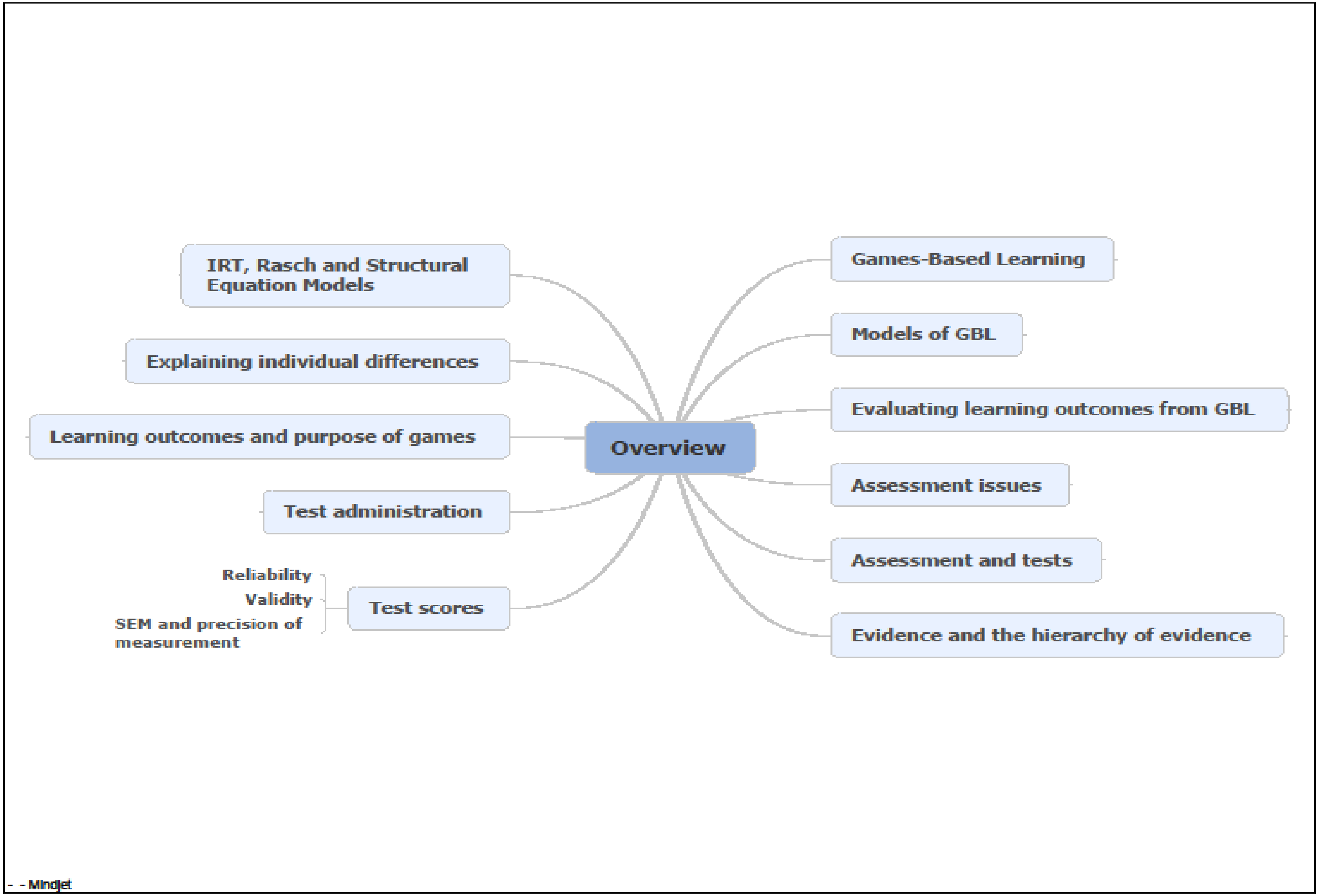
University of Strathclyde
Humanities & Social Sciences

Beyond the metrics of engagement:

Evaluation of learning outcomes from serious computer games

James Boyle
School of Psychological Sciences and Health
University of Strathclyde



Overview

- IRT, Rasch and Structural Equation Models
- Explaining individual differences
- Learning outcomes and purpose of games
- Test administration
- Test scores
  - Reliability
  - Validity
  - SEM and precision of measurement
- Games-Based Learning
- Models of GBL
- Evaluating learning outcomes from GBL
- Assessment issues
- Assessment and tests
- Evidence and the hierarchy of evidence

