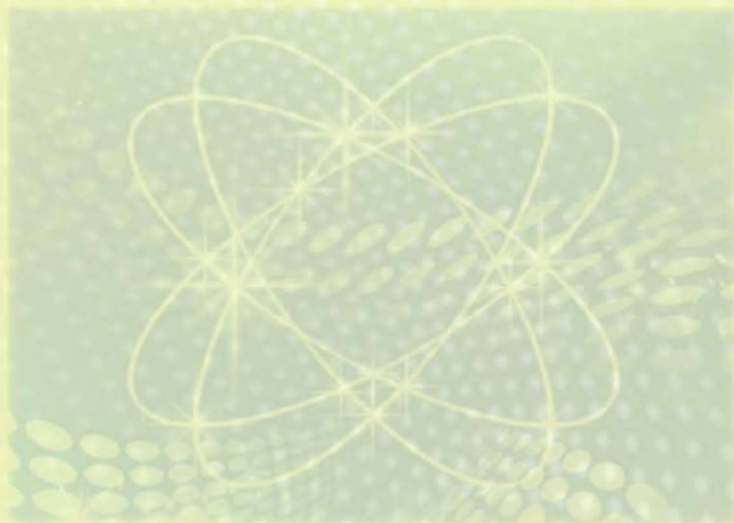


TEM和TOEFL考试内容 效度对比研究

王涛 著



天津科学技术出版社

山东省社会科学规划研究项目
专业英语四八级考试和托福考试内容效度对比研究
项目编号: 16CWZJ01

TEM和TOEFL考试内容 效度对比研究

王 涛 著

天津出版传媒集团

 天津科学技术出版社

图书在版编目 (CIP) 数据


TEM和TOEFL考试内容效度对比研究 / 王涛著. —天津: 天津科学技术出版社, 2018. 4
ISBN 978-7-5576-5035-3

I. ①T… II. ①王… III. ①大学英语水平考试—考试—内容—研究②TOEFL—考试—内容—研究 IV.
①H310.4

中国版本图书馆CIP数据核字 (2018) 第073931号

责任编辑: 布亚楠

天津出版传媒集团

 天津科学技术出版社出版

出版人: 蔡 颢

天津市西康路35号 邮编 300051

电话 (022) 23332695 (编辑部)

网址: www.tjkjcs.com.cn

新华书店经销

天津午阳印刷有限公司印刷

开本 787×1092 1/16 印张 14.5 字数 400 000

2018年4月第1版第1次印刷

定价: 38.00元

Content

Chapter 1 Introduction	1
1.1 Research Background	1
1.2 Research Purpose	3
1.3 Research Methodology	4
1.4 Layout of the book	4
Chapter 2 Literature Review	7
2.1 Basic Description	7
2.1.1 Basic Description of TEM	7
2.1.2 Basic Description of TOEFL	14
2.1.2.1 Internet-based test	15
2.1.2.2 Paper-based Test	18
2.1.2.3 TOEFL ITP Tests	19
2.1.2.4 TOEFL Junior Tests	19
2.2 Validity and Validation	20
2.2.1 Aspects of Validity Evidence	22
2.2.2 Theory-based Validity	23
2.2.3 Context Validity	25
2.2.4 Scoring Validity	26
2.2.5 Consequential Validity	27
2.2.6 Criterion-related Validity	29
2.3 The Need for Further Research	30
2.3.1 The small proportion of reading researches	30
2.3.2 The small part of comparison analysis	31
2.3.3 The limited efforts on specific reading tests	33
Chapter 3 A Proposed Framework for the Comparative Analysis	35
3.1 Basic information about TEM and TOEFL	36

