DEVELOPING RATING SCALES FOR INTEGRATED ASSESSMENT TASKS

Sathena Chan

EALTA Webinar

21 November 2017

OUTLINE

1. Definitions and examples of integrated assessment tasks
2. Important considerations for the development of a rating scale
    - characteristics of the test
    - the target construct
    - the rating approach
    - scaling descriptors
    - score reporting
    - rating training, trials, validation and improvement
3. Q&A

SOME DEFINITIONS

instructional tasks that **combine reading and writing** for various educational purposes (Ascension-Delaney, 2008, p.140)

a test that **integrates reading with writing** by having examinees read and respond to one or more source texts (Weigle, 2004, p.30)

a task which requires students to "produce written compositions that display appropriate and meaningful **uses of and orientations to source evidence**, both **conceptually** (in terms of apprehending, synthesising, and presenting source ideas) and **textually** (in terms of stylistic conventions for presenting, citing, and acknowledging sources)(Cumming et al, 2005, p.34)

COMMON INTEGRATED TASKS

- A summary task
- A response essay (reading / listening inputs)
- A situation-based integrated writing task
- A graph-writing task

SOME EXAMPLES

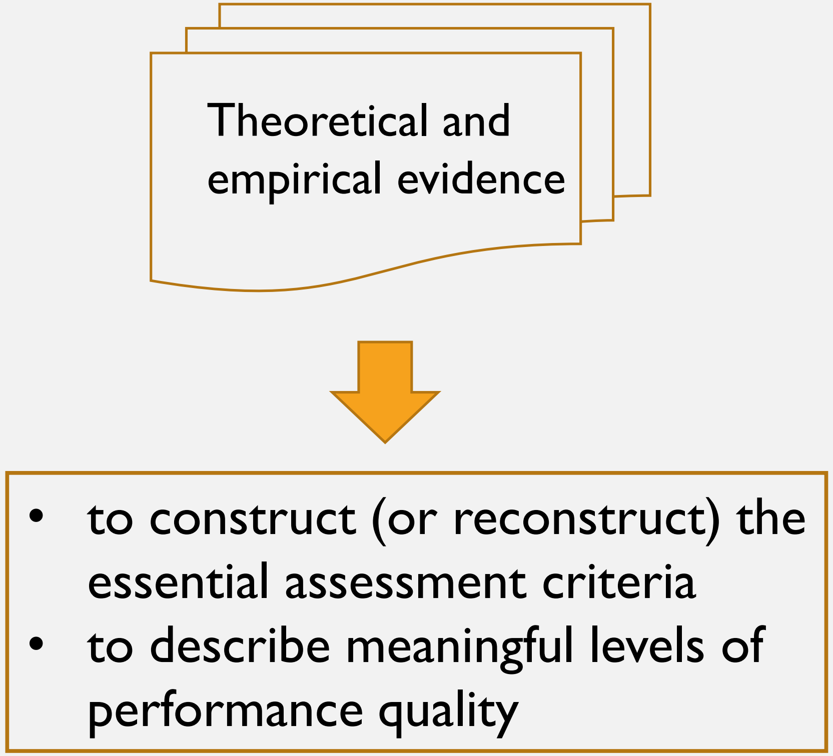
- TOEFL iBT (ETS, US):
  - Write an essay based on reading and listening tasks; 20 minutes;150-225 words; Summarise the points made in the lecture, being sure to explain how they oppose specific points made in the reading passage (https://www.ets.org/Media/Tests/TOEFL/pdf/SampleQuestions.pdf)
- General English Proficiency Test – Advanced (C1) (LTTC, Taiwan):
  - Write an essay based on two reading articles; 250 words; Summarise the main ideas of both texts and make clear your own viewpoint (https://www.lttc.ntu.edu.tw/GEPT1/Advanced/writing/writing.htm)
- Integrated Skills of English – ISE II (B2) (Trinity College London, UK):
  - Write an essay by using the information from four texts (3 written and 1 non-verbal) about an issue and its solutions; 150-180 words (http://www.trinitycollege.com/site/?id=3226)

