Principles of Test Development

Presentation to the BVM Group, Umea University

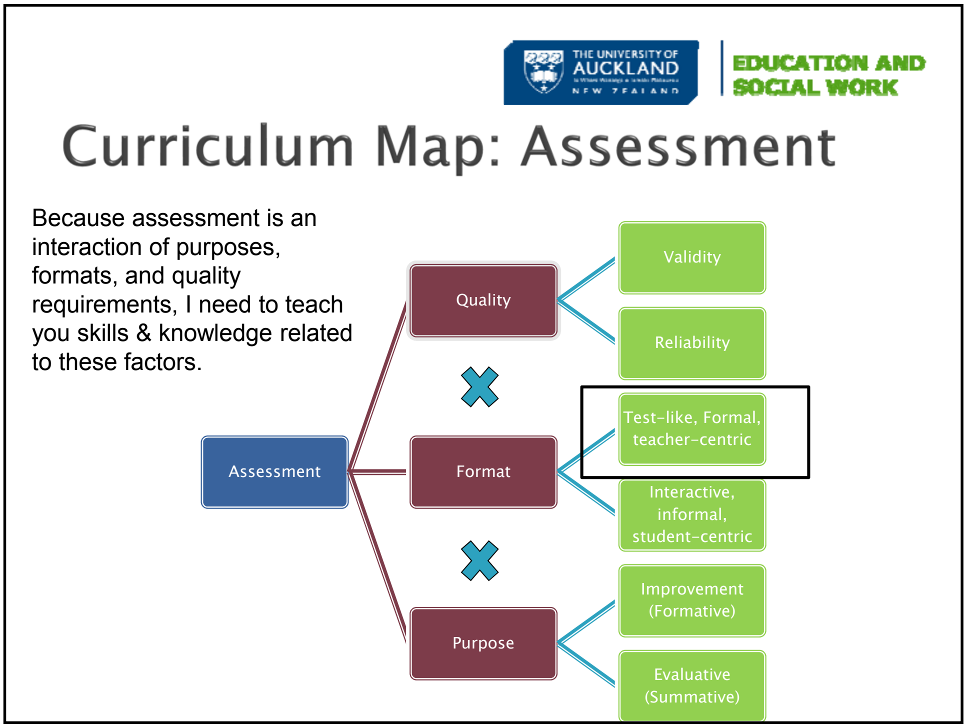
Sept 2017

Prof Gavin T L Brown

gt.brown@Auckland.ac.nz

THE UNIVERSITY OF AUCKLAND NEW ZEALAND

EDUCATION AND SOCIAL WORK



THE UNIVERSITY OF AUCKLAND NEW ZEALAND

EDUCATION AND SOCIAL WORK

Curriculum Map: Assessment

Because assessment is an interaction of purposes, formats, and quality requirements, I need to teach you skills & knowledge related to these factors.

Assessment

Quality

Validity

Reliability

Format

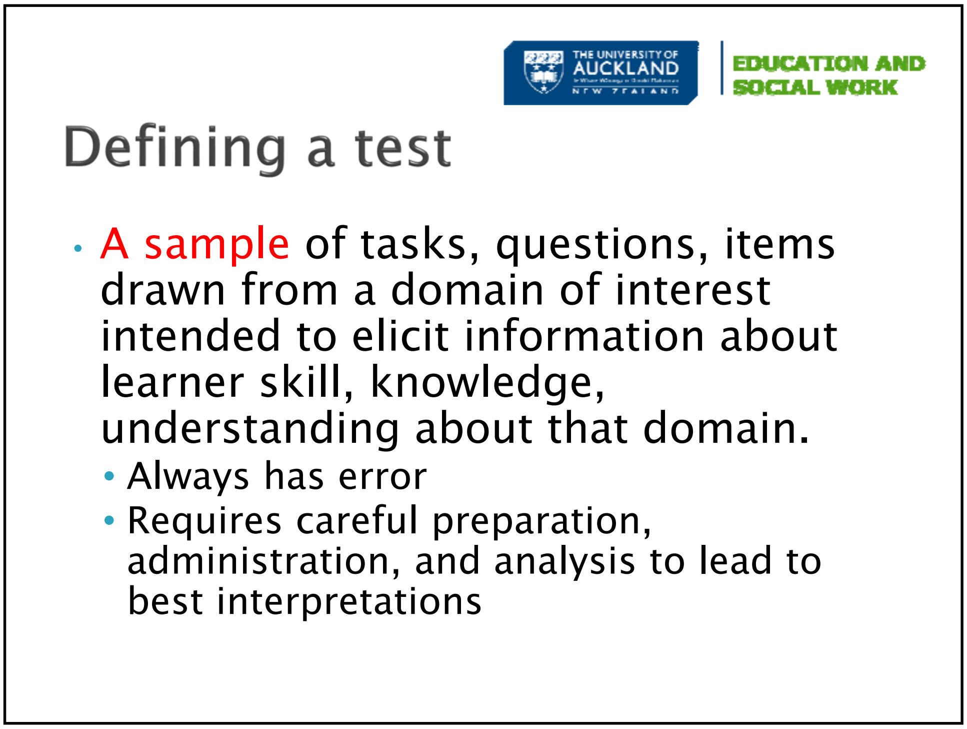
Test-like, Formal, teacher-centric

Interactive, informal, student-centric

Purpose

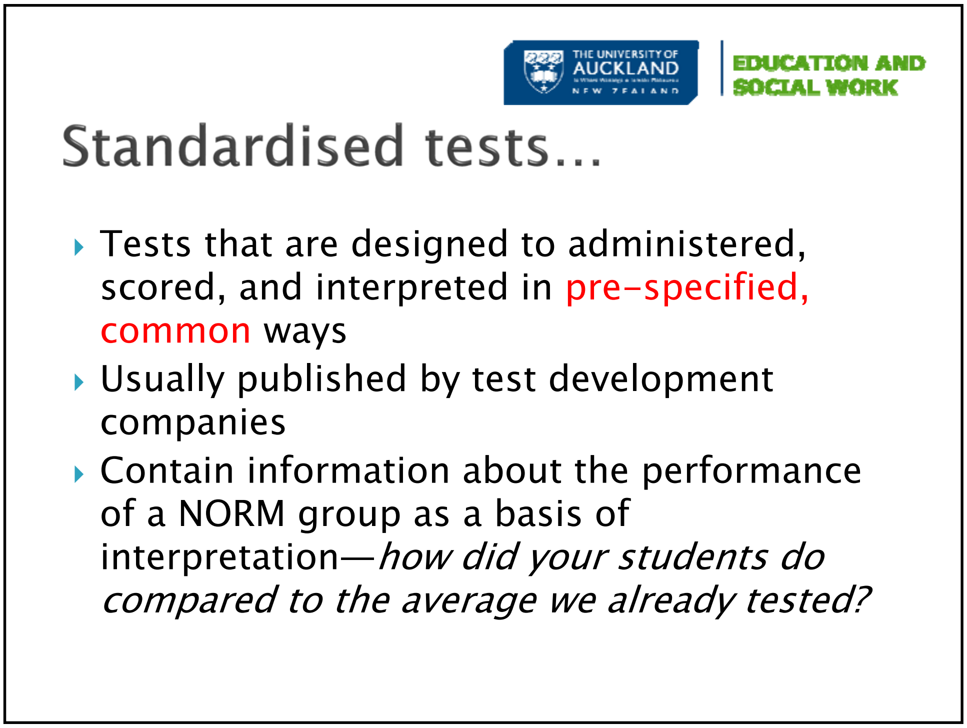
Improvement (Formative)

