Online Adaptive Item Calibration with Polytomous Items

Hao Ren, Seung Choi, and Wim J. van der Linden

Introduction

- The number of applications of computerized adaptive testing (CAT) has increased dramatically over the past decade. For example,
  - K-12 summative assessment programs
  - Diagnostic, formative, on-demand assessments
  - Patient-reported health outcomes measurement arena
- One of the consequences is an increased pressure to replenish the item pool more frequently and develop and calibrate new items more efficiently.
- More efficient item pool replenishment may necessitate smaller calibration sample sizes. With smaller sample sizes, calibration error of item parameters is likely to be substantial; consequently, the uncertainty of ability estimation is increased.

Introduction

- Incorporating the uncertainty about the ability estimation into item parameter calibration calls for a Bayesian solution.
- A Bayesian approach based on an optimized Markov chain Monte Carlo (MCMC) algorithm for dichotomous items was introduced in van der Linden and Ren (2015).
  - Rapid real-time mixing of the Markov chain and simple posterior calculations for dichotomous items using the 3PL model
  - Optimal design for field test item assignment
- The similar idea has been implemented in Bayesian CAT with dichotomous items (van der Linden & Ren, 2020) and polytomous items (Ren, Choi, & van der Linden, 2020).
  - The uncertainty about item parameters for the ability estimation.

Introduction

- Interest in CAT with polytomous items has grown considerably both in educational testing and health-outcomes measurement fields.
- Unlike dichotomous items, the information function of polytomous items tends to span a wider ability/trait range and can be multi-modal.
  - May have effect on the calibration sample size, effectiveness of optimal design, etc.
- It is worth evaluating the implementation of Bayesian optimal design for polytomous items and evaluate its behavior and performance.

Models and Methods

- Item Response Theory (IRT) Models
- Fisher's Information for Item Parameters
- Assignment of Field Test Items
- Update of Item Parameters

Item Response Theory Models

- Generalized Partial Credit Model (GPCM)

The GPCM defines the probability of receiving a score on an item in category $c = 0,1,\ldots,m-1$ as

$$P_c(\theta) = \frac{\exp[\sum_{v=0}^{c} Z_v(\theta)]}{\sum_{c=0}^{m-1} \exp[\sum_{v=0}^{c} Z_v(\theta)]}$$

with

$$Z_v = a(\theta - b_v)$$

where m is the number of response categories, θ is the examinee ability parameter, a is the discrimination parameter, and bs are step difficulty parameters and $b_0 = 0$.

Item Response Theory Models

- Graded Responses Model (GRM)

The GRM defines the probability of selecting the ordered response categories c of an item as

$$P_c(\theta) = P_c^*(\theta) - P_{c+1}^*(\theta)$$

and $P_c^*(\theta)$ is defined as

$$P_c^*(\theta) = \begin{cases} 1 & c = 0 \\ 1/(1 + \exp(-a(\theta - b_c))) & 0 < c < m \\ 0 & c = m \end{cases}$$

where b_c is category-boundary parameter.

Item Response Theory Models

- The two models are equivalent for $m = 2$; however, their item parameters are not directly comparable for $m > 2$.
- GPCM has been used to score responses to items in cognitive tests.
- GRM has been favored for fitting rating scale responses (e.g. Likert-type data).
- For practical applications, the choice between the two models has been showed to be rather inconsequential.

Fisher's Information for Item Parameters

For a polytomous item with m categories, the information matrix is given as

$$I(\eta;\theta) = \begin{bmatrix} I_{aa} & \cdots & I_{ab_{m-1}} \\ \vdots & \ddots & \vdots \\ I_{b_{m-1}a} & \cdots & I_{b_{m-1}b_{m-1}} \end{bmatrix}$$

where $\eta = (a, b_1, \ldots, b_{m-1})$, and each element of the matrix is defined as

$$I_{\eta_i\eta_j} = E\left[-\frac{\partial^2}{\partial\eta_i\partial\eta_j}\text{logLikelihood}\right]$$

