When Beta Forms AND Beta Samples Differ

Hitting the Unlucky Lottery?

Presenters

Jim Henderson

- Manager, Product Training, Ping Identity

Jill R. van den Heuvel, Ph.D.

- Senior Psychometrician, Alpine Testing Solutions

ATP Innovations in Testing 2021 Virtual

Who is Ping Identity?

Ping Identity

Alpine Testing Solutions

ASSESSMENT for a Better World

WHAT WE DO

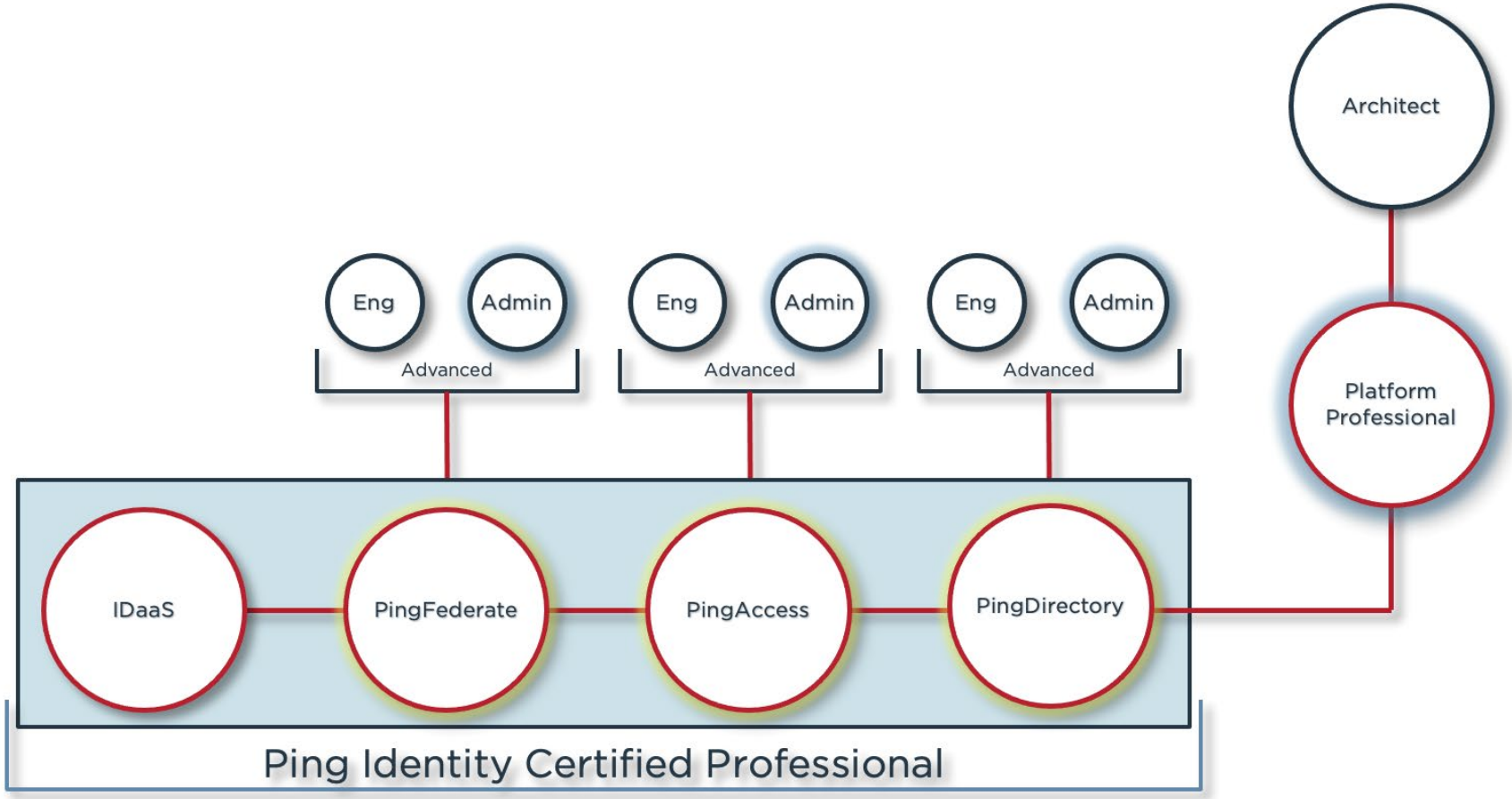
Ping Identity champions you to provide identity solutions that enable secure, extraordinary digital experiences for employees and customers.

HOW WE DO IT

We champion what makes your business and environment unique to connect users to cloud, mobile, and on-premises applications and APIs while managing identity and profile data at scale.

