs. 9407). Consequently, if the RC numbers are compared per 100 tokens, the rates of RC production are actually quite close between the two tests (0.48 vs. 0.47 and 0.55 vs. 0.94) at both levels, as shown in Table 6.

The less surprising gaps between the numbers of attempted spoken RCs and written RCs for the elementary and advanced levels could also be resolved by looking at the numbers of attempted RCs per 100 tokens (0.08 vs. 0.59 and 0.82 vs. 1.44). In that sense, the GEPT test-takers tended to be able to produce even more RCs in writing than in speaking.

Table 5. Numbers of tokens in the speaking and writing test data for each level group

Test Level		Speaking Test		Writing Test	
		Total tokens	Average tokens per person	Total tokens	Average tokens per person
Elementary	N=50	4935	98.7	3417	68.3
Intermediate	N=50	22115	442.3	9068	181.4
High-intermediate	N=40	26940	673.5	9407	235.18
Advanced	N=30	30501	1016.7	27923	930.8

Table 6. Numbers of RCs in the speaking and writing test data for each level group

Level \ Test		Speaking Test					Writing Test				
		ACC RC	INA RC	ATT RC	ATT RC per100 tokens	Accuracy Rate	ACC RC	INA RC	ATT RC	ATT RC per100 tokens	Accuracy Rate
Elementary	N =50	3	1	4	0.08	75%	17	3	20	0.59	85%
Intermediate	N =50	88	18	106	0.48	83%	40	3	43	0.47	93%
High-intermediate	N =40	122	26	148	0.55	82%	86	2	88	0.94	98%
Advanced	N =30	235	16	251	0.82	94%	390	12	402	1.44	97%

**ACC RC=Accurate RC; INA RC=Inaccurate RC; ATT RC=Attempted RC*

In terms of the accuracy rate, shown in Table 6 as well, the spoken RC accuracy rates are 75%, 83%, 82% and 94% and written RC accuracy rates are 85%, 93%, 98% and 97% at each level respectively. The RC accuracy rates generally increase as the test-takers' language proficiency gets higher. The written RC accuracy rate tends to be higher than spoken rate at each level. More details about the RC extraction are presented in Table 6.

The conformity of the RC output to the NPAH

To examine the conformity of the GEPT test-takers' RC output across levels to the NPAH, all the extracted RCs (including accurate and inaccurate ones) were categorized into the seven RC types given in Table 4. For each level group and for their speaking and writing test data respectively, the percentages of the accurate RCs by types were obtained by having the number of accurate RCs in each type divided by the total number of accurate RCs. The percentages of the attempted RCs (accurate plus inaccurate RCs) were calculated in the same manner as accurate RCs. Tables 7 and 8 summarize the calculations about the accurate and attempted RCs in the speaking test data. Tables 9 and 10 summarize the writing test data.

Table 7. Percentages of accurate RCs by types in the speaking test data

Level \ RC type	SU	DO	IO	OBL	GEN	SC	OTH
Elementary	--	100%	--	--	--	--	--
Intermediate	27%	30%	--	1%	--	--	42%
High-intermediate	50%	39%	1%	1%	--	1%	9%
Advanced	41%	37%	--	10%	--	--	12%

Table 8. Percentages of attempted RCs by types in the speaking test data

RC type Level	SU	DO	IO	OBL	GEN	SC	OTH
Elementary	--	100%	--	--	--	--	--
Intermediate	31%	25%	--	1%	--	--	42%
High-intermediate	49%	37%	1%	3%	1%	1%	8%
Advanced	41%	35%	--	11%	--	--	12%

Table 9. Percentages of accurate RCs by types in the writing test data

RC type Level	SU	DO	IO	OBL	GEN	SC	OTH
Elementary	47%	47%	--	6%	--	--	--
Intermediate	20%	15%	--	5%	--	--	60%
High-intermediate	47%	21%	--	6%	--	--	27%
Advanced	55%	22%	1%	4%	1%	1%	16%

