NOTTINGHAM TRENT UNIVERSITY

Using Assessment Criteria to Evaluate Course Quality and Inform Materials Design

Sam Barclay

BALEAP Conference 2019

University of Leeds

Overview

- Retrospective evaluation as part of course design and quality assurance
- Evaluation tools – assessment criteria
- An initial foray
- Matrix construction to facilitate formative evaluation
- Adaptations to the instrument and methodology
- Findings and implications of the current study
- Advice for implementing similar evaluations

Overview

- Retrospective evaluation as part of course design and quality assurance
- Evaluation tools – assessment criteria
- An initial foray
- Matrix construction to facilitate formative evaluation
- Adaptations to the instrument and methodology
- Findings and implications of the current study
- Advice for implementing similar evaluations

Overview

- Retrospective evaluation as part of course design and quality assurance
- Evaluation tools – assessment criteria
- An initial foray
- Matrix construction to facilitate formative evaluation
- Adaptations to the instrument and methodology
- Findings and implications of the current study
- Advice for implementing similar evaluations

Overview

- Retrospective evaluation as part of course design and quality assurance
- Evaluation tools – assessment criteria
- An initial foray
- Matrix construction to facilitate formative evaluation
- Adaptations to the instrument and methodology
- Findings and implications of the current study
- Advice for implementing similar evaluations

❖ Understanding of rationale and procedural knowledge to conduct similar analyses of your courses.

Background – Educational Evaluation

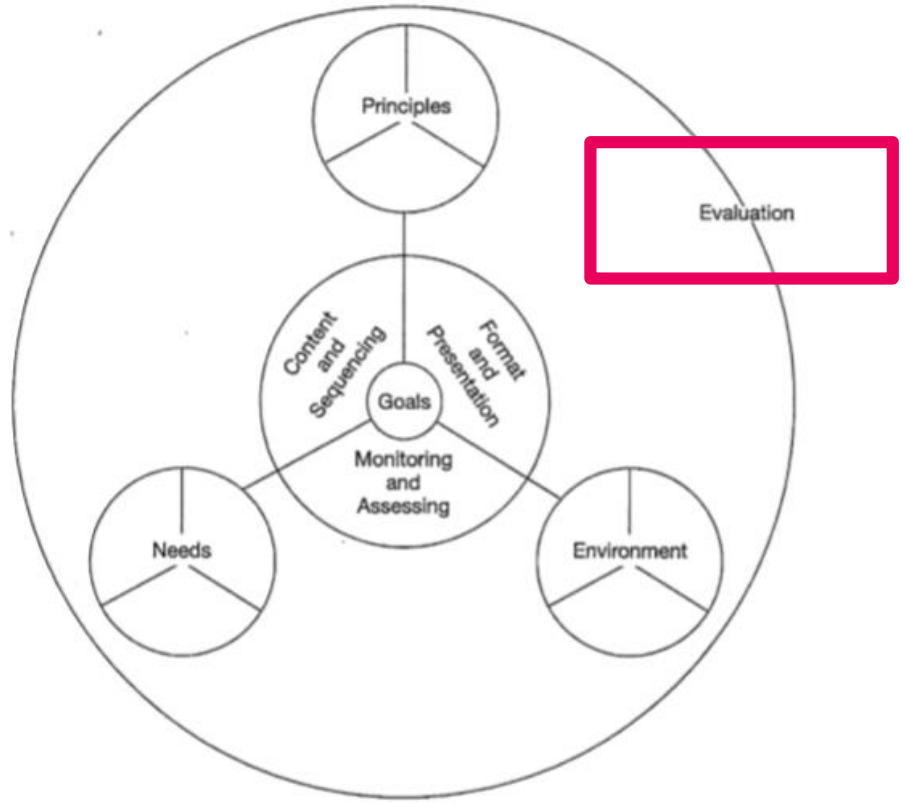
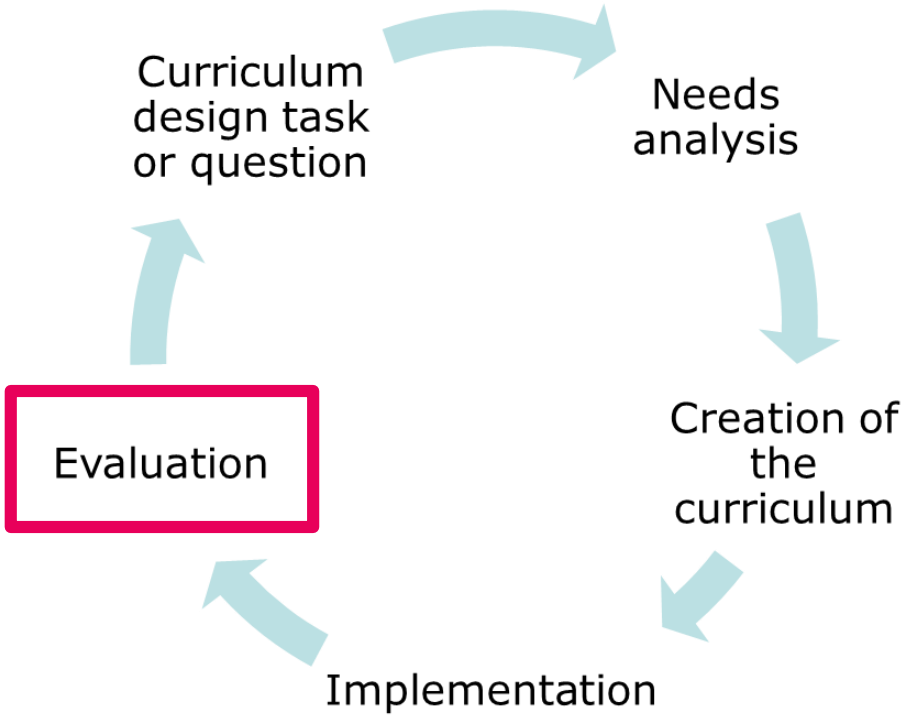


Figure 1.1 A model of the parts of the curriculum design process.

Nation & Macalister (2010)
Kostka & Bunning (2018)

- Consistent calls for iterative approach to course design.
  Bardi & Mureșan, 2012; Bocanegra-Valle, 2016; Crawford Camiciottoli, '10
- Hard with intense pre-sessional workload.

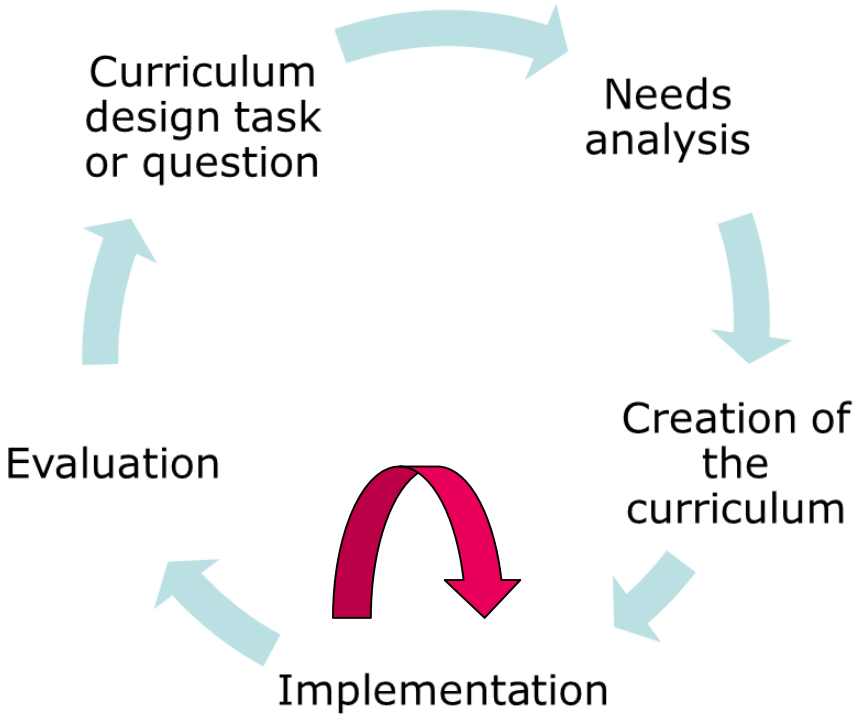
Background – Educational Evaluation



Nation & Macalister (2010)
Kostka & Bunning (2018)

- Consistent calls for iterative approach to course design.
  Bardi & Mureșan, 2012; Bocanegra-Valle, 2016; Crawford Camiciottoli, '10
- Hard with intense pre-sessional workload.

Background – Educational Evaluation



Nation & Macalister (2010)

- Consistent calls for iterative approach to course design.
  Bardi & Mureșan, 2012; Bocanegra-Valle, 2016; Crawford Camiciottoli, '10
- Hard with intense pre-sessional workload.

Background – Educational Evaluation

- Curriculum governance is a core aspect of the curriculum design process

  "[This] describes the responsibility of…educators to establish and maintain high standards of teaching and learning." (Wilkes & Bligh, 1999).

