Practitioner's Approach to Identify Item Drift in CAT

Huijuan Meng, Susan Steinkamp, Pearson
Paul Jones, Joy Matthews-Lopez, NABP

Introduction

- Item parameter drift (IPD): change in item parameters over time.
- Possible causes: changes in curriculum and training; candidates' increasing familiarity with frequently exposed items.
- Impact of IPD: affect psychometric quality of IRT applications
  - CAT: item selection; ability estimate
  - Pretest item calibration
- Evaluate IPD: maintain a stable scale and ensure the quality of item calibration

CAT Program and Data

- Data: a fixed-length CAT using 3P model
- Number of candidates: 15,000
- Test length: 150 operational (scored) items
- Number of items in the data: 1,921
- Number of items in IPD check: 1,208 items (N>=500)
- Baseline scale: item pool
- Purpose: develop procedures that can be used to efficiently identify items drifting away from the baseline scale in a real CAT data.

IPD Literature (1)

- IPD procedures have often been examined in the fixed-form test data.
- DIF, drift, IRT model misfit: all demonstrate the lack of invariance of item parameters in the data
- IPD identification in CAT research:
  - Lord's χ^2 statistic
  - CUSUM method
  - Raju's NCDIF

IPD Literature (2)

- Lord's χ^2 statistic (2P & 3P): use parameter differences and 2 sets of asymptotic variance-covariance matrices of maximum likelihood estimators for original and new item parameters; fit in the general framework of Wald test.
- CUSUM procedure: a sequential series of Wald tests, in which standardized parameter differences are sequentially added for each time period.
- Issues with Lord's χ^2 & CUSUM:
  - Unavailability of asymptotic variance-covariance matrix for original item parameter estimates
  - Impact of item sample size on the magnitude of the asymptotic matrix
- Raju's NCDIF: rely on Monte Carlo technique
  - A large number of replications—time consuming
  - Numerous item parameter sets from the asymptotic variance-covariance matrix for newly calibrated parameters—restriction can't be guaranteed

G^2 Statistic

- G^2 is a likelihood ratio chi-square statistic.

Observed proportion correct

Observed proportion incorrect

$$G^2{}_j = 2 * \sum_{h=1}^{n_g} [r_{hj} \log_e \frac{\boxed{r_{hj} / N_{hj}}}{\boxed{P_j(\hat{\bar{\theta}}_{hj})}} + (N_{hj} - r_{hj}) \log_e \frac{\boxed{(N_{hj} - r_{hj}) / N_{hj}}}{\boxed{[1 - P_j(\hat{\bar{\theta}}_{hj})]}}]$$

