1 Introduction

Since the 1970s, with rising public demand for transparent and explicit interpretations of test scores, level-based examinations have received growing attention in the field of language testing. The traditional norm-referenced approach to assessment compares test takers' performance relative to each other without establishing what they are able to do with the language. In contrast, level-based examinations divide language proficiency into defined levels which outline different degrees of achievement and identify whether test takers have attained a criterion standard. Test results are translated into proficiency statements suggesting the language activities that a test taker with a specific score is expected to be able to carry out. The proficiency statements of these level-based examinations are commonly formulated with reference to external standards, such as course objectives, national curricula, or proficiency frameworks that have already gained widespread acceptance to language levels to describe test takers' language competence and to facilitate communication between stakeholders about test objectives.

Recent advances in the fields of applied linguistics and language pedagogy have contributed to the development of numerous language proficiency frameworks in different contexts to reflect 'a hierarchical sequence of performance ranges' (Galloway 1987:27). These proficiency frameworks divide language proficiency into levels that are meaningful to their different users (Brindley 1986, 1991, Richterich and Schneider 1992, Trim 1977). The ones which have gained wide recognition and have continued to be actively used include the International Second Language Proficiency Ratings (ISLPR) Rating; later known as the Australian Second Language Proficiency (ASLPR) Scale (Ingram 1984); the American Council on the Teaching of Foreign Languages (ACTFL) Proficiency Guidelines (Hiple 1987); the Canadian Language Benchmarks (CLB; Pawlikowska-Smith 2000); and the Common European Framework of Reference for Languages (CEFR; Council of Europe 2001).

Among these frameworks, the CEFR has been the most widely used and recognised internationally 'to describe the levels of proficiency required by existing standards, tests and examinations in order to facilitate comparisons between different systems of qualifications' (Council of Europe 2001:21). In the past decade, various language testers and exam boards (e.g. Dunlea and

Matsudaira 2009, Kecker and Eckes 2010, Khalifa, ffrench and Salamoura 2010, Papageorgiou 2007, 2010, Tannenbaum and Wylie 2008, Wu and Wu 2010) followed the procedures that *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment: A Manual, Preliminary Pilot Version* (Council of Europe 2003), commonly referred as the Manual, proposed to align their exams to the CEFR (Council of Europe 2001). They attempted to describe their exams in terms of CEFR levels for the purpose of providing an easily accessible interpretation of test results to their test users and for use in seeking recognition from local governments and international professional organisations.

While the CEFR has been gaining popularity and has contributed to describing test constructs over the past decade, various case studies (e.g. Alderson (Ed) 2002, Figueras and Noijons 2009, Kecker and Eckes 2010, Khalifa et al 2010, Martyniuk and Noijons 2007, Morrow 2004, Wu and Wu 2010) have pointed to the difficulty in using the CEFR to establish proficiency bands in precise terms and call for fuller elaboration of these levels. Westhoff (2007:676) argued that 'although the CEFR descriptors tell us a lot about what learners at a certain level can do, very little is stated about what they should know . . .'. Weir (2005b:12) shared this view and commented that 'the CEFR provides little assistance in identifying the breadth and depth of productive or receptive lexis that might be needed to operate at the various levels.' He argued (2013:434) that examination boards need to 'determine what is an acceptable range for each parameter at each level of proficiency' in order to improve '. . . specifications for the different levels of proficiency which are, at best, vaguely and sparsely specified in the current Common European Frame of Reference.' Alderson, Figueras, Kuijper, Nold, Takala and Tardieu (2006:12) noted that many of the terms in the CEFR are not explicitly defined (e.g. 'long' and 'longer', 'straightforward' and 'complex'), and the CEFR provides no guidance on what structures, lexis or other linguistic features learners might be expected to cope with in order to complete test tasks at various proficiency levels. In addition to the textual features of test tasks, McNamara (1996) and Weir (1993) considered that the cognitive processes engaged by the examinees need to be given equal importance as well so that both the tasks and the conditions under which the tasks are performed can approximate to performance in the real world as closely as possible. In view of the CEFR's inherent limitations, O'Sullivan and Weir (2011) argued that considerable supplementary resources are needed to more comprehensively and explicitly define the levels as described in the CEFR. Weir (2005b:3) proposed that 'a framework is required that helps identify the elements of the context and processing and the relationships between these at varying levels of proficiency, i.e. one that addresses both situational and interactional authenticity (Bachman and Palmer 1996).' To demonstrate the extent of differentiation across exam levels, it will be necessary to identify

criterial features of the test tasks and to determine an acceptable range for relative degrees of complexity of each criterial feature at each level of proficiency for which the exam boards offer examinations.