- Educational evaluation is a core aspect of curriculum governance (Wilkes & Bligh, 99)

  "[This] is the systematic appraisal of the quality of teaching and learning. At its core, evaluation is about helping… educators improve education." (Wilkes & Bligh, 1999)

Background – Educational Evaluation

- Two types of evaluation (See Ellis, 1997):

1. **Predictive**:

   To select appropriate materials/adaptation strategy

2. **Retrospective**:

   To determine whether instruction and which activities have 'worked', and how materials should be modified in the future.

"the focus of attention has been more or less exclusively on predictive evaluation…there are very few published accounts of retrospective evaluations of course materials, and very little information about how to conduct them." (Ellis, 1997)

Background – Educational Evaluation

- Key questions relating to evaluation (Graves, 2000).
  - What is evaluated?
  - Why evaluate the course?
  - How can you evaluate the course?
  - When can you evaluate the course?
  - What is done with the results?

Background – Educational Evaluation

- Key questions relating to evaluation (Graves, 2000).
  - What is evaluated?
  - Why evaluate the course?
  - **How can you evaluate the course?**
  - When can you evaluate the course?
  - What is done with the results?

- How do you evaluate your instruction? (10 seconds)

Background – Evaluation Tools

- Retrospective evaluations typically use (Wilke & Bligh, 1999):
  1. Structural evaluation measures
     - ✓ *Attendance, engagement metrics (VLE access data), etc.*
  2. Outcome evaluation measures
     - ✓ *Tracking study, attainment data, etc.*
  3. Process evaluation:
     - ✓ *Student satisfaction, observations, etc.*
  4. Evaluation tools
     - ✓ Assessment, student journals, questionnaires, self-report, etc.

Which of these do you use to evaluate your practice? (10 secs)

Background – Evaluation Tools

- While there is not "One Best Way of conducting an evaluation", this does not mean that "anything goes". (Alderson, 1992)

- The majority of evaluations rely on perception data
  - Asking students to assess the activities they have done. (Graves, 00)
  - Conducting checklist measure and semi-structured interviews of teachers and students (Ahour and Ahmadi, 2012)

    ❖ Student feedback is affected by satisfaction and teacher popularity. Teacher feedback can also be impacted by construct irrelevant variance. So such perception data does not provide objective measure of strengths and weaknesses. (see Kiely & Rea-Dickens 2005)

Background – Evaluation Tools

- While there is not "One Best Way of conducting an evaluation", this does not mean that "anything goes". (Alderson, 1992)

- The majority of evaluations rely on perception data
  - Asking students to assess the activities they have done. (Graves, 00)
  - Checklist and semi-structured interviews of teachers and students (Ahour and Ahmadi, 2012)

    ❖It is important that "honest data is available" (Nation & Macalister, 2010). We, therefore, need a "framework for moving beyond course satisfaction feedback" (Kiely & Rea-Dickens, 2005).

Background – Assessment matrices

- One useful tool might be **assessment criteria**.

	Argument	Essay Structure	Indication of sources to be used
1	A clearly expressed statement of argument which completely addresses the title prompt.	The overall structure of the essay is clearly shown in the plan. The sequencing of the sections is logical and reflects the statement of argument. Each section has a clearly expressed and defined topic which connects logically to the argument. Each section has at least 2 relevant supporting points. The reader is well informed about the way the argument will be developed.	Each section has a clear indication of which parts of the provided source texts will be used to support points made (i.e. author, page/paragraph number). There is a clear indication of specific additional sources to be used. These sources are all appropriate.
2	The statement of argument is adequately expressed but may lack clarity. It may only partly address the title prompt.	The overall structure of the essay is quite clear although there may be occasional problems with the sequencing of the sections. The essay structure largely reflects the statement of argument. Each section has a topic which connects logically to the argument although at least one topic may lack clarity or appear similar to another. Each section has at least 1 relevant supporting point. The reader is adequately informed about the way the argument will be developed.	Each section has an indication of which of the provided source texts will be used to support points made although these references may not be specific (i.e. author, page/paragraph number may not be given). Some additional sources to be used are indicated but these may not be specific and/or appropriate.
3	The statement of argument is very poorly expressed and does not address the title prompt.	Some attempt to indicate the structure of the essay but this is not clearly shown in the plan. Sections are indicated but their sequencing has no apparent logic and may contradict the statement of argument. Although section topics may be indicated, these are poorly expressed and not clearly defined. Some sections may lack relevant supporting points. The reader is poorly informed about the way the argument will be developed.	Some sections may not include reference to the source texts to be used. Where references are made these are not specific (i.e. author, page/paragraph number are not given). There may be no indication of additional sources to be used or this may be very general.

Background – Assessment matrices

- One useful tool might be **assessment criteria**.
  - Commonly reported benefits include:
    - facilitating meaningful interpretations of writing and speaking ability (Green, 2014)
    - guiding instructional design and delivery (Arter & McTigue, 2001)
    - making the assessment process more accurate and fair
    - providing tool for self-assessment

- Can they be used as a retrospective materials evaluation tool?

Background – Assessment Criteria

Types of Assessment Criteria

- holistic: the rater makes an overall judgment about the quality of performance.
- Analytic: the rater assigns a score to each dimension separately.

Types of Retrospective Evaluation

- Macro evaluation: used to determine the efficacy of the materials as a whole. (Ellis, 1997)
- Micro evaluation: used to determine the efficacy of individual teaching activities. (Ellis, 1997)

The Project – Aims

The aims of this project were as follows:

1. To develop a systematic approach to the collection of data to empirically evaluate the materials used on a Pre-Sessional EAP course.
2. To determine the extent to which course material were constructively aligned with the assessment.

    2a. To determine the extent to which course length interacts with the efficacy of instruction.

The Project - Methodology

Prior to the use of assessment criteria:

- Adapted the assessment matrices to allow nuanced detail to emerge.

Argument

There is an effective / adequate thesis statement, but it may not fully answer the essay question. Most sections has a clear section claim.

The Project - Methodology

Prior to the use of assessment criteria:

- Adapted the assessment matrices to allow nuanced detail to emerge.

Argument

There is an effective / adequate thesis statement, but it may not fully answer the essay question. Most sections has a clear section claim.

The Project - Methodology

Prior to the use of assessment criteria:

- Adapted the assessment matrices to allow nuanced detail to emerge.

Argument
There is an effective / adequate thesis statement, but it may not fully answer the essay question. Most sections has a clear section claim.

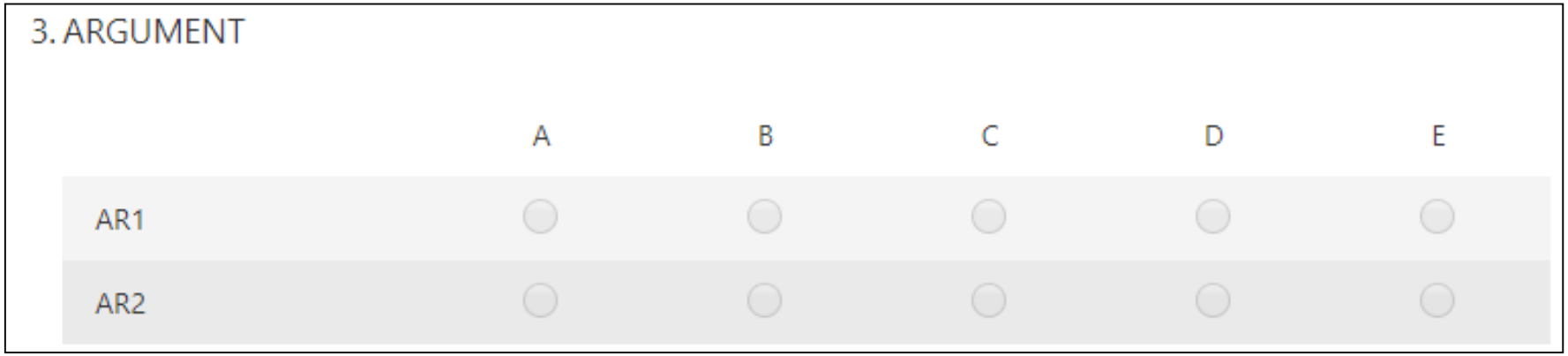
→

Argument
(AR1) There is an effective or adequate thesis statement. This may only partially answer the essay question.
(AR2) Most sections have a clear and focused section claim which are relevant to the thesis statement.

The Project - Methodology

Prior to the use of assessment criteria:

- Adapted the assessment matrices to allow nuanced detail to emerge.
- Created online forms to speed up process of assessing work. This was done with Microsoft Forms.



The Project - Methodology

Prior to the use of assessment criteria:

- Adapted the assessment matrices to allow nuanced detail to emerge.
- Created online forms to speed up assessment. This was done with Microsoft Forms.
- Conducted rater training, standardisation, and moderation.

The Project - Methodology

After to the use of assessment criteria:

- Downloaded the excel form containing all grades.
- Visually inspected the box plots of each marker for each dimension of the assessment matrix. This was done to establish outliers. Outliers were removed from further analysis.

Time limitations meant not possible to conduct intra-rater or inter-rater reliability. So this was done in an attempt to improve reliability (in addition to standardisation and moderation procedures).

The Project - Methodology

After to the use of assessment criteria:

- Downloaded the excel form containing all grades.
- Visually inspected the box plots of each marker for each dimension of the assessment matrix. This was done to establish outliers. These data were removed from further analysis.
- Descriptive statistics calculated for each subdimension.
- Compared achievement of the different lengths of courses (20 weeks, 15 weeks, 10 weeks, 6 weeks) on each subdimension.
- Inspected data to identify areas of poor performance.
- Amended materials as necessary to support problematic areas.

