Listening. Learning. Leading.®

Understanding and Detecting Data Fabrication in Large-Scale Assessments

Kentaro Yamamoto

ETS

11 May 2017, Paris, France

Impact of Data Fabrication

- Undermines reliability, validity and comparability of the survey results among subpopulations and cycles.
- Destroys the linkage between assessment objectives and data.

Stake Holders

- **Respondents** who wish a score as high as possible are the stake holders for any **individual entrance exam**.
- **Many stake holders** in large scale assessments or ”low stake tests” such as **PIAAC, PISA, TIMSS, NAEP**: sampling contractor, respondent, interviewer/proctor, administrative contractor, coder, data aggregator, data processor, and so on.

Respondents

- Omitted responses or early abandonment of survey
- Random responses (guessing, response styles)
- Intentional erroneous responses

Interviewer

- Fake interview
- Shortened data collection
- Replicating data
- Multiple cases together
- Particular responses
- Increased non-interview

Data Collection Contractor

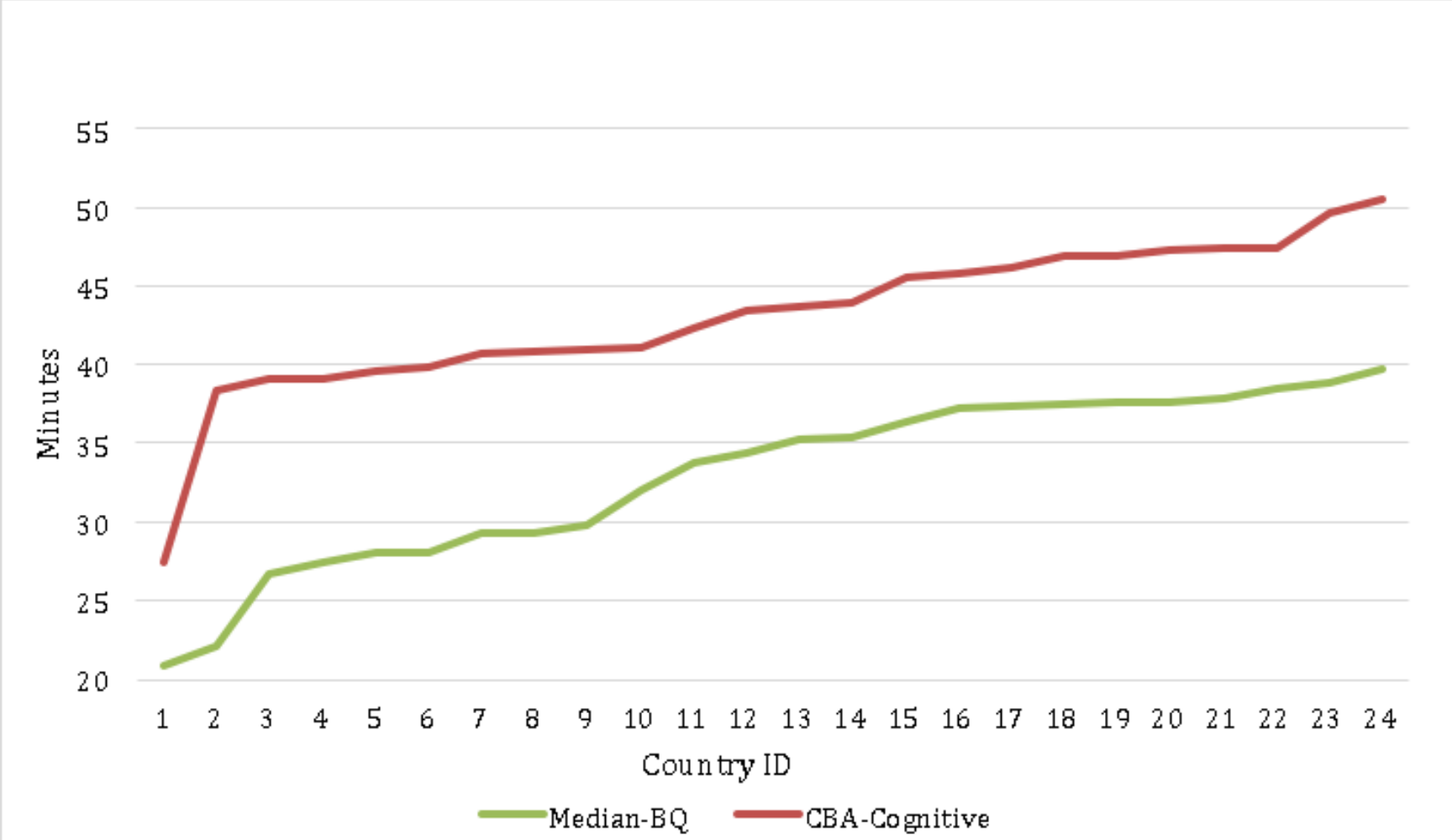
- Duplication of data
- Synthetic data
- Dropping of data
- Insufficient sampling information

