
Lihua Yao
Monterey, CA
Lihua.Yao.civ@mail.mil
Presented on
January 23, 2015
Lisbon, Portugal

The views expressed are those of the authors and not necessarily those of the Department of Defense or the United States government
Summary

• Measurement of testing data.
• Computer adaptive testing (CAT).
• Multidimensional CAT.
• Research areas in CAT and MCAT.
• A recent study comparing MCAT ability based item selection method and classification based item selection method.
Measurement of testing data

• Paper and pencil test:
  - Classical method---number of correct score; items are similar.

  • Testing has a long history in China(> 2000 years).
  • Group-administered tests by United Stated Army in World War I.

• Item response theory models:
  • Probability, Parameters, data.
  • Estimate Parameters.
  • Inferences: What the Data tells us?
Measurement of testing data

- Paper and pencil test:
  - Item response theory models:
    - Multiple choice items:
      - Rasch (Rasch 1960, Fischer 1995),
      - Three parameter logistic model (Lord 1980), with guessing parameters.
    - Constructed response items:
      - Generalized two parameter partial credit model (Masters, 1982, Muraki, 1992),
      - Graded response model (Samejima, 1969).
- Online based testing.
Computer Adaptive Testing

- Computer adaptive testing (CAT)---future of testing industry.
  - It is a computer based test that adapts to examinee’s ability.
  - Test security is better than traditional paper and pencil.
    - Examinees have different test items.
  - Convenient testing time and location.
    - For example, Screening test, unproctored test.
    - Test can be taken any time at home.
Computer Adaptive Testing

- Computer adaptive testing (CAT)
  - Reliable measures of a student’s skills while minimizing testing time.

- Items are selected that measure the examinee’s ability the best.

- Shorter test with the same precision.
Models

- **M-3PL model** (multidimensional three-parameter logistic model) for multiple choice item j,
\[
P_{j1} = P_{j1}(\theta) = P(x_j = 1 \mid \theta, \beta_j) = \beta_{3j} + \frac{1 - \beta_{3j}}{1 + e^{(-\beta_{2j} \theta^T + \beta_{1j})}}
\]

- Information Function
\[
I_j(\theta) = -E \frac{\partial^2 \log P_j}{\partial \theta^2} = \frac{(P_{j1} - \beta_{3j})^2 (1 - P_{j1})}{P_{j1} (1 - \beta_{3j})^2} \beta_{2j} \otimes \beta_{2j}
\]

- **Rasch Model**:
\[
P_{j1}(\theta) = \frac{e^{(\theta - b_j)}}{1 + e^{(\theta - b_j)}}
\]

- Information:
\[
I_j(\theta) = P_{j1}(\theta)(1 - P_{j1}(\theta))
\]

- The rectangle has a maximum area when it is a square if the parameters of rectangle is fixed.
Computer Adaptive Testing

• Information has maximum value when $P_j(\theta) = \frac{1}{2}$, which occurs when $\theta = b_j$.

• Standard error for the ability estimate is $SE(\theta) = \frac{1}{\sqrt{I(\theta)}}$.

• Computer adaptive testing is to select items that have maximum information or minimum value for the SE to have a better ability estimates.
Computer Adaptive Testing

• Components in Computer adaptive testing.
  - Item Pool with known/estimated item parameters.
    • Multiple forms/pools.
  - Initial values for the ability.
    • Prior information about the students ability.
    • No prior information.
   Select items based on certain algorithm.
   Scoring procedure to update ability estimates.
  - Stopping criteria/rules.
Computer Adaptive Testing

- Tests using Computer adaptive testing.
  - ASVAB (the Armed Services Vocational Aptitude Battery)—1976, Paper and pencil format. Has been used for selection and classification into the military.
  - CAT ASVAB research and development started in January 1979. It is the first high-stakes testing program to produce operational scores using a CAT system. Local online CAT Operational in 1990 and online through internet in 2007.
  - Examples of tests using CAT:
    - GMAT, GRE, TOEFL, NCLEX(nursing), etc.
    - More and more states using CAT for summative and diagnostic assessment.
Multidimensional Computer Adaptive Testing

