Automated approaches to establishing context validity in reading tests

Lynda Taylor and Cyril J Weir
CRELLA, University of Bedfordshire

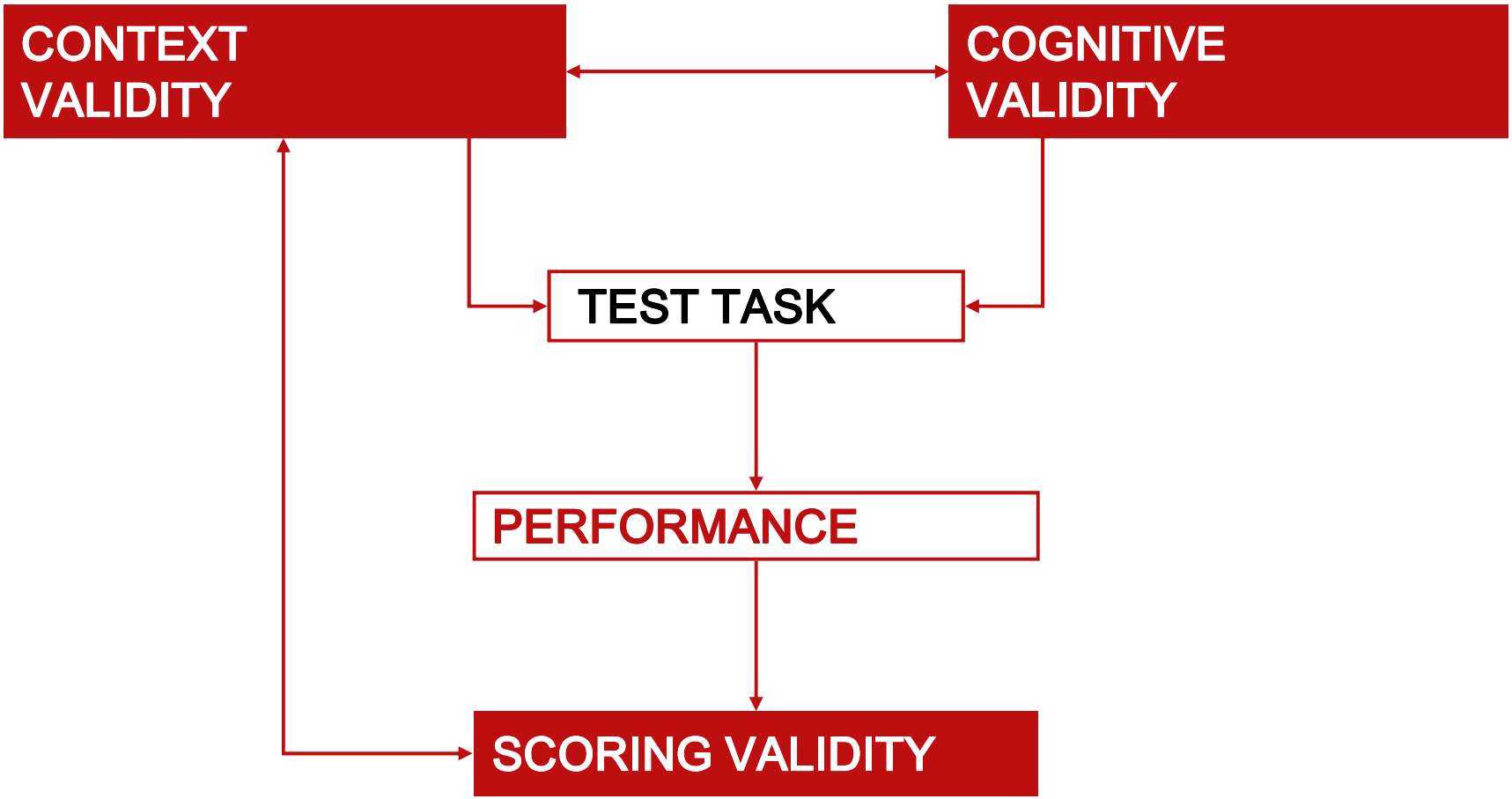
Paper given at the 9th EALTA Conference,
Innsbruck, Austria – 31 May to 3 June 2012

CRELLA
University of Bedfordshire

This presentation….

