TEM

考试效度研究

The Test for English Majors (TEM) Validation Study

上海外国语大学 TEM 考试中心

The TEM Testing Centre

Shanghai International Studies University

上海外语教育出版社

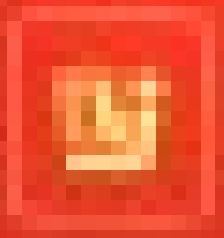
SHANGHAI FOREIGN LANGUAGE EDUCATION PRESS

1997

TEM

考试攻略研究

[redacted]



TEM 考 试 效 度 研 究

The Test for English Majors (TEM) Validation Study

上海外国语大学 TEM 考试中心

The TEM Testing Centre
Shanghai International Studies University

上海外语教育出版社

SHANGHAI FOREIGN LANGUAGE EDUCATION PRESS

1997

TEM考试效度研究

The Test for English Majors (TEM) Validation Study

总撰稿：邹申

上海外语教育出版社出版发行
（上海外国语大学内）
上海市印刷七厂印刷
新华书店上海发行所经销

开本 787×1092 1/16 20.5印张 4插页 507千字
1998年3月第1版 1998年3月第1次印刷
印数：2 000册
ISBN 7-81046-279-2
G·603 定价：28.00元.

总撰稿者:	邹　申
撰稿及审稿者:	Dr.Cyril J. Weir
	Ms. Rita Green
撰稿者:	李基安　李慧琳
(按姓氏笔划为序)	陈汉生　张艳莉
	黄素华

前　言

1993年9月，在国家教育委员会有关部门和高等院校外语专业教学指导委员会英语组的关心与支持下，在英国文化委员会的协助下，高等院校英语专业学生考试(Test for English Majors，以下简称TEM)效度研究项目在上海外国语大学英语学院TEM考试中心正式立项。项目的研究目的是通过一系列科学的测量方法，对TEM考试现有两个等级(TEM4和TEM8)进行多方位、多角度的评估，并在此基础上提出改进考试的建议，进一步提高TEM考试的科学性与合理性。

TEM考试效度研究项目历时3年。期间，项目组成员对TEM考试的内部效度、外部效度及理论效度进行了论证，开展了大量的调查与统计工作。在广泛调研的基础上，项目组完成了TEM考试效度研究报告。研究报告阐述了TEM考试的形成与发展过程，对其性质、目的、形式、内容、命题、实施、评分等等作了较为细致的介绍；研究报告总结了项目研究的经过，对TEM考试作出了客观的结论并提出了完善TEM考试的各项建议(绝大部分建议已在近年内付诸实施)；研究报告着眼于考试效度评估的科学性，调查数据翔实，评估方法先进，结论与建议部分合理，对同类研究具有较高的参考价值。

TEM考试效度研究是TEM考试发展进程中一项关键性工作。它标志着TEM考试在科学化、合理化的过程中又向前迈出了实质性的一步。这对不断提高TEM考试的权威性，使之成为衡量英语专业学生英语语言能力的有效手段，具有积极的意义。

在项目进行过程中，我们得到了有关学校、专家、教师及学生的积极配合以及大力支持，在此谨表示衷心的谢意。

上海外国语大学的朱嫣华、张伊兴、冯慎宇、何琮周等同志也参与了项目的部分工作。

上海外国语大学英语学院TEM考试中心

1997年4月

Acknowledgements

The project team wish to thank the following:

Foreign Languages Section, Department of Higher Education, SEdC;

The British Council;

Shanghai Foreign Language Education Press;

The National ELT Committee;

Members of the College of English Language and Literature at Shanghai International Studies University.

In particular our thanks go to:

Professor Dai Weidong	President of SISU;
Dr. C.J.Weir	Project external consultant (CALS, Reading University);
Ms. Rita Green	B.C. resident specialist (1993 – 1995; 1995 – , CALS, Reading University);
Professor He Zhaoxiong	Dean of CELL, SISU;
Ms. Barbara Wickham	Second Secretary, British Embassy, Beijing (– 1995); Cultural Consul, British Consulate-General, Shanghai (1995 –);
Dr. Tony Woods	Project external consultant;

and all the universities, teachers, administrators, and students who were involved in our three-year validation study.

