Providing Validity Evidence for the Engineering Students Professional Competences Test (evidence from Russia and China)

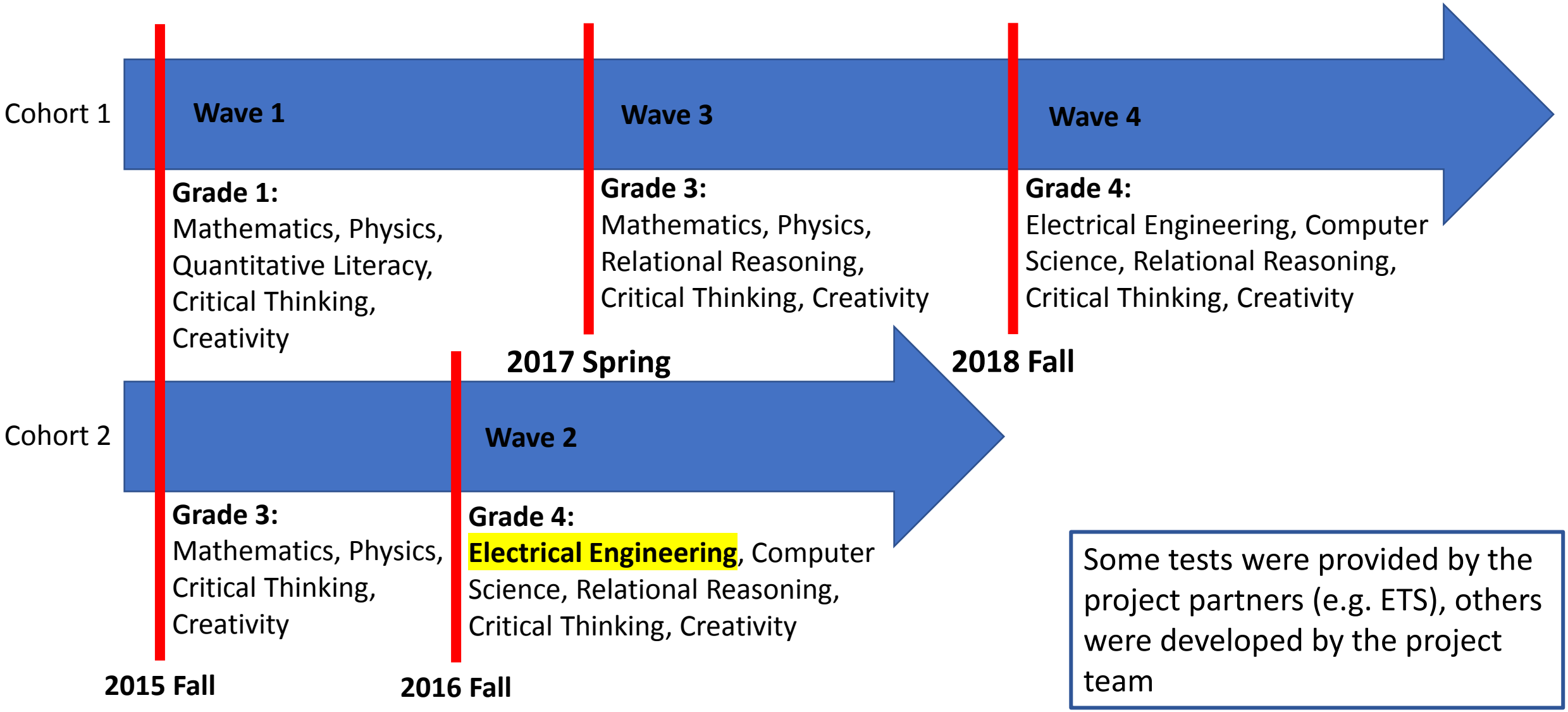
Elena Kardanova, Denis Federiakin
National Research University
Higher School of Economics, Russia
Prashant Loyalka
Stanford University , USA

ISHEL(International Study of Higher Education Learning) study

- Two main goals:
  1) Assess and compare university student skills (levels and gains) within and across countries
  2) Examine which factors help students develop skills
- Focus on engineering students (Electrical Engineering and Computer Science) in Russia and China
- Assess skills over time
  - academic skills (math, physics)
  - major-specific skills (for EE and CS majors)
  - higher order skills (critical thinking, quantitative reasoning, etc.)

