The University of Auckland
NEW ZEALAND

Better Questions for Higher Order Objectives: Improving the Quality of Assessment in Higher Education

Seminar for the Linking University with the Working World project

Chile, December 2015

Gavin T L Brown, *The University of Auckland*

gt.brown@auckland.ac.nz

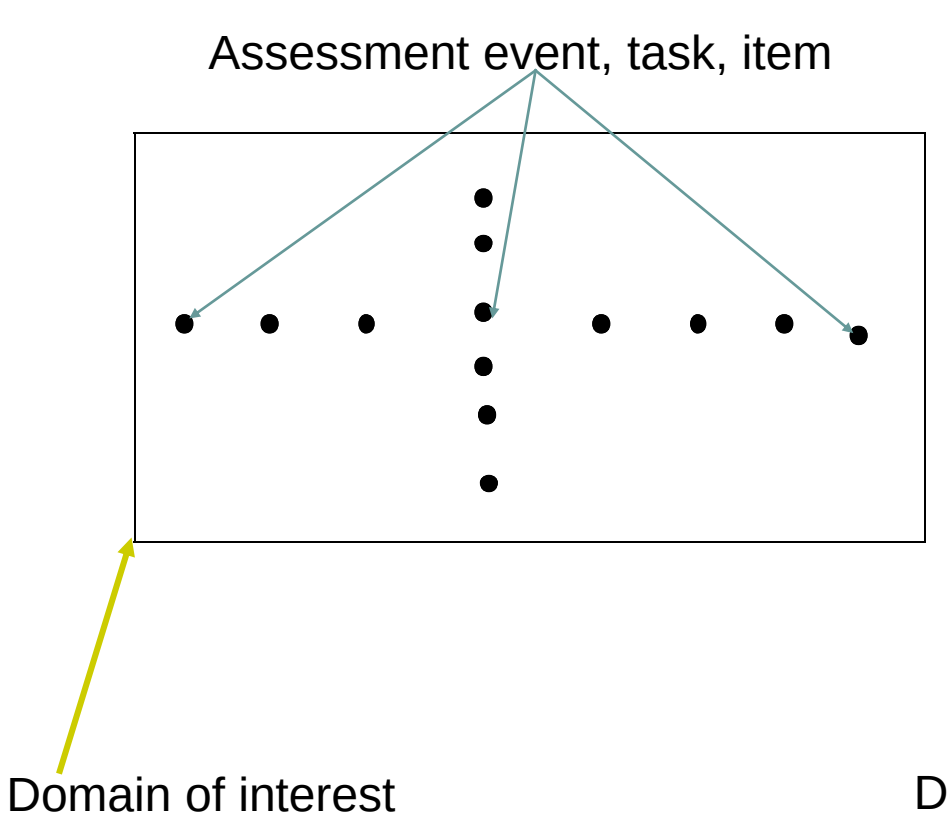
- David & Sue have given high-level picture of needing to assess complex skills that are needed for life after school…
- This talk will focus on a much more technical level at details of assessment processes
- What I will talk about is how to improve these written assessments assuming they are appropriate for the outcomes you value—don’t do them if they are not authentic to your intended learning goals

