The Talent Management Expert

Strategy. Execution. Results.

Adaptive Item Calibration and Norming: Unique Considerations of a Global Deployment

October, 2011

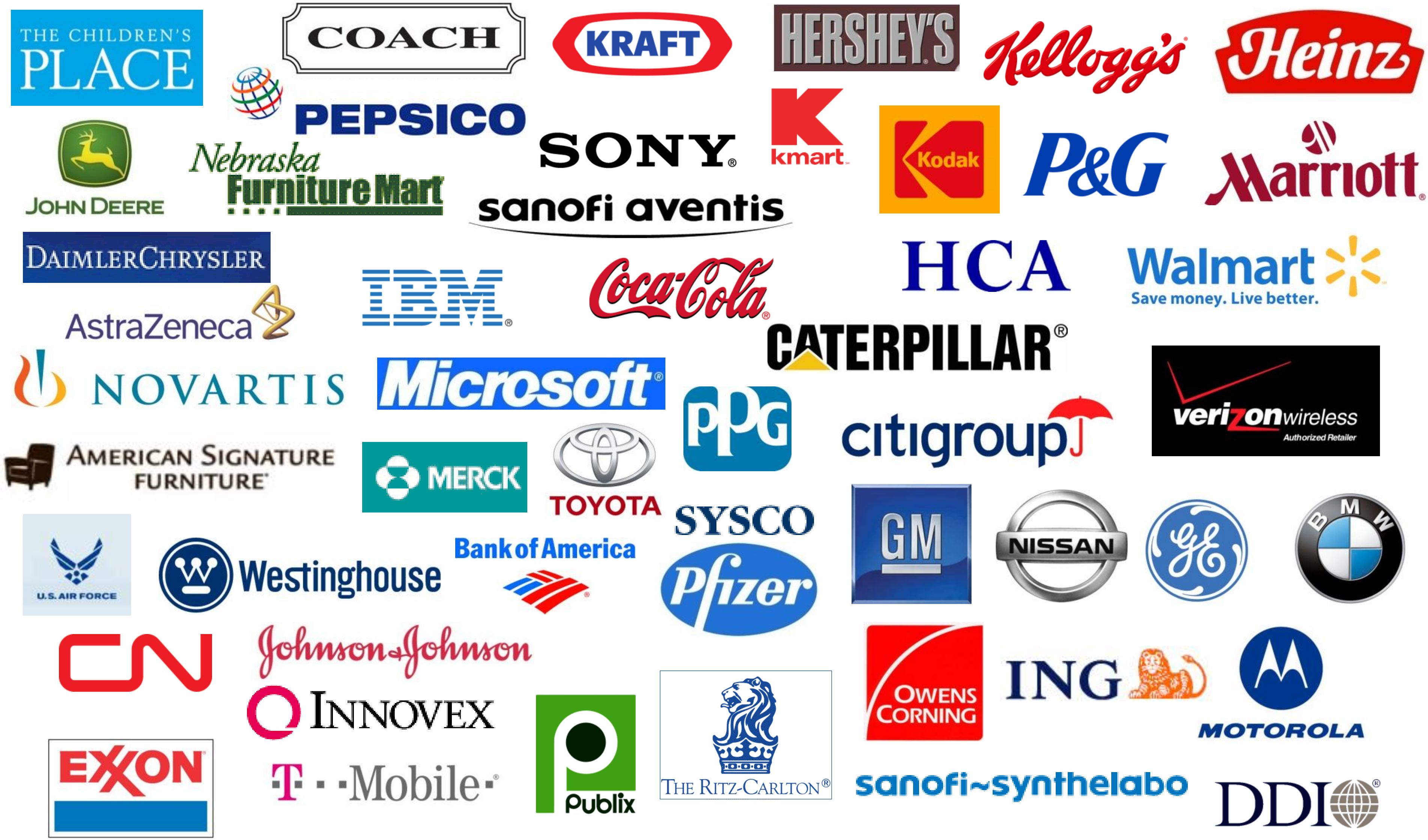
Alexander Schwall
Evan Sinar

DDI®

The Talent Management Expert

- 40 plus years, 1,600 clients
- 1,500,000 tests a year
- 1,001 associates in 42 offices in 26 countries
- 95% of our clients highly recommend us