Evaluative (Summative)



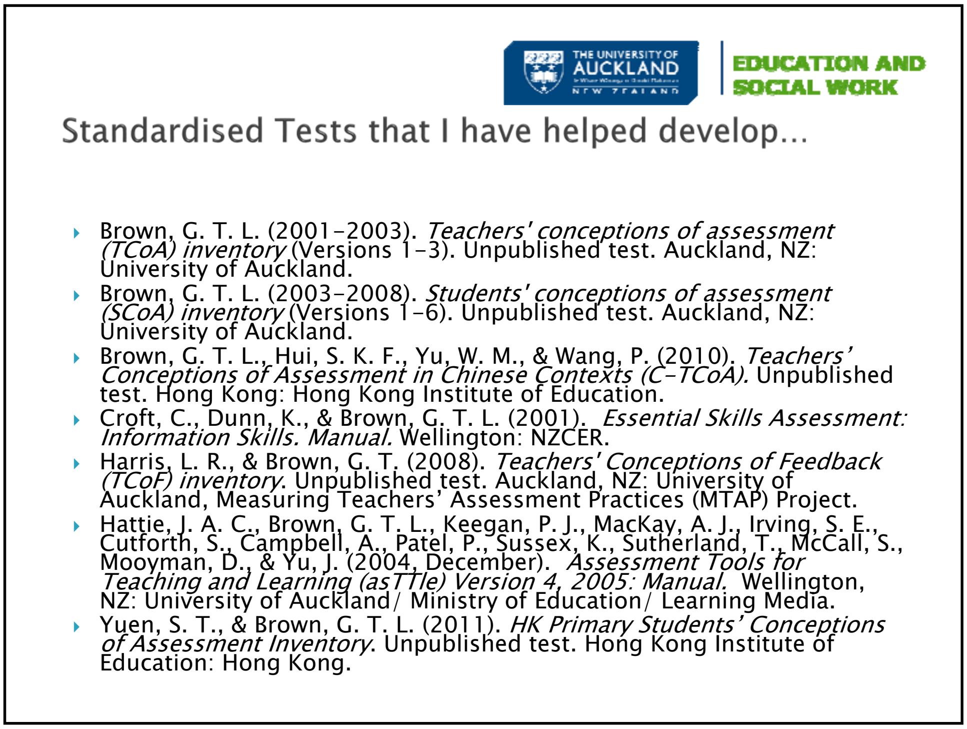
Defining a test

- A sample of tasks, questions, items drawn from a domain of interest intended to elicit information about learner skill, knowledge, understanding about that domain.
  - Always has error
  - Requires careful preparation, administration, and analysis to lead to best interpretations



Standardised tests…

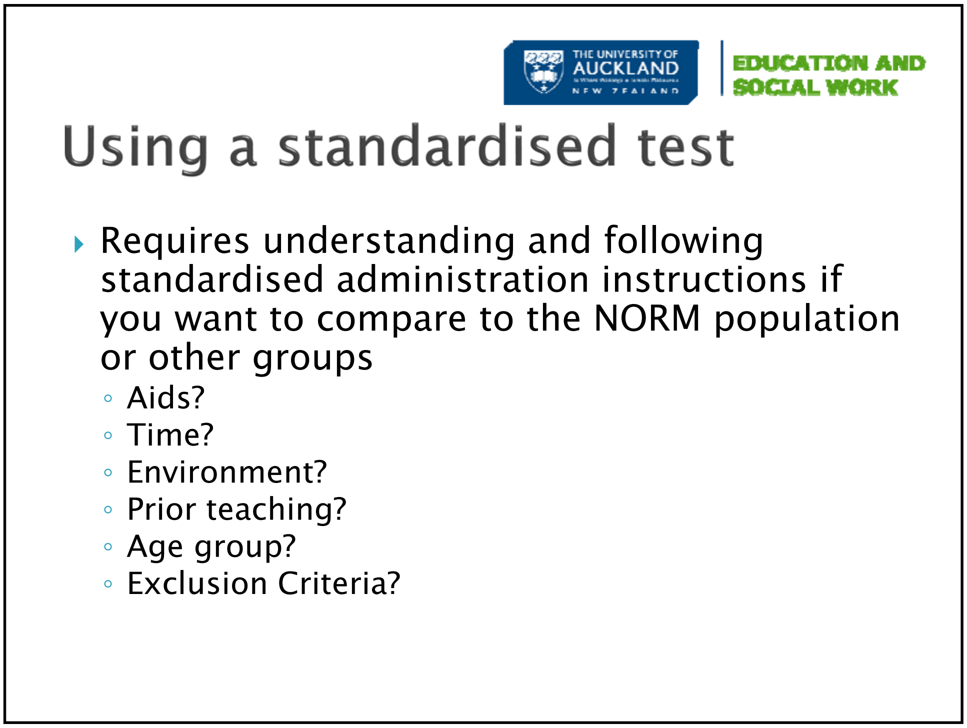
- Tests that are designed to administered, scored, and interpreted in pre-specified, common ways
- Usually published by test development companies
- Contain information about the performance of a NORM group as a basis of interpretation—*how did your students do compared to the average we already tested?*



