How is language proficiency measured: which aspects of proficiency do tests measure and how? Using test results: how to determine language requirements and interpret results?

Kuinka kielitaitoa mitataan: mitä kielitaidon osa-alueita testit mittaavat ja kuinka ne mittaavat? Testitulosten hyödyntäminen: kuinka määritellä tulosrajoja ja kuinka tulkitta tuloksia?

Testien käyttö korkeakoulujen opiskelijavalinnoissa
Seminaari 17.11.2010
Ari Huhta
Centre for Applied Language Studies
University of Jyväskylä
Outline

Examples of international English language tests
How have they changed?
What do they measure?
How do they measure?
How is their quality ensured?
How to distinguish between low vs. high quality?
Comparability of different tests – challenges
Comparison via CEFR – examples and issues
Using test results – minimum language requirements
International English language tests 1

- Educational Testing Service (ETS) tests:
  - TOEFL iBT (Test of English as a Foreign Language, Internet-based Test)
  - TOEFL PBT (TOEFL Paper-based Test)
  - (TOEIC, Test of English for International Communication)
- IELTS, International English Language Testing System
- Cambridge ESOL tests:
  - CPE, Certificate of Proficiency in English
  - CAE, Certificate in Advanced English
  - BULATS, Business Language Testing Service
  - (FCE, PET, KET)
International English language tests 2

• Pearson tests:
  • Pearson Test of English Academic
  • Pearson Test of English General

• Michigan tests:
  • ECPE, Examination for the Proficiency in English
  • (The Examination for the Certificate of Competency in English)

• (Oxford Online Placement Test)
Changes in international English tests since 1970/1980s

• More communicative (view of language)
  • major revisions (e.g., Cambridge exams; ELTS → IELTS in the 1980s; TOEFL → TOEFL iBT in the 2000s)

• Role of speaking and also writing has increased
  • more authentic tasks, more extensive, given more weight (Cambridge exams); become integral part of exams (e.g. TOEFL iBT)

• Integration of skills
  • e.g. TOEFL iBT

• Tests for specific purposes more common
  • e.g. for business, aviation purposes

• Computerisation
  • administration, even scoring (e.g. TOEFL, Pearson tests)
What do the tests measure? TOEFL iBT

Extracts:
The TOEFL iBT test measures your ability to use and understand English at the university level. And it evaluates how well you combine your listening, reading, speaking and writing skills to perform academic tasks.

Listen to lectures, classroom discussions and conversations, then answer questions.
Read passages from academic texts and answer questions.
Express an opinion on a familiar topic; speak based on reading and listening tasks.
Write essay responses based on reading and listening tasks; support an opinion in writing.
What do the tests measure? IELTS

Extracts:

Academic – tests a person’s ability to study in English at undergraduate or postgraduate level.

General Training – this module is suitable for people who are going to an English-speaking country to work or train at below undergraduate level.

Texts are taken from books, journals, magazines and newspapers and have been written for a non-specialist audience. All the topics are of general interest. They deal with issues which are interesting, recognisably appropriate and accessible to candidates entering undergraduate or postgraduate courses or seeking professional registration. The passages may be written in a variety of styles, for example narrative, descriptive or discursive/argumentative. At least one text contains detailed logical argument. Texts may contain non-verbal materials such as diagrams, graphs or illustrations.
What do the tests measure? Pearson Academic 1

The new test will measure overall English language competency, in addition to providing feedback on Reading, Writing, Listening, and Speaking skills. The PTE Academic Score Report will include the availability of Communicative Skills and Enabling Skills.

We test academic English by carefully examining where the language will be used, situations in academic settings. We defined a list of skills or abilities which are important in university academic settings and defined reading and listening texts that students are likely to encounter in academic settings. When selecting texts, item writers always select texts that students from the target population would be likely to read or hear. The international character is ensured by selecting texts and settings encountered in Australia, Canada, New Zealand, the UK and the US.
Texts used for PTE Academic are taken from real-life situations as students will encounter in an academic environment. In academic settings, students have to understand a wide variety of listening texts across different situations. Reading texts appropriate for PTE Academic include study texts of academic interest and texts related to all aspects of student life. Test materials are extracted from published sources such as textbooks or websites containing useful information for university level students. Study texts of academic interest include historical biographies and narratives, academic articles, book reviews, commentaries, editorials, critical essays, articles and reports, science reports or summaries, scientific articles written for a general academic public, and journal articles. Reading texts related to student life include instructions, course outlines, grant applications, notices and timetables. Students also need to follow different types of lectures (for example audio, video, audio-visual) delivered with different accents and at varying delivery speed. Listening texts appropriate for PTE Academic include lectures, study texts of academic interest and texts related to all aspects of student life. Listening texts are extracted from sources such as websites containing useful information for students, or based on actual speech samples collected from universities and colleges.
What do the tests measure? Cambridge CPE

use English to advise on, or talk about complex or sensitive issues

understand the finer points of documents, correspondence and reports

understand with ease virtually everything they hear and read

make accurate and complete notes during a presentation

understand colloquial asides

talk about complex and sensitive issues without awkwardness

express themselves precisely and fluently
How do tests measure proficiency?

