

國立臺灣大學管理學院資訊管理學研究所



博士論文

Department of Information Management

College of Management

National Taiwan University

Doctoral Dissertation

個人化電腦輔助出題於英文學習之研究

Personalized Computer-aided Question Generation
for English Language Learning

黃意婷

Yi-Ting Huang

指導教授：孫雅麗 博士

Advisor: Yeali S. Sun, Ph.D.

中華民國 104 年 6 月

June 2015



個人化電腦輔助出題於英文學習之研究

本論文係提交國立台灣大學
資訊管理學研究所作為完成博士
學位所需條件之一部份

研究生：黃意婷 撰
中華民國一百零四年六月



國立臺灣大學(碩、博)士學位論文 口試委員會審定書

(題目：個人化電腦輔助出題於英文學習之研究)

本論文係黃意婷君(學號 d97725008)在國立臺灣大學資訊管理學系、所完成之博士學位論文，於民國 104 年 6 月 4 日承下列考試委員審查通過及口試及格，特此證明

口試委員：

<u> </u>	<u> </u>
<u>楊控則</u>	<u>孫雅如</u>
<u>陳孟彰</u>	<u> </u>
<u>陳信平</u>	<u> </u>

所長： 黃意婷

致謝



首先我要感謝我的指導教授—孫雅麗老師和陳孟彰老師的費心指導，獲益匪淺。給予我無限的支持與包容，讓我盡情嘗試不同的研究方法，透過自身的實踐以獲得屬於自己的知識。亦感謝 Carnegie Mellon University 的 Jack Mostow 教授在我千里馬計畫期間的指導，嚴謹且一絲不苟的學術態度是我努力學習的典範。

十分感謝口試委員中央大學學習與教學所 David Wible 教授、網路與學習科技所楊接期教授與台灣大學資訊工程系陳信希教授、資訊管理系陳建錦教授，特地撥冗於論文口試時給予指正與建議，使本論文能夠更加完善。

感謝許多人對此論文的幫助：淡江大學資訊工程系郭經華教授帶領的 IWiLL 團隊，張紋禎、狀輝在實驗上的幫忙；IWiLL 18 屆、19 屆、21 屆新聞閱讀活動參加實驗的老師(高雄中正高中邱昭敏、基隆女中連珍玲、屏東女中張惠勝、大園高中陳怡蓉、景美女中劉慧平、中央附中張育偉)與同學們在資料收集上的幫忙。以及實驗室的曾雅敏、張筱珮學妹們在研究上的幫忙，和中研院資訊所的夥伴陳建名、葉基善在技術上的支援。

另外感謝系上張孟琦與王心如助教和實驗室助理黃麒曉和黃裕雯協助我行政上事務。感謝學長姊陳君銘、蕭舜文、陳力銘、李振維、黃福銘、莊明晉、林義堅等人以及學弟妹周振濤、施岱伶、陳怡寧、張淳富、吳昌桓、林意婷、林昕彥等人平日的幫忙。還有我的朋友何宛靜、林永菁、蔡佩吟、盧依麟、侯慧芳、顏勝洲、范姜士燠、王俊人、Jiyeon Kim 以及孔令鵬等人，感謝你們的陪伴與鼓勵。

最後，感謝我摯愛的家人：黃振泰先生、廖秀治女士、黃子瑜、黃子瑋、陳怡臻、黃禾勳，謝謝你們的支持與鼓勵，謹以此文獻給你們。

論文摘要

論文題目：個人化電腦輔助出題於英文學習之研究

作者：黃意婷 一百零四年五月

指導教授：孫雅麗 博士



過去幾年來，電腦輔助自動出題(Computer-aided Question Generation)研究在結合自然語言處理(Natural Language Processing)的技術和計算語言學(Computational Linguistics)的方法，受到電腦輔助語言學習(Computer-assisted Language Learning)領域中越來越多的關注。為了提供以英文為第二外語的學習者自我學習評量，本研究提出個人化方法，以判斷學習教材難易度及評估學生程度的機制，應用於電腦輔助自動出題。在判斷閱讀難易度(Reading Difficulty Estimation)部分，根據學生的學習教材，考量豐富的語言特徵以及學生語言習得年級分布(language acquisition grade distributions)，針對第二語言學習特性，提出適合第二語言學習者閱讀難易度分析；在評估學生程度(Ability Estimation)的部分，結合(Item Response Theory)和年級分布，考量受試者長期的測驗結果來估計學生實際程度。在自動出題的部分，考量單字、文法與閱讀能力有交互作用影響，提出不同難度的單字、文法與閱讀測驗出題方法，利用評估學生程度機制獲得學生的能力估計，抽取與學生程度相符的閱讀素材和考題進行測驗，考試的結果也作為下一次個人化出題參考。實驗結果顯示閱讀難易度估計和能力程度評估可以比過去相關研究還要準確；此外，透過個人化電腦輔助出題系統的協助下，學習者可以減少重複犯錯，並且有明顯的進步。

關鍵詞：電腦輔助自動出題、閱讀難易度評估、學生能力程度估計、項目反應理論、電腦輔助語言學習

THESIS ABSTRACT

Personalized Computer-aided Question Generation
for English Language Learning



By Yi-Ting Huang

DOCTOR OF PHILOSOPHY


DEPARTMENT OF INFORMATION MANAGEMENT

NATIONAL TAIWAN UNIVERSITY

June 2015

ADVISER: Dr. Yeali S. Sun

In recent years, there has been increasing attention to computer-aided question generation in the field of computer assisted language learning and Natural Language Processing (NLP). However, the previous related work often provides examinees with an exhaustive amount of questions that are not designed for any specific testing purpose. In this study, we present a personalized automatic quiz generation that generates multiple-choice questions at various difficulty levels and categories, including grammar, vocabulary, and reading comprehension. We also design a reading difficulty estimation to predict the readability of a reading material, for learners taking English as a foreign language. The proposed reading difficulty estimation is based not only on the complexity of lexical and syntactic features, but also on several novel concepts, in-



cluding the word and grammar acquisition grade distributions from several sources, word sense from WordNet, and the implicit relations between sentences. Moreover, we combine the proposed question generation with a quiz strategy for estimating a student's ability and question selection. We develop a statistical and interpretable ability estimation. This method captures the succession of learning over time and provides an explainable interpretation of a statistical measurement, based on the quantiles of acquisition distributions and Item Response Theory (IRT). The concepts behind incorrectly answered questions are reincorporated into future tests in order to improve the weaknesses of examinees. The results showed that proposed second language reading difficulty estimation outperforms other first language reading difficulty estimations and the proposed ability estimation showed more accurate and robust than other ability estimations. In an empirical study, the results showed that the subjects with the personalized automatic quiz generation corrected their mistakes more frequently than ones only with computer-aided question generation. Moreover, subjects demonstrated the most progress between the pre-test and post-test and correctly answered more difficult questions.

Keyword: Computer-aided question generation, reading difficulty estimation, ability estimation, Item Response Theory, Computer assisted language learning.

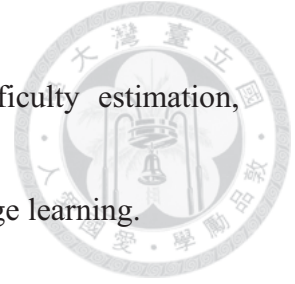
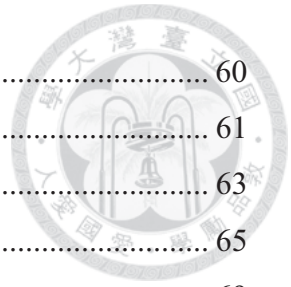


Table of Contents



口試委員會審訂書	i
致謝	ii
論文摘要	iii
THESIS ABSTRACT	iv
Table of Contents.....	vii
List of Tables.....	ix
List of Figures.....	x
Chapter 1 Introduction	1
1.1 Background.....	1
1.2 Research problem	4
1.3 Research purpose.....	5
Chapter 2 Related Work	12
2.1 Question generation.....	12
2.1.1 Computer-aided question generation for language learning.....	12
2.1.2 Question generation in natural language processing.....	16
2.1.3 The importance of the generated questions	19
2.2 Personalization	21
2.2.1 Reading difficulty estimation	21
2.2.2 Ability estimation	25
Chapter 3 Computer-aided Question Generation	29
3.1 Vocabulary question generation.....	34
3.2 Grammar question generation	37
3.3 Comprehension question generation	41
Chapter 4 Personalization	47
4.1 Reading difficulty estimation	47
5.1.1 Baseline features.....	49
5.1.2 The word acquisition grade distributions features.....	51
5.1.3 Frequency features.....	53
5.1.4 Parse features.....	55
5.1.5 The grammar acquisition grade distributions features	59

5.1.6	Semantic features.....	60
5.1.7	Relation features	61
5.1.8	Regression model	63
5.2	Ability estimation	65
5.3	Quiz Selection	69
Chapter 5	Evaluation on reading difficulty estimation	72
5.1	Data set	72
5.2	Metrics	73
5.3	Evaluation of the features	75
5.4	Optimal model selection.....	80
6.5	Reading difficulty estimation as classification	89
Chapter 6	Simulation on ability estimation.....	94
6.1	Setting.....	94
6.2	The characteristics of the proposed ability estimation	97
6.3	The comparison with other ability estimations	103
Chapter 7	An empirical Study	107
7.1	System and materials	107
7.2	Participants and procedure.....	110
7.3	The performance of the proposed ability estimation with the empirical data	112
7.4	Student performance.....	118
7.5	Unclear concept enhancement	124
7.6	User satisfaction	125
Chapter 8	Discussion and Conclusion.....	130
8.1	Summary.....	130
8.2.	Contribution.....	133
8.3	Limitations.....	138
8.4	Future applications	140
References	142



List of Tables

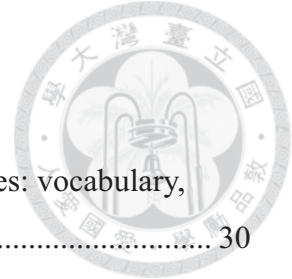


Table 1 Design of personalized questions with different question types: vocabulary, grammar, and reading comprehension questions.....	30
Table 2 Distractor templates were referred by a grammar textbook and an expert in order to ensure the disambiguation of distractors.....	39
Table 3 Results of RMSE and correlation among different feature categories.	76
Table 4 Results of the optimal model selection.....	83
Table 5 Results of the optimal model selection.....	89
Table 6 Comparison between the estimations.	93
Table 7 The results of convergence point and RMSE (each row represents the degree of difference between the initial ability and the actual ability, and each column represents the number of time periods considered by the exponential weight of the current ability)100	
Table 8 The results of RMSE between MLE, Lee (2012) and the proposed ability estimation	106
Table 9 The correlation result between the estimated ability and the post-test in the control group and the experimental group.....	113
Table 10 The mean post-test score of the subjects in different estimated ability groups between both groups and the result of ANOVA	116
Table 11 The equations among question types represent that the log odds ratio of the observation that the student i correctly answers item j is in class 1 or the student incorrectly answers item j is in class 0	118
Table 12 The results of the pretest and post-test between the control group and the experimental group.....	121
Table 13 Contingency tables for the number of correctly answered questions per difficulty level in the pretest and post-test.....	123
Table 14 The mean and standard deviation of rectification rate.....	125
Table 15 Questionnaire results.	126
Table 16 Comparison of different test environments.	134

List of Figures

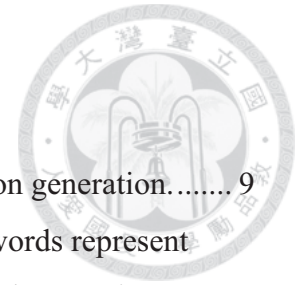


Figure 1 The architecture of the personalized computer-aided question generation.....	9
Figure 2 A paragraph and example generated questions: the bolded words represent stems, the bold italics are answers and the other plausible choices in the questions are called as distractors.....	34
Figure 3 The parse structure of the sentence “ <i>Many of the original Halloween traditions have developed today into fun activities for children</i> ”.....	41
Figure 4 A table of a database in the implemented system captures the incorrectly answered concepts of a student.	71
Figure 5 the performance of a selected model.....	87
Figure 6 The changes in the estimated ability computed from the proposed method for the different weights (n=1, n=3, n=6, n=12)	102
Figure 7 Snapshots of the system: (a) An example of a given reading materials from new online website; (b) An example of vocabulary items; (c) An example of grammar items; (d) An example of reading comprehension items; (e) An example of a score result with explicit warning.	109
Figure 8 The charts on the percentage value vary from strongly agree to the strongly disagree for item six (upper left), item seven (upper right), item eight (lower left) and item nine (lower right).....	129

Chapter 1 Introduction




1.1 Background

For many years, educational assessment has played an important role in teaching and learning (Gronlund, 1993). It can evaluate the effectiveness of teaching, diagnose the state of learning, and help the development of students' learning (Chen, Lee, & Chen, 2005; Chen & Chung, 2008; Johns, Hsingchin, & Lixun, 2008; Barla et al., 2010). With the development of computers and the Internet, Computer Adaptive Testing (CAT) is now a developing way to administer tests adapting to learners' knowledge or competence in language learning (Troubley, Heireman, & Walle, 1996). Based on adaptive tests, examinees' abilities can be more accurately measured by fewer suitable questions (Weiss & Kingsbury, 1984; Van der Linden & Glas, 2000); moreover, student performance has also been demonstrated improved (Barla et al., 2010). CAT can not only provide questions but also be combined with scaffolding hints and instructional feedback (Feng, Heffernan & Koedinger, 2010). This facilitates students learning and helps them acquire knowledge with external help. However, when a great number of reading materials is exponentially growing every day, there is room for improvement



in assessment resources because it is time-consuming and cost-intensive for human experts to manually produce questions.

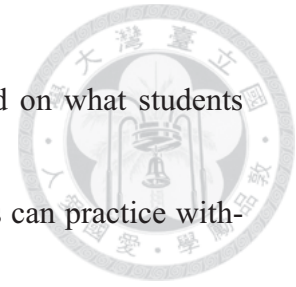
In recent years, there has been increasing attention to computer-aided question generation (also called automatic question generation or automatic quiz generation) in the field of e-learning and Natural Language Processing (NLP). It is useful in multiple subareas and has been proposed to use in generating instructions in tutoring system (Mostow & Chen, 2009), assessing domain knowledge (Mitkov & Ha, 2003), evaluating language proficiency (Brown, Frishkoff, & Eskenazi, 2005), assisting academic writing (Liu, Calvo, Aditomo, & Pizzato, 2012) and question answering (Pasca, 2011). In order to make learning environment more effectively and efficiently, many researchers have been exploring the possibility of an automatic question generation in various contexts. For example, a wide variety of applications, such as Linguistics (Mitkov, Ha, & Karamanis, 2006) and Biology (Agarwal & Mannem, 2011), identified the important concepts in textbooks and generated multiple-choice questions and gap-fill questions. In the domain of language learning, a growing number of studies (Turney, 2001; Turney, Littman, Bigham, & Shnayder, 2003; Liu, Wang, Gao, &



Huang, 2005; Sumita et al., 2005; Lee & Seneff, 2007; Lin, Sung, & Chen, 2007; Pino, Heilman, & Eskenazi, 2008; Smith, Avinesh, & Kilgarriff, 2010) are now available to not only drills and exercise, including vocabulary, grammar, reading questions, but also formal exams, including SAT (Scholastic Aptitude Test) analogy questions and TOEFL (Test of English as a Foreign Language) synonym task. To support academic writing, Liu et al. (2012) used Wikipedia and the conceptual graph structures of research papers and generated specific trigger questions for supporting literature review writing.

Several researches have addressed the benefit of facilitative learning and teaching with automatic question generation. The use of computer-aided question generation for educational purpose was motivated as research of reading comprehension consistently found that assessment is helpful in learning and enhances learners' retention of material (Anderson & Biddle, 1975). Mitkov et al. (2006) demonstrated that computer-aided question generation was more time - efficient than manual labor. Turney et al. (2003) showed that the generated SAT and TOEFL questions are comparable to that generated by experts. Liu et al. (2012) found that the generated trigger questions were more useful than manual generic questions and that the questions could prompt students to re-

flect on key concepts, because the questions were generated based on what students read. With the advantage of automatic question generation, students can practice without waiting for a teacher to compose a quiz, and teachers can spend more time on teaching; moreover, besides evaluating students' understanding, automatic question generation can be designed with additional functions.



1.2 Research problem

Recent theories on learning have focused increasing attention on understanding and measuring student ability. There is now general consensus over Vygotsky's (1978) observation that a learner's ability in the Zone of Proximal Development (ZPD)—the difference between a learner's actual ability and his or her potential development—can progress well with external help. Instructional scaffolding (Wood, Bruner, & Ross, 1976), closely related to the concept of ZPD, suggests that appropriate support during the learning process helps learners achieve their learning goals. However, effective instructional support requires identifying students' prior knowledge, tailoring assistance to meet their initial needs, and then removing this aid when they acquire sufficient


knowledge.



Even though previous studies in the field of computer-aided question generation automatically generate all possible questions based on their proposed approach in an attempt to reduce the cost of time and money of manual question generation, such exhaustive list of questions is inappropriate for language learning, because it can lead to redundant, over-simplistic test questions that are unsuitable for evaluating student progress. Moreover, it is hard to achieve meaningful test purpose and maximize examinees' learning outcomes because the personalized design (Fehr et al., 2012; Hsiao, Chang, Chen, Wu, & Lin, 2013; Wu, Su, & Liu, 2013) is still critically lacking.


1.3 Research purpose

This work is intended to provide personalized computer-aided question generation on formative assessment to assess students' receptive skills in English as a foreign or second language. It generates three question types, including vocabulary, grammar and reading comprehension, and differs from previous studies in the way learners' language proficiency levels are considered in the generating process and questions are

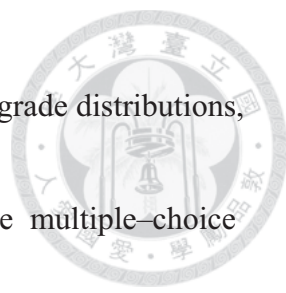


generated with difficulties. The definition of “*personalization*” refers to the adjustment to learner needs by matching the difficulty of questions to their knowledge level. In other words, questions are generated based on an individual’s ability even though students read the same learning material.

This work, the personalized computer-aided question generation, is based on the related concept to the age of acquisition (AOA). The basic idea of age of acquisition is the age at which a word, a concept, even specific knowledge is acquired. For instance, people learn some words such as “dog” and “cat” before others such as “calculus” and “statistics”. Numerous studies in psychology and cognitive science have shown the positive influence on the process of brain, such as object recognition (Urooj et al., 2013), object naming (Carrolla & Whitea, 1973; Morrison, Ellis, & Quinlan, 1992; Alario, Ferrand, Laganaro, New, Frauenfelder, & Segui, 2005; Davies, Barbón, & Cuetos, 2013), and language learning (Brysbaert, Wijnendaele, & Deyne, 2000; McDonald, 2000; Izura & Ellis, 2002; Zevin & Seidenberg, 2002). Today, with the various number of content available from the web and other digital resources, this concept can be realized with advanced technology, Information Retrieval (Baeza-Yates



& Ribeiro-Neto, 1999; Manning, Raghavan, & Schütze, 2008) and Natural Language Processing (Manning & Schütze, 1999), which counts word frequency and calculates the probability of which a word is acquired at a certain school grade when given a group of documents. With a large enough resource, such as an extensive collection of all learning materials which people read and learn, the acquisition grade distributions can be computed and implemented. For example, based on textbooks authored specifically for students in grade level six, questions can be generated based on concepts in these textbooks that were correctly answered by a student, and from this, the student can be said to either have or lack the skills at the grade level six. This implies that learning materials, such as textbook, are written with intent to represent what learners at a certain grade level learn and acquire. Two related work to this concept are a readability prediction (Kidwell, Lebanon & Collins-Thompson, 2011), which mapped a document to a numerical value corresponding to a grade level based on the distribution of acquisition age, and a word difficulty estimation (Kireyev & Landauer, 2011), which modeled language acquisition with Latent Semantic Analysis to compute the degree of knowledge of words at different learning stages.



In response to the personalized design based on the acquisition grade distributions, we propose a personalized automatic quiz generation to generate multiple-choice questions with varying difficulty, a reading difficulty estimation to predict the difficulty level of an article for English as foreign language learners, as well as an interpretable and statistical ability estimation to estimate a student's ability with inherent randomness in the acquisition process, specifically in the Web-based learning environment, as shown in Figure 1.

The purpose of personalized testing is to not only measure the achievement performance of students, but also help them improve their own learning process and correct their mistakes by understanding what they has learned and has not learned yet. Through this approach, students can read any materials online and then do more exercises to understand their strengths and to improve their weaknesses, as a strategy to guide them to language acquisition.

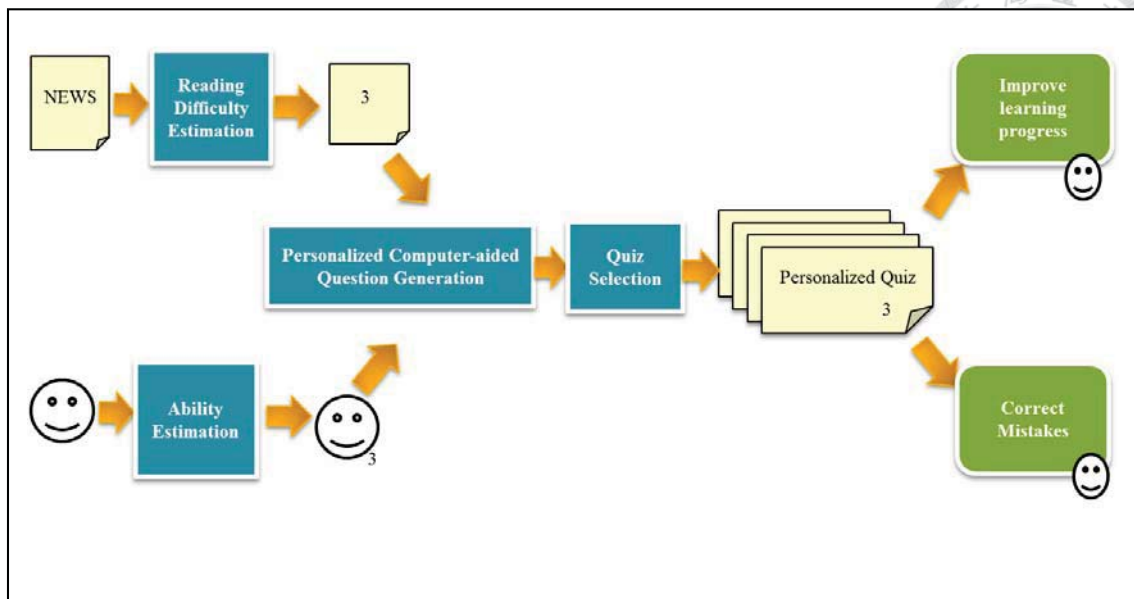


Figure 1 The architecture of the personalized computer-aided question generation.

