NATIONAL RESEARCH UNIVERSITY

39th Annual Conference IAEA 2013
International Association of Educational assessment
20-25 October 2013 | Tel Aviv , Israel

UNIVERSITAS OSTRAVIENSIS
Facultas Paedagogica

Computer Adaptive Testing Algorithm for middle school examinations in Czech Republic

Elena Kardanova, Dmitriy Abbakumov
National research university "Higher school of economics"
Moscow, Russia

Martin Malcik, Martin Rangl
University of Ostrava
Ostrava, Czech Republic

Different algorithms of computer based testing

Non-Adaptive	Adaptive
Linear: computer analogue of traditional p&p testing	Multi-stage: items are divided in several groups in accordance with their difficulty
Randomized: different test forms of fixed length are formed from an item pool	Computer Adaptive: individual set of items is selected for each examinee

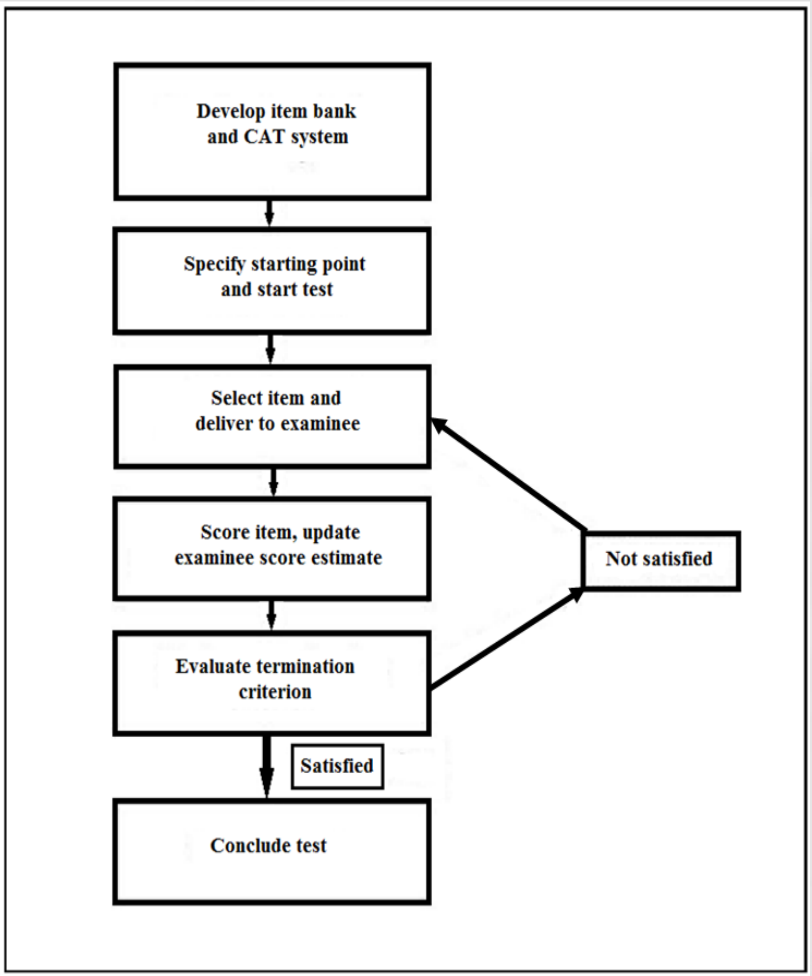
Five steps of CAT construction (Tompson N.A. and Weiss D.J., 2011)

Stage	Primary work
• Feasibility, applicability, and planning studies	Monte Carlo simulation; business case evaluation
• Develop item bank content or utilize existing bank	Item writing and review
• Pretest and calibrate item bank	Pretesting; item analysis
• Determine specifications for final CAT	Post-hoc or hybrid simulations
• Publish life CAT	Publishing abd distribution; software development

Focus of the presentation

- Item bank calibration for CAT construction
- Development of CAT algorithm using simulation study (Monte Carlo, post-hoc and hybrid simulations)

Standard CAT algorithm
(Weiss&Kingsbury, 1984; Thompson, 2007)



Five components of CAT:

(1) Calibrated item bank

(2) Starting point

(3) Item selection algorithm

(4) Scoring algorithm

(5) Termination criterion

Item bank calibration

- Item bank – a set of calibrated items, i.e. items with known parameters, that are placed on the common scale
- Usage of IRT as a poweful psychometric paradigm with many advantages for test development, item analysis, and scoring of examinees
- Pretesting of items developed. Requireement of big sample
- Item parameters must be estimated with IRT calibration software
- If there are several test forms in use, it is necessary to equate them using special procedures available in IRT

Key questions in CAT
(H.Wainer, 1990)

- How do we choose an item to start the test?
- How do we choose the next item to be administered after we have seen the examinee’s response to the current one?
- How do we know when to stop?