Certification Program

ASSESSMENT for a Better World

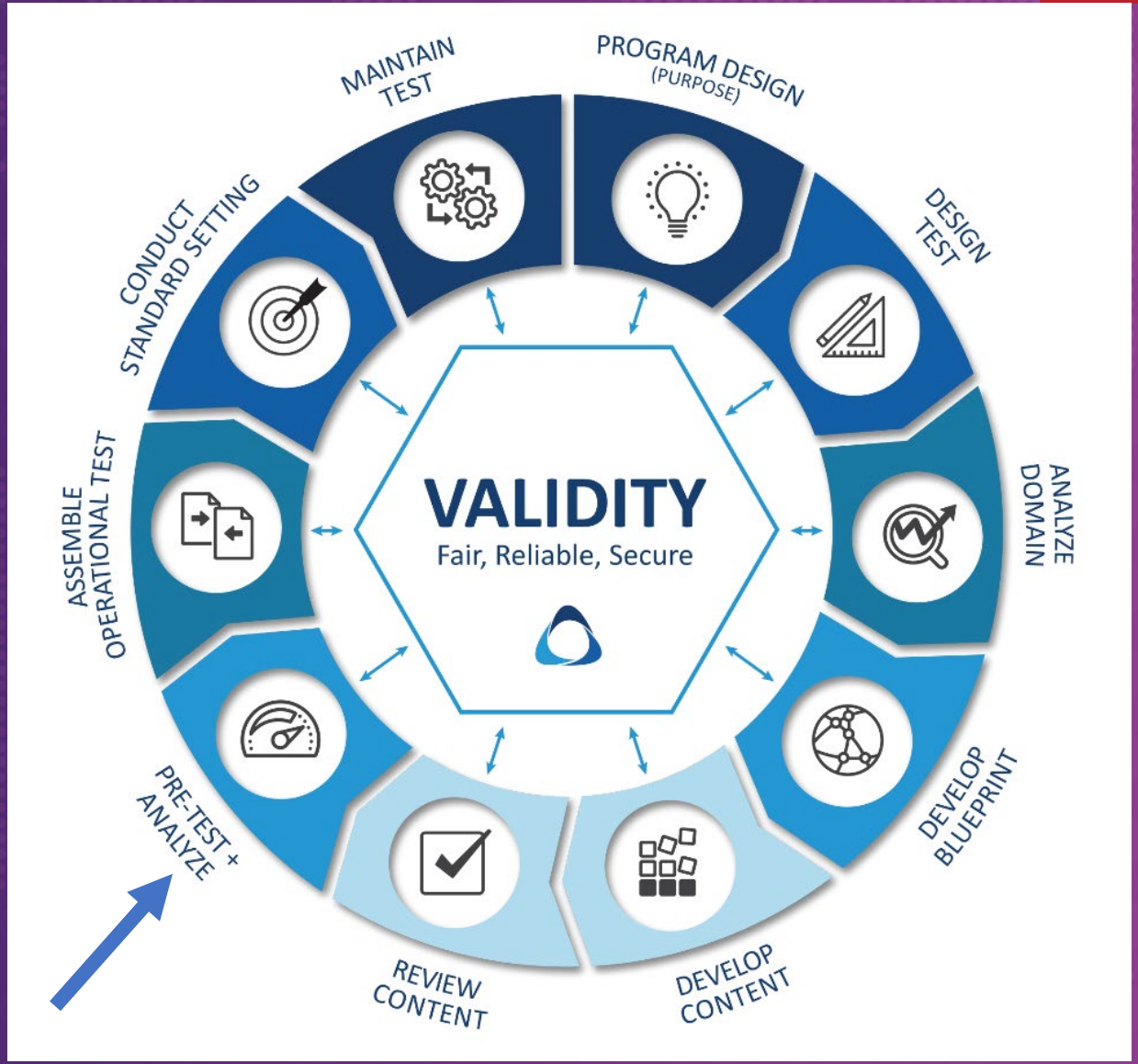


ATP Innovations in Testing 2021 Virtual

ASSESSMENT for a Better World

Ping Identity

Alpine Testing Solutions



Exam Information

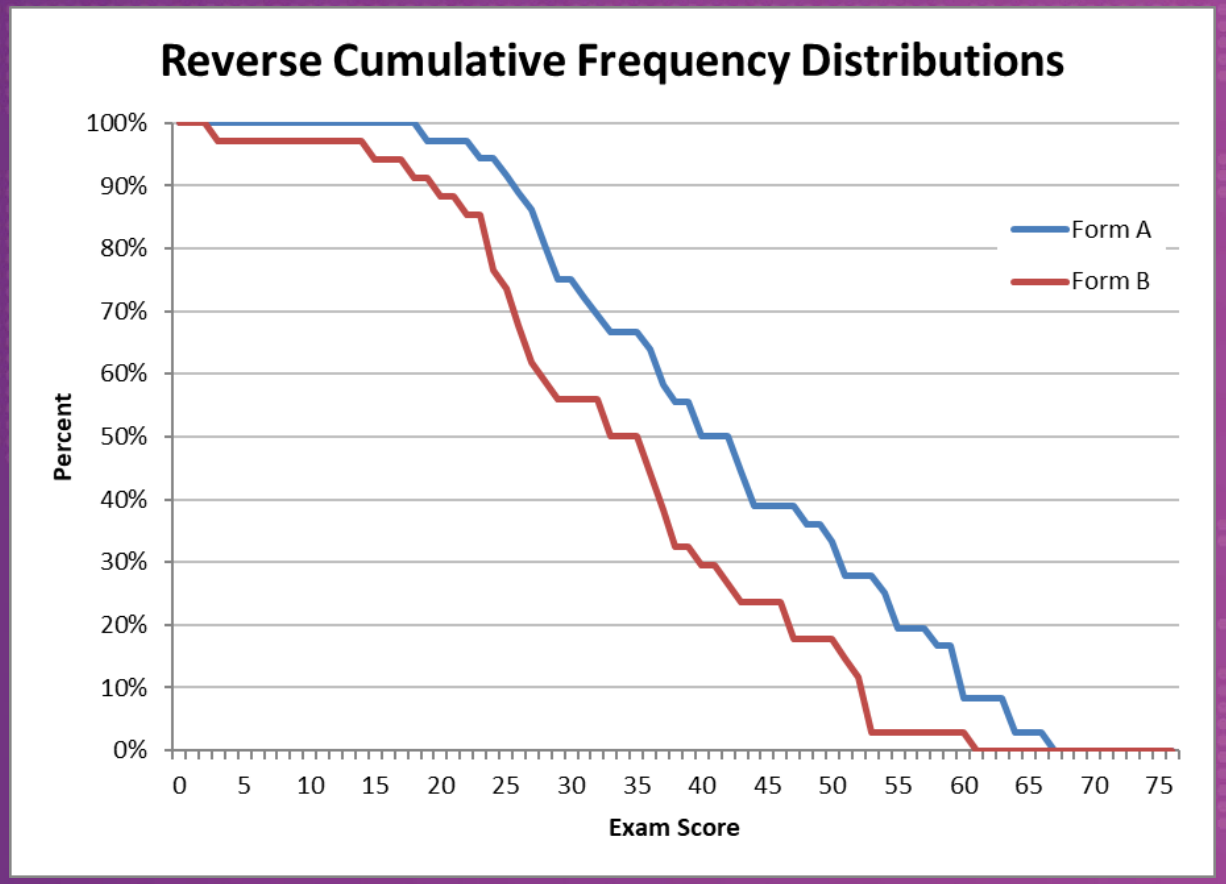
- Brand new exam and content area
- All items were newly written *(no statistics)*
- Beta forms administration allowed for statistical analyses of item performance
- Standard setting
- Live forms assembly
- Beta candidate rescore equated to live form cut score

Beta Forms

