LTF 2019

Performance based decision trees vs. traditional criteria: an exploratory study

Elena A.M. Gandini

Dr Tania Horák

School of
LANGUAGE AND
GLOBAL STUDIES

uclan
University of Central Lancashire

Presentation overview

- The impetus for the research
- The study
- The outcomes

The scenario

LANGUAGE ACADEMY

- Pre- and in-sessional EAP courses
- 4 skills tested
- Large numbers of scripts to mark
- Time pressure
- Constant changes in the team of teachers/markers
- % inexperienced teachers
- Currently using typical grid-style marking criteria
- Current criteria due for overhaul

The issues

CRITERIA

- Ambiguity of the wording
- Gradients - ‘Just filling the boxes’
- On-going attempts to improve criteria used for writing

SCORING

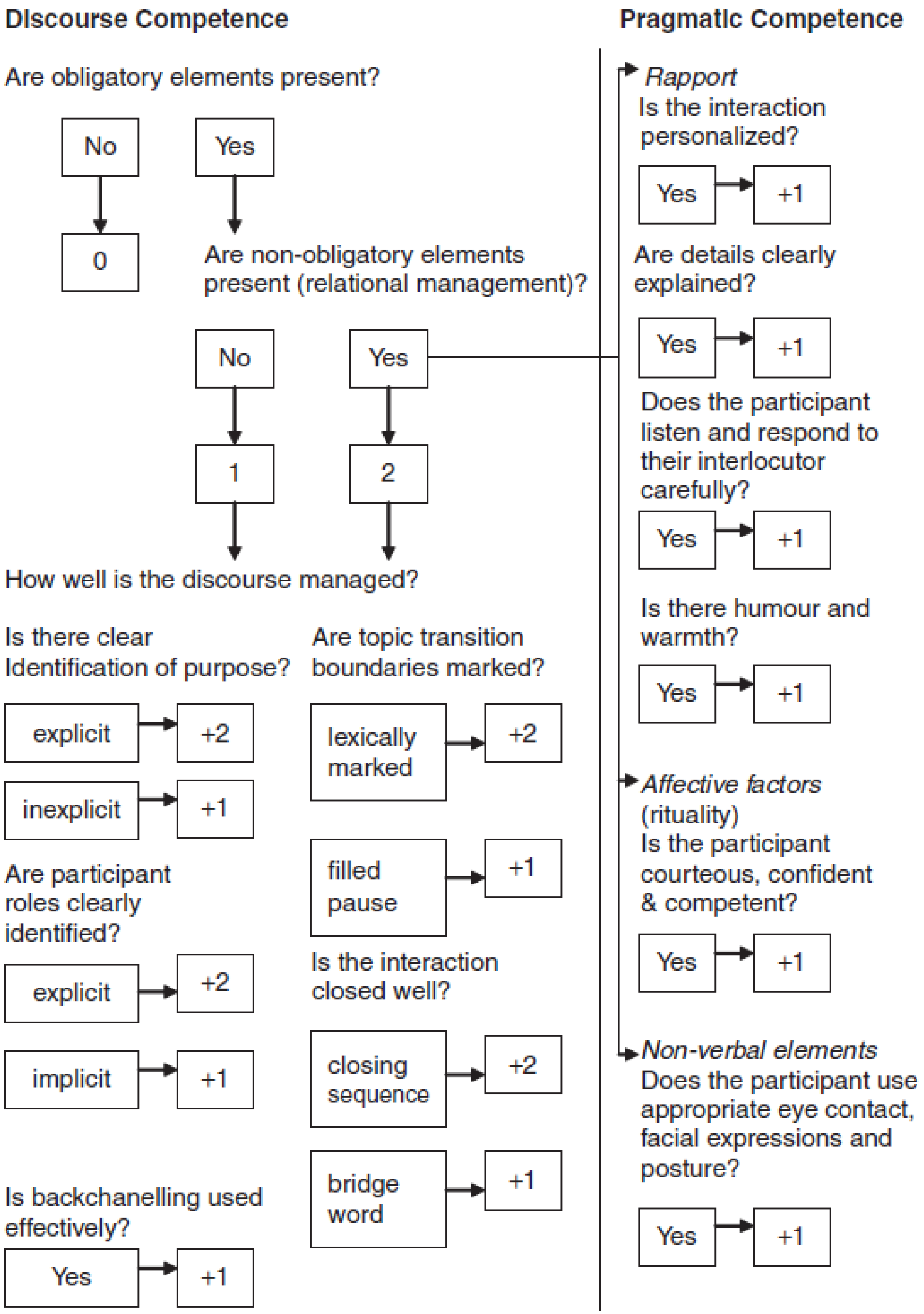
- Standardisation slippage
- Inter-rater reliability threatened
- All criteria being considered consistently?

