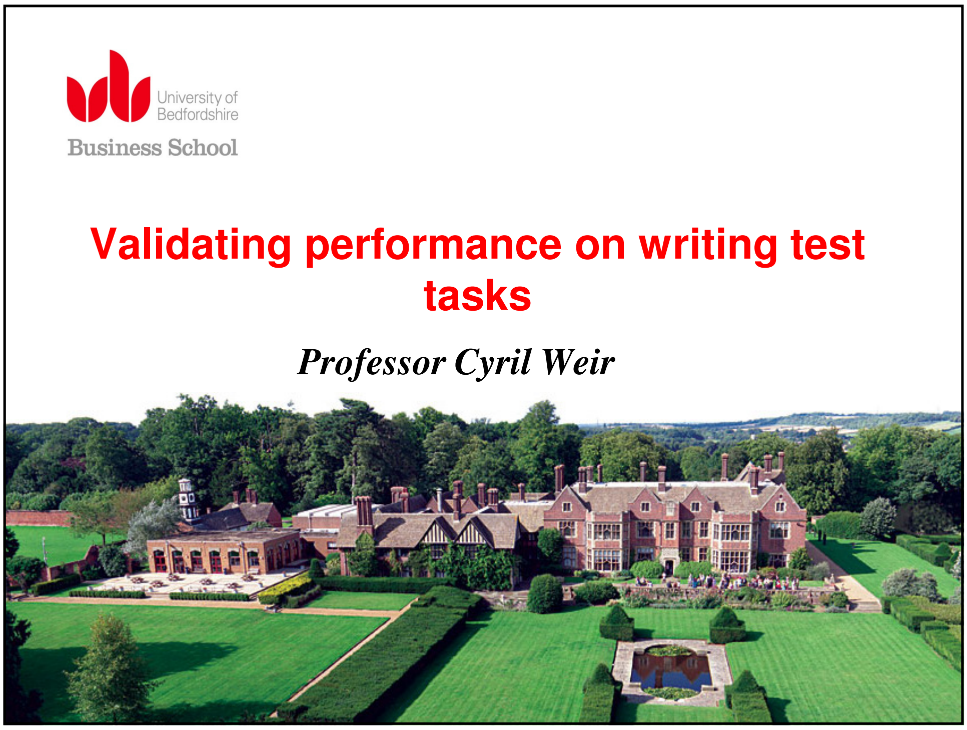
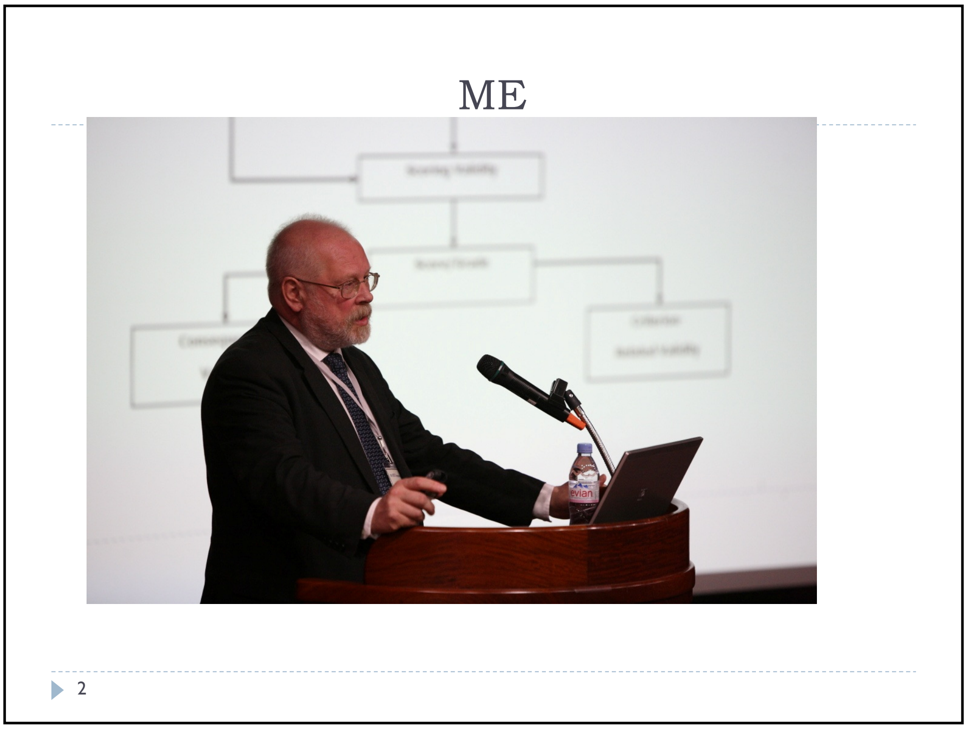


University of Bedfordshire
Business School

Validating performance on writing test tasks

Professor Cyril Weir



ME

Construct Validity

Alan Davies (1984:68) wrote in the first issue of the journal Language Testing:

…in the end no empirical study can improve a test's validity...What is most important is the preliminary thinking and the preliminary analysis as to the nature of the language learning we aim to capture.

