Standards Setting and Maintenance in TSA

Standards Setting of : Formation of Panels of Judges

- After the first year's administration of the TSA at each level (i.e. P.3 in 2004, P.6 in 2005 and S.3 in 2006)
- For each subject, **panels of judges** were established:
  - Each panel consisted experienced school teachers
  - Teachers came from a variety of school types and that schools of high, middle and low strata were equally represented
  - Curriculum Development Officers of the CDI
  - Subject Officers of the HKEAA

Standards Setting : Consensus of Expert Views

- For multiple-choice items and short answer questions, the **Angoff method** was used:
  - Estimated the probability of a minimally competent student getting each item correct
  - In the light of empirical evidence regarding actual performance levels, pooled the results, revised estimates and finally reached consensus on a cut score

- For questions that involved a holistic assessment of a single piece of work, the **Bookmark method** was used:
  - Each judge inserts a metaphorical 'bookmark' in the pile of scripts/performances to separate those deemed as meeting the standard and those not meeting the standard
  - pooled and a consensus judgment made about the final position of the 'bookmark'

Standards Setting : Final Consolidation

- Psychometric analysis was used to identify "unqualified" judges those of the lenient/harsh and/or inconsistent judges
- The ratings of judges were then pooled into a combined panel, excluding "unqualified" judges, to produce a final set
- Preliminary results were also benchmarked against international standards (as far as possible) to ensure that the standards set in Hong Kong are competitive with those of other regions

Standards Maintenance Across Years

- The current year's TSA test scores (20XX) were equated with that of the previous year (20XX – 1)
- Administer the same Research Test to a sample of students in both years:

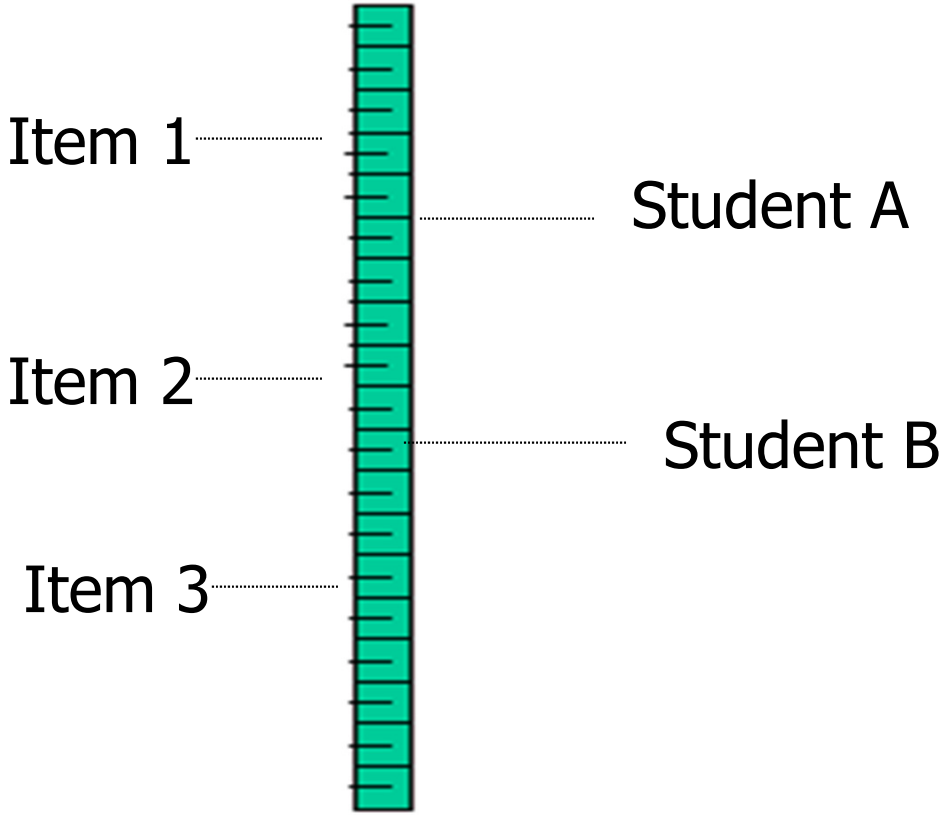
	Research Test	(20XX – 1) TSA	(20XX) TSA
(20XX – 1) Sample			
(20XX) Sample			

Note: Different shadings indicate different sets of items.