Search for alternatives which can promote better feedback and washback

Performance Decision Trees

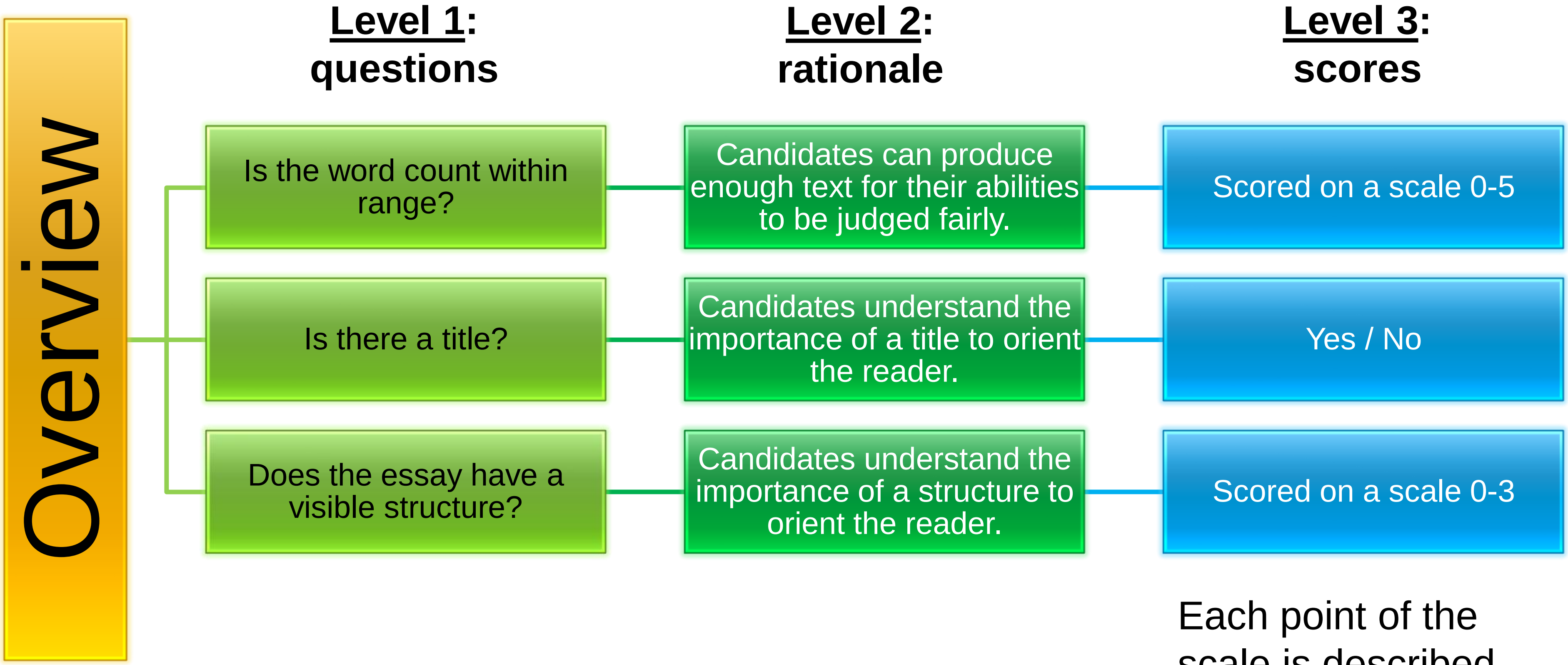
- Fulcher, Davidson & Kemp (2010)
- Two models for rating scales:
  - Measurement-driven approach
  - *Performance-driven approach*
- Main features: questions posed re certain aspects of performance
- Binary decision-making
- NB original proposed for a speaking test → ours is an adaptation of the concept for a writing test