Coder

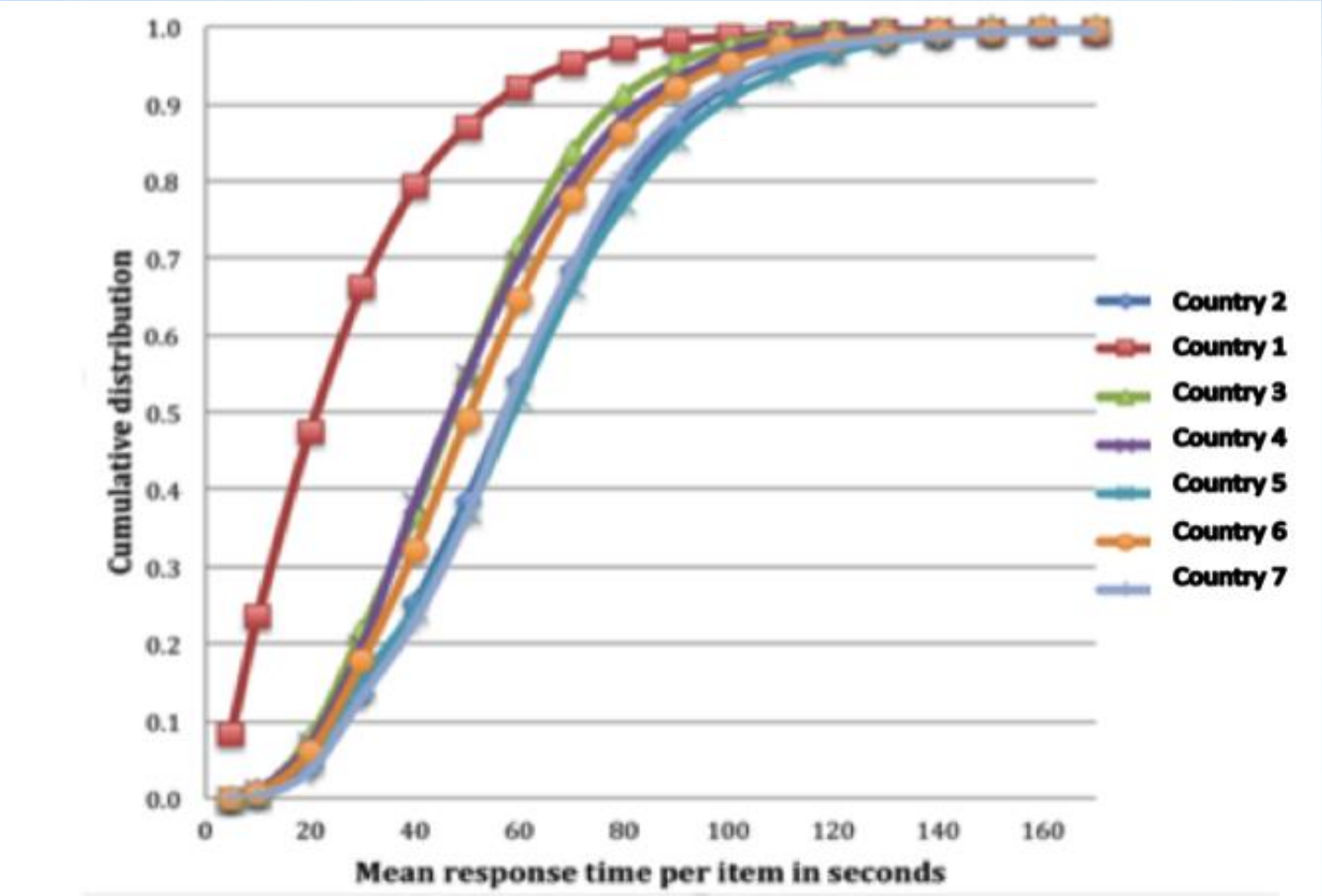
- Inaccurate coding
  - Random coding
  - Patterned coding
- Duplicate coding

Case 1: PIAAC

Median Time of BQ and Cognitive Modules by Country



Cumulative distribution of mean response time



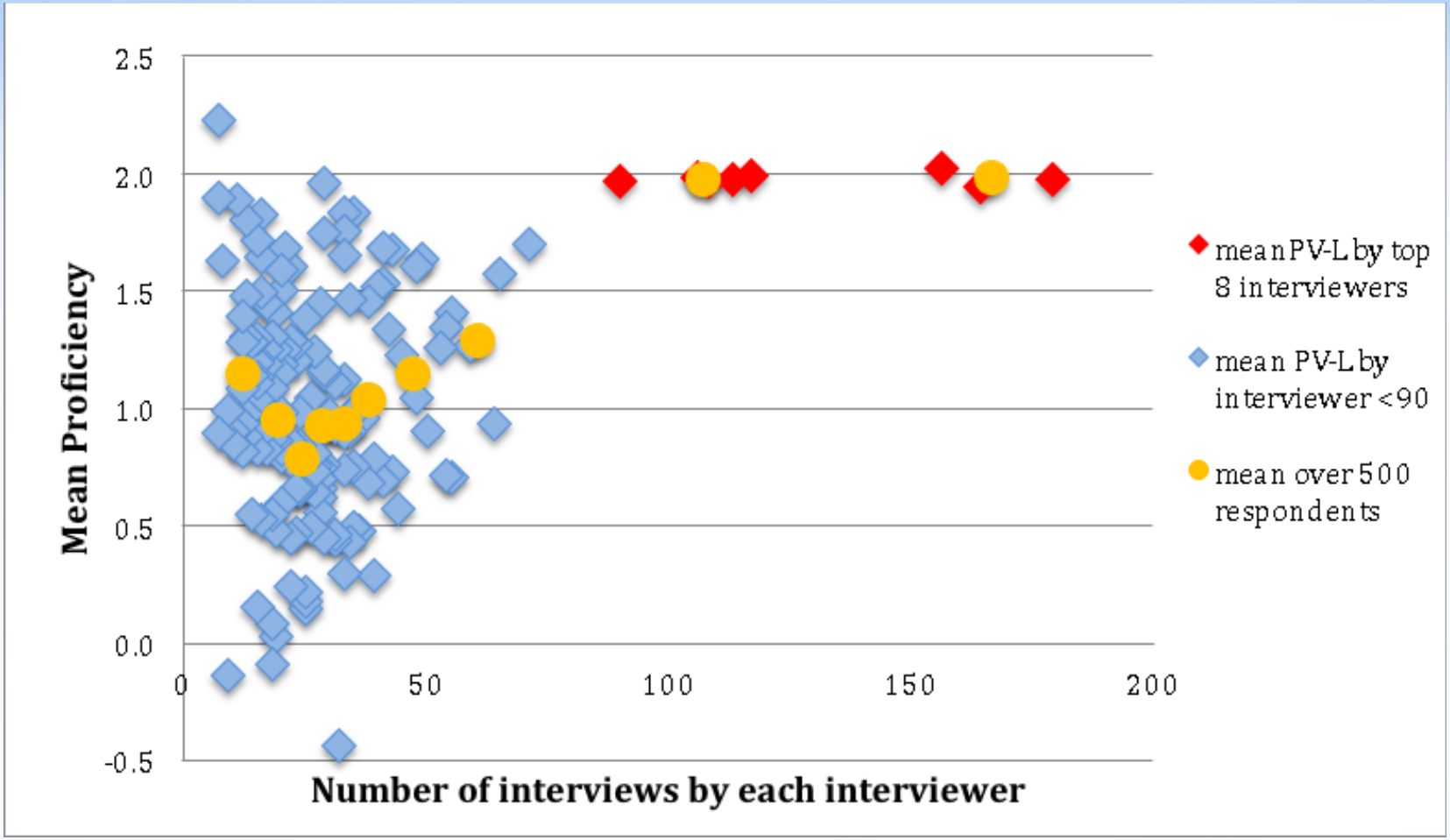
Number of items unrelated (slope is less than 0.03) to the underlying proficiency for Countries 1 and 8 in Language A