The main research questions addressed in this study are:

- (1) Does the proposed personalized design with the appropriate instructional scaffolding help students advance their learning progress?
- (2) Does the proposed personalized question selection help students correct their unclear concept?
- (3) How are students' perceptions and experiences in the proposed personalized computer-aided question generation?

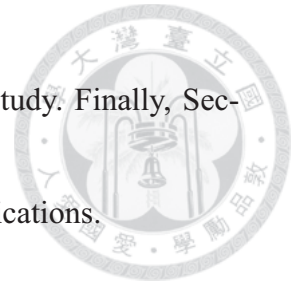
We also conduct simulation and empirical evaluations to investigate the property of the proposed personalized functions. The research questions are as the following:



- (4) What are the representative features of the proposed reading difficulty estimation in English as a foreign or second language?
- (5) How is the performance of the proposed reading difficulty estimation compared with the other reading difficulty estimation?
- (6) What are the characteristics of the proposed ability estimation based on the quantiles of acquisition grade distributions and item response theory?
- (7) How is the performance of the proposed ability estimation compared with the other ability estimations?
- (8) How is the performance of the proposed ability estimation with the empirical data in a Web-based learning environment?

The rest of this article is organized as follows. Chapter 2 describes related work. In Chapter 3, we present the design of automatic quiz generation and the mechanism for assigning question difficulty. Chapter 4 outlines the personalization framework, consisting of reading difficulty estimation, ability estimation and quiz selection. In Chapter 5 and Chapter 6, we present simulation evaluations of reading difficulty estimation and ability estimation respectively. Chapter 7 evaluates the effectiveness of

personalized computer-aided question generation in the empirical study. Finally, Section 8 summarizes with contributions, limitations, and potential applications.



Chapter 2 Related Work

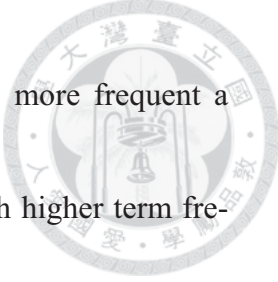


In this chapter, the background of question generation is presented, including computer-aided question generation for education purpose, and in natural language processing. Next, the related work of reading difficulty estimation is also introduced. Finally, a modern theory of testing, Item Response Theory, will be discussed.

2.1 Question generation

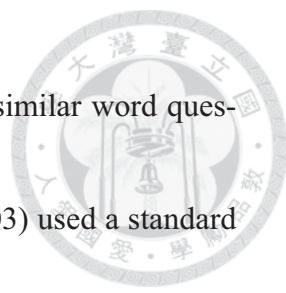
2.1.1 Computer-aided question generation for language learning

Computer-aided question generation is the task of automatically generating questions, which consists of a stem, a correct answer and distractors, when given a text. These generated questions can be used as an efficient tool for measurement and diagnostics. The first computer-aided question generation was proposed by Mitkov and Ha (2003). Multiple-choice questions are automatically generated by three components: term extraction, distractor selection and question generation. First, noun phrases are



extracted as answer candidates and sorted by term extraction. The more frequent a term appears, the more important the term becomes. The terms with higher term frequency consequently serve as answers to the generated questions. Next, WordNet (Miller, Beckwith, Fellbaum, Gross, & Miller, 1990) is consulted by the distractor selection in order to capture the semantic relation between each incorrect choice and the correct answer. Finally, the generated questions are formed by predefined syntactic templates. Most of the following studies are based on such system architecture.

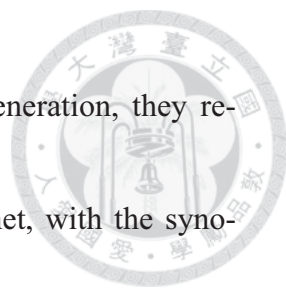
A growing number of researches are now available to shed some light on the domain of English language learning, such as vocabulary, grammar and comprehension. It is because in these question generations, linguistic characteristics are analyzed to help produce items, just like what experts do. In vocabulary assessment, Liu et al. (2005) investigated word sense disambiguation to generate vocabulary questions in terms of a specific word sense, and considered the background knowledge of first language of test-takers to select distractors. Lin et al. (2007) analyzed the semantics of words and develop algorithm to select candidates as a substitute word from WordNet (Miller et al., 1990) and filtered by web corpus searching. They presented adjective-



noun pair questions, including collocation, antonym, synonym and similar word questions in order to test students' understanding in semantic. Turney (2003) used a standard supervised machine learning approach with feature vectors based on the frequencies of patterns in a large corpus to automatically recognize analogies, synonyms, antonyms, and associations between words, and then transformed those word pairs into multiple-choice SAT (Scholastic Assessment Tests) analogy questions, TOEFL synonym questions and ESL (English as second language) synonym–antonym questions.

In grammar assessment, Chen et al. (2005) focused on automatic grammar quiz generation. Their FAST system analyzed items from the TOEFL test and collected documents from Wikipedia to generate grammar questions using a part-of-speech tagger and predefined templates. Lee and Seneff (2007) particularly discussed algorithm to generate questions for prepositions in language learning. They proposed two novel distractor selections, one is applied a collocation–based method and the other is the usage of the deletion error in a non-native corpus.

In reading comprehension assessment, the MARCT system (Yang et al., 2005) designed three question types, including true-false question, numerical information



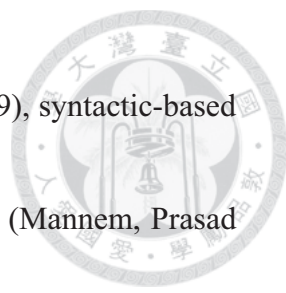
question and not-in-the-list questions. In the true-false question generation, they replaced words in a sentence, extracted from an article on the Internet, with the synonyms or antonyms by using WordNet (Miller et al., 1990). In the numerical information question generation, they listed some specific trigger words, such as “*kilogram*”, “*square foot*”, and “*foot*”, corresponding to some predefined templates, such as “*what is the weight of*”, “*how large*”, and “*how tall*”. In the not-in-the-list question generation, they used terms listed in Google Sets to identify the question type and select distractors. Unlike previous methods, Mostow and Jang (2012) designed different types of distractor to diagnose the cause of comprehension failure, including ungrammatical, nonsensical, and plausible failures. Especially, the plausible distractors considered the context in reading materials. They used a Naïve Bayes formula to score the relevance to the context in paragraph and words earlier in sentence. A student’s comprehension is judged by not only evaluating one’s vocabulary knowledge but also testing the ability to decide which word is consistent with the surrounding context.

2.1.2 Question generation in natural language processing



Question generation has been primarily concerned by the natural language processing community through the question generation workshop and the shared task in 2010 (QGSTEC 2010; Rus et al., 2010). It is an important task in many different applications including automated assessment, dialogue systems (Piwek, Prendinger, Hernault & Ishizuka, 2008), intelligent tutoring systems (Chen & Mostow, 2011), and search interfaces (Pasca, 2011). The aim of the task is to generate a series of questions based on the raw text from sentences or paragraphs. The question types includes *why*, *who*, *when*, *where*, *when*, *what*, *which*, *how many/long* and *yes/no* questions. Generally, the procedure of question generation task can be characterized in three components: content selection, the identification of a question type, and question formulation. First, the content selection identifies which part of the given text is worthy of being generated as a question. When the content is given, the identification will determine the question type. Finally, the question formulation transforms the content into a question.

Many generation approaches to *wh*-questions have been developed, inclusive of



template-based (Chen, Aist, & Mostow, 2009; Mostow & Chen, 2009), syntactic-based (Heilman & Smith, 2009; Heilman & Smith, 2010), semantic-based (Mannem, Prasad & Joshi, 2010; Yao, Bouma, & Zhang, 2012), and discourse-based approach (Prasad and Joshi, 2008; Agarwal, Shah & Mannem, 2011). To identify question type, both of template-based and syntactic-based approaches focused on lexical information and the syntactic structure of a single sentence and transformed them into questions. Chen et al. (2009) enumerated words with conditional context, temporal context and modality expression, such as “*if*”, “*after*”, and “*will*”, as criteria for selecting questioning indicators. Based on these indicators, they defined six specific rules to transform the informative sentence into questions, like “*What would happen if <x>?*” in conditional context, “*When would <x>?*” in temporal context and “*What <auxiliary-verb> <x>?*” in linguistic modality. On the other hand, Heilman and Smith (2009) analyzed the structures of sentences and proposed general-purpose rules using part-of-speech (POS) tags and category labels. The question generation, produce derived sentences from complex sentences and transform declarative sentences into questions, can generate more grammatical and readable questions rather than leading to unnatural or senseless


questions.



Since inter-sentential causal relations can also be identified by a semantic parser, such as a semantic role labeler, semantic-based question generations made use of the additional information of the semantic role labeling along with the marked relations.

Mannem, Prasad and Joshi (2010) used the predicate argument structures along with semantic roles to identify important aspects of paragraphs. For instance, the label “*ARGM-CAU*” can be seen as a cause clause marker. When the marker is recognized, a corresponding question type, like “*why*”, will be generated. Similarly, with semantic information, MrsQG system (Yao et al., 2012) transformed declarative sentences into the Minimal Recursion Semantics (MRS, Copestake, Flickinger, Pollard, & Sag, 2005), a theory of semantic representation of natural language sentences. And then MRS representations of declarative sentences were mapped to interrogative sentences.

Cross-sentence information, such as discourse relation, has been particularly influential in contributing insights into question generation in the recent year. Prasad and Joshi (2008) firstly used causal relations in the Penn Discourse Treebank (PDTB; Prasad et al., 2008) as content selection trigger. They found that the PDTB causal relations

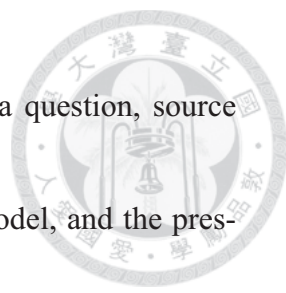


can be seen as providing the source for 71% of the *why*-questions in their experiment settings. This implied the potential for PDTB and intrigued subsequent research to follow such concept. Like Agarwal et al. (2011), they used explicit discourse connectives, such as “*because*”, “*when*”, “*although*” and “*for example*”, to select content for question formation, and construct questions involving sense disambiguation of the discourse connectives, identification of question type and applying syntactic transformations on the content.

These techniques from the field of question generation may facilitate the development of the question generations in the various forms of question types. However, unlike previous research directly related to the topic of generating questions for educational purpose, the question generation only focused on generating questions based on the given context, these related studies were not involved in the distractor selection.

2.1.3 The importance of the generated questions

While many work pertaining to computer-aided question generation have focused on the procedure of question generation and distractor selection, little work analyzed the importance of the generated questions. Heilman and Smith (2010) proposed lin-



guistic features, such as the number of tokens or noun phrases in a question, source sentence and answer phrase, the score from the n-gram language model, and the presence of questioning words or negative words, to statistically rank the quality of generated questions. Agarwal and Mannem (2011) considered lexical and syntactic features, like the similarity between sentence and the title of a given text, the presence of abbreviation, discourse connective and superlative adjective, to select the most informative sentences from a document, and generated questions on them. Chali and Hasan (2012) considered that questions associated with these topics should be generated first, so they used Latent Dirichlet Allocation (LDA) to identify the sub-topics, which are closely related to the original topic, in the given content, and next applied the Extended String Subsequence Kernel (ESSK) to calculate their similarity with the questions and computed the syntactic correctness of the questions by tree kernel. Although these output questions were improved by considering linguistic features, these studies still did not take examinees into consideration.

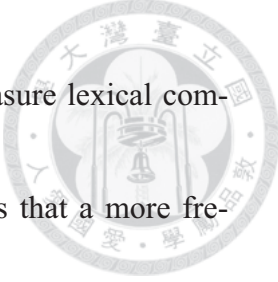


2.2 Personalization

2.2.1 Reading difficulty estimation

Reading difficulty (also called readability) is often used to estimate the reading level of a document, so that readers can choose appropriate material for their skill level. Heilman, Collins-Thompson, Callan and Eskenazi (2007) described reading difficulty as a function of mapping a document to a numerical value corresponding to a difficulty or grade level. A list of features extracted from the document usually acts as the inputs of this function, while one of the ordered difficulty grade levels is the output corresponding to a reader's reading skill.

Early related work on estimating reading difficulty only used a few simple features to measure lexical complexity, such as word frequency or the number of syllables per word. Because they took fewer features into account, most studies made assumptions on what variables affected readability, and then based their difficulty metrics on these assumptions. One example is the Dale-Chall model (Dale and Chall 1948), which determined a list of 3,000 commonly known words and then used the percentage of words to measure lexical difficulty. Another example is the Lexile Framework (Stenner,



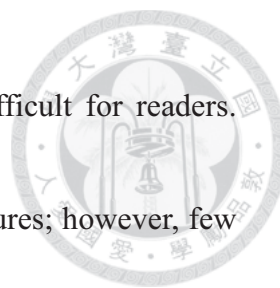
1996), which used the mean log word frequency as a feature to measure lexical complexity. Using word frequency to measure lexical difficulty assumes that a more frequent word is easier for readers. Although this assumption seems fair, since a widely used word has a higher probability to be seen and absorbed by readers, this method is not always true when there are the numerous differences in diverse words acquired by different language learners. This method is susceptible to the diverse word frequency rates found in various corpora.

More recent approaches have started to take n-gram language models into consideration to assess lexical complexity, which can measure difficulty more accurately. Collins-Thompson and Callan (2004) used the smoothed unigram language model to measure the lexical difficulty of a given document. For each document, they generated language models by levels of readability, and then calculated likelihood ratios to assign the level of difficulty; in other words, the predicted value is the level with the highest likelihood ratio of the document. Similarly, Schwarm and Ostendorf (2005) also utilized statistical language models to classify documents based on reading difficulty level, and they found that trigram models are more accurate than bigram and unigram

ones.



In addition to using fairly basic measures to calculate lexical complexity, prior studies often only calculated the mean number of words per sentence to estimate grammatical readability. Using sentence length to measure grammatical difficulty assumes that a shorter sentence is syntactically simpler than a longer one. However, long sentences are not always more difficult than shorter sentences. In response, more recent approaches have started to consider the structure of sentences when measuring grammatical complexity and making use of increasingly precise parser accuracy rates. These researches usually considered more grammatical features such as parse features per sentence in order to make a more accurate difficulty prediction. Schwarm and Ostendorf (2005) employed four grammatical features derived from syntactic parsers. These features included the average parse tree height, the average number of noun phrases, the average number of verb phrases, and the average number of subsidiary conjunctions to assess a document's readability. Similarly, Heilman et al. (2008) used grammatical features extracted from an automatic context-free grammar parse trees of sentences, and then computed the relative frequencies of partial syntactic derivations.



In their model, the more frequent sub-trees are viewed as less difficult for readers.

These approaches have investigated the effect of the sentence structures; however, few studies have been examined the effect of language learners on the grammar acquisition grade distributions.

The majority of research on reading difficulty has focused on documents written for native readers (also called first language), and comparatively little work (Heilman et al., 2007) has been done on the difficulties of documents written for second language learners. Second language learners have a distinct way to acquire second language from native speakers. As Bates (2003) pointed out, there are wide differences in the learning timelines and processing times between native and non-native readers; first language learners learn all grammar rules before formal education, whereas second language learners learn grammatical structures and vocabulary simultaneously and incrementally. Almost all first-language reading difficulty estimations focus on vocabulary features, while second-language reading difficulty estimations especially emphasize grammatical difficulty (Heilman et al., 2007). Wan, Li and Xiao (2010) found that college students in China still have difficulty reading English documents written for

native readers, even though they have learned English over a long period of time.

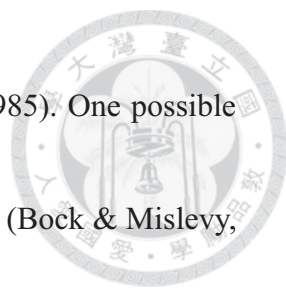
These studies indicate that it is unsuitable to apply a first-language reading difficulty

estimation directly; instead, second-language reading difficulty estimation must be de-

veloped.

2.2.2 Ability estimation

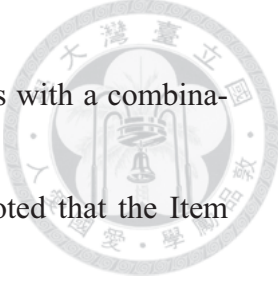
Item Response Theory (Embretson & Reise, 2000) is a modern theory of testing that examines the relationship between an examinee's responses and items related to abilities measured by the items in the test. One of the interesting characteristics of Item Response Theory is that an ability parameter and item parameters are invariant, while these parameters in Classical Test Theory (CTT) vary by sample (Crocker & Algina, 1986). Three well-known ability estimations proposed by Item Response Theory are maximum likelihood estimation (MLE), maximum a posteriori (MAP) and expected a posteriori (EAP). The procedure of MLE, an iterative process, is to find the maximum likelihood of a response to each item for an examinee. However, when an examinee correctly answers none or all of questions in a test, the MLE fails to find a convergence



point during the estimated iteration (Hambleton & Swaminathan, 1985). One possible solution to this problem involves using MAP (Baker, 1993) and EAP (Bock & Mislevy, 1982), which are variants of Bayes Modal Estimation (BME) and incorporate prior information into the likelihood function. Prior distributions can protect against outliers that may have negative influence on ability estimation. For example, Barla et al. (2010) employed EAP to score each examinee's ability for each test.

Even though Item Response Theory has been used for decades, the estimation procedure of Item Response Theory is computation-intensive. Until recently, with the rapid development of the computer industry, Item Response Theory has been increasingly used in e-learning applications as an offline service. However, Item Response Theory has till now had little application in Web-based learning environments, which is unfortunate because a real-time and online assessment would be more desirable. Fortunately, Lee (2012) proposed an alternative computational approach in which a Gaussian fitting to the posterior distribution of the estimated ability could more efficiently approximate that determined by the conventional BME approach.

In a Web-based learning environment, Computerized Adaptive Testing is usually



seen as a part of a component in the environment, providing learners with a combination of practice and measurement. But Klinkenberg et al. (2011) noted that the Item Response Theory was designed for measurement only, the reason being that the parameters of items had to be pre-calibrated in advance before items were used in a test. Generally, during the item calibration, an item should be taken by a large number of people, ideally between 200 to 1000 people, in order to estimate reliable parameters for the items (Wainer & Mislevy, 1990; Huang, 1996). This procedure is very costly and time-consuming, and also less beneficial for learning environments. It is especially impractical because the calibration had to be conducted repeatedly in order to get accurate norm referenced item parameters. Alternatively, Klinkenberg et al. (2011) introduced a new ability estimation based on Elo's (1978) rating system and an explicit scoring rule. Elo's rating system was developed for chess competitions and used to estimate the relative ability of a player. With this method, pre-calibration was no longer required, and the ability parameter was updated depending on the weighted difference between the response and the expected response. This method was employed in a Web-based monitoring system, called a computerized adaptive practice (CAP) system,

and designed for monitoring arithmetic in primary education.



Although much work has been done thus far, there are still some problems that have attracted little attention. First, although every exercise performed by a student is recorded in most of the Web-based learning environments listed above, the ability estimations of Item Response Theory only consider test responses at the time of testing, rather than incorporating a testing history. Moreover, the result of estimating an examinee's ability is often defined in terms of a norm referenced value, the interpretation of which in the most ability estimations is often defined as a number or a sign. For example, a student with the specific ability, such as level six, means he has a large proportion of knowledge similar to other students in grade level six. Unfortunately, as this definition is qualitative rather than quantitative, this approach cannot provide a quantitative result in terms of a student's understanding.

Chapter 3 Computer-aided Question Generation



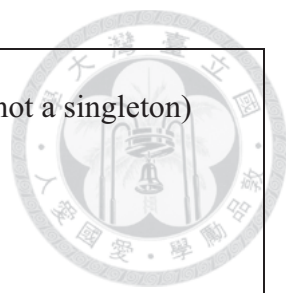
How to generate personalized questions in different question types? In this chapter, we will respectively describe the constraints on vocabulary questions, grammar questions, and reading comprehension questions.

Table 1 summarizes how to define the question difficulty and how distractors are selected and Figure 2 shows four questions (also called items) generated from a document describing the origins of Halloween.



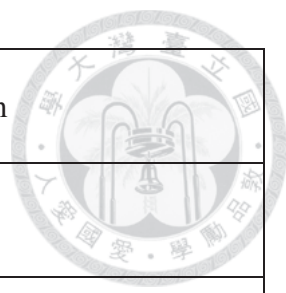
Table 1 Design of personalized questions with different question types: vocabulary, grammar, and reading comprehension questions.

	Vocabulary question	Grammar question	Independent referential question	Overall referential question
How to define question difficulty?	a graded word list	grammar frequency	reading difficulty estimation	



<p>How to select a target sentence (with answer)?</p>	<p>a word</p>	<p>a sentence</p>	<p>a referent (not a singleton)</p>	
<p>Stem template</p>	<p>In the sentence "... _____ ...", the blank can be:</p>	<p>In the Sentence, "... _____ ...", the blank can be filled in:</p>	<p>The word "[<i>target word</i>]" in this sentence "[<i>target sentence</i>]" refer to:</p>	<p>Which of the following statement is TRUE?</p>
<p>Distractor candidate source</p>	<p>words from a graded word list</p>	<p>grammar patterns defined by a grammar book</p>	<p>other noun phrases (common nouns or proper nouns) in the given article</p>	
<p>Distractor</p>	<p>word difficulty</p>	<p>disambiguation</p>	<p>non-anaphora</p>	

selection	part-of-speech		not pronoun
	word length		number
	Levenshtein distance		gender



Document

Halloween, which falls on October 31, is one of the most unusual and fun holidays in the United States. It is also one of the scariest! **It is associated with ghosts, skeletons, witches, and other scary images. ...Many of the original Halloween traditions have developed today into fun activities for children.** The most popular one is "trick or treat." On Halloween night, children dress up in costumes and go to visit their neighbors. When someone answers the door, the children cry out, "trick or treat!" What this means is, "Give us a treat, or we'll play a trick on you!"... This tradition comes from an old Irish story about a man named **Jack** who was very stingy.