3.2 Critical Features of Reading Texts.....	37
3.2.1 Text Forms	38
3.2.2 Text Length	39
3.2.3 Text Readability	39
3.2.3.1 Sentence Length	40
3.2.3.2 Percentage of Long Words	41
3.2.4 Text Topics	42
3.3 Critical Characteristics of Reading Questions	44
3.3.1 Formats of Expected Response.....	44
3.3.1.1 Selective	45
3.3.1.2 Productive.....	50
3.3.2 Testing Points	54
Chapter 4 A Quantitative and Comparative Analysis in TEM and TOEFL	58
4.1 Analysis of Critical Features of Reading Texts	58
4.1.1 At the Horizontal Level	58
4.1.1.1 Text Forms.....	58
4.1.1.2 Text Length.....	59
4.1.1.3 Text Readability.....	61
4.1.1.3.1 Sentence Length.....	62
4.1.1.3.2 Percentage of Long Words	62
4.1.1.3.3 Readability.....	63
4.1.1.4 Text Topics	64
4.1.2 At the Vertical Level.....	66
4.1.2.1 Text Length.....	66
4.1.2.2 Text Reliability	70
4.1.2.2.1 Sentence Length.....	70
4.1.2.2.2 Percentage of Long Words	71
4.1.2.2.3 Readability.....	72
4.2 Analysis of Critical Characteristics of Reading Questions.....	73
4.2.1 At the Horizontal Level	73
4.2.1.1 Formats of Expected Response	73
4.2.1.2 Testing Points	74

4.2.2 At the Vertical Level	75
4.2.2.1 Formats of Expected Response	75
4.2.2.2 Testing Points	76
Chapter 5 Conclusion	80
5.1 Major Findings	80
5.2 Implications	81
5.3 Limitations	84
5.4 Summary	85
Chapter 6 Prototyping Measures of TOEFL	87
6.1 Guiding questions for the second prototyping phase	87
6.2 Methodology	90
6.2.1 Participants	91
6.2.2 Materials	91
6.3 Prototype TOEFL measures	91
6.3.1 Test-Based Criterion Measures	93
6.3.2 Measures of Other Criteria	94
6.3.3 Sample Demographics	95
6.4 Findings	97
6.4.1 Outcomes Relevant to Evaluation	97
6.4.2 Scoring complex selected-responses tasks	97
6.4.3 Appraisals of speaking samples	100
6.4.4 Prototype TOEFL Measures	101
6.5 Summary	104
6.5.1 Outcomes Relevant to Generalizability	105
6.5.2 Reliability of the Listening Measure	107
6.5.3 Generalizability Analysis for Speaking Scores	108
6.5.4 Generalizability Analysis for Writing Scores	112
6.5.5 An Analysis of Speech Samples	119
6.5.6 Outcomes Relevant to Extrapolation	120
Chapter 7 The TOEFL Validity Argument	121
7.1 Validity argument from test design	122
7.1.1 Domain definition	123

7.1.2 Domain Analysis	124
7.1.3 Simulation of Academic Tasks	126
7.2 Evaluation	127
7.2.1 Appropriate Scoring Rubrics	128
7.2.2 Task Administration Conditions	129
7.2.3 Psychometric Quality of Norm-Referenced Scores	130
7.3 Generalization	131
7.3.1 Configuration of Tasks	132
7.3.2 Relationships with Other Tests	132
7.3.3 Task and Test Specifications	133
7.4 Explanation	134
7.4.1 Three Planes of Explanation	134
7.4.2 Relationships with Other Tests	138
Chapter 8 Assessment Behavior Difference	140
8.1 A Comparison of Assessment behavior	140
8.2 Effects of Raters' Severity and Consistency	148
8.3 The case for the construct validity of TOEFL's minitalks	157
Appendix 1: A sample of TOEFL reading test	180
Appendix 2: A sample of TEM-4 reading test	198
Appendix 3: A sample of TEM-8 reading test	207
Reference	219

Chapter 1 Introduction

1.1 Research Background

At present, TEM (4&8) (Test for English Majors-Band 4 & Band 8), TOEFL (Test of English as a Foreign Language) are the most influential English proficiency tests in China. TEM4, TEM8 and TOEFL share many characteristics, such as similar candidates (mainly college students), similar purposes (mainly for the assessment of overall English proficiency and for the study in college environment), similar size (of an extremely large scale), similar test structure (consisting of at least three subtests: reading, listening and writing), etc.

Among English proficiency tests, TOEFL is the one with the longest history, operating from 1964. TOEFL claims to provide a test that is as objective as possible so that apart from writing tasks, tasks in the other parts adopt the sole format of MCQ (Multiple Choice Questions). TEM, available from 1978, is claimed to develop largely under the influence of American TOEFL, thus inevitably sharing some advantages and disadvantages of TOEFL. Around the year 2016 when the author began writing this book, TEM4, TEM8 and TOEFL all claimed to adopt new versions in the near future as a reform effort. This turning point provides a natural and appropriate moment for a summary of the previous construction efforts (including merits and demerits) of TEM4, TEM8 and TOEFL.