*SUPER (Study for Undergraduate PERformance) project

Structure of ISHEL data gathering



Some tests were provided by the project partners (e.g. ETS), others were developed by the project team

Aim of the study

- To provide evidence regarding reliability, validity and cross-national comparability of assessment instrument that assesses and compares engineering students' professional competencies in Electrical Engineering across Russia and China for ISHEL study.

Our test development methodology

- Was based on the requirements of the Standards for Educational and Psychological Testing (AERA, APA and NCME, 2014)
- Included three stages:
  - Selection of content and sub-content areas by using expert evaluation from experts at Chinese and Russian universities
  - Collection of items for the test, which matched these content areas and verified the items based on another evaluation by local experts
  - Conducting a pilot study of a sufficient number engineering students across Russia and China and analysis of the pilot data to provide evidence that the test is reliable, cross-culturally valid, equate-able

Providing evidence in support of reliability and cross-national comparability of the EE test: stage 1

Content Areas	Mean (Russia)	Mean (China)	Mean (All)
Transistors	4.22	4.20	4.21
Basics of Digital Circuits	4.13	4.60	4.39
Fourier Analysis of Signals and Systems	4.00	4.80	4.44
Laplace and Z-transforms	4.00	4.50	4.28
Circuit Analysis	3.88	4.40	4.17
Amplifiers and Oscillators	3.88	4.60	4.28
Signal Processing	3.88	4.70	4.33
Basic Circuit and its Laws	3.86	4.40	4.18

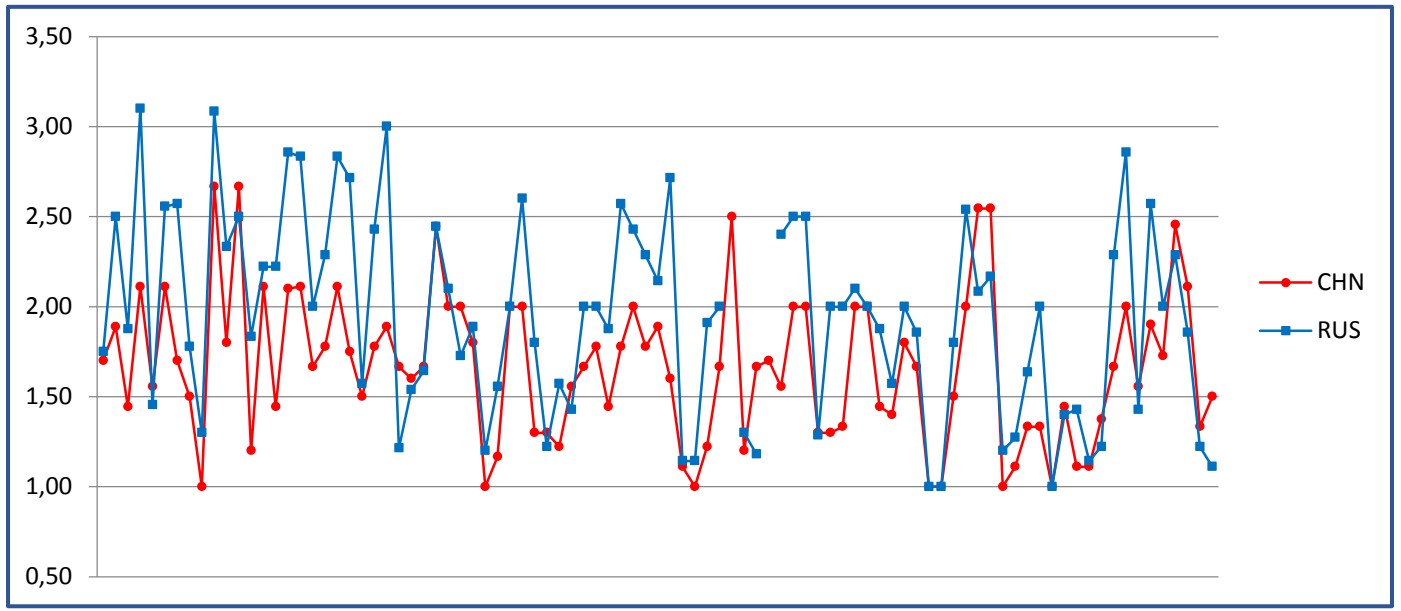
- Analysis of content and construct validity using cross-national expert evaluations of EE content areas:
  - 19 experts from a range of elite and non-elite engineering programs in China and Russia evaluated 21 content areas and 113 sub-content areas
  - 8 content areas were selected for the test