	Country 1	Country 8
Literacy	**6**	1
Numeracy	**7**	0
Problem Solving	**0**	0

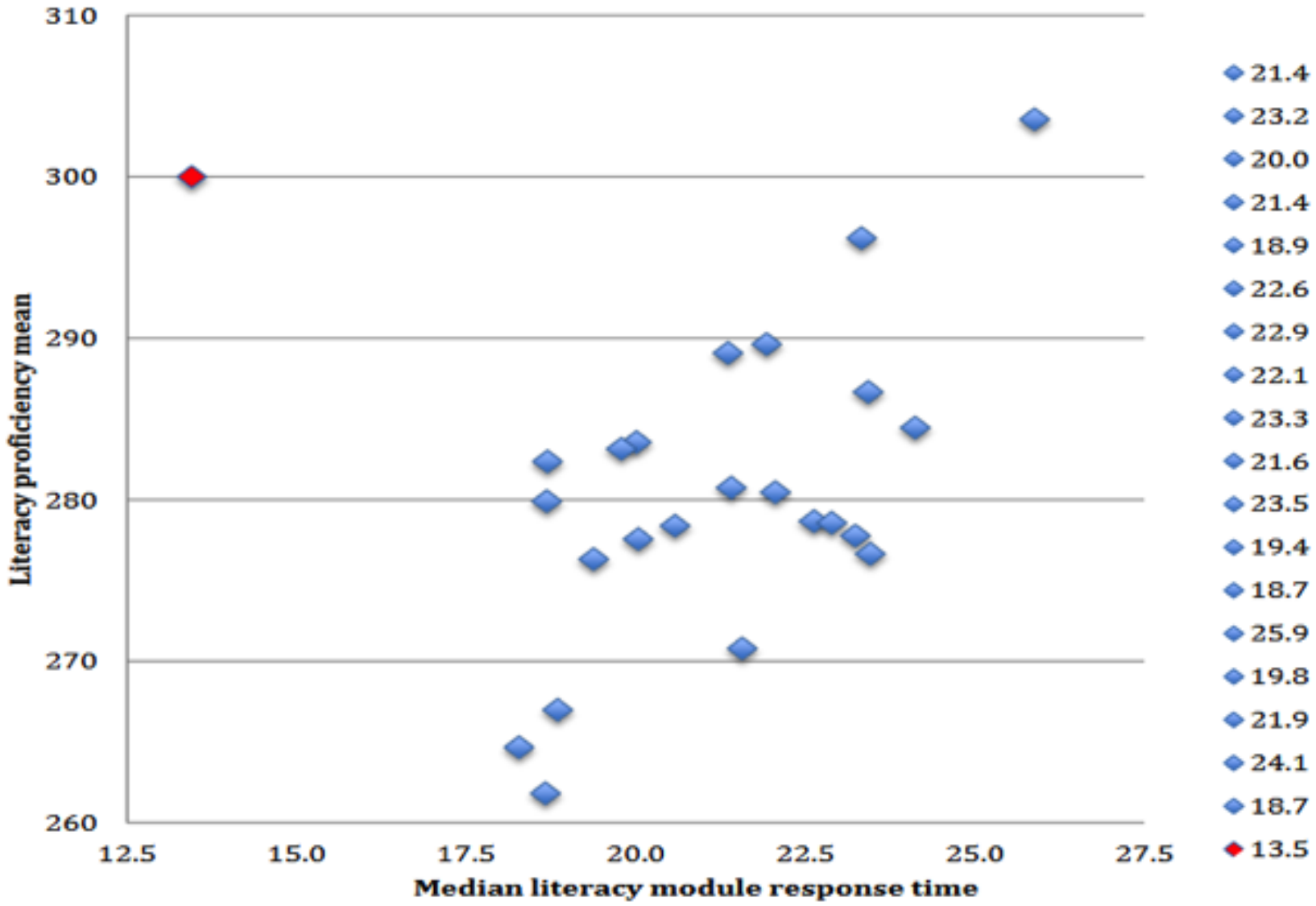
Number of items with significant deviation from international parameters in Countries 1 and 8 in Language A

	Country 1	Country 8
Literacy	**14**	6
Numeracy	**17**	3
Problem solving	**3**	0
Total	**34**	9

Mean literacy proficiency of respondents by interviewer



Mean CBA Literacy Proficiency and Median Response Time to Literacy Module by Country