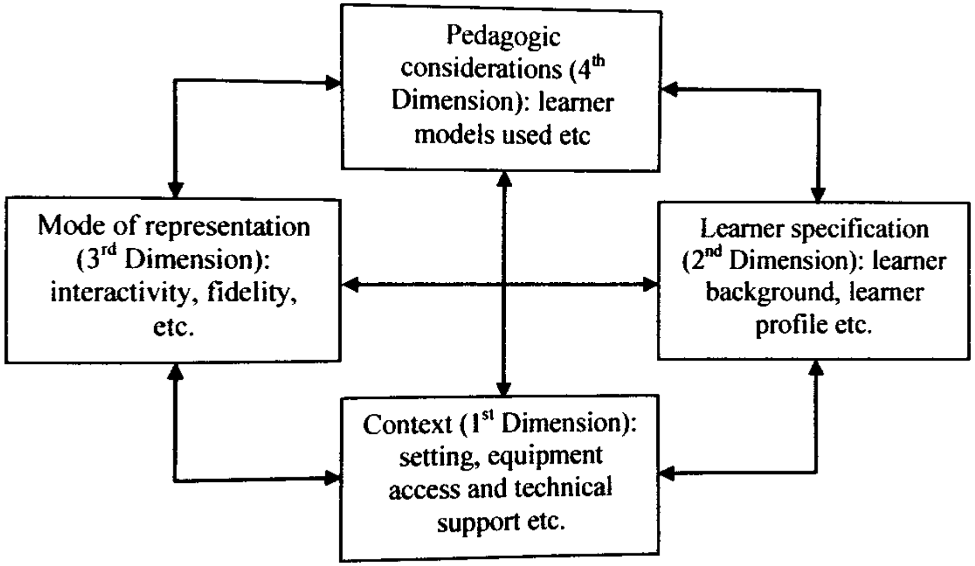
- - Mindjet

Games-Based Learning Promotes…

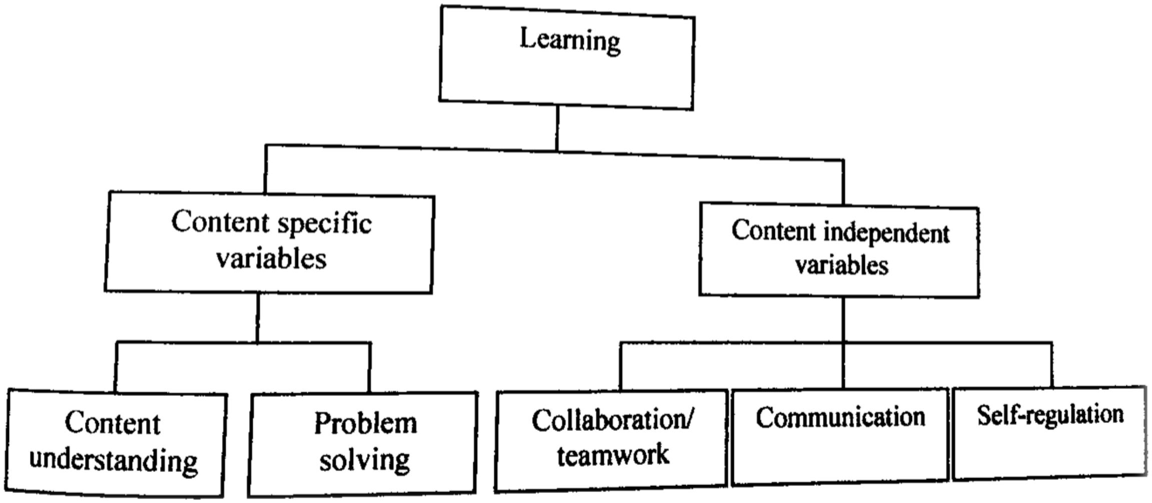
- Active learning
- Experiential learning
- Situated learning
- Problem-Based learning

via rules, constraints, challenge and feedback…

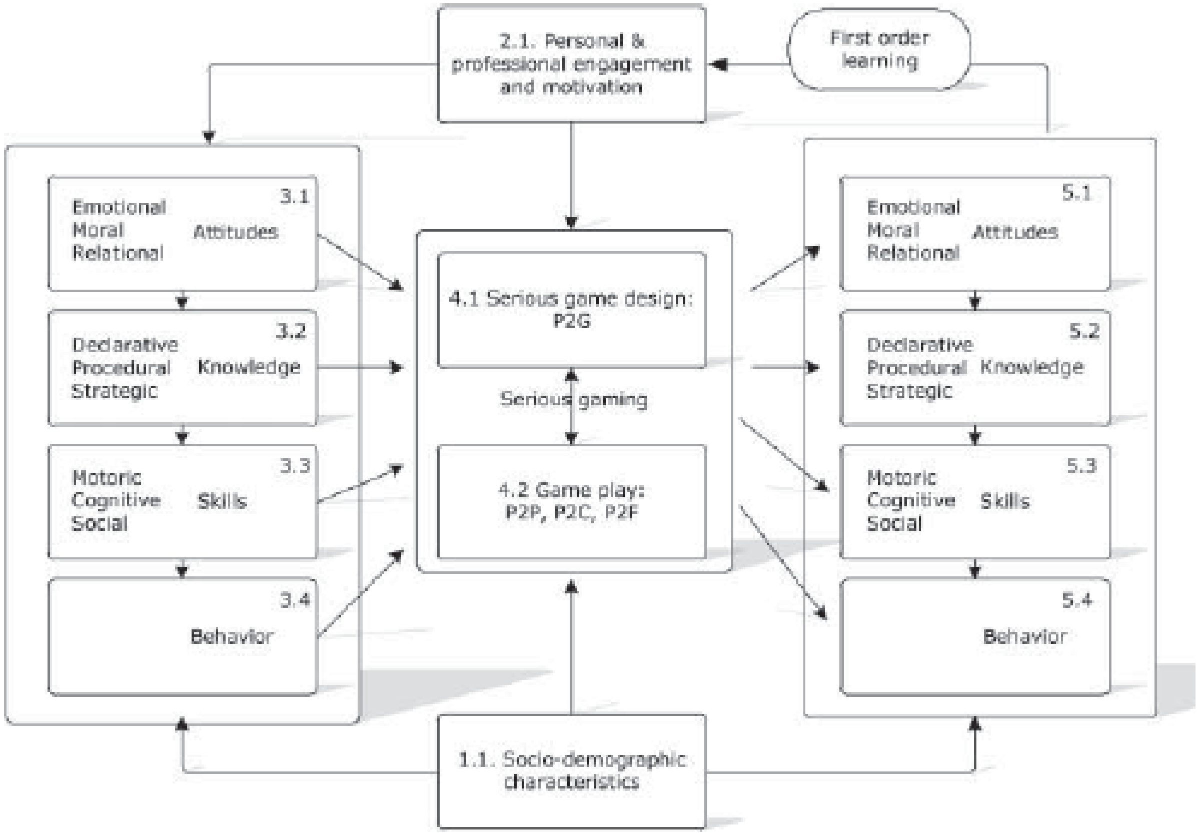
4 Dimensional Model (de Freitas & Oliver, 2006)


Pedagogic considerations (4th Dimension): learner models used etc
Mode of representation (3rd Dimension): interactivity, fidelity, etc.
Learner specification (2nd Dimension): learner background, learner profile etc.
Context (1st Dimension): setting, equipment access and technical support etc.