Providing evidence in support of reliability and cross-national comparability of the EE test: stage 2

- Analysis of content and construct validity using cross-national expert evaluations of test items:

  - item selection from four countries (Russia, China, India and USA)

  - item evaluation: feedback from the same pool of experts from China and Russia



- Consistency among the experts (Item difficulty criterion):

Correlation between Russian and Chinese experts' evaluations =0.77**

- 45 items (out of 89) were selected for piloting

**Correlation is significant at the 0.01 level

Providing evidence in support of reliability and cross-national comparability of the EE test: stage 3

- Pilot study
  - ✓ data collection from approximately 800 4th year students from engineering programs in China and Russia.
- Rasch analysis to ensure that
  - ✓ the test meets basic standards for educational measurement, and
  - ✓ it can be equated across two countries and provide comparable measurement results.
- As a result, the EE test for ISHEL main study was constructed. It was expected that it would provide reliable and comparable results for International assessment.

ISHEL: main study, wave 2

- Fall of 2016
- Random representative sample of universities in Russia and China
- Testing of year 4 students of EE departments in Russia and China

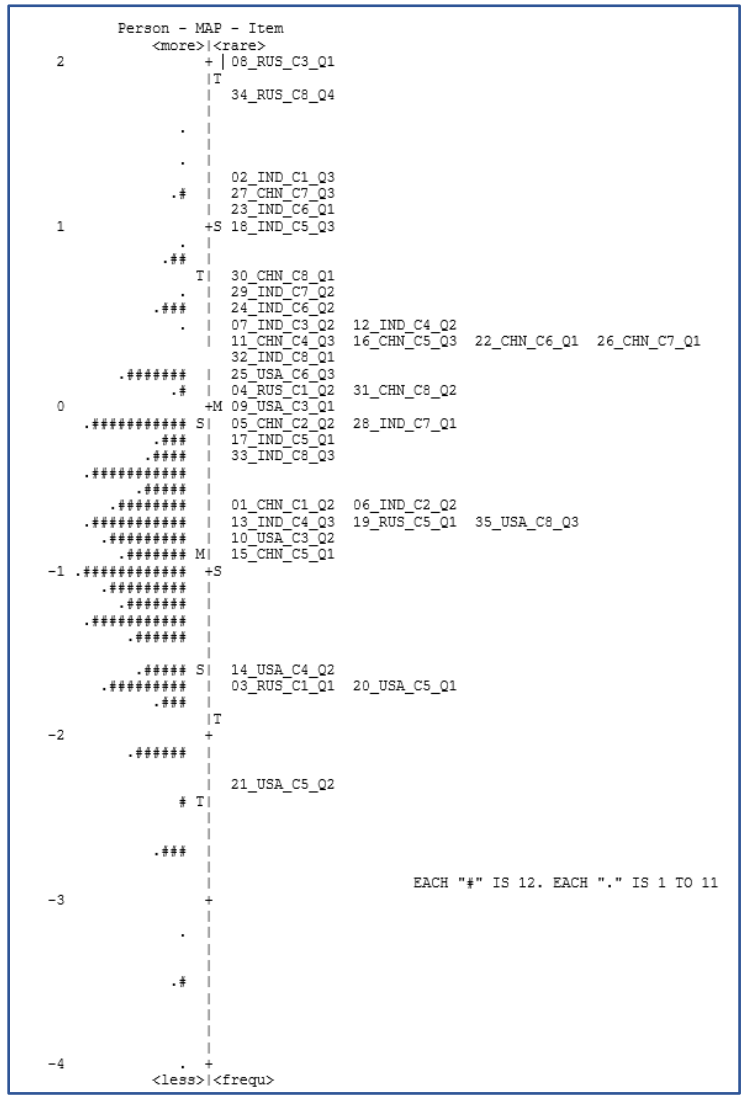
EE professional competencies test

- 1,203 students in China and 850 students in Russia
- 35 items (29 multiple-choice items with 4-5 response options and 6 short response items
- Procedure: computer-based testing, 90-minute session

Analytical approach

- The dichotomous Rasch model (Wright and Stone, 1979) was used to conduct item analysis as well as tests of dimensionality and reliability
- Winsteps, R software
- Particular attention was paid to differential item functioning (DIF) and item response time to provide evidence concerning the cross-national comparability of the test results and to ascertain the possibility of creating a common scale across the two countries.