He was so stingy that he could not enter heaven when he died. But he also could not enter hell, because he had once played a trick on the devil. All he could do was walk the earth as a ghost, carrying a lantern...

Quiz

1. In the sentence "It is _____ with ghosts, skeletons, witches, and other scary images.", the blank can be:

(1) distributed (2) *associated* (3) contributed (4) illustrated

2. In the Sentence, "Many of the original Halloween traditions _____ today into fun activities for children.", the blank can be filled in:

(1) *have developed* (2) have developing (3) is developed (4) develop

3. The word "he" in this sentence "All *he* could do was walk the earth as a ghost, carrying a lantern" refer to:

(1) ghost (2) devil (3) witch (4) *Jack*

4. Which of the following statement is TRUE?

(1) On Halloween night, neighbors dress up in costumes and go to visit their children.

(2) What this means is, "Give us a trick, or we'll play a treat on you!"

(3) But the devil also could not enter hell, because he had once played a trick on the witch.

(4) ***Jack*** was so ***stingy*** that he could not enter heaven when he died.

Figure 2 A paragraph and example generated questions: the bolded words represent stems, the bold italics are answers and the other plausible choices in the questions are called as distractors.

3.1 Vocabulary question generation

The difficulty of a vocabulary question is determined by the difficulty of the correct answer. We assume if a student selects the correct answer, he/she probably understood the question stem and distinguished the correct answer from distractors. Here, the difficulty of a word refers to word acquisition, the temporal process by which learners learn the meaning, understanding and usage of new words. For most of English as foreign language learners, the acquisition grade distributions of different words can be drawn from the inference from textbooks or a word list made by experts, because English as foreign language learners learn foreign language depending on materials they study, not the environment they live. In this study, the word difficulty is de-

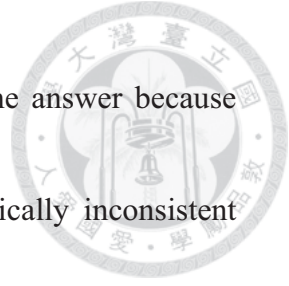
terminated by a word list made by an education organization. We adopted a wordlist from the College Entrance Examination Center (CEEC) of Taiwan



(http://www.ceec.edu.tw/research/paper_doc/ce37/5.pdf). It contains 6,480 words in English, divided into six levels, which represent the grade in which a word should be taught, as the word acquisition grade distributions. For each word from the given text, we identify its difficulty by first referencing its difficulty level from within the word list. When given the vocabulary proficiency level of a student, words with the same difficulty level in the given document are selected as the basis to form test questions.

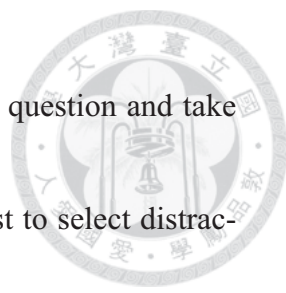
In the distractor selection, we also consult the same graded word list as the source of distractor candidates. The distractors were selected by the following criteria: word difficulty, part-of-speech (POS), word length and Levenshtein distance.

- Word difficulty: Distractors are selected with the equal difficulty for two reasons. One is for personalization. A student has personalized generated questions whose difficulty is as the same as the student's proficiency level. The other is for familiar. Choices must be familiar to students; otherwise the correct answer may be selected because students only know it.



- Part-of-speech (POS): Distractors have the same POS as the answer because this makes the target sentence grammatical, but is semantically inconsistent with the context of the target sentence. In this way, students can be tested the lexical knowledge and comprehension instead of syntax. We use Stanford POS Tagger (Toutanova, Klein, Manning, & Singer, 2003) to identify words as nouns, verbs, adjectives, or adverbs.
- Word length and Levenshtein distance: Distractors are ranked by the least small word length difference between a distractor and the correct answer and Levenshtein distance based on changing the prefix or postfix of a distractor into the correct answer. According to the (Perfetti & Hart, 2002), high-skilled students easily have confusion when words share phonological forms with other homophones. We try to catch the grapheme-phoneme by considering the word length and Levenshtein distance.

The first question in Figure 2 is a vocabulary question. When a knowledge level four student is given, difficulty level four words, e.g. “associate”, are identified by the graded word list. The sentence containing the word, “*It is associated with ghosts, skel-*

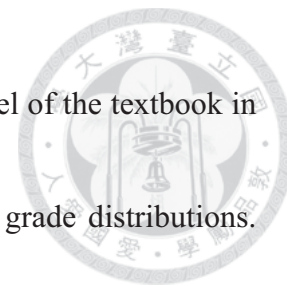


etons, witches, and other scary images”, is then extracted to form a question and take “associate” as the correct answer. We also consult the same word list to select distractors which have same difficulty (level 4) and part-of-speech (verb), and the least small distance of word length (9) and Levenshtein distance (distributed:6, illustrated:7, contributed:7).

3.2 Grammar question generation

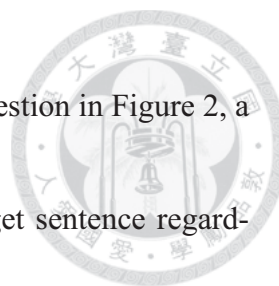
The difficulty of a grammar question, which similar as that of a vocabulary question, is determined by the difficulty of the grammar pattern of the correct answer. Unfortunately, unlike the aforementioned word list, there is no predefined grammar difficulty measure available. In addition, second language learners usually learn grammatical structures simultaneously and incrementally, while native speakers have learned all grammar rules before formal education. Second language learning materials are predominated by the well-thought learning plan. Thus, we assigned the difficulty of a grammar pattern based on the grade level of the textbook, which represents the grammar acquisition grade distributions.

The difficulties of grammar patterns rely identify the grade level of the textbook in which it frequently appears, representing the grammar acquisition grade distributions.



We manually predefined 44 grammar patterns from a grammar textbook for Taiwan high school students and automatically calculated the rate of occurrence of grammar patterns in a set of English textbooks. First, we used Stanford Parser (Klein and Manning, 2002) to produce constituent structure trees of sentences. And next Tregex (Levy & Andrew, 2006), a searching tool for matching patterns in trees, was used to recognize the instances of the target grammar patterns in the set of textbooks. Finally, we counted the frequencies of the syntactic grammar patterns in a set of corpus. This set of corpus contains 342 articles written by different authors and collected from five different publishers (including The National Institute for Compilation and Translation, Far East Book Company, Lungteng Cultural Company, San Min Book Company, and Nan-I Publishing Company).

For generating grammar distractors, we also consult the same grammar textbook and manually predefine distractor templates. These templates also need to ensure no ambiguous choices in the templates. Sometimes, not only one grammar pattern could be



correct answer in a sentence. For example, the stem in the second question in Figure 2, a distractor “develop” could be consistent with the syntax of the target sentence regardless of the global context. Thus, we referred to the grammar textbook and an expert for designing distractor templates for each grammar pattern (examples shown in Table 2).

Table 2 Distractor templates were referred by a grammar textbook and an expert in order to ensure the disambiguation of distractors.

level	function name	example answer	distractor 1	distractor 2	distractor 3
1	PerfectTense	has grown	have growing	have been grown	had grown
1	OnetheOther	one...the other	one...another	one...other	one...the others
2	TooAdjectiveTo	too happy to	too happy that	too happiest to	none of the above
2	soThat	so heavy	so heavier	so heaviest	none of the above
2	PastPerfectTense	had taken	had had taken	have taken	had been taken
3	prepVing	in helping	in being help	in helped	in being helping

4	GernudasObject	avoid taking	avoid to taking	avoid to take	avoid to took
5	Passive	is used	is using	used	will be using

6	RememberLike	remember to take	remembering to take	remember to tak- ing	none of the above
6	ModelAuxiliary	may have driven	may have driving	may has driven	may be driven

The second question in Figure 2 is a grammar question. The target testing purpose in the second question is “present perfect tense”, which is taught in the first grade. The original sentence is “Many of the original Halloween traditions have developed today into fun activities for children”. The parse structure of the original sentence is in Figure 3. The grammar pattern of this parse structure can be automatically identified by the Tregex patterns: /S.?! / < (VP < (/VB.?! / << have|has|haven't|hasn't)): /S.?! / < (VP < (VP < VBN)). When a grammar pattern is recognized (the green part of the parse tree, the difficulty degree of the grammar question is assigned based on the matched grammar pattern.

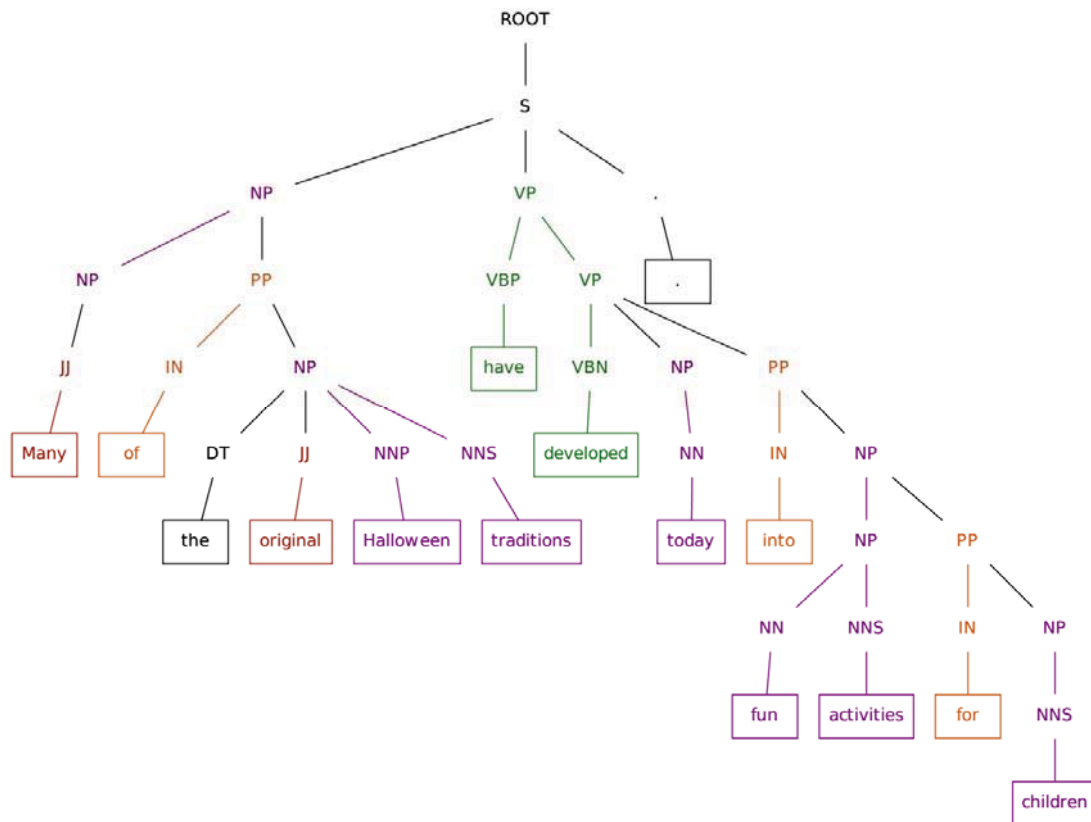
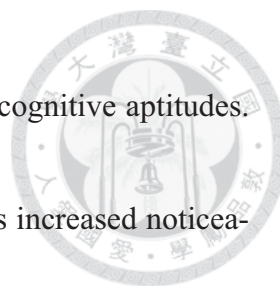


Figure 3 The parse structure of the sentence “*Many of the original Halloween traditions have developed today into fun activities for children*”.

3.3 Comprehension question generation


The difficulty of the reading comprehension questions is based on the reading level of the reading materials themselves. We assume that an examinee correctly answers a reading comprehension question because he/she could understand the whole story. The difficulty level of an article is correlated with the interaction between the



lexical, syntactic and semantic relations of the text and the reader's cognitive aptitudes.

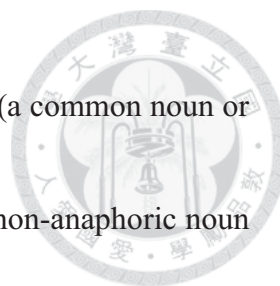
The reading level estimation of a given document in recent years has increased noticeably. Most past literature was designated for first language learners, but the learning timeline and processing between first language learners and second language learners is different. In this study, we adopt the measure of reading difficulty estimation [6] designed for English as second language learners to identify the difficulty of reading materials, as a difficulty measure for the reading comprehension questions.

Reading Comprehension relies on a highly complicated set of cognitive processes (Nation & Angell, 2006). In these processes, it is a key to make an anaphora resolution, construction-integration model and build a coherent knowledge representation (Kintsch 1998). Thus, in this work, we focus on a relation between sentences to generate two kinds of meaningful reading questions based on noun phrase coreference resolution. Similar to Mitkov and Ha (2003), who extracted nouns and noun phrases as important terminology in reading material, we also focus on the interaction of noun phrases as the test purpose. The purpose of noun phrase coreference resolution is to determine whether two expressions refer to the same entity in real life. An example is ex-



cerpted from Figure 2 (This tradition...on the devil₅). It is easy to see that *Jack*₂ means *man*₁ because of the semantic relationship between the sentences. The following *he*₃ and *he*₄ are more difficult to judge as referring to *Jack*₁ or *devil*₅ when examinees do not clearly understand the meaning of the context in the document. This information is used in this work to generate reading comprehension questions, in order to examine whether learners really understand the relationship between nouns in the given context.

There are two question types generated in the reading comprehension questions: an independent referential question for the single concept test purpose and an overall referential question for overall comprehension test purpose. When a noun phrase is selected as a target word in the stem question, it should have an anaphoric relation with the other noun phrase. In the first type, a noun phrase (a pronoun, a common noun or a proper noun) is selected as a target word in the stem question, a noun phrase (a common noun or a proper noun) will the same anaphoric relation will be chosen as the correct answer and other noun phrases (common nouns or proper nouns) will be determined as the distractors. In the second type, the same technique of the question generation applies to a sentence level. We regenerate new sentences as choices by replacing a noun (a



pronoun, a common noun or a proper noun) with an anaphoric noun (a common noun or a proper noun) as the correct answer and substituting a noun with a non-anaphoric noun as distractors.

The distractors should be satisfied with the following constraints:

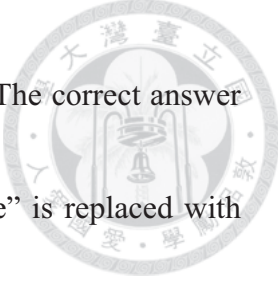
- Non-anaphoric relation: Distractors should have non-anaphoric relations. The anaphoric and non-anaphoric relations can be identified by the Stanford Coreference system (Raghunathan, Lee, Rangarajan, Chambers, Surdeanu, Jurafsky, and Manning 2010).
- Not pronoun: Pronoun is a replacement of a noun and a dependent on an antecedent (a common noun or a proper noun). Thus, distractors should be common nouns or proper nouns in order to have a clear test purpose.
- Number: Distractors should have the same number attributes (singular, plural or unknown) in order to make the sentence grammatically. For example, “devil” in the Figure 2 is singular; the number attribute of a distractor should be the same. If not, an unacceptable distractor (a plural noun or a collective noun) could violate the subject-verb agreement. The number attributes were given by



the Stanford Coreference system (Raghunathan, Lee, Rangarajan, Chambers, Surdeanu, Jurafsky, and Manning 2010), based on a dictionary, POS tags and Named Entity Recognizer (NER) tool.

- Gender: Distractors should have the same gender attributes (male, female, neutral or unknown) in order to make the sentence semantically. For example, “Jack” in the Figure 2 is male; the gender attribute of a distractor should be “male”, “neutral” rather than “female”; otherwise, students could answer the question directly instead of reading the passages. The gender attributes were assigned by the Stanford Coreference system (Raghunathan, Lee, Rangarajan, Chambers, Surdeanu, Jurafsky, and Manning 2010), which is from static lexicons.

The third question in Figure 2 independent referential question, which assesses one’s understanding of the concept of an entity involved in sentences. The word “he” in the original sentence “All he could ... a lantern” refers to “Jack”, the distractors “ghost”, “devil”, and “witch” have non-anaphoric relation, not pronouns, and are “singular” and “neutral”. The fourth question in Figure 2 the overall referential question,



which contains more than one concept that needs to be understood. The correct answer is from the sentence “He was so stingy ... died,” and the word “He” is replaced with “Jack” because they have referential relation. One of distractors is from “But he also could not ... devil,” the word “he” refers to “Jack” instead of “devil”. But we replace it with the non-anaphoric noun as a distractor. This approach further examines in the connection of concepts in the given learning material.

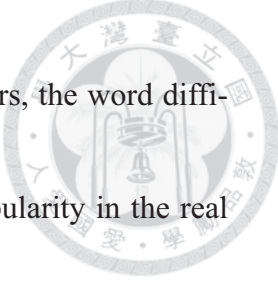
Chapter 4 Personalization



In this chapter, the personalized quiz strategy based on automatic quiz generation is presented. This personalized quiz strategy aims to achieve the following three purposes: first, we not only build a model to estimate reading difficulty, but also investigate the optimal combination of features for improving reading difficulty estimation. Next, an examinee's grade level is estimated by concerning the test responses and his or her historical data; in contrast, previous work only considered the current test responses. Finally, questions are selected with not only corresponding difficulties but also examinees' unclear concepts behind the previous incorrect responses. A student's previous mistakes are recorded and considered in advance in order to confirm whether he or she has learnt. Through the iterative practice, students' understanding will be enhanced by absorbing lots of different reading materials.

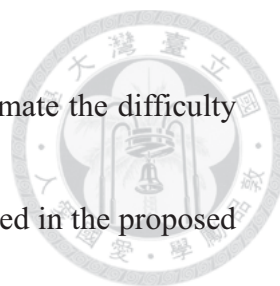
4.1 Reading difficulty estimation

As mentioned above, almost all past literature was designed for native readers, and this literature consulted word frequency from general corpora that were composed



of articles written for native readers. But for second language learners, the word difficulty depends on the structure of the material they study, not its popularity in the real world. In this section, we design a reading difficulty estimation for second language learners. We investigate the effectiveness of several meaningful lexical and grammatical features from early work, and then further consider organized grading indices of vocabulary from different sources, as well as grammar patterns collected from textbooks which represent words and grammar patterns that language learners have acquired at various grade levels, such as the word and grammar acquisition grade distributions. Furthermore, we also propose features that take into consideration word sense and coreference resolution.

Let D represents a document, while S represents the sentences in D . Suppose that D has n sentences, s_1, s_2, \dots, s_n , so that $D = \{s_1, s_2, \dots, s_n\}$. Let W be the set of words in D . Suppose D has m distinct words, w_1, w_2, \dots, w_m , so that a document $D = \{w_1, w_2, \dots, w_m\}$. We further suppose that the sentence S has k words, w_1, w_2, \dots, w_k , so that $S = \{w_1, w_2, \dots, w_k\}$, $m > k$. For a given training data set, the features are extracted and sent to a linear regression process to obtain a linear model that includes the weight of



each feature. The linear model is then applied to a document to estimate the difficulty level. In the following sections we explain and define the features used in the proposed estimation.

5.1.1 Baseline features

Word Number: A basic assumption is that a longer document is more difficult than a shorter one. Almost all prior work assumed that the number of words in a document accurately estimates reading difficulty (Flesch, 1948; Dale and Chall, 1948; Gunning, 1952; McLaughlin, 1969; Coleman and Liau, 1975; Kincaid et al., 1975). Pitler and Nenkova (2008) pointed out that this feature is significantly correlated with readability. For second language learners, we assume that a longer document takes more time to consume. Therefore, the number of words in a document is used in this study as one of the features to estimate reading difficulty. Word count is defined as follows:

$$word_number = \log|D| \quad (1)$$

Sentence length: Past studies have also taken sentence length into account, as-



suming that a shorter sentence is easier than a longer one (Flesch, 1948; Dale and Chall, 1948; Gunning, 1952; McLaughlin, 1969; Coleman and Liau, 1975; Kincaid et al., 1975). Thus for each document, we consider the average number of words per sentence as sentence length. The difficulty of a sentence is defined as follows:

$$sentence_length = \frac{word_number}{n} \quad (2)$$

Syllables: A syllable is a unit of organization for a sequence of speech sounds.

For example, the word *water* is composed of two syllables: *wa* and *ter*. Some related work has also taken syllables into consideration (Flesch, 1948; Gunning, 1952; McLaughlin, 1969; Kincaid et al., 1975). One notable example is the SMOG formula (McLaughlin, 1969), which estimates the reading difficulty of a document by only using the average number of polysyllables per sentence.

Even though syllables have proven to be a useful measure of reading difficulty for first-language users, similarities between sounds of a native speaker's mother tongue and their adopted second language can impact second-language learning. For instance, a word in an Asian language usually has one syllable, while a word in western languages usually has more than one syllable. When learning a second language, a sec-



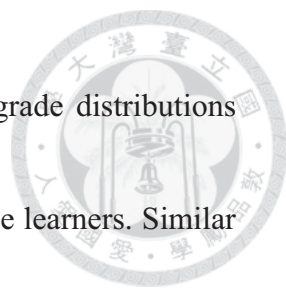
second-language learner could use similar-sounding syllables from their first language to learn vocabulary (called L1 – phonology effect hypothesis; Yamada, 2004). We assume the number of syllables in a word may affect the difficulties of documents. Thus, we find the average number of syllables of every word in a document to measure reading difficulty. The syllable difficulty of a document is defined as follows:

$$syllables = \frac{\sum_{i=0}^m word_syllables_i}{m} \quad (3)$$

where $word_syllables_i$ is the number of syllables within a word i .