Simulation studies: why and how

- Monte Carlo simulations:

  ✓ typically useful in the early stages of investigating the performance characteristics of CAT procedures when little or no data are available

  ✓ allow to quickly and efficiently vary different aspects of the data in conjunction with varying the parameters that control hypothetical CATs

  ✓ the result is the ability to answer a wide range of “what if” questions

- Post-hoc и hybrid simulations:

  ✓ allow to evaluate the various CAT testing parameters prior to live testing

  ✓require an item response matrix of real examinees responding to a CAT item bank

  ✓ the simulation uses item responses to simulate how that item bank would function if the items (for which responses are known) had been administered as a CAT

- CATSim software (Weiss&Guyer, 2010)

  ✓allows to do all kinds of simulations: Monte Carlo, post-hoc и hybrid

The method

Instrument

5th-grade

Part 1: Mathematics (30 minutes), English or German language (20 minutes)

Part 2: Czech language (30 minutes), Science (20 minutes)

Overall testing time - 100 minutes

9th-grade

Part 1: Mathematics (40 minutes), English or German language (30 minutes)

Part 2 : Czech language (40 minutes), Physics, Chemistry, Biology (30 minutes)

Overall testing time - 140 minutes

Testing procedure

- ✓ On-line testing
- ✓ 7 test forms were used
- ✓ Task order in all test forms was fixed, answer rotation was applied in MC items
- ✓ All test forms included common items

Sample

Testing Year	Schools	Students of 9th graders	Students of 5th graders
2011	494	15580	18131
2012	452	14085	9389
2013	232	6747	4733
Total		36412	32253

The method: Item bank calibration

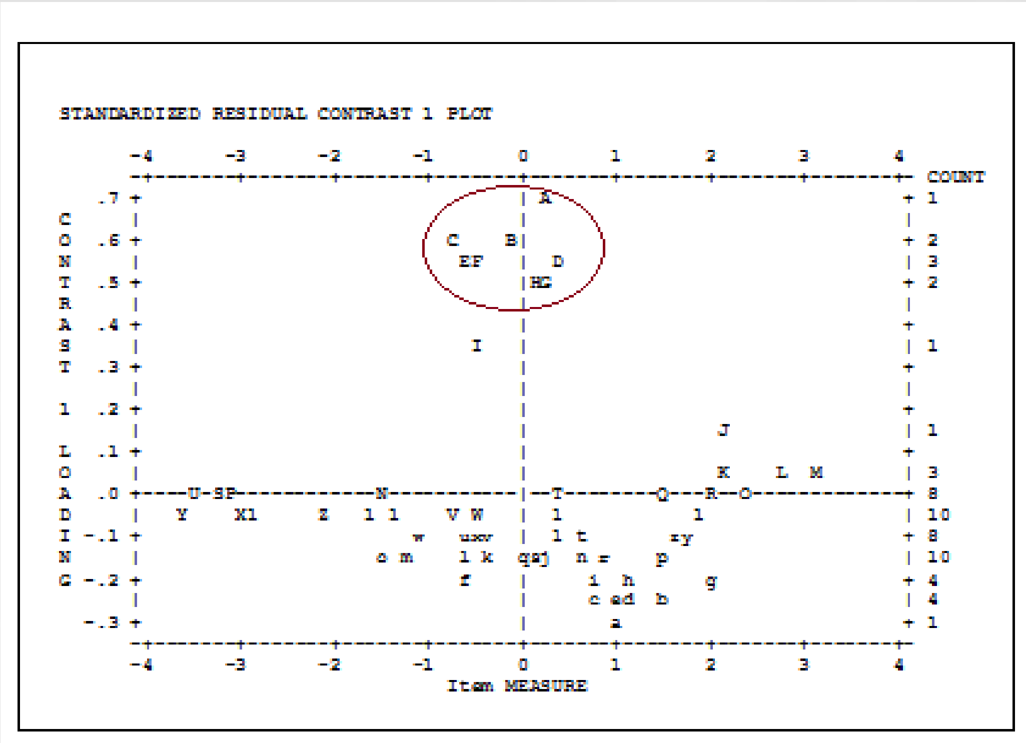
- Two stages:
  - ✓ Each of 7 test forms was calibrated separately
  - ✓ All test forms were calibrated simultaneously
- Model of measurement
  - ✓ The one-parameter dichotomous Rasch model (Wright B.D.&Stone M.N.,1979)
  - ✓ Winsteps software (Linacre J. M., 2011)
- Fit analysis
  - ✓ INFIT and OUTFIT mnsq statistics
- Dimensionality
  - ✓ Principal component analysis of the standardized residuals based on Rasch analysis (Linacre, J.M., 1998; Smith, E. V., 2002)
- DIF
  - ✓ Student's t-test and Mantel-Haenzel statistics