Preliminary analysis

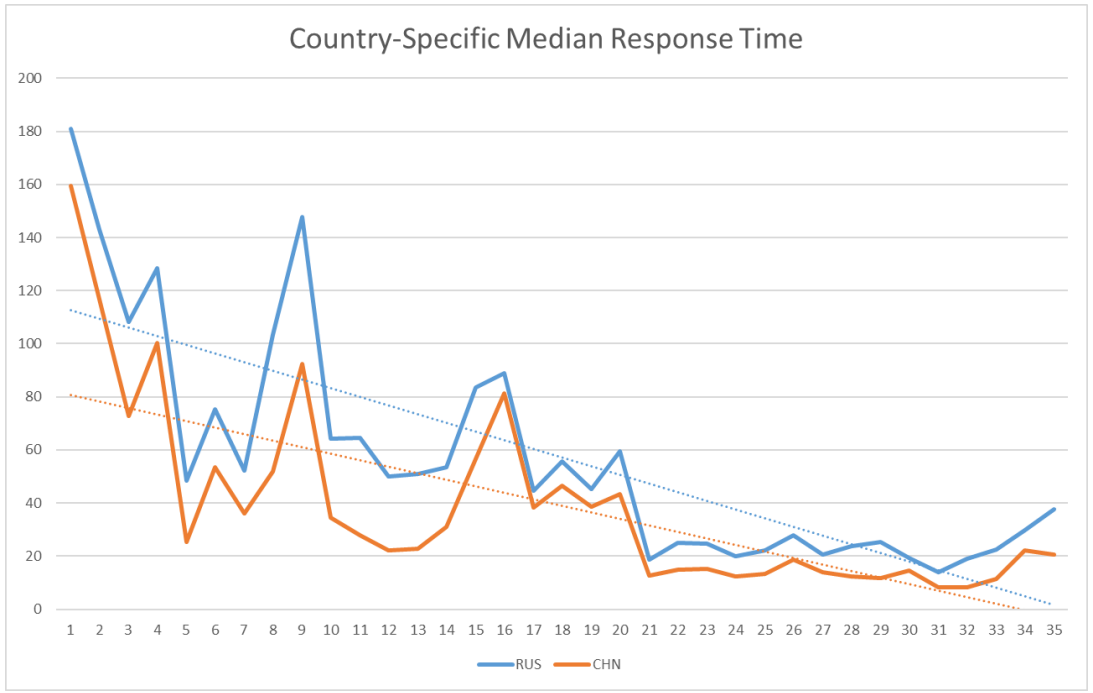


- All items fit the model
- The test is unidimensional
- The test is very difficult (Mean of the sample = - 0.87, CHN = - 0.78, RUS = - 1.05)
- Low variance of students' abilities (SD = 0.77, SD CHN = 0.84, SD RUS = 0.72)
- Reliability is not high, especially for Russia

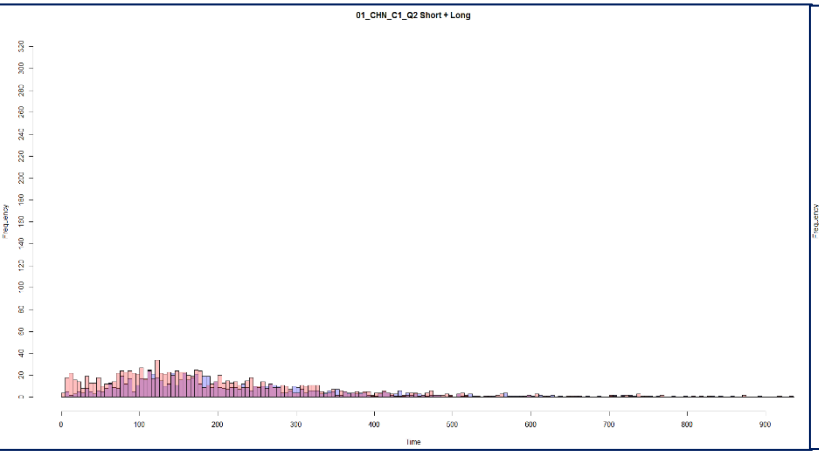
Data	Reliability (real)	Reliability (simulated data)
All	0.60	0.69
Russia	0.55	0.67
China	0.63	0.70

