Test Item Writing from the Perspective of Measurement Theory

Prof Gavin T L Brown
Quantitative Data Analysis & Research Unit
gt.brown@auckland.ac.nz

THE UNIVERSITY OF AUCKLAND
Te Whare Wānanga o Tāmaki Makaurau
NEW ZEALAND

EDUCATION AND SOCIAL WORK

THE UNIVERSITY OF AUCKLAND
Te Whare Wānanga o Tāmaki Makaurau
NEW ZEALAND

EDUCATION AND SOCIAL WORK

Published Standardised Tests that I have helped develop…

- Hattie, J. A. C., Brown, G. T. L., Keegan, P. J., et al. (2004, December). *Assessment Tools for Teaching and Learning (asTTle) Version 4, 2005: Manual.* Wellington, NZ: University of Auckland/ Ministry of Education/ Learning Media. [Mx, Rdg, Wrg Levels 2-6]
- Hattie, J. A. C., Brown, G. T. L., Keegan, P. J., et al. (2003, December). *Assessment Tools for Teaching and Learning (asTTle) Manual* (Version 3 2004). Wellington: Learning Media. [Mx Level 2-4, Rdg & Wrg Levels 2-6]
- Hattie, J. A., Brown, G. T. L., & Keegan, P. J. (2002). *Assessment Tools for Teaching and Learning (asTTle) manual: Version 2, 2003* (V2 edn.). Wellington, NZ: Learning Media. [Mx, Rdg, Wrg Levels 2-4]
- Hattie, J. A., Brown, G. T. L., & Keegan, P. J. (2002). *Assessment Tools for Teaching and Learning (asTTle) manual: English literacy 2002* (V1 edn.). Wellington, NZ: Learning Media. [Rdg & Wrg Levels 2-4]
- Croft, C., Dunn, K., & Brown, G. T. L. (2001). *Essential Skills Assessment: Information Skills. Manual.* Wellington: NZCER. [Grades 5-10]

THE UNIVERSITY OF AUCKLAND NEW ZEALAND | EDUCATION AND SOCIAL WORK

Assessment Defined

Appropriate and accurate interpretations and decisions based on appropriate collection of valid information about valued content (a domain of interest)

THE UNIVERSITY OF AUCKLAND NEW ZEALAND | EDUCATION AND SOCIAL WORK

Domain

- A theoretically defined topic or construct of interest
  - Curriculum defined
  - Classroom experience informed
- Domains are structured bodies of knowledge
  - Taxonomic structures that indicate entry points, key points, sub-fields and relationships
- May be debate as to what is in the domain depending on related fields such as
  - Teaching theory
- Hence, it can be difficult to agree upon

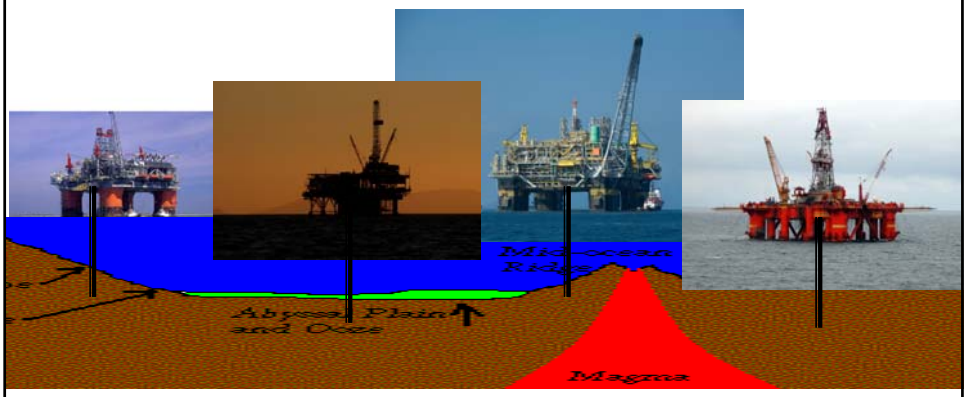
THE UNIVERSITY OF AUCKLAND NEW ZEALAND | EDUCATION AND SOCIAL WORK

Defining a test

- A sample of tasks, questions, items drawn from a domain of interest intended to elicit information about learner skill, knowledge, understanding about that domain
- In order to make inferences about:
  - Improving learning/teaching;
  - Evaluating students;
  - Evaluating schools/teachers;
  - Evaluating curriculum.