Davies (1977b, p.63) had argued earlier:

••• it is, after all, the theory on which all else rests; it is from there that the construct is set up and it is on the construct that validity, of the content and predictive kinds, is based"

O'Sullivan (2013 LTRC Korea) ...we need to put the language back into language testing

DIRECT TESTS OF WRITING

- **"The writing of Essays shews two things; what a man has to say, and how he can say it."** (Henry Latham *On the Action of Examinations Considered as a Means of Selection* 1877: 261)
- Latham made a long and detailed argument in support of the essay and among its many advantages he saw it as a good way of separating the bright from the intellectually challenged (274):

  The thinness of the soil is often displayed by the English Essay and frequently its prognostication proves correct… (276) What a good essay principally shews is a readiness in putting on paper, in a clear and orderly manner, a view that presents itself on applying the mind to a given subject..a good essay will also shew some power of seizing on important points

Direct v Indirect tasks

In the 21st century it now seems strange to train students how to improve their scores on indirect tests of writing, such as multiple choice tests of writing, as was necessary for some international high stakes tests in the not so distant past (notably TOEFL before July 1986).

White (1995: 34) is convincing on the difference:

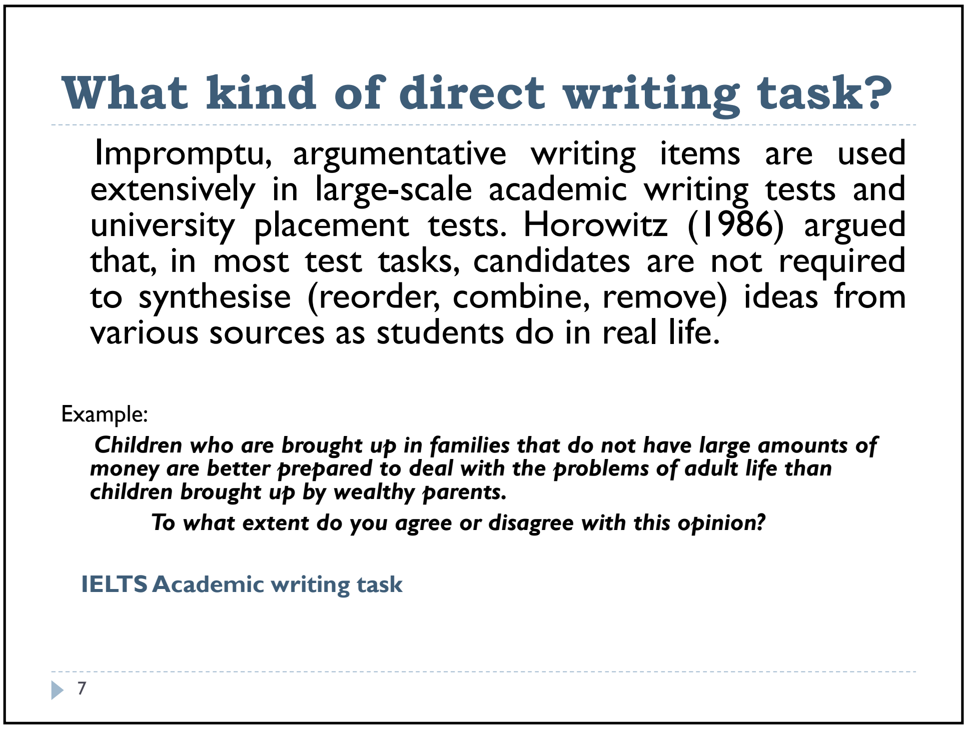
Every essay test shares the artificiality of all tests, but it does require an active response rather than the passive submission called for by multiple-choice examinations. While it is naive to imagine an essay test as a valid measure of all writing, it is disingenuous to ignore the fact that the production of writing for an essay is a wholly different activity than filling in the bubbles on an answer sheet.

Why Direct?

Hamp-Lyons (2001: 3) makes a telling case for direct writing tests:

- … multiple choice tests cannot measure the skills that most writing *teachers* identify as important to effective writing: inventing ideas and arguments; building material into a coherent and effective overall structure to convince, persuade, and teach readers; revising and editing one's own work to more closely approximate conventions of accurate and excellent text and to meet the expectations of a range of audiences …

If the purpose is to measure writing ability, examination boards should be employing writing tasks that encourage teachers ,when they prepare students for the examination, to teach them writing skills they will need in a real world context



What kind of direct writing task?

Impromptu, argumentative writing items are used extensively in large-scale academic writing tests and university placement tests. Horowitz (1986) argued that, in most test tasks, candidates are not required to synthesise (reorder, combine, remove) ideas from various sources as students do in real life.

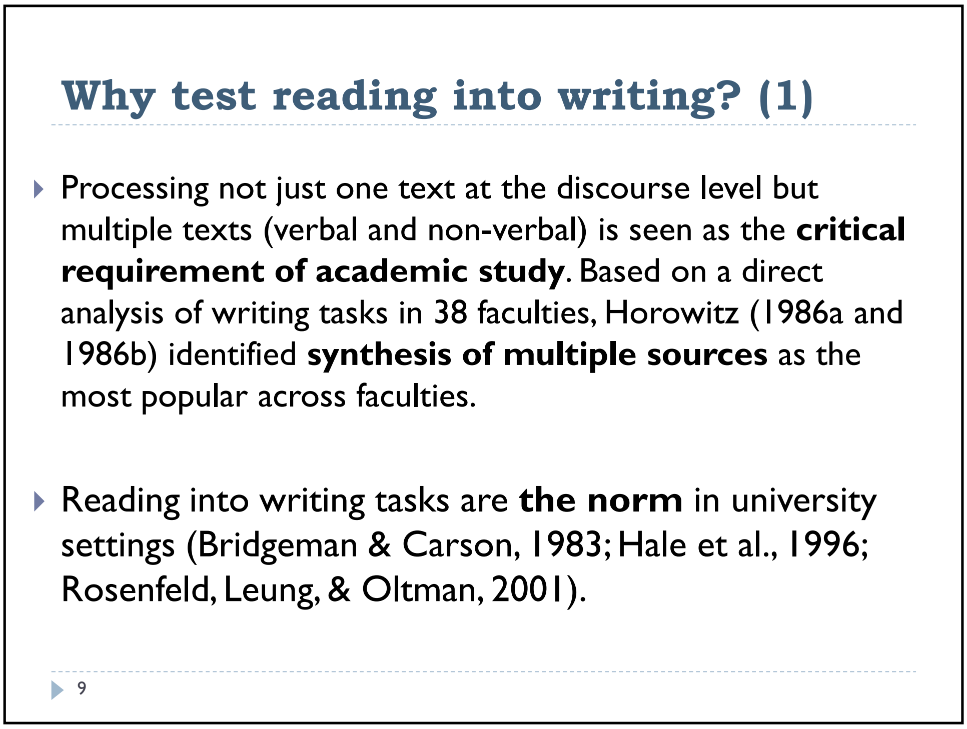
Example:

Children who are brought up in families that do not have large amounts of money are better prepared to deal with the problems of adult life than children brought up by wealthy parents.

To what extent do you agree or disagree with this opinion?

IELTS Academic writing task

7

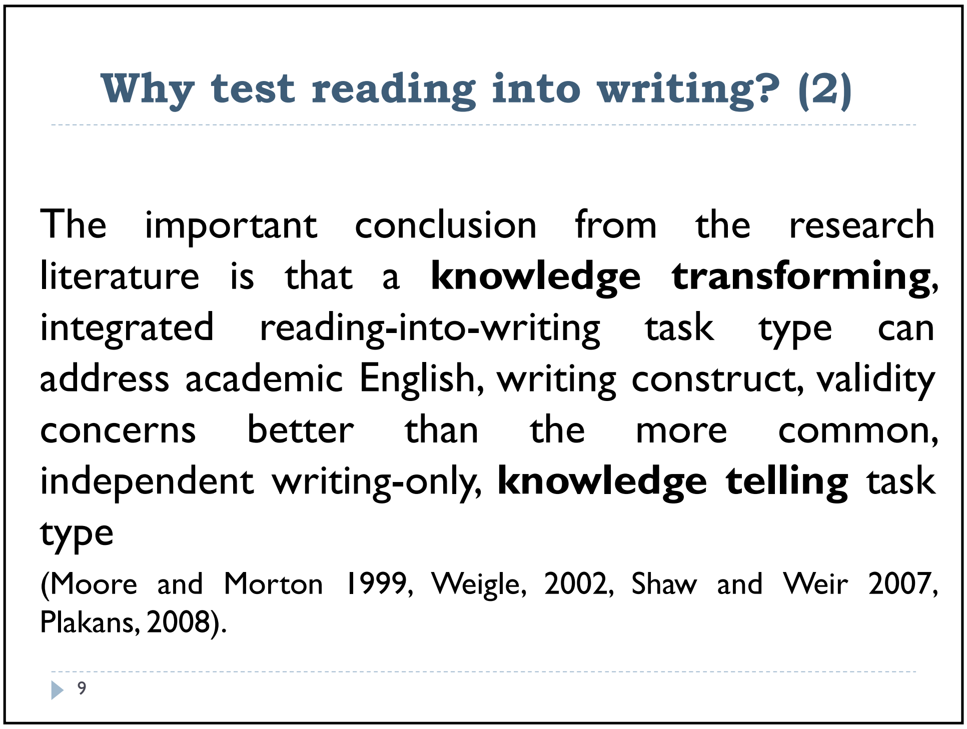


Why test reading into writing? (1)

- Processing not just one text at the discourse level but multiple texts (verbal and non-verbal) is seen as the **critical requirement of academic study**. Based on a direct analysis of writing tasks in 38 faculties, Horowitz (1986a and 1986b) identified **synthesis of multiple sources** as the most popular across faculties.

- Reading into writing tasks are **the norm** in university settings (Bridgeman & Carson, 1983; Hale et al., 1996; Rosenfeld, Leung, & Oltman, 2001).

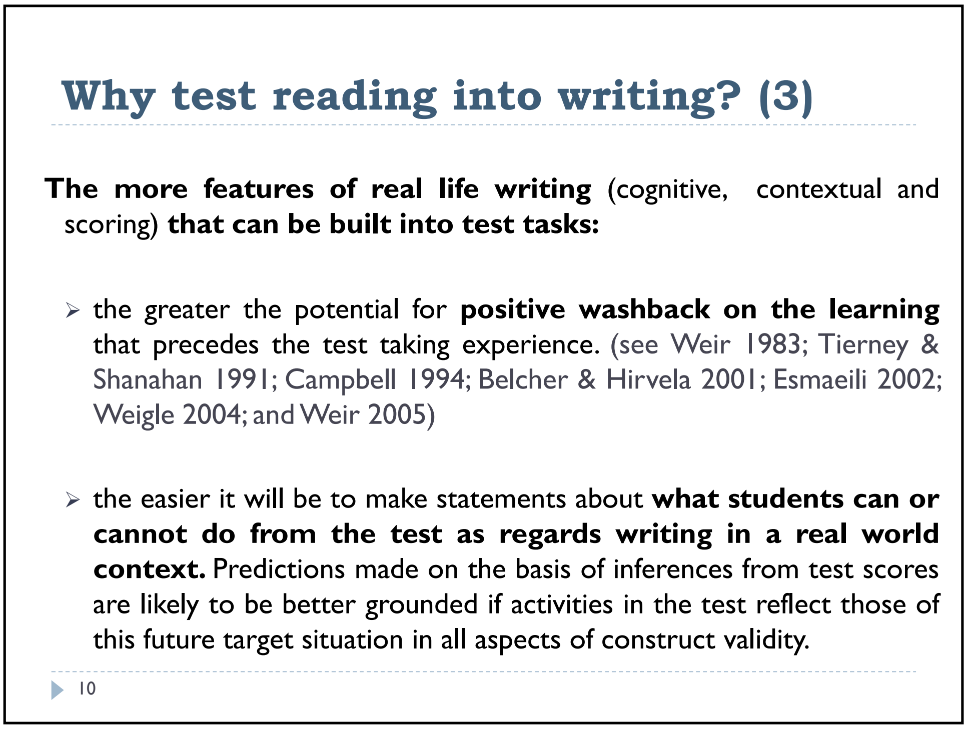
9



Why test reading into writing? (2)