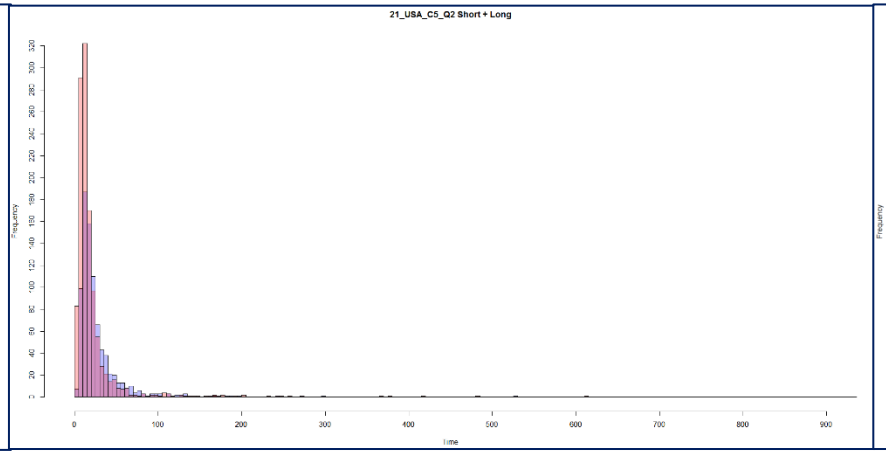
Additional information is required: response time

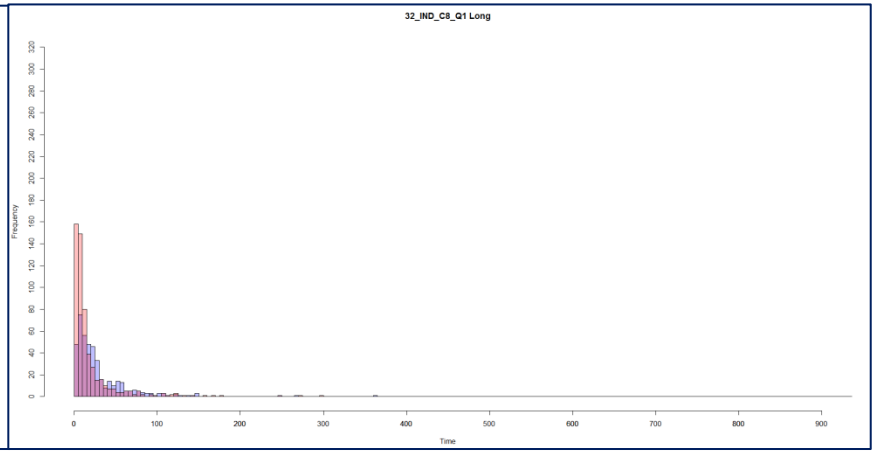
Response Time: descriptives



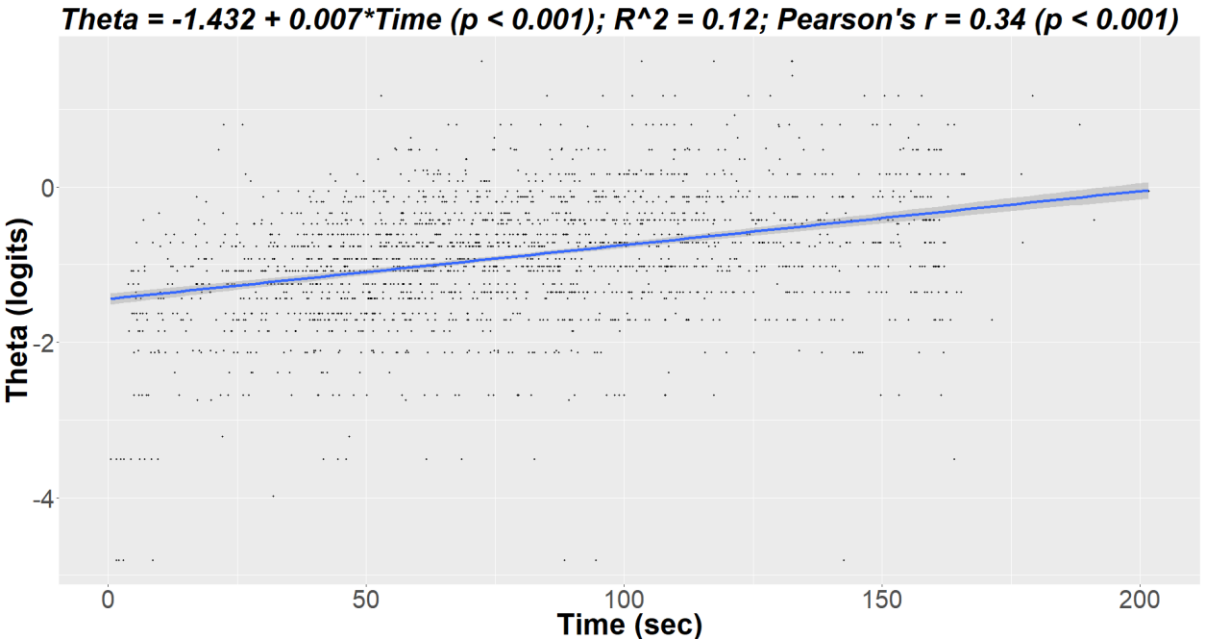
- Trend of faster responding by the end of the test for both countries (b RUS = -3.27**, b CHN = -2.45**)
- Guessing (especially, by the end of the test) is possible