The results: Calibration of test form 1

	N	Minimum	Maximum	Mean	St. Dev.
Estimated Item Measure:	59	-3,61	3,18	,00	1,60
Point-measure correlation:	59	,00	,67	,37	,14
Item Discrimination (approximates 2-PL)	59	,36	1,77	,99	,31
Item Proportion correct	59	,00	,96	,49	,26

```
-------------------------------------------------------------------------------------
|ENTRY   TOTAL  TOTAL           MODEL|   INFIT   |  OUTFIT   |PT-MEAS|                 |
|NUMBER  SCORE  COUNT  MEASURE  S.E. |MNSQ  ZSTD|MNSQ  ZSTD|CORR.   | Item          G |
|------------------------------------+----------+----------+--------+-----------------|
|    87    690   4333    2.01     .04|1.20   6.8|1.69   9.9|A .12   | 87 2995_4720 0 |
|    56    710   2071     .80     .05|1.25   9.9|1.46   9.9|B .18   | 56 2950_4771 0 |
|    91    988   2792     .77     .04|1.26   9.9|1.46   9.9|C .16   | 91 3062_4882 0 |
|    76   1226   4235    1.03     .04|1.19   9.9|1.43   9.9|D .20   | 76 2982_4707 0 |
|    86   1298   4612    1.14     .04|1.18   9.9|1.43   9.9|E .21   | 86 2993_4719 0 |
|    92    846   2782    1.04     .05|1.22   9.9|1.41   9.9|F .18   | 92 3062_4883 0 |
                                 .................................
|    26   2768   4529    -.49     .03| .81  -9.9| .75  -9.9|h .58   | 26 2917_4665 0 |
|    30   2729   4231    -.64     .04| .80  -9.9| .73  -9.9|g .58   | 30 2917_4669 0 |
|    28   2879   4321    -.76     .04| .79  -9.9| .71  -9.9|f .59   | 28 2917_4667 0 |
|    25   1965   4495     .37     .03| .79  -9.9| .73  -9.9|e .61   | 25 2917_4664 0 |
|    31   1840   3762     .16     .04| .78  -9.9| .73  -9.9|d .63   | 31 2917_4670 0 |
|    32   1860   3780     .15     .04| .78  -9.9| .72  -9.9|c .63   | 32 2917_4671 0 |
|    29   2253   4181    -.10     .03| .77  -9.9| .73  -9.9|b .62   | 29 2917_4668 0 |
|    27   1964   4314     .30     .03| .73  -9.9| .67  -9.9|a .67   | 27 2917_4666 0 |
|------------------------------------+----------+----------+---------+---------------- |
| MEAN  1839.5 3648.3    .00      .05|1.00    .5|1.00    .3|         |                 |
| S.D.  1156.9 1085.3   1.60      .02| .13   6.1| .26   6.4|         |                 |
```

Dimensionality study



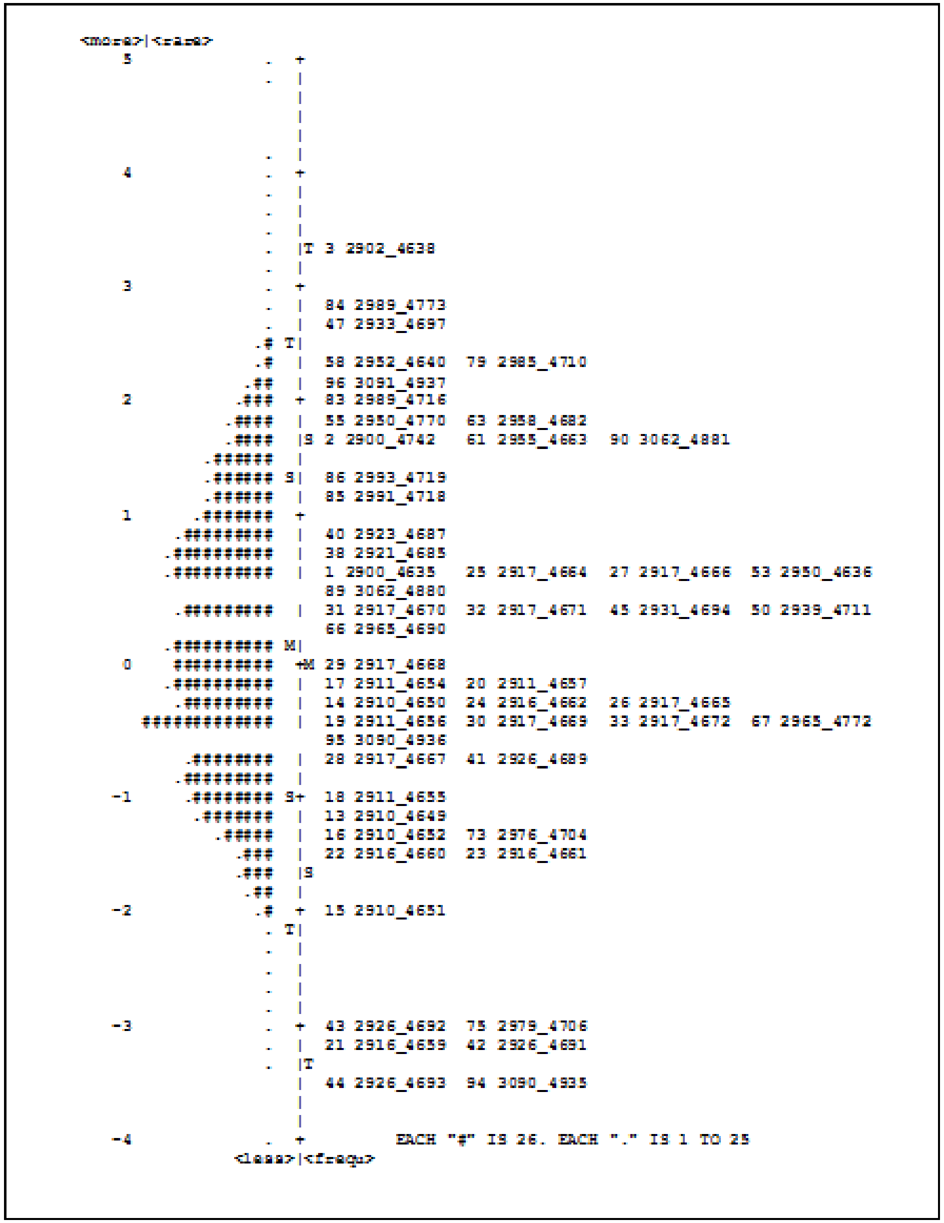
Presence of such items is a problem for P&P testing.
But it is not a problem for CAT: it is just necessary to indicate it in the content specification