- links to the EALTA conference theme of ***validity in language testing and assessment***
- relates more specifically to the ***validity of instruments and procedures***
- contributes to the ongoing debate about ***which facets are essential in establishing validity*** – **what** evidence needs to be gathered, and **how**?
- focuses upon ***contextual validity*** within a unitary concept of overall construct validity

CONSTRUCT VALIDITY



Context validity in testing reading
(what do we mean?)

CRELLA
University of Bedfordshire

Context validity in testing reading

Context validity relates to the appropriateness of both :

the linguistic and content demands of the text to be processed (i.e. read and comprehended)

and

the features of the task setting that impact on task completion (e.g. responding to comprehension questions or writing a summary)

Establishing context validity in reading tests – 3 questions

- What are the key contextual features (within-text) that we need to take account of when selecting texts for reading tests?
- How do we evaluate the relative importance / difficulty of contextual features in texts targeted at different proficiency levels to provide validity evidence?
- How might automated approaches help us to do this?

Approaches to text selection

- the traditional role of expert judgement (individual test writers) in choosing texts for reading tests
- a relative lack of quantitative tools to determine efficiently the various characteristics of the written texts used in reading tests (Biber et al 2004)
- recent advances in automated textual analysis, computational linguistics and the development of corpora
- more approaches now available focusing analytically on a wide range of individual text characteristics

Coh-Metrix and other tools (for English)

- a synthesis of the advances in these areas has been achieved in **Coh-Metrix**, a computational tool that measures cohesion and text difficulty at various levels of language, discourse, and conceptual analysis [http://cohmetrix.memphis.edu]

- also useful indices in **Vocabprofile** [Cobb, T. (2003). VocabProfile, Compleat Lexical Tutor - http://www.lextutor.ca]

- and even **Word** in Microsoft Office can be useful

- and all are publicly available!

Empiricism versus connoisseurship

- pooled expert judgement is still necessary for some decisions, e.g. cultural specificity, content knowledge, topic familiarity

- advantages of automated analysis?
  - more efficient than approaches that involve humans annotating and rating texts by hand
  - can cope with large volumes of data

Analysing some contextual parameters in reading texts *(using automated tools)*

CRELLA
University of Bedfordshire

Contextual parameters in reading

We will consider some illustrative parameters relating to:

- ***lexical complexity*** *(decoding)*
- ***structural complexity*** *(syntactic parsing)*
- ***cohesion*** *(the construction of meaning)*

… to explore how automated analysis tools can help to investigate these in reading texts.

… and to try and link up the selected parameters with relevant cognitive processing factors

Some examples to be drawn from recent studies that applied automated tools

- comparative analysis of u/grad texts and IELTS Academic Reading texts (Green, Unaldi and Weir, 2010, in *Language Testing*)
- Wu (2011) – analysis of GEPT Taiwan reading tests
- diachronic analysis of Cambridge Proficiency (CPE) reading texts used in a variety of tasks
  - translation (1913-88), summary (1930-2010), MCQ/SAQ comprehension
- comparative analysis of Cambridge Key (KET), Preliminary (PET), First (FCE), Advanced (CAE) and Proficiency (CPE) reading texts (Khalifa & Weir 2009, Weir et al 2012)

Lexical complexity
(decoding)

CRELLA
University of Bedfordshire

Some key lexical parameters

L1 Syllables per word

L2 Type token ratio

L3 Word frequency

L4 Lexical density

L5 Proportion of academic words

L1 : Average syllables per word

- the mean number of syllables per content word

  [Coh-Metrix index 56 – READASW]

  multisyllabic words take longer to read and process than monosyllabic words [Rayner & Pollatsek, 1989]

  "In general, the more syllables per word and the more words per sentence, the higher the associated grade level of the text"

  [White, S. (2011) *Understanding Adult Functional Literacy: Connecting Text Features, Task Demands, and Respondent Skills.* Taylor & Francis]

L2 : Type token ratio

- the number of unique words divided by the number of tokens of the words

[Coh-Metrix index 19 – TYPTOKc]

Each unique word in a text is a word *type.* Each instance of a particular word is a *token.*

When the type: token ratio is 1, each word occurs only once in the text; comprehension should be comparatively difficult because many unique words need to be encoded and integrated with the discourse context. A low type: token ratio indicates that words are repeated many times in the text, which should generally increase the ease and speed of text processing.

