Performance-based Language Assessment: What Is It? Does It Work?

Bordin Chinda, PhD
Chiang Mai University, Thailand
Sapporo Gakuin University, Japan

1

Outline

- Basic language testing principles
- Different types and purposes of language tests
- Performance-based language assessment
- Teachers' reactions towards performance-based language assessment
- Performance-based assessment and national policies

2

Why language assessment?

for language teachers, testing and assessment are considered as the somewhat arcane province of "experts" and of marginal relevance to everyday classroom concerns

- Language assessments play a powerful role in many people's lives (e.g. students, immigrants)
- You may be working with language assessments in your professional life (like us!)
- You may be conducting research in language study (e.g. part of an action research)

3

The criterion

- Language testing is about making **inferences**.
- Test performances are the indicators of how a person would perform **similar tasks in the real-world setting**.
- The future 'real life' language use is referred to as the **criterion**.

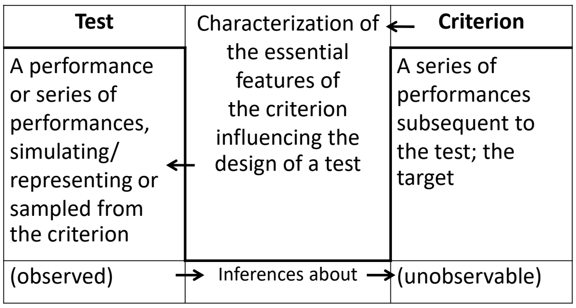
4

The criterion

- Test performances are used as the basis for making inferences about criterion performances.
- Although a criterion behaviour is the real object of interest, it can't be directly 'observed'.

5

The criterion



Test	Characterization of the essential features of the criterion influencing the design of a test	**Criterion**
A performance or series of performances, simulating/ representing or sampled from the criterion		A series of performances subsequent to the test; the target
(observed)	→ Inferences about →	(unobservable)

6

Direct VS indirect testing

- **Direct testing** requires the candidates to perform precisely the skill that we wish to measure.
- Direct testing is easier to carry out when it is intended to measure the productive skills of speaking and writing.

7

Direct VS indirect testing

- **Indirect testing** attempts to measure the abilities that underlie the skills in which we are interested.
- Indirect testing seems to offer the possibility of testing a representative sample of a finite number of abilities which underlie a potentially indefinite large number of manifestations of them.

8

Norm-referenced VS Criterion-referenced testing

Characteristic	**Norm-referenced**	**Criterion-referenced**
Type of Interpretation	Relative (A student's performance is compared to those of all other students in percentile terms.)	Absolute (A student's performance is compared only to the amount, or percentage, or material learned.)
Type of Measurement	To measure general language abilities or proficiencies	To measure specific objectives-based language points

9

Norm-referenced VS Criterion-referenced testing

Characteristic	**Norm-referenced**	**Criterion-referenced**
Purpose of Testing	Spread students out along a continuum of general abilities or proficiencies	Assess the amount of material known or learned by each student
Distribution of Scores	Normal distribution of scores around the mean	Varies; often non-normal.

10

Norm-referenced VS Criterion-referenced testing

Characteristic	**Norm-referenced**	**Criterion-referenced**
Test Structure	A few relatively long subtest with a variety of item contents	A series of short, well-defined subtests with similar item contents
Knowledge of Questions	Students have little or no idea of what content to expect in test items.	Students know exactly what content to expect in test items.

11

Traditional paper-and-pencil VS Performance testing

- **Traditional (paper-and-pencil) testing** emphasises the rank ordering of students
- in general promotes the idea of neutral, scientific measurement

12

Traditional paper-and-pencil VS Performance testing

- In traditional testing, the testing and teaching are separated activities conducted by separate groups of people of
- the students usually have no access to the criteria
- a single score is usually reported.