Interval theta mean

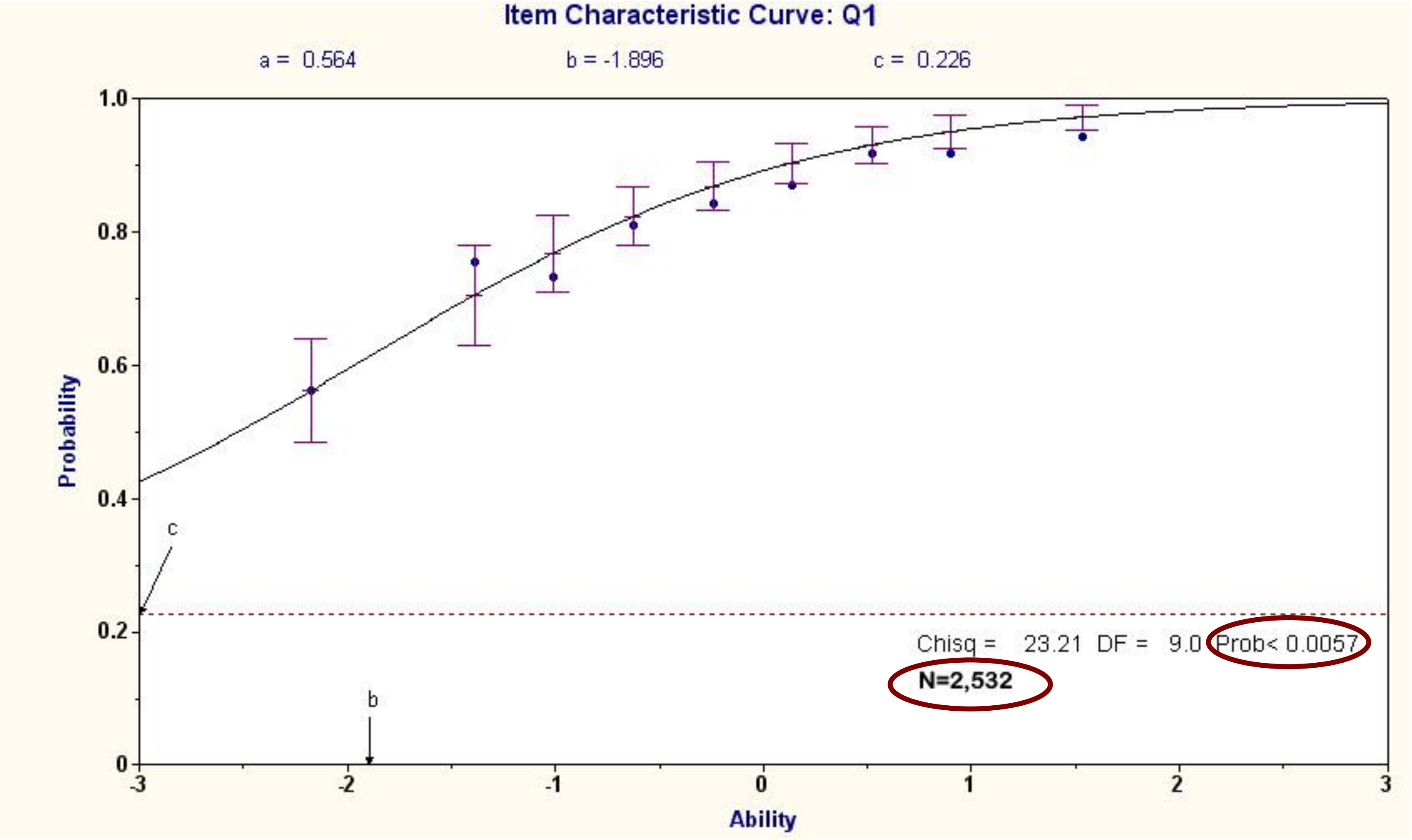
Model-based proportion correct

Model-based proportion incorrect

- G^2 can't be computed for an item with insufficient cases.
- G^2 issue: with a large sample size, an item can be flagged with even a trivial model misfit.