[Templin, M (1957) *Certain Language Skills in Children: Their development and interrelationships*. Institute of Child Welfare Monograph Series No. 26. Minneapolis, MN: University of Minnesota Press]

L3 : Word frequency

- the relative frequency of occurrence of words
  [Coh-Metrix indices 40, 41, 42, 43]

  Frequency effects have been shown to facilitate decoding:
  - frequent words are processed more quickly and understood better than infrequent ones (Haberlandt & Graesser, 1985; Just & Carpenter, 1980).
  - rapid or automatic decoding = strong predictor of L2 reading performance (Koda, 2005)
  - texts which assist such decoding (e.g., by containing a greater proportion of high frequency words) can thus be regarded as easier to process….

  The more frequent a word, the more likely it is to be processed with a fair degree of automaticity, thus increasing reading speed (even among lower level learners) and freeing working memory for higher level meaning building. (Crossley, Greenfield and McNamara, 2008)

L4 : Lexical density

- depends on distinguishing between different word types, i.e. lexical (content) and function words
  - lexical: verbs, nouns, adjectives, adverbs
  - function: auxiliaries, determiners, pronouns, prepositions, conjunctions

[VocabProfile]

Accessing the meaning of lexical items requires accessing the mental lexicon, function words can be dealt with by pattern matching. Reading focuses mainly on lexical items and readers tend to skip function words.

L5 : Proportion of academic words

- the incidence of academic words in a text

  [VocabProfile, based on AWL Coxhead, 2000]

- proved to be a good predictor of level in a study of FCE, CAE and CPE reading texts (Weir et al, 2012)

	Mean	**SD**
FCE (B2)	1.61%	1.26%
CAE (C1)	1.63%	1.41%
CPE (C2)	5.82%	2.84%

Syntactic complexity
(syntactic parsing)

CRELLA
University of Bedfordshire

Syntactic complexity

- linear processing of text in careful reading, with the reader decoding word by word
- assembly of decoded items into larger scale syntactic structure

  (Just & Carpenter, 1987; Rayner & Pollatsek, 1994)

- cognitive demands imposed vary considerably according to how complex the structure is

  (Perfetti, Landi & Oakhill, 2005)

  (Crossley ,Greenfield and McNamara, 2008)