Owing to the popularity and impact of TEM and TOEFL, their lately used papers, specimen tests, or sample tests that are issued by influential organizations are easily accessible, as compared to some less influential tests. That supplies the author with sufficient test samples as an accurate representative of tests in TEM and TOEFL.

In the field of English as a foreign language testing, much has been written on validity, that is, what to test and how to test. Nevertheless, Messick (1992) points to the fact that "... many test makers acknowledge a responsibility for providing general validity evidence of the instrumental value of the test but very few do it". This validity evidence of the instrumental value plays a crucial role in supporting the inferences made from scores during the validation process. Messick's validation framework has been used by later researchers (Guerrero 2000;

Hasselgren 2000), but serious problems in implementation can be noted. We lack an empirical, comprehensive and operationalizable framework to validate tests of national or world-wide importance. We lack instruments and procedures for applying the validation framework to high-stakes tests, and we lack different perspectives from which to examine “the relationships between the testing instrument and the construct(s) it (the test) attempts to measure” (Weir 2005) or to establish the evidence to support particular interpretations on or inferences from the scores of a test.

In the field of non-first-language test, research has been conducted (Bridgeman & Carlson 1983; The TEM Testing Center 1997; Hamp- Lyons & Kroll 1997; Cumming et al. 2000; Hyland 2002; Weigle 2002) into various tests, e.g., IELTS (International English Language Testing System), TOEFL (Test of English as a Foreign Language), GRE (Graduate Record Examination), TEM (Test for English Majors) used at tertiary level to assess test validity in English, an essential skill, of course, for those wishing to pursue postgraduate studies. The assessment of English proficiency test also clearly entails the establishment of a framework or model with effective instruments to validate tests to measure students' English proficiency at tertiary level.

It is not until recently that a comprehensive and operationalizable validation framework for validating academic writing tests (AWT), along with the frameworks for validating tests of listening, reading, speaking in terms of theory-based validity, context validity, marker reliability/scoring validity, consequential validity, and criterion-related validity, has been proposed by Weir (2005). There is still, nevertheless, a lack of relevant effective instruments and procedures based on such a framework and a comparison of validity between TEM and TEFOL.

In China, there are now nation-wide Tests for English Majors at Band 4 and Band 8 (TEM4 and TEM8) at tertiary level in both of them administered according to specifications of the ELT Syllabus for English Majors in Higher Education.

TEM4 was initiated in 1990 and TEM8 in 1991. From 1993 onwards, SISU (Shanghai International Studies University) assumed the sole responsibility for producing both TEM4 and TEM8 papers. For both TEM4 and TEM8, all the test-takers are non-native English majors in colleges and universities. The number of participants has increased greatly in recent years, along with the education development in China (with about 126,000 participants in TEM4 and 56,000 in TEM8 in 2003; 159,000 in TEM4 and 78,000 in TEM8 in 2004; 195,000

in TEM4 and 98,000 in TEM8 in 2005). The results of the TEM tests are used for graduation purposes although not all the English majors in all universities or colleges are required to participate in TEM8 as a compulsory test. In addition, TOEFL tests are widely taken as international tests for those intending to pursue postgraduate programs, but are not compulsory for undergraduate students. In addition, all students need to write a dissertation in English as a prerequisite for graduation. However, few studies have been conducted to validate either the essay in TEM 4 and TEM8 or the university graduation dissertation as effective instruments in terms of their theory-, content-, scoring-, consequential- and criterion-related validity.

This study attempts to carry out this task by developing and/or piloting instruments and procedures to compare the content validity between TEM4&8 and TOEFL in writing and reading in China.

The processes of this study can be mainly divided into three stages:

1)Review the language testing literature on validity and validation frameworks with detailed parameters for test validity;

2)Develop and/or pilot relevant instruments and procedures for applying the validation framework to the existing TEM 4&8 and TOEFL;

3)Apply validation instruments and procedures to compare the validity of TEM 4&8 and TOEFL to investigate the extent to which they are valid measures and how some of the instruments are related to the abilities the test attempts to measure.

1.2 Research Purpose

This book aims to conduct an evaluation of the quality of reading subtests in TEM 4&8 and TOEFL based on a comparative analysis of a series of critical facets of tests. Since test quality is largely determined by validity and reliability, the comparative analysis will look into potential errors of validity and reliability (especially those concerned with content validity, instrument-related reliability, and parallel or alternate forms reliability) within the three tests so as to raise constructive suggestions for improvement in this respect.