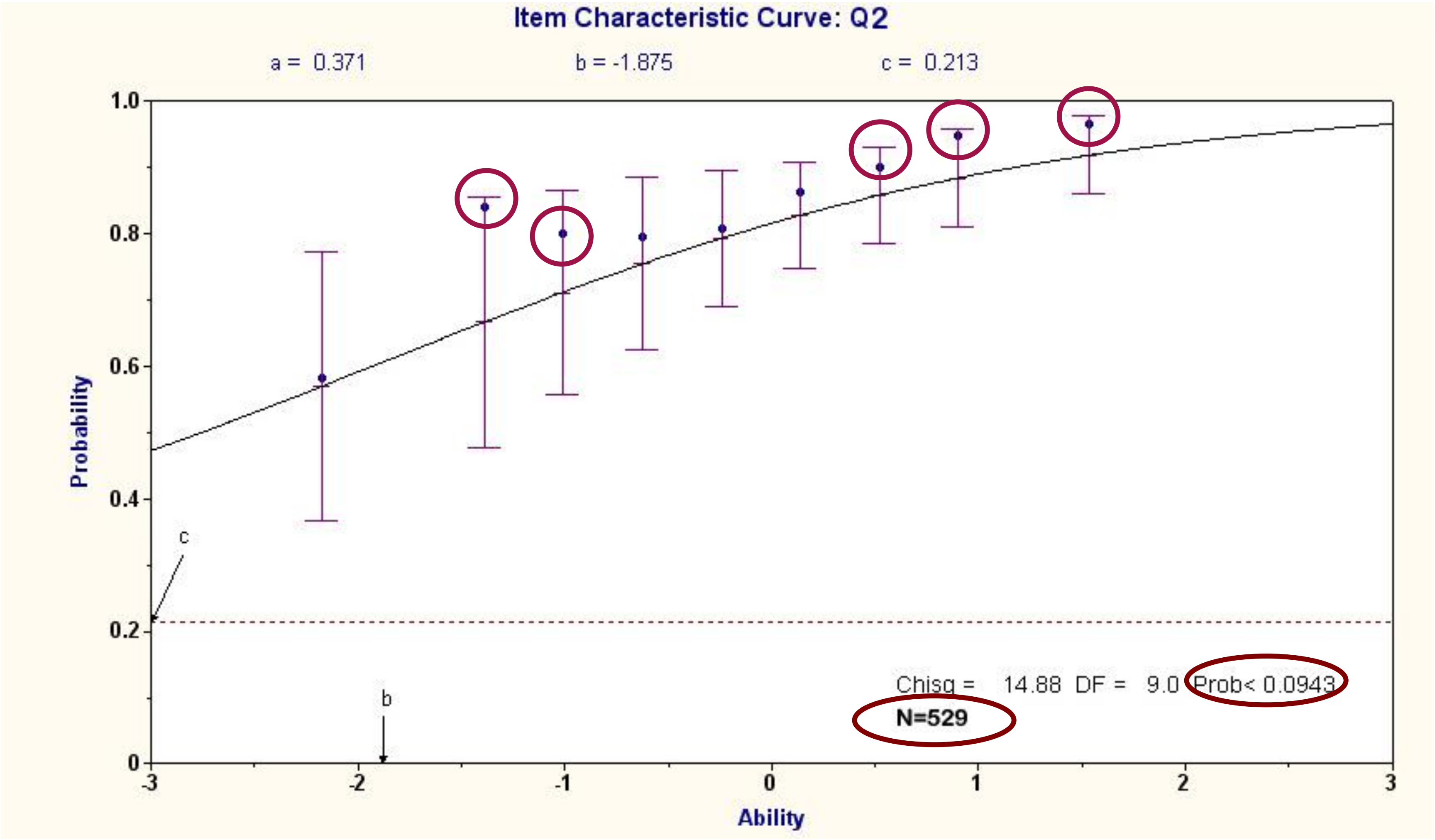
PEARSON

Item plot (1): Flagged Item (P<0.01)



PEARSON

Item plot (2): Non-flagged Item (P<0.01)



G^2 Computing and Item Plotting

- BILOG-MG plot
  - Phase 2 output: poor quality
  - IRT Graphics tool: inconvenience
- Visual inspection: subjective and time-consuming
- Quantitative evaluation: more efficient and more objective
- BILOG-MG: no detailed interim computation results for G^2
- For each item: compute a G^2, produce a plot, then use the discrepancies between observed and model-based values to refine statistical test results and to categorize items.

Initial IPD Identifications: G^2 Statistic

- G^2 comparison: BILOG-MG vs. VUE

VUE Flag	BILOG-MG Flag		
	No	Yes	Total
No	**475**	42	517
Yes	35	**656**	691
Total	510	698	**1208**

Consistency: 1131

(475+656)/1208=93.6%

Inconsistency: 77

(42+35)/1208=6.4%

- alpha=0.01: among 1,208 items, 77 (6%) are classified differently, G^2 flagging consistency rate: 94%
- Possible cause: use different interval merging methods

Further IPD Identifications (1) : Drift Category

- Indices to check item parameter drift:
  - **P-DIF**: discrepancy between <u>observed</u> and <u>model-based</u> proportion correct at each theta interval;
  - **Drift**: average of P-DIFs across all intervals;
  - **Absolute drift**: average of absolute P-DIFs across all intervals.