CON- TRAST	LOADING	MEASURE	INFIT MNSQ	OUTFIT MNSQ		ENTRY NUMBER	Item
1 1	.68	.30	.73	.67	A	27	27 2917_4666
1 1	.60	-.10	.77	.73	B	29	29 2917_4668
1 1	.59	-.76	.79	.71	C	28	28 2917_4667
1 1	.56	.37	.79	.73	D	25	25 2917_4664
1 1	.56	-.64	.80	.73	E	30	30 2917_4669
1 1	.53	-.49	.81	.75	F	26	26 2917_4665
1 1	.52	.15	.78	.72	G	32	32 2917_4671
1 1	.48	.16	.78	.73	H	31	31 2917_4670
1 1	.36	-.53	.87	.81	I	33	33 2917_4672

Summary statistics of the test form 1

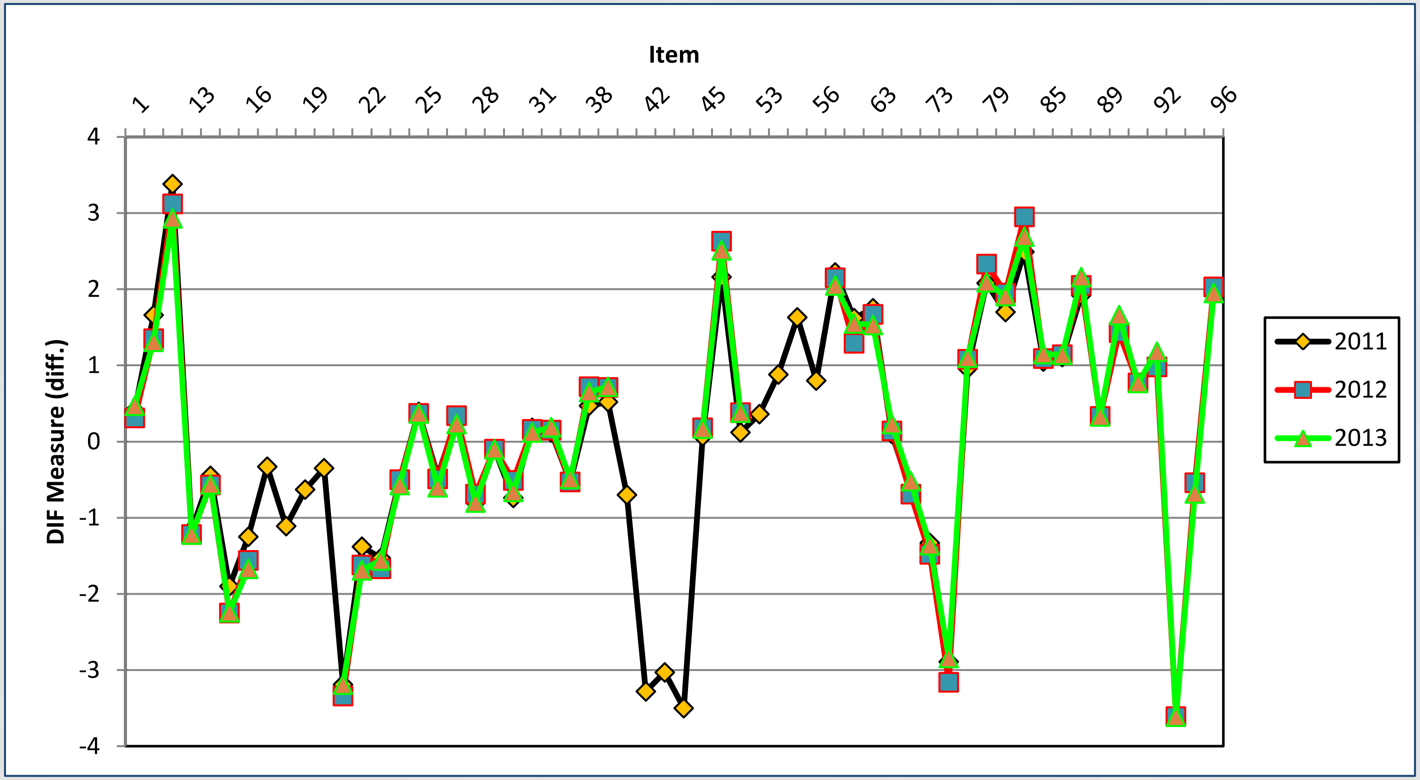
```
SUMMARY OF 5062 MEASURED (NON-EXTREME) Students
-------------------------------------------------------------------------------
|          TOTAL                         MODEL         INFIT         OUTFIT    |
|          SCORE     COUNT     MEASURE   ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN      20.1      38.2         .15     .43      1.00     .0   1.00     .0 |
| S.D.       7.8       6.7        1.14     .07       .23    1.1    .51     .9 |
| MAX.      46.0      47.0        4.95    1.52      3.39    4.3   9.64    4.7 |
| MIN.       1.0       5.0       -3.96     .35       .36   -3.2    .17   -2.0 |
|-----------------------------------------------------------------------------|
| REAL RMSE    .45 TRUE SD    1.05  SEPARATION  2.31Studen RELIABILITY  .84   |
|MODEL RMSE    .43 TRUE SD    1.06  SEPARATION  2.45  Studen RELIABILITY  .86 |
| S.E. OF Students MEAN = .02                                                  |
-------------------------------------------------------------------------------
  MINIMUM EXTREME SCORE:      5 Students
      LACKING RESPONSES:    155 Students
        VALID RESPONSES:  73.4%  (APPROXIMATE)
Students RAW SCORE-TO-MEASURE CORRELATION = .91 (approximate due to missing data)
CRONBACH ALPHA (KR-20) Students RAW SCORE "TEST" RELIABILITY = .86

     SUMMARY OF 52 MEASURED (NON-EXTREME) Item
-------------------------------------------------------------------------------
|          TOTAL                         MODEL         INFIT         OUTFIT    |
|          SCORE     COUNT     MEASURE   ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN    1952.8    3717.5         .00     .05       .99     .4    .99     .1 |
| S.D.    1168.0    1081.4        1.68     .02       .13    6.2    .26    6.2 |
| MAX.    4698.0    4974.0        3.39     .11      1.23    9.9   1.62    9.9 |
| MIN.     216.0    1838.0       -3.57     .03       .72   -9.9    .53   -9.9 |
|-----------------------------------------------------------------------------|
| REAL RMSE    .05 TRUE SD    1.68  SEPARATION 32.15  Item   RELIABILITY 1.00 |
|MODEL RMSE    .05 TRUE SD    1.68  SEPARATION 32.74  Item   RELIABILITY 1.00 |
| S.E. OF Item MEAN = .24                                                      |
-------------------------------------------------------------------------------
```