The comparison is to be conducted from two aspects—critical characteristics of reading texts and reading questions, and on two dimensions—horizontally (without reference to time) and vertically (with reference to the time of testing). Based on such a framework, statistical data are to be obtained, from which objective information concerning the quality of the two

tests. The findings from the relevant information then are assumed to have implications for various people concerned with reading testing, such as test designers, teachers, learners and prospective test takers. Of particular interest to this book is the provision of relevant and useful information for TEM test designers responsible for the reading subtests so that the quality of reading tests in TEM4 and TEM8 can be enhanced in the future.

1.3 Research Methodology

Reading subtests of TEM 4&8 and TOEFL are subjects of this book. To ensure the representativeness of the data to be collected for analysis, a wide sampling effort is made to set up a sufficiently large test bank. The test bank includes reading-related sections of 20 consecutive versions of the each of the three tests, making up 80 versions of reading subtests altogether. They are Reading Comprehension sections of TOEFL tests, Reading Comprehension sections and alternative integrative testing sections (Short Answer Questions/ Translation/ Cloze/ Error Correction) of TEM4 and TEM8 tests, and Reading sections of TOEFL tests (used tests, specimen tests, and sample tests issued by authoritative publishers like Foreign Language Teaching and Research Press in Beijing, influential training centers like Longman/ Global TOEFL Corporation, etc). The tests within the test bank are in the e-form so as to make it easier and more convenient for statistical counting and rechecking. Most of the test materials have been cross-checked to ensure the correctness of the data derived out of them in later analyses. To enable a vertical analysis later in the research, the reading subtests thus collected are listed in the chronological order with reference to the time of test. Research methodology in the quantitative analyses includes manual counting as well as statistic counting with the assistance of WORD and EXCEL software, and the electric calculator. Most statistic counts have been conducted for at least twice per point so as to avoid operation-related errors.

1.4 Layout of the book

This book is composed of seven chapters. Chapter 1, an introduction to the book, includes four sections: 1) background, which explains the context from which the writing of this book derives; 2) research purpose, which illustrates the objective and brief content of this book; 3) research methodology, which discusses the scope of sampling subjects and the methods of data collection and qualitative analyses; and 4) the layout of the book, which

provides an overview of the organization of the book and a brief introduction to the content of each chapter.

Chapter 2 does a literature review that contains three broad sections—the basic description of TEM 4&8 and TOEFL, validity and validation, and the need for further research and theoretical foundations. The need for further research in this field is detected through a literature review of the field of TEM testing as well as those fields concerned with TOEFL as a whole. The section of theoretical foundations explores into issues of feasibility and necessity of the validity comparison between TEM and TOEFL, as well as criteria of test quality. The feasibility of reading testing provides a theoretical basis for the research of this book. The necessity of reading testing provides practical significance for the writing of the book. Criteria of test quality, mainly reliability and validity, provide technical indexes which give the analyses in the book a solid theoretical edge.

Chapter 3 is a display and illustration of a proposed framework initiated by the author concerning various facets of reading tests for the conduct of quantitative comparisons of reading tests among TEM 4&8 and TOEFL in Chapter 4. Before the presentation of the framework, a brief introduction to TEM 4&8 and TOEFL provides basic information, including the structure, weighting and techniques used in these influential standardized tests. Within the second and third sections, the display and illustration of various facets of reading tests is composed of two parts—the illustration of critical facets of reading texts and the classification of critical facets of reading questions. Critical facets of reading texts under discussion include text forms, text length, text readability, sentence length, percentage of long words, and text topics. Critical facets of reading questions under illustration contain formats of expected response and testing points in focus.

Chapter 4 is a display and analysis of the data collected through the conduct of comparisons of the series of facets of reading tests in TEM 4&8 and TOEFL in accordance with the framework proposed in Chapter 3. The analysis is made from two dimensions—horizontally (without reference to time) and vertically (with regard to test time).

Chapter 5, summarizes the major findings in the previous chapter and provides implications for test designers, teachers, learners and prospective test-takers. It also points out the limitation of this research, calls on further research effort, and finally briefly restates the main research effort of this book.