GENERAL PRINCIPLES

An ‘armchair’ approach

intuitive judgements

A mixed-methods approach



Theoretical and empirical evidence

- to construct (or reconstruct) the essential assessment criteria
- to describe meaningful levels of performance quality

(Fulcher, 1996; McNamara, 1996; North, 2000; Shohamy, 1990; Upshur & Turner, 1995)

IMPORTANT CONSIDERATIONS WHEN DEVELOPING A RATING SCALE

1. What are the characteristics of the test? (e.g. purpose? level? format? stakeholder?)

2. What is the construct being measured? (e.g. the target cognitive processes? and the characteristics of test tasks?)

3. What rating approach will be used? (e.g. holistic? analytic? human scored? machine scored?)

4. What scaling descriptors will be needed? (e.g. number of levels and bands? descriptor styles?)

5. How will scores be reported? How will scores be used?

6. How will the rating scale be validated? (e.g. construct validity? reliability? practicality?)

(Knoch, 2009; Weigle, 2002)

1. CHARACTERISTICS OF THE TEST

Purpose?

Level(s)?

Format?

Stakeholders?

EXAMPLE A: ISE

- Aims of the ISE rating scale development project (see Chan et al., 2015)
  - to develop analytic criteria which address integrated reading-into-writing abilities that are not assessed on the independent writing-only task;
  - to develop a suite of level-specific scales (ISE F, I, II & III); and
  - to develop scaling descriptors for 4 possible bands (i.e. band 1, 2, 3 & 4) within each ISE level
- Features of the examination to be considered
  - High-stakes (for visa application purposes)
  - Validity and reliability
  - Stakeholders (e.g. Trinity, Test takers, Teachers, Regulatory authority)
  - Score reporting (e.g. technical mechanisms, marketing needs)

EXAMPLE B:
UOB ACADEMIC READING-INTO-WRITING TEST

- Aims
  - To offer **prompt assessment** and **targeted intervention** at an early stage in students' academic careers
  - Practical
  - Cost effective
  - Tailored to assessing the language related academic study skills
  - Lead to targeted support for students
- Features of the test to be considered
  - C1 level: post university entry
  - Low-stakes
  - Stakeholders (e.g. the University, students, lecturers, language centre)
  - Score reporting (for decision making & diagnostic purposes)

2. THE TARGET CONSTRUCT

The features and demands of the test tasks?

The target cognitive processes?

CONTEXTUAL PARAMETERS FOR READING-INTO-WRITING TASKS

Overall task setting

- Time and length
- Purpose
- Topic domain
- Genre
- Scope of interaction between input and output
- Language functions to perform
- Clarity of intended reader
- Knowledge of criteria

Input text features

- Input format (e.g. single, multiple, text-based, graph based inputs)
- Genre (e.g., articles, reports, case studies, abstracts, proposals)
- Non-verbal input (e.g., diagram, tables, charts)
- Discourse mode
- Concreteness of ideas
- Explicitness of textual organisation
- Lexical complexity
- Syntactic complexity
- Degree of cohesion

(Chan, 2018; Shaw and Weir; 2007)

EXTERNAL REFERENCES

The CEFR descriptors (CoE, 2001)

- Only 4 of the 12 scales which focus on reading and writing make specific reference to integrated use of skills
  - *overall written production* (p.61)
  - *reports and essays* (p.62)
  - overall *reading comprehension* (p.69)
  - *processing text* (p.96)
- You may also check *the interaction scales* on speaking

PROCESSING TEXT SCALE

A1:

- Can copy out single words and short texts presented in standard printed format.

A2:

- Can copy out short texts in printed or clearly handwritten format.
- Can pick out and reproduce key words and phrases or short sentences from a short text within the learner's limited competence and experience.

B1

- Can collate short pieces of information from several sources and summarise them for somebody else.
- Can paraphrase short written passages in a simple fashion, using the original text wording and ordering.