13

Traditional paper-and-pencil VS Performance testing

- **Performance/alternative/classroom-based assessment** is based on an investigation of developmental sequences in student learning, a sampling of **genuine performances** that reveal the underlying thinking processes, and the provision of **an opportunity for further learning**

14

Traditional paper-and-pencil VS Performance testing

- In performance assessment, assessment and teaching are integrated with active participation of the students

15

Traditional paper-and-pencil VS Performance testing

Performance-based assessment	***Traditional testing***
Fluency-focused	Accuracy-focused
Individual-focused	Group- or ‘norm’-focused
Achievement/progress focused	Proficiency-focused
Process-focused	Product-focused
Teachers’/student’s voices	Rule-makers’ voices
Leads to assessment **FOR** learning	Leads to ‘teaching to the test’

16

Test purposes

- Achievement test
- Proficiency test
- Diagnostic test
- Placement test

17

Achievement test

- **Achievement tests** are directly related to language courses to establish how successful individual students, group of students, or the courses themselves have been in achieving objectives
- There are two kinds: final achievement tests and progress achievement tests

18

Achievement test

- Final achievement tests are those administered at the end of a course of study
- The content of these tests must be related to the courses with which they are concerned

19

Achievement test

- Progress achievement tests are intended to measure the progress that students are making.
- These tests should be related to the objectives of the course.

20

Proficiency test

- **Proficiency tests** are designed to measure people's ability in language regardless of any training they may have had in that language
- The content of a proficiency test, therefore, is not based on the content or objectives of language course

21

Placement test

- **Placement tests** are intended to provide information will help to place students at the stage of the teaching program most appropriate to their abilities.
- These tests are used to assign students to classes at different levels.

22

Diagnostic test

- **Diagnostic tests** are used to identify learners' strengths and weaknesses.
- They are intended primarily to ascertain what learning still needs to take place.

23

Assessment Purposes

Summative

VS

Formative

24

Summative Assessment

- Provides a public outcome statement on performance at the end of a period of instruction for the benefits of all interested stakeholders
- Normally by a test or examination procedure

25

Formative Assessment

- Attend to the process of a program to provide immediate feedback which could lead to improvement

26

Assessment Purposes

	Formative	**Summative**
Purpose	To monitor and guide a process while it is still in progress	To judge the success of a process at its completion
Time	During the process	At the end of the process

27

Assessment Purposes

	Formative	**Summative**
Type	Informal observation, quizzes, homework, worksheets	Formal tests, projects, and term papers
Use	Improve and change a process while it is still going on	Judge the overall success of a process; e.g. grades

28

Assessment AND Learning

Assessment **OF** Learning

VS

Assessment **FOR** Learning

29

Assessment OF Learning

- Intended to certify learning and report to parents and students about students' progress
- Typically done at the end of something (e.g. a course)
- Takes the form of tests or exams
- The results are expressed symbolically, generally as grades

30

Assessment OF Learning

- Assesses the quantity and accuracy of student work
- Indicate which students are doing well and which ones are doing poorly

31

Assessment FOR Learning

- Teachers collect a wide range of data to modify the learning work for students
- Uses the insights that come from the process to design the next steps in instruction
- Marking highlights each student's strengths and weaknesses
- Provides students with feedback that will further their learning

32

Formative – HOW?!?!

Four types of action

- Questioning
- Feedback through marking
- Peer-and self-assessment by students
- The formative use of summative test

33

Questioning

- effort has to be spent in framing questions which explore issues that are important to the development of students' understanding
- wait time has to be increased (to several seconds) to allow students time to think
  - everyone should be expected to have an answer and contribute to the discussion
  - all answers (+/-) can be used to develop understanding
  - the aim is thoughtful improvement (rather than getting it right first time)

34

Questioning

- follow-up activities have to be rich – extend students' understanding
- the only point of asking questions is to raise issues about which the teacher needs information or about which the students need to think

35

Feedback by marking

- identifies what has been done well and what sill needs improvement
- gives guidance on how to make that improvement
- provides opportunities for students to follow up comments

36

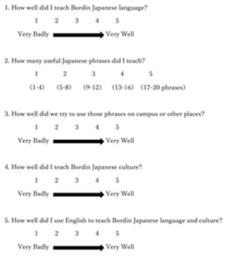
Peer- and self-assessment

- The criteria for evaluation must be made transparent to students
- Students
  - should be taught the habits and skills of collaboration in peer-assessment
  - should be encouraged to keep in mind the aims of their work and assess their own progress

37

Peer- and self-assessment

Teaching Bordin Japanese language and culture "Self-Assessment"



1. How well did I teach Bordin Japanese language?

1 2 3 4 5

Very Badly → Very Well

2. How many useful Japanese phrases did I teach?

1 2 3 4 5

(1-4) (5-8) (9-12) (13-16) (17-20 phrases)

3. How well did we try to use those phrases on campus or other places?

1 2 3 4 5

Very Badly → Very Well

4. How well did I teach Bordin Japanese culture?

1 2 3 4 5

Very Badly → Very Well

5. How well did I use English to teach Bordin Japanese language and culture?

1 2 3 4 5

Very Badly → Very Well

38

Communicative language testing

Communicative test-setting requirements

- The communication that is required of the students should be meaningful to the students as individuals.
- It should include functions of the language that are useful to them.
- In order for communication to be meaningful, it will probably be necessary to create a situation that is as authentic as possible.