DDI: Clients in various industries



Agenda

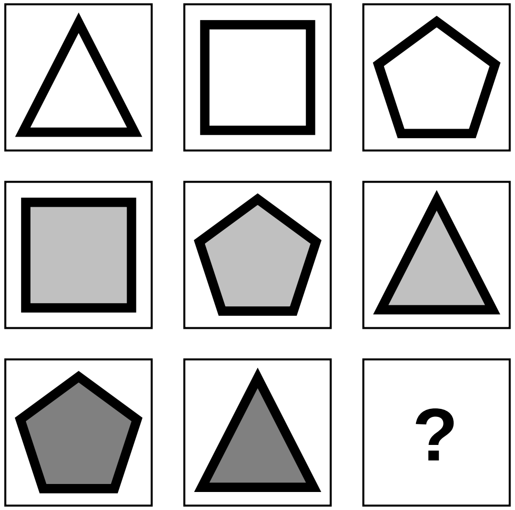
Testing requirements/goals of global testing clients

How adaptive testing can help meet these requirement

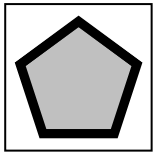
The Adaptive Reasoning Test

- Figural Reasoning Test
- About 500 items in pool
- Fixed Length Test
- Administered for selection of external candidates for all roles from administrative positions to management
- Over 250.000 test administrations per year
- Administered word-wide in over 90 countries