B2

- Can summarise a wide range of factual and imaginative texts, commenting on and discussing contrasting points of view and the main themes.
- Can summarise extracts from news items, interviews or documentaries containing opinions, argument and discussion.
- Can summarise the plot and sequence of events in a film or play.

C1:

- Can summarise long, demanding texts.

C2:

- Can summarise information from different sources, reconstructing arguments and accounts in a coherent presentation of the overall result.

READING-INTO-WRITING PROCESSES

- Conceptualisation (e.g. task representation, macro-planning)
- Meaning construction (e.g. mining/selecting relevant ideas, connecting ideas across texts)
- Organisation (e.g. organising ideas into a compositional structure)
- Translation/Execution (e.g. transforming the language used in the source text)
- Monitoring and Revising

(Chan, 2018; Field, 2004; Flower & Hayes, 1983; Hayes, 1996; Shaw & Weir, 2007; Spivey and King, 1989)

3. SCORING APPROACH

holistic?

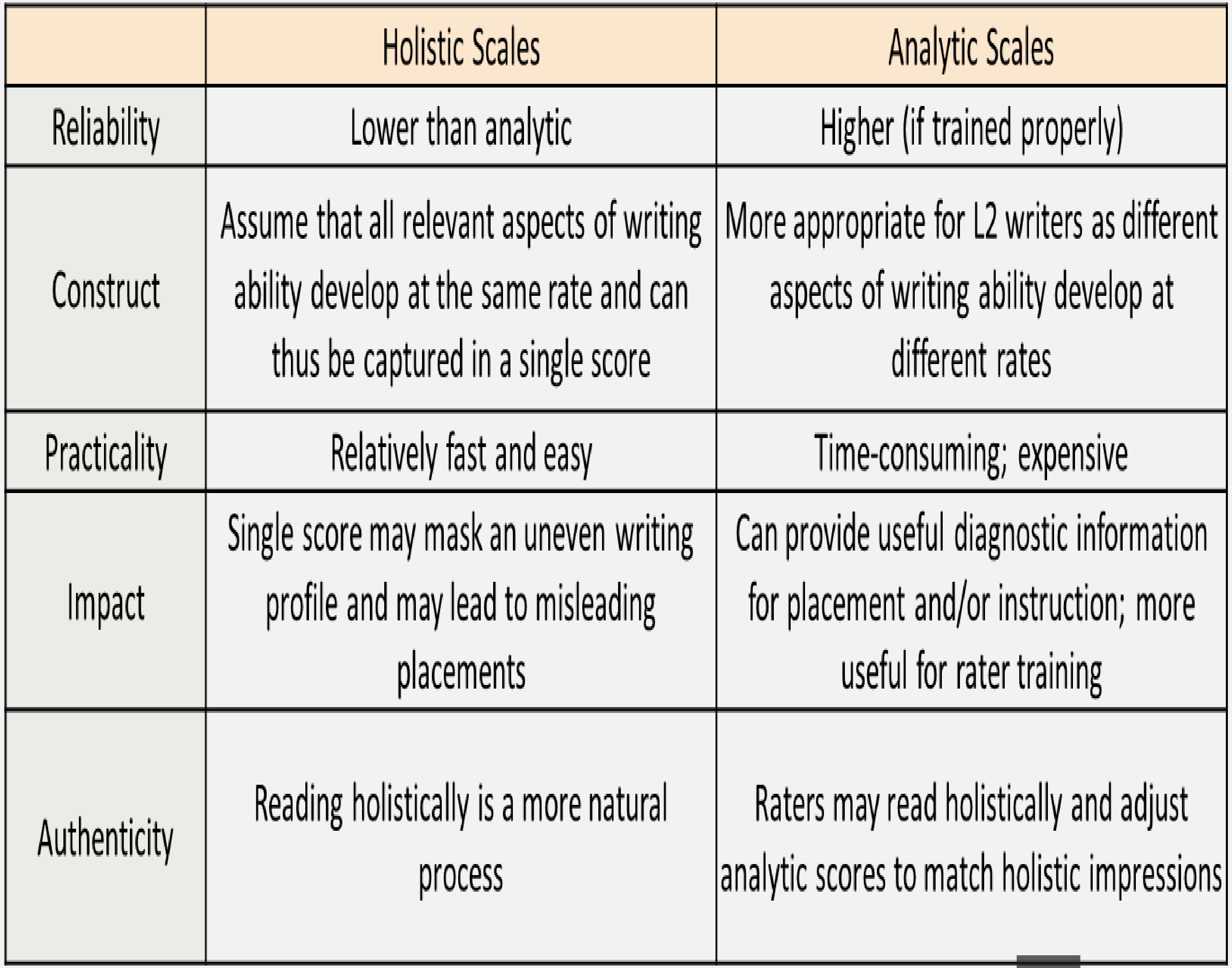
analytic?