The important conclusion from the research literature is that a **knowledge transforming**, integrated reading-into-writing task type can address academic English, writing construct, validity concerns better than the more common, independent writing-only, **knowledge telling** task type

(Moore and Morton 1999, Weigle, 2002, Shaw and Weir 2007, Plakans, 2008).



Why test reading into writing? (3)

The more features of real life writing (cognitive, contextual and scoring) **that can be built into test tasks:**

- the greater the potential for **positive washback on the learning** that precedes the test taking experience. (see Weir 1983; Tierney & Shanahan 1991; Campbell 1994; Belcher & Hirvela 2001; Esmaeili 2002; Weigle 2004; and Weir 2005)
- the easier it will be to make statements about **what students can or cannot do from the test as regards writing in a real world context.** Predictions made on the basis of inferences from test scores are likely to be better grounded if activities in the test reflect those of this future target situation in all aspects of construct validity.

Integrated tasks: not a new idea

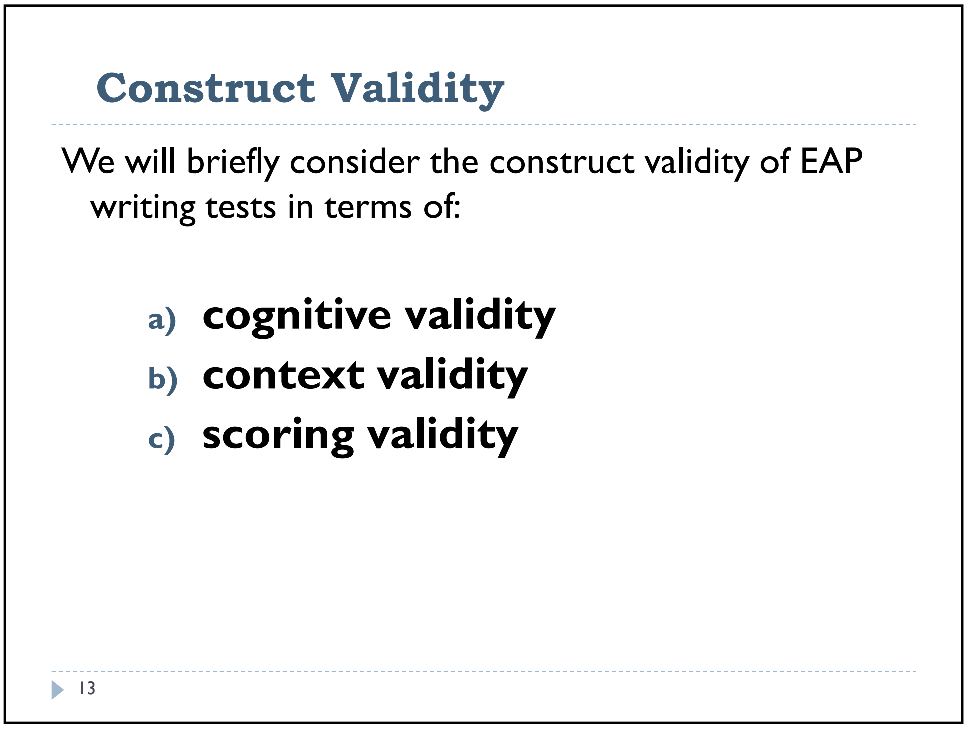
But besides affording very ample time, I would also allow candidates while writing their essay in the examination room to have access to some standard authorities on their subject… there is now no object in forcing men to carry a number of details in their heads…the range of subjects which can be given for essays is very much extended

Latham *On the Action of Examinations Considered as a Means of Selection* (**1877**: 282):

Practicality: a stumbling block?

“ It may be said that practical inconvenience would be found in supplying access to books of reference if the number of candidates were large and suggests this real world task is best saved for the few distinguished candidates”.