Table 10. Percentages of attempted RCs by types in the writing test data

RC type Level	SU	DO	IO	OBL	GEN	SC	OTH
Elementary	40%	55%	--	5%	--	--	--
Intermediate	21%	16%	--	5%	--	--	58%
High-intermediate	45%	20%	--	7%	--	--	27%
Advanced	54%	22%	1%	5%	1%	1%	16%

The percentages in Tables 7, 8, 9 and 10 reveal there is a tendency that the higher the test-takers' proficiency level was, the more frequently they used SU relatives, no matter in terms of accurate or attempted RCs in the speaking or writing test. In addition, the elementary test-takers used DO relatives exclusively in the speaking test and used DO relatives generally more frequently than other types in the writing test. All of the above-mentioned evidence indicates that the GEPT test-takers' RC output goes against the NPAH predictions about SU and DO, because, according to the NPAH, SU relatives are supposedly more likely to be used at the earlier stages of development in English.

However, Tables 7-10 also show that there is generally an increasing use of OBL relatives from elementary to advanced level, indicating OBL relatives were used more frequently at the later stages of development in English in this case, as predicted by the NPAH.

The use of IO, GEN and SC relatives appears to be very limited in the study (accounting for only 1% whenever it appears) so these three types will not be discussed here for their rare appearance in the study. On the other hand, it is worth noticing that the percentages of OTH relatives appear to be the highest at the intermediate level in all four tables (42%, 42%, 60% and 58%), but the percentages drop dramatically when it comes to the high-intermediate and advanced levels, since the use of SU and DO comes into very big play at these two levels along with a few other types. Therefore, it might be concluded that the total use of RCs did increase with the proficiency level and the increasing number of RCs at the higher levels belong mostly to the SU and DO types (as shown in Tables 7 to 10).

Another way to test the conformity of the GEPT test-takers' RC output to the NPAH is to check whether there were fewer errors in the less marked forms (in this case, SU and DO) than in the marked ones (in this case, OBL). As shown in Table 11, the RC accuracy rates by types were calculated by having the number of accurate RCs divided by the number of attempted RC for each type. It is found that both SU and DO have approximately the same accuracy rates and are higher than the OBL accuracy rate at each level, either in the speaking or writing test, which supports the claim that OBL did occur at the later stages of the RC development, but the relationship between SU and DO here is not clear. Meanwhile, the accuracy rates for each RC type generally increased from the lowest to the highest level, suggesting that fewer errors were made in each type as the test-takers' language proficiency got higher.

In summary, the GEPT test-takers' language output did not entirely follow the NPAH predictions in that SU and DO relatives were not used as predicted, but the use of OBL relatives seemed to have followed the predictions to appear at the later stages of development in English.

Table 11. RC accuracy rates by types in the speaking and writing tests over the four levels

RC type Level	**Speaking**			**Writing**		
	SU	**DO**	**OBL**	**SU**	**DO**	**OBL**
Elementary	0%	75%	0%	(100%)*	73%	(100%)*
Intermediate	73%	96%	(100%)*	89%	86%	(100%)*
High-intermediate	85%	85%	20%	100%	100%	83%
Advanced	93%	97%	85%	98%	97%	89%

The parentheses around 100% in this table indicate that there was only one such occurrence in that category, which happened to be accurate.

Correlation between the RC statistics and sub-test scores across levels

Some correlational analysis was also conducted to investigate the relationship between the GEPT test-takers' RC production and their sub-test scores on listening, reading, speaking and writing. However, as shown in Table 1, not all the parts in the speaking and writing tests for the elementary, intermediate and high-intermediate levels were included for extracting the RC attempts, but the speaking and writing test scores used in the correlational analysis reflect the test-taker's overall performance on all of the parts in the speaking and writing tests. As a result, it would be safe to say that the correlational analysis was conducted between the test-takers' RC production and their "total scores on each component" of the listening, reading, speaking and writing test respectively. The results are presented in Tables 12-15 for each of the four levels.