Recognising the need to validate how the constructs of level-based exams may differ according to learners' level of language proficiency, the present study aimed to identify parameters that are useful for developing operationalisable specifications for different levels of reading proficiency and to establish an empirical framework enabling test validation and comparison of examinations developed by different exam boards aiming at the same level. The scope of the study is limited to CEFR B1 and B2 levels. This study applied Weir's (2005a) socio-cognitive validation framework to collect validity evidence of different test levels in terms of contextual parameters, cognitive processing skills, and test results. It focuses on the *cross-level* relationships between two CEFR-aligned reading tests, i.e. the General English Proficiency Test (GEPT) and the core Cambridge English examinations at the B1 and B2 levels.

The GEPT is a 5-level English as a Foreign Language (EFL) testing system, developed by The Language Training and Testing Center (LTTC), Taiwan, in accordance with Taiwan's national education framework. The LTTC, originally named The English Training Center, was established in 1951 to provide training in English for government-sponsored personnel preparing to go to the United States under technical assistance programs in place at that time. In 1986, the Center was registered with the Ministry of Education in Taiwan as a non-profit educational foundation. The LTTC now offers language training and testing in English, Japanese, French, German, and Spanish. In March 1998, the Ministry of Education (MOE) in Taiwan promulgated the *Towards A Learning Society* (邁向學習社會) white paper to promote lifelong learning. Under this policy in 1999, the MOE lent its support to the LTTC to develop the GEPT in order to enhance citizens' motivation for learning English by providing accessible attainment targets for English learners in Taiwan. Test content at the first two levels of the GEPT, i.e. Elementary and Intermediate, is guided by the national curriculum objectives of junior high schools and senior high schools, respectively. The three upper levels of the GEPT, i.e. High-Intermediate, Advanced, and Superior, for which no national curriculum exists, were developed based on the expectations of stakeholders in English education in Taiwan as identified through textbook analysis, needs analysis, and teachers' forums. Items and content for each GEPT level are designed to match specific level criteria which include a general level description of the overall English proficiency expected at that level and specific skill-level descriptors for the listening, reading, writing, and speaking components.

In 2004, the Executive Yuan, the highest administrative body in the government (comparable to the cabinet in other countries), approved 'measures

to enhance the English proficiency of civil servants (提升公務人員英語能力改進措施)', a plan undertaken under the policy 'Challenge 2008-National Development Plan (挑戰 2008 國家發展重點計畫)' (2002), and called for 50% of civil servants to pass the GEPT Elementary or Intermediate levels, or other certified equivalent English exams, within three years. To provide information for interpreting scores from different tests, Taiwan's MOE decided to adopt the CEFR as an international yardstick to benchmark test results. The CEFR, which divides communicative proficiency into six levels arranged in three bands – Basic User (A1 and A2), Independent User (B1 and B2), and Proficient User (C1 and C2), is intended to 'provide a common basis for elaboration of language syllabuses, curriculum guidelines, examinations, textbooks, etc.' (Council of Europe 2001:1) and has been used in Europe and beyond (e.g. Korea and Canada) to describe curricular aims and learner attainment, as well as to interpret test performance; therefore, the Ministry considered that the framework suited its need to set English proficiency targets for EFL learners in Taiwan and establish a common platform for comparisons of standards with foreign language educational systems in other countries. Since 2005, the MOE has required all major English exams administered in Taiwan to be mapped against the CEFR. The LTTC thus followed the procedures proposed by the Manual (Council of Europe 2003) to relate the GEPT to the CEFR levels (Wu and Wu 2010). The results showed that the Elementary, Intermediate, High-Intermediate, and Advanced levels of the GEPT reading tests are situated at CEFR A2, B1, B2, and C1 levels, respectively.

The core Cambridge English examinations, developed by Cambridge English Language Assessment, formerly named University of Cambridge ESOL (English for Speakers of Other Languages) Examinations, were used as external criterion measures to provide evidence of criterion-related validity for the GEPT level tests in this study. They were selected because they are among the few examinations which have made claims about the relationships of their examinations to the levels of the CEFR. The core Cambridge English examinations 'already ha[ve] an established connection with the CEFR' (Khalifa et al 2010:98), and is 'among a relatively small number of examination[s]' that have applied all three procedures, i.e. 'Specification of the content and purpose', 'Standardisation of interpretation of CEFR levels', and 'Empirical validation studies', recommended by the Manual (Council of Europe 2003) to link with the CEFR (Taylor and Jones 2006:4).