39

Communicative language testing

Communicative test-setting requirements

- The students should encounter unpredictable language input and be put in a position where they must produce creative language output.
- Just like real life, students should be using all four language skills.

40

Communicative language testing

Communicative test-setting requirements
Meaningful communication
Authentic situation
Unpredictable language input
Creative language output
All language skills

41

Communicative language testing

Bases for ratings
Success in getting meaning across
Use focus rather than **usage**
New components to be rated

42

Performance assessment

- driving test
- Olympic diving
- an artist’s portfolio
- a surgeon’s 1st triple bypass operation
- use of a language (e.g., English)

43

Performance assessment

- the ability of candidates to perform particular tasks is assessed
- Tasks are designed to measure learners’ productive language skills through performances which allow learners to exhibit the kinds of language skills that may be required in a real-world context.

44

Performance assessment

There are three factors distinguishing performance tests from traditional tests of second language:

- there is a performance by the candidate
- the performance is judged using an agreed set of criteria (rubrics)
- there is a degree of authenticity of the assessment tasks

45

Performance assessment

A language test is said to be authentic when it mirrors as exactly as possible the content and skills under test, e.g.

- conversation test to test for conversation skills
- an essay test to test for writing essay skills

46

Performance assessment

Another important characteristic of performance-based assessment is a new type of interaction **between the rater and the scale (rubrics)**

47

Performance assessment

RATER
↓ ➔ RATING (SCORE)
SCALE
↓
PERFORMANCE
↑
INSTRUMENT
↑
CANDIDATE

48

Performance assessment

- In marking any performance-based assessment tasks, the markers/raters, or teachers in classrooms, are required to make **more complicated judgements** than the right-wrong decisions in multiple-choice where the candidate's responses can be marked as either 'correct' or 'incorrect'

49

Assessment criteria

- A rating scale (or rubrics in Am. Engl.) is a scale for the description of language proficiency consisting of a series of constructed levels against which a language learner's performance is judged.
- The levels or bands are commonly characterised in terms of what the learners can do with the language and their mastery of linguistic features.

50

Assessment criteria

- Rating scales also represent the most concrete statement of the construct being measured. The statements in rating scales are commonly referred to as descriptors which describe the level of performance required of candidates at each point on a proficiency scale.

51

Analytic vs. holistic rating scales

- With an analytic scale, raters are asked to judge several components of a performance separately as traits, criteria, or dimensions of performance. These components are divided so that they can be judged separately rather than giving a single score for the entire performance.

52

Analytic vs. holistic rating scales

Analytic scales are used when

- planning instruction to show relative strengths and weaknesses of a performance
- teaching students the nature of a quality performance
- giving detailed feedback,
- knowing how to precisely describe quality is more important than speed

53

Analytic vs. holistic rating scales

- with a holistic scale, raters are asked to give a judgement on a candidate's performance as a whole, or in other words, a single score for an entire performance based on an overall impression of a candidate's work

54

Analytic vs. holistic rating scales

Holistic scales are used when

- speed of scoring is more important than knowing precisely how to describe quality,
- the performances are simple
- when a quick snapshot of overall achievement is the objective

55

Analytic rating scales

	Beginning 1	Developing 2	Accomplished 3	Exemplary 4	Score
Criteria 1	Description reflecting beginning level of performance	Description reflecting movement toward mastery level of performance	Description reflecting achievement of mastery level of performance	Description reflecting highest level of performance	
Criteria 2					
Criteria 3					

56

Rating scales

Each criterion statement should be modified to describe each level of the performance's attribute(s).

The choice of words that describe the changing values of the attribute usually depends on verbal qualifiers.