Table 12 shows that there is no correlation between the elementary test-takers' RC production and sub-test scores, which is obvious on the ground that the elementary test-takers produced very limited RCs in both the speaking and writing tests.

Table 12. Correlation between the RC statistics and sub-test scores (Elementary)

	S-Acc RC	S-Att RC	W-Acc RC	W-Att RC	L-Score	R-Score	S-Score	W-Score
S-Acc RC	1.000							
S-Att RC	.857**	1.000						
W-Acc RC	-.003	-.043	1.000					
W-Att RC	-.027	.047	.928**	1.000				
L-Score	.024	.092	.033	.096	1.000			
R-Score	-.066	-.013	.129	.193	.635**	1.000		
S-Score	-.116	.012	.134	.151	.300**	.263	1.000	
W-Score	-.122	-.180	.221	.214	.141	.425**	.272	1.000

*p<.05 **p<.01 (Acc RC: Accurate RC; Att RC: Attempted RC)

Table 13 shows that at the intermediate level, there is a correlation between the test-takers' RC attempts in the speaking test and their speaking test scores (r=.318), while their accurate and attempted RCs in the writing test are correlated with their reading test scores respectively (r=.328 and .293). Their spoken and written RCs are also correlated with each other, either in terms of accurate or attempted ones (r=.360 and .365). However, the correlations here are weak due to the low correlation coefficients between the variables.

Table 13. Correlation between the RC statistics and sub-test scores (Intermediate)

	S-Acc RC	S-Att RC	W-Acc RC	W-Att RC	L-Score	R-Score	S-Score	W-Score
S-Acc RC	1.000							
S-Att RC	.97**	1.000						
W-Acc RC	.360*	.404**	1.000					
W-Att RC	.318*	.365*	.976**	1.000				
L-Score	.109	.046	.096	.016	1.000			
R-Score	.012	.000	.328*	.293*	.413**	1.000		
S-Score	.270	.318*	.187	.166	.249	.167	1.000	
W-Score	.177	.211	.276	.233	.133	.431**	.151	1.000

*p<.05 **p<.01 (Acc RC: Accurate RC; Att RC: Attempted RC)

Table 14 shows that the high-intermediate test-takers' accurate RCs in the speaking test are correlated with their listening and speaking test scores respectively (r=.314 and .332). Also, their RC attempts in the speaking test are correlated with their listening scores (r=.346). Again, these correlations are not strong due to the low correlation coefficients between the variables.

Table 14. Correlation between the RC statistics and sub-test scores (High-intermediate)

	S-Acc RC	S-Att RC	W-Acc RC	W-Att RC	L-Score	R-Score	S-Score	W-Score
S-Acc RC	1.000							
S-Att RC	.924**	1.000						
W-Acc RC	.250	.182	1.000					
W-Att RC	.251	.179	.994**	1.000				
L-Score	.314*	.346*	-.059	-.033	1.000			
R-Score	.169	.226	.092	.101	.645**	1.000		
S-Score	.332*	.290	-.143	-.105	.438**	.249	1.000	
W-Score	.082	.125	-.085	-.091	.482**	.426**	.534**	1.000

*p<.05 **p<.01 (Acc RC: Accurate RC; Att RC: Attempted RC)

In Table 15 for the advanced test-takers, correlations can be found only between their speaking test scores and their accurate and attempted RCs in the speaking test respectively (r=.452 and .516). However, the correlations are stronger in this level group compared to the previous two levels.