5.1.2 The word acquisition grade distributions features


It is crucial to understand when a word is acquired by target readers. Kireyev and Landauer (2011) have tried using latent semantic analysis to capture word difficulty. Even though there is no existing dictionary presenting the word acquisition distribution in school grades, second language learners learn vocabulary in a limited range, which is usually decided by experts or teachers. Similar to Wan et al. (2010), we build two dictionaries from educational grading indices made by human experts for determining



the word grading. This helps better identify the word acquisition grade distributions resulting from random draws from the population of second language learners. Similar to Section 4.1, two resources, the General English Proficiency Test Reference Vocabulary and the Vocabulary Quotient, are used to estimate the word acquisition grade distributions in our study.

GEPT Word Lists: The General English Proficiency Test (GEPT; Wu and Liao, 2010) is designed to evaluate student proficiency in English as a second language. It provides a reference vocabulary list with about 8,000 words divided into three word levels: elementary (**gept1**), intermediate (**gept2**) and high-intermediate (**gept3**). Some words not found in the GEPT word list are attributed to the out of GEPT word list (**gept0**). For each word from a document, we identify its vocabulary difficulty by searching for the word's level from the GEPT word lists, counting the number of distinct words in each level, and finally normalizing by the total number of distinct words in each level.

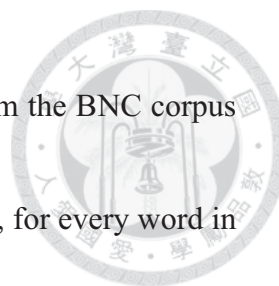
Age of Word Acquisition: In addition to the GEPT, we also collected a word list from an organization, Vocabulary Quotient (VQ; Ho and Huong, 2011). This organiza-



tion collected more than 10,000 words and labeled them in reference to other educational institutions, such as the Elementary School Reference Vocabulary and the Junior High School English Reference Vocabulary texts made by the Ministry of Education of Taiwan, and the High School English Reference vocabulary text made by the College Entrance Examination Center of Taiwan. The word list is divided into fourteen levels (**vk3—vk16**), which represent the words learned by second language learners from elementary school to college. Just as with the GEPT list, some words are still absent from the Vocabulary Quotient word list; these words are attributed to out of vocabulary list (**vk0**). For each word from a document, we identify its difficulty by first referencing its difficulty level from within those word lists, and after counting the number of distinct words in each level, normalizing by the total number of distinct words in each level.

5.1.3 Frequency features

Besides the word acquisition grade distributions features, word frequency is another approach to estimating word difficulty. Word frequency is based on the assumption that more frequent words are easier to identify. In order to compare the feature, the



word acquisition grade distributions, we find its word frequency from the BNC corpus and also use a Google search result count as an alternative frequency, for every word in a document.

Word Frequency in BNC Corpus: The British National Corpus (BNC; Lou and Guy, 1998) is a 100 million word collection of written and spoken language from a wide range of sources, designed to represent a wide cross-section of British English from the later 20th century. For each word in a document, we calculate the distinct word frequency (wf) that refers to the times it appears in the BNC corpus. Word frequency is defined as follows:

$$wf_i = \frac{n_i}{|d_j|} \quad (4)$$

where n_i is the number of occurrences of the considered distinct word w_i in document d_j , and the denominator is the sum of the number of occurrences of all distinct words in document d_j , that is, the size of the document $|d_j|$. For each word in a given document, we also calculate the average number of log word frequency. The document's difficulty value based on word frequency in the BNC corpus is defined as follows:

$$bnc_frequency = \log \frac{\sum_{i=0}^m wf_i}{m} \quad (5)$$



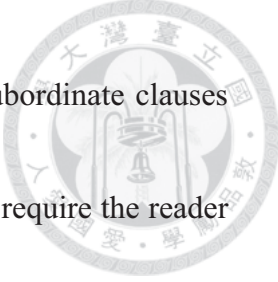
Google Search Result Count: For a given query, Google will return a list of documents containing the queried words and a search result count. We use the search result count as a measure of word frequency, like the word frequency from a corpus. For each word in a given document, we also calculate the average number of log word frequency. The document's difficulty value based on word frequency from Google is defined as follows:

$$google_search_count = \log \frac{\sum_{i=0}^m google_i}{m} \quad (6)$$

where $google_i$ is the search result count of a word i from Google.

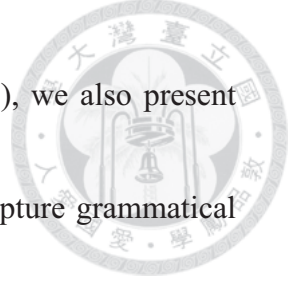
5.1.4 Parse features

Syntactic constructions affect the understanding of a sentence. This assumes that the more complicated a sentence, the greater its difficulty. Schwarm and Ostendorf (2005) proposed four syntactic features for their measure of reading difficulty: the average parse tree height, the average number of noun phrases per sentence, the average



number of verb phrases per sentence, and the average number of subordinate clauses per sentence (SBAR). Because sentences with multiple noun phrases require the reader to remember more entities, Barzilay and Lapata (2008) found that documents written for adults tended to contain more noun phrases than those written for children. In addition, while including more verb phrases in each sentence increases sentence complexity, adults might prefer to have related clauses explicitly grouped together. Pitler and Nenkova (2008) have also found a strong correlation between readability and the number of verb phrases. These works show that the more complicated the parse features in a document, the more likely it was written for adults. Hence, we also examine the influence of parse features for second language learners.

Prepositions are a class of words that indicate relationships between nouns, pronouns and other words in a sentence. Prepositions can be divided into two kinds: simple prepositions and compound prepositions. Simple prepositions are single word prepositions, while compound prepositions are more than one word. We assume that more prepositional phrases in a sentence also increase its complexity, and second language learners might be confused by complex prepositional phrases. Thus, in addition



to the parsing features proposed by Schwarm and Ostendorf (2005), we also present the average number of prepositional phrases as a new feature to capture grammatical complexity.

Thus, from the outline above, for a document we consider the following syntactic features from parse results generated by a Stanford parser (Klein and Manning, 2003): the average parse tree height, the average number of noun phrases, the average number of verb phrases, the average number of SBAR and the average number of prepositional phrases.

Average Parse Tree Height: Suppose the height of a parse tree of a sentence is h .

The average parse tree height difficulty of a document is defined as follows:

$$tree_height = \frac{\sum_{i=0}^n h_i}{n} \quad (7)$$

Average Number of Noun Phrases: Suppose a sentence has np_i noun phrases.

The average noun phrase difficulty of a document is defined as follows:

$$np = \frac{\sum_{i=0}^n np_i}{n} \quad (8)$$



Average Number of Verb Phrases: Suppose a sentence has vp_i verb phrases. The average verb phrase difficulty of a document is defined as follows:

$$vp = \frac{\sum_{i=0}^n vp_i}{n} \quad (9)$$

Average Number of SBAR: Subsidiary conjunctions (SBAR), for example, *because*, *unless*, *even though*, and *until*, are placed at the beginning of a subordinate clause that links the subordinate clause and the dominant clause. SBAR is an indicator to measure sentence complexity. The SBAR difficulty of a document is defined as follows:

$$sbar = \frac{\sum_{i=0}^n sbar_i}{n} \quad (10)$$

Average Number of Prepositional Phrases: Suppose a sentence has pp_i prepositional phrases. The average number of the prepositional difficulty of a document is defined as follows:

$$pp = \frac{\sum_{i=0}^n pp_i}{n} \quad (11)$$

5.1.5 The grammar acquisition grade distributions features



In Heilman et al. (2007), they found that grammatical features played an important role in reading difficulty estimation for second-language learners. A model with complex syntactic grammatical feature sets achieved more accurate results than simpler models. In their work, they examined the ratio of grammatical occurrence per 100 words: both the passive voice and past participle had obvious differences between the lowest and highest levels in the second-language corpus. Thus, we measure grammatical difficulty as a linguistic processing factor in estimating reading difficulty for second language learners.

Grading Index of Grammar (grammar1—grammar6): To decide the grammatical difficulty level of a document, the same method described in Section 4.2. We first collected sentences from the six versions of second-language textbooks and parsed the sentences to find their grammar patterns, for a total of 44 grammar patterns. Manually identifying these grammar patterns allows the parse tool to then automatically find these same patterns within a given document. Next, using this parse tree structure

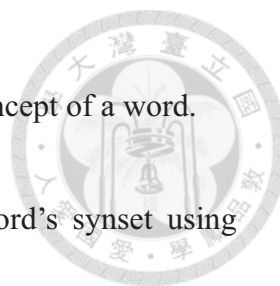
searching tool (Levy and Andrew, 2006), the grammatical structures were assigned to the textbook grade in which they frequently appear.



5.1.6 Semantic features

For any given word, its meaning may vary broadly depending on the context. For example, the word “*bank*” has two distinct meanings (also called two senses), “*financial institution*” and “*sloping mound*”, not to mention its other colloquial uses. For both the word acquisition grade distributions and frequency features, we assume that a word only has one sense, because this still results in accurate performance with many language technologies, such as information retrieval or text classification. However, it cannot be claimed that a second language learner having learned a word knows every sense of the word. Therefore, we designed semantic features to identify word senses in a document.

Average Number of WordNet Synsets: We adopted WordNet (Miller et al., 1990) as a resource for understanding the senses in a word. WordNet is a large lexical database of English. The database contains 155,287 words, with each word annotated with a set of senses. The average noun has 1.23 senses and the average verb has 2.16 senses.

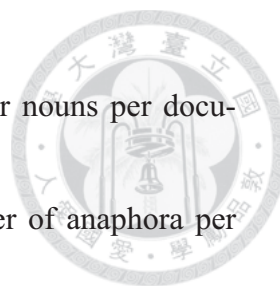


A set of near-synonyms is defined as a synset, which represents a concept of a word.

For each word in a document, we total the number of a word's synset using WordNet. To determine the representation of this feature, we develop seven categories (**wordnet1**—**wordnet7**) to represent the number of synsets of each word in a document. Here, suppose a word has ws_i synsets. The number is normalized as two square roots and then rounded down to an integer as a feature index. For example, if the number of synsets of a word is 17, it is attributed to **wordnet4**. If the number of synsets of a word is greater than 49, it is assigned to **wordnet7**. Finally, we count the number of distinct words in each WordNet category and normalize by the total number of distinct words.

5.1.7 Relation features

Coreference is a grammatical relation that presents two referring expressions that refer to the same entity. This entity is called an antecedent, and the referring expression is called an anaphora. We assume that coreference represents the implicit relations between sentences. When second language learners recognize the coreferent relation well, they might be able to understand the reading material more clearly. For a document, we



count the number of pronouns per document, the number of proper nouns per document, the number of antecedents per document, the average number of anaphora per coreference chain and the average distance between anaphora and antecedents per chains.

Average Number of Pronouns: We assume that the greater the number of pronouns in a document, the more entities the reader needs to remember, and this increases reading difficulty. Thus, we total the average number of pronoun in a document.

Average Number of Proper Nouns: If a sentence contains more than one proper noun, a reader must remember more objects in a document. Barzilay and Lapata (2008) found that documents written for adults tended to contain more entities than those written for children. Hence, we count the average number of proper nouns in a document.

The Number of Antecedents per Document: Antecedents represent real entities mentioned in the document. Similar to the average number of proper nouns, we assume that if a document contains less entity, the document is easier to read. We total the number of antecedents as the number of entities to capture this idea.



The Average Number of Anaphora per Coreference Chain (corefer_chain):

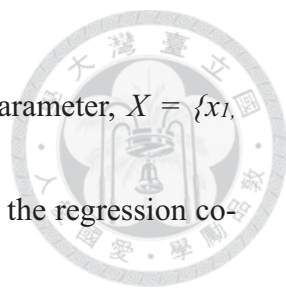
We assume that with more anaphora per coreference chain, second language learners need more knowledge to resolve them; consequently, we count the average number of anaphora per chain.

The Average Distance between Anaphora and Antecedents per Chain (co-refer_distance): This captures the distance between antecedents and anaphora. We assume if an antecedent and anaphora are in the same sentence, the sentence will be easy to understand. In contrast, if they are several sentences apart, it is probable that the document is more complex to read.

5.1.8 Regression model

Linear regression is an approach to modeling the relationship between a scalar variable Y and variables denoted X . A prediction of a given document is the inner product of a vector of feature values for the document and a vector of regression coefficients estimated from the training data.

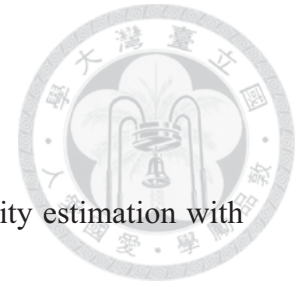
$$Y = \alpha + \sum_{i=1}^n \beta_i X_i + \varepsilon, \quad i = 1, 2, \dots, n \quad (12)$$



where Y is the difficulty value of a document, α is the intercept parameter, $X = \{x_1, x_2, \dots, x_n\}$ represents the feature values, $\beta = \{\beta_1, \beta_2, \dots, \beta_n\}$ refers to the regression coefficient for each feature value i , and lastly ε is an unobserved random variable that represents noise in the linear relationship between the dependent variable and regressors.

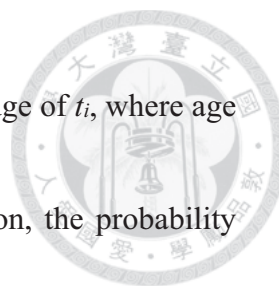
The primary reason for adopting linear regression for a reading difficulty model is that the output scores are continuous and related with each other, whereas the outputs of other methods, such as classification, are discrete and unrelated between levels. Kate et al. (2010) evaluated the performance of reading difficulty among several machine learning methods and reported that all of the regression family outperformed the baseline. In our study, the readability of a text is represented by a score or a class, which is typically indicated in terms of school grades. Overall, the content difficulty of textbooks increases incrementally. Thus, we opt for linear regression as our model, as we assert that our estimated results are correlated.

5.2 Ability estimation



In this section, we propose an interpretable and statistical ability estimation with inherent randomness in the acquisition process, specifically in the Web-based learning environment. This model draws a connection between students' abilities and the acquisition grade distributions. For a student who is said to be a grade level six in this work, our method is able to estimate how much the student has acquired as a certain percentage of the knowledge in a population when he correctly answers a certain percentage of items on a test.

We propose the following interpretation of the quantitative definition: an examinee is said to have ability θ if s percent of items in a test $T = (t_1, \dots, t_m)$ have been correctly answered each by r percent of the population. We first consider that each item t_i in a test T has been correctly answered by r percent of the population. In general, there is a specific knowledge behind each tested item t_i . The difficulty level of the specific knowledge represents the age at which most people have acquired knowledge of t_i . Most people understand some knowledge at an early age, whereas some understand this knowledge later in life. Here, we precisely denote the level the specific knowledge



by the age at which r percent of the population has acquired knowledge of t_i , where age refer to school grades. When given a knowledge t_i and a population, the probability distribution of grade acquisition $p_t(\theta)$ can be calculated. Let the quantile function q_t of the cumulative distribution function correspond to the grade acquisition distribution p_t . In other words, $q_t(r)$ represents the grade at which r percent of the population has acquired knowledge of t . This assumes a normal distribution,

$$q_t(r) = \mu_t + \Phi^{-1}(r)\sigma_t \quad (13)$$

where μ_t and σ_t represent the mean and standard deviation of the distribution p_t , and $\Phi^{-1}(r)$ is a quantile function representing the probability of exactly r to fall inside the interval of the distribution. When an examinee correctly responds to the item t_i , the examinee's ability is regarded as a school grade. To investigate the distribution of the grade level of a test T , we collect the grade level values generated from each quantile function $q_t(r)$ as the distribution of knowledge acquisition within a single test f_Q .

In practice, this is time consuming and costly to find the distribution p_t for each item t_i known in advance. Fortunately, under Item Response Theory (Embretson & Reise, 2000), a response of an examinee to an item is modeled by a mathematical item



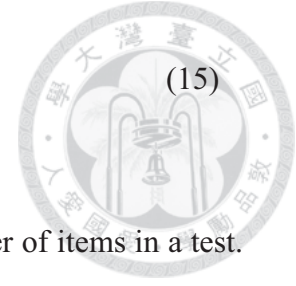
response function, known as the item characteristic curve. The item characteristic curve is a mathematical family model that describes the probability of a correct response between an examinee's ability and the item parameters. These models employ one or more parameters, such as an item difficulty parameter and an item discrimination parameter, to define a particular cumulative form. When given the item parameters, the grade level at which r percent of the population correctly responds to item t can be inferred. Take one-parameter logistic model as an example,

$$q_t(r) = \ln\left(\frac{r}{1-r}\right) + b \quad (14)$$

where variable b as item difficulty.

Estimating an examinee's ability through a test relies on the test responses of the test. We consider a percentage of correct responses in a test as variable s and define the s th quantile of the distribution of knowledge acquisition in a test f_Q as the examinee's ability. The distribution of the s th quantile of f_Q , where s percent of items in a test have been correctly answered by r percent of the population, can be performed using a standard formula for normal approximation of order statistics (David & Nagaraja, 2003):

$$q_T(r,s) \sim N(F_Q^{-1}(s), \frac{s(1-s)}{m[f_Q(F_Q^{-1}(s))]^2}) \quad (15)$$



where F_Q is the cumulative distribution function and m is the number of items in a test.

This result is more certain of the estimated grade level assigned to a large sample item size. In cases where an examinee correctly answered all items or no item, a smooth constant c is used ($c=0.01$ in this study).

When given an examinee's responses in a test, the current examinee's ability θ_t can be described by the distribution (3) in which r percent of the population correctly answer s percent of items. We also consider an examinee's history record, and employ Exponential Moving Average (EMA; Brown, 2004) to combine this history with the current ability, transformed by the following formula:

$$ability_t = \alpha \times \theta_t + (1 - \alpha) \times ability_{t-1} \quad (16)$$

where θ_t is the current ability in time t obtained from the mean of the equation (3), $ability_{t-1}$ is the past estimated ability in the time $t-1$ as history records, and $ability_t$ is the final estimated ability in time t after the combination of the current ability and the past estimated ability with EMA. Additionally, $\alpha = 2/(n+1)$ is a smoothing constant repre-

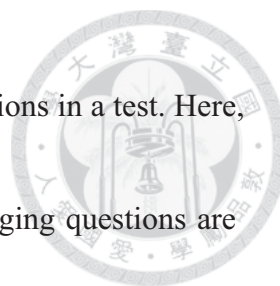
sented as an exponential weight, and n represents the period as the length of the moving window.



5.3 Quiz Selection


The purpose of personalized computer-aided question generation is to help students improve their own learning progress and correct their previous mistakes. Thus, this section presents the quiz strategy to select questions.

When given a learner's ability θ , it is critical to determine how to best form a test from a series of questions which match learner's ability. In (Barla et al., 2010), the researchers selected history-based questions consisting of recently used questions and correctly answered questions. Similarly, in this study, a test is composed of not only fit questions (a question's difficulty level is equal to a learner's grade level) and history-based questions (a question's difficulty level is easier than a learner's grade level) but also challenging questions (a question's difficulty level is more difficult than a learner's grade level). The purpose of this test is not only to measure student's proficiency but also to review the previous relevant knowledge and stimulate the future lessons.



Like Barla et al. (2010), we define probability values to assign questions in a test. Here, the percentage of history-based questions, fit questions and challenging questions are 20%, 60% and 20%, respectively.

When questions are incorrectly answered, they are stored in the system. Incorrectly answered vocabulary answers are represented as the concept of the items, and incorrectly answered grammar patterns are the concept of the items. Figure 4 presents the incorrectly answered concepts of each student in the database table of the implemented system. During the next iteration of the test, if there is any similar question based on the same concept, that question will be selected first. The goal of this design is to enhance learners' understanding and improve their proficiency. When students read other reading materials, the system generates similar questions based on the same concepts.



username	type	target	times
candy401888@yahoo.com.tw	vocabulary	sending	1
candy401888@yahoo.com.tw	vocabulary	battle	1
candy401888@yahoo.com.tw	vocabulary	study	1
candy401888@yahoo.com.tw	vocabulary	possible	1
candy401888@yahoo.com.tw	vocabulary	heavy	1
candy401888@yahoo.com.tw	vocabulary	family	2
candy401888@yahoo.com.tw	vocabulary	drink	1
candy401888@yahoo.com.tw	grammar	infinitive	1
candy401888@yahoo.com.tw	grammar	Passive	1

Figure 4 A table of a database in the implemented system captures the incorrectly answered concepts of a student.

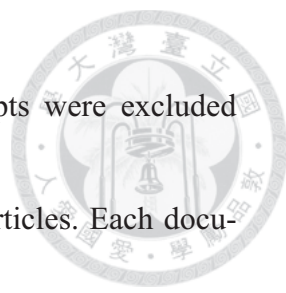
Chapter 5 Evaluation on reading difficulty estimation



In this section, the proposed features and model of the proposed reading difficulty estimation are evaluated. To determine how each feature contributes to an accurate readability judgment, we first conducted an experiment to test the performance of each feature and feature categories using linear regression algorithms. Next, an optimal feature set is determined by model selection, in order to investigate how to best combine the features that improve reading difficulty estimation. Finally, the proposed estimation was also modeled as a multiclass classification and compared to other related work. These experiments are described in the following subsections.

5.1 Data set

Our experiment used data from senior high school English textbooks designed for Chinese students in Taiwan to learn English as a foreign language. We gathered 342 documents from five different publishers (including The National Institute for Compilation and Translation, Far East Book Company, Lungteng Cultural Company, San Min



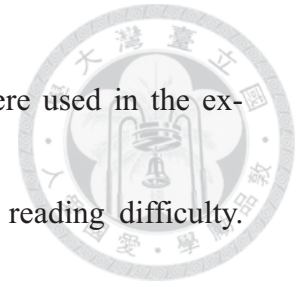
Book Company, and Nan-I Publishing Company). Poets and scripts were excluded from the data set because their formats are different from normal articles. Each document was graded using six levels, ranging from one to six. These levels indicate the semester grade levels of senior high school students, and were used as the gold standard in this work.

During data preprocessing, a majority of words were adopted directly and stop words were also used. Even though some words are derivative of the same root, they are acquired and used in different ages. For example, in our scenario, the word *promise* is taught in the second semester, while the word *promising* is in the fourth semester. This suggests that the different forms of words are represented as different meanings. Thus, initially, words were used directly in our study; otherwise, some words were lemmatized when necessary. For the same reasons, stop words are retained in the proposed estimation except for frequency features.