THE UNIVERSITY OF AUCKLAND NEW ZEALAND | EDUCATION AND SOCIAL WORK

Sampling a domain

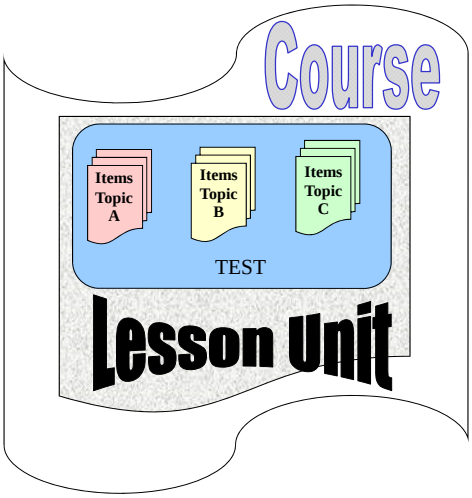


Test=Sample

THE UNIVERSITY OF AUCKLAND NEW ZEALAND | EDUCATION AND SOCIAL WORK

- Is the information sufficient & representative in order to dependably generalise to the domain we're interested in?
- What are the limits of the information's meaning?
- What does knowing the answers to these items tell us about person's knowledge of chapter and ability to function in domain?

- **HENCE, Multiple items to generate a robust estimate of competence in a domain**



PS: What implication if all the marks came from one section?

THE UNIVERSITY OF AUCKLAND NEW ZEALAND | EDUCATION AND SOCIAL WORK

Basic Principles of Test Design

- Know your domain—identify, describe what you want to teach and learn
  - reading is not one thing;
  - maths is not arithmetic
- Select rich ideas for important content
  - What are you really testing?
    - Generally content * cognitive function * difficulty

THE UNIVERSITY OF AUCKLAND NEW ZEALAND | EDUCATION AND SOCIAL WORK

Basic Principles of Test Design

- Consider cognitive demand
  - SOLO taxonomy; Bloom's revised taxonomy
- Consider logistics
  - Item formats (multiple styles),
  - Time, length;
  - Difficulty
- Agenda: enough information to be confident of any conclusions you want to draw
  - How many times does a student have to do something for us to conclude that she knows/can do it?

THE UNIVERSITY OF AUCKLAND NEW ZEALAND | EDUCATION AND SOCIAL WORK

Test Blueprint/Template

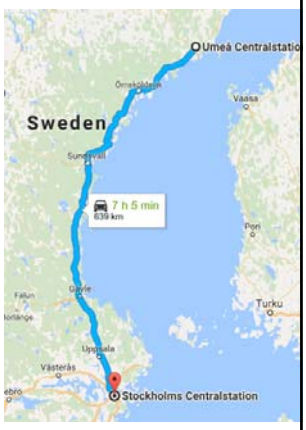
- A powerful way to organise writing items and reporting results
  - A test is a structured sample to permit valid inferences and actions about those content areas

	Style		Cognitive Demand	
Content Areas	Selected	Constructed	Surface	Deep
1.				
2.				
3.				
4.				

THE UNIVERSITY OF AUCKLAND NEW ZEALAND | EDUCATION AND SOCIAL WORK

Progress

- The point of testing is not just to describe the lay of the land
- It is also to identify where on the journey a student is so that appropriate responses can be made
  - Signposts guide appropriate actions



Gävle kommun

Uppsala KOMMUN

THE UNIVERSITY OF AUCKLAND NEW ZEALAND | EDUCATION AND SOCIAL WORK

Educational challenge

- We are trying to encourage growth, development
- Young people are not always stable in 'learning' something
- The things we want students to learn are both simple and complex; and sometimes never easy
- Our estimates are not usually accurate
  - Multiple estimates, multiple judges, multiple opportunities, multiple methods

THE UNIVERSITY OF AUCKLAND NEW ZEALAND | EDUCATION AND SOCIAL WORK

The point of test items

- To generate evidence of competence around key parts of the domain to permit legitimate decision making that will be acceptable to students, parents, administrators, colleagues, funders, etc.
- Enough items of different types to eliminate alternative explanations
- Aligned to curricular practices and goals
- Challenging but doable

THE UNIVERSITY OF AUCKLAND NEW ZEALAND | EDUCATION AND SOCIAL WORK

Assumptions in Measurement

- If something exists we can measure it
- If we can't measure it, we may have failed in our ingenuity and ability to measure—it doesn't mean the thing doesn't exist
- We have an unfortunate tendency to treat as valuable only the things we can measure
  - We tend to treat as value-less anything we can't measure even if it does have value
- We have unfortunate tendency to settle for the easy to test parts as a proxy for the hard stuff