Table 15. Correlation between the RC statistics and sub-test scores (Advanced)

	S-Acc RC	S-Att RC	W-Acc RC	W-Att RC	L-Score	R-Score	S-Score	W-Score
S-Acc RC	1.000							
S-Att RC	.981**	1.000						
W-Acc RC	.283	.344	1.000					
W-Att RC	.254	.317	.993**	1.000				
L-Score	.000	.067	.273	.270	1.000			
R-Score	.128	.150	.150	.152	.416**	1.000		
S-Score	.452*	.516*	.068	.044	.175	.093	1.000	
W-Score	.111	.100	.304	.288	.133	.303	.109	1.000

*p<.05 **p<.01 (*Acc RC: Accurate RC; Att RC: Attempted RC*)

Generally speaking, the correlations between the GEPT test-takers' RC production and their sub-test scores are relatively weak. A rough pattern that could be concluded here is that it seems the higher the group's proficiency level is, the more likely their spoken RC production is correlated with their speaking test scores. In other words, if the test-takers in the advanced level group use more RCs in their test responses, they are more likely to get higher scores on the speaking test. On the other hand, there is no such relationship between advanced test-takers' written RC production and their writing test scores.

Qualitative analysis

The quantitative analysis in the previous sections concludes that the GEPT test-takers' language output did not entirely follow the NPAH predictions and their RC production was weakly correlated with some of their GEPT sub-test scores. However, to catch the nuances that the grouped data might not be able to show in the quantitative analysis, a qualitative analysis was conducted on the contents of a few test-takers' speaking and writing test responses.

Due to its very limited RC production, the elementary level was excluded from the qualitative analysis. For the other three levels, one or two test-takers each with the highest and lowest test scores (lowest but still higher than the passing scores) for each level were chosen for further analysis. Moreover, test-takers whose speaking or writing test responses did not show a single RC were also taken into consideration for how they managed to pass the speaking or writing test, especially at higher levels. The background information of the test-takers under investigation is listed in Table 16.

Table 16. Background information of the test-takers for qualitative analysis

Scores Level	Code	Age	Sex	LS	RS	SS	WS	S- Acc RC	S- Att RC	W- Acc RC	W- Att RC
Intermediate	**A**	18	F	115	114	100	100	5	6	5	5
	B	17	F	80	81	80	80	1	1	1	1
	C	15	M	115	75	80	90	0	0	0	0
	D	14	M	117	87	80	80	0	0	0	0
	E	18	F	104	117	80	80	0	0	0	0
	F	17	F	99	90	80	80	0	0	0	0
High-intermediate	**G**	16	F	120	120	100	100	7	9	3	3
	H	17	F	120	115	100	100	2	2	0	0
	I	17	F	91	77	80	84	2	2	1	1
	J	20	M	77	88	80	84	3	3	7	7
	K	17	M	91	101	80	84	0	0	2	2
	L	19	F	109	104	80	84	0	0	6	7
	M	18	F	104	96	80	84	0	0	1	1
	N	18	M	88	101	80	84	0	0	1	1
	O	15	M	96	77	80	84	0	0	1	1
	P	15	M	109	93	100	88	6	6	0	0
	Q	37	F	109	104	80	84	5	5	0	0
	R	17	M	99	77	80	84	2	2	0	0
	S	16	F	96	80	80	84	1	2	0	0
Advanced	**T**	28	M	111	105	4.0	4	14	14	18	18
	U	15	M	114	93	4.0	4	15	17	35	38
	V	14	F	86	74	3.0	3	9	9	11	11
	W	39	F	80	83	4.0	3	13	13	5	5
	X	15	F	102	74	3.0	3	0	0	10	10

LS: listening sub-test score; RS: reading sub-test score; SS: speaking sub-test score; WS: writing sub-test score

(1) Analysis on the intermediate level: According to the quantitative analysis, the intermediate test-takers' spoken and written RC production are correlated with their corresponding speaking and writing sub-test scores. Therefore, it is expected to see test-taker A produced more spoken and written RCs than test-taker B, since the former got perfect scores on the speaking and writing tests (even the listening and reading test scores are among the highest); while the latter got only passing scores on the speaking and writing tests (and also the lowest combined scores for all the four tests among the 50 test-takers at this level).