The idea


Discourse Competence
Are obligatory elements present?
No
Yes
0
Are non-obligatory elements present (relational management)?
No
Yes
1
2
How well is the discourse managed?
Is there clear Identification of purpose?
explicit
+2
inexplicit
+1
Are participant roles clearly identified?
explicit
+2
implicit
+1
Is backchannelling used effectively?
Yes
+1
Are topic transition boundaries marked?
lexically marked
+2
filled pause
+1
Is the interaction closed well?
closing sequence
+2
bridge word
+1
Pragmatic Competence
Rapport
Is the interaction personalized?
Yes
+1
Are details clearly explained?
Yes
+1
Does the participant listen and respond to their interlocutor carefully?
Yes
+1
Is there humour and warmth?
Yes
+1
Affective factors (rituality)
Is the participant courteous, confident & competent?
Yes
+1
Non-verbal elements
Does the participant use appropriate eye contact, facial expressions and posture?
Yes
+1

Fulcher, Davidson & Kemp (2010)

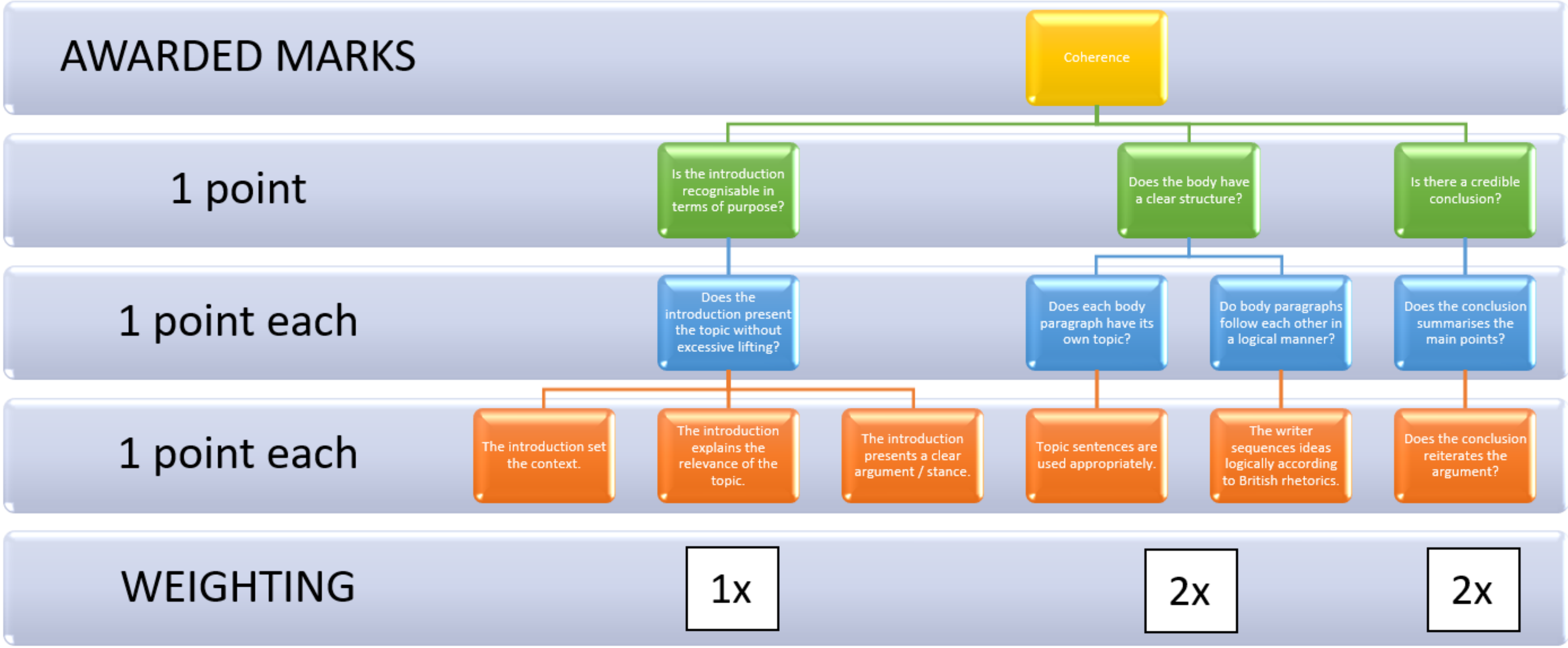
Different attempts:



Overview

Level 1: questions	Level 2: rationale	Level 3: scores
Is the word count within range?	Candidates can produce enough text for their abilities to be judged fairly.	Scored on a scale 0-5
Is there a title?	Candidates understand the importance of a title to orient the reader.	Yes / No
Does the essay have a visible structure?	Candidates understand the importance of a structure to orient the reader.	Scored on a scale 0-3

Each point of the scale is described and explained.