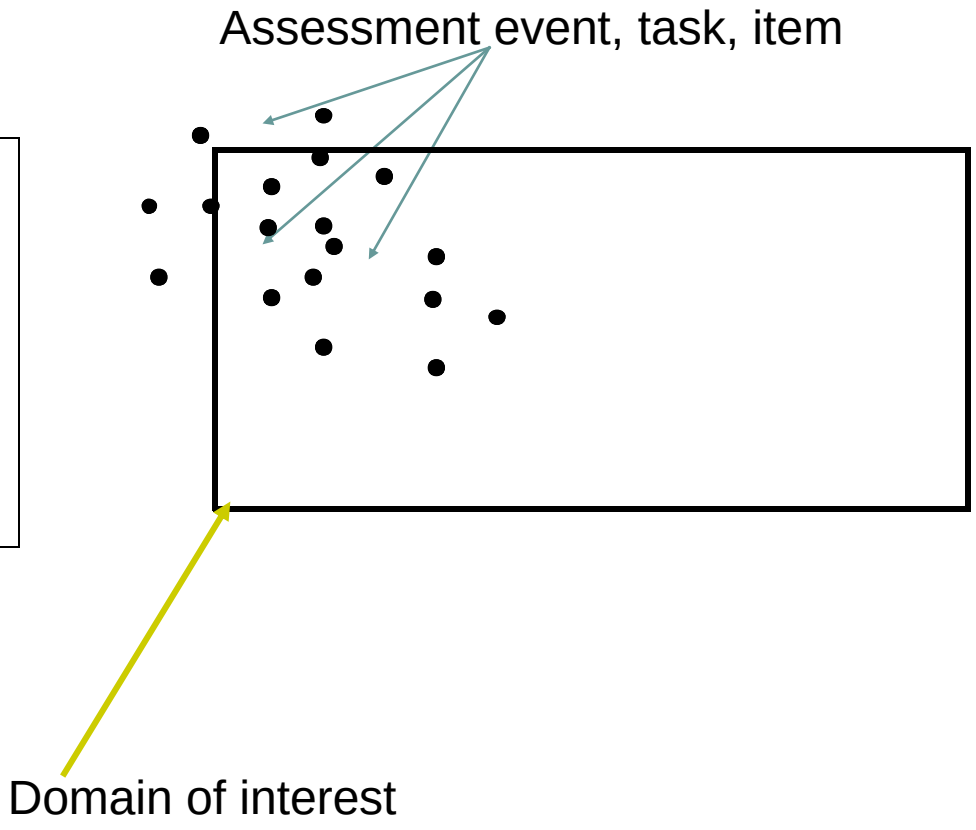
Authenticity in Written Assessment

- Accuracy + Efficiency + Validity
  - How do we assess efficiently many different topics with so many students and ensure that the questions focus on what we really care about?
    - Multiple Choice Questions
- Higher-order, deeper thinking
  - How do we ensure our questioning and task design get at skills beyond remember and recognise?
    - The SOLO Taxonomy

Alignment to Curricular Goals

- If you want assessment to be aligned to curricular goals, the assessment must be within the curriculum
  - Systematic and Aligned is ESSENTIAL





MULTIPLE CHOICE ITEMS

- Use where the task calls for a single, clear answer to a question.
- When well designed, emphasize critical thinking and reasoning, rather than factual recall.
- Use when the range of possible correct answers is too broad – to focus thinking
- Use to remove load of writing; not thinking

Writing M-C items: rules for stems

- Keep clear & concise – “specific is terrific!”
- Not too long to read
- Avoid negatively worded questions.
  - Emphasise ***NOT*** if you must ask a negative question
- Check the answer is not elsewhere in paper
- Avoid clues in grammar (*a/an; is/are etc*)
- Use interrogatives (*What is the name of this tool?)* or imperatives *(State the advantages of the …)* rather than sentence completion

Writing M-C items: rules for answers

- Only one correct answer – the key.
- Answer is actually correct.
  Check, check, re-check!
- Answer is sufficient to answer the question.
- No pattern of correct answers.
- Should not repeat words in stem.
- Use typical errors students make.

Multiple Choice Distractors or *wrong answers*

- Plausible
  - not silly or plainly wrong
  - Connected to
    - a commonly held misunderstanding, or
    - An overgeneralisation or a narrowing of application
- Similarity to each other and answer
  - Similar length
  - Similar style as answer
  - Match the grammar or style of stem or question
- Attract guessers & those who have imperfect or weak knowledge
- Arrange in a logical order – alphabetical, numerical, time series …
- Avoid implausible qualifiers – e.g., ***never***, ***always***
- Avoid ***all of the above***, ***none of the above***

Number of response alternatives

- Typically three, four or five
- Four or five favoured over three – reduces guessing, increases discrimination
- Four is most common, but 3 is defensible
  - Use three if entirely appropriate – *an acre is* ***larger/smaller/equal to*** *a hectare?*

Test of Objective Evidence

- Each of the questions in the following set has a logical or “best” answer from its corresponding multiple choice answer set. Best answer means the answer has the highest probability of being the correct one in accordance with the information at your disposal. There is no particular clue in the spelling of the words and there are no hidden meanings. Please record your eight answers.