Such a difference was easily observed by comparing the contents of the two test-takers' test responses. Test-taker A not only used more RCs than test-taker B, she also used more types of RCs than test-taker B. Compared to test-taker B, who used only one SU relative each in the speaking and writing tests, test-taker A used SU, DO and OTH in both tests (with one inaccurate RC).

On the other hand, a total of four test-takers (C, D, E and F) at the intermediate level did not use a single RC either in the speaking or writing test, but they still managed to pass the level. Some of these test-takers even got pretty good scores on the reading and listening tests, such as test-taker

E. She got 104 and 117 on the listening and reading tests respectively; but only passing scores on the speaking and writing tests. Even without the production of relative clauses, these four test-takers still demonstrated their ability to use compound sentences as well as subordinate clauses such as noun clauses and adverbial clauses. Although there is no evidence to show these four test-takers were unable to use relative clauses at this stage, it still makes sense to speculate that their receptive knowledge of RC (listening and reading) at this level does not necessarily turn into active use of RC (speaking and writing).

In general, the qualitative analysis seems to suggest that the GEPT test-takers' RC production at the intermediate level might not yet play a crucial part in reflecting their language proficiency.

(2) Analysis on the high-intermediate level: All the test-takers at the high-intermediate level demonstrated an ability to produce RCs on the ground that each one of them made at least one RC either in the speaking or writing test, which is a significant progress over their intermediate counterparts. Based on the quantitative results about the high-intermediate level, there is a correlation between the test-takers' RC production in the speaking test and their speaking test scores. The 13 test-takers' (G-S) score and RC information in Table 16 seems to match such analysis quite well in that those producing zero RC in the speaking test (K-O) did get only passing scores (80) on the speaking test.

However, the two test-takers holding the highest scores on the four tests show quite different patterns of RC use, which deserves special attention. Test-taker G, who got perfect scores on all the four tests, made a total of 12 RC attempts (including 2 inaccurate ones) in both the speaking and writing tests, with the RC types ranging from SU, DO, OBL to OTH. However, test-taker H, who also got nearly all perfect scores on the four tests (except a score of 115 on reading), demonstrated only two RCs (one SU and one DO) in the speaking test and even no RCs at all in the writing test. Since both test-takers' spoken and written production met the GEPT criteria for obtaining the same perfect scores, it is suggested that the RC production is not necessarily responsible for the quality of a test-taker's language output at the GEPT high-intermediate level. This assumption is also supported by looking at the test performance of two of the lowest score-holders at this level, I and J. Both of them scored slightly above the passing scores on the speaking and writing tests (also slightly above the passing scores on the listening and reading tests), but test-taker J, especially, used a total of 10 RCs in the speaking and writing tests, including SU, DO and OTH.

In a similar vein, a close look was given to the 5 test-takers with no spoken RC production (K-O) and the 4 more test-takers with no written RC production (P-S) out of the 40 high-intermediate test-takers. It seems that, just as mentioned earlier, test-takers K-O simply got passing scores on the speaking test, test-takers P-S showed a similar pattern on the writing test as well, suggesting that even though the RC production might not be a deciding factor for ensuring the quality of a test-taker's language output, the lack of such production is often associated with some inadequacies, as judged by the GEPT criteria for the high-intermediate level, in one's speaking or writing performance.

(3) Analysis on the advanced level: Like the high-intermediate test-takers, all the test-takers at the advanced level demonstrated an ability to produce RCs on both the speaking and writing tests, except test-taker X, who made zero RC on the speaking test (but 10 RCs on the writing test, including SU, DO and GEN). With a significantly larger amount of elicited tokens than the other three levels, it is no surprise that the numbers of RCs produced by the advanced test-takers are significantly higher than the other three levels either in the speaking or writing test. However, since it is clear that the formation of RCs never constitutes a problem for all the advanced learners, it is the quality, not the quantity of the RC production that deserves even more attention.