Acronyms

AIID	alpha if item deleted
BC	British Council
BFSU	Beijing Foreign Studies University
BNU	Beijing Normal University
CALS	Centre for Applied Language Studies
CELL	College of English Language and Literature
CET	College English Test
CITC	corrected item-total correlation
DONGNAN	Dongnan University, Nanjing, China
ECIU	East China Industrial University, Shanghai
ECNU	East China Normal University, Shanghai
ELT	English language teaching
ELTS	English Language Testing System
FACMEAN	mean facility value
FUDAN	Fudan University, Shanghai
FV	facility value
GIFL	Guangzhou Institute of Foreign Languages
IDP	item discrimination pattern
JIAO TONG	Shanghai Jiaotong University
LFLI	Luoyang Foreign Languages Institute
MAX	maximum score
MCQ	multiple-choice question
MIN	minimum score
MINS	minutes
n/NO	number
ODA	Overseas Development Administration
PAR	parallel tests
PCA	principal component analysis
QINGHUA	Qinghua University, Beijing
SD	standard deviation
SEdC	State Education Commission, China
SISU	Shanghai International Studies University

SMI	Shanghai Maritime Institute
SPSS	Statistical Package for Social Sciences
Std Dev	standard deviation
SUDA	Suzhou University
TEEP	Test of English for Educational Purposes
TEM	Test for English Majors

Executive Summary

* Full specifications for the TEM tests have been developed.
* Sample tests are now available.
* An item writer training pack has been developed. Training will be conducted on a more regular basis in the future.
* A system has been established for trialling and moderating test items. These are now trialled a year in advance with 25% extra items as a safety margin.
* Appropriate marking schemes and marker standardization procedures have now been developed. They need to be vigorously implemented before each test administration.
* A system for monitoring markers has been established but more effort needs to be made to improve it further.
* A set of administrative procedures for running the tests has been developed but there is evidence that more attention needs to be paid to their implementation.
* A computer based system for statistical analysis of test data has been developed.
* Test development and statistical expertise has been developed in SISU. Efforts need to be made to sustain and further develop the testing team.
* An annual test report is now made of TEM4 and TEM8.
* The internal statistical analysis has shown:

    * Both TEM4 and TEM8 have exhibited moderate internal consistency with TEM4 at .85 and TEM8 at .80;
    * Components in each test can be differentiated from each other;
    * TEM4 is performing acceptably. It has now a 60% pass rate at a 60% cut-off point;
    * A large proportion of TEM8 candidates are not operating at the level of TEM8 with a 30% pass rate at a 60% cut-off point; TEM8 matches the specifications, therefore either changes need to be made to the syllabus or the students need to be better prepared for the test.

* The external statistical analysis has shown:

    * Only a moderate correlation between TEM4 and the parallel version PAR4 (.63) and between TEM8 and PAR8 (.74). Efforts need to be made to improve this situation;
    * A low correlation between TEM4 and the external criterion TEEP (.49), but a higher

one between TEM8 and TEEP at .78 for whose candidates it was more appropriate;

* A correlation of .47 was obtained between TEM4 teachers' assessments and the students' total scores;
* TEM4 appears to be a reasonable indicator of future English language performance (.71) in TEM8;
* Feedback from the questionnaires indicates the tests have high validity in the view of the teachers:

TEM4

More than 80% of the teachers felt that TEM4 was a good test of their students' English proficiency with 70% indicating it had a positive effect on teaching;

TEM8

More than 80% of the teachers felt that TEM8 was a satisfactory test of their students' English proficiency and more than 70% felt it had a good effect on teaching.

There is however still room for improvement.

* Feedback from the questionnaires indicates that TEM4 candidates express higher satisfaction with the test than TEM8. The low pass rate of the latter may have contributed to this.
* A number of components of both TEM4 and TEM8 can be reduced in size without affecting their internal consistency or external validity.
* There must be some concern over the low pass rate in dictation (TEM4) and listening, gap-filling and proofreading (TEM8). There is evidence from the questionnaire data that candidates are insufficiently trained in these components and attention needs to be given to this in teaching. There is also evidence that these components can be improved by the test developers.