THE UNIVERSITY OF AUCKLAND NEW ZEALAND | EDUCATION AND SOCIAL WORK

Standardised Tests that I have helped develop…

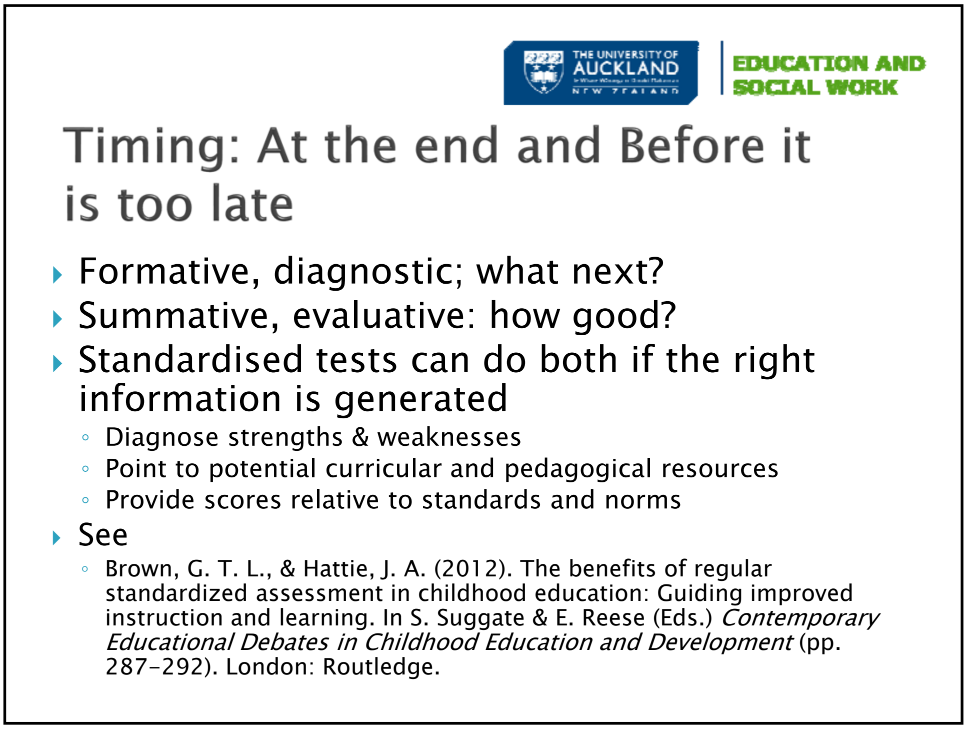
- Brown, G. T. L. (2001-2003). *Teachers' conceptions of assessment (TCoA) inventory* (Versions 1-3). Unpublished test. Auckland, NZ: University of Auckland.
- Brown, G. T. L. (2003-2008). *Students' conceptions of assessment (SCoA) inventory* (Versions 1-6). Unpublished test. Auckland, NZ: University of Auckland.
- Brown, G. T. L., Hui, S. K. F., Yu, W. M., & Wang, P. (2010). *Teachers' Conceptions of Assessment in Chinese Contexts (C-TCoA).* Unpublished test. Hong Kong: Hong Kong Institute of Education.
- Croft, C., Dunn, K., & Brown, G. T. L. (2001). *Essential Skills Assessment: Information Skills. Manual.* Wellington: NZCER.
- Harris, L. R., & Brown, G. T. (2008). *Teachers' Conceptions of Feedback (TCoF) inventory.* Unpublished test. Auckland, NZ: University of Auckland, Measuring Teachers' Assessment Practices (MTAP) Project.
- Hattie, J. A. C., Brown, G. T. L., Keegan, P. J., MacKay, A. J., Irving, S. E., Cutforth, S., Campbell, A., Patel, P., Sussex, K., Sutherland, T., McCall, S., Mooyman, D., & Yu, J. (2004, December). *Assessment Tools for Teaching and Learning (asTTle) Version 4, 2005: Manual.* Wellington, NZ: University of Auckland/ Ministry of Education/ Learning Media.
- Yuen, S. T., & Brown, G. T. L. (2011). *HK Primary Students' Conceptions of Assessment Inventory.* Unpublished test. Hong Kong Institute of Education: Hong Kong.



THE UNIVERSITY OF AUCKLAND NEW ZEALAND | EDUCATION AND SOCIAL WORK

Using a standardised test

- Requires understanding and following standardised administration instructions if you want to compare to the NORM population or other groups
  - Aids?
  - Time?
  - Environment?
  - Prior teaching?
  - Age group?
  - Exclusion Criteria?

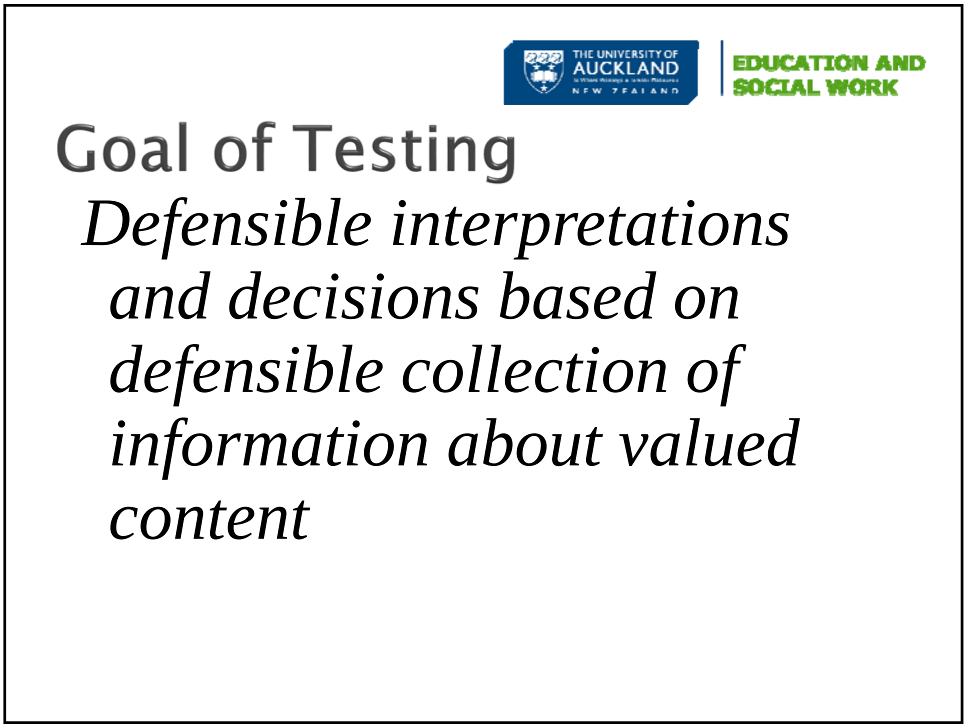


THE UNIVERSITY OF AUCKLAND NEW ZEALAND

EDUCATION AND SOCIAL WORK

Timing: At the end and Before it is too late

- Formative, diagnostic; what next?
- Summative, evaluative: how good?
- Standardised tests can do both if the right information is generated
  - Diagnose strengths & weaknesses
  - Point to potential curricular and pedagogical resources
  - Provide scores relative to standards and norms
- See
  - Brown, G. T. L., & Hattie, J. A. (2012). The benefits of regular standardized assessment in childhood education: Guiding improved instruction and learning. In S. Suggate & E. Reese (Eds.) *Contemporary Educational Debates in Childhood Education and Development* (pp. 287-292). London: Routledge.



THE UNIVERSITY OF AUCKLAND NEW ZEALAND

EDUCATION AND SOCIAL WORK

Goal of Testing

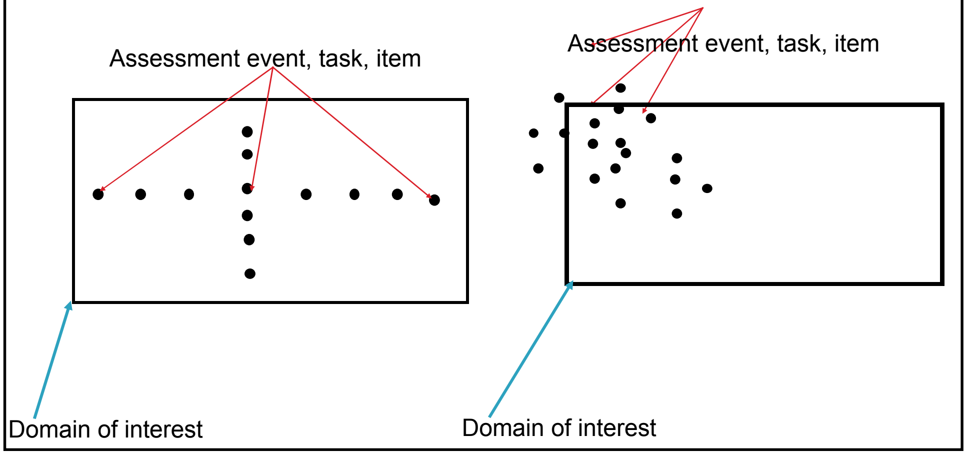
Defensible interpretations and decisions based on defensible collection of information about valued content

Message 1

Development of standardized tests is an art form not an exact science. But we have developed conventions and standards. It requires professional judgement of the developers informed by expert comment, user feedback, student response, and statistics.