Drift Category	VUE G^2 Flag		Total
	No	Yes	
OK	**361 (76%)**	**112 (24%)**	473
E	**10**	111	121
EE	0	45	45
EEE	0	13	13
H	0	12	12
HH	0	5	5
HHH	0	2	2
V	**145**	344	489
VV	**1**	45	46
VVV	0	2	2
Total	517	691	1208

473 (39%) drift OK

179 (15%) getting easier

19 (2%) getting harder

537 (44%) Mixed directions

404 (75%): easier

133 (25%): harder

Further IPD Identifications (2) Standard Indices

- **Standard P-DIF:** P-DIF / standard error of model-fit ICC value at each interval

$$Standard\ Error\ (P(\bar{\hat{\theta}}_{hj})) = \sqrt{P(\bar{\hat{\theta}}_{hj}) * (1 - P(\bar{\hat{\theta}}_{hj})) / N_{hj}}$$

- **Standard Drift Flag (Yes/No)**:

  Yes: standard P-DIF mean <= -1.645 or >= + 1.645

- **Absolute Standard Drift Flag (Yes/No)**:

  Yes: absolute standard P-DIF mean >= +1.645

- **Lower Asymptote Flag (Yes/No)**:

  Yes: 2 lower standard P-DIF values >= +2 or <= -2

- **Upper Asymptote Flag (Yes/No)**:

  Yes: 2 upper standard P-DIF values >= +2 or <= -2

- **Medium and large drift Flag (Yes/No)**:

  Yes: drift category is NOT OK, E, H, or V

Further IPD Identifications (3) Final Classification

- Each of the **1208** items is placed under one of two categories: Recalibration or Anchor.
- **Decision rule**:

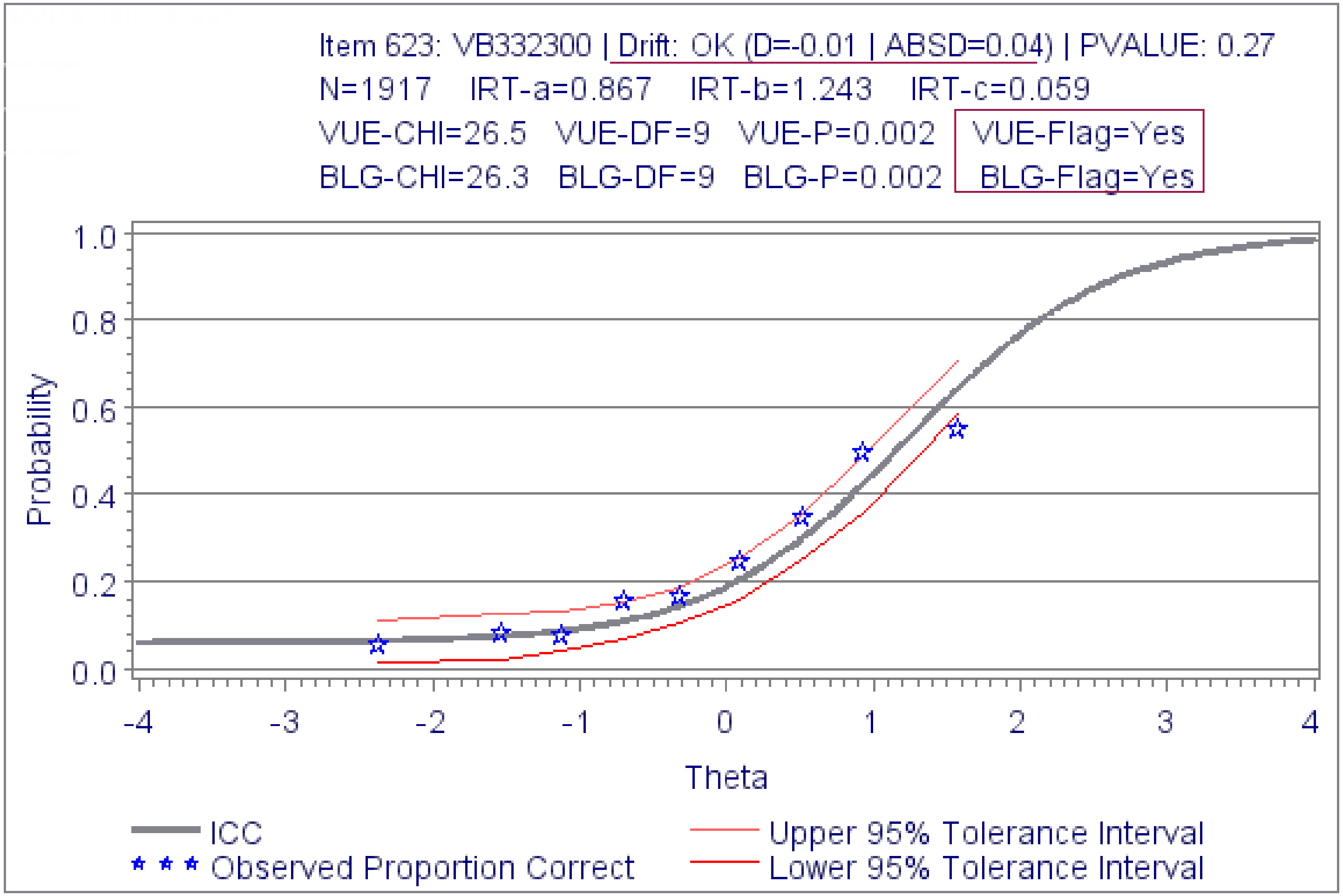
  If (VUE G^2 flag = Yes or BILOG-MG G^2 = Yes or Drift Category ≠ OK) and (any of the drift flags = Yes) **then** the item is placed in the **Recalibration** category (509), **otherwise** it is placed in the **Anchor** category (699).

Category	Anchor		Recalibration		Total
	G^2 No	G^2 Yes	G^2 No	G^2 Yes	
OK	358	59	3	53	473
E	10	30	0	81	121
EE	0	0	0	45	45
EEE	0	0	0	13	13
H	0	2	0	10	12
HH	0	0	0	5	5
HHH	0	0	0	2	2
V	136	104	9	240	489
VV	0	0	1	45	46
VVV	0	0	0	2	2
Total	504 (72%)	195 (28%)	13 (3%)	496 (97%)	1208