Conclusions and Recommendations

1 Test Specifications

The test specifications have now been finalised in the light of the evidence emerging from the validation study and will provide a valuable guide for teachers, students, items writers and the test development team. [See 3.1 & Appendix V]

2 Test Size

The study showed that the size of a number of components in both TEM4 and TEM8 can be reduced without affecting the level of internal consistency or external validity. [See 4.1.1.3 – 4.1.1.6; 4.2.1.2 – 4.2.1.4]

3 Pass Rate

Just over 60% of candidates pass TEM4, which is satisfactory. [See 4.1.3] The current pass rate for TEM8 [See 4.2.3] is barely half of that (30% in 1995). This suggests a serious mismatch between the difficulty of the test and the preparation of the candidates. An examination of the tests' specification has convinced the TEM Project Team that the test is for the most part written correctly to meet its stated objectives. There is clear evidence from the questionnaires to suggest that the candidates may not all be adequately trained to meet the demands of the test.

Under no circumstances should the pass rate be increased simply by decreasing the pass mark. There can be no justification for this and such a step would leave the test open to serious criticism.

Assuming the quality of individual items can be assured, there are two ways out of the dilemma. Either the test specification could be altered to demand a lower level of performance or the current student population needs to be better prepared to meet the present standard set.

4 Item Writing

The procedures for item writing need to be tightened up. Reading and listening passages need to be mindmapped so items test main ideas and important details. [See 4.2.1.1; 4.2.1.4 & Appendix XVI]

Training packs for designing reading and listening items were developed in SISU during the consultancy visit in October 1995. [See 6.2 & Appendix XVII] These need to be operationalised to further enhance the validity of the items in the TEM tests.

The attrition rate of item writers still needs attention. However, there are signs of stability,

but the situation needs to be kept under constant scrutiny. [See 6.2]

Item writing workshops took place during the October 1995 consultancy visit in SISU. [See 6.5.1.4 & Appendix XX] These now need to be expanded to include item writers from outside Shanghai.

5 Trialling

It is clear that weak items in the past have had a deleterious effect on the measured performance of those close to the pass grade boundary. Ensuring that such items are eliminated in the future will in itself tend to increase the pass rate. [See 4.1.1 & 4.2.1]

The test has suffered in the past from a failure to ensure that items included in the final version have been systematically trialled and analyzed. A system has now been set up whereby all the items destined for a test are trialled at least a year in advance together with an additional 25% as a safety margin. The procedures for analysing items and for selecting those which function well have now been established. [See 6.1－6.4; 6.7]

Item analysis will enable items in the listening and reading sections to be ordered in terms of difficulty which again should serve to improve the performance of candidates. [See 6.2－6.4]

6 Analysis and Computing

With the attachment of two members of the team to Reading in the summer of 1995, a strong basis was established for the development of analytic and computing skills at TEM. During the October 1995 visit an additional member of the team was brought up to the same standard as the others. [See 6.1 & 6.7]

The team is now capable of using SPSS to obtain all the statistics necessary for the administration and analysis of the tests with the exception of test equating.

7 Test Equating

Any decisions on the merits of equating must be taken with due regard to the availability of a sufficient number of suitable students and to the existence of two forms of the test in which all the items are of sufficient quality. Over time the latter will be assured when systems that have been developed in the validation study are in place and are being properly monitored. As regards the number of students required it is recommended that test equating should be carried out on not less than 400 students from the middle and top ranges. This makes it an unlikely prospect for TEM8 given the small test population. It may be considered at some time in the future for TEM4. At such a stage the statistical software developed for CET (Central Statistical Package) should be available for the TEM tests.