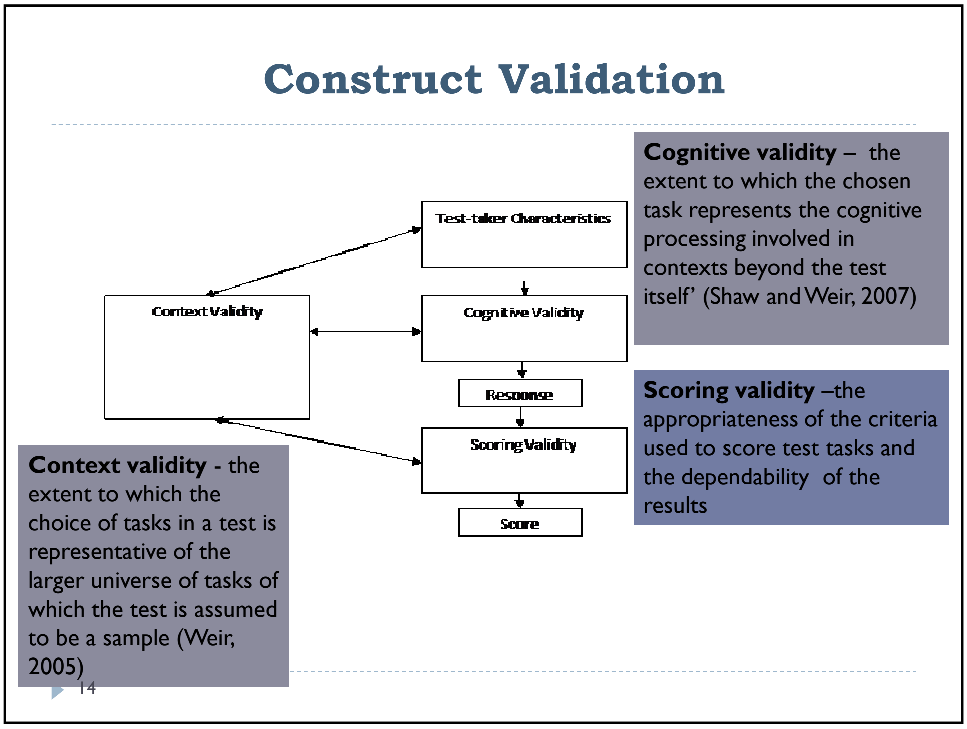
But with computerised tests….



Construct Validity

We will briefly consider the construct validity of EAP writing tests in terms of:

a) **cognitive validity**
b) **context validity**
c) **scoring validity**



Construct Validation

Cognitive validity – the extent to which the chosen task represents the cognitive processing involved in contexts beyond the test itself' (Shaw and Weir, 2007)

Test-taker Characteristics

Context Validity

Cognitive Validity

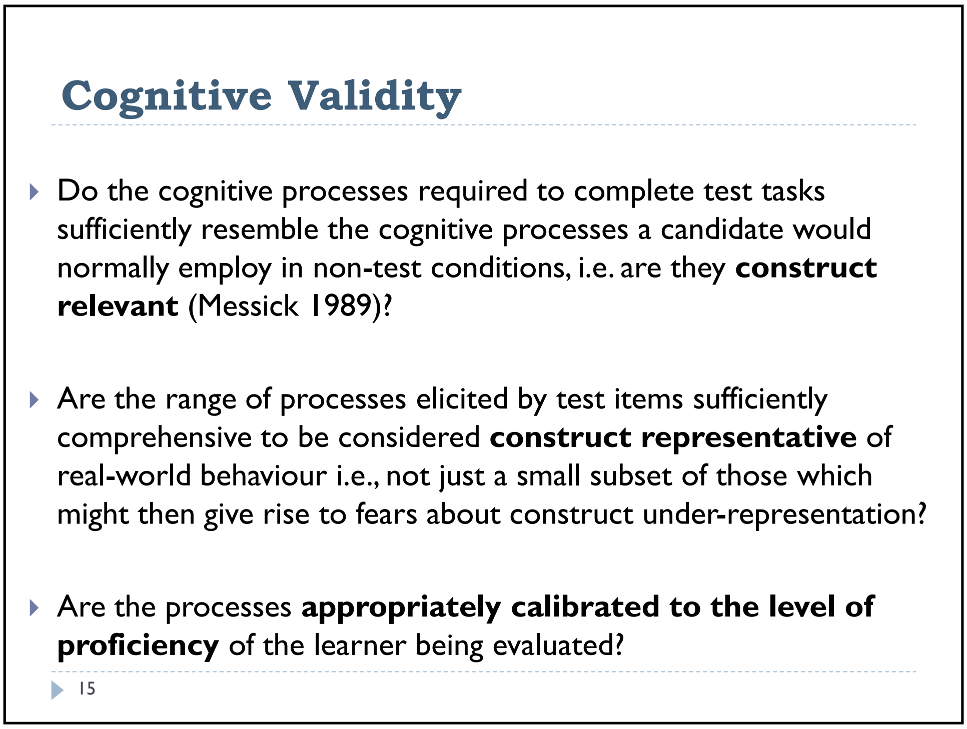
Response

Scoring Validity

Score

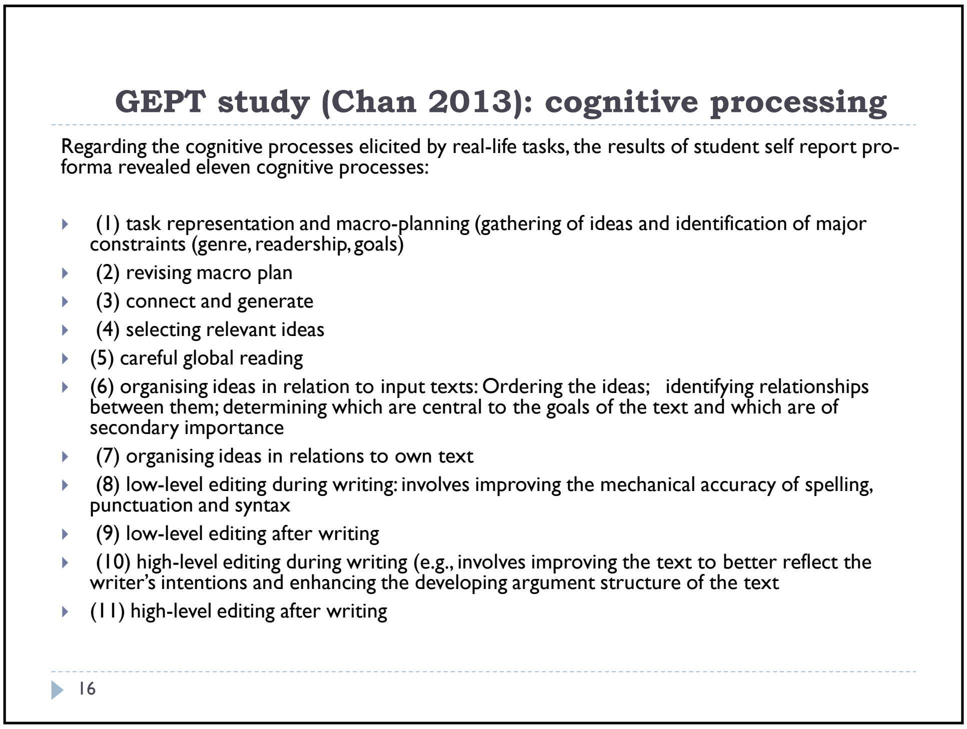
Scoring validity –the appropriateness of the criteria used to score test tasks and the dependability of the results