Response Time: How do students use time?



	Unstandardized B	Standard Error	Significance
Constant	-1.674	0.062	p<0.001
Time	0.007	0.001	p<0.001
Country (1 = CHN, 0 = RUS)	0.254	0.076	p<0.005
Time*Country Interaction	0.002	0.001	p<0.05

Low ability levels correspond to small average item response time – this clue supports presence of fast random guessing

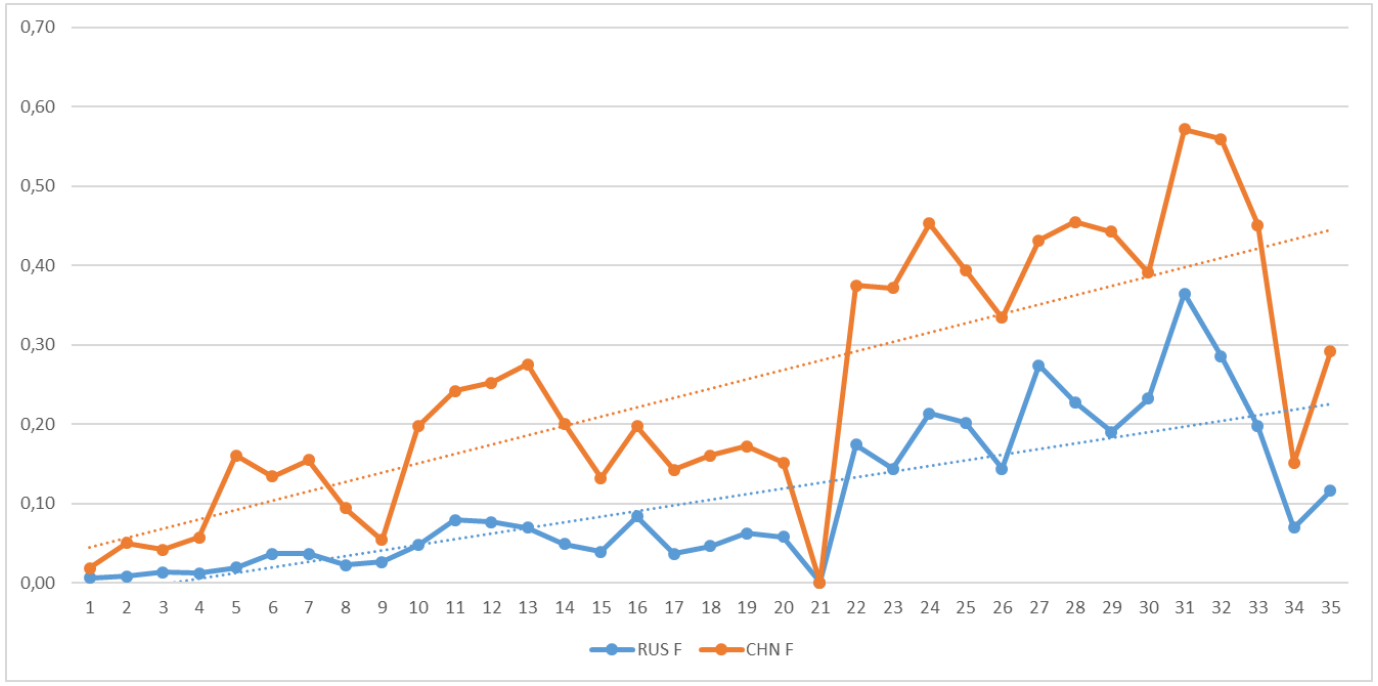
For Chinese sample connection between average response time and ability level is higher (as well as reliability)

Different strategies of guessing?