Typically a variety of contexts, situations, texts, task types
Often try to simulate real-life tasks (cf. communicative tests)
Computers increasingly used
  • test delivery
  • scoring
  • authenticity?
Ensuring quality in major English language testing systems

- Multi-stage, iterative process that combines theoretical considerations and empirical / psychometric research
- Sari Luoma’s PhD thesis 2001 illustrates this by using TOEFL and IELTS as examples

“The aim of the thesis is to clarify the principles by which test developers can and should build quality into their tests and the practical activities that these principles entail.”


http://www.solki.jyu.fi/vanhat/Luoma_Sari_2001_PhD_manuscript1.pdf
Grounding: test purpose, development brief, resources, constraints

Initial test development
- define construct
- develop specifications to operationalize construct
- develop and revise tasks
- develop and revise assessment criteria

Validation
record process and products of initial test development
- evaluate products against aims and state of theory
- examine values
- identify plausible rival hypotheses
- plan future validation research

Pre-publication piloting
- pre-testing on various sizes of groups
- data analysis

Development of admin procedures
- selection and training of raters
- training of admin staff
- creation of admin documents

Review and revision of test materials
- tasks
- scoring rubrics and criteria
- administration instructions
Operational test development
- develop new tasks and items:
  - commissioning
  - vetting
  - editing
  - pretest construction
  - pretesting
  - reviewing
  - accepting, revising, rejecting
  - store new items in item bank
  - assemble new forms

Operational administration
- administer the test according to set procedures
- gather performances
- get performances rated according to set procedures
- report scores

Monitoring & maintenance
- analyse and archive tasks for future use
- monitor rater and test form reliability
- obtain regular feedback
- evaluate the test’s performance and any need for revision

Ongoing empirical validation
- assess appropriacy of proposed test use
- gather, analyse and interpret evidence for score meaning
- develop specific validity cases
- test rival hypotheses
- propose revisions
- assess impact
- examine values
How to distinguish between low and high quality in language testing?
Quality does not come easily

High-quality language testing requires expertise

What to look for in the institution, company, board, etc responsible for the test?

• Does it have experts in all the key areas of test development (e.g. content / language / language learning / item writing / psychometric experts)?
  • Can you find that information?

• Developing expertise takes years to develop: a company, institution, etc of whom nobody in language testing has heard about before is unlikely to be able to provide high-quality service
Quality does not come cheaply

High-quality language testing requires resources

An institution, company or board that maintains an even moderately large-scale testing system needs to have

• enough staff for the various activities
• enough funding (public or private)

The number of institutions etc that accept the test is usually also a sign of quality or lack of it
Quality needs to be demonstrated

All tests claim to be of high quality but not all tests can show that they are

Look for publicly available information:
- documents, brochures, demos, sample material, etc that inform different user groups of the tests and how to interpret test scores
- reports on test results from different administrations (e.g. score distributions overall and by major test taker groups, reliabilities)
- research into different aspects of tests and test use
Examples of research to back claims about quality: IELTS

Predictive validity of the IELTS Listening test as an indicator of student coping ability in English-medium undergraduate courses in Spain

Examining academic spoken genres in university classrooms and their implications for the IELTS speaking test

Examiner use and views of the revised IELTS pronunciation descriptors

To what extent does IELTS encourage communicative language teaching in Chinese IELTS classrooms?
Examples of research to back claims about quality: TOEFL

Test Takers’ Attitudes About the TOEFL iBT

Does Content Knowledge Affect TOEFL iBT Reading Performance?

Toward an Understanding of the Role of Speech Recognition in Non-Native Speech Assessment

The Impact of Changes in the TOEFL Examination on Teaching and Learning in Central and Eastern Europe
Examples of research to back claims about quality: Pearson Academic

Investigating the Lexical Validity of PTE Academic

Aligning PTE Academic Test Scores to the CEF

Standardizing Rater Performance

Two experiments on automatic scoring of spoken language proficiency
Comparability of different tests: challenges

• Content & target groups
• Skills / areas covered / emphasised (underlying theory)
• Difficulty
• Testing methods
• Rating criteria & scales
• Procedures of test development and analysis
• Overall quality of procedures and outcomes
Comparability: content & target groups

Test of general language vs. Test for specific purposes
• General English vs. Academic English vs. Engineering English?
  • PTE General vs. PTE Academic / TOEFLiBT / IELTS vs. Aviation English tests
• Test of English for medical purposes vs. Test of English for military purposes