Chapter 6, reports prototyping measures of TOEFL. In this part, we discuss the questions

that motivated research during this phase in terms of their relevance to the assumptions underlying inferences and warrants in the evolving interpretative argument. We describe the design and methodology of the central study. The data and responses collected in this study provided grist for many different analyses and for other studies. The findings of these analyses are reported here as they pertain to the inferences in the interpretative argument. We conclude with a summary of how the results of these studies contributed to an emerging consensus about the design of a new TOEFL.

Chapter 7 summarizes the TOEFL validity argument. These perspectives on validation have been reflected throughout this book: The theoretical and empirical work pertaining to the TOEFL revision has been discussed in view of the evidence it provides for TOEFL interpretation and use. However, the approach taken in this volume also contributes toward the evolving conception of validation by presenting the research in terms of its role in an interpretive argument (Kane, 1992, 2001, 2004, 2006). The point of organizing research around an interpretive argument is to make clear the intended interpretations and uses of test scores. In other words, the interpretive argument defines what the validity argument is about. Before synthesizing the research, however, the role of the validity argument at this stage of the TOEFL's design needs to be clarified.

Chapter 2 Literature Review

The literature review below aims to reveal three essential issues concerned with the writing of this book: basic description of these three tests, validity and validation and the need for further study.

2.1 Basic Description

2.1.1 Basic Description of TEM

Test purpose: The purpose of the TEM is to measure the English proficiency of Chinese university undergraduates majoring in English Language and Literature and to examine whether these students meet the required levels of English language abilities as specified in the National College English Teaching Syllabus for English Majors (hereafter the Syllabus, NACFLT, 2000). Since the Syllabus divides the four-year English major undergraduate programme into the foundation stage (the first and second year) and the advanced stage (the third and fourth year), the TEM test battery correspondingly consists of TEM4 and TEM4 Oral, assessing students' English proficiency at the end of the foundation stage, and TEM8 and TEM8 Oral, assessing students' English proficiency at the end of the advanced stage.

Length and administration: The TEM is administered by the National Advisory Committee for Foreign Language Teaching (NACFLT) on behalf of the Higher Education Department, Ministry of Education, People's Republic of China. The three tests in the battery are all administered once a year with TEM4 in April, TEM8 in March, TEM4 Oral in May and TEM8 Oral in December. The total test time is 135 minutes for TEM4 and 195 minutes for TEM8. Each oral test takes approximately 25 minutes to complete.

Scores: TEM test scores are reported to the Academic Affairs Office of the participating universities. In the case of TEM4 and TEM8, individual test takers scoring 60 or above receive a certificate from the NACFLT on which their level of performance is reported, including 'excellent' (score 80 or above), 'good' (score between 70 and 79) and 'pass' (score between 60 and 69). Neither composite scores nor section scores are reported to test takers. They can, however, check their composite scores through the Academic Affairs Office of their

university. For the two oral tests, test takers who pass the tests are awarded a separate certificate from the NACFLT on which the same three levels are reported: ‘excellent’, ‘good’ and ‘pass’. The levels are converted from the average of the total raw scores awarded by two TEM authorized oral examiners.

Author/publisher and contact information: The National Advisory Committee for Foreign Language Teaching. Shanghai International Studies University, No. 550, West Dalian Road, Hongkou District, Shanghai 200083, People’s Republic of China. Tel: +86-21-35372000.

Price: The prices are 45 RMB for TEM4, 50 RMB for TEM8, 80 RMB for TEM4 Oral, and 90 RMB for TEM8 Oral (at the time of writing, the exchange rate of USD vs. RMB was roughly 1:7).

The TEM is a criterion-referenced English language test specifically targeted at university undergraduates majoring in English Language and Literature in China (NACFLT, 2004a, 2004b). TEM4 and TEM8 were officially launched in 1992 following the publication of the first national teaching syllabus for English majors for the foundation stage in 1989 and that for the advanced stage in 1990. Due to its rapid economic and social development and its increasing integration into the international community, China in the early 1990s was acutely short of university graduates proficient in the English language. As part of the country’s strategy for the new century, English language degree programs flourished and kept expanding, as was evidenced by both the growing number of universities offering degrees in English and the surging number of students majoring in English (see Dai & Hu, 2009). As a result, an English language test designed specifically for this group of English learners was badly needed. The debut of TEM4 and TEM8 was therefore met with ‘immediate consent and enthusiastic support among ELT specialists in China at that time’ (The TEM Test Centre, 1997, p. 1). The two oral tests, TEM4 Oral and TEM8 Oral, were subsequently launched in 1999 and 2003 after a task force carefully investigated the feasibility of their operationalization (see Wen & Zhao, 1995). Among the three tests in the test battery, only TEM4 is required of all English major undergraduates while TEM8 and the two oral tests are optional. Decisions concerning penalties for those failing TEM4 are, however, left in the hands of participating universities.