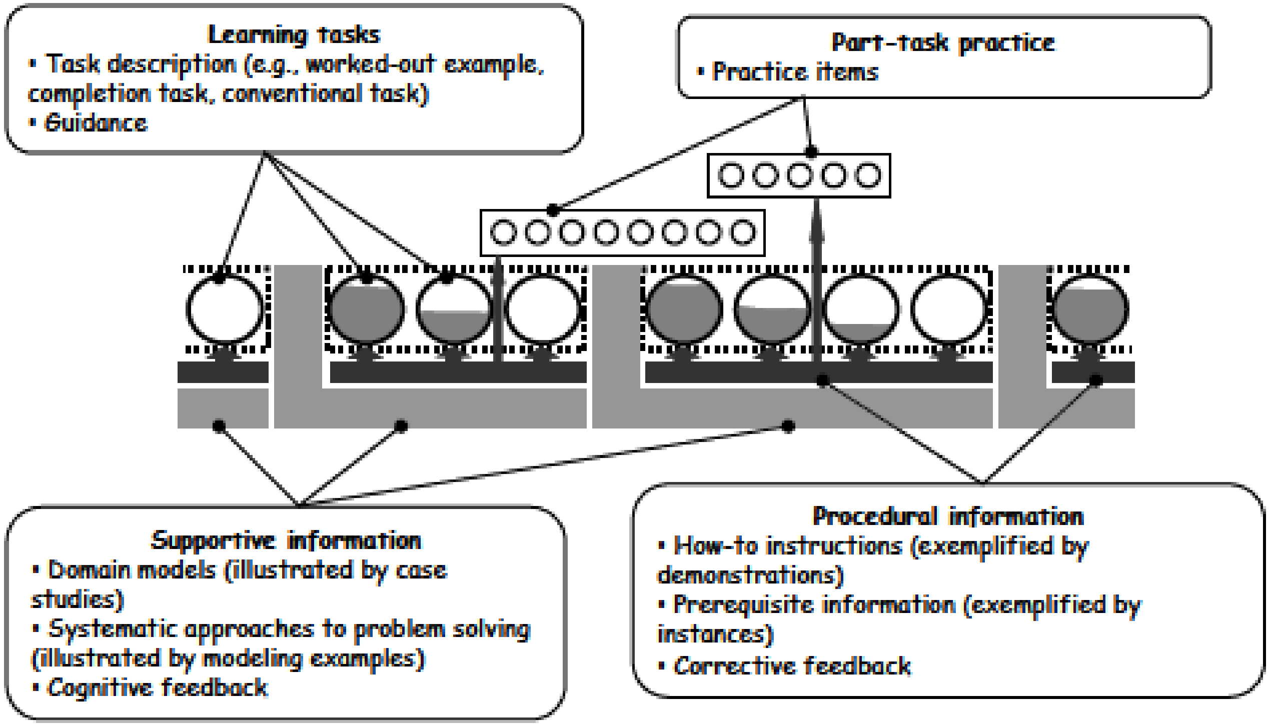
CRESST Model (Bayer & Mayer, 1999)



Comparative model for games-based learning: (Mayer et al., 2013)



Four-Component Instructional Design Model (van Merrienboer & Kester, *in press*)



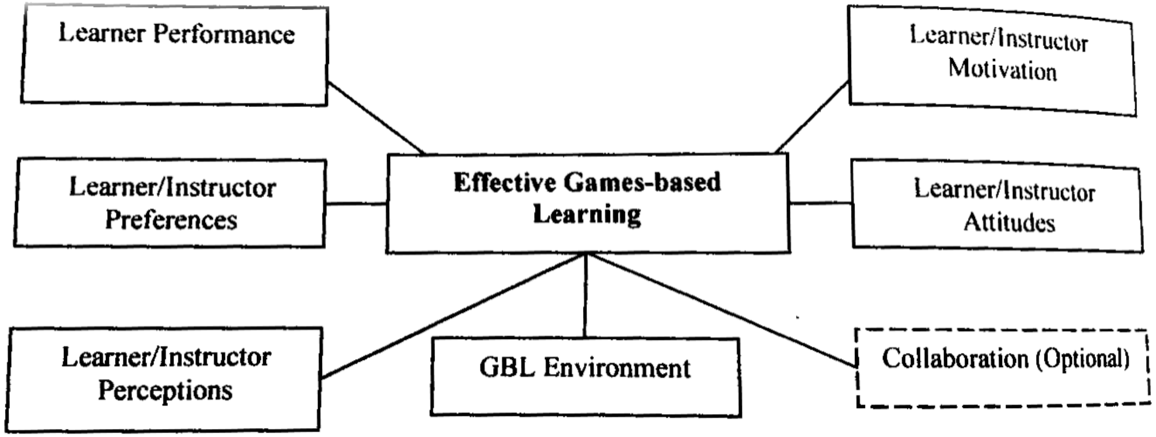
Learning tasks
- Task description (e.g., worked-out example, completion task, conventional task)
- Guidance