Alignment to Curricular Goals

- If you want assessment to be aligned to curricular goals, the assessment must be within the curriculum
  - *Which is better?*



But what about difficulty?

- Don’t just group content but also think about easy-medium-hard within an appropriate range for the teaching you are doing

PLANNING: test template/blueprint or table of specifications

Tests have specified

- Time limits
- Content coverage
- Number of items
- Response formats
- Language

We think about what can go wrong and prepare for it

Content Areas	Selected	Constructed	Surface	Deep
1.				
2.				
3.				
4.				

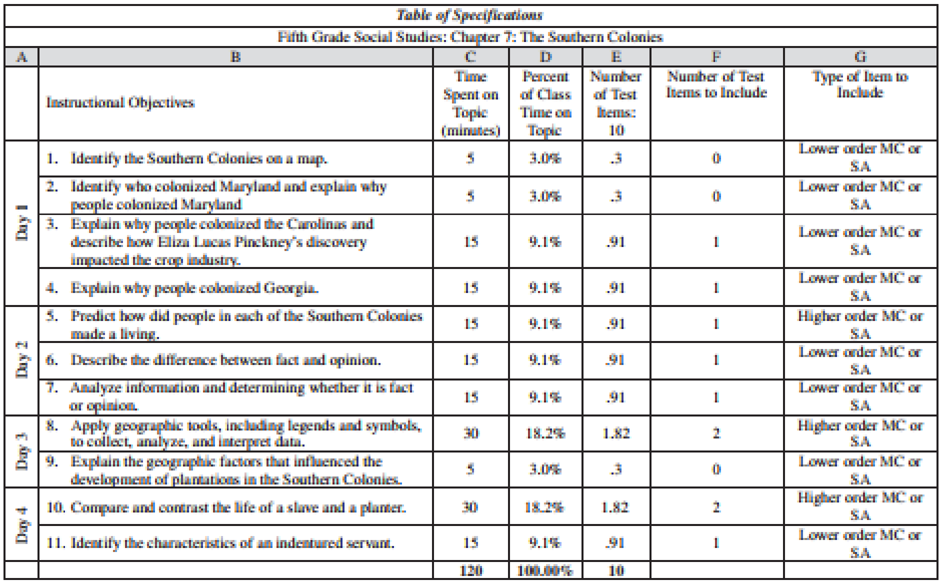
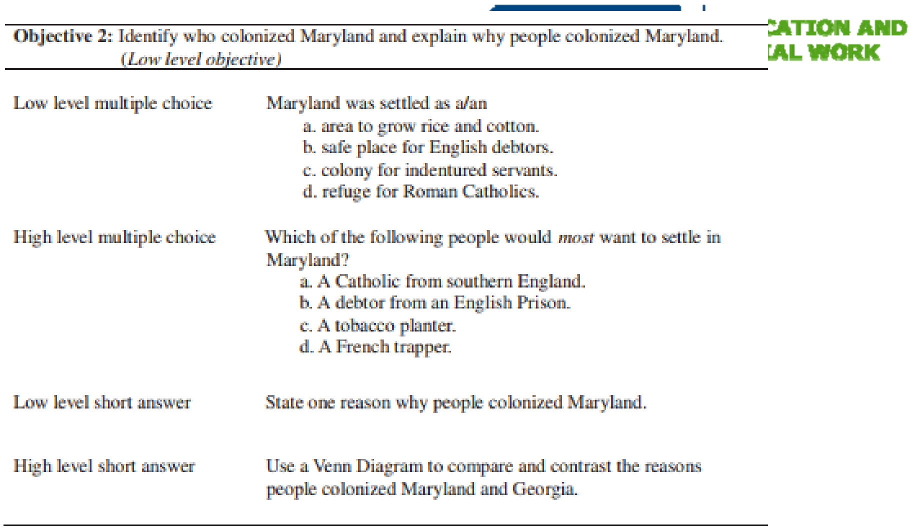


Table of Specifications

Fifth Grade Social Studies: Chapter 7: The Southern Colonies

A	B	C	D	E	F	G
	Instructional Objectives	Time Spent on Topic (minutes)	Percent of Class Time on Topic	Number of Test Items: 10	Number of Test Items to Include	Type of Item to Include
Day 1	1. Identify the Southern Colonies on a map.	5	3.0%	.3	0	Lower order MC or SA
	2. Identify who colonized Maryland and explain why people colonized Maryland	5	3.0%	.3	0	Lower order MC or SA
	3. Explain why people colonized the Carolinas and describe how Eliza Lucas Pinckney's discovery impacted the crop industry.	15	9.1%	.91	1	Lower order MC or SA
	4. Explain why people colonized Georgia.	15	9.1%	.91	1	Lower order MC or SA
Day 2	5. Predict how did people in each of the Southern Colonies made a living.	15	9.1%	.91	1	Higher order MC or SA
	6. Describe the difference between fact and opinion.	15	9.1%	.91	1	Lower order MC or SA
	7. Analyze information and determining whether it is fact or opinion.	15	9.1%	.91	1	Lower order MC or SA
Day 3	8. Apply geographic tools, including legends and symbols, to collect, analyze, and interpret data.	30	18.2%	1.82	2	Higher order MC or SA
	9. Explain the geographic factors that influenced the development of plantations in the Southern Colonies.	5	3.0%	.3	0	Lower order MC or SA
Day 4	10. Compare and contrast the life of a slave and a planter.	30	18.2%	1.82	2	Higher order MC or SA
	11. Identify the characteristics of an indentured servant.	15	9.1%	.91	1	Lower order MC or SA
		120	100.00%	10		

DiDonato-Barnes, N., Fives, H., & Krause, E. S. (2014). Using a Table of Specifications to improve teacher-constructed traditional tests: An experimental design. *Assessment in Education: Principles, Policy & Practice, 21*(1), 90-108. doi:10.1080/0969594X.2013.808173



Objective 2: Identify who colonized Maryland and explain why people colonized Maryland. (*Low level objective*)

Low level multiple choice	Maryland was settled as a/an a. area to grow rice and cotton. b. safe place for English debtors. c. colony for indentured servants. d. refuge for Roman Catholics.
High level multiple choice	Which of the following people would *most* want to settle in Maryland? a. A Catholic from southern England. b. A debtor from an English Prison. c. A tobacco planter. d. A French trapper.
Low level short answer	State one reason why people colonized Maryland.
High level short answer	Use a Venn Diagram to compare and contrast the reasons people colonized Maryland and Georgia.