504+195=699 **13+496=509**

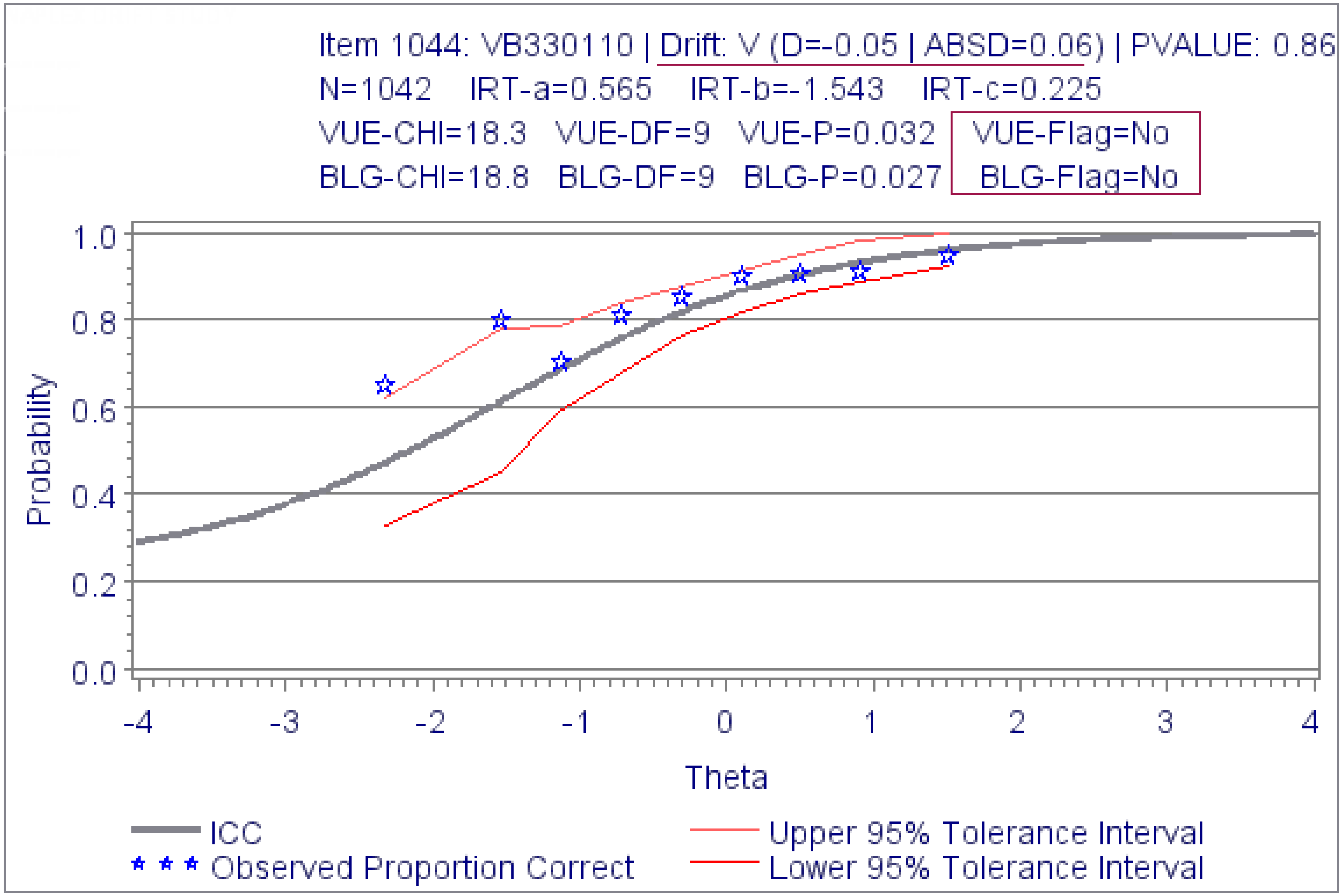
PEARSON

Item plot (1): Anchor Item (Flagged by G²)