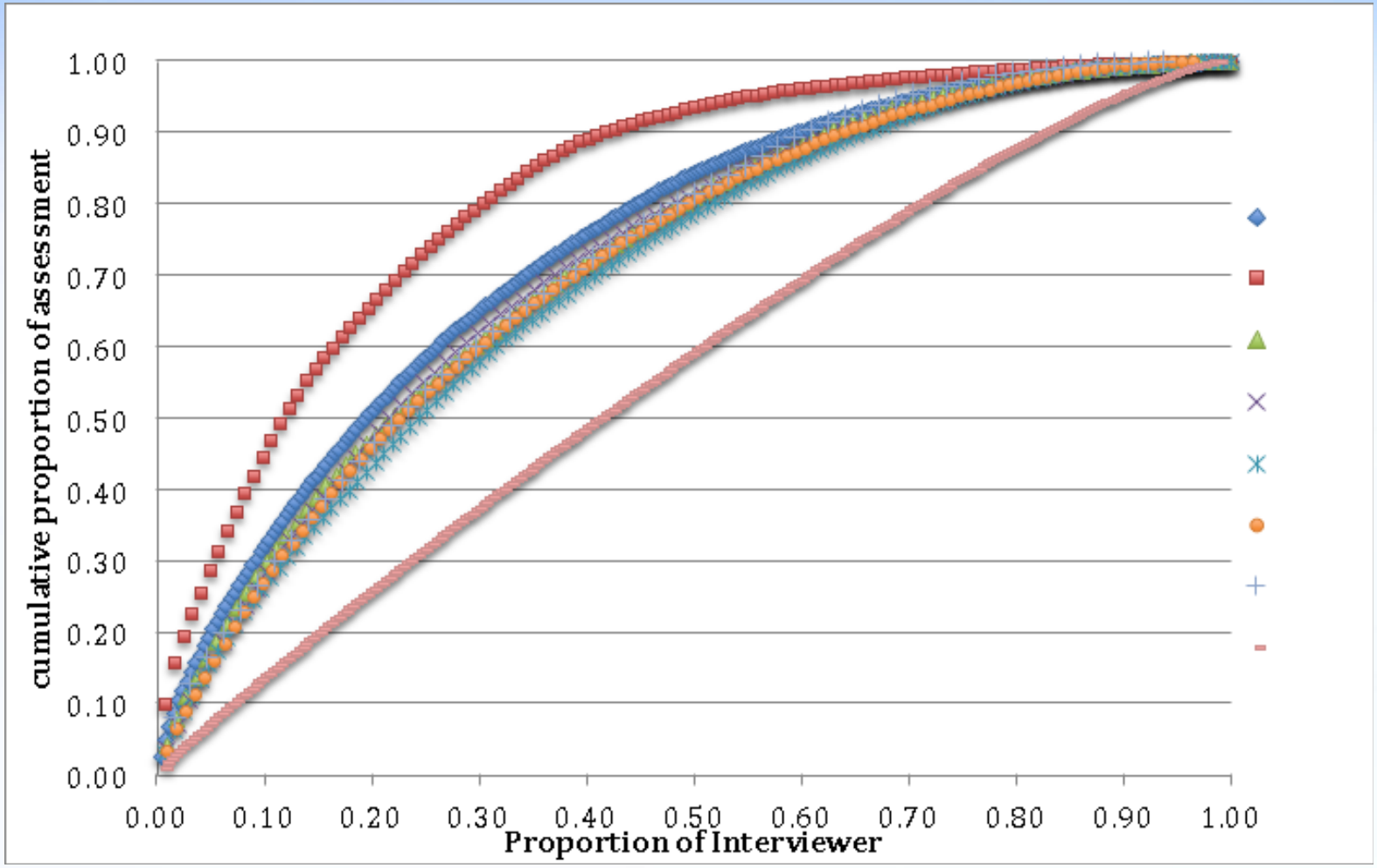
Summary

- Duplicated data
- Short response time
- Lack of variance of performance data within interviewer
- Geographically localized anomaly
- Incongruity of responses within a respondent
- Lack of overall comparability of data within a country
- Lack of comparability of data across countries

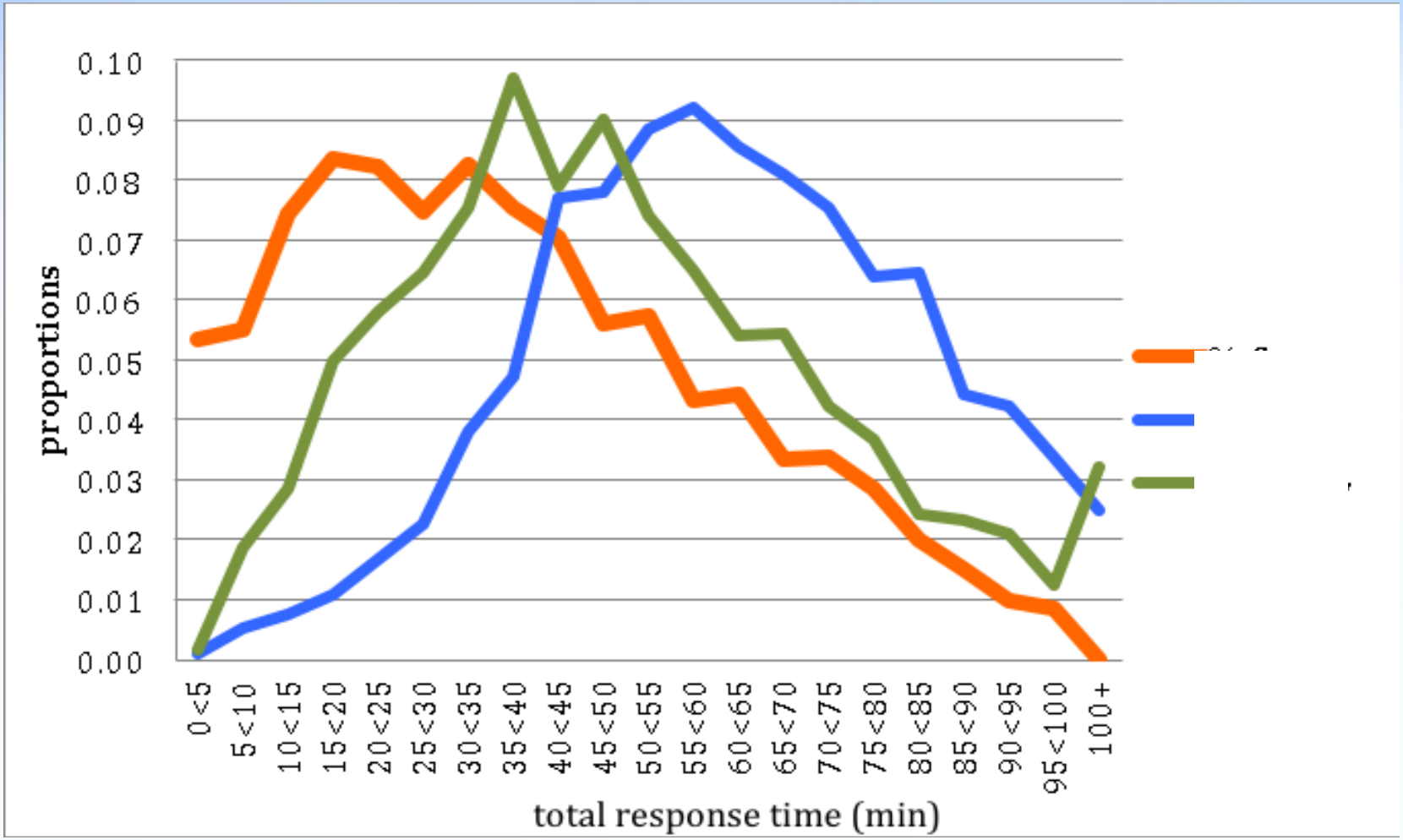
Resulted in elimination of data from entire region, respondents with average response time less than 10 seconds, and all duplicated cases. Altogether 1220 cases were eliminated.