- Multidimensional CAT, and software SimuMCAT (Yao, 2011)
  - MCAT yields better precision than UCAT.
    - Correlation information between content areas are used.
    - Bayesian estimates for the multidimensional ability after each step. Strong prior improve both content domain scores and overall scores.
    - Incorporate content constraints, exposure control, item response time, precision, etc., simultaneously.
• Multidimensional CAT, and software SimuMCAT (Yao, 2011)

• Short response time is desirable and items in different content areas require different response time. For example, an Arithmetic Reasoning item may need 79 seconds, while a Word Knowledge item may need 14 seconds.
Multidimensional Computer Adaptive Testing

- Components/Steps MCAT:
  - Item pool.
  - Initial ability assignment.
  - Item selection based on certain rules (Yao 2012, Psychometrika)
  - Update ability estimates based on the selected items - MLE, MAP, EAP.
  - Stopping rules:
Multidimensional Computer Adaptive Testing

• MCAT:
  - Item selection based on certain rules (Yao 2012, Psychometrika)----information is a matrix.
    • Maximum Determinant of information matrix.
    • Minimum the error variances for the composite score.
    • Maximum Kullback–Leibler information.
    • Angle method.
    • General model.
    • Maximum reduction of the SEE (standard error estimates).
Multidimensional Computer Adaptive Testing

• MCAT
  - Stopping rules:
    • Fixed length and variable length test (Yao, 2013, APM) with different stopping rules.
    • Precision (standard error and predicted standard error reduction).
    • Classification as the stopping rule.
    • Certain criteria for content domain SEE or composite score SEE.
Multidimensional Computer Adaptive Testing

• MCAT
  - Content constraints
    • Upper limit and lower limit for each content area.
  - Exposure control
    • Sympson-Hetter (Yao, 2013, JEM)
    • Fixed rate.
  - Stopping rule, Content Constraints, and Item exposure control can be incorporated in to an index (MPI) which will be used as a weight assigned to each item in the pool at each selection steps.
Multidimensional Computer Adaptive Testing

- Multidimensional CAT, and software SimuMCAT (Yao, 2011)
  - Test measures overall/content domain scores well, has short response time, and satisfy all constraints.
- Better Precision for the composite score AFQT (Armed Forces Qualification Test).

\[
\theta_{AFQT} = 1.18(w_{AR}\theta_{AR} + w_{MK}\theta_{MK} + w_{PC}\theta_{PC} + w_{MK}\theta_{WK}) - 0.28.
\]
Multidimensional Computer Adaptive Testing

- Multidimensional CAT, and software SimuMCAT (Yao, 2011)
- Short response time.
  - Response times for AR (Arithmetic Reasoning), WK (Word Knowledge), PC (Paragraph Comprehension), and MK (Mathematical Knowledge) are 79, 14, 69 and 42 seconds, respectively.
Multidimensional Computer Adaptive Testing

- Multidimensional CAT, and software SimuMCAT (Yao, 2011)

Select item $m=j$ such that,

$$\tilde{w} \left[ I_{j-1}^m \left( \tilde{\theta}^{j-1} \right) \right]^{-1} (\tilde{w})^T + W_{\text{time}} \times \text{Time}_m$$

has a minimum value or

$$\frac{\text{SEE}_{j-1} \left( \tilde{\theta}^{j-1} \right) - \tilde{w} \left[ I_{j-1}^m \left( \tilde{\theta}^{j-1} \right) \right]^{-1} (\tilde{w})^T}{\text{Time}_m W_{\text{time}}}$$

has a maximum value
Multidimensional Computer Adaptive Testing

- Multidimensional CAT, and software SimuMCAT (Yao, 2011)
  - Results: we are able to identify conditions or compose a test that meet the required precision (or Type I and Type II error for classification purpose) and has test response time less than 15 minutes.
- CAT ASVAB (Armed Services Vocational Aptitude Battery)
  - “ICAST Screening Test”: 4, 8, 3, 5 items for AR, WK, PC, and MK, respectively.
  - Regular full length test: 15, 15, 10, 15 items for AR, WK, PC, and MK, respectively.
Issues in MCAT compared to CAT