Fast guessing: 10 seconds tailoring

- All “too fast” responses were treated as omitted data
- Cut off value of 10 sec
- 6% of Russian data and 14% of Chinese data were tailored

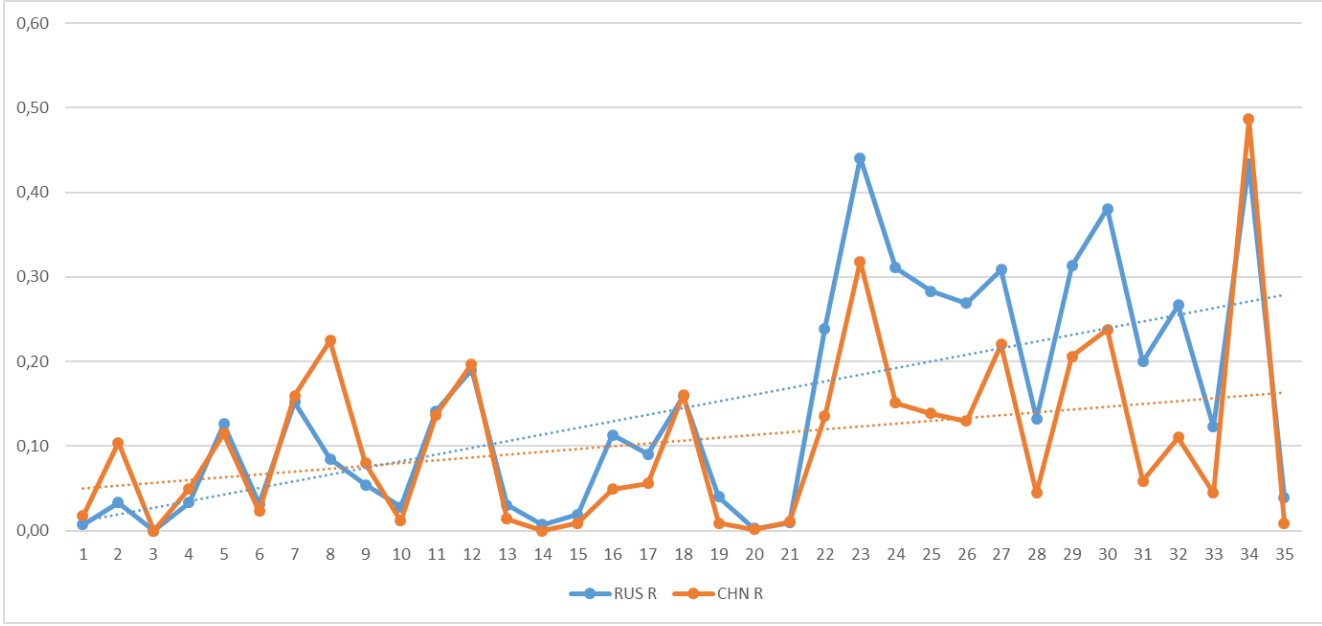
Data	Reliability	
	Initial	After 10 sec tailoring
All	0.60	0.66 (+0.06)
Russia	0.55	0.58 (+0.03)
China	0.63	0.71 (+0.08)