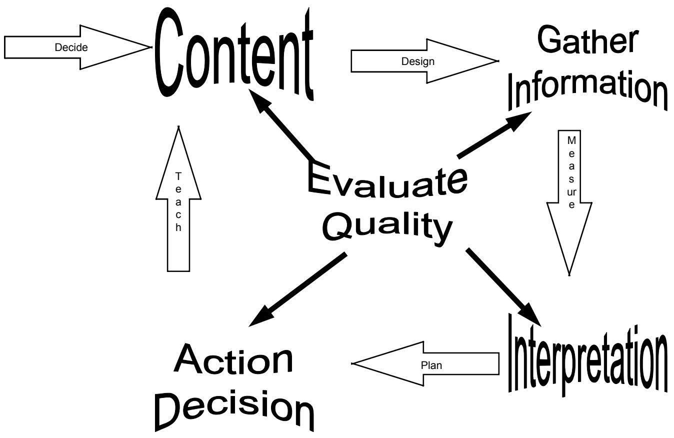
significant differences in the quality of TCE (i.e., assessment or test items adequately assess the subject matter that was taught) scores (treatment group scored higher than the comparison group)…This finding provides empirical support that the Table of Specifications can help teachers choose test items that adequately assess the subject matter that was taught.

Message 2

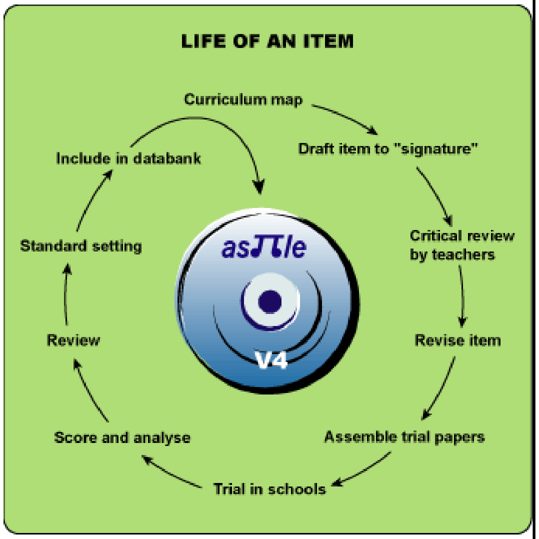
Standardized tests can be used to inform classroom-based decision making; teachers need to know how the tests permit that.

The quality of design and development lends credibility.

Assessment Processes: Technological Approach



Development Cycle of a Standardised Test



LIFE OF AN ITEM

Curriculum map

Draft item to "signature"

Critical review by teachers

Revise item

Assemble trial papers

Trial in schools

Score and analyse

Review

Standard setting

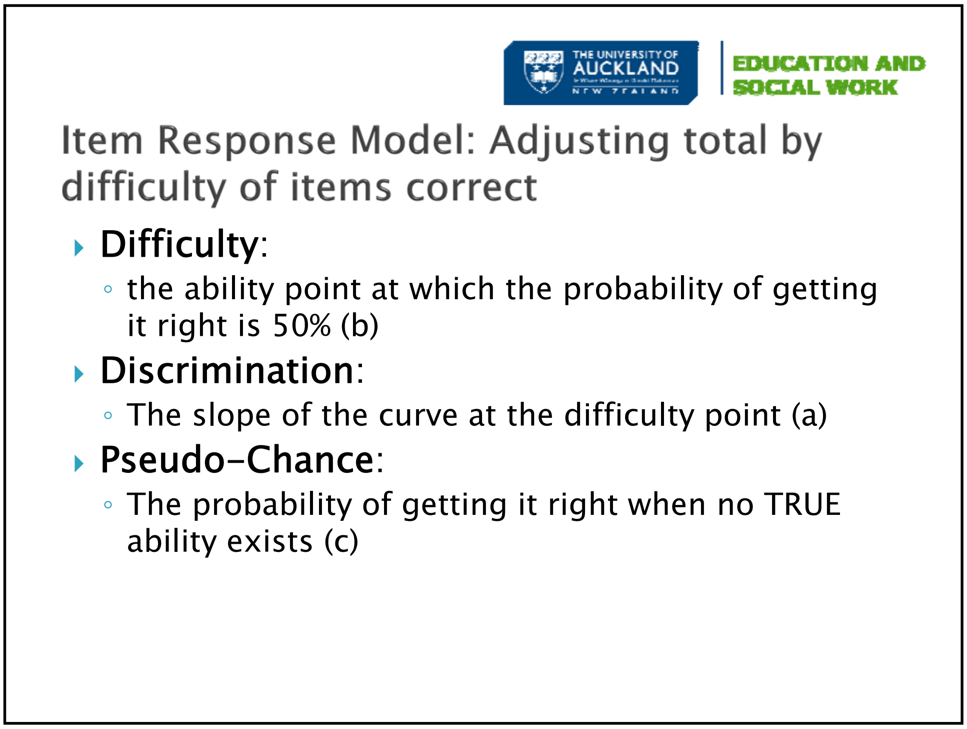
Include in databank

Classical Test Theory: Approach to creating a score

The sum of all items answered correctly (divided by max items for %)

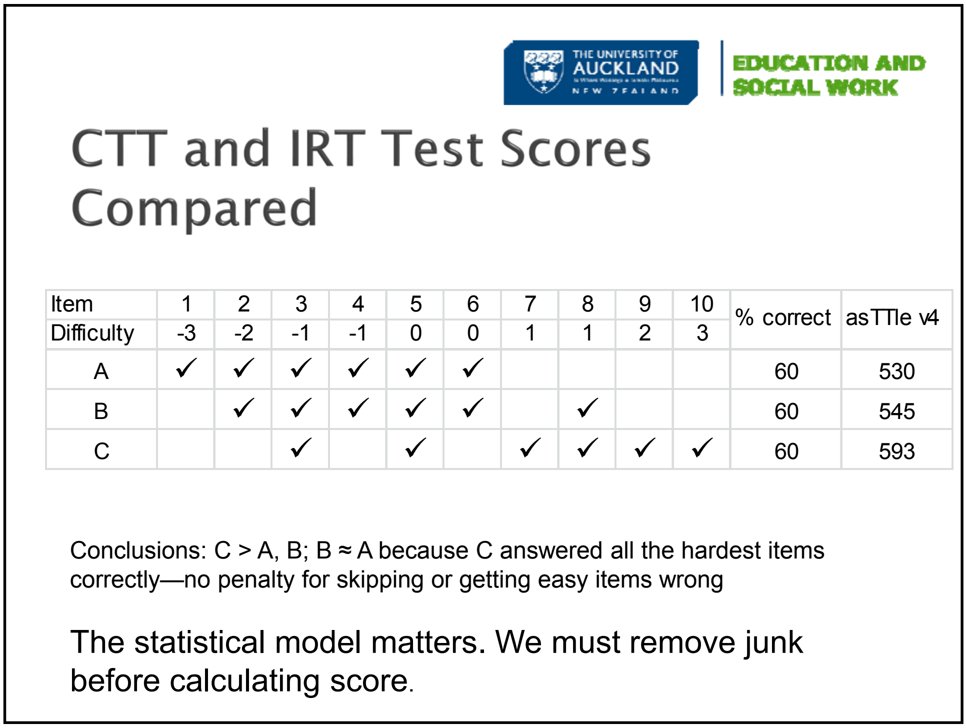
Key item information

- How hard is the item?➔Item Difficulty (*p*): % of people who answered correctly
  - Make sure items not too hard or too easy
- Who gets the item right?➔Item Discrimination
  - Correlation between each item and the total score without the item in it
  - Ideally, look for $r > .20$