Case 2: PIAAC

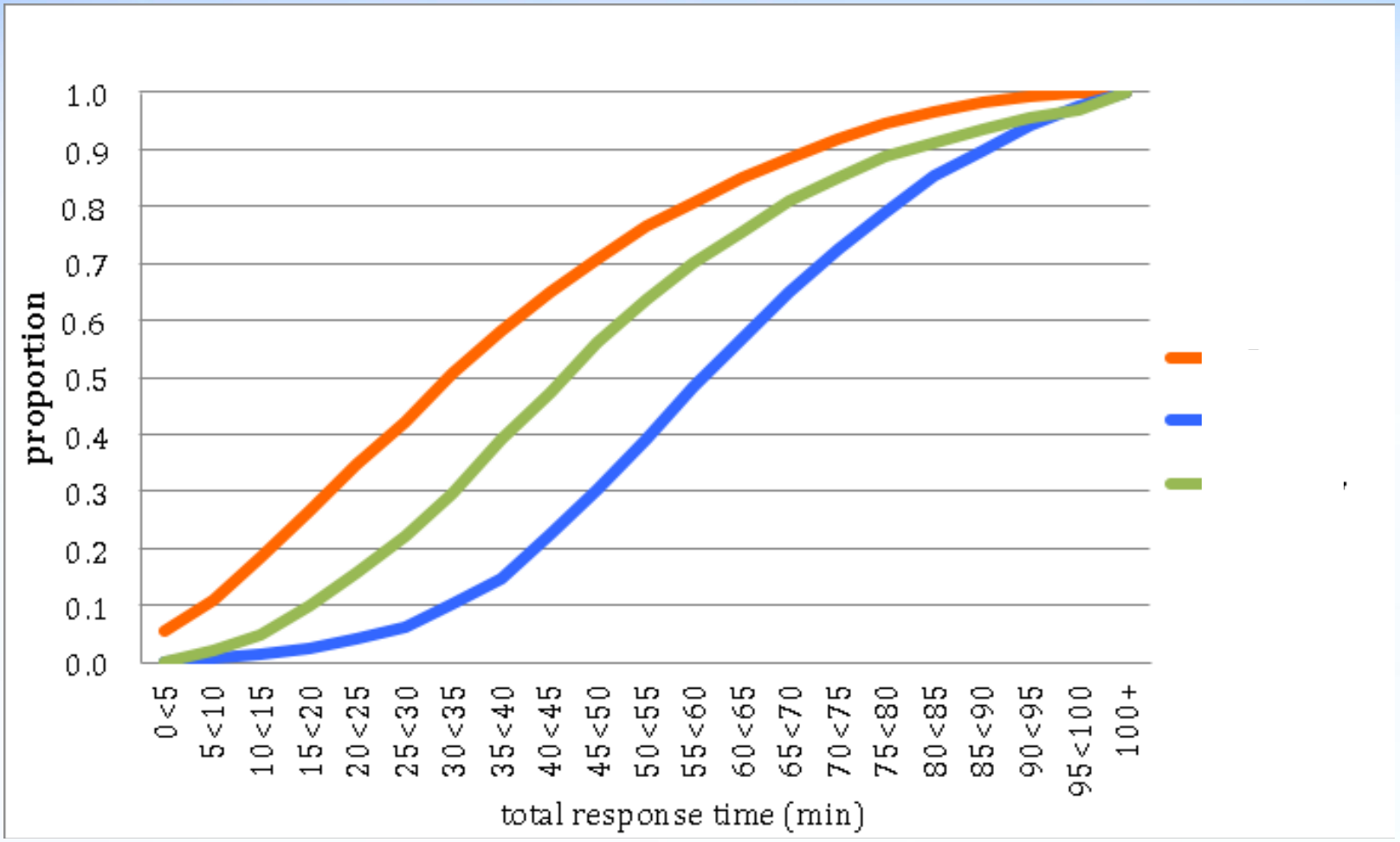
Cumulative Proportion of Assessment by Interviewer



Distribution of respondents based on the total time on CBA cognitive items (min)



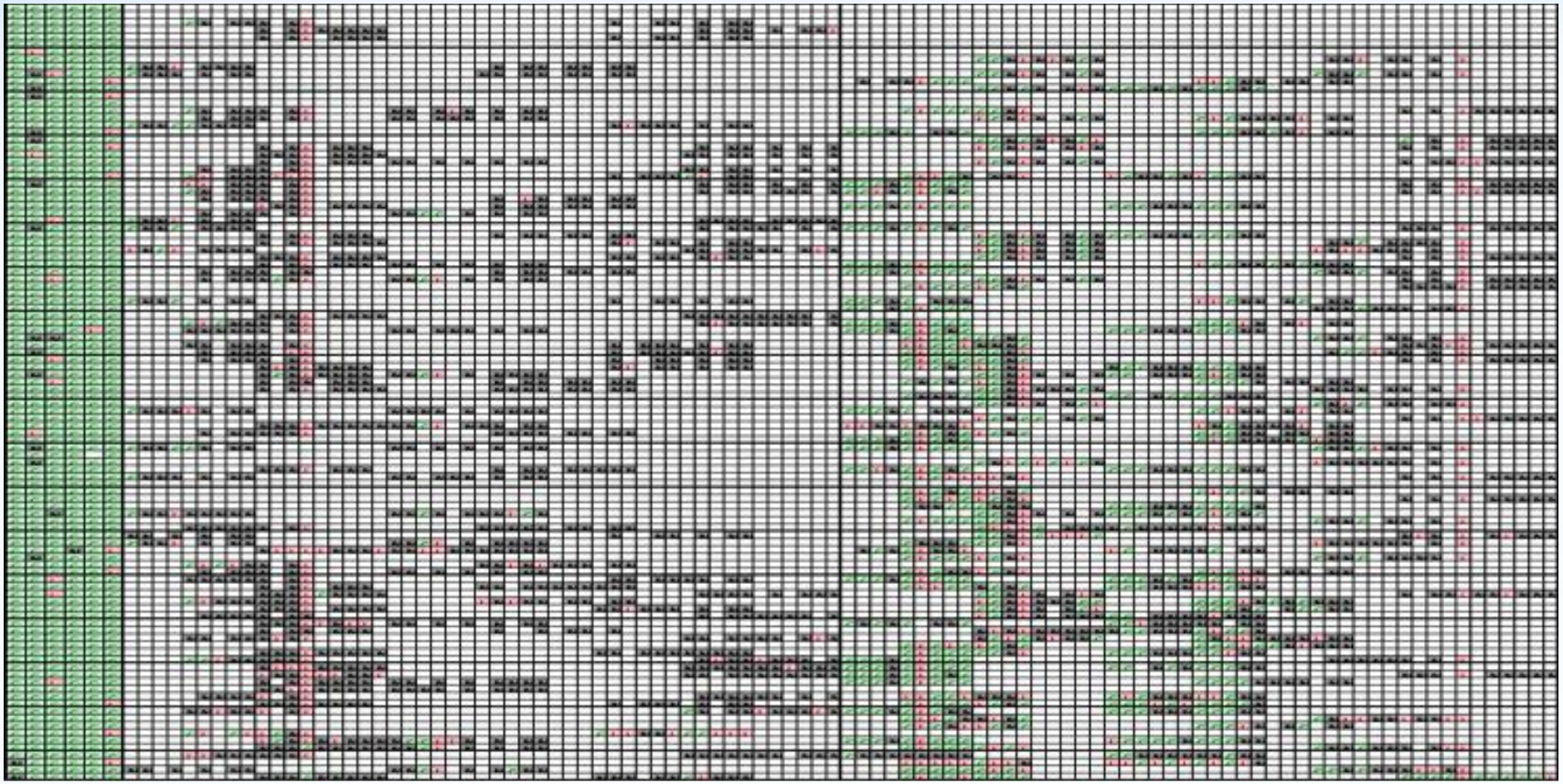
Cumulative distribution of respondents based on the total time on CBA cognitive items (min)