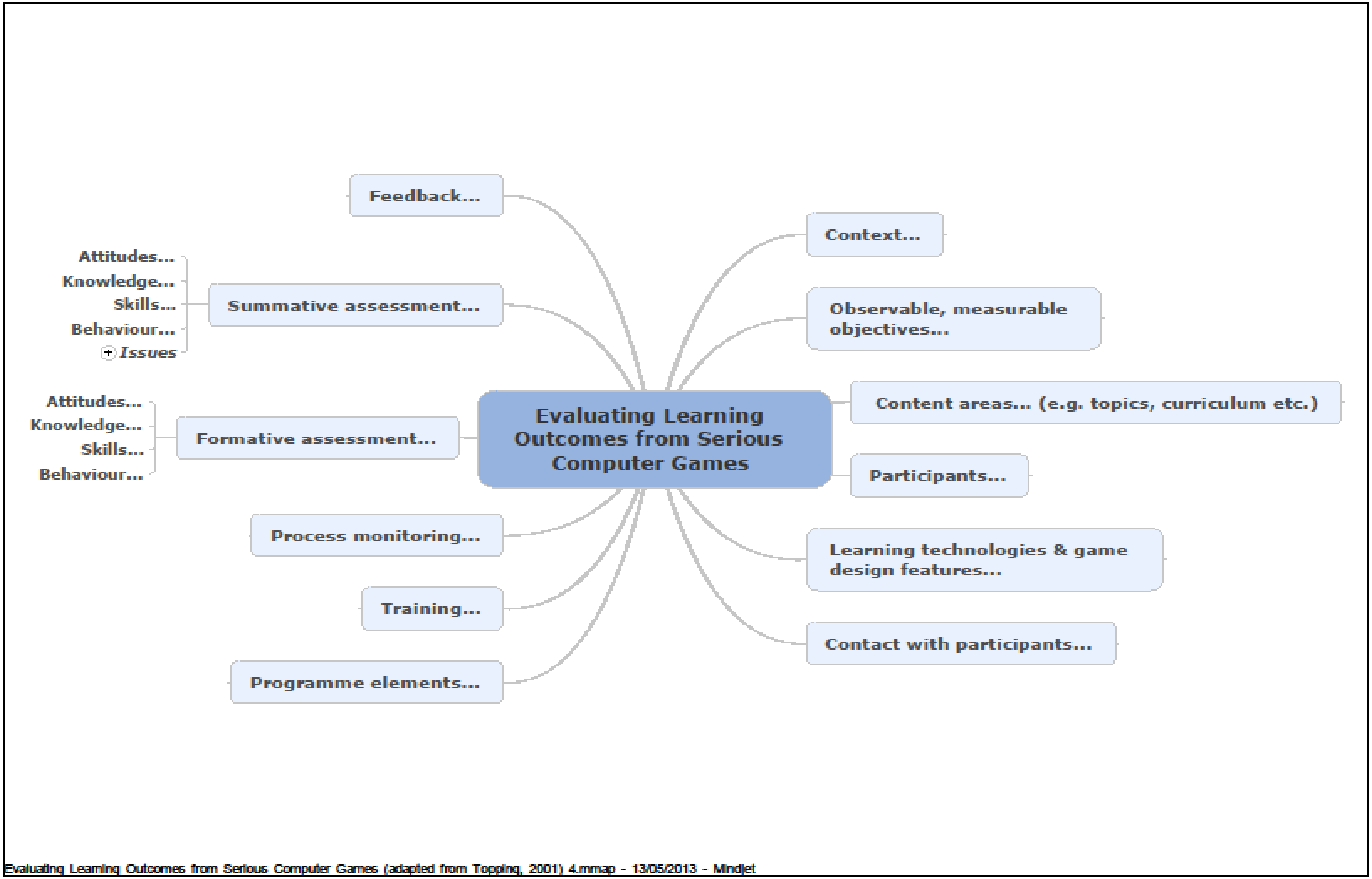
Part-task practice
- Practice items

Supportive information
- Domain models (illustrated by case studies)
- Systematic approaches to problem solving (illustrated by modeling examples)
- Cognitive feedback

Procedural information
- How-to instructions (exemplified by demonstrations)
- Prerequisite information (exemplified by instances)
- Corrective feedback

Effective Games-Based Learning Model (Connolly, Stansfield & Hainey, 2009)


Learner Performance
Learner/Instructor Motivation
Learner/Instructor Preferences
Effective Games-based Learning
Learner/Instructor Attitudes
Learner/Instructor Perceptions
GBL Environment
Collaboration (Optional)



Evaluating Learning Outcomes from Serious Computer Games

Feedback…

Summative assessment…

- Attitudes…
- Knowledge…
- Skills…
- Behaviour…
- *Issues*

Formative assessment…

- Attitudes…
- Knowledge…
- Skills…
- Behaviour…

Process monitoring…

Training…

Programme elements…

Context…

Observable, measurable objectives…

Content areas… (e.g. topics, curriculum etc.)

Participants…

Learning technologies & game design features…

Contact with participants…

Evaluating Learning Outcomes from Serious Computer Games (adapted from Topping, 2001) 4.mmap - 13/05/2013 - Mindjet

Assessment (Boyle & Fisher, 2007)

- Helps us to understand student performance (e.g. strengths & weaknesses, motivation and engagement, strategies used) and optimise learning experience
- Helps us to audit and evaluate programmes and teaching
- Provides researchers and instructors with data to feedback to users and stakeholders to inform decision-making

Underpins evidence-based practice…

…and Tests

- A form of systematic assessment, with standardized procedures, from which numerical scores are obtained
- Test use is generally quicker and simpler than other forms of assessment such as interviewing or observation
- The individual test taker is the principal source of data, Test use compares the test taker with others, whereas assessment in general may also consider the uniqueness of the individual, group or context
- Test use is highly structured, with objective procedures, whereas assessment more generally may be less structured and more subjective

Types of Tests (Rust & Golombok, 2000)

- *Norm-referenced* versus *criterion-referenced*…

- *Knowledge-based* (tests of maximum performance measuring ability, attainment, achievement and aptitude) versus *person-based* (tests of typical performance measuring personality, attitudes, moods, motivation, self-esteem, social competence)

- *Objective* (with pre-specified scoring criteria) versus *open-ended* (e.g. essays, projective tests)