human scored?

machine scored?

TYPES OF RATING SCALES

- **Holistic**: Give a single score to the written response as a whole
- **Analytic**: Give a single score for each rating category; students receive several scores for a response



	Holistic Scales	Analytic Scales
Reliability	Lower than analytic	Higher (if trained properly)
Construct	Assume that all relevant aspects of writing ability develop at the same rate and can thus be captured in a single score	More appropriate for L2 writers as different aspects of writing ability develop at different rates
Practicality	Relatively fast and easy	Time-consuming; expensive
Impact	Single score may mask an uneven writing profile and may lead to misleading placements	Can provide useful diagnostic information for placement and/or instruction; more useful for rater training
Authenticity	Reading holistically is a more natural process	Raters may read holistically and adjust analytic scores to match holistic impressions

(Weigle, 2002, p.121)

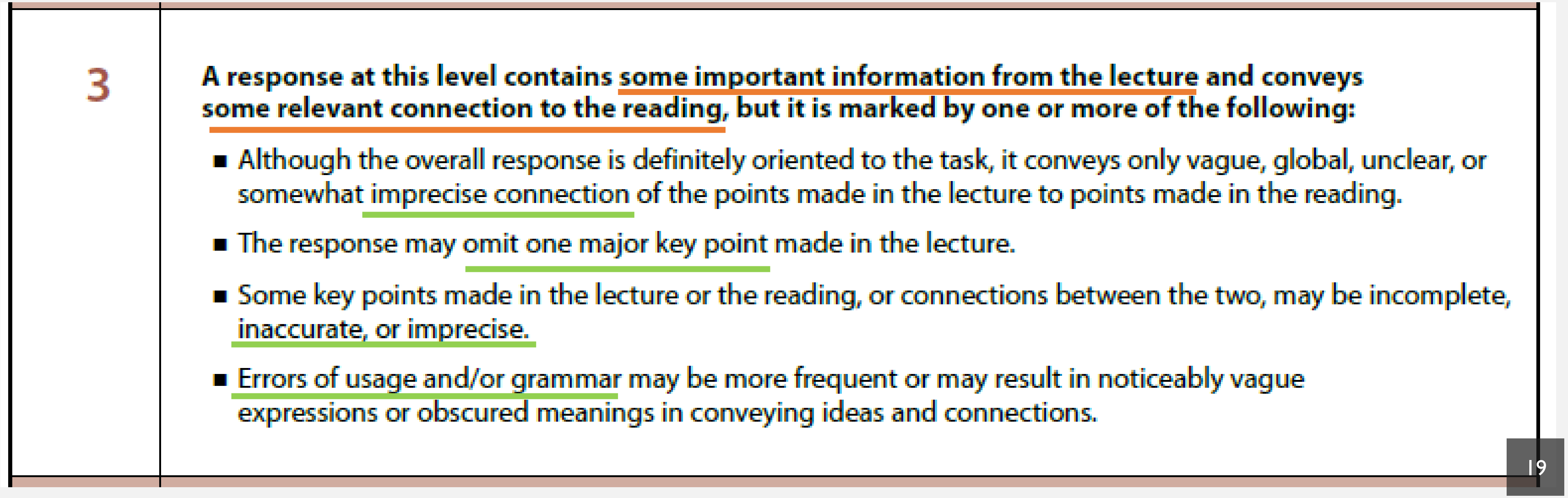
RATING SCALE FEATURES FOR INTEGRATED TASKS

- It is important that the rating scales measure the construct of skill integration and provide a working definition for the users of the scale
- To account for the transformation that has taken place in the language from source text to the final written product
  - Content
  - Language
  - Organisational structure
  - Cohesion
  - Acknowledgement of sources

(Knoch & Sitajalabhorn, 2013)

HOLISTIC SCALES - EXAMPLE

- TOEFL iBT integrated writing rubrics (1-5) http://www.ets.org/s/toefl/pdf/toefl_writing_rubrics.pdf



3	**A response at this level contains some important information from the lecture and conveys some relevant connection to the reading, but it is marked by one or more of the following:** ■ Although the overall response is definitely oriented to the task, it conveys only vague, global, unclear, or somewhat imprecise connection of the points made in the lecture to points made in the reading. ■ The response may omit one major key point made in the lecture. ■ Some key points made in the lecture or the reading, or connections between the two, may be incomplete, inaccurate, or imprecise. ■ Errors of usage and/or grammar may be more frequent or may result in noticeably vague expressions or obscured meanings in conveying ideas and connections.

4. SCALING DESCRIPTORS

- number of levels?
- number of bands?
- descriptor styles?

USEFUL MATERIALS

The following materials are useful when writing descriptors:

- the CEFR descriptors
- previous rating scales of the test
- other established rating scales