Fisher's Information for Item Parameters

For GPCM, the detailed formula can be written as

$$I_{aa} = \sum_{c=0}^{m-1} P_c \cdot T_c^2 - \left[\sum_{c=0}^{m-1} T_c \cdot P_c\right]^2$$

where $T_c = \sum_{v=0}^{c}(\theta - b_v)$. For $k = 1, \ldots, m-1$,

$$I_{ab_k} = a\left\{\left(\sum_{c=k}^{m-1} P_c\right) \cdot \left(\sum_{c=0}^{k-1} P_c \cdot T_c\right) - \left(\sum_{c=0}^{k-1} P_c\right) \cdot \left(\sum_{c=k}^{m-1} P_c \cdot T_c\right)\right\}$$

Fisher's Information for Item Parameters

For $k = 1, \ldots, m-1$,

$$I_{b_k b_k} = a^2 \left(\sum_{c=0}^{k-1} P_c \right) \cdot \left(\sum_{c=k}^{m-1} P_c \right)$$

For $k < l$,

$$I_{b_k b_l} = a^2 \left(\sum_{c=0}^{k-1} P_c \right) \cdot \left(\sum_{c=l}^{m-1} P_c \right)$$

and

$$I_{b_l b_k} = I_{b_k b_l}$$

Fisher's Information for Item Parameters

- The formulas are for one item and one test taker. For multiple items and test takers, the information equals the sum across all items and all test takers.

- For GRM, there are similar formulas which also have clean and closed forms.

Fisher's Information for Item Parameters

The Bayesian version of the Information matrix is defined as

$$I_B = \int I(\eta;\theta) \cdot f(\eta) \cdot f(\theta) d\eta d\theta$$

where the $f(\eta)$ and $f(\theta)$ are posterior distribution of η and θ. It can be calculated using the draws from the posterior distribution of θs and the draws from the updated posterior distribution of η as

$$I_B \approx (ST)^{-1} \sum_{s=1}^{S} \sum_{t=1}^{T} I\left(\eta^{(t)};\theta^{(s)}\right)$$

Assignment of Field Test Items

Suppose the field test items have already been updated (*b-1*) times, and another test taker *j* is ready to take one field test item, then one of the available field test items is selected according to the following optimality criteria.

- D-optimality

$$D = \det\left(I^{(b-1)} + I_j\right) - \det\left(I^{(b-1)}\right)$$

- A-optimality

$$A = \text{trace}\left[\left(I^{(b-1)}\right)^{-1} - \left(I^{(b-1)} + I_j\right)\right]^{-1}$$

- T-optimality

$$T = \text{trace}\left(I^{(b-1)} + I_j\right) - \text{trace}\left(I^{(b-1)}\right)$$

Assignment of Field Test Items

- E-optimality

$$E = \max \text{Eigen}\left[\left(I^{(b-1)}\right)^{-1}\right] - \max \text{Eigen}\left[\left(I^{(b-1)} + I_j\right)^{-1}\right]$$

- W-optimality

$$W = \det\left[\left(I^{(b-1)}\right)_b^{-1}\right] - \det\left[\left(I^{(b-1)} + I_j\right)_b^{-1}\right]$$

where the subscript "b" means only focus on parameter bs.

- Random

Item Parameter Update

After a batch of responses for one field test item is collected, the field test item parameters are updated once. The similar algorithm used in van der Linden & Ren (2015) was used. It is a modified Gibbs algorithm and includes two main steps:

1. Sampling the posterior distributions of θs.
   - Resampling the draws saved after the CAT with operational items.
2. Sampling the conditional posterior distribution of item parameters
   - Metropolis-Hastings step to generate one sample of item parameters.

A vector of draws sampled from the stationary part of the Markov chain for item parameters is saved for future field test item assignment and parameter update.

Simulation Study

Adaptive tests were simulated from a pool of 150 operational items and 50 field-test items. The adaptive test was organized as follow:

- Test length was 15.
- Maximum information criterion was used for item selection.
- EAP was used for ability estimation.
- The simulee’s true ability was generated from standard normal distribution.

Simulation Study

- Field-test items were assigned either after the 5^{th} or 15^{th} operational item.
- The draws from the posterior distribution of θ after the 5th / 15th operational item were used for the field-test item selection.
- The field-test parameters were updated after batches of 20, 50, and 100 responses, using the draws of θ at the end of the test.
- A field-test item was retired after the collection of 500 or 1000 responses. The calibrated item was then replaced with a new field-test item.
- The whole simulation was stopped when 50 field-test items had finished their calibration.

Simulation Study

Observe that

- The field-test pool always has 50 active field-test items.
- Thus, there always were 50 items competing with each other to meet the optimality criterion.