57

Rating scales

Three verbal qualifier 'scales' are commonly used:

- amount (e.g.,few, some, all)
- frequency (e.g., rarely, sometimes, often), and
- intensity (e.g. little, some, very).

58

Rater Training

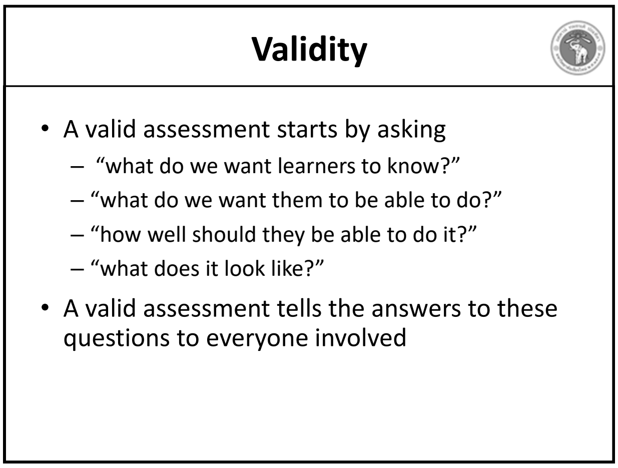
- a rater training prepares raters for the task of judging candidate performance
- it mainly involves the process of the familiarising raters with the test format, test tasks, rubrics, and exemplar performances at each criterion level
- it reduces extreme differences in severity between raters and makes raters more internally self-consistent

59

Reliability

- A reliable assessment is one that elicits similar performances all the time:
  - from different people with the same behaviour
  - on different occasions
  - on same task-type but different tasks
  - when different people rate

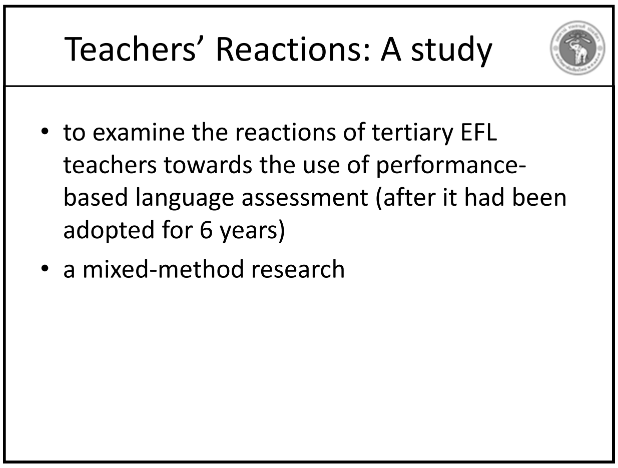
60



Validity

- A valid assessment starts by asking
  - “what do we want learners to know?”
  - “what do we want them to be able to do?”
  - “how well should they be able to do it?”
  - “what does it look like?”
- A valid assessment tells the answers to these questions to everyone involved

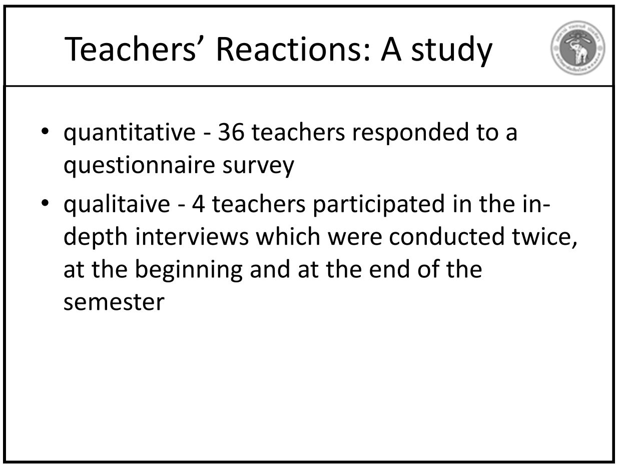
61



Teachers’ Reactions: A study

- to examine the reactions of tertiary EFL teachers towards the use of performance-based language assessment (after it had been adopted for 6 years)
- a mixed-method research