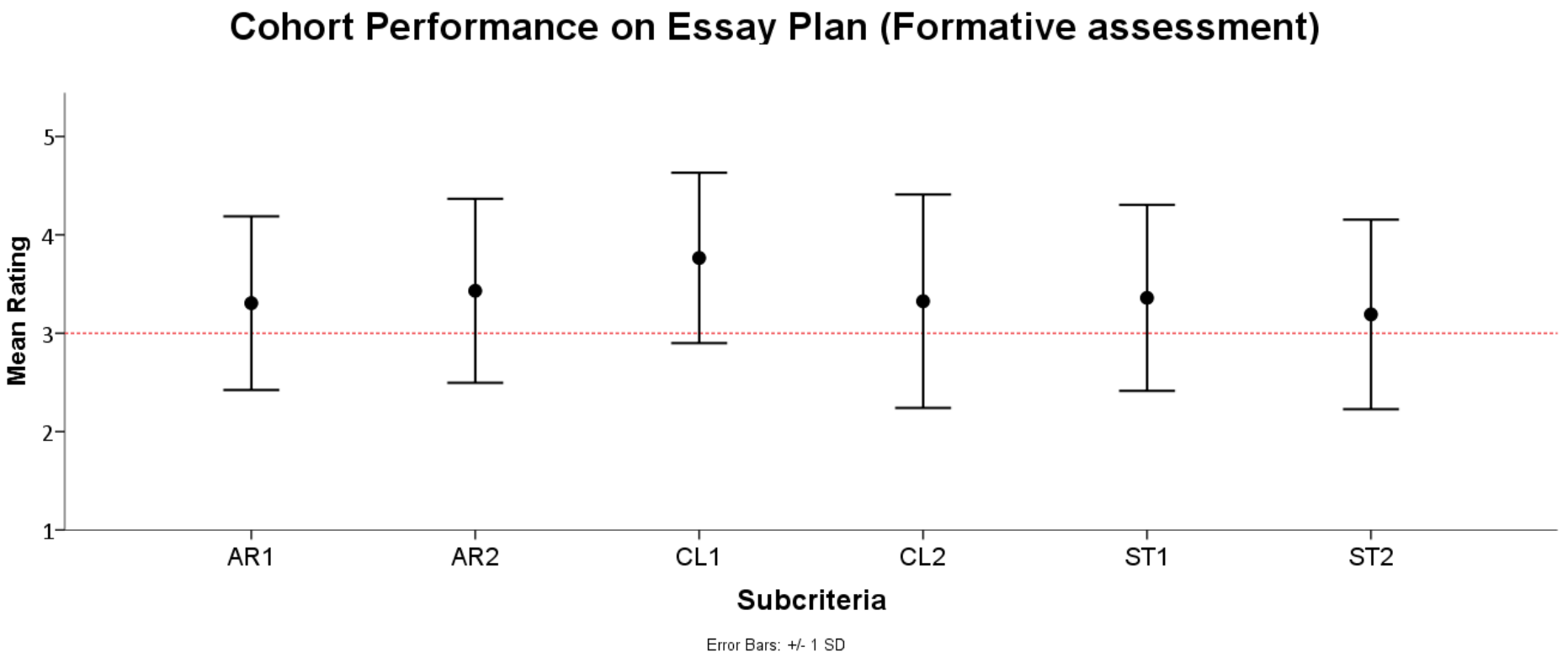
The Project - Methodology

- 190 completed assessment matrices from pre-sessional EAP course at NTU considered.

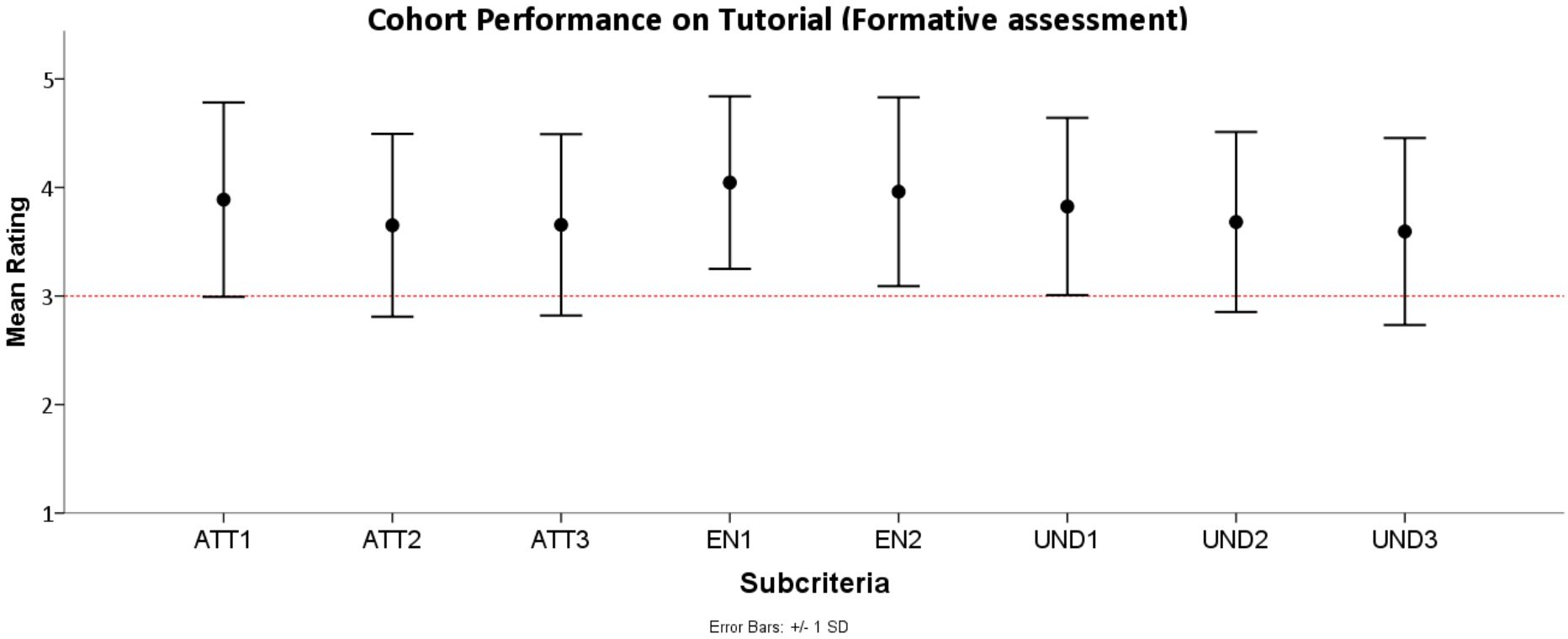
Course length	Male	Female	Total
20 weeks	10	5	15
15 weeks	11	20	31
10 weeks	29	38	67
6 weeks	24	53	77
Total	74	116	190

- Matrices for formative and summative assessment considered (Coursework essay: plan, tutorial, final draft; Presentation; Writing test)

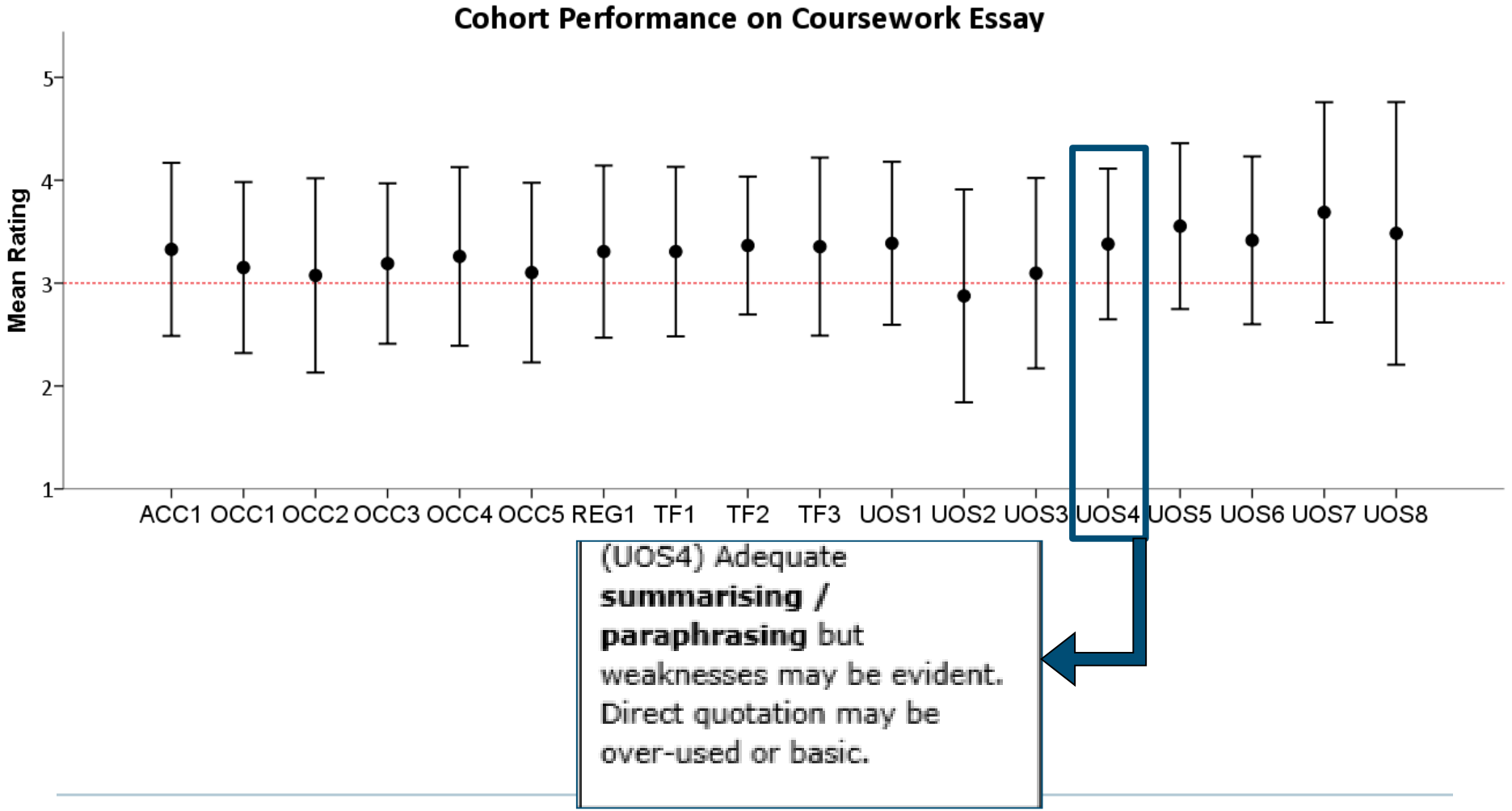
Results – Whole Cohort on **Essay Plan**