Context validity - the extent to which the choice of tasks in a test is representative of the larger universe of tasks of which the test is assumed to be a sample (Weir, 2005)



Cognitive Validity

- Do the cognitive processes required to complete test tasks sufficiently resemble the cognitive processes a candidate would normally employ in non-test conditions, i.e. are they **construct relevant** (Messick 1989)?
- Are the range of processes elicited by test items sufficiently comprehensive to be considered **construct representative** of real-world behaviour i.e., not just a small subset of those which might then give rise to fears about construct under-representation?
- Are the processes **appropriately calibrated to the level of proficiency** of the learner being evaluated?



GEPT study (Chan 2013): cognitive processing

Regarding the cognitive processes elicited by real-life tasks, the results of student self report pro-forma revealed eleven cognitive processes:

- (1) task representation and macro-planning (gathering of ideas and identification of major constraints (genre, readership, goals)
- (2) revising macro plan
- (3) connect and generate
- (4) selecting relevant ideas
- (5) careful global reading
- (6) organising ideas in relation to input texts: Ordering the ideas; identifying relationships between them; determining which are central to the goals of the text and which are of secondary importance
- (7) organising ideas in relations to own text
- (8) low-level editing during writing: involves improving the mechanical accuracy of spelling, punctuation and syntax
- (9) low-level editing after writing
- (10) high-level editing during writing (e.g., involves improving the text to better reflect the writer's intentions and enhancing the developing argument structure of the text
- (11) high-level editing after writing

Results (Chan 2013)

The analysis comparing the extent to which the eleven processes were employed between the real-life and test conditions reveals positive results for the cognitive validity of the reading-into-writing test tasks. Both Test Task A (essay task with multiple verbal inputs) and Test Task B (essay task with multiple verbal and non-verbal inputs) were able to elicit from participants most of the cognitive processes in a similar manner to the way participants employed the processes on the real-life tasks

Context Validity

- Cognitive processing in a writing test never occurs in a vacuum but is activated **in response to the contextual parameters** set out in the wording of the writing task.

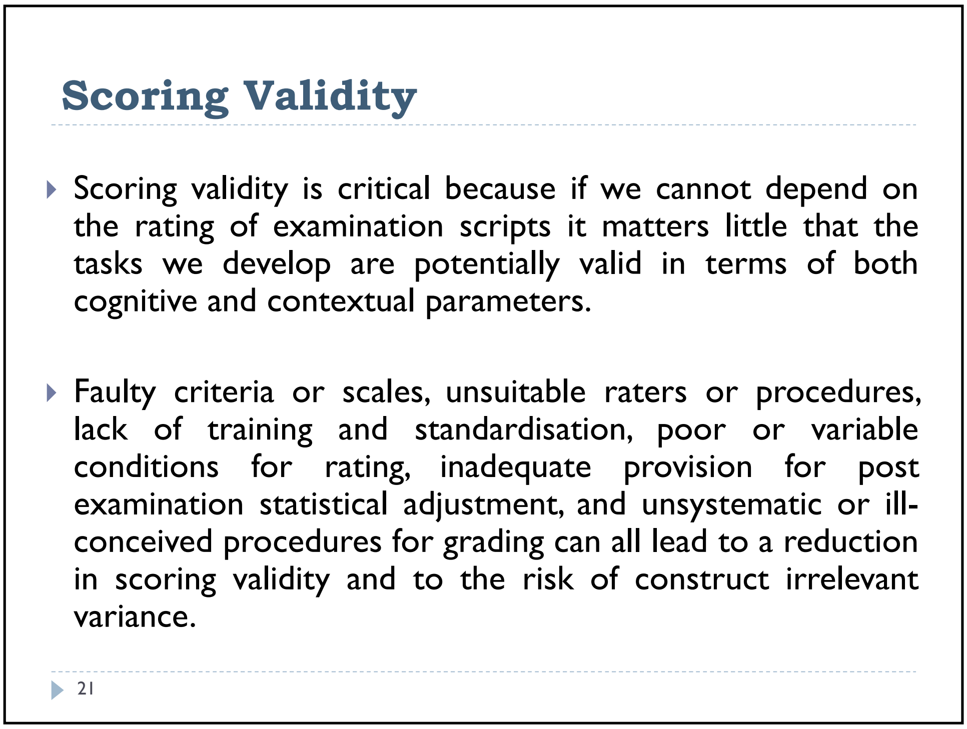
- Context validity for a writing task addresses **the particular performance conditions, the setting under which the task is to be performed** (e.g. purpose of the task, time available, length required, specified addressee, known marking criteria as well as **the linguistic demands inherent in the successful performance of the task**)

Context validity

- Are the characteristics of the test task an adequate and comprehensive representation of those that would be normally encountered in **the real life context**?
- Are they appropriately calibrated to **the level of proficiency** of the learner being evaluated?