Some key syntactic parameters

S1 Sentence length

S2 Readability formulae

S3 Higher level constituents

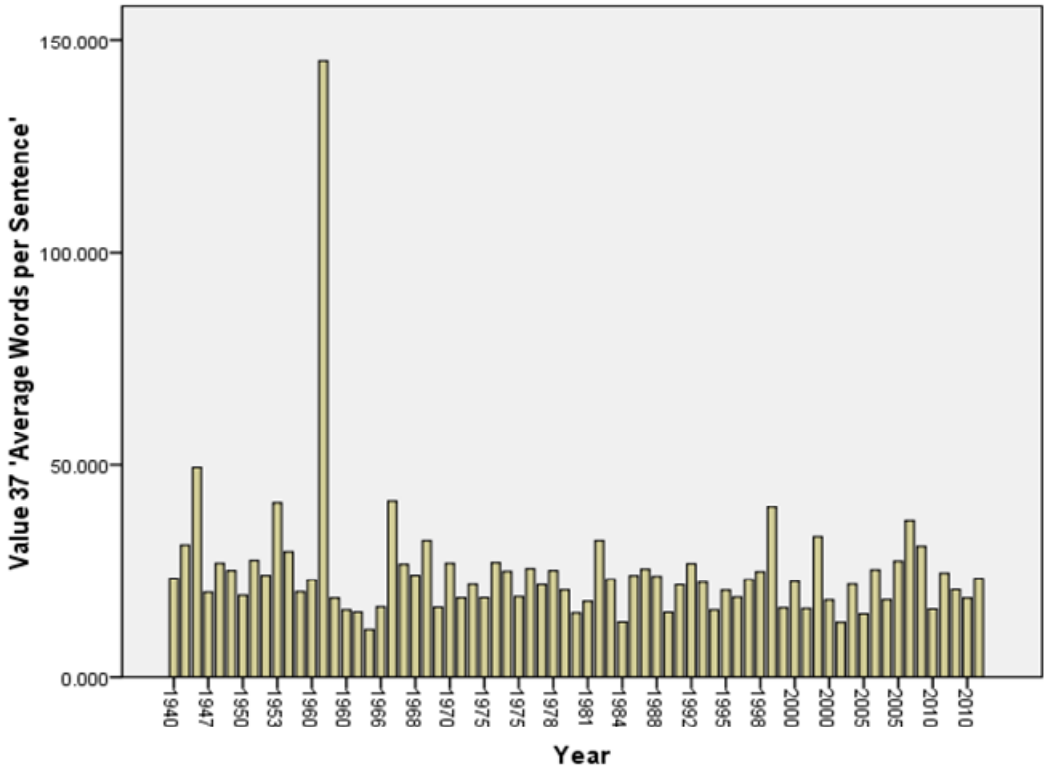
S1 : Sentence length in Cambridge ESOL reading papers

[using Word - or Coh-Metrix index 57]

Main Suite Level	Average number of words per sentence	Range
KET (A2)	13.2	8 - 17
PET (B1)	14.9	10 - 20
FCE (B2)	18.4	11 - 25
CAE (C1)	18.6	13 - 27
CPE (C2)	19.6	13 - 30

S1 : Sentence length
(a diachronic perspective on CPE)

SAQ



Short answer question (SAQ)
CPE 1960 outlier sentence

Strangely enough there was not a sound in the house as having opened the street-door with his latchkey as he was in the habit of doing every evening at about this time, he walked into the lighted hall after shutting the door behind him with the customary click, noticing while his hands were occupied with the mechanical movements of hanging his hat and coat on the stand against the wall, that the light on the upper landing of the stairs was, for some reason, perhaps a perfectly trivial one, not on as it usually was, before moving, again habitually, to the door of the sitting-room at the foot of the stairs where, with his hand on the knob, he suddenly let the incipient feeling of alarm at the back of his mind take rigid hold of him with the discovery that the door was locked.

(sentence of 145 words!)

S2 : Readability formulae

- are long-established and widespread in use
- rely heavily on **word length** and **sentence length** (simple and shallow metrics?)
- ignore many language and discourse components that are theoretically expected to influence reading and comprehension difficulty

…nevertheless…

- texts with **longer words** and **lengthier sentences** are more difficult to read
  - longer words tend to be less frequent in the language - Zipf's (1949) law, and infrequent words take more time to access and interpret during reading (Just & Carpenter, 1980)
  - longer sentences place more demands on working memory
  - real-time processing means holding information in your head until you can parse sentences syntactically
  - the longer the sentence, the more difficult this may be

S2 : Difficulty/readability estimates in Cambridge ESOL reading papers

[using Word or Coh-Metrix indices 59 & 60]

Main Suite level	Flesch reading ease score	Flesch-Kincaid grade level	Flesch-Kincaid range
KET (A2)	78.3	5.5	2 – 7.4
PET (B1)	64.7	7.9	5 – 10.1
FCE (B2)	66.5	8.4	5 – 12.3
CAE (C1)	58.4	9.6	5.7 - 16
CPE (C2)	57.7	9.9	5.6 – 16.1

S3 : Higher level constituents

- the number of main verbs in a sentence is broadly indicative of the number of clauses - thus of complex syntactic composition
- the more complex the syntactic composition, the greater the load on cognitive processing
- the more clauses you have to process in a sentence, the more propositions you have to hold in working memory and link together

[Coh-Metrix index 51 – SYNHw]

Cohesion (and coherence)
(the construction of meaning)

CRELLA
University of Bedfordshire

Cohesion (and coherence)

Cohesion is an objective property of the explicit language and text. There are explicit features, words, phrases, or sentences that guide the reader in interpreting the substantive ideas in the text, in connecting ideas with other ideas, and in connecting ideas to higher level global units (e.g., topics and themes). These cohesive devices cue the reader on how to form a coherent representation.