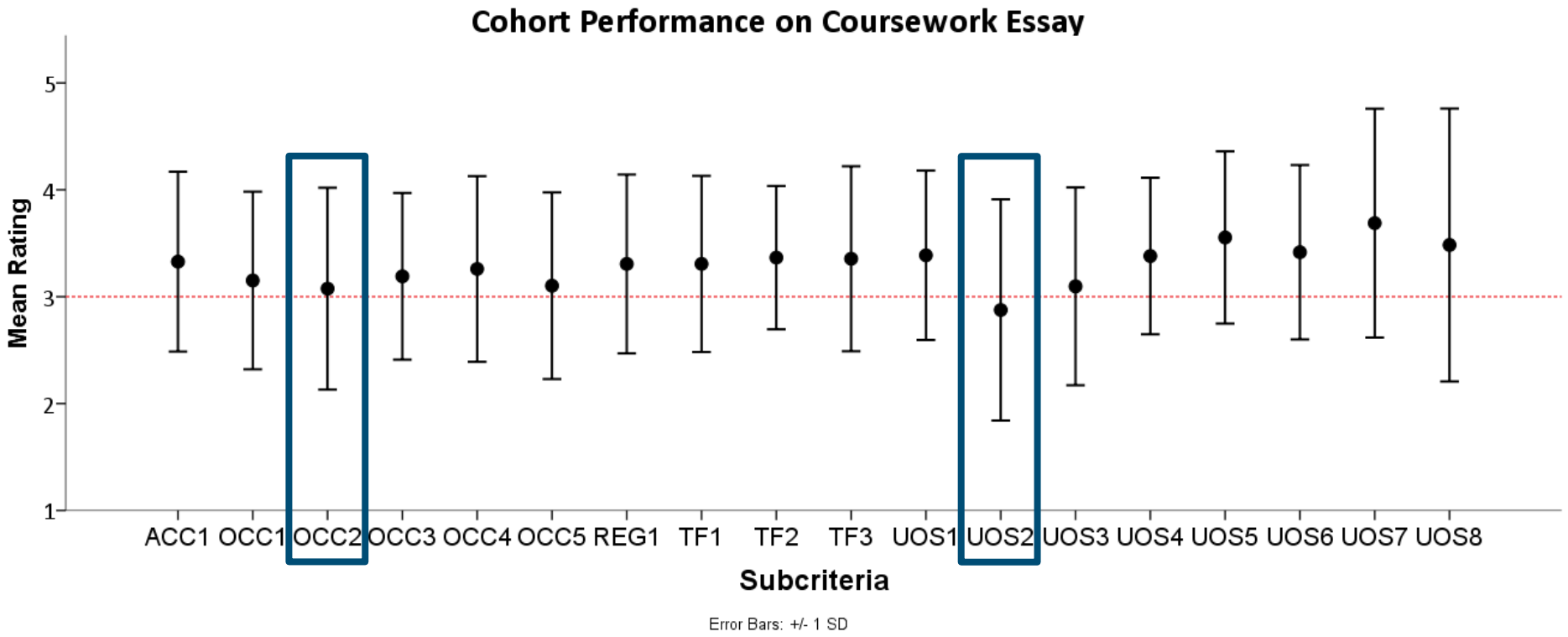
Results – Whole Cohort on **Essay Tutorial**



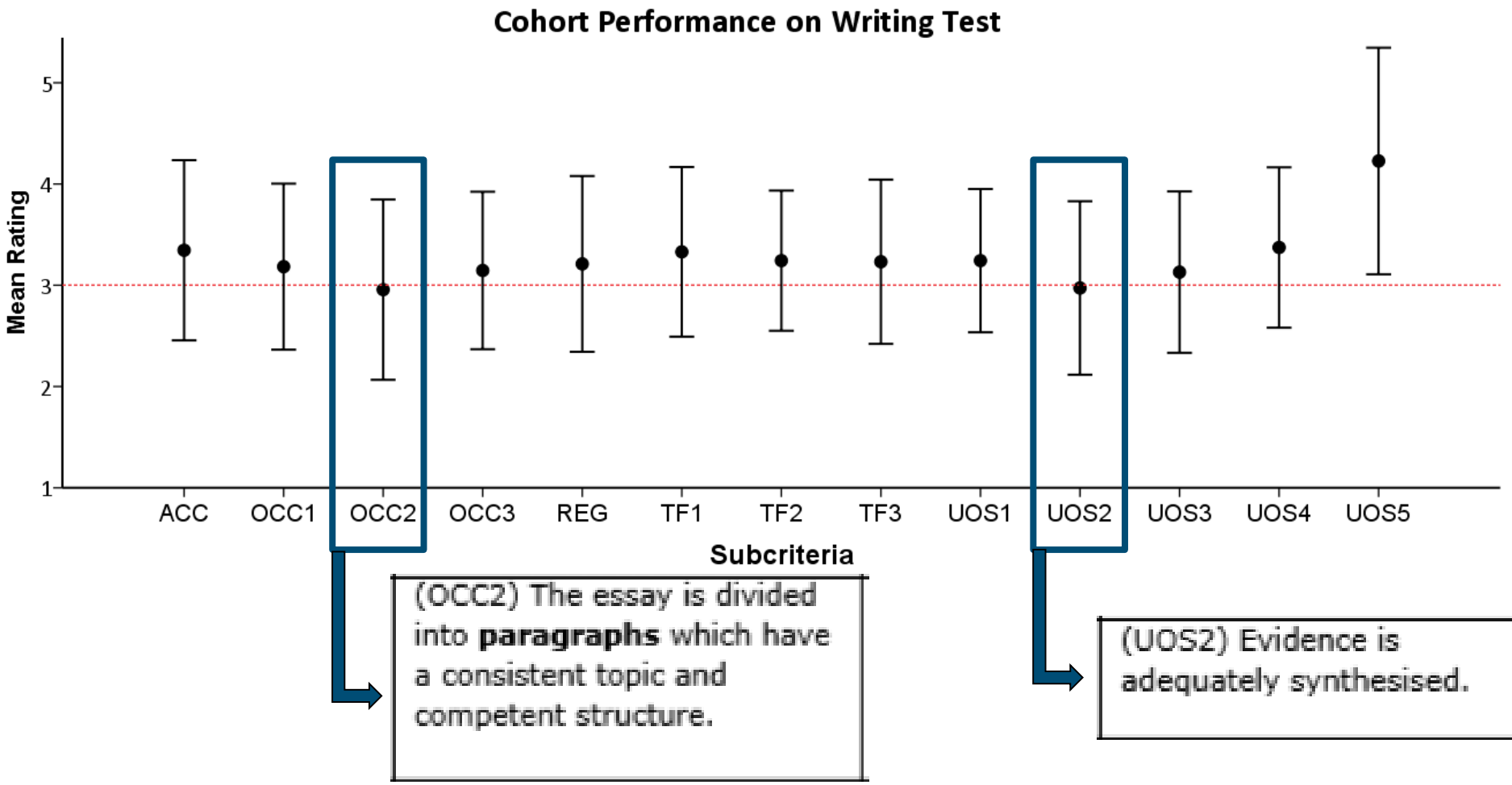
Results – Whole Cohort on Coursework Essay