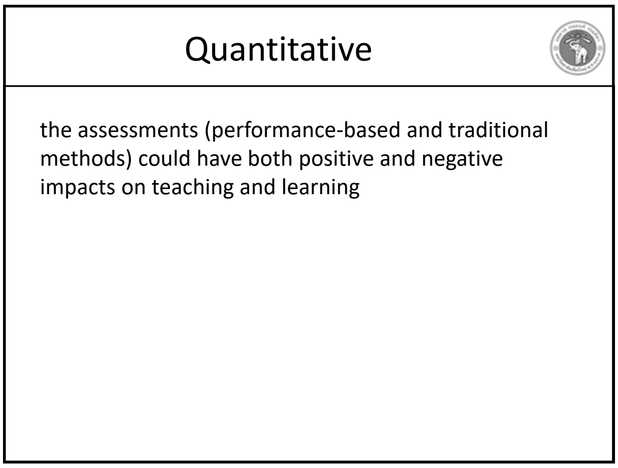
62



Teachers’ Reactions: A study

- quantitative - 36 teachers responded to a questionnaire survey
- qualitaive - 4 teachers participated in the in-depth interviews which were conducted twice, at the beginning and at the end of the semester

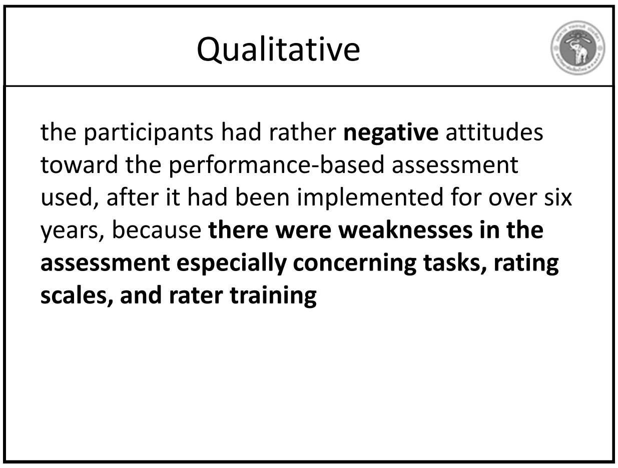
63



Quantitative

the assessments (performance-based and traditional methods) could have both positive and negative impacts on teaching and learning

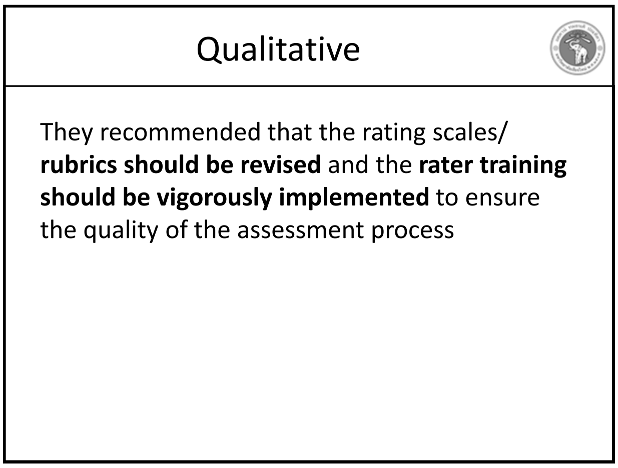
64



Qualitative

the participants had rather **negative** attitudes toward the performance-based assessment used, after it had been implemented for over six years, because **there were weaknesses in the assessment especially concerning tasks, rating scales, and rater training**

65



Qualitative

They recommended that the rating scales/ **rubrics should be revised** and the **rater training should be vigorously implemented** to ensure the quality of the assessment process

66

Language Test and Policies

- assessment comes in all shapes and sizes, ranging from international monitoring exercises to work with individual pupils in the classroom
- assessment, therefore, has been viewed as a powerful tool and used by authorities **to create change**

67

National Policies

2016

- Office of Higher Education Commission (now Ministry of Higher Education, Science, Research and Innovation) implemented the English language education policies aiming at raising the standards of English language teaching at the tertiary level in Thailand.
- One of the five sections outlined in the policy, which all higher education institutions have to take serious action, states that **ALL tertiary students are required to take an English proficiency test**.

68

National Policies

- The test, which could be locally developed, **has to be equivalent** to the "Common European Framework of Reference for Languages (CEFR) or other standards".
- The results of the test could be recorded on the transcript or in a form of a test certificate.

69

Locally developed VS Commercial Tests

Why do we need to develop a test as there are commercially available tests (e.g. TOEIC)?