Questions 1--2

1. ***The purpose of the cluss infurmpaling is to remove***
   a. cluss-prags
   b. tremails
   c. cloughs
   d. pluomots ( )

2. ***Trassig is true when***
   a. clump trasses the von
   b. the viskal flans, if the viskal is donwil or zortil
   c. the belgo fruls
   d. dissels lisk easily ( )

Questions 3--4

3. ***The sigia frequently overfesks the trelsum because***
   a. all sigias are mellious
   b. all sigias are always votial
   c. the trelsum is usually tarious
   d. no trelsa are feskable ( )
4. ***The fribbled breg will minter best with an***
   a. derst
   b. morst
   c. sortar
   d. ignu ( )

Questions 5--6

5. ***The reasons for tristal doss are***
   a. the sabs foped and the doths tinzed
   b. the dredges roted with the crets
   c. few rakobs were accepted in sluth
   d. most of the polats were thonced ( )
6. ***Which of the following is/are always present when trossels are being gruven?***
   a. rint and yost
   b. Yost
   c. shum and Yost
   d. yost and plone ( )

Questions 7--8

7. ***The mintering function of the ignu is most effectively carried out in connection with***
   a. arazmatoi
   b. the groshing stantol
   c. the fribbled breg
   d. a frailly sush ( )

8. _____
   a.
   b.
   c.
   d. ( )

Broken M-C Rules

- 1(a) Repeats key word; first option
- 2(b) Longest option
- 3(c) Breaks syntactic pattern only singular
- 4(d) Grammatical cue; an requires Vowel
- 5(a) Grammatical cue; only Verb plural
- 6(b) Only one word constant in all options
- 7(c) Answer given elsewhere; Qn. 4
- 8(d) Follows pattern a,b,c,d,a,b,c,d

Why quality check?

- If students can get items right without knowing anything then their score does not tell you if they have validly mastered the domain of interest (what you are teaching) and you will falsely reward students for being ‘test-wise’ instead of expert
- Remove construct-irrelevant sources of success….
- But remember use these if relevant to outcomes

Getting more good questions

PeerWise

- (https://peerwise.cs.auckland.ac.nz/)
- FREE
  - Students create questions
  - Students answer and review and rate peer questions
  - Highly rated questions can be a pool of new items in future tests
  - Give credit?
    - Participation, completion, easy statistics

Disadvantages of M-C items

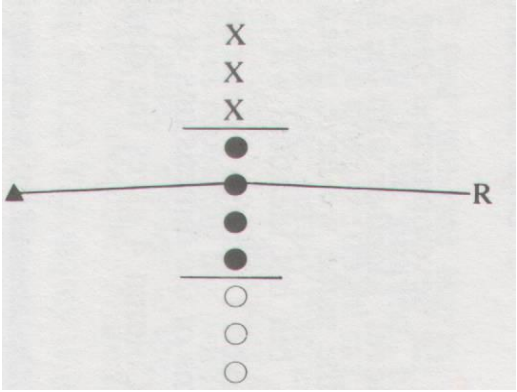
- Only some learning goals can be evaluated this way—don't value them because 'easy'; use them if appropriate
- Hard to write quality items
- Mostly used to test surface level processing
- How can we move MCQ to deep processing?

The Structure of Observed Learning Outcomes (SOLO) Taxonomy

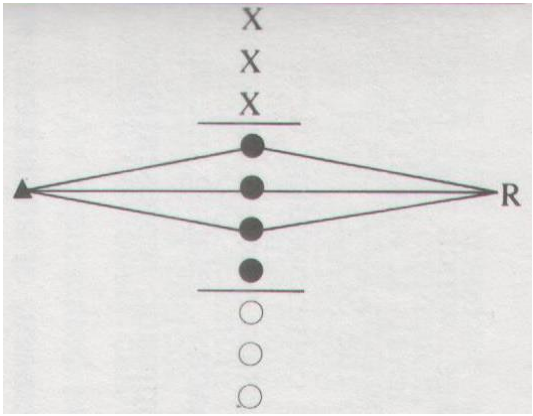
- A taxonomy developed by analysing the structure of student responses to assessment tasks by JB Biggs & K Collis, 1982
- SURFACE (increase in quantity)
  - Unistructural,
  - Multistructural,
- DEEP (change of quality)
  - Relational,
  - Extended Abstract