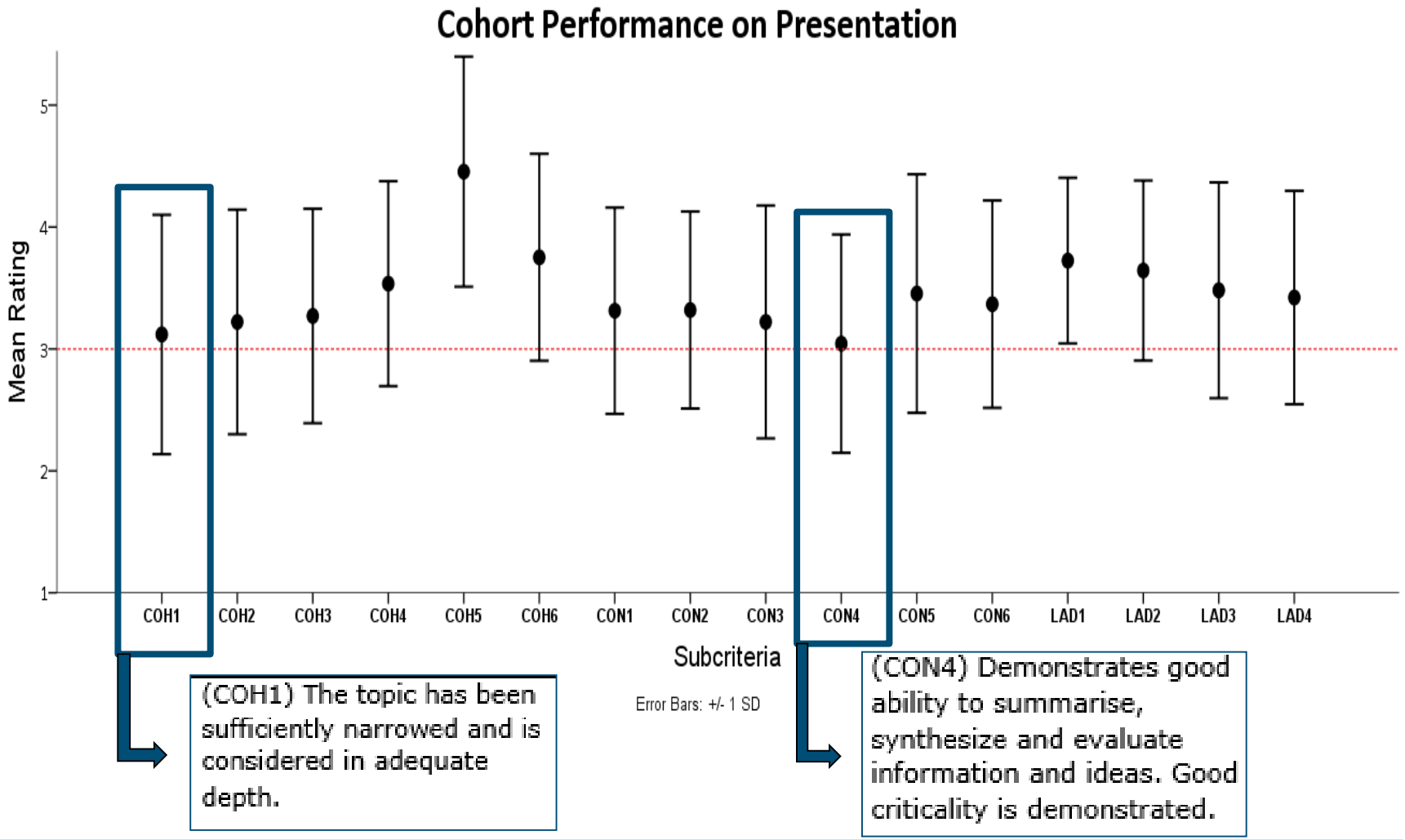
Results – Whole Cohort on Coursework Essay



Results – Whole Cohort on Timed Writing



Results – Whole Cohort on Presentation



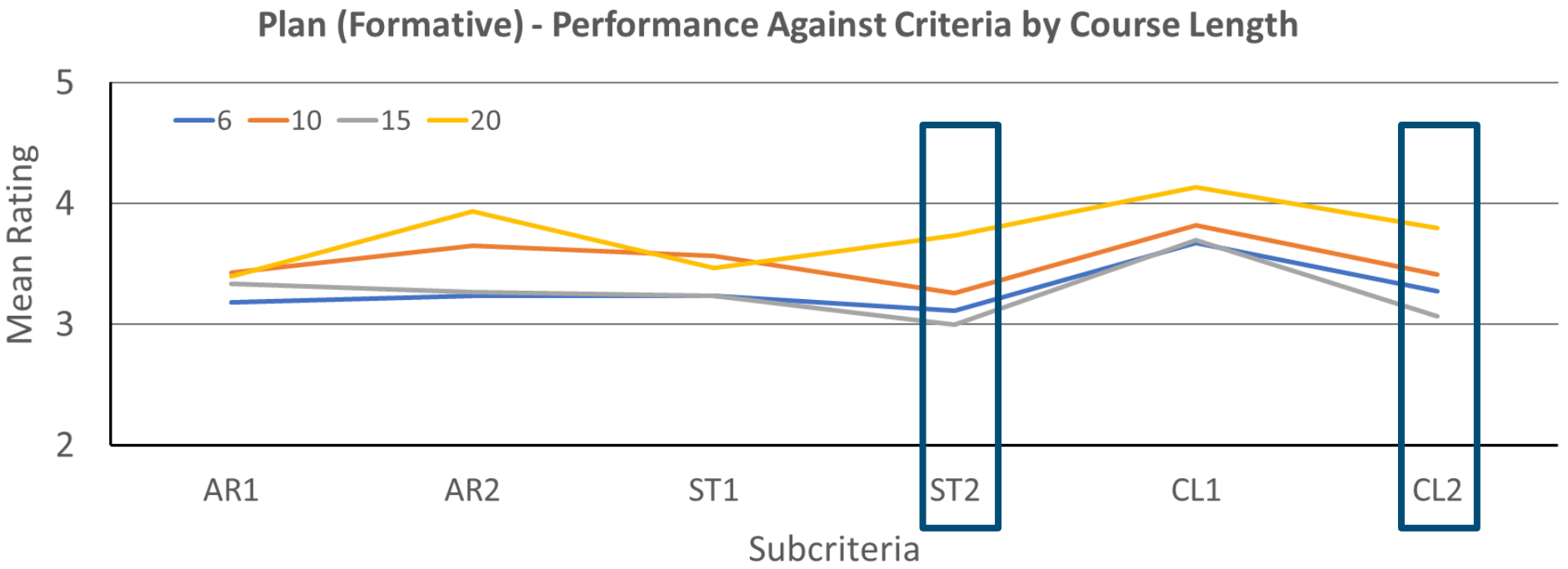
The Project – Aims

The aims of this project were as follows:

1. To develop a systematic approach to the collection of data to empirically evaluate the materials used on a Pre-Sessional EAP course.
2. To determine the extent to which course material was constructively aligned with the assessment.

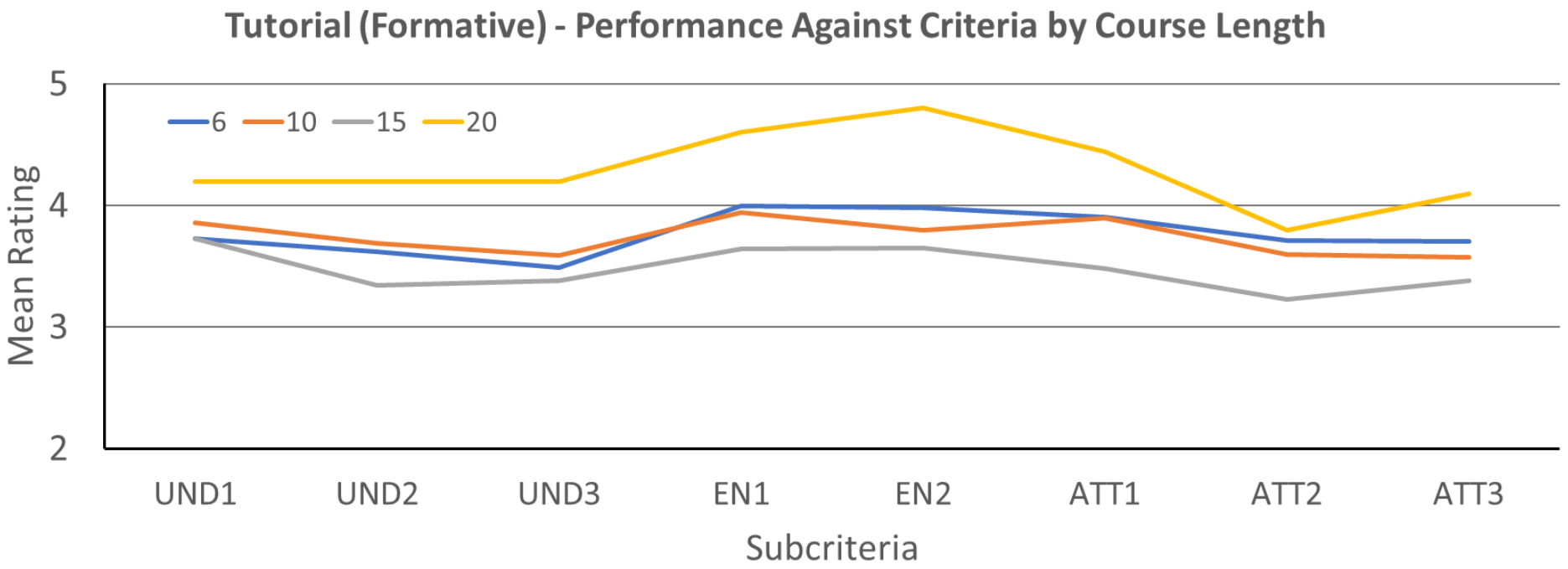
   2a. To determine the extent to which course length interacts with the efficacy of instruction.