The University of Cambridge formed the University of Cambridge Local Examinations Syndicate (UCLES), now Cambridge English Language Assessment, over 150 years ago. Its aims were to raise standards in education by administering exams for people who were not members of the University. Cambridge English Language Assessment provides a variety of examinations covering a wide range of subjects and levels. The five levels of the core Cambridge English examinations are: *Cambridge English: Key* (KET;

also known as Key English Test), *Cambridge English: Preliminary* (PET; also known as Preliminary English Test), *Cambridge English: First* (FCE; also known as First Certificate in English), *Cambridge English: Advanced* (CAE; also known as Certificate in Advanced English), and *Cambridge English: Proficiency* (CPE; also known as Certificate of Proficiency in English). The CPE was first administered in 1913. Following the CPE, UCLES launched the Lower Certificate in English (renamed as FCE in 1975) in 1939, PET in 1980, CAE in 1991 and KET in 1994. These five tests correspond to the Association of Language Testers in Europe (ALTE) Levels 1 to 5 and CEFR A2, B1, B2, C1, and C2 levels, respectively. The five levels reflect the levels of language ability familiar to English language teachers around the world and have been described as 'natural levels' (North 2006:8).

A systematic comparison of the GEPT and the core Cambridge English examinations could potentially provide a more grounded specification of proficiency levels at CEFR B1 and B2 than is currently available and in so doing elaborate an efficient methodology for such comparisons that other examination boards might find useful. It would also provide the LTTC and Cambridge English Language Assessment with validity evidence relating to the constructs underlying their English language assessments at these levels.

The main questions that this study addresses are:

> Research Question 1: Is a GEPT reading test designed to measure at CEFR B2 level more difficult than a GEPT reading test designed to measure at CEFR B1 level in terms of test results, contextual parameters, and cognitive processing skills?
>
> Research Question 2: Are GEPT reading tests at CEFR B1 and B2 levels comparable to alternative CEFR-linked measures in terms of test results, contextual parameters, and cognitive processing skills?

This introductory chapter has provided an outline of the background to the study, the research objectives, and the research questions. Chapter 2 presents a review of related literature on vertical scaling, horizontal comparison of test scores on different tests at an equivalent level, content-based approaches to defining and comparing proficiency levels, and test comparability. A review of vertical scaling includes research on linking different levels of a multilevel exam onto the same vertical scale to provide direction in the construction of data collection and procedures for validation of vertical differentiation of a level-based test, followed by a brief discussion of how scores on a different test at an equivalent level can be used as an external criterion-related check on the validity of a defined level of difficulty. To sort through features that different language exams adopt to define levels of proficiency, the literature on CEFR alignment, CEFR linking studies, language proficiency scales which have gained wide recognition and have continued to

be actively used, contextual impacts on reading performance, and cognitive processing in reading, are surveyed. The literature survey on CEFR alignment covers alignment procedures and CEFR linking studies to provide the background to and justification for the present study. This chapter concludes with a discussion of issues involved in comparing examinations.

Chapter 3 discusses the research methodology used in this study. To answer Research Question 1, the research design and procedures for vertically linking scores from different test levels onto a common score scale are described in order to examine whether difficulty increases as the test level advances. To answer Research Question 2, empirical procedures for comparing two different reading tests targeting the same proficiency level are explained to assess whether two reading tests, provided by different exam boards at the same CEFR level, are comparable in terms of test takers' performance. In addition, qualitative and quantitative procedures to analyse contextual features and cognitive operations involved when test takers are responding to the reading tests are presented to answer both Research Questions 1 and 2.

Chapter 4 reports results of the validation of the GEPT level framework in terms of test difficulty, which addresses Research Question 1. Results from vertically linking different levels of the GEPT onto a common score scale are presented, and qualitative and quantitative analyses of contextual features and cognitive processes are discussed.

Chapter 5 reports results from empirical validation comparing two CEFR-aligned tests at the same proficiency level to answer Research Question 2. Results from the empirical comparison between scores from the GEPT and Cambridge English reading tests at CEFR B1 and B2 levels, respectively, are presented. Relationships between test performance and results from qualitative and quantitative analyses of contextual features and cognitive processes are discussed.

Finally, in Chapter 6, a summary of the findings is presented; the implications for test theory, for test design, for CEFR alignment procedures, and for teaching and course designers are discussed. Limitations of the present study are considered, and suggestions for future research are put forward.