AWARDED MARKS

Coherence

1 point

Is the introduction recognisable in terms of purpose?

Does the body have a clear structure?

Is there a credible conclusion?

1 point each

Does the introduction present the topic without excessive lifting?

Does each body paragraph have its own topic?

Do body paragraphs follow each other in a logical manner?

Does the conclusion summarises the main points?

1 point each

The introduction set the context.

The introduction explains the relevance of the topic.

The introduction presents a clear argument / stance.

Topic sentences are used appropriately.

The writer sequences ideas logically according to British rhetorics.

Does the conclusion reiterates the argument?

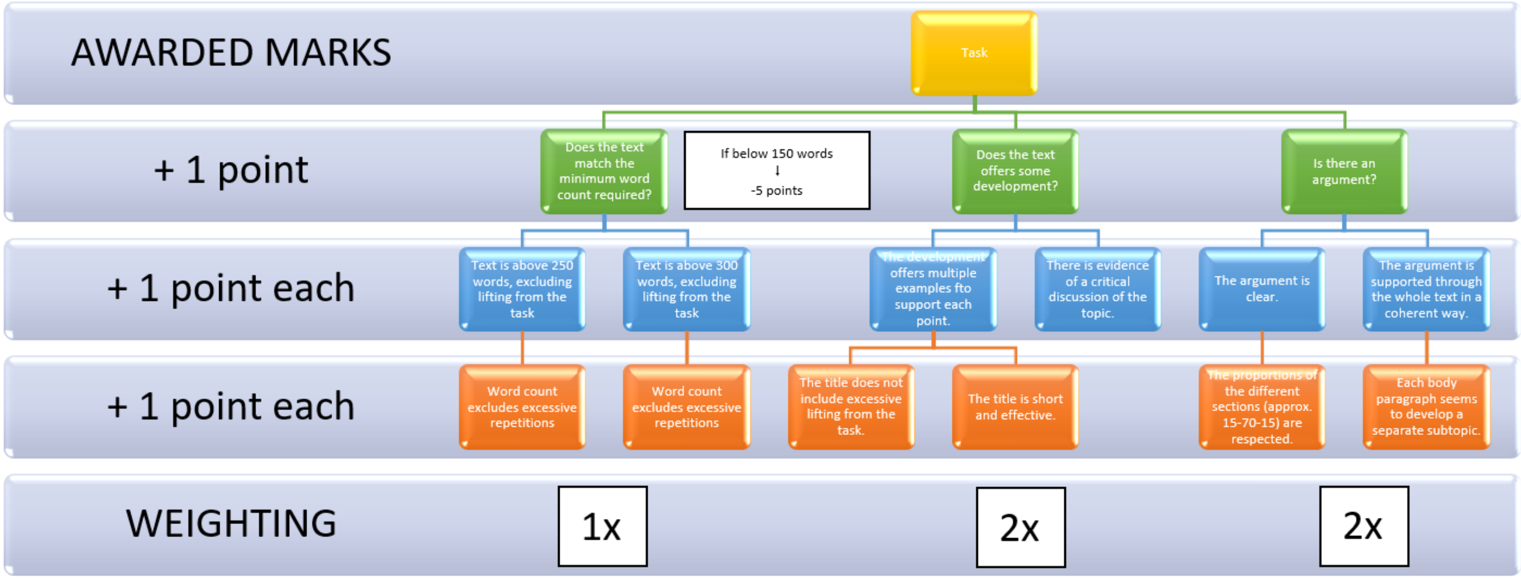
WEIGHTING

1x

2x

2x

TASK	To what extent does the text match the task requirements?						
	5	4	3	2	1	0	
• Does the text match the required wordcount?							• The text is below the required wordcount.
• Is there a clear argument / stance?							• No clear argument can be identified.
• Is the topic well developed throughout?							• Very little to no development of the topic.



AWARDED MARKS

Task

+ 1 point

Does the text match the minimum word count required?