Item Response Model: Adjusting total by difficulty of items correct

- **Difficulty**:
  - the ability point at which the probability of getting it right is 50% (b)
- **Discrimination**:
  - The slope of the curve at the difficulty point (a)
- **Pseudo-Chance**:
  - The probability of getting it right when no TRUE ability exists (c)

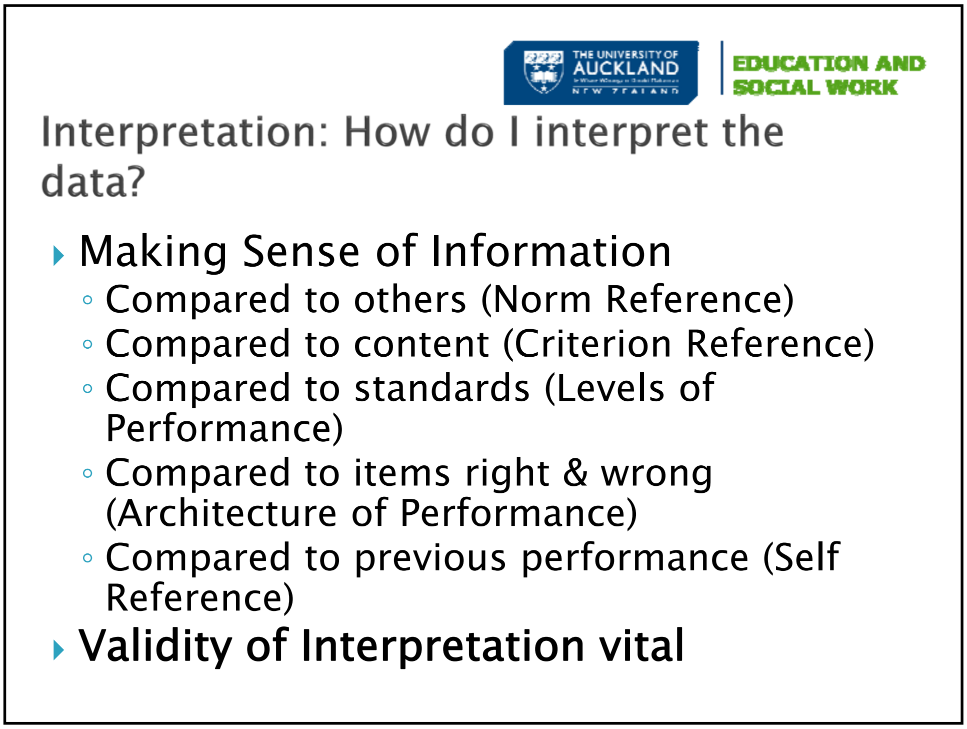


CTT and IRT Test Scores Compared

Item Difficulty	1 -3	2 -2	3 -1	4 -1	5 0	6 0	7 1	8 1	9 2	10 3	% correct	asTTle v4
A	✓	✓	✓	✓	✓	✓					60	530
B		✓	✓	✓	✓	✓		✓			60	545
C			✓		✓		✓	✓	✓	✓	60	593

Conclusions: C > A, B; B ≈ A because C answered all the hardest items correctly—no penalty for skipping or getting easy items wrong

The statistical model matters. We must remove junk before calculating score.

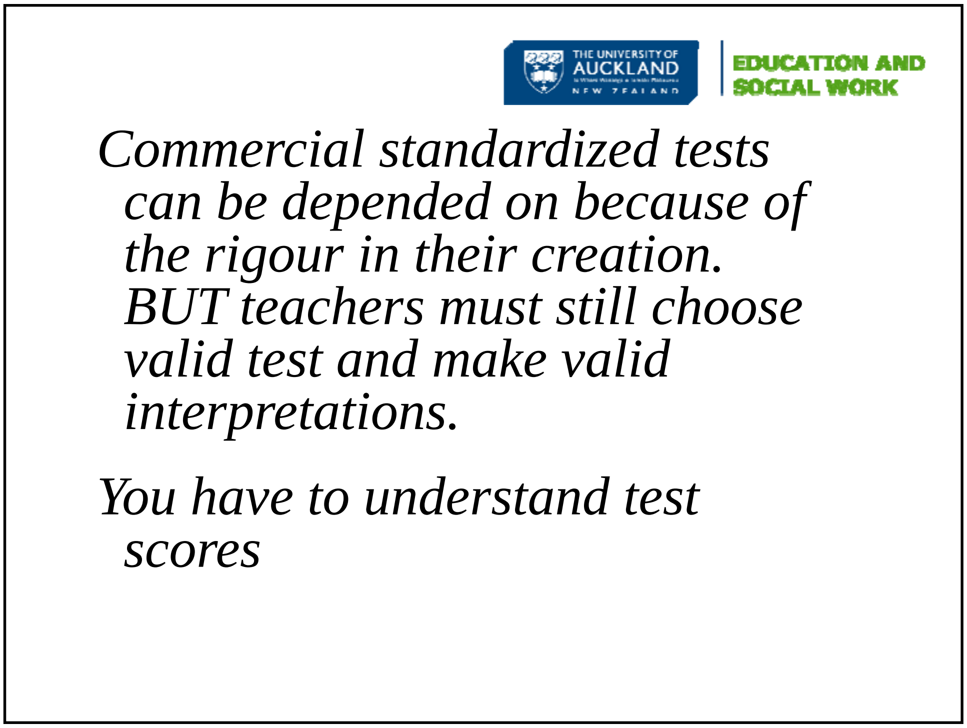


THE UNIVERSITY OF AUCKLAND NEW ZEALAND

EDUCATION AND SOCIAL WORK

Interpretation: How do I interpret the data?

- Making Sense of Information
  - Compared to others (Norm Reference)
  - Compared to content (Criterion Reference)
  - Compared to standards (Levels of Performance)
  - Compared to items right & wrong (Architecture of Performance)
  - Compared to previous performance (Self Reference)
- **Validity of Interpretation vital**