Salvia & Ysseldyke (2004): Making sense of test scores

- Performance Standards (basis for comparisons): *comparative* between a learner and others or *absolute* between test-retest for an individual
- Informational Context: performance must be related to other information gathered: ***importance of multiple sources of information***

Assessment & Test Use: How trustworthy you regard ‘evidence’ to be depends on

1. What kind of scientist you are (applied/research)…

2. What questions you need answered…

3. What evidence might be available…

4. Who the stakeholders are…

5. *Beliefs, values, knowledge & impact…* the problem of ***‘reflexivity***’, the impact upon our own behaviour and of our theories about the world and how other people behave

6. ***Your model of science or epistemology…***

Naturalistic & Interpretative

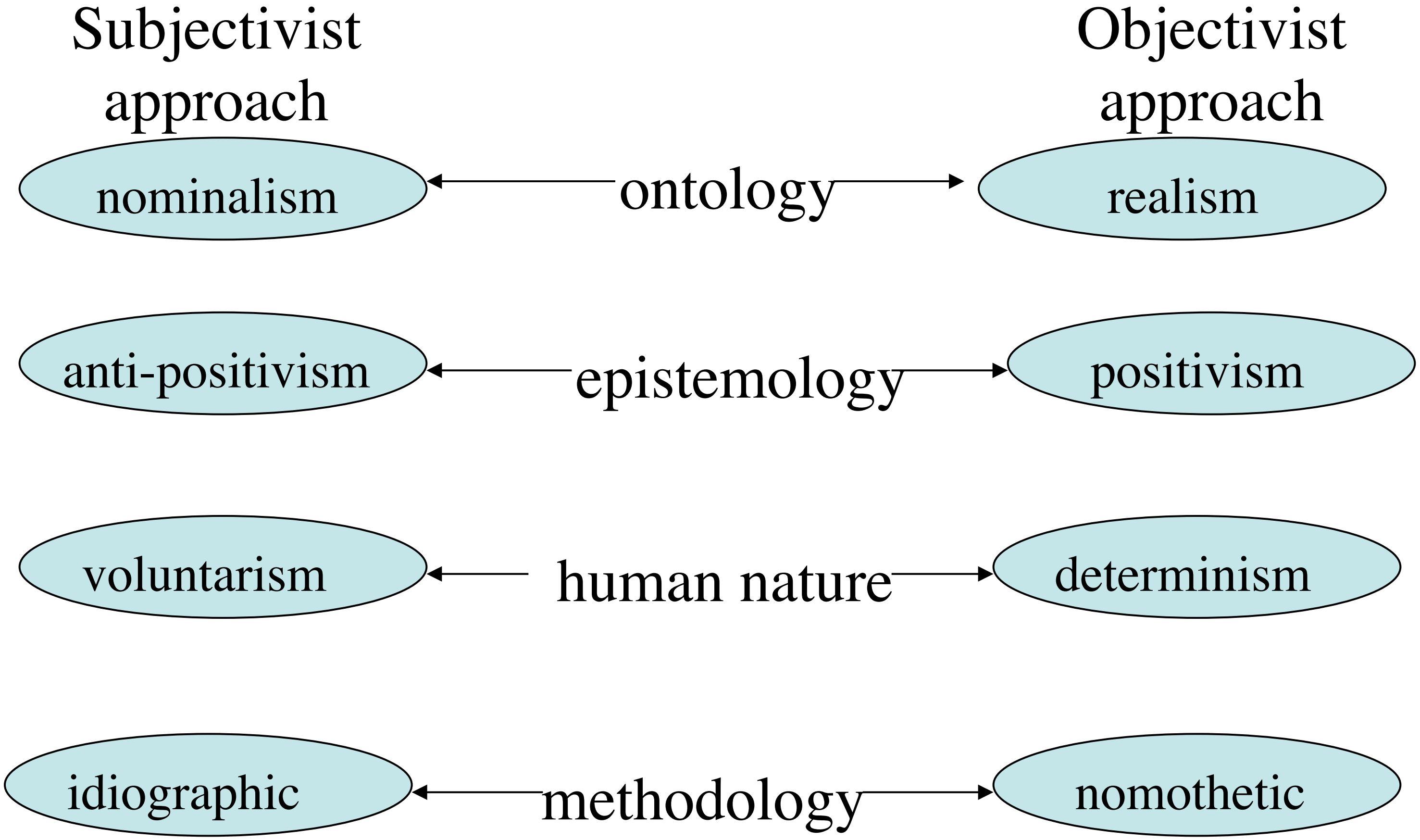
- ‘Controls in’ context
- May be problems in generalizing
- Interactions between investigator & subjects accepted
- ‘Action Research’ qualitative approach but may involve quantitative methods

Scientific & Positivist

- ‘Controls out’ effects of context
- Deals in generalizable data
- Investigator & subjects are independent
- Quantitative methods are used
- Deductive approach (a priori theory and hypothesis-testing)

Multiple methods:
Reduction of inappropriate uncertainty; triangulation and cross-validation enhance interpretability of results; complementary process model to investigate threats to validity, such as differential attrition

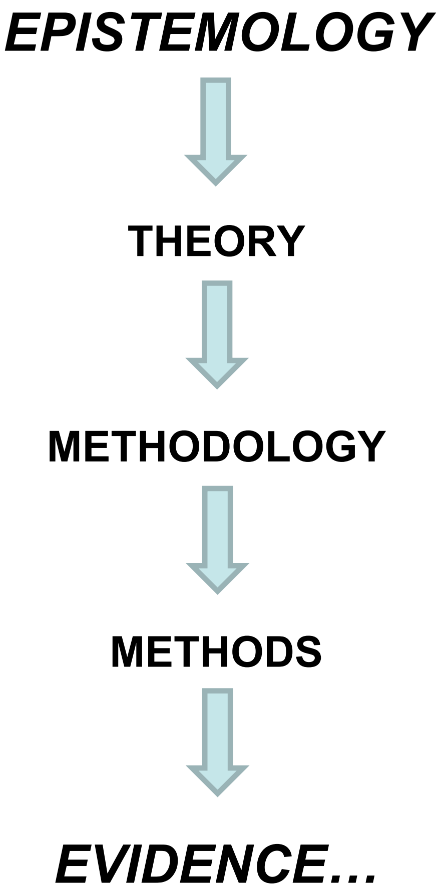
Opposing Views of Nature of Enquiry in Science…