- empirical evidence in the literature (e.g. language features which are salient to each level of proficiency)
- analysis of actual scripts
  - to identify features that differentiate strong performance from weak performance at each level
  - to identify excerpts to exemplify the descriptors (for training purposes)

FEATURES OF INTEGRATED PERFORMANCE

	Higher-scoring performance	Lower-scoring performance
Content	• included important ideas from all sources • summarised source information	• included ideas mainly from a single source • made declarations based on personal knowledge
Language	• showed evidence of paraphrases	• consisted direct copying of words and phrases
Organisation	• had a more appropriate organisation	• tended to follow the structure of one of the sources
Acknowledgement of sources	• indicated sources of information	• inappropriate source use

(Cumming et al., 2005; Flower et al., 1990; Plakans & Gebril, 2013; Watanabe, 2001)

EXAMPLE A. ISE

- The rating scale has 4 analytic criteria
  - *Reading for Writing*
  - *Task Fulfilment*
  - *Organisation and Structure*
  - *Language Control*
- Each analytic category has several sub-categories, e.g.
  - *Reading for Writing*
    - Understanding of source materials
    - Selection of relevant content from source texts
    - Ability to identify common themes and links within and across the multiple texts
    - Adaptation of content to suit the purpose of writing
    - Use of paraphrasing/summarising

(http://www.trinitycollege.com/site/?id=3634)

EXAMPLE A. ISE

- Each analytic category has 4 bands of descriptors:

  4 = Strong performance at the level (possibly above the level)

  3 = Good performance at the level

  2 = Adequate performance at the level

  1 = Inadequate performance at the level (below the level)

(http://www.trinitycollege.com/site/?id=3634)

EXAMPLE A.
ISE II (B2) BAND 3
READING FOR WRITING DESCRIPTORS

- Full and accurate understanding of the essential meaning of most source materials demonstrated
- An appropriate and accurate selection of relevant content from the source texts (i.e., most relevant ideas are selected and most ideas selected are relevant)
- Good ability to identify common themes and links within and across the multiple texts and the writers' stances
- A good adaptation of content to suit the purpose for writing (e.g., apply the content of the source texts appropriately to offer solutions, offer some evaluation of the ideas based on the purpose for writing)
- Good paraphrasing/summarising skills of factual ideas, opinions, argument and/or discussion demonstrated (with very limited lifting and few disconnected ideas)