THE UNIVERSITY OF AUCKLAND NEW ZEALAND | EDUCATION AND SOCIAL WORK

Assumptions in Measurement

- All measures have some error
  - The length of a certain platinum bar in Paris is a metre
  - It is supposed to be 1/10,000,000th of the distance from equator to pole
  - **but** it is short by 0.2 mm according to satellite surveys
- Less error in measures of physical phenomena and more in social phenomena

THE UNIVERSITY OF AUCKLAND NEW ZEALAND | EDUCATION AND SOCIAL WORK

Assumptions in Measurement

- Reality has patterns partly because random chance has patterns in it
  - Toss a coin enough times and you will get patterns like this:

T H, T T H H, T T T H H H, T T T T H H H H

OR

T T T T T T T, H H H H H H H

  - If you measure something often enough, one of your results will appear to be non-chance, even though it actually is a chance event

THE UNIVERSITY OF AUCKLAND NEW ZEALAND | EDUCATION AND SOCIAL WORK

Assumptions in Measurement

- Chance plays a significant part in the results we generate
  - Sometimes the result we get could occur by chance anyway; just because something happens doesn't mean it would not have occurred anyway
  - If it is relatively unlikely to occur by chance then we say it is "*Statistically significant*"
    - So we create tables of how often things occur by chance as a reference point
  - We can estimate the probability (*p*) of something happening by chance and use this to determine whether our result is real or a chance artefact

THE UNIVERSITY OF AUCKLAND NEW ZEALAND | EDUCATION AND SOCIAL WORK

Measurement

- Properties are measured with tools that are calibrated into numerical scales
- Ratio Scale
  - E.g., length➔metres; weight➔kilograms;
  - The distance between any two markers is identical
  - There is a non-arbitrary zero-value
    - (there is a real world state that can be described as lacking any of the property of interest; but you can't have less than ZERO)
  - The point on the scale is a RATIO of a base unit and addition of the ratio values is possible
    - Full statistical analyses can be done

THE UNIVERSITY OF AUCKLAND NEW ZEALAND | EDUCATION AND SOCIAL WORK

Other types of scales

- Interval scale
  - Temperature in Celsius
    - (each point is 1/100th of distance between freezing and boiling point of fresh water at sea level)
  - Equal size distances between scale points
    - (1 to 5 is the same distance as 6 to 10)
  - Differences can be ratios:
    - 10 to 20 is twice the distance of 1 to 5
  - Zero point is arbitrary; negative scale points possible
  - Standard arithmetic can be applied to such scales especially concerning centre of distributions

THE UNIVERSITY OF AUCKLAND NEW ZEALAND | EDUCATION AND SOCIAL WORK

Other types of scales

- Ordinal scale
  - Rank ordered objects
    - 1st, 2nd, 3rd in a race
    - Tall, Taller, Tallest (not 1.80m; 2.00m; 2.26m)
    - Neutral, slightly agree, moderately agree, strongly agree
  - Distance between ranks or orders not necessarily equal
  - Challenge of how to do mathematical operations with some that is not continuous (non-parametric statistics)
    - We need evidence that distances are at least approximately equal to use rank orders—what evidence would be good?

THE UNIVERSITY OF AUCKLAND NEW ZEALAND | EDUCATION AND SOCIAL WORK

Other types of scales

- Nominal scale
  - Categorical or classification naming of objects that are qualitatively different to each other
    - Igneous, sedimentary, metamorphic
      - These can all weigh the same or have the same length but on critical features they are not identical
    - Male, female
    - Agree, disagree
    - Right, wrong
  - We can count the frequency of each category and compare the distribution of frequencies in samples

THE UNIVERSITY OF AUCKLAND NEW ZEALAND | EDUCATION AND SOCIAL WORK

Measuring Human mental properties

- Problem of measuring stuff 'in-the-head'
  - Our mental actions are difficulty to observe directly
- How do we know how much *xxx* you have?
  - Measure *xxx* with a recognised tool
- Measuring Tools for Social Science
  - Answers to Questions (paper or oral)
  - Self-reports
  - Observational Check Lists
- What scale properties do tools like these have?

THE UNIVERSITY OF AUCKLAND
NEW ZEALAND
EDUCATION AND SOCIAL WORK

Measuring Human mental properties

- What kind of SCALE is the sum of items answered correctly?
- What kind of mathematical analysis can we use with that type of scale?
- We calculate means and standard deviations of test scores—do we have the right scale properties for this?
- What type of mathematical or statistical methods do we need to take into account chance factors?