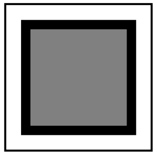
Matrix item



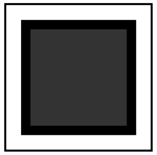
A



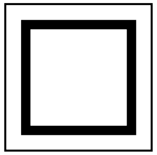
B



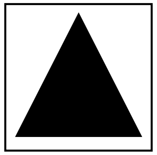
C



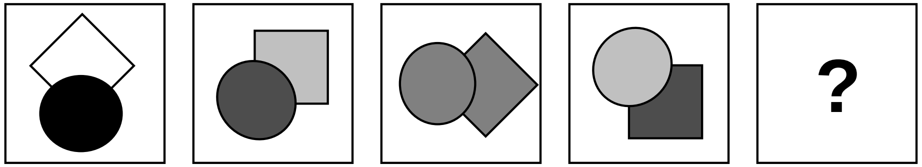
D



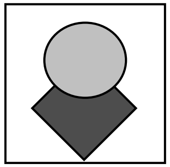
E



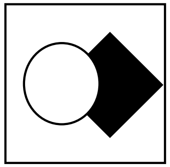
Series item



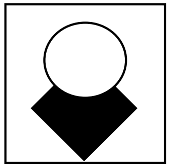
A



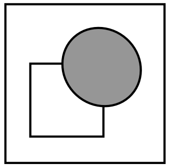
B



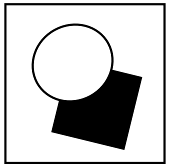
C



D



E



Testing needs of DDI's global clients

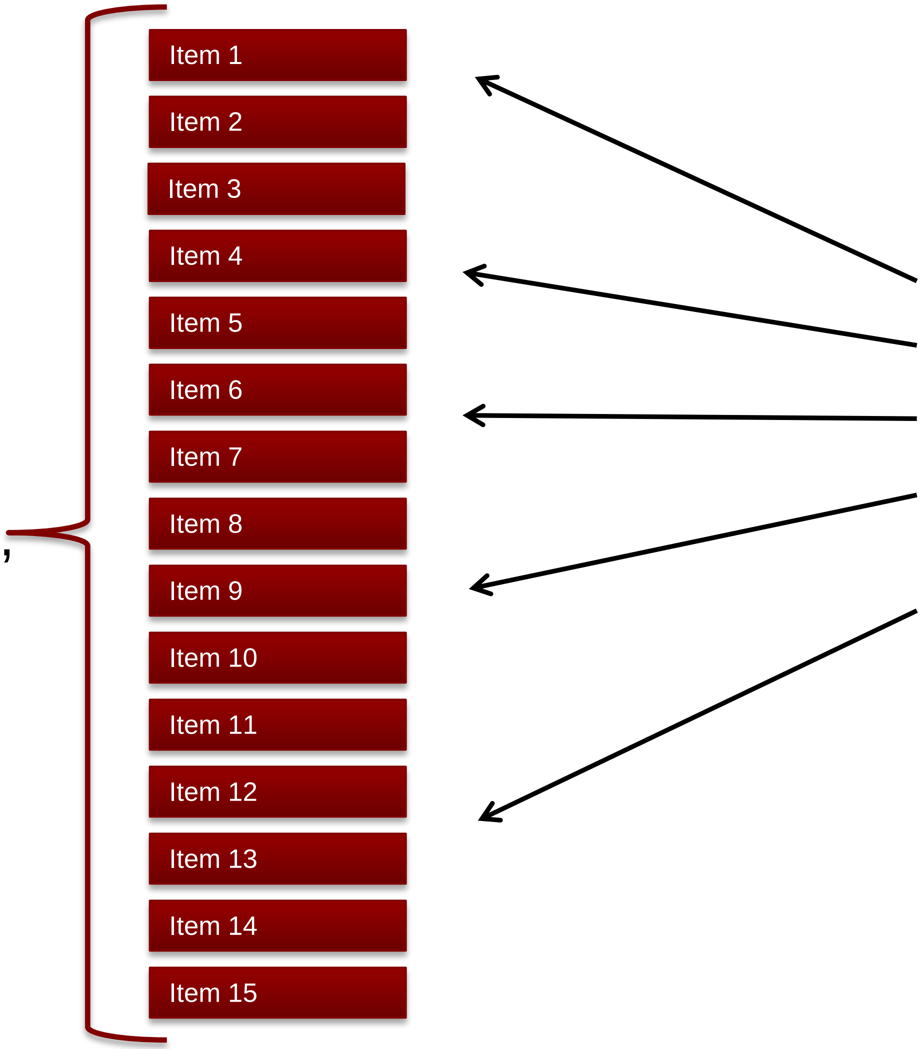
1. Test needs to be deployed as a Unproctored Internet-based Test (UIT)
2. Test needs to be short and minimally disruptive to the recruiting process
3. Test needs to be deployable with minimal translation and produce equivalent scores across all regions/countries.

Challenges of a Global Testing Environment

- High exposure of test items
  - Relative high exposure of best performing items despite exposure control
  - Propensity of cheating in some countries
- Limited Opportunity to collect data for experimental item calibration
  - No tolerance for calibration data collection among testing clients
  - No tolerance for calibration among test takers

How can adaptive testing help address these challenges?

Administration of experimental items



Items presented, total 15

Item 1

Item 2

Item 3

Item 4

Item 5

Item 6

Item 7

Item 8

Item 9

Item 10

Item 11

Item 12

Item 13

Item 14

Item 15

Experimental items can appear in any of the 15 positions

Approach to item calibration: Common Item Equating

- Use of common item equating (Yu & Osborn-Popp, 2005)
- Experimental items are assumed to be unique to each “test form”
- Live Items are assumed to be “Anchor Items” (Rizopoulos, 2011) shared among test takers
- Calibration performed using ltm package in R

Data Structure for calibration

	Exp 1	Exp 2	Exp 3	Exp …	Exp 112	Live 1	Live 2	Live 3	Live 4	Live 5	Live 6	Live 7	Live 8	Live 9	Live …	Live 500
P1	1					1		0			1		1			
P2		1			1		1	1			0	0	1			
P3		1				1			1					0		0
P4		0	1					1	0			1				
P5			0			0					1		1			
P6	1										1	1				0
…		0			0	0			1					1		
Pn	1						1				0		0			