144 Responses collected by the Interviewer #92

Core | Literacy | Numeracy

144 respondents



Summary

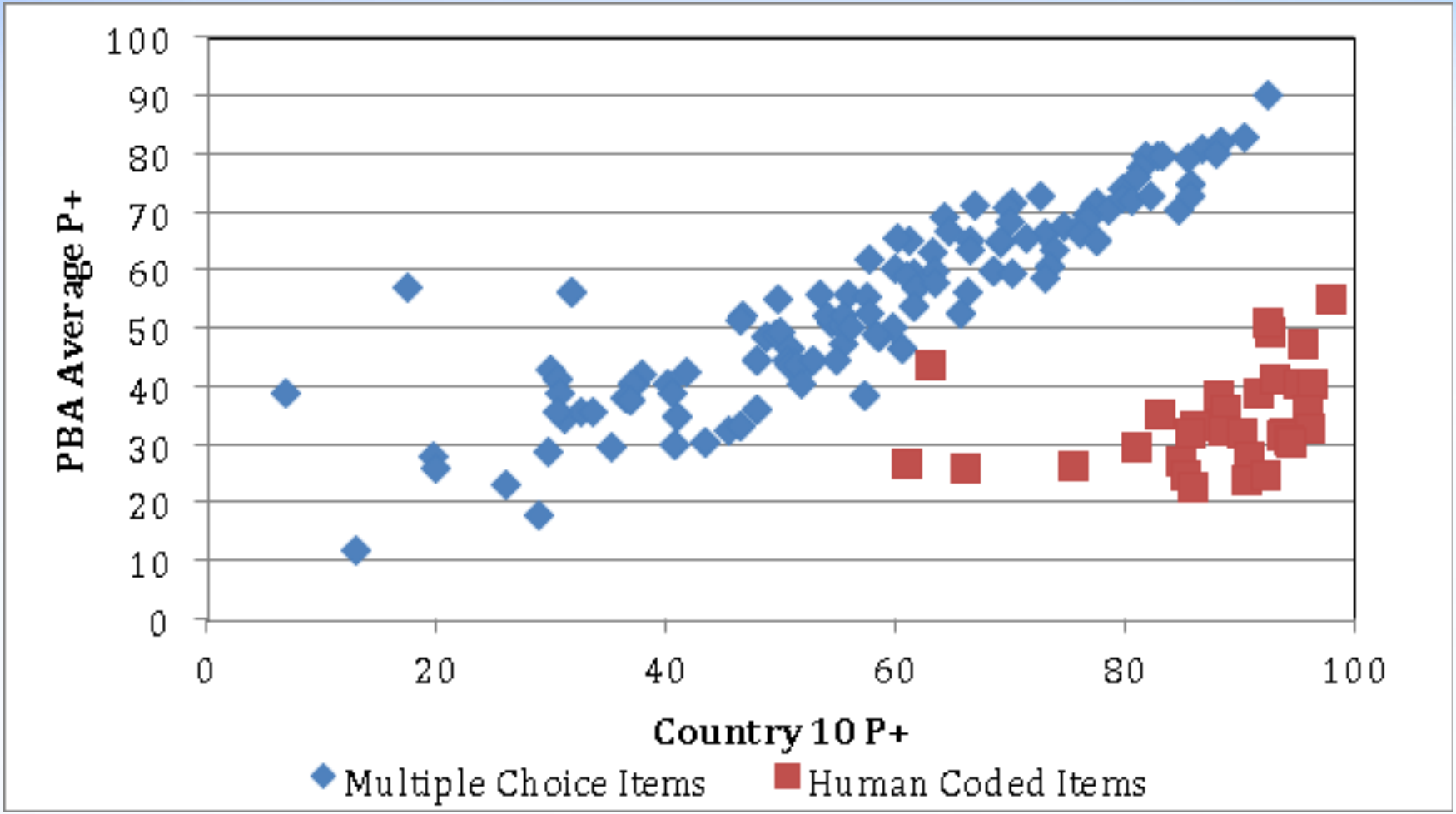
- 20% of data was collected by 6 interviewers
- Short response time interacted with interviewer ID
- Lack of variance of performance data within interviewer
- Large number of non-response within a few interviewers
- Incongruity of responses across domains within a respondent
- Lack of overall comparability of data within a country
- Lack of comparability of data across countries
- BQ data did not show anomaly

Resulted in elimination of cognitive responses of 1042 cases collected by 7 interviewers.