Results by Course Length – Essay Plan



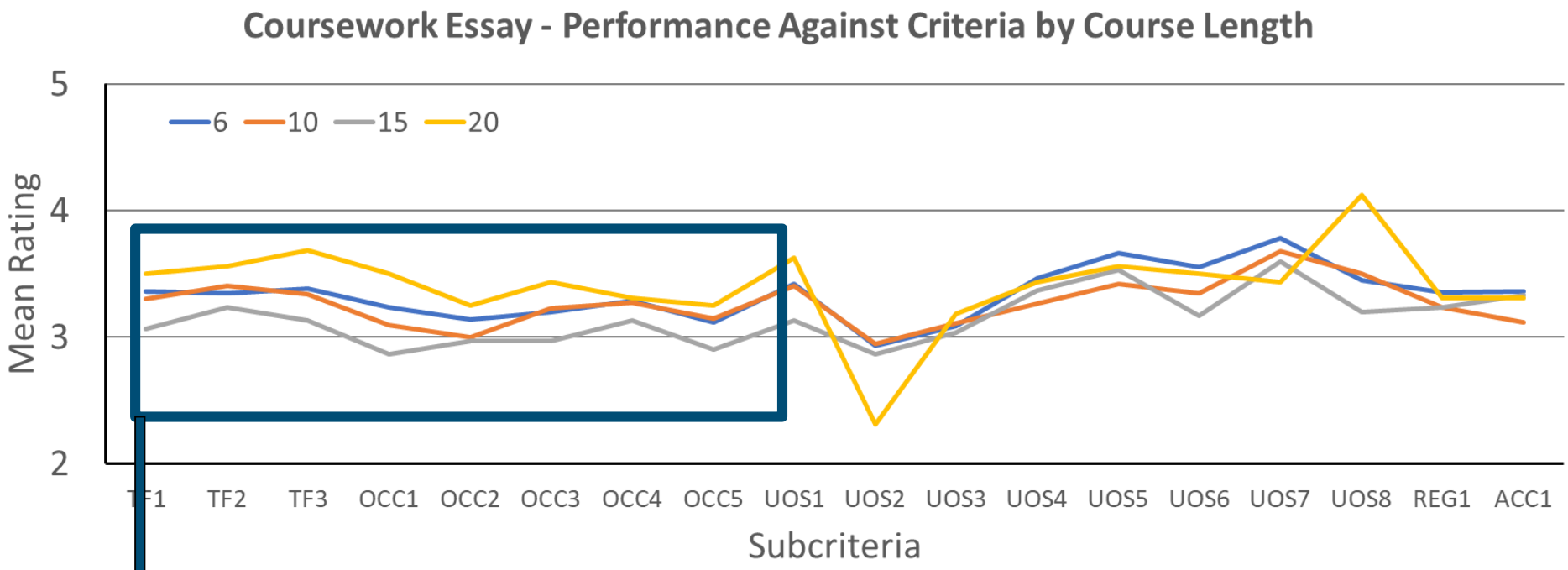
- 20 week students performing well.
- 15 week students seem to need more support with structure, and selecting source material in particular.

Results by Course Length – Essay Tutorial



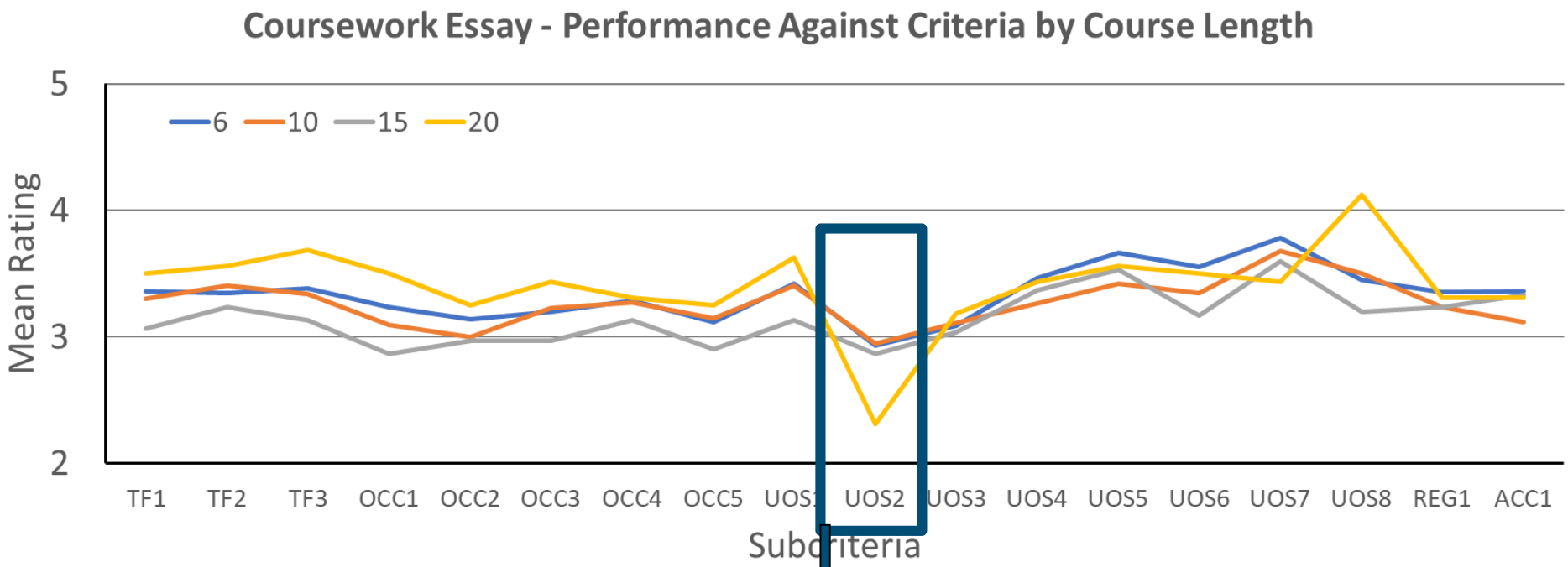
- In general, all students performing well.
- Differential attainment suggests we need to look at materials and support mechanisms for the 15-week students.

Results by Course Length – Coursework Essay



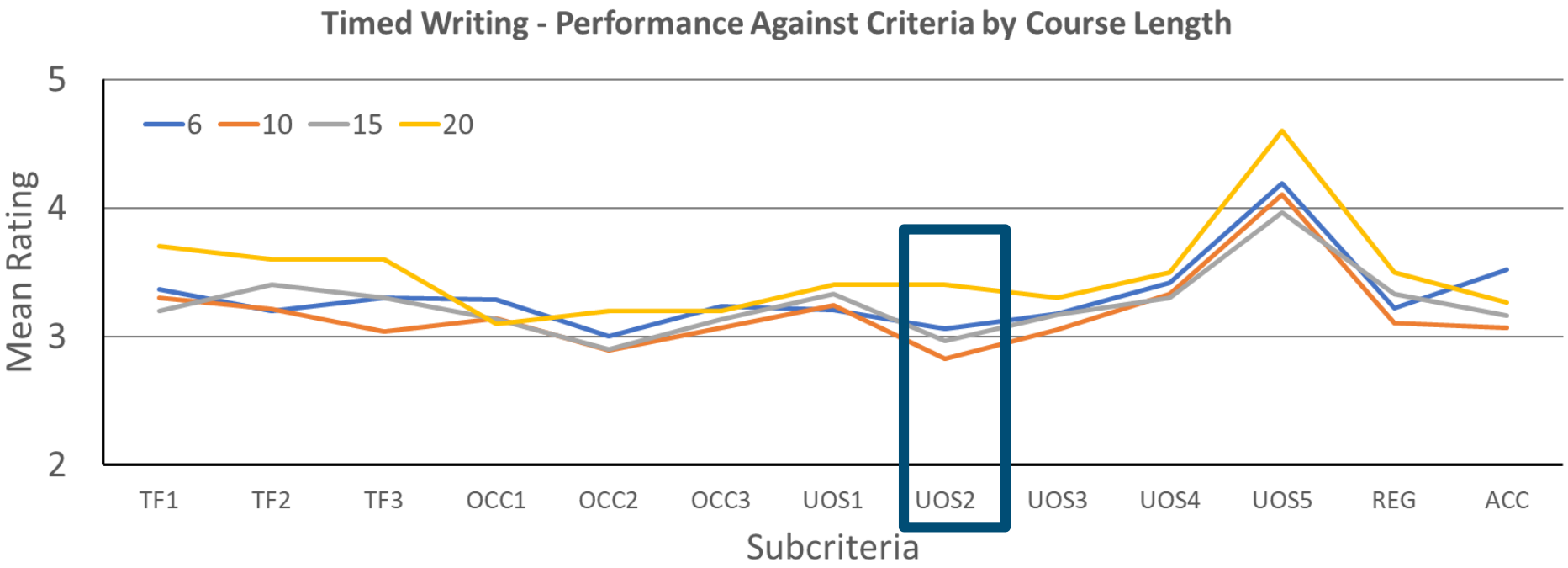
15 week students seem to struggle with organisation, cohesion, and coherence. Need to look at materials and emphasise during teacher induction.