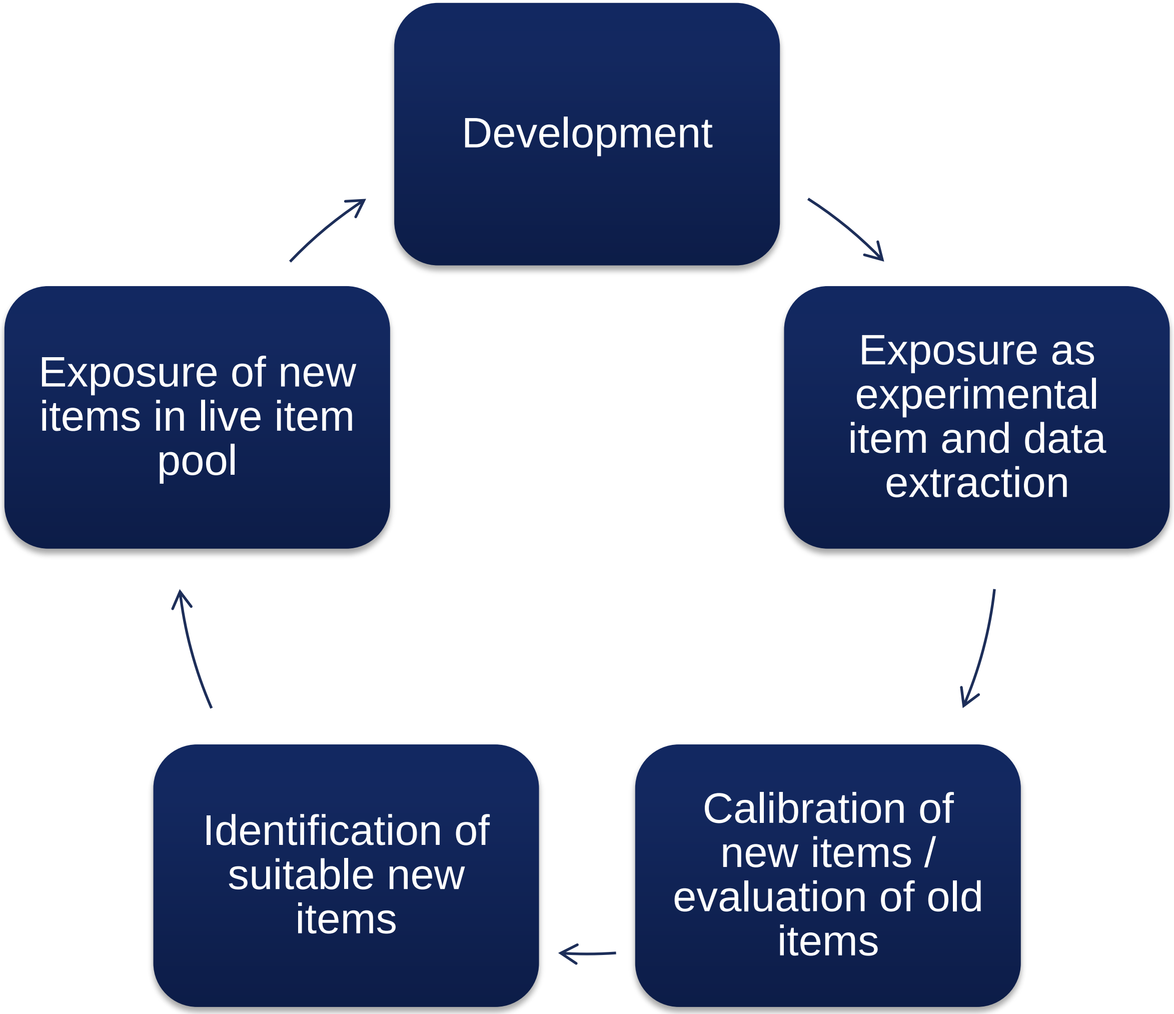
Item Constraining

- Old (i.e. previously calibrated) items are constraint in terms of their parameters
- New items are calibrated in reference to the new items

Advantages of Common Item Equating

- Experimental items can be co-administered in small quantities with live items
- New items are tethered to old items: difficulty of item pool is not drifting*
- Test taking experience is minimally affected by administration of experimental items
- No administrative cost to data collection
- Test taker is motivated and engaged when taking the experimental item

Experimental Item Life-Cycle


Development
Exposure as experimental item and data extraction
Calibration of new items / evaluation of old items
Identification of suitable new items
Exposure of new items in live item pool

Keeping the item pool “fresh”

Identification of suitable new items

- Analysis of difficulty band coverage
- Frame of reference training for item writers to target certain difficulty areas.
- Result: the ART is fed by an evolving item pool with a blend of established a newly calibrated items.