- Item pool is larger----items in different content domains are put together.
  - Item selection time is longer.
  - Find a solution. For example, divide item pools into smaller forms.
- Multidimensional ability estimates after each item selection takes longer time to compute than unidimensional ability estimates.
  - Find solution or improvement.
Issues in MCAT compared to CAT

• For unidimensional CAT, a sequential Bayesian procedure (Owen, 1969, 1975) is used to update ability using the scored response—it is computationally efficient than other Bayesian estimators.

• MAP is used to compute the final ability estimates using all responses----the order of item administration does not affect MAP but affect Owen estimator.

• Similar method for MCAT??
Issues in MCAT compared to CAT

- Long Item selection time and ability computation time is relatively speaking---not that long!
  - 4-dimensional and 900 items in the pool, 36 items used around 0.1-0.3 seconds for each examinee.
Models

- Software SimuMCAT (Yao, 2011): at www.BMIRT.com
  - M-3PL model for multiple choice items:
  - M-2PPC model for polytomously scored items.
  - Between-item (Simple structured item).
  - Within-item (Complex structured item).
  - Passage.
  - All previously described item selection methods and stopping rules.
Research Areas in CAT or MCAT

• Item Pool

  - Item pool generation: optimal item pool size and item characteristics.
    • Integer programming method (Boekkooi-Timminga, 1991, Veldkamap & van der Linden, 1999)—shadow test, optimal software, computational extensive.
    • Simulation approach (Reckase, 2003, 2007).
      - Within-item----No research yet.
      - Between-item.
    - Calibration of item pool.
      • Replace retired items.
      • Replace retired forms.
Research Areas in CAT or MCAT

• Item selection Algorithms.
• Stopping rules.
• Ability estimates.
• Cheating detection.
A Study comparing two types of item selection methods

• Motivation of the study:
  - The purpose of the test is to classify examinees into categories for their composite scores based on pre-fixed cut points.
  - Good classification accuracy but not necessarily precise score estimates.
A Study comparing two types of item selection methods

- Item selection methods
  - Select items based on the current ability estimates—AB
  - Select items based on classification at cut points—CB

- Models compared
  - UIRT
    - Bifactor model, with first dimensional measuring overall ability.
      - Five-dimensional.
      - Seven-dimensional.
A Study comparing two types of item selection methods

- MIRT
  - Four-dimensional IRT
  - Composite score=linear weighted sum of the content domain score.
    \[ \theta_{\alpha} = \sum_{i=1}^{D} \theta_i w_i \]
    - Prefixed weight
    - Optimal weight.
A Study comparing two types of item selection methods

- Item Pool: 253 items in total with four content areas AR, WK, PC and MK.

- Sample:
  - Generate responses:
    - 253 four-dimensional simple structured item parameters.
    - Four sets of size 3000 of four-dimensional Normal P1-P4.
A Study comparing two types of item selection methods

- BMIRT calibration:

  • One-dimensional ability and one-dimensional item parameters, used as true abilities and item pool, respectively.

  • Five-dimensional ability and five-dimensional item parameters used as true abilities and item pool, respectively.

  • Seven-dimensional ability and seven-dimensional item parameters used as true abilities and item pool, respectively.

  • Four-dimensional ability and four-dimensional item parameters used as the item pool. True values.
A Study comparing two types of item selection methods

• Results:
  - four-dimensional model D4 performed the best followed by unidimensional model D1, followed by D5.
  - For D4, optimal weight is slightly better than prefixed weight.
  - AB method or CB method using two or three cut points for D4 model performed similar.
  - To get a good classification rate, CB methods using equal or more cut points than the defined categories are desirable.
  - CB is better than AB for all models.
A Study comparing two types of item selection methods

• Results:
  - D4 CB method with two or three cut points has the smallest misclassification rate, Chi-square skewness, test overlap rate.
Thank you!

Free download software at

www.BMIRT.com