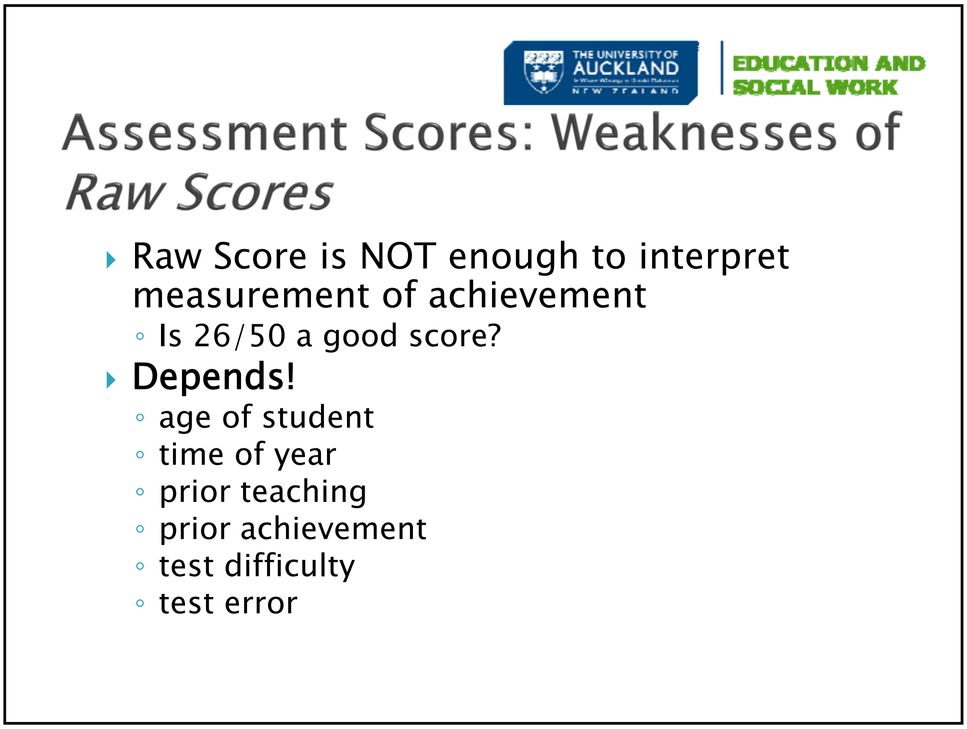


THE UNIVERSITY OF AUCKLAND NEW ZEALAND

EDUCATION AND SOCIAL WORK

Commercial standardized tests can be depended on because of the rigour in their creation. BUT teachers must still choose valid test and make valid interpretations.

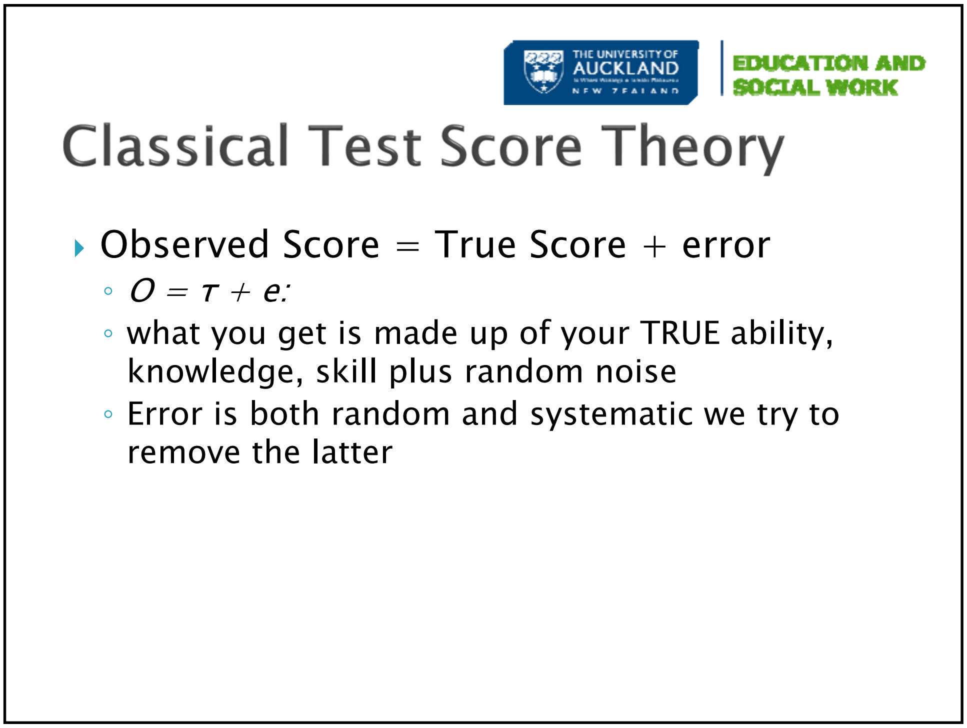
You have to understand test scores



THE UNIVERSITY OF AUCKLAND NEW ZEALAND | EDUCATION AND SOCIAL WORK

Assessment Scores: Weaknesses of *Raw Scores*

- Raw Score is NOT enough to interpret measurement of achievement
  - Is 26/50 a good score?
- **Depends!**
  - age of student
  - time of year
  - prior teaching
  - prior achievement
  - test difficulty
  - test error



THE UNIVERSITY OF AUCKLAND NEW ZEALAND | EDUCATION AND SOCIAL WORK

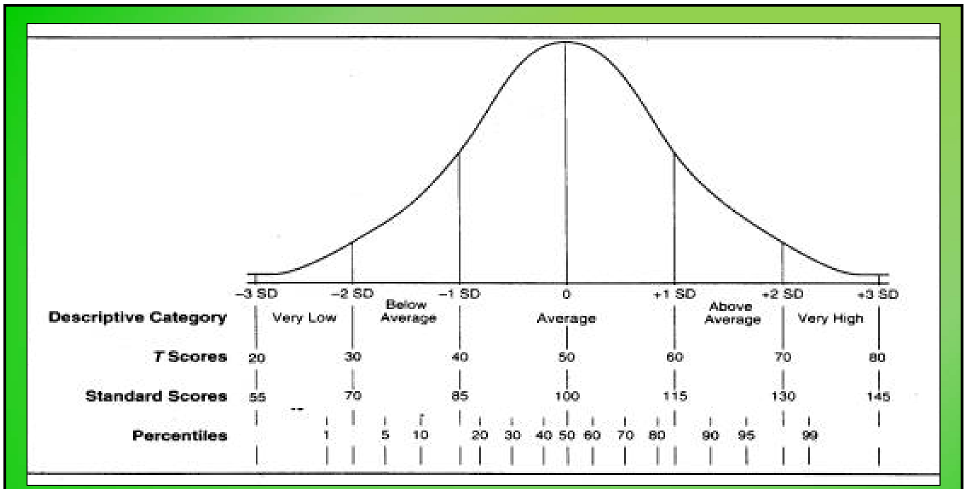
Classical Test Score Theory

- Observed Score = True Score + error
  - $O = \tau + e$:
  - what you get is made up of your TRUE ability, knowledge, skill plus random noise
  - Error is both random and systematic we try to remove the latter