Table 1: Average Bias of Field-Test Parameters (Calibration Size: 1000)

Item Position		5th			15th		
Batch Size		20	50	100	20	50	100
a	D	-0.145	-0.164	-0.159	-0.215	-0.168	-0.171
	A	-0.173	-0.157	-0.166	-0.137	-0.161	-0.149
	T	-0.160	-0.199	-0.191	-0.158	-0.186	-0.208
	E	-0.143	-0.115	-0.158	-0.147	-0.138	-0.162
	W	-0.148	-0.123	-0.140	-0.153	-0.136	-0.146
	R	-0.100	-0.147	-0.140	-0.120	-0.143	-0.128
b	D	-0.057	0.009	0.011	-0.039	-0.004	-0.005
	A	-0.127	-0.046	-0.072	-0.118	-0.013	0.001
	T	-0.047	-0.318	-0.100	-0.088	-0.419	-0.092
	E	0.208	-0.008	-0.046	-0.066	-0.070	-0.076
	W	-0.117	-0.013	-0.027	-0.254	-0.044	-0.013
	R	-0.065	-0.062	-0.032	-0.021	-0.029	-0.045

Table 1: Average Bias of Field-Test Parameters (Calibration Size: 500)

Item Position		5th			15th		
Batch Size		20	50	100	20	50	100
a	D	-0.129	-0.159	-0.106	-0.075	-0.130	-0.115
	A	-0.153	-0.167	-0.170	-0.136	-0.151	-0.143
	T	-0.194	-0.203	-0.203	-0.149	-0.187	-0.207
	E	-0.164	-0.181	-0.161	-0.147	-0.186	-0.172
	W	-0.147	-0.145	-0.137	-0.116	-0.164	-0.162
	R	-0.071	-0.109	-0.123	-0.130	-0.128	-0.119
b	D	0.020	0.007	-0.010	-0.019	-0.018	-0.019
	A	-0.052	-0.081	-0.026	-0.055	-0.041	-0.032
	T	-0.144	-0.019	-0.068	-0.010	-0.087	-0.026
	E	-0.072	-0.063	-0.042	-0.067	-0.107	-0.044
	W	0.040	-0.097	-0.026	0.876	-0.122	-0.063
	R	-0.066	-0.052	-0.010	-0.052	-0.033	-0.039

Table 2: Average RMSE of Field-Test Parameters (Calibration Size: 1000)

Item Position		5th			15th		
Batch Size		20	50	100	20	50	100
a	D	0.232	0.224	0.193	0.260	0.226	0.194
	A	0.229	0.195	0.187	0.204	0.197	0.185
	T	0.231	0.238	0.231	0.226	0.228	0.246
	E	0.204	0.168	0.186	0.213	0.180	0.191
	W	0.204	0.180	0.167	0.211	0.181	0.189
	R	0.217	0.174	0.169	0.222	0.176	0.171
b	D	0.594	0.561	0.258	0.571	0.384	0.303
	A	1.075	0.390	0.433	0.872	0.715	0.327
	T	0.805	2.467	0.551	1.558	3.909	0.668
	E	2.858	0.520	0.405	0.947	0.474	0.394
	W	1.263	0.730	0.329	0.971	0.612	0.568
	R	0.838	0.439	0.248	0.867	0.427	0.242

Table 2: Average RMSE of Field-Test Parameters (Calibration Size: 500)