After about two decades of development, the TEM has grown into one of the predominant English tests in China, winning extensive recognition from the test takers, relevant institutions and society at large and playing an increasingly important role in English

language teaching and learning at the tertiary level. The number of English major undergraduates taking the tests has also been expanding steadily and rapidly over the years. Statistics from the test centre show that the number of TEM4 test takers soared from 8554 when the test was initially launched in 1992 to 270,000 in 2010. A similar growth pattern can be identified for TEM8 whose test population increased from 4613 in 1992 to 189,000 in 2010. The test population for the two oral tests has been quite stable in recent years with around 20,000 students taking TEM4 Oral and 10,000 students taking TEM8 Oral each year.

In the course of its development in the past 20 years, the TEM has undergone two major revisions. The first TEM syllabus was officially published in 1994 in accordance with the requirements set in the 1989 Syllabus. A comprehensive validation study of the TEM from 1993 to 1996 brought about a number of changes to the TEM, the most noteworthy among which were the publication of the revised test syllabuses in 1997 and the establishment of quality control procedures for the design, development and administration of the TEM (The TEM Test Centre, 1997). The second major revision of the TEM took place in 2004 as a response to a further revision of the teaching syllabus in 2000. The revised version of the TEM features more integrative tasks, using lengthier and more authentic input materials for listening and reading. In comparison, the two oral tests have not undergone as many changes as TEM4 and TEM8. A major reform was to supplement the tape-mediated format with computer-based spoken English tests in 2008.

In TEM 4&8, listening, reading, writing and speaking are tested through a variety of test methods, ranging from the traditional single-answer four-option multiple choice question (hereinafter MCQ) to more integrative formats such as dictation, note-taking, and gap-filling. Translation is also tested in TEM8 as a major language skill. In addition, contributory language knowledge and skills are tested through such tasks as Grammar and Vocabulary, Cloze, and Proofreading. From 2005, a new component General Knowledge was added to TEM8 to better reflect the requirements of the teaching syllabus (Zou, 2003). For the constructed-response tasks of writing, translation and speaking, the TEM adopts the analytic approach of marking. Writing performance is judged against the criteria of content (relevance and completeness) and language (grammar and vocabulary, and appropriateness). Translation performance is evaluated against the criteria of faithfulness to the source text and quality of language. The assessment criteria for speaking are effectiveness of task completion (20% for each task) and quality of language (20% for pronunciation and intonation, and 20% for

grammar and vocabulary).

The appraisal of the TEM is largely based on the test qualities and practices prescribed in Standards for Educational and Psychological Testing (hereafter the Standards) (AERA/APA/NCME, 1999). As a large-scale standardized English language test which is designed, administered and used in the Chinese context, the appraisal will not be confined to examining the validity and reliability of the TEM alone but will also include practicality. Furthermore, due to the fact that the TEM serves the dual purpose of both measuring students' English language abilities and more positively affecting English teaching and learning (NACFLT, 2004a, 2004b) and also because the TEM enjoys wide social recognition as a major benchmark of English proficiency, an important part of the appraisal will be extended to the washback and impact of the TEM.

Shortly after TEM4 and TEM8 were officially launched, a Sino-British cooperative validation study was conducted from 1993 to 1996, which investigated reliability and other aspects of TEM validity (The TEM Test Centre, 1997). The internal consistency coefficients were reported to be acceptably high with TEM4 at 0.854 ($n = 13,675$) and TEM8 at 0.801 ($n = 6325$). Principal component analyses were run on both the item level and the subcomponent level, suggesting a general competence factor and a comprehension versus production factor. The study concluded that 'the TEM tests are reasonably reliable and valid tests that are set at an appropriate (difficulty) level as defined in the test specification' (The TEM Test Centre, 1997, p. 63). Furthermore, the 3-year validation effort highlighted a number of problems facing the TEM and proposed ways to address these problems in the test's future development. A valuable outcome of this study, as explained earlier, was the establishment of a host of standardized quality control procedures for the TEM.