Subjectivist approach		Objectivist approach
nominalism	← ontology →	realism
anti-positivism	← epistemology →	positivism
voluntarism	← human nature →	determinism
idiographic	← methodology →	nomothetic

(Burrell & Morgan, 1979)

Crotty (1998): Underpinnings of ‘evidence’



EPISTEMOLOGY

THEORY

METHODOLOGY

METHODS

EVIDENCE…

Hierarchy of Evidence *(Sackett et al., 2000)* Dealing with 'Threats to Validity':

1.	A	Systematic reviews/ meta-analyses of RCTs
	B	Single RCTs
	C	Experimental designs
2.	A	Cohort control studies
	B	Case-control studies
3.	A	Consensus conference
	B	Expert opinion
	C	Observational study
	D	Other types of study eg. Interview based, local audit
	E	Quasi-experimental, qualitative design
4.		Personal communication

Test Scores as Dependent Variables…

- Reliability
- Validity
- Adequacy of Standardisation
- Date of Standardisation
- *Standard Scores are interval scales and are easy to interpret*

Reliability

- **Reliability**: how accurate or sensitive a test is

- A reliability coefficient indicates the proportion of variability in a set of scores that reflects true differences among individuals

- The reliability indicates how much error is to be expected too.

- Treated like correlations – the higher the reliability the better (up to +1.0). Values above 0.7 reflect acceptable levels of reliability.

Three Main Types of Reliability

- Test-retest reliability
- Inter-rater reliability
- Internal consistency reliability

Test-re-test reliability

- An index of stability. The test is administered to a large number of participants (at least 100) and re-administered usually two weeks later. The correlation between the two sets of scores is the test-re-test reliability coefficient.

- Values of 0.8 or more essential, 0.9 for ability tests. Note also that the more time that passes between the administration of the two forms the greater the likelihood of change in true scores, plus there may be problems of attrition.

- Note also that effects of maturation and of practice may inflate the size of the coefficient.

Inter-rater reliability

- **Inter-rater reliability:** allows generalisation to different scorers. Two testers score a set of tests independently. The correlation between their scores for each set is the reliability coefficient for scorers. Should be 0.80 or better. (A second approach is to calculate the percentage of point-to-point agreement)

Internal consistency reliability: allows generalisation to different forms of the test

- **Same/alternate/parallel form reliability:** where two equivalent forms of the test are developed (A and B) and a large sample of participants receive both forms, in counterbalanced order.

- Scores from forms A and B are correlated and this is the same/alternate form reliability coefficient. A coefficient of 0.9 or better is required. (Note that the more time that passes between the administration of the two forms the greater the likelihood of change in true scores.)

- **Split-half reliability**: which requires only one form of the test. The test is administered to a group of participants and after administration, two alternate forms of the test created (e.g. odd versus even items), each containing half of the items. The correlation between the two ‘halves’ is called the split-half reliability coefficient.

- Cronbach’s (1951) Coefficient alpha can be used to provide the average split-half correlation based on all possible divisions of a test into two parts. Note if inter-item correlations are too high, some of the items might be redundant.

Validity

- Does a test measure what we want it to measure?

- We must be able to demonstrate that the test actually measures what we think it measures.

Seven types of Validity

- **Face validity**: if test appears to be measuring what it claims to measure

- **Faith validity:** belief that a test is valid (based on no specific criteria)

- **Content validity**: the extent to which the test's items actually represent the domain or universe to be measured (e.g. in the case of tests of ability and attainment, where the items are judged by experts in the field to be suitable for their purpose). Associated with the appropriateness of the items used in the test, are the completeness of the item sample in relation to the item universe, and the way in which the items assess the content. Need therefore for developers to conceptualise the content of a test precisely.

- **Consequential validity:** addresses the intended and unintended consequences of test interpretation and use

- **Construct validity:** the extent to which a procedure or test measures a theoretical trait or characteristic (e.g. if the test results fit hypotheses concerning the nature of the test variable)

- **Criterion related validity:** the extent to which a person's performance on a criterion measure can be estimated from the person's performance on the test being validated.
  - ✓ May be: **concurrent validity** (e.g. if test can be shown to correlate highly with another test of the same variable administered at the same time) or **predictive validity** (e.g. if the test score will predict performance at a later time).

- **Differential validity:** the test's ability to predict one criterion better than another

Construct validity is of central importance…

- Construct validity tells us about the extent to which a procedure or test measures a key theoretical concept or a construct from factor analysis
- If two tests measuring the same constructs have good construct validity then we would expect scores from the two tests to be strongly correlated (‘convergent validity’)
- Conversely, scores from two tests measuring different constructs should not be highly correlated (‘divergent validity’)
  - ✓ Important therefore in test development to correlate scores on the test with those from a benchmark test
- Differential scores which are a function of age or some other relevant group variable can also provide evidence for construct validity

Modes of Test Administration and Serious Games

- **Open:** Open tests can be accessed freely over the internet. There is no requirement to identify the test taker, and they can be completed without any supervision by the test administrator.

- **Controlled:** In controlled tests, test takers are required to register before they can complete the test, however, overall supervision is not necessary.

- **Supervised:** In supervised tests, test takers are required to register and confirm their identity. A degree of overall supervision and support is necessary.