The test form 1 variable map



```
<more>|<rare>
  5             .  +
                .  |
                   |
                   |
                   |
                .  |
  4             .  +
                .  |
                .  |
                .  |
                .  |T 3 2902_4638
                .  |
  3             .  +
                .  |  84 2989_4773
                .  |  47 2933_4697
               .# T|
               .#  |  58 2952_4640  79 2985_4710
              .##  |  96 3091_4937
  2          .###  +  83 2989_4716
            .####  |  55 2950_4770  63 2958_4682
            .####  |S 2 2900_4742   61 2955_4663  90 3062_4881
          .######  |
          .###### S|  86 2993_4719
          .######  |  85 2991_4718
  1      .#######  +
       .#########  |  40 2923_4687
      .##########  |  38 2921_4685
      .##########  |  1 2900_4635   25 2917_4664  27 2917_4666  53 2950_4636
                      89 3062_4880
       .#########  |  31 2917_4670  32 2917_4671  45 2931_4694  50 2939_4711
                      66 2965_4690
      .########## M|
  0    ##########  +M 29 2917_4668
      .##########  |  17 2911_4654  20 2911_4657
       .#########  |  14 2910_4650  24 2916_4662  26 2917_4665
    #############  |  19 2911_4656  30 2917_4669  33 2917_4672  67 2965_4772
                      95 3090_4936
        .########  |  28 2917_4667  41 2926_4689
       .#########  |
 -1     .######## S+  18 2911_4655
         .#######  |  13 2910_4649
           .#####  |  16 2910_4652  73 2976_4704
             .###  |  22 2916_4660  23 2916_4661
             .###  |S
              .##  |
 -2            .#  +  15 2910_4651
                . T|
                .  |
                .  |
                .  |
                .  |
 -3             .  +  43 2926_4692  75 2979_4706
                .  |  21 2916_4659  42 2926_4691
                .  |T
                   |  44 2926_4693  94 3090_4935
                   |
                   |
 -4             .  +        EACH "#" IS 26. EACH "." IS 1 TO 25
              <less>|<frequ>
```

DIF analysis across years



Students Label		N	Minimum	Maximum	Mean	Std. Deviation
2011	Estimated Students Measure:	2242	-4,03	3,99	,01	1,03
2012	Estimated Students Measure:	2002	-3,15	4,13	,03	1,02
2013	Estimated Students Measure:	978	-3,04	3,35	,06	1,05