The quantitative analysis on the advanced learners' RC production and test scores shows that there is quite a strong correlation between the test-takers' spoken RC production and their speaking test scores. This relationship seems to apply quite well to the five test-takers (T-X) in Table 16 in that the three test-takers with higher speaking test scores (T, U and W) produced more spoken RCs than V and X, who got lower rankings on the speaking test. Both test-takers T and U are also among the highest score-holders in terms of the listening, reading and writing tests; while V and W got the lowest combined scores on listening and reading among the 30 test-takers as well as a passing score (3.0) on the writing test. However, no matter it is the highest test score-holders (such as T and U) or the lowest test score-holders (such as V and W) or a test-taker simply with no spoken RC (such as X) that are being taken into consideration, it is easily seen that each one of them produced at least two to three or even four types of RCs in either the speaking or writing test, not to mention the much higher accuracy and productivity of these RCs. In that sense, the association between the test-takers' RC production and their test scores at this level does not seem to have much meaning as compared to the lower levels, because the RC construction tends to become an underlying structure that a qualifying advanced learner should have acquired. It is likely that the quality of the spoken and written production of an advanced test-taker is determined by criteria other than the ability of making RCs.

Conclusion

This present study employed both quantitative and qualitative analysis on 170 GEPT elementary, intermediate, high-intermediate and advanced test-takers' RC production in parts of their speaking and writing test responses in order to test whether these test-takers' developmental progress of relativization follows the predictions of the NPAH from level to level. Some minor details will also be finalized here in this section with the major conclusion of the study.

The quantitative analysis concludes that the GEPT test-takers' language output did not entirely follow the NPAH predictions in that SU and DO relatives were not used in a way that is predicted, but the use of OBL relatives seemed to have followed the predictions to appear at the later stages of development in English. In addition, the correlational analysis shows that the GEPT test-takers' RC production and their sub-test scores are weakly correlated in that the higher the group's proficiency level is, the more likely their spoken RC production is correlated with their overall speaking test scores.

On the other hand, the qualitative analysis conducted on the intermediate, high-intermediate and advanced levels suggests that test-takers' RC production at different levels seems to weigh differently in relation to their language proficiency. At the intermediate level, the GEPT test-takers' RC production might not yet play a crucial part in reflecting their language proficiency. However, at the high-intermediate level, the lack of RC production is often associated with less adequate performance in either the speaking or writing test. Finally at the advanced level, the RC construction is considered to be an underlying structure that a qualifying advanced learner should have acquired so the quality of the spoken and written production at this level is probably determined by criteria other than the ability of making RCs.

In general, the GEPT test-takers did show progress in their RC production from level to level in terms of their RC attempts, accuracy and types, even though this progress did not entirely follow the developmental sequence as predicted by the NPAH. Such a mismatch, however, should probably be attributed to, as mentioned in the literature review section, the test-takers' L1 influence (in this case, mostly Mandarin Chinese) or other linguistic features (which is beyond the scope of this present study) rather than the test itself.

In addition, it is quite natural to see weak correlations between the GEPT test-takers' RC production and their speaking or writing test scores, since the GEPT scoring criteria for speaking and writing evaluates not only the test-taker's lexical and grammatical use, but also, more importantly, the content relevance and adequacy as well as organization (such as coherence and cohesion)[3].

As a result, it is concluded that the four levels of the GEPT test differentiates its test-takers in terms of RC production in the following ways: It is expected to see test-takers who passed only the elementary level produced very limited RCs. Those who passed the intermediate level might still be unstable (or unable) in producing RCs but they could still manage to get their meaning across without RCs. As for the high-intermediate level, test-takers who passed this level are believed to be able to produce RCs but not without some slight inadequacies. If test-takers managed to pass the advanced level, then supposedly they should have no problem producing RCs and their language proficiency is higher than simply knowing how to make correct RCs. In other words, the GEPT test works as a reliable assessment tool for anchoring test-takers' RC development over its four levels; namely, elementary, intermediate, high-intermediate and advanced levels.