Adult learners’ tests vs. Young learners’ tests
• First Certificate vs. Young Learners’ Test of English

Also concerns frameworks such as the CEFR
• mainly general, mainly for adults
Comparability: Skills / areas covered or emphasized

Spoken tests vs. Written tests
• old TOEFL vs. Cambridge exams
• old vs. new TOEFL (PBT vs. iBT)
Comparability: difficulty

- Easiest / most common approach to comparing different tests
- Often the only approach !?
- Why so common?
  - Empirical difficulty is fairly straightforward to find out
  - Direct comparison of two or more tests or via a common yardstick
    - English-speaking union’s scale
    - CEFR scale
  - Standard setting usually focuses only on difficulty
- Content etc often ignored
  - Difficulty is based on so many different factors
  - Difficult for whom?
Comparability: testing methods

Paper-based vs. Tape-mediated / Computer mediated vs. Totally computerised test
• IELTS vs. TOEFLiBT vs. Pearson Academic

Receptive test methods vs. Productive test methods
Use of only a few vs. many different test methods

Test methods affect test results to some extent
• familiarity with the methods used
• use of very common vs. very unusual methods
Comparability: rating criteria & scales

Tests of writing & speaking

Holistic vs. Analytic rating

Only a few vs. many criteria and scales

Different criteria
  • Accuracy of language vs. Fluency, communication
Comparability: test development and analyses

Nature and scope of overall testing framework, test specifications and item writing guidelines

Pilot testing vs. No pilot testing

Comparability, equating etc of supposedly parallel test versions?

• stability / comparability of test results across test administrations
  • See also criterion vs. norm-referenced tests (Finnish Matriculation examination)

Item analyses?

Other test analyses?
Comparability: Overall quality

All aspects of testing process affect quality
• planning, development, training, administration, analyses, infrastructure, research, ...

Quality of language tests worldwide varies enormously from useless, even harmful thrash to high-quality examples of good professional and ethical practice
Comparability via Common European Framework of Reference

- Linkage to the CEFR seems to be a must for most language tests, at least in Europe
- Linkage to the 6-point CEFR scale of proficiency
<table>
<thead>
<tr>
<th>Level</th>
<th>Test</th>
<th>Language Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>C2</td>
<td>ENGLISH TEST Z</td>
<td>GERMAN TEST C</td>
</tr>
<tr>
<td>C1</td>
<td>ENGLISH TEST Y</td>
<td>GERMAN TEST A</td>
</tr>
<tr>
<td>B2</td>
<td>ENGLISH TEST X</td>
<td>GERMAN TEST B</td>
</tr>
<tr>
<td>B1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Quality & accuracy of linkage

... varies

Claims & educated guesses

Empirical linkage
• Council of Europe’s Manual
• Other guidelines on linking & standard setting

→ Does the testing organisation provide any evidence about its claims on linkage with the CEFR? (empirical studies, preferably more than one type of linking / standard setting)
UNIVERSITY OF JYVÄSKYLÄ

Linking with CEFR – IELTS

http://www.ielts.org/researchers/common_european_framework.aspx
Linking with CEFR – Cambridge ESOL

http://www.cambridgeesol.org/exams/exams-info/cefr.html
大学: Jyväskylä 大学

标题: 链接 CEFR - TOEFL iBT


<table>
<thead>
<tr>
<th>Test Section</th>
<th>Total Score Scale Range</th>
<th>A1</th>
<th>A2</th>
<th>Minimum Score</th>
<th>B1</th>
<th>B2</th>
<th>C1</th>
<th>C2</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOEFL iBT Total</td>
<td>0 – 120</td>
<td></td>
<td></td>
<td>57 – 86</td>
<td>8</td>
<td>22</td>
<td>28</td>
<td>29</td>
</tr>
<tr>
<td>TOEFL iBT Reading</td>
<td>0 – 30</td>
<td></td>
<td></td>
<td>8</td>
<td>13</td>
<td>21</td>
<td>26</td>
<td></td>
</tr>
<tr>
<td>TOEFL iBT Listening</td>
<td>0 – 30</td>
<td></td>
<td></td>
<td>13</td>
<td>19</td>
<td>23</td>
<td>28</td>
<td></td>
</tr>
<tr>
<td>TOEFL iBT Speaking</td>
<td>0 – 30</td>
<td>8</td>
<td>11</td>
<td>17</td>
<td>21</td>
<td>28</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TOEFL iBT Writing</td>
<td>0 – 30</td>
<td></td>
<td></td>
<td>11</td>
<td>17</td>
<td>21</td>
<td>28</td>
<td></td>
</tr>
</tbody>
</table>
UNIVERSITY OF JYVÄSKYLÄ