(http://www.trinitycollege.com/site/?id=3634)

EXAMPLE B.
UOB ACADEMIC READING-INTO-WRITING TEST

- The scale has 3 analytic criteria
  - Relevance and adequacy of content (coverage of main ideas)
  - (Compositional) Organisation (cohesion and coherence)
  - Language (source use, choice and control of lexis, grammar)
- Each criterion has three bands to identify students who need
  - Green (Band A): no additional systematic remedial intervention
  - Amber (Band B): some additional systematic remedial intervention
  - Red (Band C): a substantial level of systematic remedial intervention

EXAMPLE B.
RELEVANCE AND ADEQUACY OF CONTENT

- This refers to the extent to which the writer has responded appropriately to the task. It covers the need to address the **4 essential points** required for the essay as stated in the task rubric, i.e., providing an appropriate title; identifying the issue; summarising the main ideas and explaining which main idea is most significant and why.
- It also deals with the **communicative effect** of the writing on the reader (i.e. awareness of writer-reader relationship and appropriate level of formality).

EXAMPLE B.
RELEVANCE AND ADEQUACY OF CONTENT

Band	Score	**Descriptor of performance quality**	Meaning
A	**3**	• Relevant and fully adequate response to the task. • All 4 key points required in the task included and expanded appropriately. • Achieves desired communicative effect on target reader.	an *adequate* performance, by a student who should not need additional EL/study skills support post entry
B	**2**	• Partially successful response to the task. • One or two key points inadequately covered or omitted, and/or some irrelevant material included. • May fail to communicate clearly to target reader and/or achieve the desired effect.	a *below adequate* performance, by a student who will benefit from some targeted EL/study skills intervention post entry
C	**1**	• Limited response to the task. • More than 2 key points omitted and/or considerable irrelevance/repetition, possibly due to misinterpretation of the task. • Fails to achieve the desired effect because considerable effort will be required of the reader.	a *significantly weak* performance, by a student who will be a high-priority candidate for substantial EL/study skills intervention post entry.
U	0	• Chunks of language have been copied / plagiarised • Too little language to form judgement	As above

5. SCORE REPORTING

How will scores be reported?

- an overall score
- analytic scores
- a certificate
- a diagnostic profile http://www.trinitycollege.com/site/?id=3484

How will scores be used?

- high-stakes vs low-stakes
- decision making
- feedback to students

6. TRIALS, VALIDATION AND IMPROVEMENT

- Analysis of rating data
- Raters’ processes
- Analysis of writing scripts
- Feedback, e.g. practicality, usefulness, clarity

ANALYSIS OF RATING DATA

- Approaches
  - Classical Test Theory (CTT)
  - Multi-Faceted Rasch Analysis (Linacre, 2006)
- Aspects:
  - Discrimination of the rating scale
  - Rater reliability
  - Variation in ratings, e.g. severity, variations across versions/modes

ANALYSIS OF WRITING SCRIPTS

Ranking and commentary by raters, e.g.

1. ranking the scripts into several piles e.g. *at the level* and *below the level*
2. selecting extracts from the script pool to exemplify the ‘at the level’ and ‘*below the level*’ performance
3. providing a rationale for the selection

Automated text analysis tools

- VOCABPROFILE ENGLISH - Compleat Lexical Tutor
- Coh-Metrix
- Text Inspector

RATER'S PROCESSES

Common procedures

- reading the task instruction
- reading the rubrics
- reading the source texts
- identifying relevant parts in the source texts
- reading the script
- assigning a score to each criterion
- checking the source texts
- checking the rubrics and reconsidering the assigned scores

(Chan et al., 2015, p.28)