Summary for 7 test forms

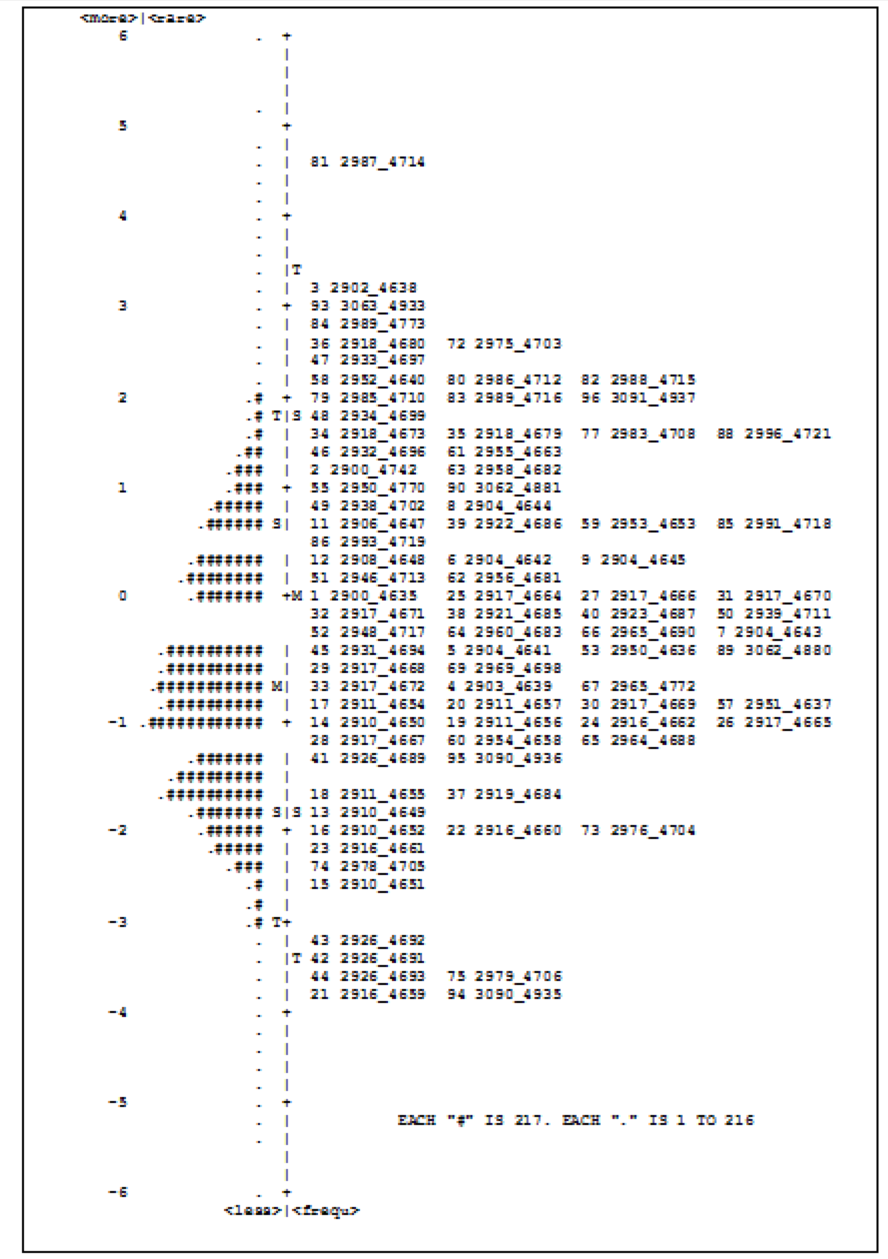
	Number of examinees	Number of items	Number of items left	Reliability	Error of measurement
Test form 1	5222	58	52	0.86	0.43
Test form 2	5203	55	51	0.88	0.45
Test form 3	5210	34	31	0.82	0.52
Test form 4	5244	50	43	0.80	0.47
Test form 5	5202	47	40	0.85	0.49
Test form 6	5186	58	54	0.87	0.43
Test form 7	5222	34	30	0.74	0.53

- The same items demonstrated poor fit across all test forms. There were 13 items in total that were needed to be deleted

- All other items have satisfactory psychometric characteristics and are functioning by similar way for three years