Challenges set by testing clients

1. Test needs to be deployed as a Unproctored Internet-based Test (UIT)
2. Test needs to be short and minimally disruptive to the recruiting process
3. Test needs to be deployable with minimal translation and produce equivalent scores across all regions/countries.

Challenges of a global participant population

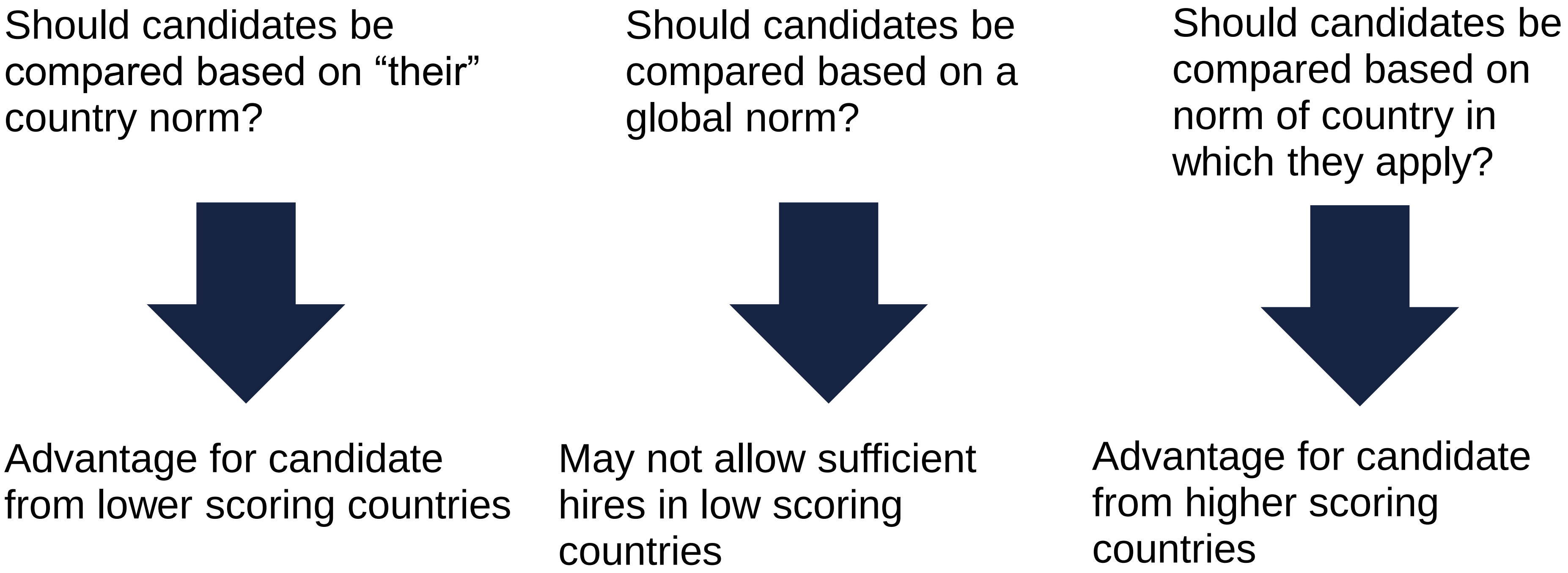
- Median US = 0.12 vs.
  - Median China = 0.63
  - Median Japan = 0.56
  - Median Percentile Singapore = 0.35
  - Median UK = 0.21

What causes score differences and what are the implications?

- Score differences based on caliber of candidates (all country population means are similar)
- Are score difference based on country specific differences in targeted construct
- Currently: score differences require country specific norms

Trade offs between norming approaches

However, test may be used for cross country comparisons: candidate apply in countries other than their land of origin.



Should candidates be compared based on “their” country norm?

Advantage for candidate from lower scoring countries

Should candidates be compared based on a global norm?

May not allow sufficient hires in low scoring countries

Should candidates be compared based on norm of country in which they apply?

Advantage for candidate from higher scoring countries

Summary

- CAT has allow us to
  - Globally deploy a UIT
  - Administer a high volume of GMA tests
  - Provide good protection against cheating
  - Offer a brief but valid test

- Still more research necessary to eliminate score differences

Thank You!