Results by Course Length – Coursework Essay



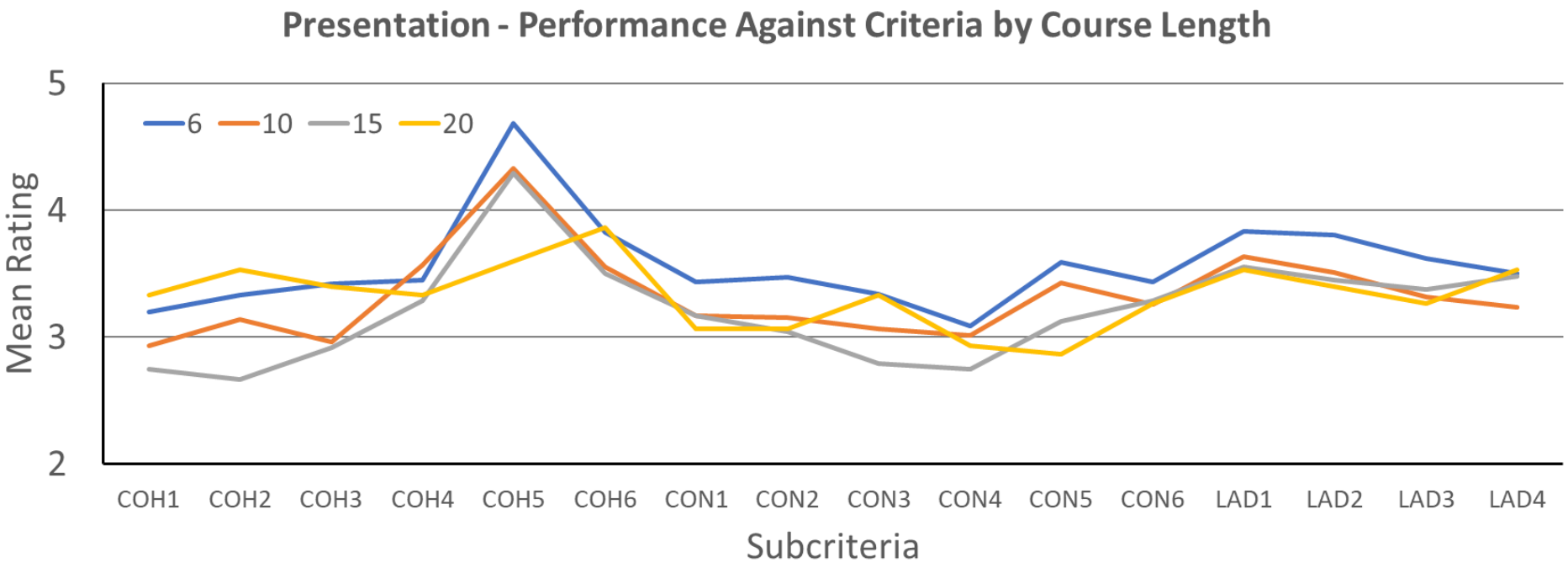
20-week students seem to struggle particularly with synthesis. This needs attention.

Results by Course Length – Timed Writing



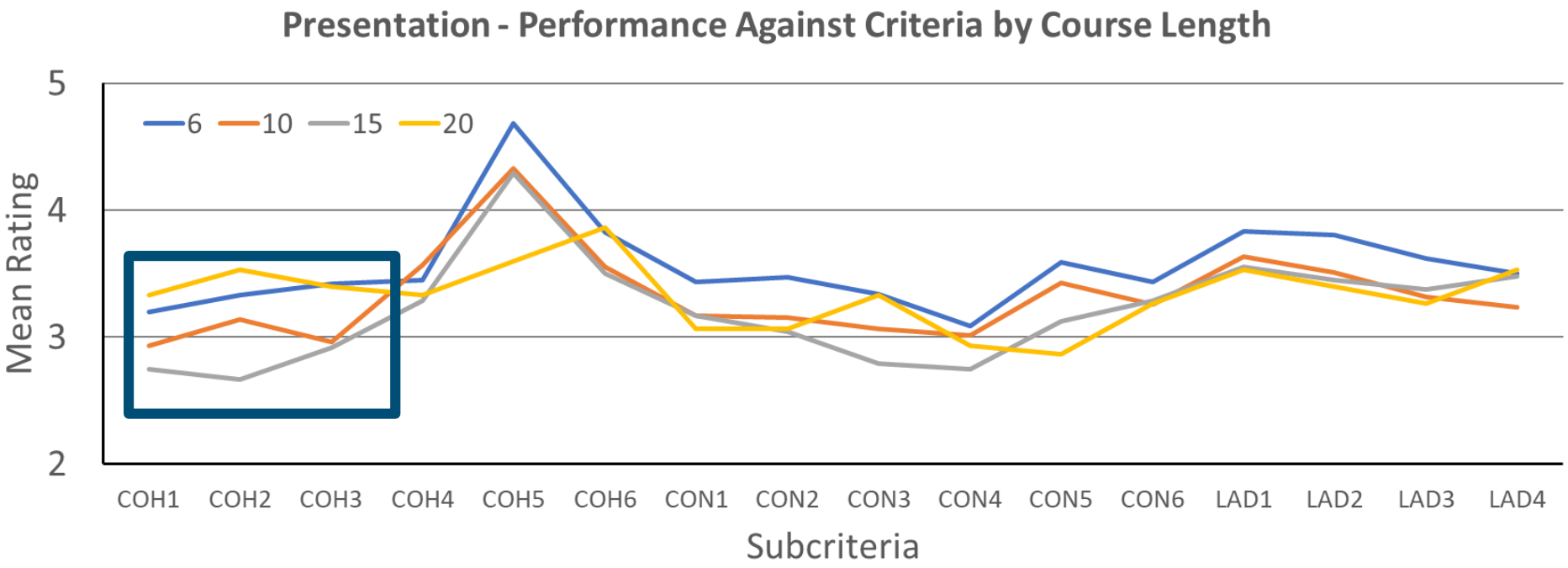
- Interesting that longer-course students are now synthesising better. Perhaps issue is not synthesis per se, but number of sources (fewer sources on writing test)

Results by Course Length - Presentation



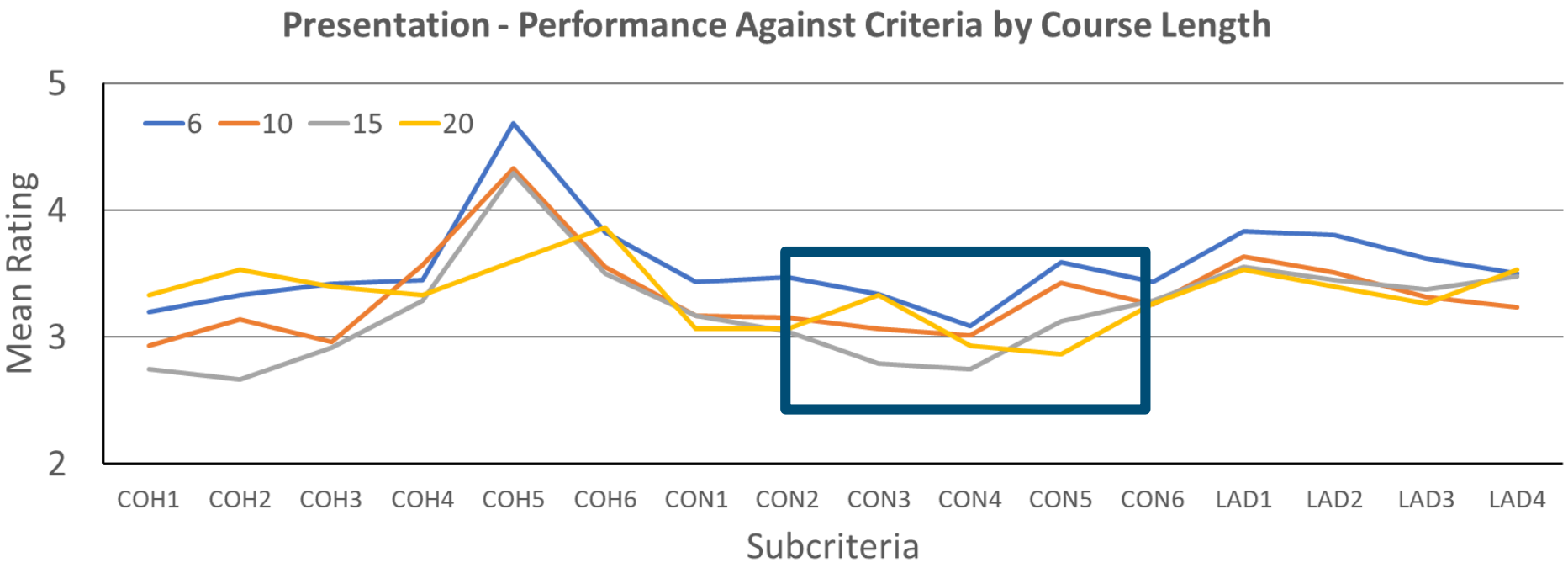
- In general, shorter course students outperforming longer-course students.