- "High stake" VS "Low stake" use of test scores
- Budgets

70

Case 1: Local Test

Chiang Mai University

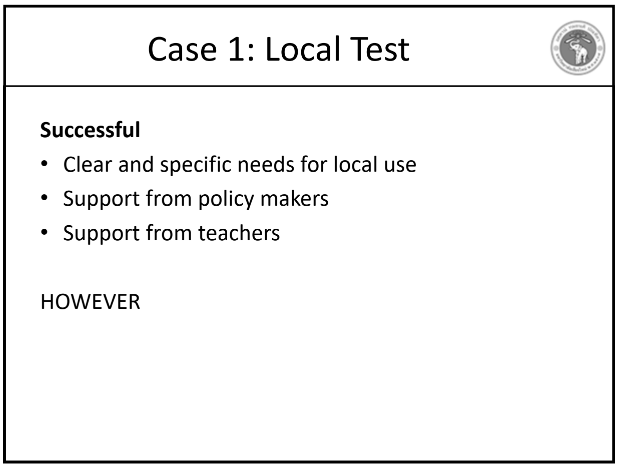
- School leaving English proficiency test
- CEFR B1 level
- Low stake use of test scores
- Budgets
- Locally developed proficiency test (computer-delivered)

71

Case 1: Local Test

- Has been implemented for 3 years
- Replace the old placement test (starting next academic year)
- Compare English proficiency (pre and post) - for policy makers
- Commissioned to write more items every year

72



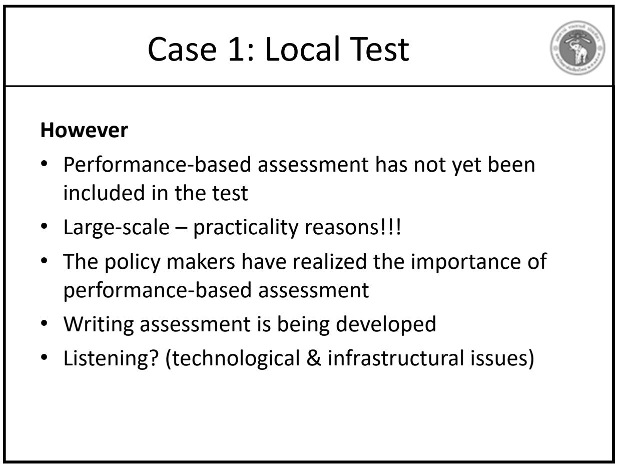
Case 1: Local Test

Successful

- Clear and specific needs for local use
- Support from policy makers
- Support from teachers

HOWEVER

73

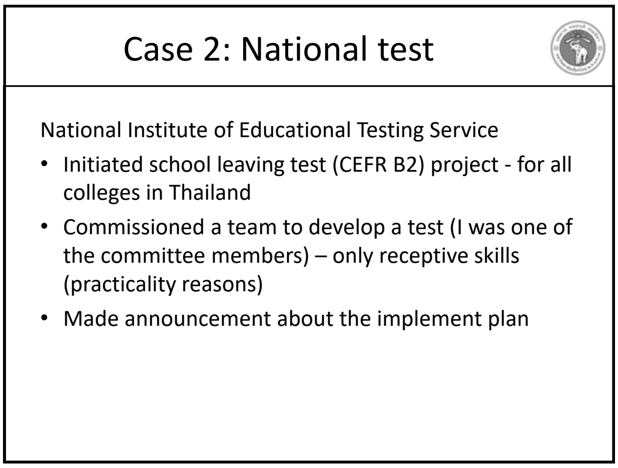


Case 1: Local Test

However

- Performance-based assessment has not yet been included in the test
- Large-scale – practicality reasons!!!
- The policy makers have realized the importance of performance-based assessment
- Writing assessment is being developed
- Listening? (technological & infrastructural issues)