Acknowledgements

This research project was sponsored by the Language Training and Testing Center (LTTC) in Taiwan for the year of 2013-14. The sponsorship included a research grant and access to the GEPT test data for research, for which I am deeply grateful. The gratitude also goes to the anonymous LTTC reviewers for their valuable feedback to the final report of the project.

[3] For more details about the GEPT scoring criteria, please refer to the LTTC's official website at https://www.gept.org.tw/About/gept_01_02.asp

References

Comrie, B. (2007). The acquisition of relative clauses in relation to language typology. *Studies in Second Language Acquisition*, 29(2),301-309.

Croteau, K.C. (1995). Second language acquisition of relative clause structures by learners of Italian. In F. R. Eckman, D. Highland, P.W. Lee, J. Mileham, & R.R. Weber (Eds.), *Second language acquisition theory and pedagogy* (pp. 115-128). Mahwah, NJ: Erlbaum.

Dasinger, L., & Toupin, C. (1994). The development of relative clause functions in narratives. In R. Berman & D. I. Slobin (Eds.), *Relating events in narratives: A crosslinguistic developmental study* (pp. 457-514). Mahwah, NJ: Erlbaum.

Diessel, H. (2004). *The acquisition of complex sentences*. New York: Cambridge University Press.

Diessel, H., & Tomasello, M. (2005). A new look at the acquisition of relative clauses. *Language*, 81 (4), 882-906.

Doughty, C. J. (1991). Second language instruction does make a difference: Evidence from an empirical study of SL relativization. *Studies in Second Language Acquisition*, 13 (4), 431-469.

Eckman, F.R., Bell, L., & Nelson, D. (1988).On the generalization of relative clause instruction in the acquisition of English as a second language. *Applied Linguistics*, 9, 1-20.

Ellis. R. (1994) *The study of second language acquisition*. Oxford: Oxford University Press.

Gass, S., & Selinker, L. (2001). *Second language acquisition: An introductory course* (2nded.). Mahwah, NJ: Lawrence Erlbaum Associates.

Jeon, K.S., & Kim, H.Y. (2007). Development of relativization in Korean as a foreign language: The noun phrase accessibility hierarchy in head-internal and head-external relative clauses. *Studies in Second Language Acquisition*, 29(2), 253-276.

Keenan, E., & Comrie, B. (1977). Noun phrase accessibility and Universal Grammar. *Linguistic Inquiry,* 8, 63-99.

Matthews, S., & Yip, V. (2002). Relative clauses in early bilingual development: Transfer and universals. In A. Giacalone Ramat (Ed.), *Typology and second language acquisition* (pp. 39-81). Berlin: Mouton de Gruyter.

Ozeki, H., & Shirai, Y. (2005). Semantic bias in the acquisition of relative clauses in Japanese. In A. Burgos, M.R. Clark-Cotton, & S. Ha (Eds.), *Proceedings of the 29th Annual Boston University Conference on Language Development: Vol. 2* (pp. 459-470). Somerville, MA: Cascadilla Press.

Ozeki, H., & Shirai, Y. (2007). Does the noun phrase accessibility hierarchy predict the difficulty order in the acquisition of Japanese relative clauses? *Studies in Second Language Acquisition*, 29(2),169-196.

Shirai, Y., & Kurono, A. (1998). The acquisition of tense-aspect marking in Japanese as a second language. *Language Learning*, 48 (2), 245-279.

Yip, V., & Matthews, S. (2007). Relative clauses in Cantonese-English bilingual children: Typological challenges and processing motivations. *Studies in Second Language Acquisition*, 29(2), 277-300.