Results by Course Length - Presentation



- In general, shorter course students outperforming longer-course students.
- 15-week students seemed to struggle with interpretation, stance, and focusing topic..

Results by Course Length - Presentation



- In general, shorter course students outperforming longer-course students.
- 15-week students seemed to struggle with interpretation, stance, and focusing topic..
- Again, longer-course students struggled synthesising many sources.

Discussion – Aim 1

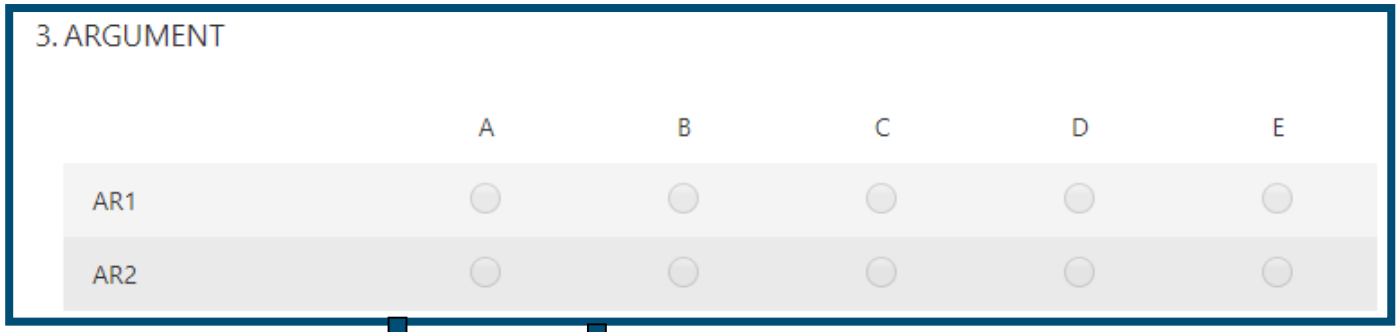
Aim One:

To develop a systematic approach to the collection of data to empirically evaluate the materials used on a Pre-Sessional EAP course.

- Analysis easy to perform and provided meaningful evidence to adapt materials.
- Evaluation often results in little change (Nation & Macalister, 2010)
  → objective data harder to ignore and more actionable than perception data.
- Actually reduced teacher work load and did not negatively impact quality of feedback to students.

Discussion – Aim 1

This is the only input



Staff

Data automatically concatenated onto grades collection spreadsheet to allow management staff to calculate results.

Student

Data pulled, using macro, automatically create an individual student record. Automatically sent to students upon completion.

Discussion – Aim 1

STUDENT NUMBER:	EXAMPLE						
Coursework Essay							
	AR1	AR2	ST1	ST2	CL1	CL2	AR
Final Plan	A	B	C	C	A	B	B
	UND1	UND2	UND3	EN1	EN2	EN3	ATT1
Viva	A	A	A	C	D	C	B
	TF1	TF2	TF3	OCC1	OCC2	OCC3	OCC4
Final Draft	C	C	D	D	D	D	D
Presentation							
	ENG1	ENG2	ENG3	ENG4	ENG		
Group work and Padlet	B	C	C	C	C		
	COH1	COH2	COH3	COH4	COH5	COH6	CON1
Final Presentation	C	B	D	A	A	C	D
Writing	TF1	TF2	TF3	OCC1	OCC2	OCC3	UOS1
	D	D	D	E	D	D	D

Discussion – Aim 1

COURSEWORK ESSAY ASSESSMENT COMPONENT 1: ESSAY PLAN			
	Argument	**Structure**	**Clarity**
A	(AR1) There is an expert or skilful thesis statement which completely answers the essay question.	(ST1) The sequencing of the sections is sophisticated or accomplished. It reflects the thesis statement and will allow the student to synthesise multiple sources.	(CL1) The plan is very easy to understand and is expertly or skilfully formatted in a consistent manner. Overall, the plan will be very easy to use when the student is writing the essay.
	(AR2) Each section has a clear and focused section claim which is relevant to the thesis statement.	(ST2) Each section has exceptional supporting detail. This is taken from a range of appropriate sources. The quantity and quality of supporting evidence will allow the student to write a sophisitcated or accomplished answer to the prompt and meet the word limit. Sophisticated or skilful understanding of which areas need to be expanded may be indicated.	(CL2) Indicated source extracts are specific and will be easy to find at a later date.
B/C	(AR1) There is an effective or adequate thesis statement. This may only partially answer the essay question.	(ST1) The structure of the essay is solid or adequate although there may be some problems with the sequencing of the sections. The essay structure largely reflects the thesis statement, but may not allow the student to synthesise multiple sources.	(CL1) The plan is quite easy to understand and effectively or adequately formatted, although there may be some inconsistency. The plan will be quite easy to use when the student is writing the essay.
	(AR2) Most sections have a clear and focused section claim which are relevant to the thesis statement. [AND/OR] Sections may lack clarity and focus although generally connect to the thesis statement.	(ST2) Each section has solid or adequate supporting detail. This is taken from multiple sources, although the student may rely on one (or two) key sources. The quantity and quality of supporting evidence will allow the student to write a solid or adequate answer to the prompt, but that answer might be over or under length. Good or satisfactory understanding of which areas need to be expanded may be indicated.	(CL2) Indicated source extracts are generally specific and will be easy to find at a later date, although this might not always be the case.
D/E	(AR1) The thesis statement is ineffective or poor and does not address the essay question.	(ST1) The structure of the essay is inappropriate or poor. There is little or no descernable logical order underpinning the sequencing of sections. The essay structure does not reflect the thesis statement. It will not allow the student to synthesise multiple	(CL1) The plan is difficult to understand and is inadequately or poorly formatted. There may be an inconsistent format which affects the effectiveness of the plan. Overall, the plan will not be easy to use when the student is writing the essay.
	(AR2) Although sections may be indicated, these are inadequately or pooly focused and (some) may not connect to the thesis statement.	(ST2) Each section has ineffective or poor supporting detail. This is taken from a limited range of sources - the student relies on one (or two) key sources. The quantity and quality of supporting evidence is such that the student will write an ineffective or poor answer to the prompt, and that answer might be over or under length. Inadequate or poor understanding of which areas need to be expanded may be indicated.	(CL2) Indicated source extracts are generally unspecific and will be difficult to find at a later date.

Discussion – Aim 1

Advice

- Develop a system that allow expeditious evaluation without increasing workload associated with rating and reporting.
- Important to separate sub-criteria to improve formative evaluative power of the summative assessment tool. Trade off with time needed to rate (Wolf & Stevens, 2007) – piloting is essential to ensure tool is practical.

Discussion – Aim 2

Aim Two

To determine the extent to which course material was constructively aligned with the assessment.

- In general, materials seem to be *working*. Sub-criteria associated with task(s) and learning objectives. Students largely met these. Shows good constructive alignment.

- Analysis highlighted areas in need of support. These activities will be discussed with teachers, cross-referenced to records of work, and amendments made. Course design is iterative (Brown, 2009; Hyland, 2006) so need follow up next academic year.

Discussion – Aim 2a

Aim 2a

To determine the extent to which course length interacts with the efficacy of instruction.

- In general, longer course students seemed to struggle with more complex tasks – synthesis, interpretation, cohesion. Suggests need earlier consistent focus on these areas.

- Longer-course students performing well on language-related sub-criteria (this is a positive change from previous analyses).

Limitations

- No inter-rater reliability analysis, but tried to adequately control given environmental constraints.
- Data need triangulation against teacher reports, records of work, and observation before conclusions can be drawn, but current set is a good place to start!
- Other grouping variables – L1, destination course, gender – to check for differential functioning of materials.

Conclusions

1. Use of a summative assessment matrices can facilitate formative materials evaluation.
2. Provides meaningful data that encourages evidence-based, iterative, principled, course design.
3. Each assessment sub-criterion equates to micro evaluation. This allows for nuanced materials / instruction amendments.
4. Analysis is quick to conduct, so can be utilised in the busy pre-sessional calendar.

samuel.barclay@ntu.ac.uk
@samuelcbarclay

Inter-rater and Intra-rater reliability

- "No evaluation is ever objective…the best we can hope for is pooled intersubjectivity and reduced or neutralised partiality." (Alderson, 1992)
- "Ideally, an assessment should be independent of who does the scoring and the results similar no matter when and where the assessment is carried out, but this is hardly obtainable." (Jonsson & Svingby 2007)
- Variations in raters' judgments can occur either across raters, known as inter-rater reliability, or in the consistency of one single rater, called intra-rater reliability. There are several factors that can influence the judgment of an assessor...Besides the more obvious reasons for disagreement, like differences in experience or lack of agreed-upon scoring routines, it has been reported that things like teachers' attitudes regarding students' ethnicity, as well as the content, may also influence the ratings of students work (Davidson, Howell, and Hoekema, 2000)"