Linking with CEFR – Pearson Academic

http://www.pearsonpte.com/Testme/Pages/Scores.aspx
Direct score comparisons – TOEFL iBT & IELTS

- Study of 1100 test takers who took both tests

<table>
<thead>
<tr>
<th>TOTAL SCORES</th>
<th>TOEFL</th>
<th>IELTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>35</td>
<td>5.0</td>
<td></td>
</tr>
<tr>
<td>46</td>
<td>5.5</td>
<td></td>
</tr>
<tr>
<td>60</td>
<td>6.0</td>
<td></td>
</tr>
<tr>
<td>79</td>
<td>6.5</td>
<td></td>
</tr>
<tr>
<td>94</td>
<td>7.0</td>
<td></td>
</tr>
<tr>
<td>102</td>
<td>7.5</td>
<td></td>
</tr>
<tr>
<td>110</td>
<td>8.0</td>
<td></td>
</tr>
<tr>
<td>115</td>
<td>8.5</td>
<td></td>
</tr>
</tbody>
</table>
Issues with linkage to the CEFR

• Different examination bodies appear to have somewhat different views of how their examinations relate to the CEFR and each other

• TOEFL iBT & Pearson Academic vs. IELTS & Cambridge ESOL

• See John de Jong’s presentation at www.ealta.eu.org → Resources → Conference presentations at the 2009 conference in Turku
Using test results: how to determine what is enough? 1

How to decide on what score to require on each language test accepted by the educational institution?

Depends partly on the consequences of decisions
  • is the goal to minimise the number of false positives, i.e., decisions to admit applicants who turn out to have too low proficiency?
  • is the goal to minimise false negatives, i.e. decisions to reject applicants with adequate proficiency?

Depends partly on resources to deal with false positives
  • is language tuition available for those who turn out to need to improve their proficiency?
Using test results: how to determine what is enough? 2

Analyse carefully the language requirements of the programme / subject matter / department / faculty in question
• e.g. analyses of text books, interviews with teaching staff
• do this overall and skill by skill
Compare those with what you know about the tests that you accept as proof of proficiency

Short-term solution: study what minimum scores other similar programmes etc require in other countries and use the same
<table>
<thead>
<tr>
<th>Band</th>
<th>Linguistically demanding academic courses</th>
<th>Linguistically less demanding academic courses</th>
<th>Linguistically demanding training courses</th>
<th>Linguistically less demanding training courses</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.5 – 9.0</td>
<td>Acceptable</td>
<td>Acceptable</td>
<td>Acceptable</td>
<td>Acceptable</td>
</tr>
<tr>
<td>7.0</td>
<td>Probably acceptable</td>
<td>Acceptable</td>
<td>Acceptable</td>
<td>Acceptable</td>
</tr>
<tr>
<td>6.5</td>
<td>English study needed</td>
<td>Probably acceptable</td>
<td>Acceptable</td>
<td>Acceptable</td>
</tr>
<tr>
<td>6.0</td>
<td>English study needed</td>
<td>English study needed</td>
<td>Probably acceptable</td>
<td>Acceptable</td>
</tr>
<tr>
<td>5.5</td>
<td>English study needed</td>
<td>English study needed</td>
<td>English study needed</td>
<td>Probably acceptable</td>
</tr>
</tbody>
</table>
Examples of current requirements: Finland

Aalto University (for master’s or doctoral studies):
• IELTS: 6.5 (and 5.5 in writing)
• TOEFL: 580 / 237 / 92 (and 22 iBT / 4.0 PBT writing)
• CAE or CPE: pass grade A, B or C

Åbo Akademi (for master’s studies):
• IELTS Academic: 6.5 (minimum 5.5 in all sections)
• TOEFL: 575 / 90 (min 52 PBT or 17 iBT)

Savonia Polytechnic (degree programme in English):
• IELTS Academic: 6.0
• TOEFL: 550 / 213 / 79-80
• some national (Finnish) tests
Examples of current requirements: USA

For direct entry to an undergraduate degree:

- **TOEFL**
  - Low (e.g. Indiana Institute of Technology): 500 / 173 / 61
  - High (e.g. Alfred University): 550 / 213 / 80
- **IELTS**: from 5.0 (Elmira College) to 6.0 – 7.0 (Florida Southern College)

For direct entry to a Master’s degree:

- Low (Baldwin-Wallace College): 523 / 193 / 69
- High (Alfred University): 590 / 23 / 90

www.universitiesintheusa.com/english-level-requirements.html
Using test results: how to determine what is enough? 3

Long-term solution: monitor systematically if the minimum scores you decided for each test work in your institution

• Requires years of monitoring
• Requires enough information (group level & case studies)
Kiitos
Tack
Thank you

ari.huhta@jyu.fi