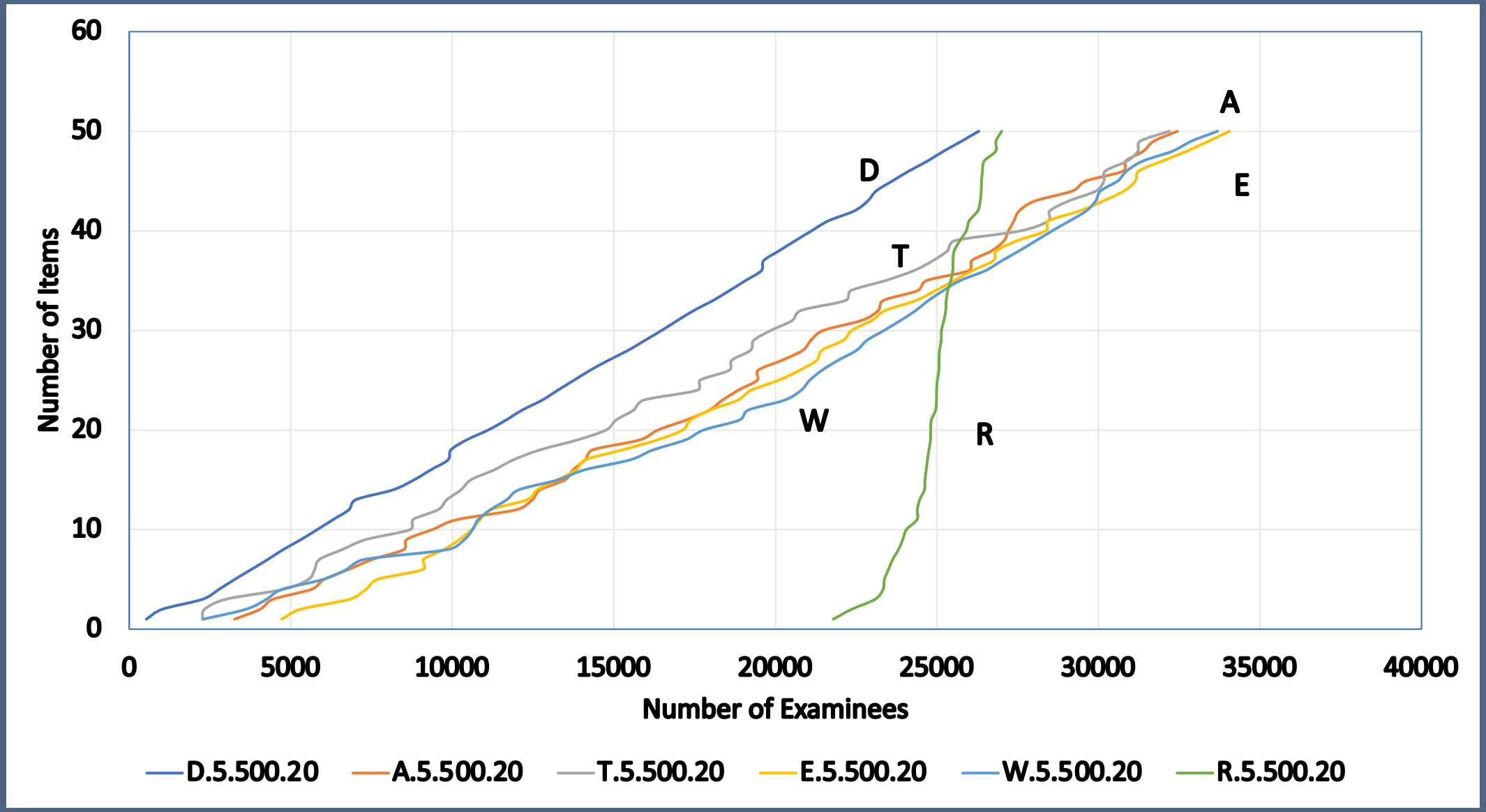
Item Position		5^{th}			15^{th}		
Batch Size		20	50	100	20	50	100
a	D	0.216	0.208	0.148	0.233	0.177	0.152
	A	0.228	0.207	0.203	0.219	0.189	0.183
	T	0.273	0.243	0.241	0.232	0.237	0.263
	E	0.219	0.216	0.188	0.213	0.214	0.210
	W	0.221	0.184	0.173	0.190	0.195	0.198
	R	0.163	0.162	0.166	0.222	0.172	0.153
b	D	0.490	0.283	0.279	0.448	0.312	0.324
	A	1.126	0.591	0.336	0.847	0.529	0.324
	T	1.192	0.898	0.541	1.121	0.747	0.534
	E	1.162	0.662	0.387	1.114	0.619	0.343
	W	1.295	0.473	0.318	10.549	0.665	0.380
	R	0.815	0.404	0.229	0.705	0.398	0.274

Table 3: Average Posterior SD of Field-Test Parameters (Calibration Size: 1000)