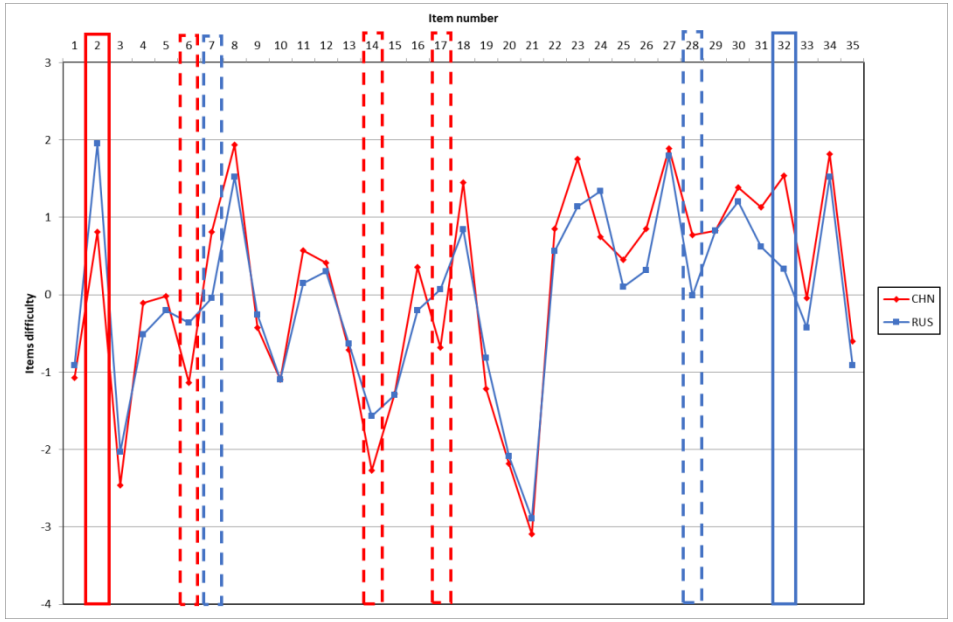


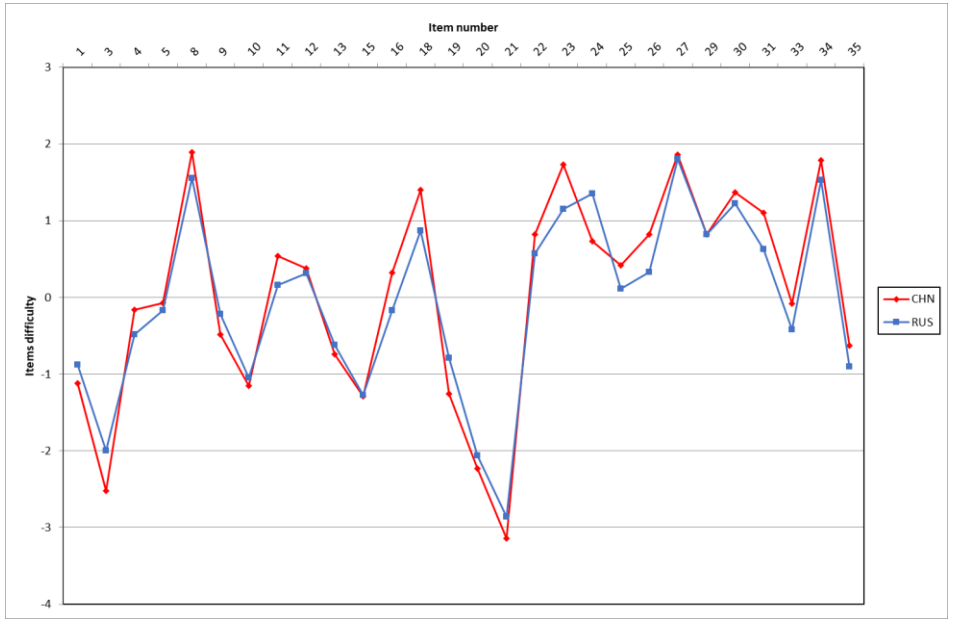
Rundom guessing: “smart” tailoring

- All “too fast” responses were treated as omitted data: cut off value 10 sec
- Additional “smart” tailoring: all responses with P(U=1) < 0.25 & t(U) < 30 sec were treated as omitted (15% of Russian data and 21% of Chinese data)
- Tailored data is more reliable, so item analysis should be done with this data

Data	Reliability		
	Initial	After 10 sec tailoring	After smart tailoring
All	0.60	0.66 (+0.06)	0.71 (+0.05)
Russia	0.55	0.58 (+0.03)	0.64 (+0.06)
China	0.63	0.71 (+0.07)	0.74 (+0.03)







Main test analysis

- All items fit the model
- The test is unidimensional
- Cross-national equivalence
  - ✓DIF-analysis revealed 7 items with DIF: 4 in favor of China and 3 in favor of Russia
  - ✓The rest 28 items can be used for linking
  - ✓Common scale can be constructed

Conclusions

- The test itself is valid for cross-national comparisons
- Reliability troubles are caused by external factors (first of all, guessing)
- Students from China and Russia use different guessing strategies in terms of time
- Several items exhibit DIF, however, IRT-modeling allowed us to build a common cross-nationally equivalent scale for comparison of educational achievements based on the fair items
- Future research: How to estimate students ability taking into account response time?

Thank you for your attention!