Rasch Modeling: Basics

- Each person is characterized by an ability index
- Each item is characterized by a difficulty index (Note: for a polytomous item; i.e., an item with full marks > 1, a set of difficulty indices is used instead)
- Both of them can be expressed by numbers along ONE single line
- The difference between these two numbers → The probability of observing a particular scored response

Same Ruler for Items and Students



Item Difficulties and Student Abilities

Student <----> Athlete

Item <----> Hurdle

Note:

Both athlete's ability and difficulty of the hurdle are measured in the same unit/ same ruler; i.e. the height

Pass/Fail: Jump over a specific height

IRT Analysis: Modeling Formula

$$\Pr\{x_{ni}\} = \frac{\exp[x_{ni}(\beta_n - \delta_i) - \sum_{k=0}^{x_{ni}} \tau_{ki}]}{\gamma_{ni}}$$

Student n with ability β_n
Item i with difficulty indices: δ_i and τ_{ki}
x_{ni} is the actual score obtained by Student n on Item i

Principles:
Given a set of student responses to a test $\{x_{ni}\}$
β_n, δ_i and τ_{ki} are estimated to be values, which maximize the probability (or likelihood) for obtaining the observed responses; i.e., Maximum Likelihood Estimation (MLE)

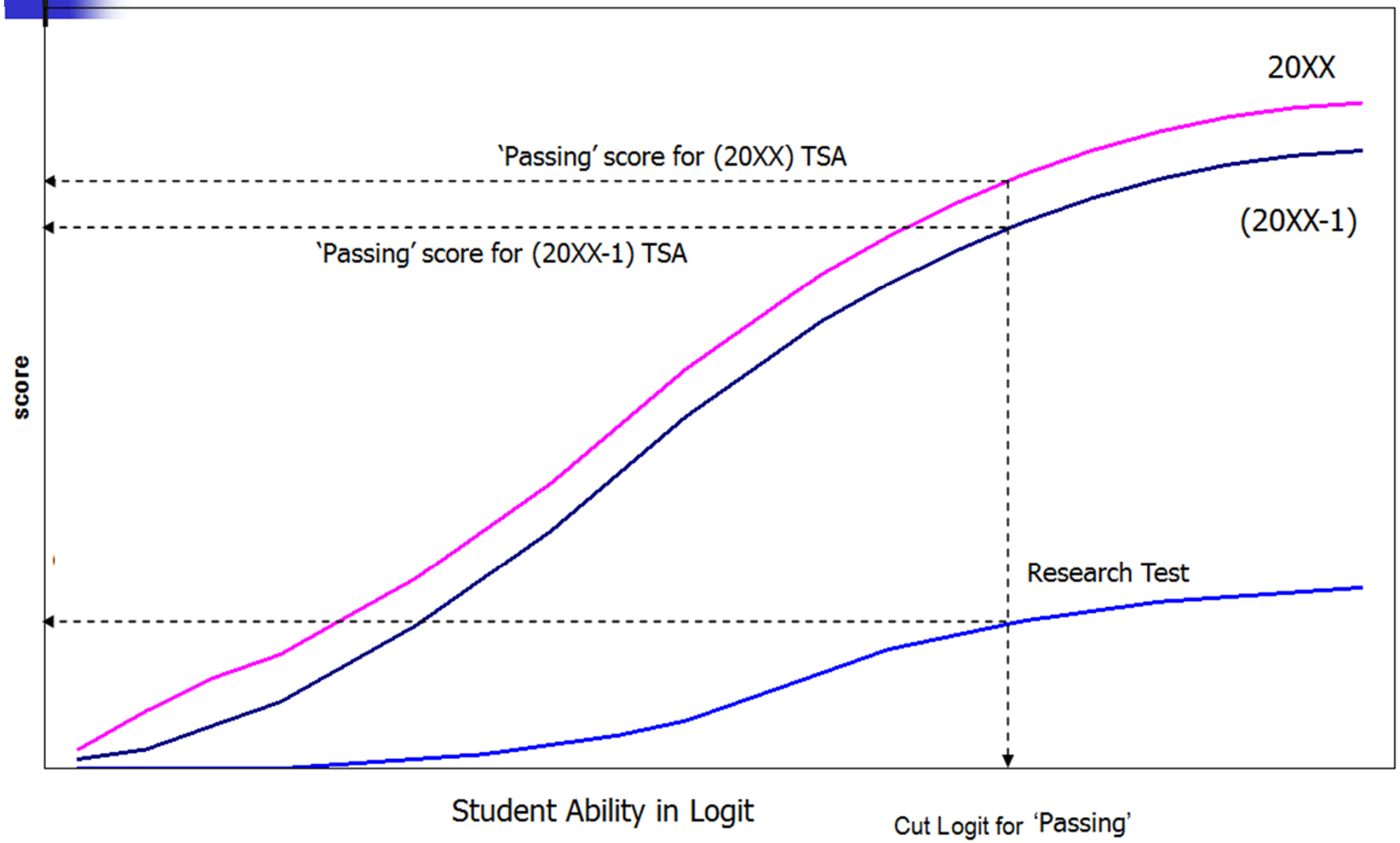
Test Equating using IRT

- Implement test equating:
- Based on exam results (x_{ni}), student abilities (β_n) and item difficulties (δ_i, and τ_{ki}) could be estimated
- With respect to a student with a specific ability (β), the expected mark of an item i for the student can be derived:

$$E(X_i) = \sum_{x'=1}^{m_i} x' P(X_i = x' | \beta)$$

- The expected of the whole subject for the student can be derived by accumulating his/her expected mark of each item

Graph: Test Equating for Standards Maintenance



Thank YOU!!!