- **Managed:** In managed modes of assessment, the test administrator controls the testing environment: he/she introduces and administers the test, and is available for giving any necessary support and to answer any appropriate questions.

Assessment Strategies…

- *Norm-referenced*: use when you wish to compare the test-taker's knowledge base with normative standards

- *Criterion-referenced*: use when you wish to assess the content of a test-taker's knowledge base within a specific domain (i.e. when you want to measure mastery of information or skills: what the test-taker can or can't do in terms of absolute standards).

- *Curriculum-based*: use when you wish to assess the test-taker's instructional needs based upon his/her on-going performance on the existing content of curriculum.

- *Dynamic*: use when you wish to find out more about the appropriateness of the strategies used by the test-taker, how they react to suggestions made by an instructor, and identify barriers to learning.

So how trustworthy are test scores?

- Different sources of error can lead a test to be unreliable:
  - ✓ Measurement error
  - ✓ Scoring error
  - ✓ Situational factors (e.g. fire alarm)
  - ✓ Item sampling

- Reliability estimates are sample specific and can change with variability. This has implications for how reliability estimates are interpreted

- Tests can be reliable but not very good an discriminating at the top and bottom of the score range.

- NB: A test can be reliable but not accurate (valid). Just because we always get a similar result, does not mean it is a good measure of a construct.

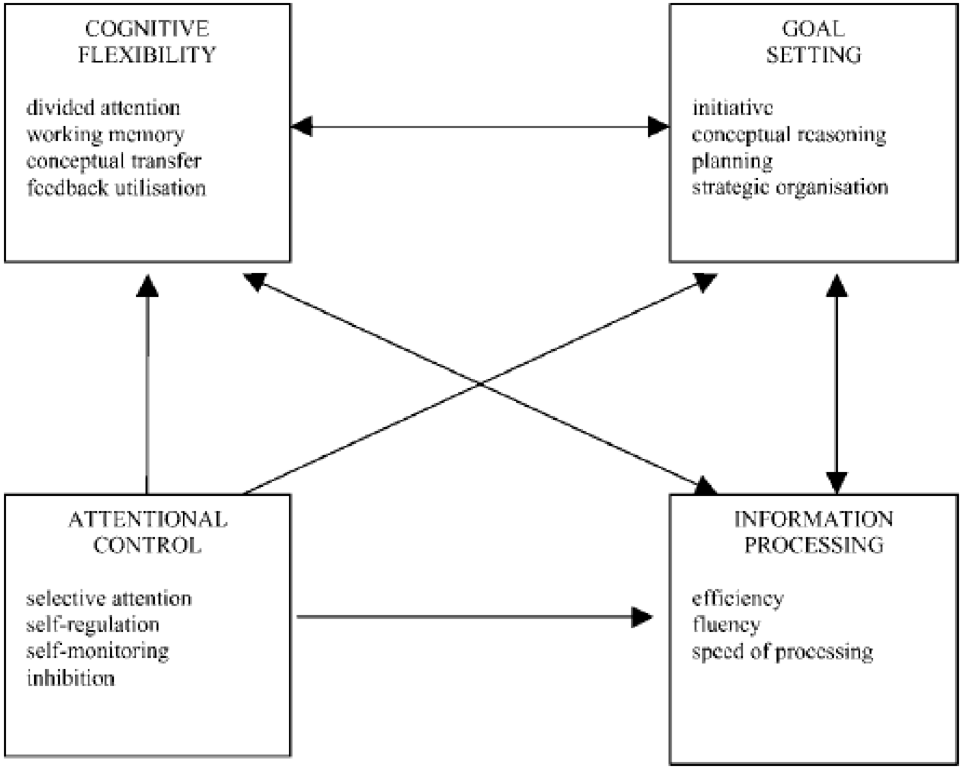
SEM and precision of measurement

- All tests are subject to measurement error, but we can calculate Confidence Intervals (CI) to account for this error using the **Standard Error of Measurement (SEM)** to determine the range of scores within which the 'true' score falls, given the obtained score, at various degrees of probability.
- The SEM can also be used to set up confidence limits to evaluate the significance of re-test scores.
- This tells us whether we have ***reliable change****: this is different from statistical significance (e.g. in an experiment or meta-analysis)*
- The SEM is calculated using the following expression:
  - $SEM = SD \times \sqrt{1-r} \times \sqrt{2}$
  - where SD is the standard deviation and r is the reliability estimate from the test manual

Primary purpose of game by learning and behavioural outcome (Boyle et al., 2012)

Outcomes of playing game	Entertainment game	Game for learning	Serious game	Total
Affective and motivational outcomes	26 (14)	5 (4)	2 (0)	**33 (18)**
Behaviour change	7 (3)	4 (3)	2 (2)	**13 (8)**
Knowledge acquisition/content understanding	3 (3)	26 (12)	3 (2)	**32 (17)**
Motor skills	1 (0)	5 (2)	2 (2)	**8 (4)**
Perceptual and cognitive skills	13 (9)	7 (4)	0 (0)	**20 (13)**
Physiological outcomes	11 (6)	0 (0)	0 (0)	**11 (6)**
Social/soft skill outcomes	6 (1)	2 (1)	3 (2)	**11 (4)**
Various	1 (0)	0 (0)	0 (0)	**1 (0)**
Grand Total	**68 (36)**	**49 (26)**	**12 (8)**	**129 (70)**

Explaining Individual Differences: Anderson's (2002) Model of Executive Functions



Explaining Individual Differences…

- The attentional control components include *selective attention* (the ability to select specific stimuli to attend to and to concentrate for sustained periods of time), *self-regulation* and *self-monitoring* (the ability to initiate, monitor and terminate processes) and *inhibition* (the ability to inhibit both irrelevant information and ‘prepotent’ responses which are not appropriate to the task in hand).