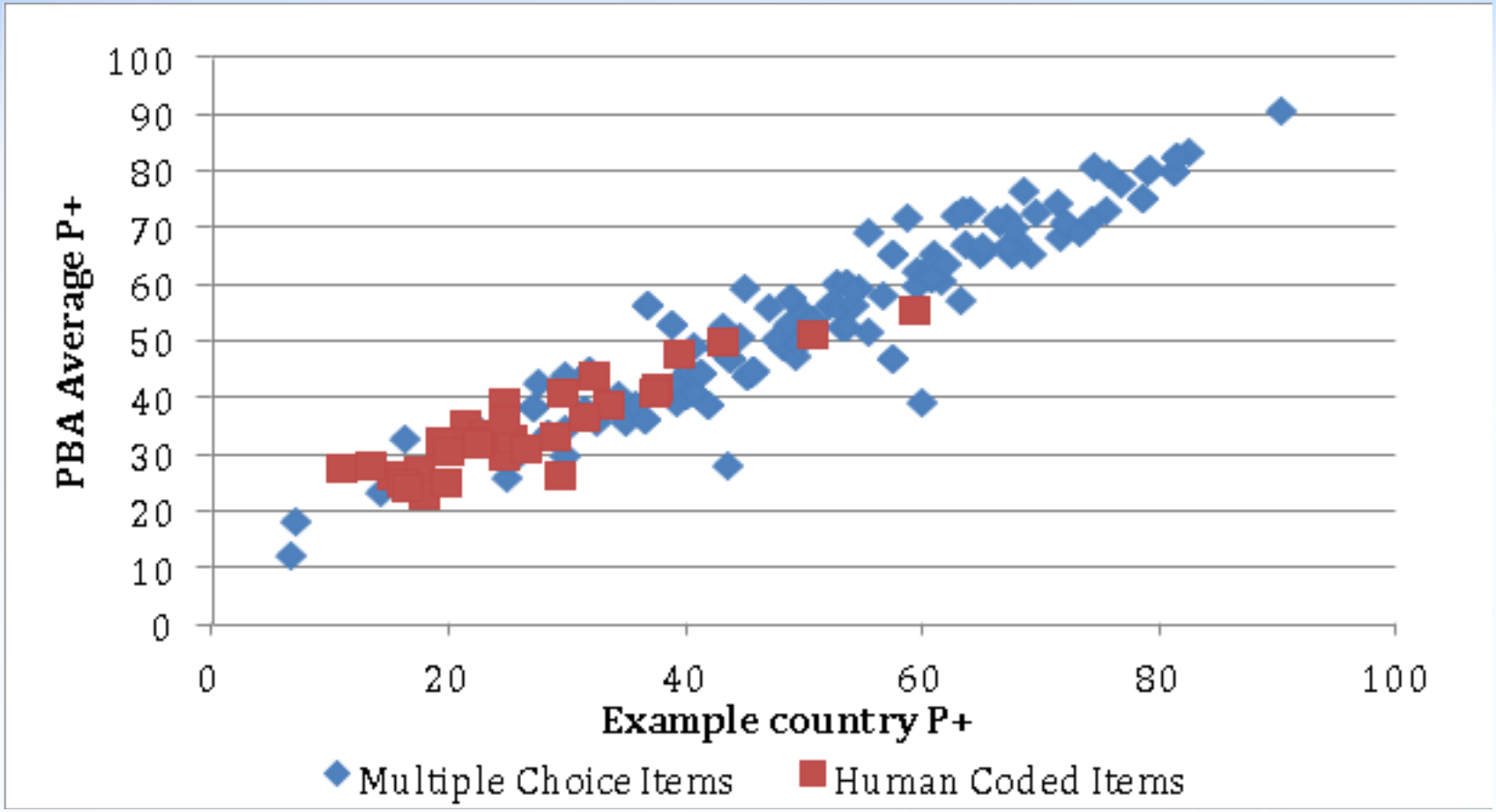
Case 3: PISA

- As of 2015 PISA is administered primarily as a computer-based assessment, all previous PISA data is paper-based
- A few countries chose a paper-based assessment in 2015
- Case 3 administered PBA in two languages
- About 35% to 45% of items require human coding

Item difficulties for Country 10 compared to the average P+ across PBA countries



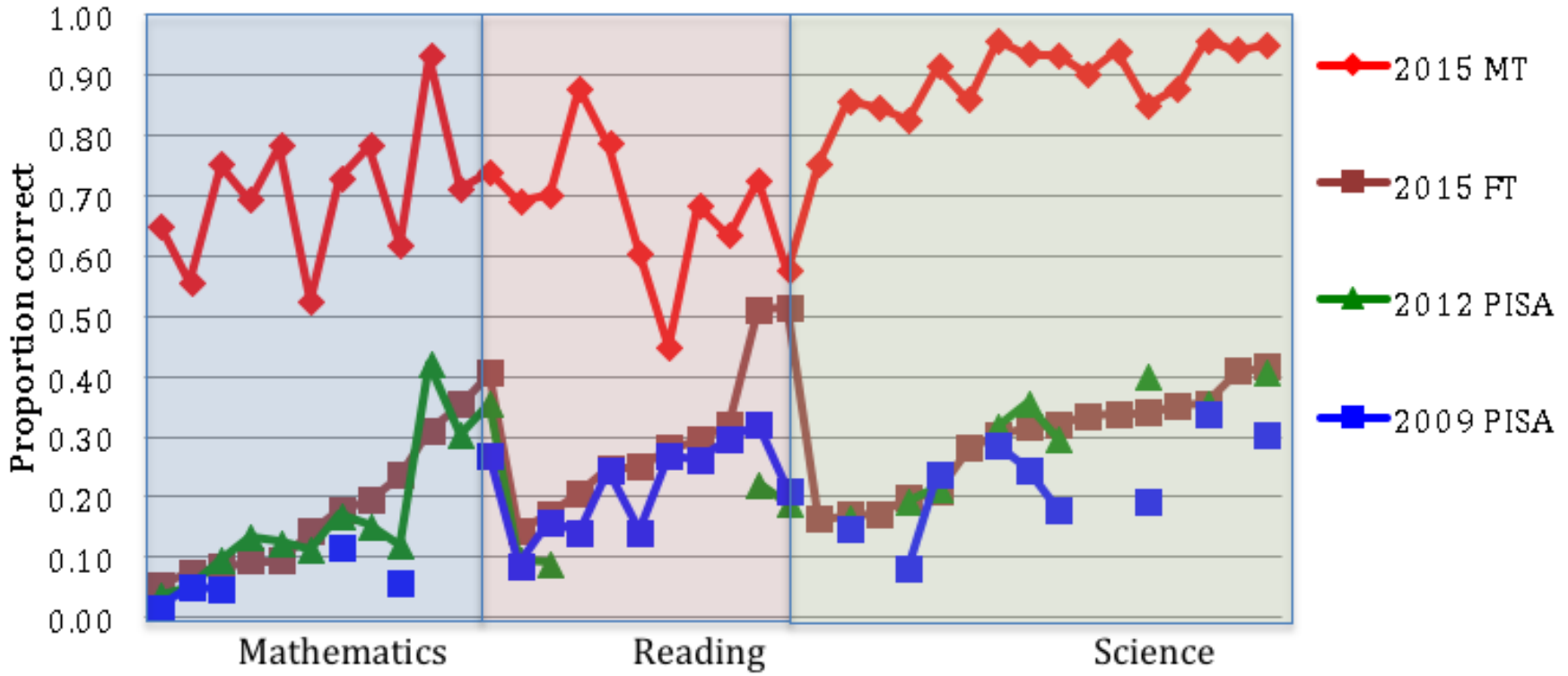
Item difficulties for a country compared to the average P+ across PBA countries