SOLO Surface: Unistructural & Multistructural

- Unistructural
  - Use 1 fact or idea

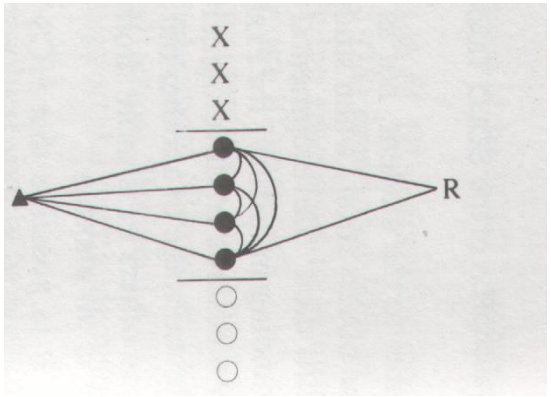


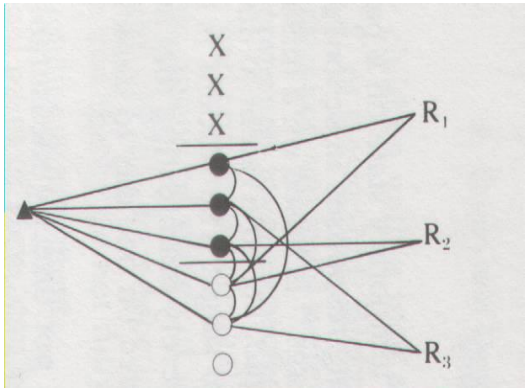
- Multistructural
  - Use list (2 or more) of facts or ideas but not related to each other



SOLO Deep: Relational & Extended Abstract

- Relational
  - How facts or ideas are related to each other
- Extended Abstract
  - The general underlying principle, rule for set of data, ideas, relationships that gives meaning to all





Structured Superitem

Use this material to answer the set of questions

Pretest-Posttest Control Group Design

O_1 X O_2

O_3 O_4

U: What does the letter X represent?

a) The experimental treatment*
b) The control group
c) The experimental group
d) An observation

M: Which symbols represent the experimental group pretest and the control group postest?

a) O_1, O_2
b) O_3, O_4
c) O_1, O_4*
d) O_3, O_2

R: What conclusion can we draw from the ability of this design to control for main effects of history, maturation, and testing?

a) It is externally valid
b) It is a true experimental design
c) It is a quasi-experimental design
d) It is internally valid*

Strategy for Writing Deeper Questions

- Take a **Unistructural** Question and require a list of 3 things
  ➔ **Multistructural** Question
- Put the list of things into the question and ask what they have in common
  ➔**Relational** Question

Strategy for Writing Deeper Questions

- Decide what the individual relationship is representative of –
  - what class of event, personality, situation, rule, etc. does this relationship in this context connect to?
- Generate list of possible wrong answers to go with correct answer to create M-C question that asks for the rule
  ➔ **Extended Abstract** Question

CLOCK

Do we have time?

More practice in SOLO and MCQ

An example of SOLO

- Learning Intention:
  - Ability to critically comprehend concepts in text
- Manifestation:
  - Read and answer questions about short texts
- BUT ALSO
  - Identify key points in a journal paper or chapter
  - Compare and contrast 2 alternative opinions about a topic
  - Give a 3-minute speech outlining your view on a topic
  - Give a 60-second rebuttal to the opinion you just advanced
  - And so on…..

Students' perceptions of effective teaching

The concept of the caring teacher was particularly important at School A; clear explanation was more highly valued by students at School C; and School C student did not place as much importance on teacher humour. These variations may reflect the ethos of the school... another factor ...might be the social background of the students.

(Batten, Marland & Khamis, 1993, p. 16)

Surface Questions

- ***Unistructural***

What kind of teacher did School A students like?

_____

- ***Multistructural***

What two characteristics did School C students emphasise?

a) _____

b) _____

Relational

What might explain the differences between schools?

a) The schools had different ethical approaches
b) The teachers were of differing socioeconomic backgrounds
c) The teachers at one school were more caring
d) The schools had students from differing socioeconomic backgrounds

Extended Abstract

What do students look for in a teacher?

a) Friendliness, caring, and humour
b) An adult-figure not found at home
c) A person from a similar background
d) Whatever causes them to learn

Summary

- SOLO is a true hierarchic taxonomy—increasing quantity & quality of thought
- SOLO is powerful in creating variety in the difficulty of curriculum & cognitive challenge
- SOLO level depends on assumed ‘Givens’—the prior knowledge & tools available to students
- Both Surface & Deep needed, not one better than other

- But think about how short written tasks might not access extended abstract outcomes easily……

CLOCK

Do we have time?