5.2 Metrics

In the evaluation of features and the optimal model selection, the root mean

squared error (RMSE) and Pearson's correlation coefficient (r) were used in the experiments in order to evaluate the effectiveness of the estimated reading difficulty.



RMSE measures the averaged erroneous value between ground truths and estimated responses. It is an averaged distance for measuring how far estimated responses approach ground truths; the lower the RMSE, the better the estimation. The Pearson's correlation coefficient (r) measures the trends between the ground truth and the generated results. It represents the strength of the linear relationship between two random variables. A high correlation shows that simple documents are estimated as having a low difficulty value, while difficult documents are predicted as having a high difficulty value.

In the evaluation of reading difficulty estimation as classification, not only the RMSE and correlation coefficient described in the previous section, but also accuracy and trend accuracy in direction (TAD) are adopted as measurements in the evaluation. Accuracy is defined as the proportion of correctness of the generated results comparing with the ground truth. The TAD is used in the performance of trend forecasting (Zhang et al. 2005). The result can be interpreted as the proportion of the same direction be-



tween the estimated results and the gold truth. To employ the measurement, the TAD was modified as:

$$TAD = \frac{\sum_i^n \sum_j^n D(i, j)}{n \times (n-1)} \times 100\% \quad , \quad D(i, j) = \begin{cases} 1, & \text{where } (y_i - y_j)(g_i - g_j) > 0 \\ 0, & \text{otherwise} \end{cases} \quad (17)$$

where y_i is the estimated level, g_i is the gold truth, and n is sample size.

5.3 Evaluation of the features

Five-fold cross-validation was employed. The data was first split into five sets. One set was used as held-out data to predict the reading difficulties of documents, while the rest of the data was used as training data to build a regression model. Each fold was further repeated five times by changing the pairing of training and testing.


To understand the impact of features and feature categories, we reported the RMSE scores and the correlations among different feature categories. Table 3 summarizes RMSE and the correlations among different feature categories. The rows of a block represent the different feature categories in Section 5.1. The baseline features have the best results, leading to an RMSE of 0.98 and correlation of 0.82. This was

composed of word number, sentence length, and syllables, and indicates that they represent how a majority of experts design learning materials.



Table 3 Results of RMSE and correlation among different feature categories.


Categories	Features	RMSE	r
Baseline	baseline-only	0.98	0.82
Word (AOA)	gept-only	1.24	0.68
Word (AOA)	vq-only	1.25	0.67
Relation	coreference-only	1.48	0.48
Parse	parse-only	1.50	0.46
Grammar (AOA)	grammar-only	1.61	0.31
Semantic	wordnet-only	1.60	0.33
Frequency	bnc_frequency	2.22	0.12
Frequency	google_search_count	4.50	-0.04



The second and third feature categories were GEPT and VQ, both of which are the word acquisition grade distributions features. They represent when non-native readers learn words at different ages. It is highly likely that the word acquisition grade distributions is also an important factor in analyzing reading difficulty for non-native readers. Even though GEPT and VQ features are derived from different resources and classified by fine-coarse grade respectively, both of their performances were high correlated with ground truth and the RMSE values were less or equal to 1.25. Surprisingly, the GEPT word list, only divided into three levels, performs better than VQ, which categorizes words acquired by non-native readers from elementary schools to universities.

The fourth feature category was coreference features. The coreference features captured the inter-relationship between noun phrases. While baseline, GEPT, and VQ features only took explicit words into consideration, coreference features considered not only noun phrase types but also implicit interaction between noun phrases. Feng, Jansche, Huenerfauth, and Elhadad (2010) first proposed coreference and discourse features. In their investigation, noun phrase features and coreference inference features

slightly improved. The result was fairly consistent in the second-language materials.



The fifth and sixth features were parse and grammar features. Both features analyze sentence structures in a document. While parse features were automatically extracted from a parser, grammar features were identified from a tree structure search tool based on manual grammatical patterns. The results of our grammar features were consistent with Heilman et al. (2007). In their study, vocabulary-based features produced more accurate results than grammar-based features alone, but complex grammar features performed better than simple ones. Even though we collected more than forty grammatical patterns from six grades in textbooks (more than the Heilman method), the results indicated that parse features were slightly better than grammar features. It is possible that parse features could be more robust than grammar features.

The next features are semantic features. Unfortunately, calculating the number of senses for each word seemed to have little impact on reading difficulty estimation. No matter how many senses a word has, the most important factor is whether the readers understand the specific sense of words in a document or not. The better solution might first determine each sense of word in a document, and then assess when reader had

learned those meanings. This may represent a new research problem for future studies.



The last remaining features were frequency features derived from the BNC corpus and Google search engine. Surprisingly, frequency features were not good indicators for estimating reading difficulty. This went against Tanaka-Ishii, Tezuka, Terada (2010), who used the log frequency obtained from corpora as features to predict document reading difficulty. One explanation for this is that the format of features and the method in (Tanaka-Ishii et al., 2010) may be very different from this study. Another explanation is the possibility that lower and higher word frequencies are counteracted by the summarization of word frequencies.

The results indicated that the lexical-based feature categories (the baseline and the word acquisition grade distributions features) produced more accurate results than grammar-based feature categories (coreference, parse and grammar features) alone.

5.4 Optimal model selection



To investigate how combining features improves reading difficulty estimation, the forward selection were used to select the best subset of features for linear regression, and Bayesian information criterion (BIC; Schwarz, 1978) were applied to decide the best regression model. If a regression model employs every available feature (47 in total), it becomes sensitive to training data. In contrast, if a model was not designed well, its performance with the testing data should be poor. This section examines how this study identified an appropriate model with features that play important roles in determining reading difficulty.

The forward selection was employed to evaluate the optimal model; it starts with the intercept and adds at each step the features that most improve. The detailed rules for this process are as follows:

Step 1. The first feature with the highest Pearson Correlation Coefficient value was selected to the best model.

Step 2. The next selected feature had the highest semi-partial correlation, and is added into the model.



Step 3. After adding the new feature in step 2, the squared multiple correlation coefficient of new the model was calculated (R^2).

Step 4. To test whether the new feature contributes significantly to the model, the difference between the new R^2 value and old R^2 value was evaluated using Analysis of Variance (ANOVA).

Step 5. If the incremental difference in Step 4 significantly improved, the new feature stayed in the model; otherwise it was removed.

This process from step 2 to step 5 is repeated until the addition of further features produces no significant improvement.

Bayesian information criterion was used to select the best model based on estimating the Kullback-Leibler divergence between a true model and a proposed model, incorporating sample size. This process also introduces a penalty term for the number of parameters in a model. BIC is denoted as:

$$BIC = n \times \ln\left(\frac{RSS}{n}\right) + \ln(n) \times k \quad (18)$$

where RSS is the residual sum of squares from the regression model, k denotes the number of model parameters, and n is the sample size. Information criteria tend to pe-

nalize complex models, giving preference to simpler models in selection.



Table 4 summarizes the top performance of each selective model and the full model. As shown in the second row of the table, the number of words in a document was the first feature with the highest validity, and thus this feature was involved in the first model. The remaining features are added in turn to the model, according to their significant individual contributions, as described in the second column of


Table 4. As shown, when `gept1` was added to the model, the results greatly improved; the RMSE of the second model dropped to 0.96 and the correlation rose to



0.82. This represents a positive contribution for the word acquisition grade distributions. Thus, these results imply that when given a document, the estimated level can be accurately calculated by using the number of words and the proportion of the `gept1` word list combined into the regression model.

Table 4 Results of the optimal model selection.

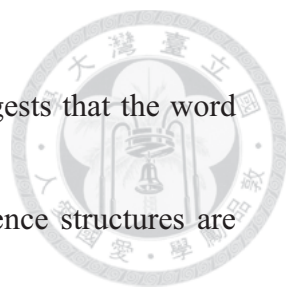
Model	Added Feature	RMSE	r	BIC	RSS
1	<code>word_number</code>	1.25	0.67	157.50	532.87
2	<code>gept1</code>	0.96	0.82	-20.19	311.58
3	<code>tree_height</code>	0.87	0.86	-87.77	251.39
4	<code>vq12</code>	0.85	0.86	-99.52	238.79
5	<code>vq13</code>	0.84	0.87	-106.52	229.99
6	<code>proper_noun</code>	0.84	0.87	-104.46	227.47



7	vq15	0.84	0.87	-101.14	225.80
8	vq5	0.85	0.86	-97.86	224.12
9	antecedent	0.85	0.86	-94.88	222.26
10	vq11	0.85	0.86	-91.40	220.74
	all	1.51	0.64	121.91	208.15


Based on BIC values, the best model was a combination of the following features: the number of words, gept1, tree height, vq12 and vq13. Our results show that the fifth model in

Table 4 had the least difference between the gold truth and the estimated levels: the RMSE is as small as 0.84 and the correlation turned out to be closer to 1 at 0.87. With the other features added, the performance remained steady until the seventh mod-



el. After that, the performance began to decrease. This finding suggests that the word acquisition grade distributions and the average complexity of sentence structures are important factors in reading difficulty and should be taken into consideration. Except for the average tree height, the *gept1* word list refers to words that users have already learned, while the *vq12* and *vq13* word lists contain vocabulary that are currently being acquired, corresponding to the specific readers' ability. From these results, we conclude that for non-native readers, previously learned vocabulary, current new vocabulary, and the complexity of sentence structure lead to a successful reading difficulty estimation.

To better compare these potential models, the performance of selected models is presented in Figure 5. The upper half of the figure illustrates the RMSE among the models with increased feature numbers, while the lower half of the figure shows the correlation between the models and the ground truth. Initially, the RMSE value decreases and correlation sharply rises as features are added. After identifying the most accurate model, the performance of both measures levels off. This can be seen as a great advantage of model selection, since a small number of identified features



achieves a satisfying outcome. Until the Google search result variable was added to the model (the 29th model), the RMSE rapidly increased and the correlation significantly declined. This implies that frequency in a large corpus, such as Google, might not be as useful as the word acquisition grade distributions in reading difficulty estimation. These results reinforce the assumptions of previous studies (Huang, Chang, Sun and Chen, 2011), where the word acquisition grade distributions has more relative importance than frequency within corpora. After the 29th model, the performance fluctuated and worsened, compared to previous models in both measurements. This indicates that performance becomes unstable if the model over-fits. This error may be due to the fact that these models capture idiosyncrasies of the training data rather than generalities.

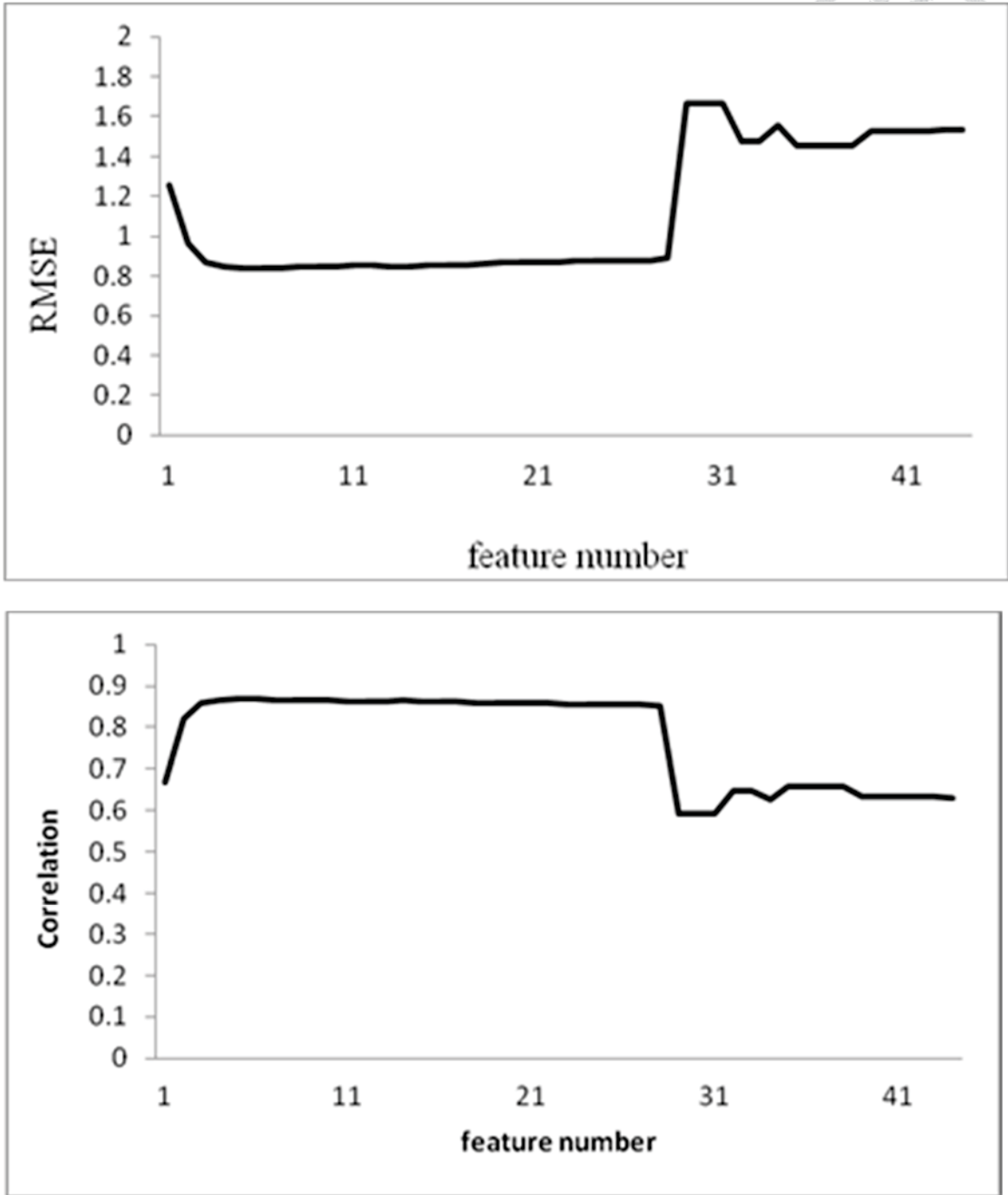
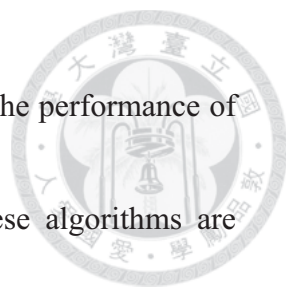


Figure 5 the performance of a selected model.



To understand the impact of feature sets, we also investigated the performance of several regression algorithms with two different feature sets. These algorithms are available in the WEKA package (Bouckaert, Frank, Hall, Holmes, Pfahringer, Reutemann, ... and Sonnenburg, 2010), including Support Vector Regression (SVR; EL-Manzalawy and Honavar, 2005), Sequential Minimal Optimization for Regression (SMOreg; Shevade, Keerthi, Bhattacharyya and Murthy, 2000), Pace Regression (Wang and Witten, 2002), and linear regression. All parameters were used as a default setup. Adopting regression as a reading difficulty model assumes that the output scores are continuous and related with each other. Table 5 shows the results of these regressions. All methods with only the optimal features outperformed those with all features. This indicates that the optimal feature set could help regression estimates. In addition, the results of the linear regression outperformed those of SVR. This finding is in contrast with Kate et al. (2010), which noted similar performance among regression algorithms. SMOreg, which is a SVR improved by Sequential Minimal Optimization, had the best performance among the models with all features; however, linear regression with optimal features matched the results of SMOreg and Pace Regression. This sup-

ports our contribution; when the optimal feature set is identified, the performance among various regression techniques is similar.

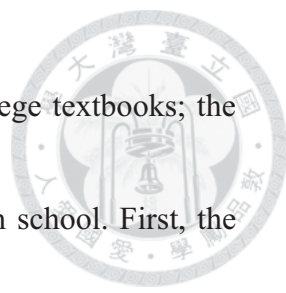


Table 5 Results of the optimal model selection.

Method	All Features		Optimal Features	
	RMSE	r	RMSE	r
Support Vector Regression (nu-SVR)	1.66	0.22	1.24	0.60
Support Vector Regression (epsilon-SVR)	1.65	0.25	1.42	0.59
SMO for Regression (SMOreg)	1.25	0.73	0.84	0.87
Pace Regression	1.95	0.52	0.84	0.87
Linear Regression (Proposed)	1.51	0.64	0.84	0.87

6.5 Reading difficulty estimation as classification

The proposed model can also be modeled as a multiclass classification. The labels were determined by eight levels ranging from zero to seven: zero represents the documents under the specific readers' ability such as elementary textbooks; seven repre-



sents the documents above the specific readers' ability such as college textbooks; the remaining levels are the same as the semester grades of senior high school. First, the thresholds were found and the estimated levels were assigned to the closest level corresponding to the threshold. The minimum threshold was assigned the minimum value from the training dataset; likewise, the maximum threshold was selected based on the maximum value from the training dataset.

To understand the performance of the proposed estimation compared with other studies, our experiment also compared the estimated levels within the Flesch Reading Ease (Flesch 1948), Flesch–Kincaid Grade Level (Kincaid et al. 1975), Coleman-Liau (Coleman and Liau 1975), Lexile (Stenner 1996), and the Heilman method (Heilman et al. 2007). The Flesch–Kincaid Grade Level, Flesch Reading Ease and Coleman-Liau were duplicated, while Lexile and the Heilman method are available online. All of these methods are designed for native readers. In the training phase, the output score generated from each document by those estimations is found, like the procedure of the proposed method, as well as the threshold between each level of the other estimations. During the testing phase, the estimated levels of testing documents were determined

using these thresholds.



For accuracy and RMSE, we expect that the proposed estimation will obviously produce a more accurate reading difficulty prediction than other estimations. Through TAD, we expect that the proposed estimation will be consistent with the ground truth, although it might tend to predict easy documents with a lower grade and difficult documents with higher grades. In the correlation coefficient, we expect that the proposed estimation will report a particularly high correlation than other estimations. This may suggest that the relationship between the proposed estimation and the ground truth is stronger than others. In summary, existing difficulty estimation methods will perform poorly for second language learners, which may due to the different and insufficient features used.

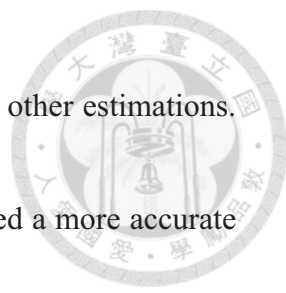



Table 6 shows the results between the proposed estimation and other estimations. For accuracy and RMSE, the proposed estimation obviously produced a more accurate reading difficulty prediction than other estimations. When the proposed estimation fails to predict the correct reading difficulty, its error ranges are almost within one grade; by comparison, the error ranges of Flesch–Kincaid Grade Level, the Lexile and the Heilman method were between one to two grades, and Flesch Reading Ease and Coleman–Liau had an even wider error range. Through TAD, the proposed estimation was consistent with the ground truth, although it might tend to predict easy documents with a lower grade and difficult documents with higher grades. In contrast, the results of the other method are fluctuant. In the correlation coefficient, all estimations are positively correlated. The proposed estimation reported a particularly high correlation at 0.87, whereas the other estimations were at <0.5 . This suggests that the relationship between the proposed estimation and the ground truth is stronger than others. In summary, existing difficulty estimation methods perform poorly for non-native readers, which may be due to the different and insufficient features used.

Table 6 Comparison between the estimations.



Estimations	RMSE	<i>r</i>	Accuracy	TAD
Flesch Reading Ease	2.17	0.27	0.28	0.40
Flesch–Kincaid Grade Level	1.85	0.48	0.26	0.49
Coleman–Liau	2.16	0.31	0.24	0.41
Heilmen	1.84	0.41	0.26	0.43
Lexile	1.76	0.46	0.33	0.49
Model 5 = word_number+gept1+tree_height+vq12+vq13	1.01	0.87	0.42	0.68

Chapter 6 Simulation on ability estimation



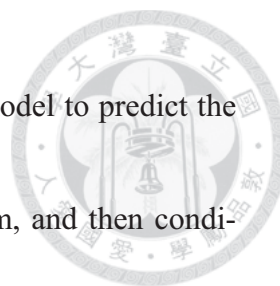
In this chapter, the proposed ability estimation is evaluated by a simulation study.

To investigate the characteristics of the proposed method, we first analyze the convergence speed and the error distance between the ground truth and the estimated ability.

Next, an example, which presents the benefits of taking historical data into consideration, is shown. Finally, the proposed ability estimation was compared with other related work. The details of the experimental designs are described in the following subsections.

6.1 Setting

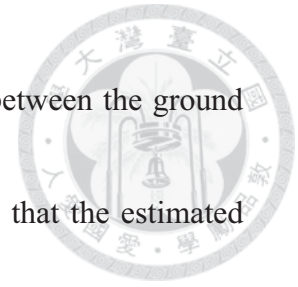
To understand the performance of the proposed method, we conducted a simulation. According to a one-parameter logistic model in Item Response Theory (Emberson & Resise, 2000), the probability of correct response is 0.5 when an item difficulty is equal to an examinee's ability. In the simulation, we referred to this probability for setting the variable r . Moreover, the item response model also provides information in



the estimation of the variable s . We used a one-parameter logistic model to predict the probability of a correct response when given the ability and an item, and then conditionally randomly sampled the variable $s \sim N(\text{given ability}, 0.2)$.

In each simulation, ten items were generated according to an examinee's ability at the time. The distribution of difficulty of these items acts as a normal distribution. For example, given an examinee's ability $\theta=3$, the difficulties of a test are $\{2, 2, 3, 3, 3, 3, 3, 3, 4, 4\}$. Ability and difficulty in this study range from one to six, corresponding to the school grades. In practice, an examinee's school grade is considered as their initial ability, and the ability is updated by responses in each test. Thus, the simulation starts with any grade ranging from one to six in order to simulate different grade students with various abilities, updates the estimated ability and then terminates 100 iterations after the convergence point. We found the convergence point and then counted the Root Mean Square Error (RMSE) during the 100 iterations. The definition of the convergence point is determined by computing the difference between the estimated ability and the ground truth, and the difference value is continuously four times smaller than a threshold ($thd = 0.25$ in the simulation). Each simulation was processed 1000

times. RMSE is used here, which represents the average distance between the ground truth and the generated results. The smaller RMSE value indicates that the estimated ability is close to the ground truth. In addition, we also discuss the parameter α in equation (16). The parameter is presented in terms of n time periods and represents the weight of the observation at the present time. The variable n was set from one to twelve.