Attention paid to

- accuracy of source use
- relevance of source use
- adequacy of source use (i.e. is it enough?)
- clarity of source information
- appropriateness of textual borrowing strategies (Is this good paraphrasing? Is it patch writing? . . .etc.)
- effectiveness of source use (i.e. does it really support the ideas? is it in the right place?)
- overuse of source materials

(Gebril and Plakans, 2014, p.63)

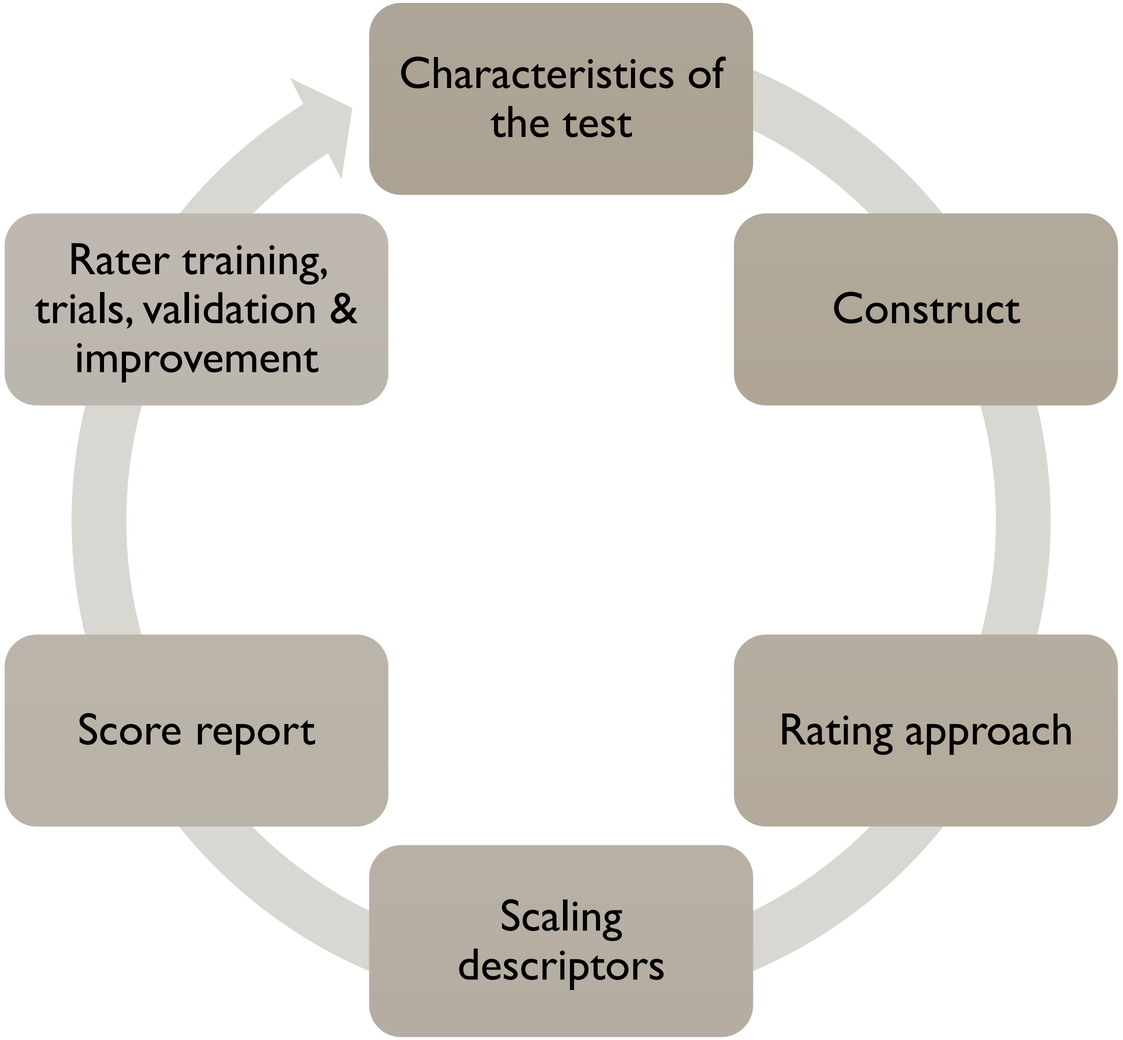
FEEDBACK – CHALLENGES AND SOLUTIONS

- Task specificity (need to be familiar with the source texts) + time consuming
- Locating source information
  - Hard to distinguish language cited from source materials vs language produced by the writers
- Quality of source use
  - Difficult to determine different levels of text integration
- Textual borrowing/citation mechanics
  - how much copying is allowed?
  - how many quotations (reflecting inappropriate textual practices) are allowed?
- Difficult terminology and inconsistent adjectives

- To provide a list of the relevant ideas in the rater training pack
- To quantify some of the descriptors to make the indicated requirements more transparent
  - inadequate selection of relevant content from the source texts MEANS fewer than half of the relevant ideas are selected
- To provide examples of direct quotes, paraphrases and summaries at each level
- Be specific about expectations
  - e.g. You will fail if you copy chunks (i.e. more than 3 continuous words in a sentence) from the articles

Chan, et al., 2015; Cumming, et al. 2001; Gebril and Plakans, 2014)

RECAP: IMPORTANT CONSIDERATIONS WHEN DEVELOPING A RATING SCALE



Characteristics of the test

Construct

Rating approach

Scaling descriptors

Score report

Rater training, trials, validation & improvement

USEFUL REFERENCES

- Chan, S. H. C. (2017). Using keystroke logging to understand writers' processes on a reading-into-writing test. *Language Testing in Asia., 7(10), 1-27.*