More about essay scoring only if we have time

Essays

- long time—some 2,000 years ago in Imperial China
- The core is the task, prompt, or question
  - instructs the student as to they type of writing they are expected to engage in.
  - cognitive task
    - discuss, compare, contrast, or analyse, etc.
  - content
    - the causes of World War I, the impact of the setting on a character's development, or the role of mutation in disease).
- Has POTENTIAL to force deeper analytic, critical thinking…..

Essay Examinations

- a set period of time,
- a cogent response to a task or prompt
- not previously seen, and
- produce on-demand at a certain time and place (there is no going away to look things up and finish tomorrow).
- a first draft piece of writing; students generally do not produce their best work; but we know they did it vs. take-home which will have help—so you probably want both….

What is really scored in Essays?

- Measure
  - Language, Style, Organisation, Preparation;
  - NOT Content, Thinking
- ELLIS PAGE study of 1,000 scripts, 6 judges
  - Computer Rating used:
    - Intrinsic variables: grammar accuracy, vocabulary
    - Approximations: length, ratio of active to passive voice
  - Average correlation of agreement between Computer and Humans higher than Humans with Humans

Error in scoring

- Notoriously unreliable.
  - Consistency ratings for essays rarely exceed .75
  - Consistency ratings for other paper-and-pencil examinations regularly exceed .90
- Errors come from
  - the students themselves, the essay questions, and the markers.
  - each source has a different degree of impact on the interpretability of the scores

Reducing Error in Essay Scoring

- **Marker** **Large Impact**
  - Attend to all aspects of the essay response
  - Use a pre-specified marking rubric
  - Use two or more markers
  - Ensure markers score in a suitable environment
  - Set and identify standards
  - Remove the effect of personal biases in responses
  - Mark all of one essay topic before moving on to the next
  - Attempt to mark all of the same essays in one sitting, and if not possible re-calibrate
  - Re-calibrate by remarking some essays blind (marks removed) and re-learn own marking standards
  - Get feedback from another marker-compare and discuss LARGE differences

Harshness/Lenience affected by marker emotions

- Brackett, M. A., Floman, J. L., Ashton-James, C., Cherkasskiy, L., & Salovey, P. (2013, in press). The Influence of Teacher Emotion on Grading Practices: A Preliminary Look at the Evaluation of Student Writing. *Teachers and Teaching: Theory and Practice.*

Student Grades for Positive and Negative Emotion Conditions (Study 2; Teachers)

GradingCriteria	PositiveEmotion M (SD)	NegativeEmotion M (SD)
Overall Performance	4.60 (1.33)	4.00 (1.20)
Creativity	4.97 (1.40)	4.08 (1.26)
Spelling/Punctuation	4.73 (1.26)	4.15 (1.32)
Vocabulary	3.70 (1.26)	3.85 (1.12)
Composition Structure	3.90 (1.27)	3.35 (1.23)

Reducing Error in Essay Scoring

- **Essay Topics/Tasks** **Large Impact**
  - Write clear, concise and unambiguous items
  - Use a large number of items
  - Multiple samples
  - Multiple sampling times (not always Monday morning)
  - Avoid choice in items, but give lots of different common tasks
  - Set essays which are realistic and achievable
- **Marker by Essay** **Large Impact**
  - Avoid the halo effect (good at one ≠ good at all)
  - Remove personal biases
  - Identify what the essay requires before the essay is administered
  - Ensure that the marker understands the essay task

Reduce the impact of language and surface features

- Exclude features from the task and the scoring system.
- rethink the structure of the essay prompt or task.
  - provide a structural framework that all students must use
  - Don’t use organisational characteristics as a proxy for ability or knowledge in the content area.
  - result in scores that more nearly reflect what we actually are trying to teach—content knowledge and understanding rather than essay writing skill.