6.2 The characteristics of the proposed ability estimation



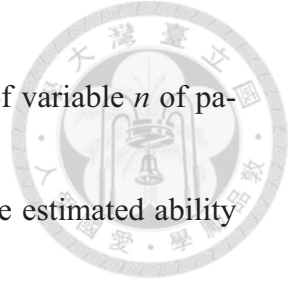
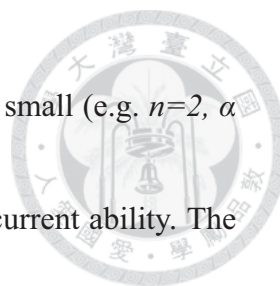


Table 7 shows the average convergence points in the number of variable n of parameter α in equation (16) over the degree of difference between the estimated ability and ground truth, and the results of RMSE during the 100 iterations after the convergence points. It is clear that the proposed method can successfully estimate abilities in the finite iterations. Specifically, an examinee's ability can be estimated more precisely when he or she continues to have more tests. Furthermore, the error distances between the estimated abilities and the ground truths are low enough to be acceptable after convergence. That is, an examinee's ability can be steadily measured during a long-term observation.

The parameter $\alpha = 2/(n+1)$ in the equation (16) is an exponential weight of the current ability, and n represents the number of time periods, such as times or days, taken into consideration. When $n=1$, it represents that an examinee's ability only considers the current estimated ability without the history record. In Table 7, the values in screentone present that the average convergence points are fewer than the points generated from $n=1$. This result shows that the estimated abilities are quickly found and the error distances decrease when considering the history record. In particular, it is ap-



parent when the initial grade is equal to the ground truth. When n is small (e.g. $n=2$, $\alpha = 2/3$; $n=3$, $\alpha = 1/2$), the estimated ability is mainly decided by the current ability. The convergence points are smallest and the RMSE is slightly smaller than one generated from $n=1$. In contrast, when n increases, the estimated ability is principally composed of abilities from the past to now. If an examinee's initial ability is not close to his or her actual ability, it takes more information to accurately estimate. Although it takes time, the RMSE is clearly shrinking.

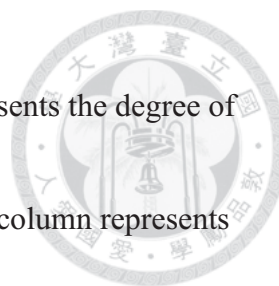


Table 7 The results of convergence point and RMSE (each row represents the degree of difference between the initial ability and the actual ability, and each column represents the number of time periods considered by the exponential weight of the current ability)

d \ n												
	1	2	3	4	5	6	7	8	9	10	11	12
0	20.61	13.88	11.72	11.53	10.98	10.90	10.26	10.52	10.16	10.35	10.18	10.04
1	21.96	16.17	15.74	16.31	17.40	19.07	20.43	22.29	23.98	25.45	26.92	28.42
2	22.91	18.08	18.54	19.91	21.90	24.18	26.64	29.06	31.50	33.53	35.62	38.58
3	23.86	19.67	19.91	21.91	24.59	27.62	30.33	32.90	35.74	38.43	41.52	44.13
4	24.30	20.73	21.52	23.51	26.71	29.68	32.96	36.00	40.19	42.83	45.45	48.65
5	24.50	21.41	22.66	25.22	29.10	31.92	35.97	38.22	42.62	46.40	49.18	53.12
RMSE	0.39	0.32	0.28	0.26	0.24	0.23	0.22	0.21	0.20	0.19	0.19	0.18

Consider a dramatic example to explain the properties of the proposed method. Assume that a first grade student, whose real ability is the sixth grade, learns and has a test in a web-based learning system once a day.

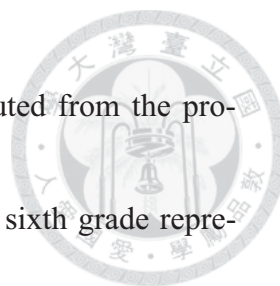


Figure 6 illustrates the changes in the estimated ability computed from the proposed method in different weights. The black horizontal line at the sixth grade represents the student's actual ability as the ground truth. The other curves depict the estimated abilities under the different weights: a red dotted line, $n=1$; a green solid line, $n=3$; a purple solid line, $n=6$; and a blue solid line, $n=12$. The mark labels on each line are the convergence points (the value is continuously four times smaller than $thd = 0.25$). It is clear that the estimated abilities are converging as n decreases in size. Although these estimated abilities are estimated using few iterations when $n=1$, the red-dotted line drastically fluctuates after the convergence point. In other words, if the ability estimation only takes the current responses into consideration, instead of past performance, the variance of every estimated ability may be large. In this situation, question selection in a test using inaccurate ability estimation could result in confusion by the examinee. In contrast, the estimated error gradually decreases when $n>1$, even though the estimated abilities when $n=1$ take more time to estimate. In this situation, the students' abilities were gradually updated and the difficulties of items incrementally increased. This is thus a trade-off problem between speed and precision.

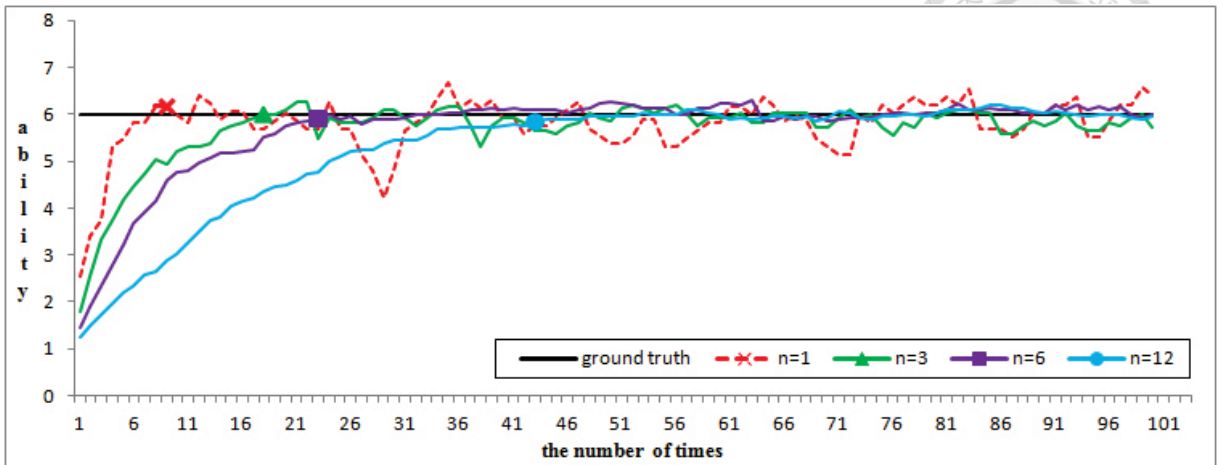


Figure 6 The changes in the estimated ability computed from the proposed method for the different weights ($n=1$, $n=3$, $n=6$, $n=12$)

6.3 The comparison with other ability estimations



To understand the performance of the proposed ability estimation, we compare our results ($n=1$ used in this section) to those of MLE (Emberson & Resise, 2000) and Lee (2012). One of the typical ability estimations in Item Response Theory is MLE in which the estimated ability is obtained by multiplying the item response function of each item and finding the highest possibility of which is the maximum likelihood estimate of a student ability using the Newton-Raphson method. Lee (2012) extended BME in Item Response Theory and proposed a conventional approach to approximate the posterior distribution of the student's ability obtained from the subsequent responses.

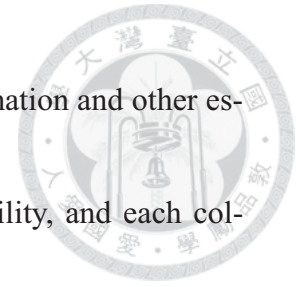


Table 8 shows the results of RMSE between the proposed estimation and other estimations. Each row represents the degree of simulated student ability, and each column represents the given difficulty of a test. When the difficulty levels of items were equal to the abilities of simulated students (shown as in the diagonals of the matrixes), the estimated results between MLE and Lee (2012) were similar, but these estimated by the proposed method was more close to the ground truth. With the increase in difference between the student abilities and item difficulties, it was obvious that the proposed estimation produced more accurate estimated abilities than other estimations. When questions were more difficult (the upper-right of the matrixes) or easier (the bottom-left of the matrixes) than the abilities of students, all of these methods failed to estimate the correct student abilities because the uncertainty among responses was unpredictable. But the error ranges of the proposed method were mostly within two grade; by comparison, the error ranges of MLE and Lee's method were from four to five grades. This demonstrates that the proposed method is robust especially when a student's ability is unknown. Moreover, note that the proposed method used in the section did not incorporate historical data during the estimation. It means that the estimated

abilities will be obtained more accurately if both of the current responses and the past performance are used in the ability estimation, as the previous section shown.

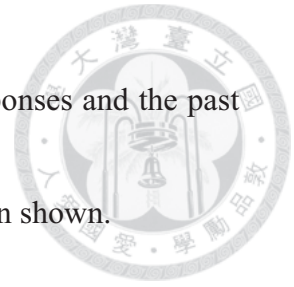
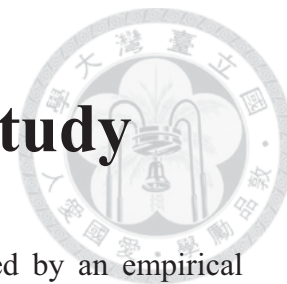




Table 8 The results of RMSE between MLE, Lee (2012) and the proposed ability estimation

MLE							Lee (2012)							The proposed method						
t \ s	1	2	3	4	5	6	t \ s	1	2	3	4	5	6	t \ s	1	2	3	4	5	6
1	0.22	1.00	2.01	2.99	4.04	5.13	1	0.21	1.01	2.04	3.05	4.13	5.17	1	0.13	0.52	1.04	1.51	1.95	2.18
2	1.00	0.23	1.00	2.02	3.03	4.04	2	1.01	0.22	1.01	2.05	3.11	4.15	2	0.51	0.13	0.52	1.04	1.54	1.95
3	2.00	0.99	0.22	1.01	2.01	3.03	3	2.03	1.00	0.21	1.02	2.04	3.11	3	1.03	0.52	0.13	0.53	1.03	1.53
4	2.96	1.99	1.00	0.23	1.03	2.01	4	3.05	2.02	1.01	0.22	1.04	2.05	4	1.50	1.02	0.51	0.13	0.53	1.03
5	3.98	3.01	1.98	1.00	0.24	1.01	5	4.09	3.07	2.01	1.01	0.23	1.02	5	1.93	1.53	1.01	0.52	0.13	0.52
6	4.91	3.93	2.98	2.00	1.00	0.23	6	4.74	3.78	2.84	1.87	0.89	0.11	6	2.16	1.92	1.51	1.03	0.52	0.13


Chapter 7 An empirical Study



In this chapter, the proposed ability estimation was examined by an empirical study. To investigate the performance of the proposed method, we will examine the correlation between the estimated abilities and real data; moreover, we explore the students' performance on the post-test and responses among the different ability groups. Next, the students' performance was analyzed whether or not appropriate instructional scaffolding could help students advance their learning; furthermore, we also analyze whether or not unclear concept will be enhanced by the proposed personalized computer-aided question generation. Finally, user satisfaction will be investigated by a questionnaire.

7.1 System and materials

The proposed system will be implemented and named as AutoQuiz. It will provide English language learners with computer-aided question generation. AutoQuiz will be integrated on the IWiLL learning platform (Kuo et al., 2002), which offers learners an online English reading and writing environment. Given the grade level of a



student, an article from an online news website is selected (see Figure 7a). After reading the article, the examinee will be given a test, consisting of ten vocabulary items (see Figure 7b), five grammar items (see Figure 7c), and three reading comprehension items (see Figure 7d). These items are generated automatically and respectively based on his/her vocabulary, grammar and reading comprehension levels. When the examinee finishes the test, the score and the incorrect responses will be shown (see Figure 7e). In addition, the system also shows an explicit warning near questions that are incorrectly answered (see the frame in Figure 7e). In order to encourage examinees to find the answer by themselves, the explicit warning shows the number of mistakes made rather than the answer, for any questions answered incorrectly less than three times (after which, the warning will reveal the correct answer). Finally, an error report button is designed to allow students to report any questionable items (see the circle in Figure 7b, Figure 7c, and Figure 7d), which experts will then check and remove if necessary.

A total of 2,481 items, composed of vocabulary, grammar, and reading comprehension, were automatically generated based on 72 news stories as reading materials.



These news articles were collected from several global and local online news websites:

Time For Kids (the estimated grade 1-4), Voice of America (the estimated grade 1-6),

China Post Online (the estimated grade 1-6), Yahoo! News (the estimated grade 5-6),

Student Times (the estimated grade 3), and CNN (the estimated grade 5-6).

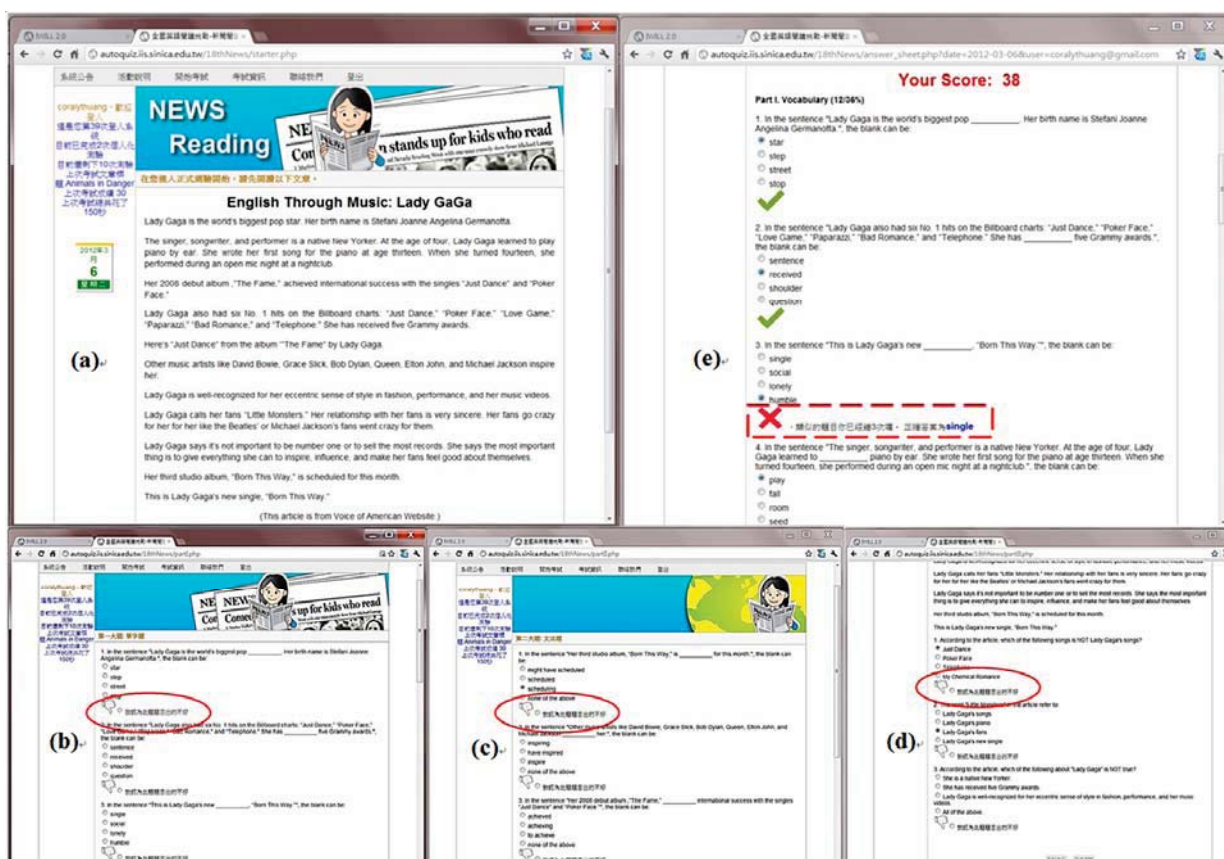


Figure 7 Snapshots of the system: (a) An example of a given reading materials from new online website; (b) An example of vocabulary items; (c) An example of grammar items; (d) An example of reading comprehension items; (e) An example of a score result with explicit warning.

7.2 Participants and procedure

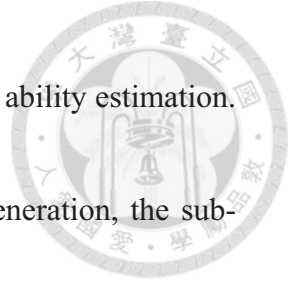


The participants in this study were the second grade students of senior high schools in Taiwan, who take English as a foreign language (EFL). During the experiment, the subjects were asked to participate in twelve activities, consisting of reading an article and then taking a test. Each test was composed of ten vocabulary questions, five grammar questions and three reading comprehension questions. After each activity, the proficiency levels of the subjects in the experimental group are estimated. The grade level in this study is defined from one to six, corresponding to the six semesters of Taiwanese senior high school. In addition, there were a pre-test and a post-test for evaluating learner's proficiency. They were from the College Entrance Examination and had similar degree of difficulties.

There are two investigations in the empirical study, one is to validate the accuracy of the proposed ability estimation with real data, and the other is to evaluate the performance of the proposed personalized computer-aided question generation. the participants in this study will be divided into two groups: a control group (C1: 30 students) where ability is estimated only based on current responses, and an experimental group

(E1: 47 students) that incorporates the history record into the current ability estimation.

In the investigation of the personalized computer-aided question generation, the subjects are divided into two groups: a control group with general automatic quiz generation (questions are generated according to their grades in the school, as the scenario in the traditional classroom; C2: 21 students), and an experimental group with personalized automatic quiz generation (questions are generated depending on their language proficiency; E2: 72 students). Noticeably, the subjects in each group are different person.



7.3 The performance of the proposed ability estimation with the empirical data



To validate the accuracy of the proposed ability estimation, the subjects' abilities in the two groups will be estimated, one's is only based on current responses (C1) and the other incorporates the history record into the current ability estimation (E1). Table 9 reports the Pearson's correlation coefficient between the estimated abilities (the estimated grade is rounded by the estimated score) and the post-test scores among the three quiz types. All of the measures are significantly positively correlated. The results in the experimental group ranged from *0.44* to *0.69*, while ones in the control group ranged from *0.47* to *0.54*. Most of the correlation values in the experimental group are higher than the values in the control group; this suggests that estimating ability with the history record leads to a clearer relationship between the estimated ability and the ground truth.



Table 9 The correlation result between the estimated ability and the post-test in the control group and the experimental group

	vocabulary		grammar		reading comprehension	
	score	grade	score	grade	score	grade
Control group	0.47*	0.49**	0.54**	0.51**	0.54**	0.47*
Experimental group	0.51***	0.44**	0.55***	0.55***	0.69***	0.65***

p<0.05, **p<0.01, *p<0.001*

Comparing the post-test score in each estimated ability (grade) is another way to assess the accuracy of the proposed ability estimation. If the estimated abilities are accurate, the subject performance of each ability will differ from that of other abilities.




Table 10 presents the mean post-test score of the subjects of different estimated abilities between the control group and the experimental group. Intuitively, a subject estimated a higher ability should have higher post-test score than one estimated a lower ability. One-way Analysis of Variance revealed that there were differences in the estimated vocabulary ability ($F=5.75, p=0.001$), the estimated grammar ability ($F=4.71, p=0.003$) and the estimated reading comprehension ability ($F=5.98, p<0.001$) in the experimental group, while there were no statistical differences between the estimated vocabulary and grammar ability in the control group. Noticeably, although the estimated reading comprehension ability in the control group has a significant difference, the mean scores among every ability fluctuated. The bolded values in

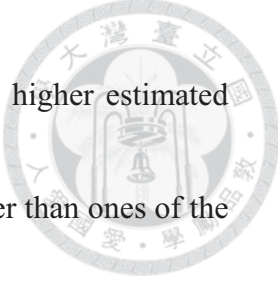


Table 10 are unreasonable, because the averaged scores of the higher estimated abilities (grade 2, grade 4 and grade 5) in the control group were lower than ones of the lower estimated abilities (grade 1 and grade 3). Though there was an unreasonable value for grade 6 of the estimated vocabulary ability in the experimental group, this is likely because only two students were assigned to grade 6. This sample size is likely unrepresentative. Moreover, in the experimental group, a Bonferroni post hoc test indicated that the performance of the estimated ability 1 and 2 were significantly different from the estimated ability 5 and 6. This indicates that the proposed ability estimation can effectively distinguish higher ability examinees from lower ones.

Table 10 The mean post-test score of the subjects in different estimated ability groups

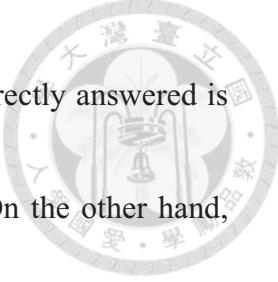


between both groups and the result of ANOVA

Estimated ability	Control group			Experimental group		
	vocabulary	grammar	reading	vocabulary	grammar	reading
1	-	37.50	46.80	-	-	37.67
2	48.33	47.00	40.00	23.00	34.33	46.63
3	38.00	51.40	52.57	52.86	52.80	53.50
4	54.40	41.40	41.00	62.33	54.94	64.50
5	61.22	62.83	32.67	69.71	66.81	66.90
6	65.83	65.56	70.18	57.67	72.00	78.00
F score	2.67	2.54	6.12***	5.75***	4.71**	5.98***

** $p < 0.01$, *** $p < 0.001$

To evaluate the validity of the proposed ability estimation, a logistic regression was performed. Table 11 shows the equations using the ability of a student i and the difficulty of a question j on the log odds ratio of the observation, which the student i correctly answers question j is in class 1 or the student incorrectly answers question j is



in class θ . Generally, the probability of which a question can be correctly answered is relatively higher, when the ability of a student is more advanced. On the other hand, the more difficult a question is, the lower the probability of which a student correctly answered a question is. If the observed abilities in the empirical study are precisely estimated, the relationship between the estimated abilities and dichotomous outcome will be explainable. The results showed that the regression coefficients for the ability of each student among these three question types are positive and the coefficient values for the difficulty of each questions among these types are negative. Even though the values among three question types were slightly different, all of them had the same influence on the dependent variable. This supports the assumption which the estimated abilities of students were so accurate that they, with advanced proficiencies, could correctly respond more difficult questions.