If below 150 words
↓
-5 points

Does the text offers some development?

Is there an argument?

+ 1 point each

Text is above 250 words, excluding lifting from the task

Text is above 300 words, excluding lifting from the task

The development offers multiple examples fto support each point.

There is evidence of a critical discussion of the topic.

The argument is clear.

The argument is supported through the whole text in a coherent way.

+ 1 point each

Word count excludes excessive repetitions

Word count excludes excessive repetitions

The title does not include excessive lifting from the task.

The title is short and effective.

The proportions of the different sections (approx. 15-70-15) are respected.

Each body paragraph seems to develop a separate subtopic.

WEIGHTING

1x

2x

2x

	Word count	**Development**	**Argument**	**TOTAL Weighted**
TASK TOTAL				**/25**

Overview - presentation
Is there a relevant title?
Is the classical essay structure obvious (i.e. introduction, body paragraphs, conclusion)?
Is the body of the essay divided into multiple (2 or more) paragraphs?
Are words written sitting on the line (aiding easy reading)?
Are individual words easily distinguishable one from the next?
Is the type of genre clearly recognisable from the task (i.e. this is an essay, not a report, or letter)?
Record your score for this section on the Excel sheet
Content/ Argumentation
Task fulfilment - Which of the following best describes this essay? Pick *one* ○ The candidate has addressed the task fully ○ The candidate has mostly addressed the task but some parts maybe a little off topic. ○ The essay bears little relation to the set task - it is mostly off topic ○ The essay is entirely irrelevant to the set task - it is totally off topic
Have the main points of the candidate's argument been developed effectively through the use of examples and /or evidence?
Does the candidate take a clear stance on the topic? Is their argument clear?
Record your score for this section on the Excel sheet
Coherence (overall sense)
Does the introduction set out what will be discussed?
Does each of the body paragraphs have its own specific topic?

Performance Decision Trees

- Adaptation: initial trials – branching harder to navigate
- Outcome: a checklist rather than a tree

- Aim: more guided marking, all component considered
- Benefit: usable with students, facilitates feedback

Our hypothesis

- Analytic marking will increase both inter-rater & intra-rater reliability (Goulden 1994)
- Value of analytic scales: Bacha (2001)

- A Decision-tree-influenced / Checklist-style criteria will improve quality and consistency of feedback to Ss
- And will improve washback – components of good performance easier to identify and teach/ study

Research questions

- How would raters perceive the two types of criteria when asked to implement both?
- Would raters recognise any advantages to checklist-style criteria?
- Would inexperienced raters be provided with more guidance by the checklist-style criteria?

Criteria production

- Key unique elements in the extant grid-style criteria identified (colour coded)
- List of key elements produced
- Rewordings of single feature (e.g. basic / good / excellent command; basic / wide repertoire, etc.) ignored
- Questions formed around them
- Checked again against grid
- Grouped according to outcomes of previous stage of research re. categories

Criteria comparison

GRID

3 main categories:

- Content & Appropriacy
- Organisation & Cohesion
- Language (Grammar, Vocabulary, Spelling & Punctuation)

CHECKLIST

7 main categories:

- Overall presentation
- Content & Argumentation
- Cohesion
- Coherence
- Accuracy / Appropriacy
- Range
- Academic register and style

NB

No new components in the checklist, just re-organised

The study

- Very small-scale pilot
- Convenience sample – 25 participants
  - 77.77% return rate (28/32)
  - 3 respondents did not follow the instructions → discarded
- Profile: Qualified EFL tutors – range of experience
- Each pp graded 8 essays
  - 4 using grid-style criteria
  - 4 using checklist-style criteria
  - 1 benchmarked essay for reference

The study

Each essay marked using both styles of criteria

10 pt. max ↰

↱ *5 pt. max*
in each category

	Grid (holistic)	**Checklist (analytic)**
Group 1 (n=14)	Essays 1-4	Essays 5-8
Group 2 (n=11)	Essays 5-8	Essays 1-4

The data

QUALITATIVE:

QUESTIONNAIRES

- Open Qs: perceptions of each method (Barkaoui 2007)
- MCQs: bio data, teaching and examining experience (for comparison purposes)

QUANTITATIVE:

SCORE SHEETS

- Grid
- Checklist

Analysis

QUALITATIVE

- Basic thematic analysis on questionnaires
- Raters grouped according to the bio data:
    - EAP experienced
    - General EFL experienced
    - Inexperienced

QUANTITATIVE

- Analysis of scores in Excel and SPSS (still ongoing)
- Comparison of scores (grid vs checklist)
- Score range
- Inter-rater reliability
- Raters grouped according to:
    - Experience level
    - Criteria style preference

Results

Preferences:

44% Grid

Holistic marking

It’s something I’m more familiar with

44% Checklist

Reminds the marker of everything they need to consider

More systematic

12% Uncertain

Grid for an overall grade, checklist to identify specific features in the writing

Results - PPs' perceptions

Key themes – Reported of **both** styles!