- Problems with these attentional control components will be associated with impulsivity, problems in task completion, poor error correction and inappropriate responding which would have an adverse impact upon performance in serious computer games.

Information-processing components

- The information-processing components include *fluency*, *efficiency* and *speed of processing.* These components reflect the individual's reaction times to stimuli as a function of efficiency of the brain and nervous system.

- Problems here in slow reaction times will also adversely affect performance in computer games.

Cognitive Flexibility

- The cognitive flexibility components include *divided attention* (the ability to split attentional processes and resources across different sources of information and responses at the same time), *working memory* (a limited-capacity and time-limited active memory system with the ability to store and simultaneously process information), *conceptual transfer* (the ability to transfer a solution from one problem to another and hence to develop alternative strategies) and *feedback utilisation* (the ability to adapt to and learn from feedback).

- Problems with these components will be associated with persisting with unsuccessful strategies and hence in problems in learning from experience.

Goal-Setting Components

- Goal-setting components include *initiative* (the ability to generate new ideas), *conceptual reasoning* (problem-solving abilities), *planning* (the ability to devise the implementation of actions) and *strategic organisation* (the ability to coordinate strategies).

- Problems with these components will be associated with the use of existing strategies as a result of poor problem-solving rather than generating more effective strategies.

IRT, Rasch Models and SEM

- Finally, Item Response Theory (IRT), Rasch Models and Structural Equation Modelling offer powerful tools for measuring learning outcomes from serious computer games
  - see Hung et al. (2012) Computers & Education 59 (2012) 762–773
  - IRT (or latent trait models) provides estimates of the relationship between a test-taker's response to an item and his/ her level of underlying ability on the latent variable
  - Rasch scaling is a form of IRT which transforms ordinal items to a logit scale which is an interval scale of measurement

Reeve (2002)

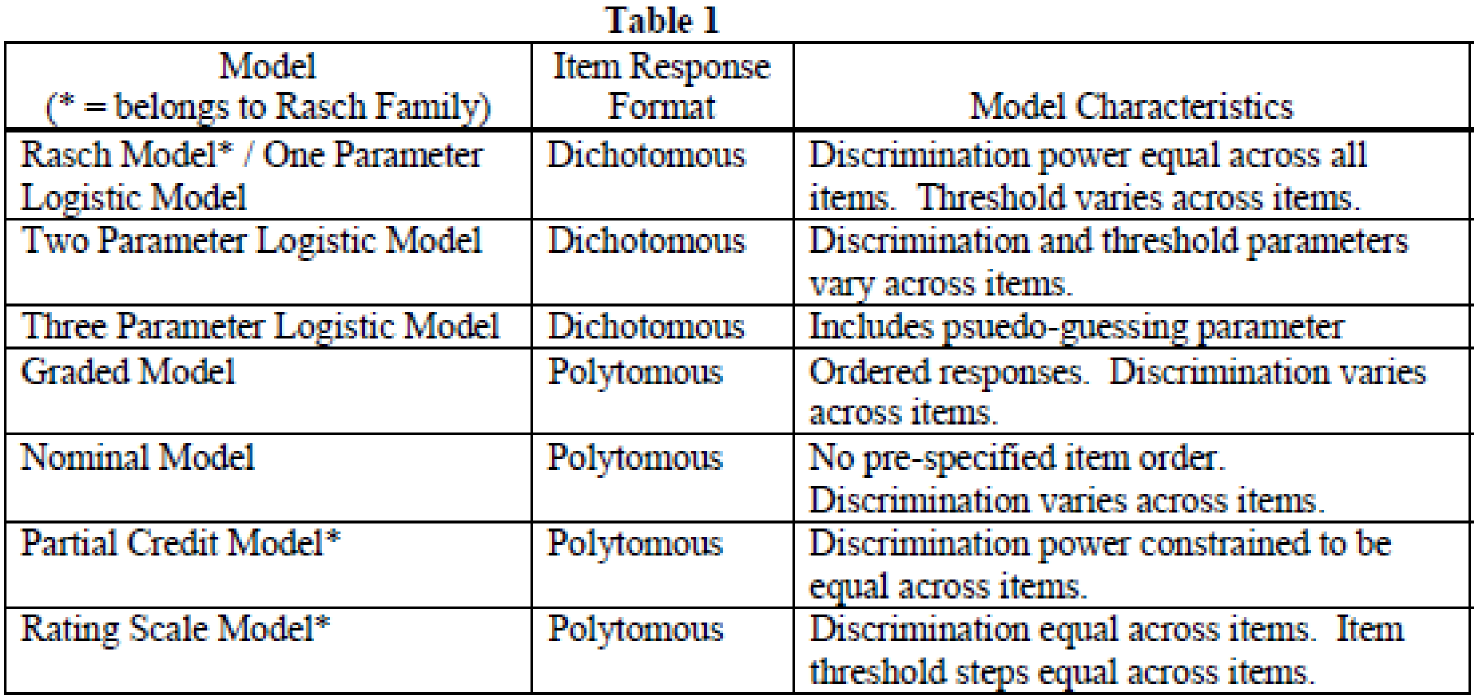


Table 1

Model (* = belongs to Rasch Family)	Item Response Format	Model Characteristics
Rasch Model* / One Parameter Logistic Model	Dichotomous	Discrimination power equal across all items. Threshold varies across items.
Two Parameter Logistic Model	Dichotomous	Discrimination and threshold parameters vary across items.
Three Parameter Logistic Model	Dichotomous	Includes psuedo-guessing parameter
Graded Model	Polytomous	Ordered responses. Discrimination varies across items.
Nominal Model	Polytomous	No pre-specified item order. Discrimination varies across items.
Partial Credit Model*	Polytomous	Discrimination power constrained to be equal across items.
Rating Scale Model*	Polytomous	Discrimination equal across items. Item threshold steps equal across items.