THE UNIVERSITY OF AUCKLAND NEW ZEALAND | EDUCATION AND SOCIAL WORK

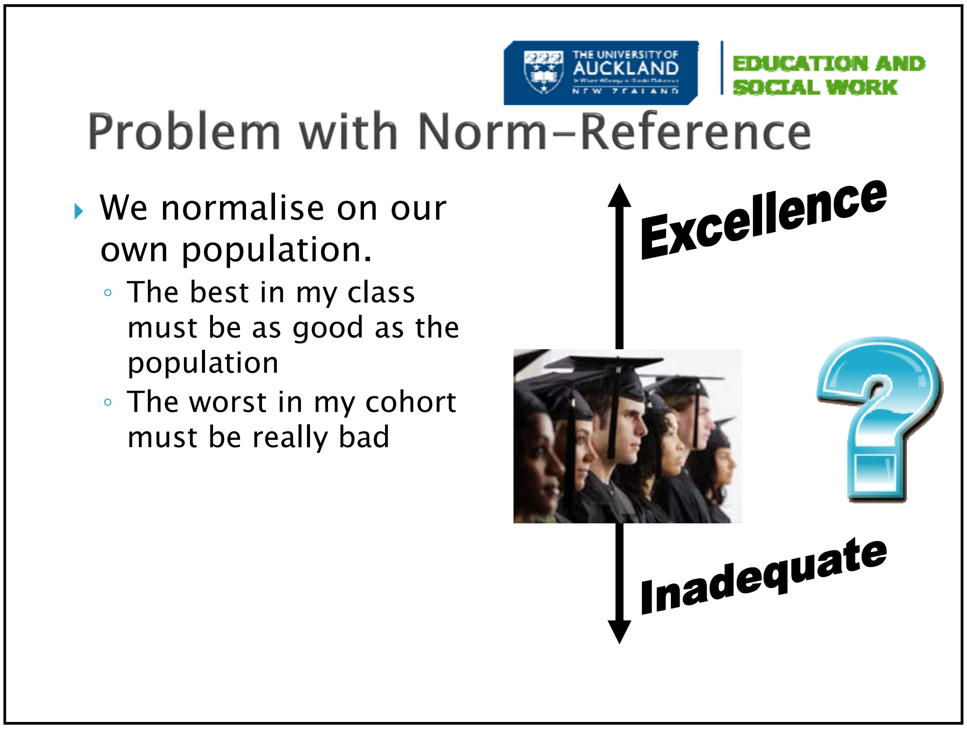
Rank Order: Norm-Referenced Interpretation

- **Well established**: Instructors can order people by competence, ability, performance
- **Assumption**: Position is relatively stable;
- **Consequence**: Rank resists instruction, so why bother?



Norm-referenced scores »

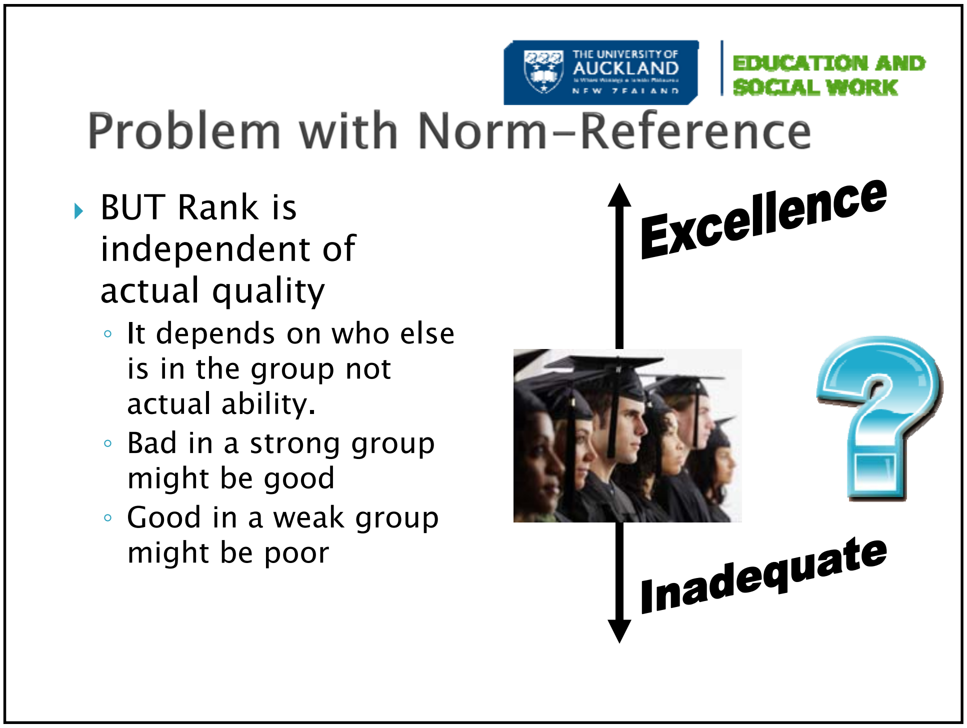
THE UNIVERSITY OF AUCKLAND NEW ZEALAND | EDUCATION AND SOCIAL WORK



Problem with Norm-Reference

- We normalise on our own population.
  - The best in my class must be as good as the population
  - The worst in my cohort must be really bad

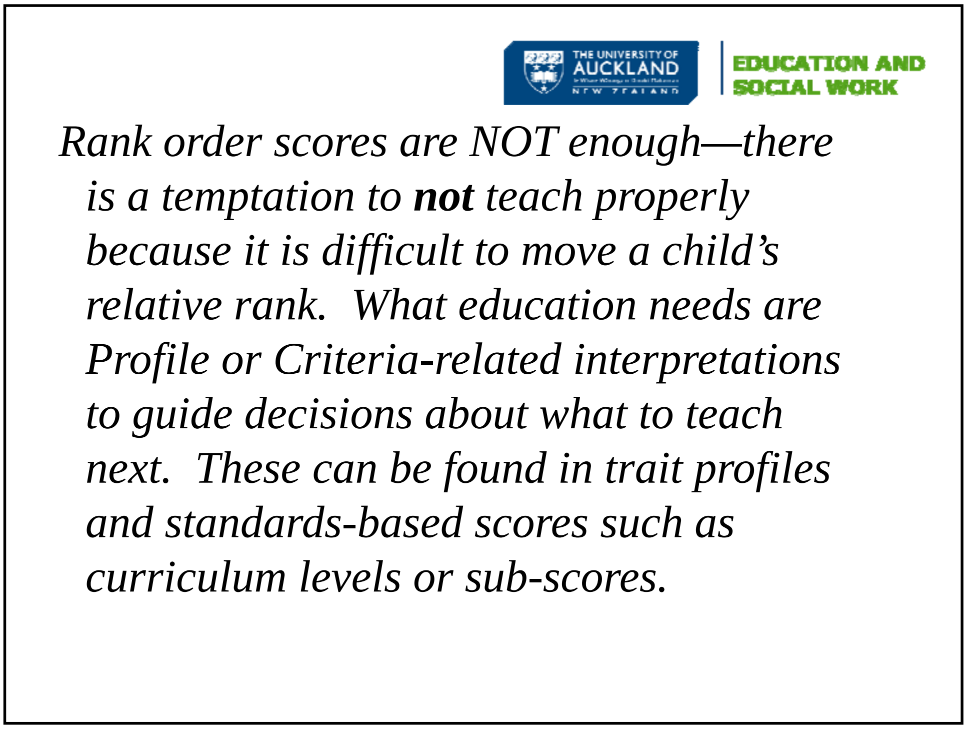
Excellence
Inadequate



Problem with Norm-Reference

- BUT Rank is independent of actual quality
  - It depends on who else is in the group not actual ability.
  - Bad in a strong group might be good
  - Good in a weak group might be poor

Excellence
Inadequate



Rank order scores are NOT enough—there is a temptation to ***not*** *teach properly because it is difficult to move a child's relative rank. What education needs are Profile or Criteria-related interpretations to guide decisions about what to teach next. These can be found in trait profiles and standards-based scores such as curriculum levels or sub-scores.*