GEPT Study Chan 2013: contextual parameters

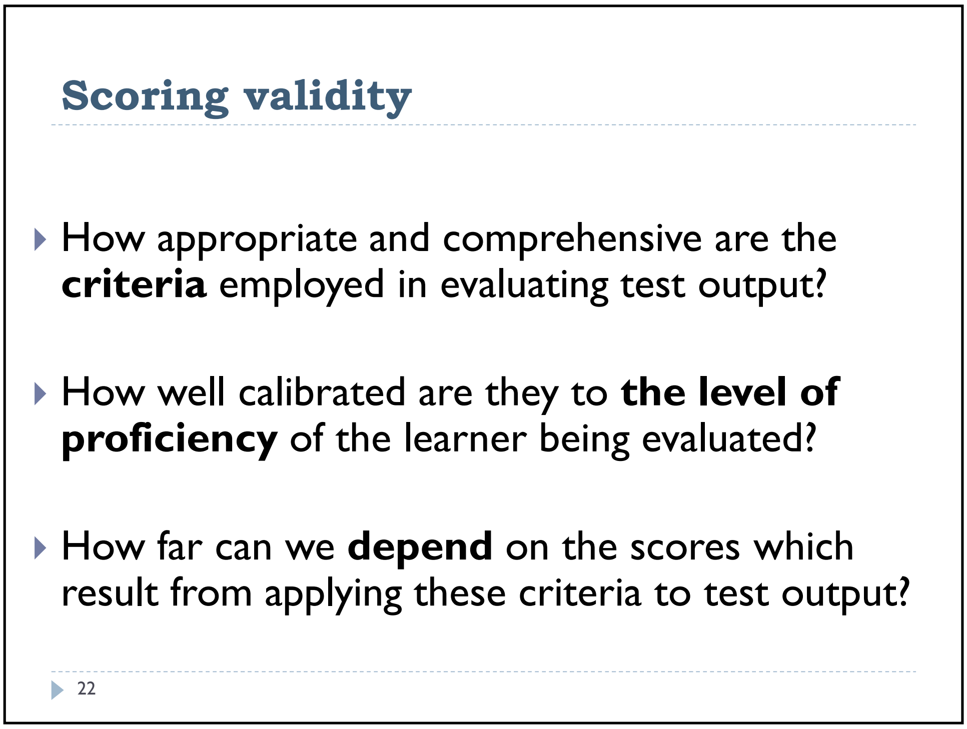
- The difficulty level between the real-life input texts and test task input texts was similar in terms of most of the lexical, syntactic and cohesion automated indices investigated in the study
- Based on the expert judgement analysis, the two reading-into-writing test tasks resembled the *overall task setting* of the real-life tasks in a number of important ways e.g., clarity of task purpose, genre required, language functions, textual organisation, intended audience, cultural specificity, knowledge of assessment criteria…



Scoring Validity

- Scoring validity is critical because if we cannot depend on the rating of examination scripts it matters little that the tasks we develop are potentially valid in terms of both cognitive and contextual parameters.

- Faulty criteria or scales, unsuitable raters or procedures, lack of training and standardisation, poor or variable conditions for rating, inadequate provision for post examination statistical adjustment, and unsystematic or ill-conceived procedures for grading can all lead to a reduction in scoring validity and to the risk of construct irrelevant variance.



Scoring validity

- How appropriate and comprehensive are the **criteria** employed in evaluating test output?

- How well calibrated are they to **the level of proficiency** of the learner being evaluated?

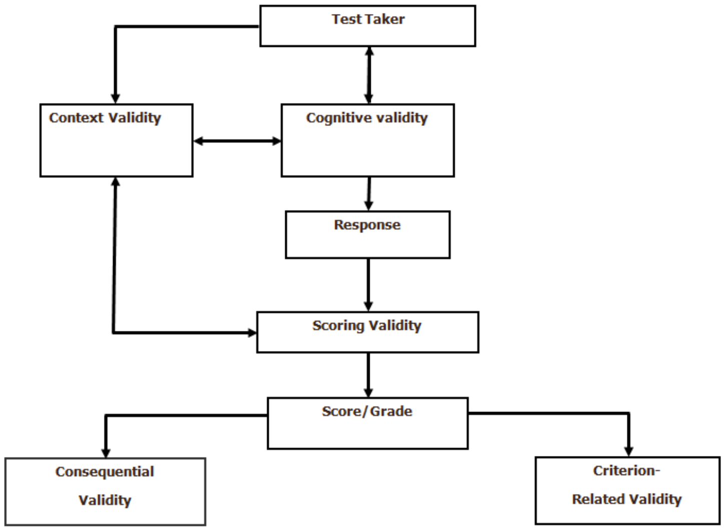
- How far can we **depend** on the scores which result from applying these criteria to test output?

Scoring Validity

- Experimental generalisability studies were carried out as part of the IELTS Speaking and Writing Revision Projects to investigate the reliability of ratings (Shaw, 2004; Taylor & Jones, 2001). More recent G-studies based on examiner certification data showed coefficients of 0.83-0.86 for Speaking and 0.81-0.89 for Writing.