- Chan, S. H. C. (2018 in publication). *Defining Integrated Reading-into-Writing Constructs: Evidence at the B2 C1 Interface.* English Profile Series Studies. Cambridge University Press.
- Chan, S. H. C., Inoue, C. and Taylor, L. (2015) Developing rubrics to assess the skill of reading-into-writing: a case study. *Assessing Writing*, 26, 20-37.
- Cumming, A. (2013). Assessing integrated writing tasks for academic purposes: Promises and perils. Language Assessment Quarterly, *10*(1), 1–8.
- Cumming, A., Kantor, R., Baba, K., Eouanzoui, K., Erdosy, U., & James, M. (2006). Analysis of discourse features and verification of scoring levels for independent and integrated tasks for the new TOEFL. Princeton, NJ: ETS (TOEFL Monograph No. MS-30 Rm 05-13).
- Cumming, A., Kantor, R., Baba, K., Erdosy, U., Eouanzoui, K., & James, M. (2005). Differences in written discourse in independent and integrated prototype tasks for next generation TOEFL. Assessing Writing, 10, 5–43
- Gebril, A., & Plakans, L. (2014). Assembling validity evidence for assessing academic writing: Rater reactions to integrated tasks. Assessing Writing, 21, 56–73.
- Knoch, U. (2011). Rating scales for diagnostic assessment of writing: What should they look like and where should the criteria come from? Assessing Writing, 16, 81–96
- Knoch, U., & Sitajalabhorn, W. (2013). A closer look at integrated writing tasks: Towards a more focused definition for assessment purposes. *Assessing Writing*, *18*(4), 300–308.
- Shaw, S., & Weir, C. J. (2007). *Examining writing: Research and practice in assessing second language writing, Studies in Language Testing 26*. Cambridge: UCLES/Cambridge University Press.
- Plakans, L. (2013). Writing scale development and use in a language program. TESOL Journal, 4, 151–163.
- Weigle, S. (2004). Integrating reading and writing in a competency test for non-native speakers of English. Assessing Writing, 9, 27–55.
- Weigle, S. C., & Parker, K. (2012). Source text borrowing in an integrated reading/writing assessment. *Journal of Second Language Writing*, *21*(2), 118-133.

Thank you!

sathena.chan@beds.ac.uk