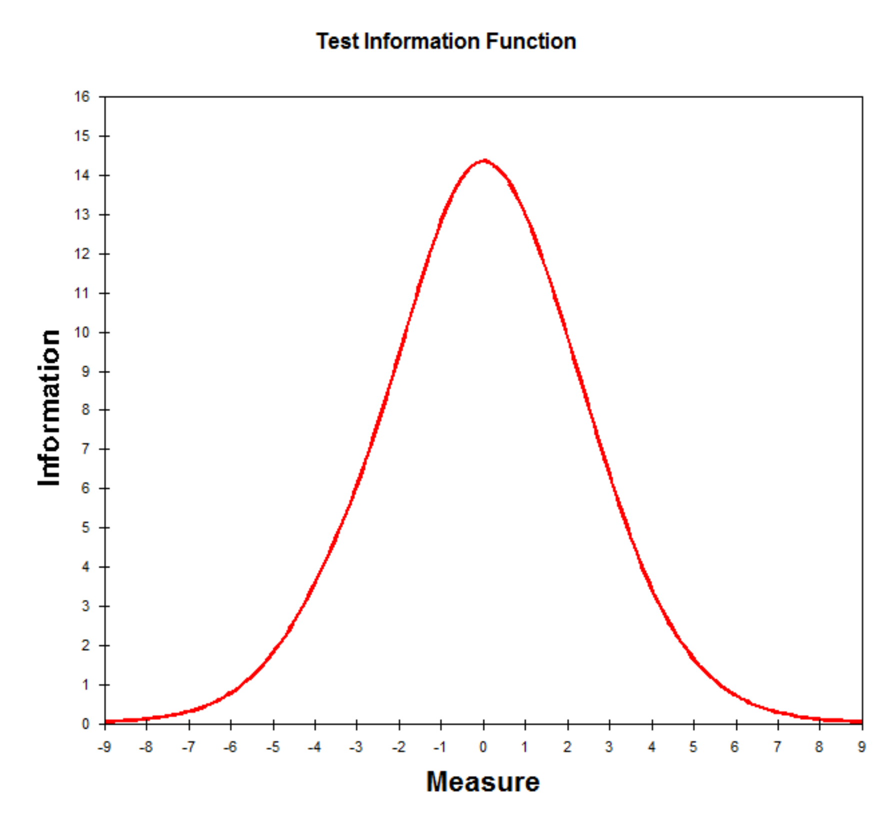
Conclusion: all test forms can be used for item bank construction

Simultaneous calibration



```
      <more>|<rare>
 6            .  +
                 |
                 |
                 |
              .  |
 5               +
              .  |
              .  |  81 2987_4714
              .  |
              .  |
 4            .  +
              .  |
              .  |
              .  |T
              .  |  3 2902_4638
 3            .  +  93 3063_4933
              .  |  84 2989_4773
              .  |  36 2918_4680  72 2975_4703
              .  |  47 2933_4697
              .  |  58 2952_4640  80 2986_4712  82 2988_4715
 2           .#  +  79 2985_4710  83 2989_4716  96 3091_4937
             .# T|S 48 2934_4699
             .#  |  34 2918_4673  35 2918_4679  77 2983_4708  88 2996_4721
            .##  |  46 2932_4696  61 2955_4663
           .###  |  2 2900_4742   63 2958_4682
 1         .###  +  55 2950_4770  90 3062_4881
         .#####  |  49 2938_4702  8 2904_4644
        .###### S|  11 2906_4647  39 2922_4686  59 2953_4653  85 2991_4718
                    86 2993_4719
       .#######  |  12 2908_4648  6 2904_4642   9 2904_4645
      .########  |  51 2946_4713  62 2956_4681
 0     .#######  +M 1 2900_4635   25 2917_4664  27 2917_4666  31 2917_4670
                    32 2917_4671  38 2921_4685  40 2923_4687  50 2939_4711
                    52 2948_4717  64 2960_4683  66 2965_4690  7 2904_4643
    .##########  |  45 2931_4694  5 2904_4641   53 2950_4636  89 3062_4880
    .##########  |  29 2917_4668  69 2969_4698
   .########### M|  33 2917_4672  4 2903_4639   67 2965_4772
    .##########  |  17 2911_4654  20 2911_4657  30 2917_4669  57 2951_4637
-1 .############ +  14 2910_4650  19 2911_4656  24 2916_4662  26 2917_4665
                    28 2917_4667  60 2954_4658  65 2964_4688
       .#######  |  41 2926_4689  95 3090_4936
     .#########  |
    .##########  |  18 2911_4655  37 2919_4684
       .####### S|S 13 2910_4649
-2      .######  +  16 2910_4652  22 2916_4660  73 2976_4704
         .#####  |  23 2916_4661
           .###  |  74 2978_4705
             .#  |  15 2910_4651
             .#  |
-3           .# T+
              .  |  43 2926_4692
              .  |T 42 2926_4691
              .  |  44 2926_4693  75 2979_4706
              .  |  21 2916_4659  94 3090_4935
-4            .  +
              .  |
              .  |
              .  |
              .  |
-5            .  +
              .  |          EACH "#" IS 217. EACH "." IS 1 TO 216
              .  |
                 |
                 |
-6            .  +
         <less>|<frequ>
```

The total number of items: 85
The total number of examinees: 36490
Mean error of measurement: 0.45
Separation index: 2.27



Simulation study 1: is it possible to realize CAT with this item bank

Set of options

- ✓ Sample size: 5000
- ✓ Ability distribution: β-distribution, $\alpha = 1$, $\beta = 1$, [-3 , 3]
- ✓ Starting rule: Initial level of difficulty was chosen randomly from the interval [-1; 1]
- ✓ Ability estimatiom method: WML
- ✓ Selection of the next item: maximum of the information function on the current value of ability estimation
- ✓ The termination criteria:
  1) the change in measurement standard errors is equal or less than 0,001 logit.
  2) + standard error of the ability estimates is less or equal to 0.35 logits
  3) + the minimum 40 and the maximum 45 numbers of items constraints

Correlations between ability parameters

	Generated	Fullbank
Generated	1	
Full bank	0.982	1
CAT 1	0.962	0.974
CAT 2	0.959	0.971
CAT 3	0.964	0.980