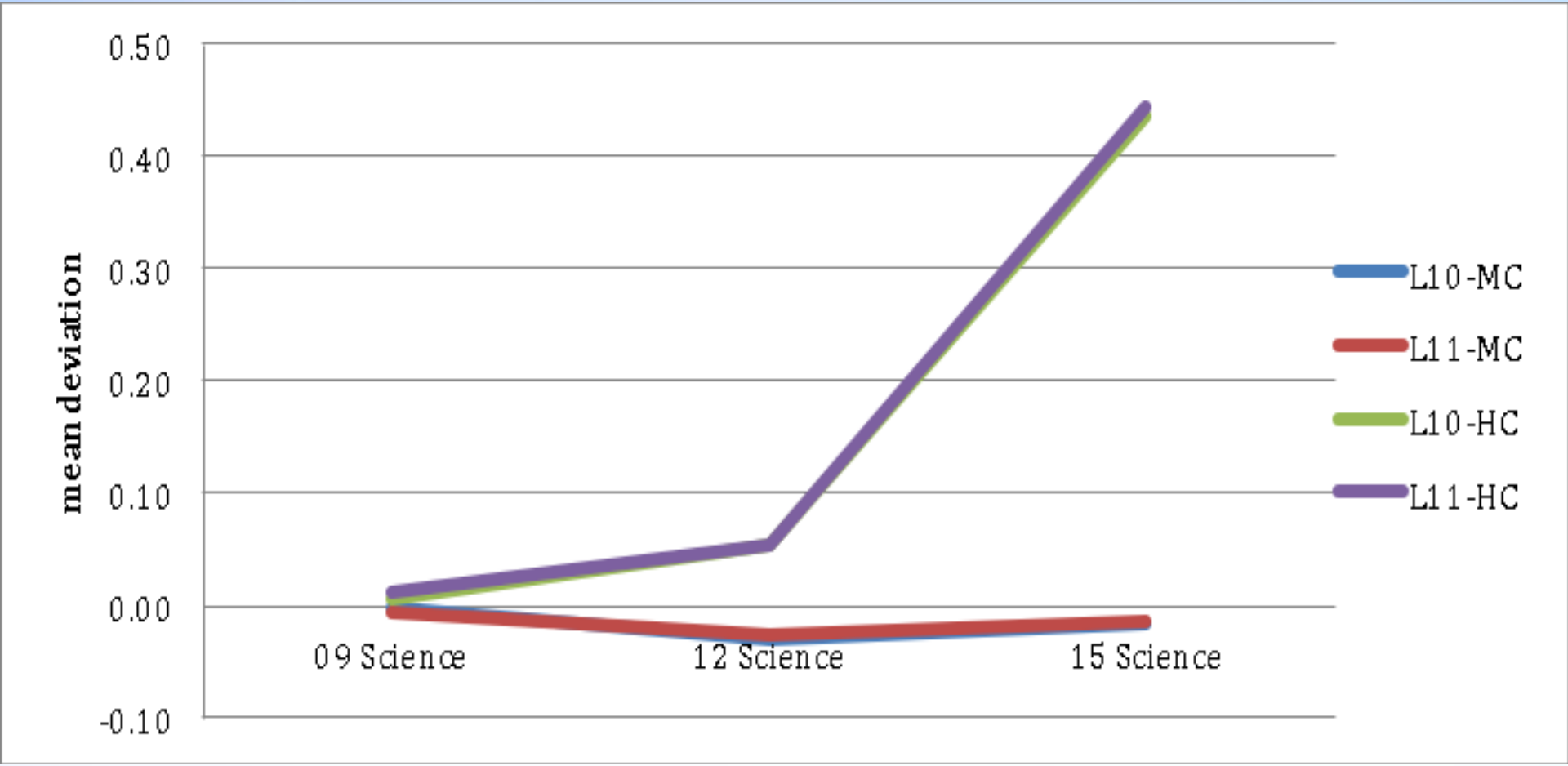
Human-coded vs. multiple-choice response statistics in Country 10 vs. other PBA countries

	Average proportion correct (P+) across items by domain and type		Average correlation (R) across items by domain and type	
	Country 10	Mean of other PBA countries	Country 10	Mean of other PBA countries
Math (MC)	.56	.46	0.93	0.95
Math (HC)	.75	.26	0.53	0.95
Read (MC)	.58	.53	0.90	0.92
Read (HC)	.73	.43	0.65	0.92
Sci (MC)	.58	.55	0.90	0.92
Sci (HC)	.88	.30	0.33	0.80

Comparisons with Past Data



Mean deviation of science items from international parameters across Languages C and D in Country 10



Summary

- Proportion of correct responses (P+) on human coded items differed from multiple choice items, from other countries, from historical data within a country, in all domains
- Especially notable is the disappearance of non-response
- International IRT parameters fit very poorly for the human coded items in all domains
- Same anomaly was found in both language tests

- Evaluation of coding accuracy was conducted on selected items in randomly sampled booklets by cApStAn
- Results showed coding leniency by 25%
- Data still contains 30-40% increase of proportion of correct responses compared to the 2015 Field Trial, PISA 2012 and PISA 2009.

Result

- After all findings were considered, it was decided to delete all human-coded responses from Country 10's data and use only multiple-choice items for the analysis.
- This reduced the reliability of the survey and framework representation by items because nearly 35% of cognitive responses were eliminated from consideration in comparison to other countries.

Future of Detection of Data Fabrication

- Increased usage of log files
- Cumulative database provides better basis to compare congruency of data in multiple references
- Automatizing procedure is necessary to increase detection speed, reduce efforts and costs
- Fabrication of any part of the data reduces the credibility and validity of a survey altogether; this is every stakeholder's concern and should be eradicated.

Thank you!