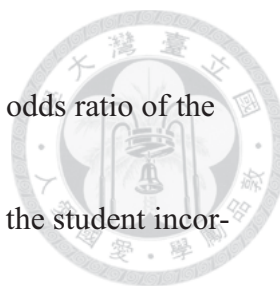


Table 11 The equations among question types represent that the log odds ratio of the observation that the student i correctly answers item j is in class 1 or the student incorrectly answers item j is in class 0 .

Question types	Equations
vocabulary	$\ln(p_{ij} / 1 - p_{ij}) = -1.554 + 1.129 \text{student}_i - 0.321 \text{question}_j$
grammar	$\ln(p_{ij} / 1 - p_{ij}) = -1.518 + 0.859 \text{student}_i - 1.321 \text{question}_j$
reading comprehension	$\ln(p_{ij} / 1 - p_{ij}) = -0.178 + 0.898 \text{student}_i - 0.783 \text{question}_j$

7.4 Student performance

To understand the influence of a personalized automatic quiz generation, we evaluate the effects of tests on student performance. The scores in the post-test between the experimental group (E2) and control group (C2) were calculated and compared. In keeping with the previous results, the estimated subjects' abilities in the experimental group were more accurate than those in the control group. We assume that appropriate instructional scaffolding could help students advance their learning, when effectively identifying their abilities.



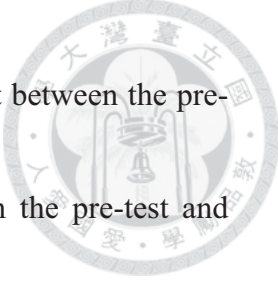


Table 12 presents the descriptive statistic and results of a T-test between the pre-test and post-test. The results of the independent T-test ($p=0.92$ in the pre-test and $p=0.51$ in the post-test) showed a similar effect on the post-test between the experimental group and the control group. One explanation for the results may be rooted in the short time (only five weeks) allowed for the treatment in the experiment, while Klinkenberg et al. (2011) conducted one-year experiment and Barla et al. (2010) employed their method for a winter term course. However, it is noticeable that the average score of the experimental group in the pretest was lower than the control group, but that of the experimental group in the post-test made great progress and surpassed the control group. Additionally, the paired sample T-test showed a significant effect of the pre-test and the post-test in the experimental group ($p<0.001$), while the performance of the control group had no statistically significant effect ($p>0.05$). This indicates that the subjects in the experimental group with an appropriate support can exceed the past themselves when successfully recognizing their learning status.

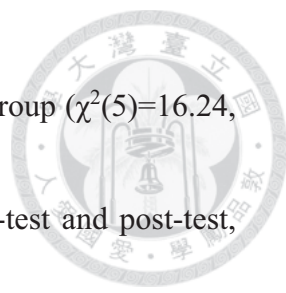


Table 12 The results of the pretest and post-test between the control group and the experimental group

	Pretest		Post-test		Paired sample
	mean	std.	mean	std.	t-test
Control group	53.23	19.35	56.70	17.99	1.57
Experimental group	52.83	16.67	59.28	16.01	3.71***
independent t-test	0.20		0.66		

*** $p < 0.001$

To further investigate the learning effectiveness, we studied the difference of student performance in each difficulty level between the pre-test and post-test. The number of correctly answered questions among the six difficulty levels in the pre-test and the post-test were computed. The tests are comprised of 28 items among six difficulty levels (six, three, six, three, seven and three questions per respective level, corresponding to levels one through six). A Chi-Square test for homogeneity of proportions was conducted to analyze the proportion between the pre-test and post-test. Table 13 presents two contingency tables respectively in the control group and the second



graders of the experimental group. The results of the experimental group ($\chi^2(5)=16.24$, $p<0.01$) show the significant different proportions between the pre-test and post-test, while the control group ($\chi^2(5)=7.46$, $p>0.05$) has a similar percentage among the six difficulty levels. This change reveals that the adaptive test affects the ability of the students in the experimental group. To further investigate the difference in the experimental group, a posteriori comparison reveals that the number of correctly answered questions with level two and level six in the post-test were statistically higher than those in the pre-test, whereas the number of questions with level one and level four in the post-test were significantly lower than those in the pre-test. This suggests that the number questions with higher difficulty level that were correctly answered increased after the personalized quiz strategy.



Table 13 Contingency tables for the number of correctly answered questions per difficulty level in the pretest and post-test.

Difficulty Level	1	2	3	4	5	6
The number of questions	6	3	6	3	7	3
Control group	69 (23.8%)	27 (9.3%)	63 (21.7%)	36 (12.4%)	68 (23.4%)	27 (9.3%)
Pretest						
Post-test	73 (21.3%)	50 (14.6%)	72 (21.0%)	33 (9.6%)		44 (12.8%)
Experimental group	248 (24.8%)	99 (9.9%)	209 (20.9%)	129 (12.9%)	206 (20.6%)	108 (10.8%)
Pretest						
Post-test	234 (20.5%)	147 (13.1%)	253 (22.6%)	106 (9.5%)	236 (21.1%)	142 (12.7%)

7.5 Unclear concept enhancement



The aim of the quiz strategy is to enhance students' understanding of unclear concepts behind incorrect responses. We measured the rate at which students successfully corrected their mistakes on repeated concepts (denoted as the rectification rate) in the experimental group (E2) and control group (C2), in order to determine the effect of generating items with repeated concepts and an appropriate difficulty. To make comparisons, the independent-samples *t*-test and the Mann-Whitney U test were both performed. Ideally, the distribution between the two groups is a normal distribution, and thereby uses a *t*-test. However, because of unequal sample sizes, the nonparametric method is complementary. The results of the rectification rate in the two groups can be seen in

Table 14. Here, the results suggest that the rectification rate in the experimental group was on average significantly higher than in the control group ($t=6.597$, $p<0.001$ in the independent-samples *t*-test and $Z=-5.974$, $p<0.001$ in the Mann-Whitney U test). Moreover, the subjects in the experimental group were more than half as likely to correct unclear concepts and answer similar questions correctly. This indicates that a

personalized approach would help learners correct previous mistakes.

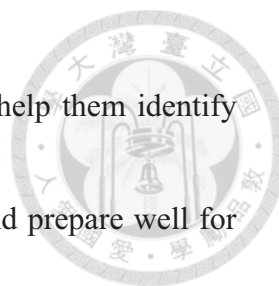


Table 14 The mean and standard deviation of rectification rate.

Group	Mean	Std.
Experimental group	<i>0.542</i>	<i>0.290</i>
Control group	<i>0.115</i>	<i>0.111</i>

7.6 User satisfaction

In terms of evaluating the performance of the automatic question generation, six questions in the questionnaire concerning the subjects' perception will be investigated. Subjects in the experimental group (E2) will fill out a questionnaire that elicited information concerning the examinees' experience and the quality of the generated questions. Questions in the questionnaire will be taken from (Wilson, Boyd, Chen, & Jamal, 2010). A five-point Likert scale will be employed. From the expectation of the results, most of the questions will score good results. Table 15 displays the detailed questions and shows their mean score and standard deviation. From the results, the quality of the interface and the functionality of the generated questions have high agreement. Most



subjects agreed that the adaptive question selection strategy could help them identify strengths and weaknesses, so that they could improve their skills and prepare well for exams. Item six, item seven and item eight assessed the quality of the generated questions among three categories, and item nine asked the subjects to self-assess their English ability after using the adaptive test environment.

Table 15 Questionnaire results.

Items	Mean	SD
1 The news interface is easy to use (Wilson et al., 2011).	3.89	0.99
2 The test interface is easy to use (Wilson et al., 2011).	3.86	0.95
3 Taking the quiz has helped me to evaluate my strengths and weaknesses (Wilson et al., 2011).	4.00	0.67
4 Taking the quiz has helped me to identify areas of knowledge that need improvement (Wilson et al., 2011).	4.03	0.64
5 Taking the quiz is useful preparation for exams (Wilson et al., 2011).	3.89	0.7
6a I clearly understood the vocabulary questions on the quiz	3.27	0.99

(Wilson et al., 2011).



6b	I clearly understood the grammar questions on the quiz (Wilson et al., 2011).	3.46	0.99
6c	I clearly understood the reading comprehension questions on the quiz (Wilson et al., 2011).	3.38	0.95
7a	Compare to the traditional manual questions, I can accept the quality of the vocabulary questions on the quiz.	3.57	0.99
7b	Compare to the traditional manual questions, I can accept the quality of the grammar questions on the quiz.	3.38	1.11
7c	Compare to the traditional manual questions, I can accept the quality of the reading comprehension questions on the quiz.	3.59	1.04
8a	Compare to the traditional manual questions, I agree with the quality of the vocabulary questions are comparable.	3.59	0.96
8b	Compare to the traditional manual questions, I agree with the quality of the grammar questions are comparable.	3.43	1.04
8c	Compare to the traditional manual questions, I agree with the	3.46	1.07

quality of the reading comprehension questions are comparable.



9a	I feel that I have made a progress in the vocabulary skills.	3.62	0.79
9b	I feel that I have made a progress in the grammar skills.	3.41	0.76
9c	I feel that I have made a progress in the reading comprehension skills.	3.81	0.81

Figure 8 displays charts of these items, with responses ranging from strongly agree (5) to strongly disagree (1) for questions on vocabulary, grammar and reading comprehension items. More than 80% of the participants understood the generated questions and agreed these questions were acceptable. Compared to traditional manual questions, automatic generated items were viewed as acceptable, especially for vocabulary items, which 92% of subjects believed were close to the traditional items. This information supports the performance of the proposed automatic question generation and represents the usefulness of the generated questions. Finally, the results show that more than 90% of examinees felt that their English had progressed.

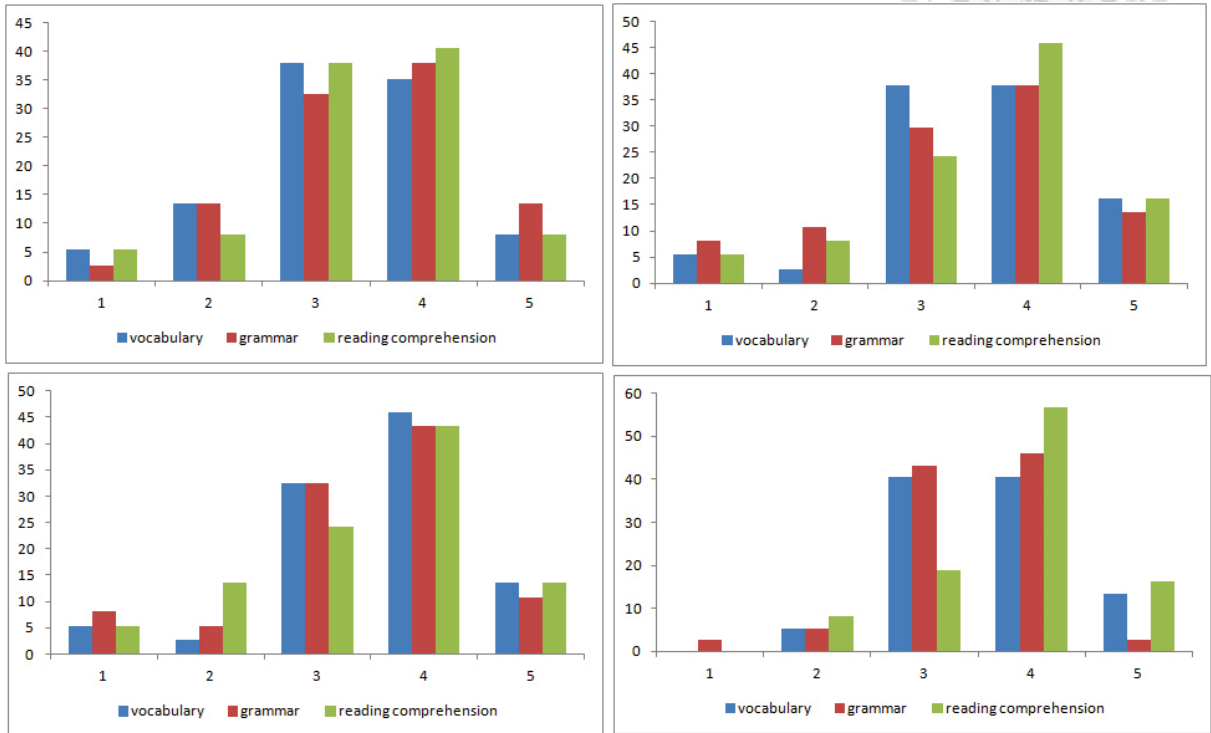


Figure 8 The charts on the percentage value vary from strongly agree to the strongly disagree for item six (upper left), item seven (upper right), item eight (lower left) and item nine (lower right).

Chapter 8 Discussion and Conclusion

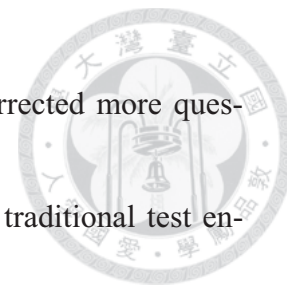


8.1 Summary

This work presents an adaptive test environment in order to enhance English as foreign language learners improving their understanding. We propose a personalized automatic quiz generation model to generate multiple-choice questions with varying difficulty and select questions depending on a student's estimated proficiency level and unclear concepts behind incorrect responses. We also present a reading difficulty estimation, which designed for English as foreign language learners. By Bayesian Information Criterion (BIC), we investigate the optimal combination of features for improving reading difficulty estimation. These features were extracted and sent to a linear regression model to estimate a reading level of a document. Finally, a novel and interpretable statistical ability estimation is presented based on the quantiles of acquisition grade distributions and Item Response Theory, and considers long-term observation as a student's estimated ability. The results in the empirical study showed:

- (1) The proposed personalized design with the appropriate instructional scaffolding helped students advance their learning progress.

(2) The students with our proposed personalized method corrected more questions which they answer incorrectly than students in the traditional test environment do.



(3) The questionnaire results showed that the proposed personalized method can identify the students' knowledge which needed to be improved and help students understand their strengths and weaknesses; furthermore, most subjects will agree that the proposed system is of functionality and quality

The proposed reading difficulty model not only inherently employed the complexity of lexical and syntactic features, but also newly introduced some meaningful new features such as the word and grammar acquisition grade distributions, word sense, and co-referential relations. The results from the evaluations reported:

(4) The representative features of the proposed reading difficulty estimation showed that the word acquisition grade distributions particularly plays an important role for reading materials written for English as foreign language learners.

(5) The results of the proposed reading difficulty estimation were better than the

previous work.



This work develops a statistical and interpretable method of estimated ability that captures the succession of learning over time in a Web-based test environment. Moreover, it provides an explainable interpretation of the statistical measurement based on the quantiles of acquisition grade distributions and Item Response Theory. The results from the simulation demonstrated:

- (6) The proposed ability estimation based on the grade distributions was robust especially when the responses were uncertain.
- (7) The result from proposed ability estimation was more accurate than the other ability estimations and can provide a better understanding of student competence.
- (8) The empirical results revealed that the correlation values between the estimated abilities which incorporating this testing history were higher than the values that only consider the test responses at the current test. Moreover, students who were estimated as advanced graders will show significantly higher post-test scores and better responses than ones who were estimated as

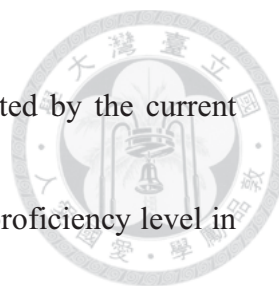
basic graders.



8.2. Contribution

To our knowledge, the work is the first empirical study to analyze the student performance with automatically generated questions and a personalized test strategy.

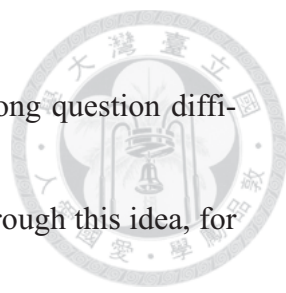
Table 16 Comparison of different test environments provides a comparison of the proposed system with previous test environments. In the traditional test environment, examinees in the same grade or class usually take the same tests, which was previously made by experts. In the adaptive test environment (e.g., Barla et al., 2010), tests are likewise made beforehand by experts, but examinees in the same grade or class could have different tests depending on their ability. With automatic question generation (e.g., Mitkov & Ha, 2003), tests save both time and production costs; nevertheless, they are usually not designed for any test purpose. In our method, questions and the difficulty of questions are not only generated automatically, but are also provided to examinees depending on various abilities and their previous mistakes. The examinee's performance is recorded in the system and the concepts behind incorrectly answered ques-



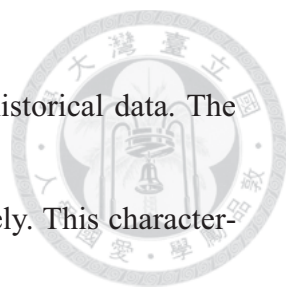
tions are reincorporated into future tests. Their abilities are estimated by the current responses incorporating testing history. Additionally, the estimated proficiency level in this study corresponds to an explicit grade level in a school, whereas that in the ability estimation of the traditional adaptive test environment is a point on an implicit scale. This retains the advantage of the adaptive test environment and automatic question generation. It offers students an effective approach to automatically measure their understanding and clear their incorrect concepts; moreover, it reduces teachers' burden on question generation. Teachers can take more time to teach and assist students.

Table 16 Comparison of different test environments.

Comparison	Automation	Personalization
The traditional test environment	No	No
The adaptive test environment	No	Yes
The automatic question test environment	Yes	No
The proposed adaptive test environment	Yes	Yes

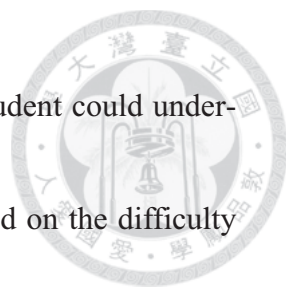


This work is the first to mathematically draw connections among question difficulties, ability estimation and the acquisition grade distributions. Through this idea, for example, the estimated ability represents a student as grade level six because he or she answered correctly 90 percent of items in a test with the difficulty level which normally distributed in level six and this behavior is equal to 80 percent of the population (assume $s=90\%$ and $r=80\%$). Unlike the traditional approaches, which focused on norm referenced item parameter scale for an individual item, the ability which estimated by the proposed method is explainable that the ability scale are based on the school grade of the most of people acquire these knowledge. In addition, for the proposed method, an examinee's ability is estimated from all responses of questions in a test; in contrast, for the traditional approaches, the ability was determined by an individual question. This point is similar to that of Classical Test Theory (Crocker & Algina, 1986), which considered all responses in a test as an examinee's observed scores. But the result of Classical Test Theory is sample-dependent; instead, the estimated result from the proposed method is stable due to estimating based on the acquisition grade distributions. Moreover, our estimated ability is obtained from the weighted



combination of an examinee's current performance and his or her historical data. The much historical data allows the ability to be estimated more accurately. This characteristic remains the advantage of BME (Bock & Mislevy, 1982; Baker, 1993; Lee, 2012), which considers the successive change in the ability level within a learning session, and achieves more accurate results than the BME. Finally, the experimental sample in this study was drawn from the student population with varied abilities, whereas the parameters in Lee's research (2012) were estimated on the student population with similar knowledge level. Even though the characteristics of Item Response Theory are robust enough to use the same student population without losing any generality, it would be better to acquire parameters from different student populations.

Several implications can be drawn from this study, if learners could learn English with this learning environment. First, it would provide a personalized learning environment. Students with different abilities could practice adaptive exercises with appropriate difficulties and repeatedly unclear concepts. This could be used as a qualitative guideline for identifying the current learning status of students for providing instructional supports, which could in turn enhance what students do not acquire yet. For ex-



ample, when the estimated ability of a student is determined, the student could understand his or her learning status because the ability is estimated based on the difficulty levels of words he or she acquired. It is easier for students to see the extent of their proficiency in the different levels. Moreover, the system records students' behavior, teachers can use this information to clear up misunderstanding that students have. Second, it would take away the barrier of the physical academic textbooks. By having online resource be available and updated every day, learners would be able to learn something new every time they want. Finally, the framework of this system could be used as a quantitative purpose for adapting the different learning environment for offering flexible measurement, which could set different values in these two parameters r and s depending on the various conditions. A good example is native speakers versus second language learners. In this way, teachers could adjust the parameters of the proposed ability estimation to the test purpose, regarding a qualified ability corresponding to the age which the certain percent of a population have acquired.

8.3 Limitations

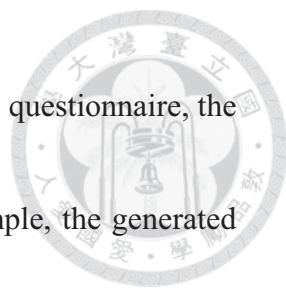


Limitations of our evaluations itself leave ample room for future research.

One limitation concerns the difficulty of reading comprehension questions used in the study. To develop this question type, we only took the predicted value from the proposed reading difficulty estimation as consideration. It should identify other criteria for the characteristics of anaphoric relations, e.g. the forward reference or the backward reference, or the most frequent mistakes in the coreference resolution.

One of the limitations in our current research is the limited question types even though vocabulary, grammar and reading comprehension (referential) questions were proposed in this study. It will be desirable to see more different generated questions types in the future work. Moreover, because of the limited number of question types, it is difficult to identify students' incorrect responses in reading comprehension questions. Although these questions are classified into various difficulties, it could be insufficient to investigate students' understanding. One possible solution is to observe and learn from data; however, it requires researchers or students to label and define this resource.

Future work should evaluate the personalized questions on additional criteria.



Even though these questions were evaluated with empirical data, e.g. questionnaire, the quality of generated questions could be examined further. For example, the generated questions could be evaluated by a comprehensive sample of experts: Is a generated question acceptable? One criterion is psychometric reliability: how well does performance on a question correlate with performance on other questions with the same difficulty? Another idea is to design how to filter an invalid generated question automatically.