- The inter-rater reliability indices for the GEPT writing and speaking tests are between 0.89 and 0.90,

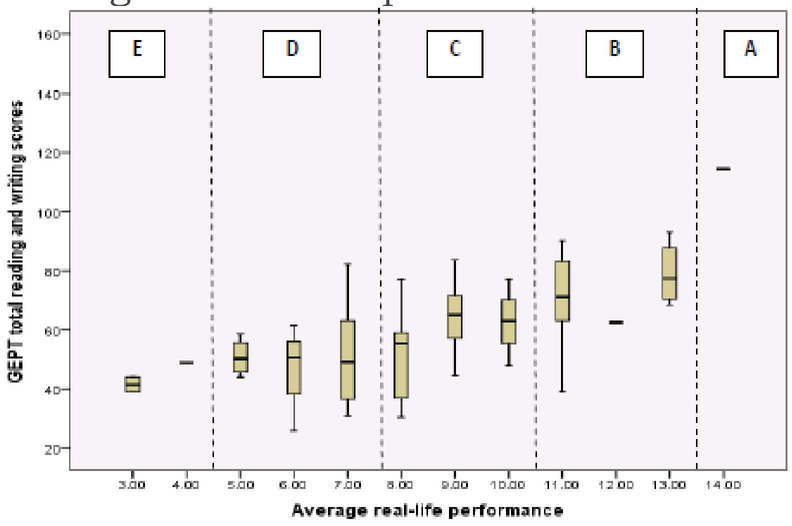
The socio-cognitive framework for test development and validation



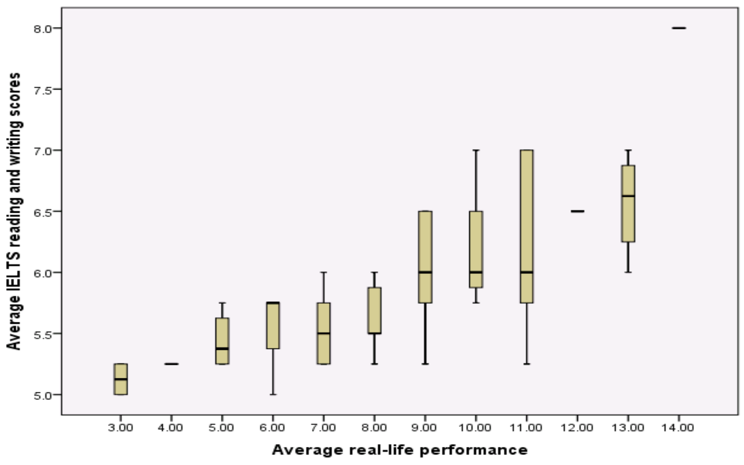
Criterion-related validity

Criterion-related validity is a form of external evidence, which is defined as ‘a predominantly quantitative and a posteriori concept, concerned with the extent to which test scores correlate with a suitable external criterion of performance with established properties’ (Weir, 2005:35).

Relationships between GEPT reading and writing and real-life performance



Relationships between IELTS reading and writing and Real-life performance



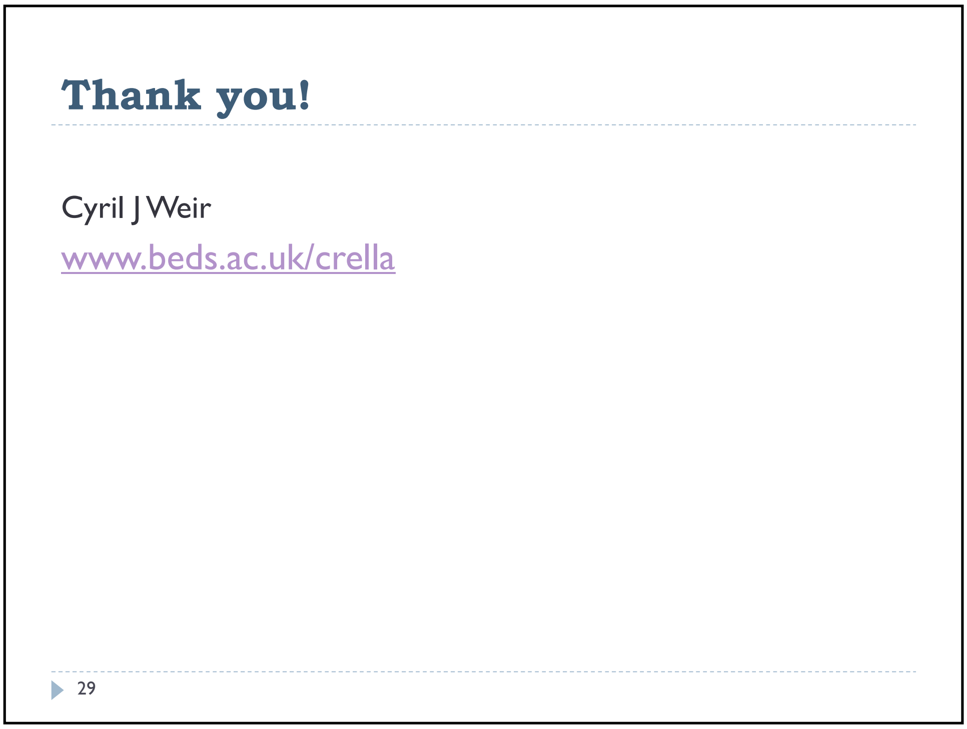
Predictive Validity

GEPT reading and writing scores correlated with the participants' real-life academic performances at .529 (p<.01), explaining 27.98% of the variance of the real-life performances.

IELTS reading and writing scores correlated with the participants' real-life academic performances at .602 (p<.01), explaining 36.23% of the variance of the real-life performances

A final question

Why are these reliability coefficients higher than in past, when .25-.35 was regarded as a good correlation between tests and real life performance (e.g., see Pollitt 1988 Predictive validity. *ELTS Research Report 1 (ii)*, 62-65)?



Thank you!

Cyril J Weir

www.beds.ac.uk/crella

29