- Cognitive load – too much info to juggle
- Reduces subjectivity
- Quicker
- Visual effect / presentation
- Detailed / Not detailed enough

Cognitive load

PP18 - Checklist:

'*Too many questions to consider and to keep in mind. For every essay I had to read them again*'

PP17 - Grid

'*Too wordy and tiring to read / remember*'

Time

PP8 – Checklist:

‘I think it is easier and it doesn’t take as much time to mark an essay as does the grid-style.’

PP26 – Grid:

‘Time-saving’

PP21 – Grid:

*‘Much quicke*r'

Subjectivity/1

PP1 – Grid:

'*I felt the grid was easier to use and less subjective*.'

PP25 – Grid:

‘It gives … a more comprehensive overview of the criteria … making marking less arbitrary.’

Subjectivity/2

PP4 – Checklist:

‘*Helps you to remember aspects you might have forgotten about ... and so this reduces subjectivity’*

PP13 – Checklist:

‘I felt that using the check-list style marking criteria leaves less room for interpretation and subjectivity on the examiner’s part’

PP18 – Checklist:

‘*Possibly fairer and more objective’*

Problems with the wording

PP23 – Checklist:

‘*Some of the sections had overlapping questions that had already been addressed in previous sections, or the wording was very similar*.’

PP23 – Grid:

‘Some of the sections within the criteria … had very similar wording, which made it hard to match students’ varying skills to a grade.’

Systematicity of the checklist

PP14:

‘Helps markers think about the micro details of each criteria section and what contributes to each – e.g. which elements exactly contribute to accuracy’

PP17:

‘Easier to focus on criteria in turn’

Emotional level

PP9 - Checklist:

‘I didn’t feel “responsible” for the final result. The questions led me so I might have been less biased’

PP24 - Checklist:

‘It seems more thorough and precise. I felt more confident with the marks I gave’

PP28 – Checklist:

‘Helped me to be confident because I could justify my marks’

Potential for feedback of the checklist

PP4 – Checklist:

‘*Might seem more transparent to students / stakeholders*’

PP7 – Checklist:

‘*Good for diagnostic info for student*’

Quantitative results

- Checklist ratings are generally higher than grid scores
- Range of scores on some essays remains quite wide with both criteria (due to a lack of standardisation)
- Mini pilots on the previous versions of the criteria suggest that through standardisation checklist / decision tree can improve inter-rater reliability (this time seems rather poor – as per our expectations)
- Will be compared with new results from the future stages of the project

[Cf Barkoui, 2007 – effects of different scales on rating]

Limitations

- No standardisation training
- Checklist criteria still require interpretation
- Limited number of essays marked
- Time burden on research participants

The outcomes – going further…

- Standardisation training still vital for final version
- Resistance to change – many experienced raters familiar with grid-style

- Support for new style checklist (both inexperienced and experienced raters)
- Potential for better feedback to Ss recognised

The outcomes – going further…

- Proposed amendments:
  - Reduce the level interpretation – points per question
  - Provide greater clarity about the exam scenario – pps claimed to use different rating styles (high v low stakes)

- Continue to investigate how raters interpret the two set of criteria

References

- Bacha, N (2001) Writing evaluation: what can analytic versus holistic essay scoring tell us? *System,* 28, 371-383
- Fulcher, G, Davidson, F & Kemp, J (2010) Effective rating scale development for speaking tests: Performance decision trees, *Language Testing,* 28 (1) 5-29.
- Goulden, N. R. (1994). Relationship of analytic and holistic methods to rater’s scores for speeches. *The Journal of Research and Development in Education*, 27, 73–82.

Tania
thorak@uclan.ac.uk

Elena
eamgandini@uclan.ac.uk