Item Position		5th			15th		
Batch Size		20	50	100	20	50	100
a	D	0.247	0.156	0.111	0.232	0.154	0.110
	A	0.197	0.136	0.097	0.198	0.135	0.099
	T	0.207	0.137	0.111	0.204	0.141	0.111
	E	0.201	0.143	0.100	0.196	0.137	0.097
	W	0.197	0.149	0.107	0.197	0.147	0.108
	R	0.237	0.151	0.108	0.233	0.150	0.108
b	D	0.547	0.374	0.253	0.577	0.364	0.262
	A	0.806	0.432	0.321	0.787	0.485	0.305
	T	0.836	0.703	0.461	0.938	0.770	0.436
	E	0.941	0.445	0.319	0.797	0.442	0.322
	W	0.930	0.492	0.312	0.852	0.489	0.343
	R	0.664	0.387	0.261	0.682	0.383	0.254

Table 3: Average Posterior SD of Field-Test Parameters (Calibration Size: 500)

Item Position		5th			15th		
Batch Size		20	50	100	20	50	100
a	D	0.253	0.156	0.117	0.263	0.160	0.115
	A	0.200	0.137	0.103	0.205	0.136	0.104
	T	0.196	0.152	0.113	0.218	0.146	0.115
	E	0.200	0.131	0.100	0.198	0.130	0.103
	W	0.198	0.145	0.109	0.212	0.137	0.107
	R	0.235	0.154	0.112	0.228	0.151	0.112
b	D	0.556	0.356	0.249	0.534	0.352	0.257
	A	0.826	0.459	0.317	0.758	0.461	0.315
	T	0.857	0.672	0.439	0.866	0.613	0.439
	E	0.875	0.500	0.318	0.821	0.489	0.326
	W	0.935	0.464	0.323	1.392	0.496	0.325
	R	0.661	0.386	0.255	0.646	0.379	0.256


60
50
40
30
20
10
0
Number of Items
0
5000
10000
15000
20000
25000
30000
35000
40000
Number of Examinees
D
A
E
T
W
R
D.5.500.20
A.5.500.20
T.5.500.20
E.5.500.20
W.5.500.20
R.5.500.20

Discussion

- Large batch size helped reducing the average RMSE and posterior SD of parameter b.
- Reducing sample size from 1000 to 500 did not have significant impact on the calibration results.
- Under the Bayesian framework, assigning the field test items at early stage of operational CAT did not have significant impact on the calibration results compared to assign the field test item at the end of operational CAT.
  - A lack of effect most likely due to the fact that the posterior sample of the ability parameters at the end of the test were used for the calibration.
- The negative bias of *a* parameter calibration result may indicate that the standard normal distribution as true ability/trait distribution did not provide enough discrimination information for polytomous items given the width of their information functions.

Discussion

- Among the criteria, D-optimality and Random had similar performance on the calibration results of *b* parameter in terms of RMSE and average posterior SD;
- As demonstrated by the figure, optimal design resulted in early calibration of some of the more informative items while with random assignment we had to wait for the first calibrated item until the end of the study.
- Optimal design can be further embedded in the shadow-test approach to provide overall control across different aspects of a test.

Discussion

- This study showed the feasibility of the implementation for online item calibration with polytomous items and the proposed Bayesian framework.
  - Computation time for item selection (field test item assignment) is low enough for application in real-world adaptive testing.
- Computation time for item parameter update is about 2~3 second.
  - The configuration in this study was very conservative. The performance can be tuned if real-time online update is desired.
- Periodical monitoring is recommended for the cases of non-convergence or extreme values.

References

- van der Linden, W. J., & Ren, H. (2015). Optimal Bayesian Adaptive Design for Test-Item Calibration. *Psychometrika*, *80*(2), 263–288. https://doi.org/10.1007/s11336-013-9391-8
- Ren, H., van der Linden, W. J., & Diao, Q. (2017). Continuous Online Item Calibration: Parameter Recovery and Item Utilization. *Psychometrika*, *82*(2), 498–522. https://doi.org/10.1007/s11336-017-9553-1
- van der Linden, W. J., & Ren, H. (2020). A Fast and Simple Algorithm for Bayesian Adaptive Testing. Journal of Educational and Behavioral Statistics, 45(1), 58–85. https://doi.org/10.3102/1076998619858970
- Ren, H., Choi, S. W., & van der Linden, W. J. (2020). Bayesian adaptive testing with polytomous items. *Behaviormetrika*, *47*(2), 427-449. https://doi.org/10.1007/s41237-020-00114-8
- van der Linden, W.J., & Jiang, B. (2020). A Shadow-Test Approach to Adaptive Item Calibration. *Psychometrika* 85, 301–321. https://doi.org/10.1007/s11336-020-09703-8

Thanks!