74

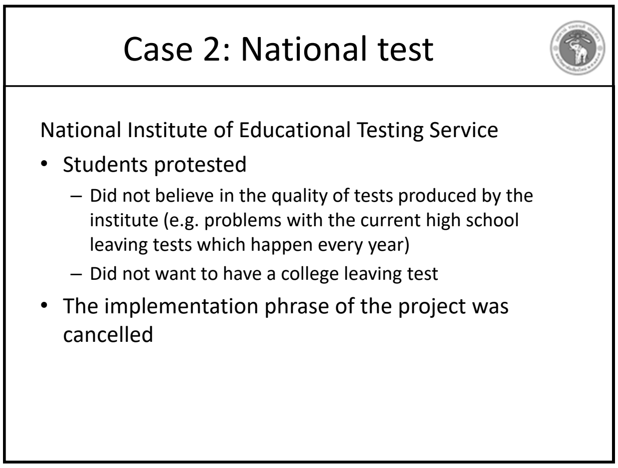


Case 2: National test

National Institute of Educational Testing Service

- Initiated school leaving test (CEFR B2) project - for all colleges in Thailand
- Commissioned a team to develop a test (I was one of the committee members) – only receptive skills (practicality reasons)
- Made announcement about the implement plan

75

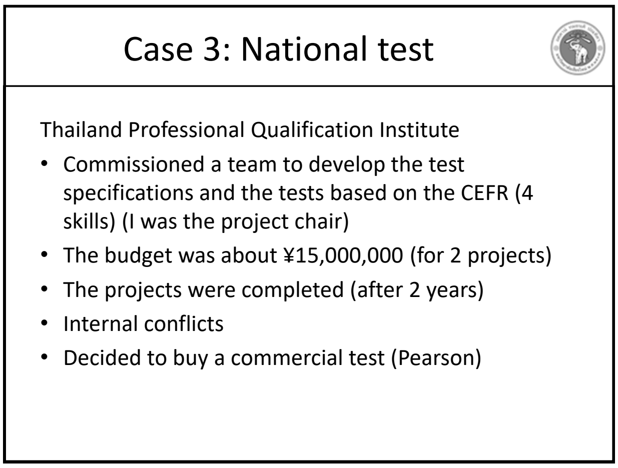


Case 2: National test

National Institute of Educational Testing Service

- Students protested
  - Did not believe in the quality of tests produced by the institute (e.g. problems with the current high school leaving tests which happen every year)
  - Did not want to have a college leaving test
- The implementation phrase of the project was cancelled

76

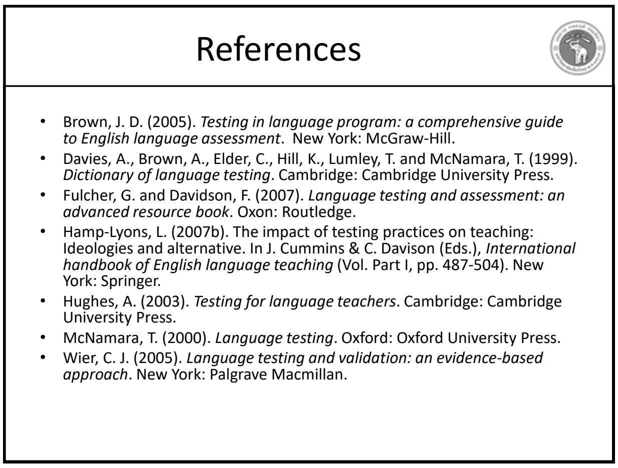


Case 3: National test

Thailand Professional Qualification Institute

- Commissioned a team to develop the test specifications and the tests based on the CEFR (4 skills) (I was the project chair)
- The budget was about ¥15,000,000 (for 2 projects)
- The projects were completed (after 2 years)
- Internal conflicts
- Decided to buy a commercial test (Pearson)

77



References

- Brown, J. D. (2005). *Testing in language program: a comprehensive guide to English language assessment*. New York: McGraw-Hill.
- Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T. and McNamara, T. (1999). *Dictionary of language testing*. Cambridge: Cambridge University Press.
- Fulcher, G. and Davidson, F. (2007). *Language testing and assessment: an advanced resource book*. Oxon: Routledge.
- Hamp-Lyons, L. (2007b). The impact of testing practices on teaching: Ideologies and alternative. In J. Cummins & C. Davison (Eds.), *International handbook of English language teaching* (Vol. Part I, pp. 487-504). New York: Springer.
- Hughes, A. (2003). *Testing for language teachers*. Cambridge: Cambridge University Press.
- McNamara, T. (2000). *Language testing*. Oxford: Oxford University Press.
- Wier, C. J. (2005). *Language testing and validation: an evidence-based approach*. New York: Palgrave Macmillan.

78