Simulations results

Parameter	**Full bank**	**CAT 1**	**CAT 2**	**CAT 3**
SE Mean	0.312	0.503	0.541	0.455
SE SD	0.111	0.340	0.320	0.19
Number of items per one examinee				
Mean		47	35	41

Simulation study 2: a set of CAT rules for the given item bank

Set of options

- ✓ Real sample : 36490 examinees
- ✓ Starting rule, parameter estimation method and item selection rule are the same
- ✓ Termination criteria:

(1) the level of the standard error of measurement equal or less than 0.350 logit or all possible items have been used

(2) + the minimum 40 and the maximum 45 numbers of items constraints

Correlations between ability parameters

	Fullbank
CAT 1	0.992
CAT 2	0.993

Simulation study 2 results:

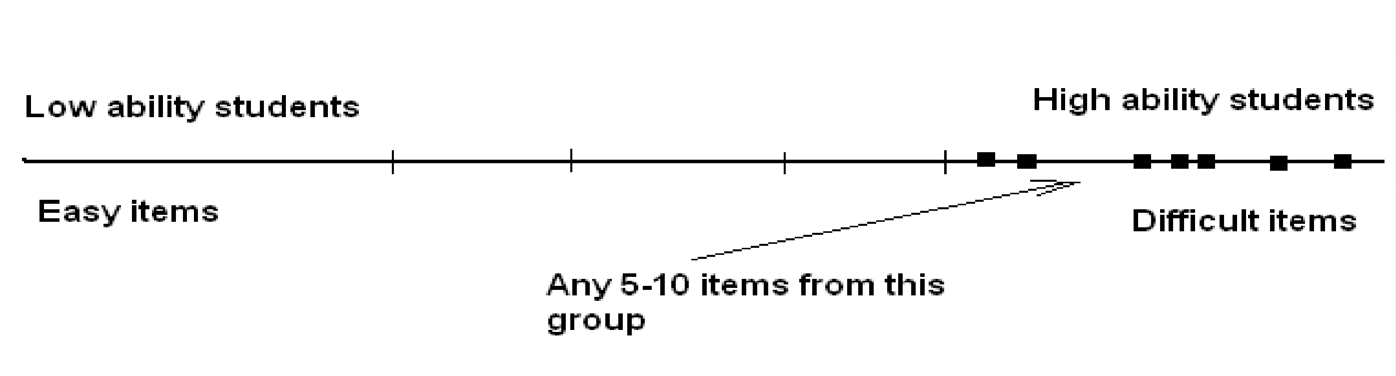
	Non-CAT	CAT 1	CAT 2
SE Mean	0.47	0.386	0.382
SE SD	-	0.191	0.194
SE Min	-	0.342	0.326
SE Max	-	1.465	1.469
NI Mean	43	44	41
NI SD	-	15	2
NI Min	-	34	40
NI Max	-	85	45

Simulation studies conclusions

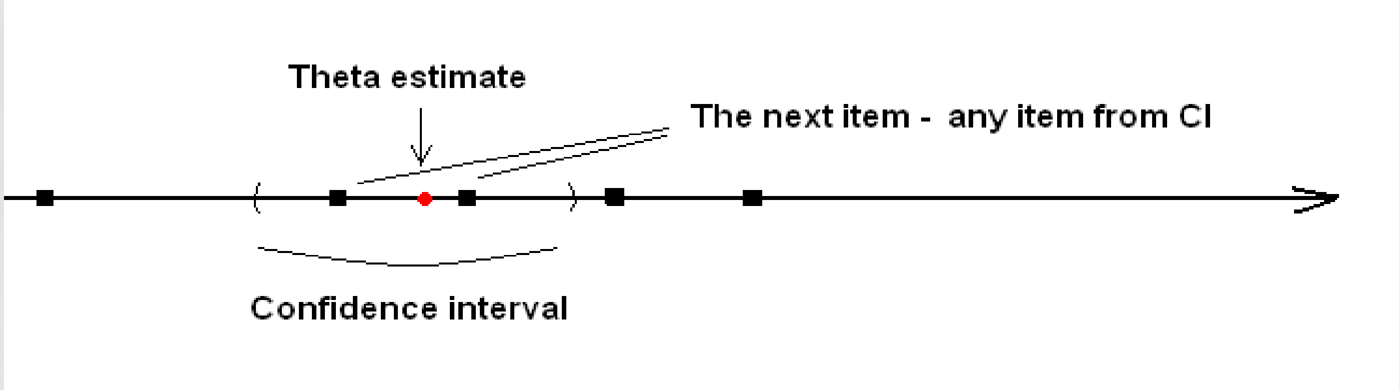
- Confirmation of the item bank applicability to implement the CAT algorithm
- CAT can provide substantial reduction in the standard error of measurement in comparing with non-adaptive testing
- Limitation of the minimum and maximum numbers of items does not result in loss the quality students' estimation
- The optimal termination criteria were determined

Conclusion: CAT algorithm

How to start?



How to continue?



How to stop?

Termination criteria: the level of the standard error of measurement equal or less than 0.350 logit and the number of items is in the range from 40 to 45.

Thank you for your attention

ekardanova@hse.ru