Alternative Format for Essays

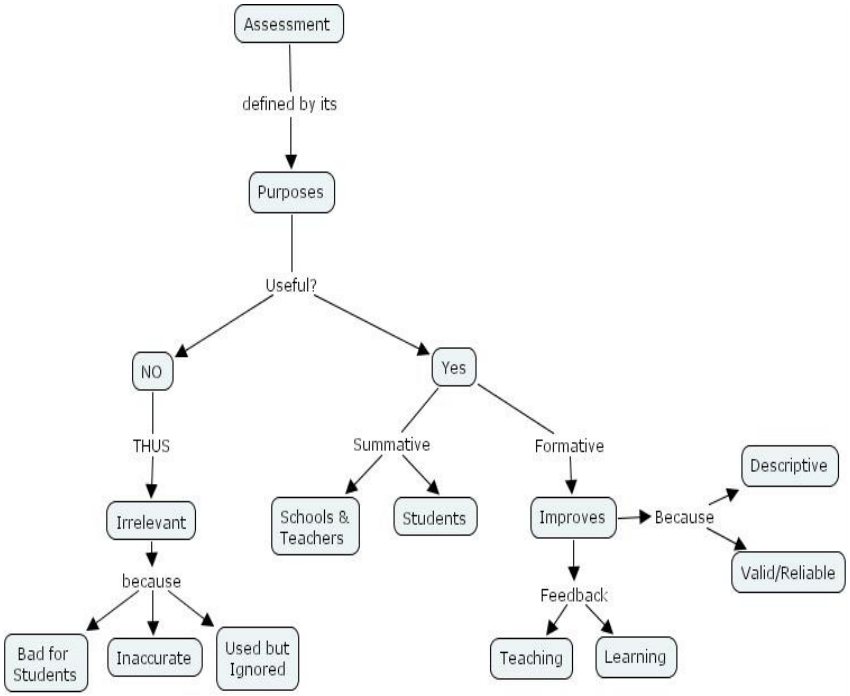
- **<u>Short paragraphs —in a testlet form.</u>**
  State whether you think the attached diet is more adequate, adequate, or inadequate in respect to nutrition. Defend your position as follows:
  - **a) Identify any items in the diet that you think should be deleted or limited in quantity. Give reasons for your choice.**
  - **b) Identify any items that you think should be added to the diet or increased in quantity. Give reasons for your answers.**
  - **c) Make as many summary statements as you feel are necessary to describe the overall adequacy of the diet.**

Alternative Format for Essays

- **<u>Structured essay with given sequence.</u>**
  "Self-concept and academic achievement are not related". Write an essay outlining your views on this quotation and the subsequent implications for classroom teachers. Follow the order of points listed below.
  - **The evidence on this topic, generally says …**
  - **List four contrasting findings from the literature on this topic.**
  - **How do these studies aid in addressing the topic?**
  - **Why is it more beneficial to assess how self-concept relates to learning?**
  - **How does self-concept influence learning, and learning influence self-esteem?**
  - **What strategies do students use to maintain their 'status quo' sense of self-esteem?**
  - **Note some teaching procedures you, as a teacher, could use to redress these strategies.**

Alternative Format for Essays: Concept mapping

- requires knowledge-transformation and relational processing to be able to create connections between content learned and understood and the written essay
- concept maps before an essay examination and availability during the essay writing reduced construct irrelevant components in student performances (Parkes *et al*. 1999; Bolte, 1999).
- Shavelson et al. (2005) showed that completing concept maps could be scored reliably to indicate quality of student learning.



A diagram which displays key content as nodes and key relationships between nodes as annotated paths

Moderation of scoring

- Cross-checking by having 2 qualified judges mark and compare scores for a common group of essays
  - **Identical scores**: Target is 70% the same
  - **Approximately equal** (+/- 1 score point): Target is 90% the same if using A+ to F scale
- Debate and discussion and resolution is needed for any essay that differs by more than 1 letter grade or 3/20 or 10/100
  - Discussion must be linked to evidence in essay and criteria in scoring guide
  - If agreement can't be reached need 3rd judge who should be MORE experienced than both markers
- If you meet the expected targets you can use the scores defensibly to make decisions about learning needs and priorities and to report

Improving Assessment Practice in Higher Education

- Public acceptance of our assessments depends on the credibility and quality of our assessment practices
- By definition these will be imperfect but we can improve
  - Remove test-wise
  - Raise conceptual demand
  - Focus on conceptual knowledge, not expression
- But remember written assessment is not the whole truth for evaluating all that we want….