8 Validation Study Statistical Results

Both TEM4-95 and TEM8-95 exhibited moderate internal consistency estimates. [See 4.1.3; 4.2.3 & Appendices IX, X]

The paired components factor analyses also provided some evidence that the components were differentiated from each other. [See 4.1.2.2; 4.2.2.2 & Appendices Ⅸ, Ⅹ]

In the case of TEM4-95, the correlation estimates with external test measures are unreliable because of the narrow variability in performance of the relatively good students available for this part of the study. [See 5.1.1.3; 5.1.2.4 & Tables 1 and 3, Appendix Ⅺ]

In the case of TEM8-95, correlations with external test measures were moderate (.7+). [See 5.1.1.4; 5.1.2.4 & Tables 2 and 4, Appendix Ⅺ]

An analysis run on a total of 446 completed assessments resulted in a correlation of .4715 between the total TEM4 score and the teachers' subjective assessments. [See 5.1.3]However, this correlation is attenuated by the aggregation of the different internal scales of different teachers. In other words, a more accurate estimate of the true correlation would be obtained by working on an individual teacher basis and then averaging the resulting statistics across teachers. Unfortunately, that specific level of information was not available in the data set.

The correlation between SISU's TEM8-95 students' scores and their TEM4-93 scores was at .7122, which indicated that the earlier TEM4 performance was a moderate indicator of later performance on TEM8. [See 5.2]

The major weakness in the TEM test is the assessment of the subjectively scored components, writing, translation and to a lesser extent dictation. Work still needs to be done to improve the standardisation of examiners. A standardisation procedure for the assessment of writing was introduced and marker reliability study set up. [See 6.5 & Appendices ⅩⅤ - ⅩⅥ] The results of the latter are encouraging and compare favourably with the results in the ELTS validation study.

9 Questionnaire Data

TEM4 Teachers

* 81% felt TEM4 was a good test of their students' English proficiency;
* 78% felt TEM4 was a good indicator of their students' performance;
* 75% felt the number of marks allocated to TEM4 components were appropriate;
* 70% felt TEM4 had a good effect on teaching.

TEM8 Teachers

* 86% felt TEM8 was a satisfactory test of their students' English proficiency;
* 80% felt the TEM8 components were an adequate measure of their students' skills;
* 76% felt TEM8 had a good effect on teaching;
* 72% felt the number of marks allocated to TEM8 components were appropriate.

[See 5.3.1.2; 5.3.2.2 and Appendices Ⅺ - ⅩⅥ]

10 Test Itself

An overall satisfaction was expressed with the TEM tests particularly by candidates and teachers in TEM4. [See 5.3.1.2] There was less satisfaction with TEM8 where a majority of the candidates felt they had done badly in many components. [See 5.3.2.2] There is evidence

from the questionnaires to suggest that this may be because TEM8 candidates are insufficiently trained in some of the components.

Teachers particularly need to focus on those skills required for TEM4 dictation and TEM8 listening, gap-filling and proofreading. In addition steps should be taken to improve these components by the testing team.

Listening presents some problems especially in TEM8:

* Need to improve playback facilities—there is strong evidence that this is interfering with performance; [See 4.1.1.3; 4.2.1.1]
* Need to improve listenability by recording in front of a live audience to make the performance more authentic; [See 4.2.1.1]
* In TEM8, gap-filling should come immediately after listening to the tape; [See 4.2.1.3]
* Number of items in gap-filling can be reduced from 20 to 10.

These measures in TEM8 might bring the means of both listening (40% in 1995) and gap-filling (39% in 1995) more in line with the average for reading (61% in 1995).

Improvements in the listenability of the recording and playback equipment might also help with dictation in TEM4 where around 1,500 candidates (one tenth of the 1995 test population) scored zero. It has the lowest mean score in TEM4 (47%). [See 4.1.2]

Proofreading in TEM8 is far too difficult, with the mean at 34%, which is the lowest in TEM8. A more principled basis for deleting the items in proofreading is needed. Deletions should be based on common errors students make in writing rather than idiosyncratic whim. There is no justification for such a large number of items in so short a passage. Statistical data from the study supports a reduction in the number of items from 20 to 10. [See 4.2.1.3]

In TEM8 serious thought needs to be given to increasing the number of writing tasks to two to enhance validity and reliability in testing this skill. [See 4.2.1.6]

Contents