The coherence relations are constructed in the mind of the reader and depend on the skills and knowledge that the reader brings to the situation…coherence is a psychological construct, whereas cohesion is a textual construct

[Graesser et al 2004: 193]

Cohesion

- two forms of textual cohesion can be estimated :
  - referential cohesion (the extent to which words in the text co-refer)
  - conceptual cohesion (the degree of similarity between concepts in different parts of a text)

[Coh-Metrix indices 7, 8, 9, 10, 13, 14]

- Coh-Metrix also provides an index relating to degree of concreteness/abstractness in a text

[Coh-Metrix indices 45]

… however…

- the recent literature suggests that different forms of cohesion are not always positively correlated with grade-level bands (Graesser et al 2011)

- text cohesion has a small variation over grade level, with a slight decrease for referential cohesion within most text genres and a slight increase for causal cohesion

- in studies carried out by CRELLA, cohesion indices did not clearly relate to different levels of text either in reading or writing.

The value of automated tools

- importance of analysing texts at multiple language-discourse levels:
  - *words, syntax, explicit textbase, the mental model, the discourse genre and rhetorical structure*
- Coh-Metrix and similar automated tools
  - provide a convenient, rapid check of similarities and differences between texts and should facilitate the development of more materials that are genuinely comparable
- fewer issues of reliability than judgements about textual features by test writers adopting a more traditional checklist approach
- but also complements expert judgement (connoisseurship)

Some considerations?

- which of the analytical features (measures/indices) available are really informative and useful?
- how do we select those parameters in which to invest a measure of meaningfulness or confidence?
- can we convincingly link each parameter to current understanding and research evidence on cognitive processing in reading?

Results of one-way ANOVA among FCE, CAE and CPE texts (Weir et al 2012)

		FCE (N=48)		CAE (N=49)		CPE (N=69)		Kruskal-Wallis Test
		Mean	SD	Mean	SD	Mean	SD	
1	Argument Overlap, adjacent, unweighted	0.50	0.12	0.45	0.16	0.38	0.24	.000
2	Anaphor reference, adjacent, unweighted	0.53	0.14	0.41	0.16	0.37	0.20	.000
3	Anaphor reference, all distances	0.30	0.12	0.20	0.12	0.18	0.13	.000
4	Number of negations, incidence score	5.83	3.70	6.44	4.72	8.90	6.25	.015
5	Logical operator incidence score	35.72	8.31	38.58	11.42	44.87	15.22	.001
6	LSA, Sentence to sentence adjacent mean	0.13	0.03	0.15	0.05	0.18	0.07	.000
7	Average words per sentence	18.68	3.03	20.42	4.01	23.39	7.25	.000
8	Average Syllables per word	1.42	0.07	1.54	0.11	1.55	0.13	.000
9	Mean number of modifiers per noun-phrase	0.75	0.16	0.92	0.20	0.90	0.16	.000
10	Higher level constituents per word	0.76	0.04	0.72	0.04	0.72	0.03	.000
11	Mean number words before main verb of main clause in sentences	4.07	1.12	4.40	1.30	5.23	2.15	.014
12	Type-token ratio for all content words	0.75	0.04	0.75	0.08	0.80	0.07	.000
13	Celex, logarithm, mean for content words (0-6)	2.35	0.11	2.21	0.15	2.18	0.13	.000
14	Sentence syntax similarity, all, across paragraphs	0.09	0.02	0.08	0.02	0.08	0.02	.003
15	Prop. content words over-lapping between adjacent sentences	0.08	0.03	0.07	0.03	0.07	0.03	.007
16	Concreteness, minimum in sentence for content words	170.92	15.53	167.55	16.72	181.72	21.35	.000
17	AWL	1.61	1.26	1.64	1.41	5.02	2.84	.000
18	Offlist >15k	0.67	0.59	1.05	0.91	1.61	1.64	.001

Application in testing and beyond

- to ensure greater consistency when evaluating contextual characteristics of texts in a test
- to contribute to test validity and test form comparability
- to support test writer training

and maybe also…

- to assist EL reading materials producers?
- to enhance specification within CEFR?