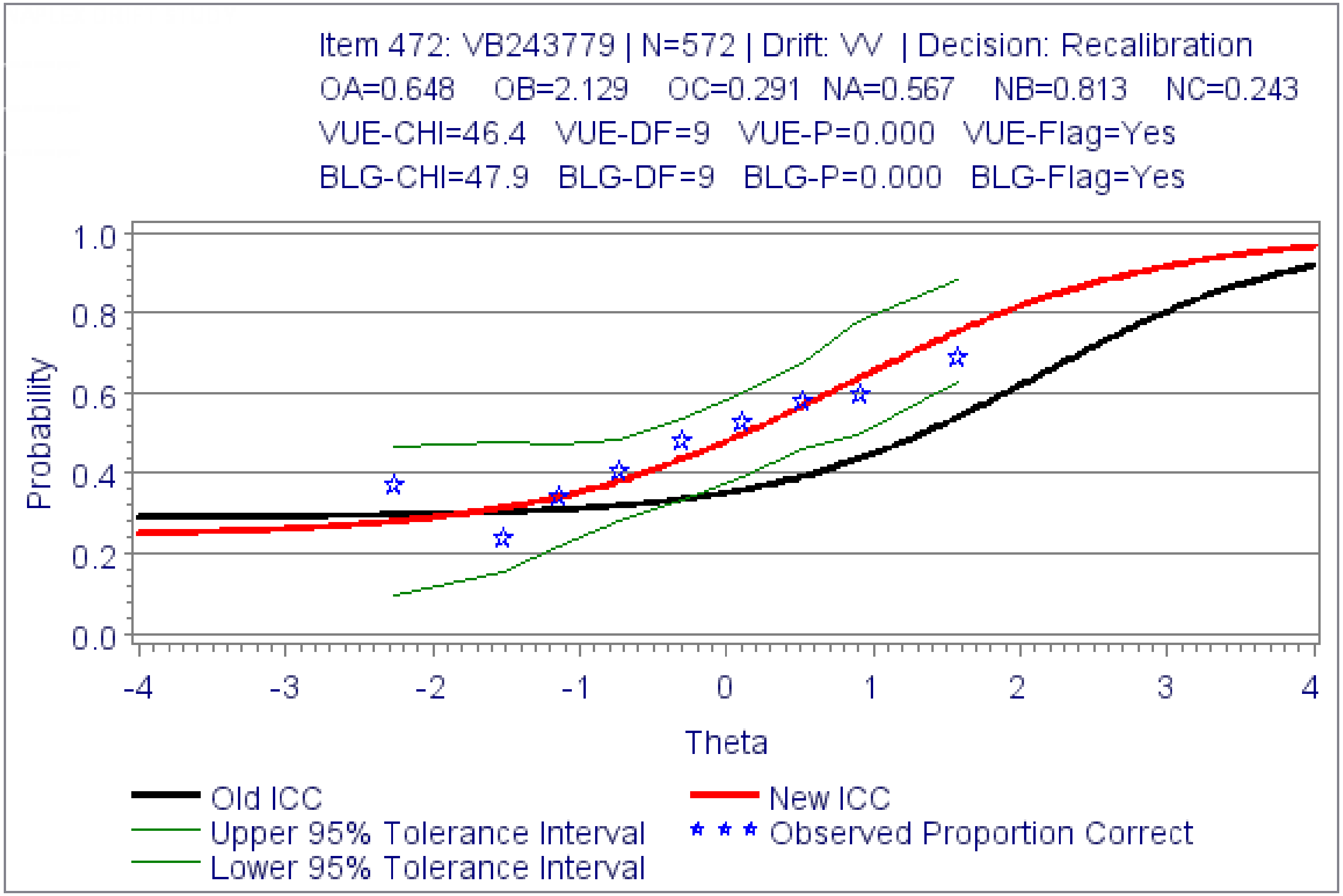
PEARSON

Item plot (2): Recalibration Item (Not flagged by G^2)



PEARSON

Item plot (3): Recalibration Item ICC



PEARSON

Summary

- Using both G^2 statistic and criteria derived from the discrepancies between <u>observed</u> and <u>model-based</u> proportion correct, we check parameter drift for 1,208 operational items.
- Plots for those items have been produced and scanned; in general, the real data support our final classification of items and recalibration outcomes.
- Although the results can be confounded by item model misfit in original data calibration, it is still considered as a practical way of identifying drift items in a real CAT data.
- A simulation study should be conducted to further examine the accuracy of this approach.
- Finally, we will not completely replace parameters for all flagged items with newly calibrated values; instead, we have procedures to determine whether using recalibration results for an item directly or updating an item parameters by reconciling original and new values.

Questions? Comments?

Thank You!

huijuan.meng@pearson.com