Another limitation is that the distribution of item difficulties of questions in a test was assumed as a normal distribution. Even though teachers usually design a combination of difficulties of question in a test which is similar to a normal distribution, some questions are uniformly generated. One of possible solution is that the item discrimination parameter and the guessing parameter described in three-parameter logistic model of Item Response Theory might be taken into consideration. The item characteristic curve could accurately model the probability of a correct response between an examinee's ability and the item parameters. This concern would be much more desirable to address in the future.

Additional limitation is that this approach only focuses on English learning. The

personalized framework may be applied to other language learning field, but other disciplines, such as mathematics, need to be redesigned.

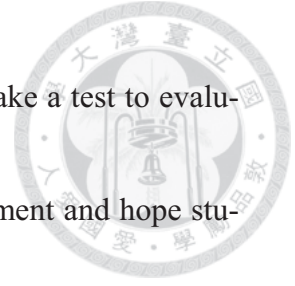


8.4 Future applications

One possible use is human-assisted machine generation of personalized question generation, for example, with the human editing or selecting among candidate questions generated automatically, thereby reducing the amount of human effort currently required to compose questions, and producing them more systematically. Further research might extend the framework for automatic use. One thing for the further development is to design the automatic evaluation of generated questions. If a generated question is reported as an unacceptable question, it should be removed and used to improve the algorithm.

Another potential application is adaptive test based on Big Data (Long & Siemens, 2011). With the emergence of abundant online learning materials and electronic textbooks, it is highly practical to employ the proposed framework of personalized automatic question generation in the future. We can imagine a scenario in which English as

a foreign language learner read up-to-date news and immediately take a test to evaluate himself. We look forward to a fast adoption of learning environment and hope students and teachers will have the benefits of this work.



References



- [1] Agarwal, M. & Mannem, P. (2011). Automatic gap-fill question generation from text books. *Proceedings of the 6th Workshop on Innovative Use of NLP for Building Educational Applications*, 56–64.
- [2] Agarwal, M., Shah, R., & Mannem, P. (2011). Automatic question generation using discourse cues. *Proceedings of the 6th Workshop on Innovative Use of NLP for Building Educational Applications*, 1–9.
- [3] Alario, F. X., Ferrand, L., Laganaro, M., New, B., Frauenfelder, U. H. & Segui, J. (2005). Predictors of picture naming speed. *Behavior Research Methods, Instruments, & Computers*, 36, 140-155.
- [4] Anderson, R. C., & Biddle, W. B. (1975). On asking people questions about what they are reading. *Psychology of learning and motivation*, 9, 89-132.
- [5] Baeza-Yates, R. & Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. New York: ACM Press.
- [6] Baker, F. B. (1993). Equating tests under the nominal response model. *Applied Psychological Measurement*, 17, 239-251.



- [7] Bates, E. (2003). On the nature and nurture of language. Retrieved November 24, 2011, from <http://crl.ucsd.edu/bates/papers/pdf/bates-inpress.pdf>
- [8] Barla, M., Bielikova, M., Ezzeddinne, A. B., Kramar, T., Simko, M. & Vozar, O. (2010). On the impact of adaptive test question selection for learning efficiency. *Computer & Education*, 55(2), 846–857.
- [9] Barzilay, R. and Lapata, M. (2008). Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1), 1-34.
- [10] Brown, R. G. (2004). *Smoothing, forecasting and prediction of discrete time series*. New York: Dover Publications.
- [11] Brown, J., Frishkoff, G. & Eskenazi, M. (2005). Automatic question generation for vocabulary assessment. *Proceedings of Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 819-826.
- [12] Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, 6, 431-444.



- [13] Barzilay, R. and Lapata, M. (2008). Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1), 1-34.
- [14] Brown, R. G. (2004). *Smoothing, forecasting and prediction of discrete time series*. Dover Publications.
- [15] Brysbaert, M., Wijnendaele, I. V., & Deyne, S. D. (2000). Age-of-acquisition effects in semantic processing tasks. *Acta Psychologica*, 104(2), 215-226.
- [16] Carrolla, J. B. & Whitea, M. N. (1973). Word frequency and age of acquisition as determiners of picture-naming latency. *Quarterly Journal of Experimental Psychology*, 25(1), 85-95.
- [17] Chali, Y. & Hasan, S. A. (2012). Towards Automatic Topical Question Generation. *Proceedings of the 24th International Conference on Computational Linguistics*, 475–492.
- [18] Chen, C. M., Lee, H. M., & Chen, Y. H. (2005). Personalized e-learning system using item response theory. *Computers & Education*, 44(3), 237–255.
- [19] Chen, C. M., & Chung, C. J. (2008). Personalized mobile English vocabulary learning system based on item response theory and learning memory cycle.



- Computers & Education*, 51(2), 624–645.
- [20] Chen, C. Y., Ko, M. H., Wu, T. W. & Chang, J. S. (2005). FAST : Free Assistant of Structural Tests. *Proceedings of the Computational Linguistics and Speech Processing (ROCLING 2005)*.
- [21] Chen, W., & Mostow, J. (2011). A Tale of Two Tasks: Detecting Children’s Off-Task Speech in a Reading Tutor. *Proceedings of the 12th Annual Conference of the International Speech Communication Association*, 1621-124.
- [22] Chen, W., Aist, G., & Mostow, J. (2009). Generating Questions Automatically from Informational Text. *Proceedings of AIED 2009 Workshop on Question Generation*, 17-24.
- [23] Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart & Winston.
- [24] Coleman, M. and Liao, T. L. (1975). A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283–284.
- [25] Collins-Thompson, K. and Callan, J. (2004). A Language Modeling Approach to Predicting Reading Difficulty. *Proceedings of the Human Language Tech-*



nology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL2004)

[26] Collins-Thompson, K., Bennett, P. N., White, R. W., Chica, S. Sontag, D.

(2011). *Personalizing Web Search Results by Reading Level. Proceedings of*

CIKM2011

[27] Copestake, A., Flickinger, D., Pollard, C. & Sag, I. A. (2005). Minimal

Recursion Semantics: An Introduction, *Research on Language and*

Computation, 3, 281-332.

[28] Dale, E. and Chall, J. S. (1948). A Formula for Predicting Readability. *Edu-*

cational Research Bulletin, 27(1).

[29] David, H. A. & Nagaraja, H. N. (2003), *Order statistics*. Marblehead, MA:

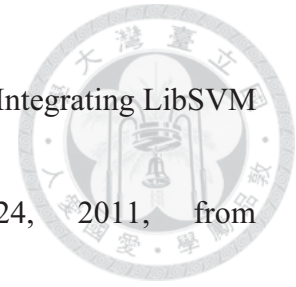
Wiley.

[30] Davies, R., Barbón, A., & Cuetos, F. (2013). Lexical and semantic

age-of-acquisition effects on word naming in Spanish. *Memory & Cognition*,

41(2), 297-311.

[31] EL-Manzalawy, Y. and Honavar, V. (2005). {WLSVM}: Integrating LibSVM into Weka Environment. Retrieved November 24, 2011, from <http://www.cs.iastate.edu/~yasser/wlsvm/>



[32] Elo, A. (1978). *The rating of chessplayers, past and present*. New York: Arco Publishers.

[33] Embertson, S., & Resise, S. (2000). *Item response theory for psychologists*. New Jersey, USA: Lawrence Erlbaum.

[34] Fehr, C. N., Davison, M. L., Graves, M. F., Sales, G. C., Seipel, B., & Sekhran – Sharma, S. (2012). The effects of individualized, online vocabulary instruction on picture vocabulary scores: an efficacy study, *Computer Assisted Language Learning*, 25(1), 87–102.

[35] Feng, L., Jansche, M., Huenerfauth, M., and Elhadad, N. (2010). A Comparison of Features for Automatic Readability Assessment. *Proceedings of International Conference on Computational Linguistics*, 276-284.

[36] Flesch, R. (1948). A new readability yardstick. *Journal of applied psychology*, 32(3), 221-233.



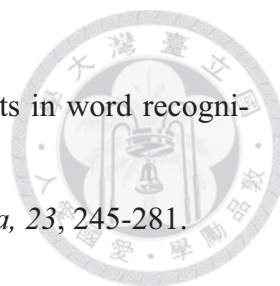
- [37] Gronlund, N. (1993). *How to make achievement tests and assessments*. New York: Allyn and Bacon.
- [38] Gunning, R. (1952). *The technique of clear writing*. McGraw-Hill, 1952.
- [39] Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: principles and applications*. Boston, MA: Kluwer-Nijhoff.
- [40] Heilman, M., Collins-Thompson, K., Callan, J., and Eskenazi, M. (2007). Combining Lexical and Grammatical Features to Improve Readability Measures for First and Second Language Texts. *Proceedings of the Human Language Technology Conference*, 460-467.
- [41] Heilman, M., Collins-Thompson, K. and Eskenazi, M. (2008). An Analysis of Statistical Models and Features for Reading Difficulty Prediction. *Proceedings of the Third ACL Workshop on Innovative Use of NLP for Building Educational Applications*, 71–79.
- [42] Heilman, M. & Smith, N. A. (2009). *Question generation via overgenerating transformations and ranking*. Technical report, Language Technologies Institute, Carnegie Mellon University Technical Report CMU–LTI–09–013,

Retrieved from

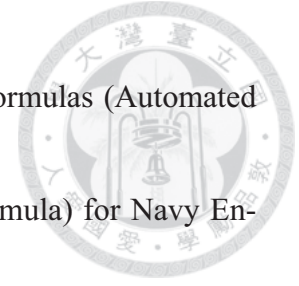
<http://www.cs.cmu.edu/~nasmith/papers/heilman+smith.tr09.pdf>.



- [43] Heilman, M. & Smith, N. A. (2010). Good question! statistical ranking for question generation. *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 609–617.
- [44] Hsiao, H. S., Chang, C. S., Chen, C. J., Wu, C. H. & Lin, C. Y. (2013). The influence of Chinese character handwriting diagnosis and remedial instruction system on learners of Chinese as a foreign language, *Computer Assisted Language Learning*, DOI: 10.1080/09588221.2013.818562.
- [45] Ho, H. and Huong, C. (2011). A Multiple Aspects Quantitative Indicator for Ability of English Vocabulary: Vocabulary Quotient. *Journal of Educational Technology Development and Exchange*, 4(1), 15-26.
- [46] Huang, S. X. (1996) A content-balanced adaptive testing algorithm for computer-based training systems. *Intelligent Tutoring Systems*, 306–314.



- [47] Izura, C., & Ellis, A. W. (2002). Age of acquisition effects in word recognition and production in first and second languages. *Psicologica*, 23, 245-281.
- [48] Johns, T. F., Hsingchin, L., & Lixun, W. (2008). Integrating corpus - based CALL programs in teaching English through children's literature, *Computer Assisted Language Learning*, 21(5), 483-506.
- [49] Kate, R. J., Luo, X., Patwardhan, S., Franz, M., Florian, R., Mooney, R. J., Roukos, S., Welty, C. (2010). Learning to Predict Readability using Diverse Linguistic Features. *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 546-554.
- [50] Kidwell, P., Lebanon, G., and Collins-Thompson, K. (2009). Statistical estimation of word acquisition with application to readability prediction. *Proceedings of Empirical Methods in Natural Language Processing*, 900-909.
- [51] Kidwell, P., Lebanon, G., & Collins-Thompson, K. (2011). Statistical Estimation of Word Acquisition With Application to Readability Prediction. *Journal of the American Statistical Association*, 106(493), 21-30.
- [52] Kincaid, J. Peter; Fishburne, Lieutenant Robert P., Jr.; Rogers, Richard L.;



- Chissom, Brad S. (1975). Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. *Branch Report*. Virginia: National Technical Information Service.
- [53] Kintsch, W. (1998). *Comprehension: A Paradigm for Cognition*. Cambridge: Cambridge University Press.
- [54] Kireyev, K. & Landauer, T. K. (2011). Word maturity: computational modeling of word knowledge. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, 299–308.
- [55] Klein, D. and Manning, C. D. (2003). Accurate Unlexicalized Parsing. *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, . 423-430.
- [56] Klinkenberg, S., Straatemeier, M., & van der Maas, H.L.J. (2011). Computer adaptive practice of Maths ability using a new item response model for on the fly ability and difficulty estimation. *Computers & Education*, 57(2), 1813-1824.
- [57] Kuo, C. H., Wible, D., Chen, M. C., Sung, L. C., Tsao, N. L. & Chio, C. L.



- (2002). Design and implementation of an intelligent Web-based interactive language learning system. *Journal of Educational Computing Research*, 27(3), 785–788.
- [58] Lin, Y. C., Sung, L. C. & Chen, M. C. (2007). An automatic multiple-choice question generation scheme for English adjective understanding. *Proceedings of the 15th International Conference on Computers in Education*, 137-142.
- [59] Lee, J. & Seneff, S. (2007). Automatic generation of cloze items for prepositions. *Proceeding of INTERSPEECH 2007*, 2173–2176
- [60] Lee, Y. J. (2012). Developing an efficient computational method that estimates the ability of students in a Web-based learning environment. *Computer and Education*, 58(1), 579-589.
- [61] Levy, R. & Andrew G. (2006). Tregex and Tsurgeon: tools for querying and manipulating tree data structures. *Proceeding of 5th International Conference on Language Resources and Evaluation*.
- [62] Lin, Y. C., Sung, L. C. & Chen, M. C. (2007). An automatic multiple-choice question generation scheme for English adjective understanding. *Proceedings*



- of the 15th International Conference on Computers in Education, 137–142.
- [63] Liu, C. L., Wang, C. H., Gao, Z. M., & Huang, S. M. (2005). Applications of lexical information for algorithmically composing multiple-choice cloze items. *Proceedings of the Second Workshop on Building Educational Applications Using Natural Language Processing*, 1–8.
- [64] Liu, M., Calvo, R. A., Aditomo, A., & Pizzato, L. A. (2012). Using Wikipedia and conceptual graph structures to generate questions for academic writing support. *IEEE Transactions on learning technologies*, 5(3), 251-263.
- [65] Long, P. and Siemens, G. (2011). Penetrating the Fog: Analytics in Learning and Education. *Educause Review*, 46(5), 31-40.
- [66] Lou, B and Guy, A. (1998). *The BNC handbook: exploring the British National Corpus*. Edinburgh: Edinburgh University Press.
- [67] Manning, C. D. & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.
- [68] Manning, C. D., Raghavan, P. D., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge: Cambridge University Press.



- [69] Mannem, P., Prasad, R., & Joshi, A. (2010). Question generation from paragraphs at UPenn: QGSTEC system description, *Proceedings of the Third Workshop on Question Generation*.
- [70] McDonald, J. L. (2000). Grammaticality judgments in a second language: Influences of age of acquisition and native language. *Applied Psycholinguistics*, 21(3), 395-423.
- [71] McLaughlin, G. (1969). SMOG grading: A new readability formula. *Journal of Reading*, 12(8), 639-646.
- [72] Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. J. (1990). WordNet: An on-line lexical database. *International Journal of Lexicography*, 3(4), 235-244.
- [73] Mitkov, R. & Ha, L. A. (2003). Computer-aided generation of multiple-choice tests. *Proceeding of the Workshop on Building Educational Applications Using Natural Language Processing*, 17-22.
- [74] Mitkov, R., Ha, L. A., & Karamanis, N. (2006). A computer-aided environment for generating multiple-choice test items. *Natural Language*



- Engineering*, 12(2), 177-194.
- [75] Mostow, J. & Chen, W. (2009). Generating instruction automatically for the reading strategy of selfquestioning. *Proceeding of Artificial Intelligence in Education*, 465-472.
- [76] Mostow, J. & Jang, H. (2012). Generating diagnostic multiple choice comprehension cloze questions. *Proceeding of the 7th Workshop on Innovative Use of NLP for Building Educational Applications*, 136–146.
- [77] Morrison, C. M., Ellis, A. W. & Quinlan, M. (1992). Age of acquisition, not word frequency, affects object naming, not object recognition. *Memory & Cognition*, 20(6), 705-714.
- [78] Nation, K. & Angell, P. (2006). Learning to read and learning to comprehend. *London Review of Education*, 4(1), 77-87.
- [79] Prasad, R., & Joshi, A. (2008). A Discourse-based Approach to Generating Why-Questions from Texts. *In Proceedings of the Workshop on the Question Generation Shared Task and Evaluation Challenge*.
- [80] Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A. K., &



- Webber, B. L. (2008). The Penn Discourse TreeBank 2.0. *In Proceedings of LREC*.
- [81] Pasca, M. (2011). Asking what no one has asked before: using phrase similarities to generate synthetic web search queries. *Proceedings of the 20th ACM international conference on Information and knowledge management (CIKM '11)*, 1347-1352.
- [82] Perfetti, C. A., & Hart, L. (2002). The lexical quality hypothesis. *Precursors of functional literacy*, 11, 67-86.
- [83] Pino, J., Heilman, M., & Eskenazi, M. (2008). A Selection Strategy to Improve Cloze Question Quality, *Proceedings of ITS Workshop on Intelligent Tutoring Systems for Ill-Defined Domains*, 22-32.
- [84] Pitler, E. and Nenkova, A. (2008). Revisiting Readability: A Unified Framework for Predicting Text Quality. *Proceedings of Conference on Empirical Methods in Natural Language Processing*, 186-195.
- [85] Piwek, P., Prendinger, H., Hernault, H., & Ishizuka, M. (2008). Generating Questions: An Inclusive Characterization and a Dialogue-based Application. *In*



Proceedings of Workshop on the Question Generation Shared Task and Evaluation Challenge, 25-26.

- [86] Raghunathan, K., Lee, H., Rangarajan, S., Chambers, N., Surdeanu, M., Jurafsky, D. & Manning, C. (2010). A multi-pass sieve for coreference resolution. *Proceeding of the Empirical Methods on Natural Language Processing*, 492–501.
- [87] Rus, V., Wyse, B., Piwek, P., Lintean, M., Stoyanchev, S., & Moldovan, C. (2010). The First Question Generation Shared Task Evaluation Challenge. *In Proceedings of the Sixth International Natural Language Generation Conference*.
- [88] Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6 (2), 461–464.
- [89] Schwarm, S. and Ostendorf, M. (2005). Reading Level Assessment Using Support Vector Machines and Statistical Language Models. *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, 523-530.
- [90] Stenner, A. J. (1996). Measuring Reading Comprehension with the Lexile



- Framework. *Fourth North American Conference on Adolescent/Adult Literacy*.
- [91] Sumita, E., Sugaya, F., & Yamamoto, S. (2005). Measuring Non-native Speakers Proficiency of English by Using a Test with Automatically-Generated Fill-in-the-Blank Questions, *Proceedings of the 2nd Workshop on Building Educational Applications using NLP*, 61-68.
- [92] Smith, S., Avinesh, P.V.S. & Kilgarriff, A. (2010). Gap-fill Tests for Language Learners: Corpus-Driven Item Generation. *Proceedings of the 8th International Conference on Natural Language Processing*.
- [93] Tanaka-Ishii, K., Tezuka, S., Terada, T. (2010). Sorting Texts by Readability. *Computational Linguistics*, 36(2), 203-227.
- [94] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, 58(1), 267-288.
- [95] Troubleyn, K., Heireman, K. & Walle, A. N. (1996). ATLAS: Computerised second language proficiency testing, *Computer Assisted Language Learning*, 9(4), 359 – 366.
- [96] Turney, P. D. (2001). Mining the Web for Synonyms: PMI-IR versus LSA on



- TOEFL. *Proceedings of the 12th European Conference on Machine Learning*, 491-502.
- [97] Turney, P. D., Littman, M. L., Bigham, J., & Shnayder, V. (2003). Combining Independent Modules to Solve Multiple-choice Synonym and Analogy Problems. *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, 482-489.
- [98] Wainer H. & Mislevy R.J. (1990). *Computerized adaptive testing: a primer*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- [99] Wan, X., Li, H., & Xiao, J. (2010). EUSUM: Extracting Easy-to-Understand English Summaries for Non-Native Readers. *Proceedings of the 33th Annual International ACM SIGIR Conference*, 491-498.
- [100] Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, 21, 361-375.
- [101] Wilson, K., Boyd, C., Chen, L. & Jamal, S. (2010). Improving student performance in a first-year geography course: Examining the importance of com-



- puter-assisted formative assessment. *Computers & Education*, 57 (2), 1493–1500.
- [102] Wood, D., Bruner, J.S., & Ross, G. (1976). The role of tutoring and problem solving. *Journal of Child Psychology and Psychiatry*, 17, 89-100.
- [103] Wu, C. H., Su, H. Y., & Liu, C. H. (2012). Efficient personalized mispronunciation detection of Taiwanese – accented English speech based on unsupervised model adaptation and dynamic sentence selection, *Computer Assisted Language Learning*, DOI: 10.1080/09588221.2012.687383.
- [104] Wu, R. Y. F., and Liao, C. H. Y. (2010). Establishing a Common Score Scale for the GEPT Elementary, Intermediate, and High-Intermediate Level Listening and Reading Tests. *Proceedings of International Conference on English Language Teaching and Testing*. Taipei: Language Training and Testing Center.
- [105] Urooj, U., Cornelissen, P. L., Simpson, M. I., Wheat, K. L., Woods, W., Barca, L. & Ellis, A. W. (2013). Interactions between visual and semantic processing during object recognition revealed by modulatory effects of age of acquisition. *NeuroImage*.



- [106] Van der Linden, W. J., & Glas, C. A. W. (2000). *Computerized adaptive testing: Theory and practice*. Boston, MA: Kluwer.
- [107] Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological process*. Cambridge, MA.: Harvard University Press.
- [108] Yamada, J. (2004). An L1-script-transfer-effect fallacy: a rejoinder to Wang et al. (2003). *Cognition*, 93, 127-132.
- [109] Yang, Y. C., Yang, C. F, Chang, C. M. & Chang, J. S. (2005). Computered-aid reading comprehension automatic quiz generation. *Proceedings of the Computational Linguistics and Speech Processing*.
- [110] Yao, X., Bouma, G., & Zhang, Y. (2012). Semantics-based Question Generation and Implementation. *Dialogue and Discourse*, 3(2), 11-42.
- [111] Zhang, D., Jiang, Q. and Li, X. (2005). Application of Neural Networks in Financial Data Mining. *World Academy of Science, Engineering and Technology*, 1, 136-139.
- [112] Zevin, J. D., & Seidenberg, M. S. (2002). Age of acquisition effects in word reading and other tasks. *Journal of Memory and Language*, 47 (1), 1-29.