References

- Biber, D., Conrad, S., Reppen, R., Byrd, P., Helt, M., Clark, V., Cortes, V., Csomay, E.C. and Urzua, A. (2004) *Representing Language Use in the University: Analysis of the TOEFL 2000 Spoken and Written Academic Language Corpus.* (ETS TOEFL Monograph Series, MS-25). Princeton, NJ: Educational Testing Service.
- Coxhead, A. (2000) A new academic word list. *TESOL Quarterly* 34 (2), 213-38.
- Crossley, S., Greenfield, J. & McNamara, D.S. (2008) Assessing text readability using cognitively based indices. *TESOL Quarterly* 42 (3), 475-493.
- Graesser, A.C., McNamara, D.S., Louwerse, M.M. & Cai, Z. (2004) Coh-Metrix: Analysis of text on cohesion and language. *Behaviour Research Methods, Instruments, and Computers* 36, 193-202.
- Graesser , A.C., McNamara, D.S. & Kulikowich, J.M. (2011) Coh-Metrix: Providing multilevel analyses of text characteristics. Educational Researcher 40 (5) , 223-234.
- Green, A., Unaldi, A. & Weir, C.J. (2010) Empiricism versus connoisseurship: establishing the appropriacy of texts for testing reading for academic purposes. *Language Testing 27* (3), 1-21.
- Haberlandt, K.F. & Graesser, A.C. (1985) Component processes in text comprehension and some of their interactions. *Journal of Experimental Psychology: General* 114, 357-374.
- Just, M.A. & Carpenter, P.A. (1980) A theory of reading: from eye fixations to comprehension. *Psychological Review* 87 (4), 329-54.
- Just, M.A. & Carpenter, P.A. (1987) *The Psychology of Reading and Language Comprehension*. Boston: Allyn and Bacon.
- Khalifa, H. & Weir, C.J. (2009) *Examining Reading: Research and practice in assessing second language reading*. Cambridge: Cambridge University Press.
- Koda, K. (2005) *Insights into Second Language Reading*. New York: Cambridge University Press.
- Perfetti, C.A., Landi , N. & Oakhill , J. (2005) The acquisition of reading comprehension skill. In M.J. Snowling & C. Hulme (Eds.) *The Science of Reading: A Handbook*. Oxford: Blackwell. (pp.227-247).
- Rayner, K. & Pollatsek, A. (1989) *The Psychology of Reading*. Englewood Cliffs, UK: Prentice Hall.
- Rayner , K. & Pollatsek, A. (1994) *The Psychology of Reading* (2nd revised edn). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Shaw, S.D. & Weir, C.J. (2007) *Examining Writing: Research and practice in assessing second language writing*. Cambridge: Cambridge University Press.
- Templin, M (1957) *Certain Language Skills in Children: Their development and interrelationships*. Institute of Child Welfare Monograph Series No. 26. Minneapolis, MN: University of Minnesota Press
- Weir, C.J., Green, A., Chan, S., Taylor, L., Field, J., Nakatsuhara, F. & Bax, S. (2012) *Textual features of CAE reading texts compared with IELTS reading and essential undergraduate texts*. Interim report prepared for Cambridge ESOL Examinations.
- White, S. (2011) *Understanding Adult Functional Literacy: Connecting Text Features, Task Demands, and Respondent Skills.* Taylor & Francis.
- Wu, R. Y. (2011) *Establishing the validity of the General English Proficiency Test reading component through a critical evaluation on alignment with the Common European Framework of Reference*. Unpublished PhD dissertation, University of Bedfordshire.

Thank you very much for your interest and attention!

CRELLA
University of Bedfordshire