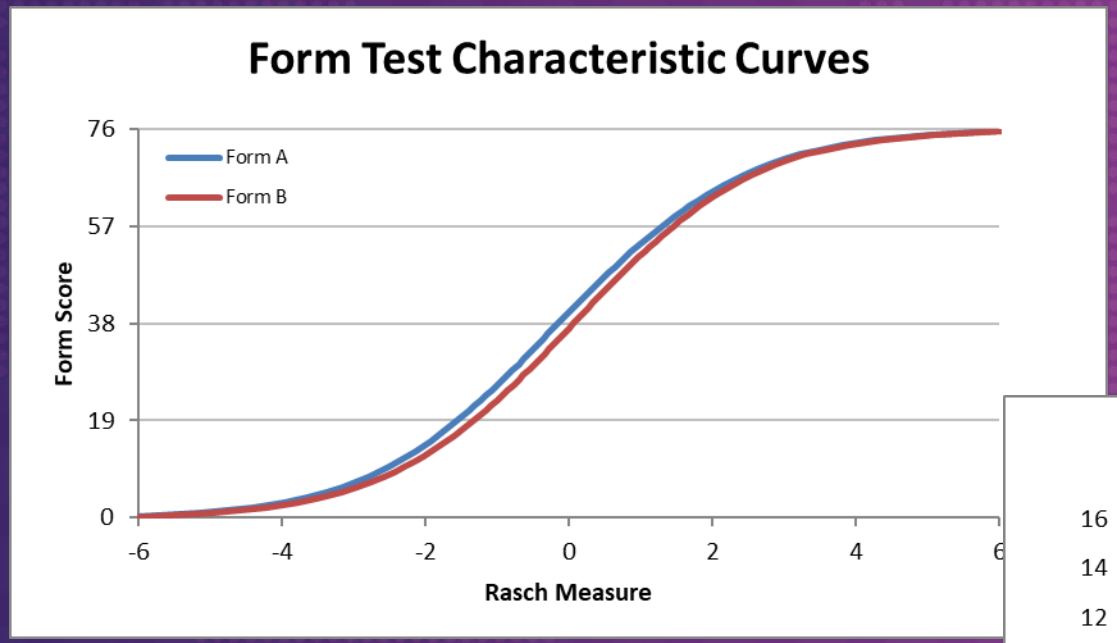
- **All 132 items developed were used in the beta**
- **76 items per beta form (2 beta forms)**
  - Content proportionally aligned with the live form blueprint
  - Item types balanced (multiple-choice, choose all that apply)
- **20 common items across forms**
  - Content approximately proportionally aligned with the live form blueprint
- **One month administration window**
- **Candidates were randomly assigned to Form A or Form B**

Beta Form Performance

Health Check	Form A	Form B
Candidate Count	36	34
Number of Items	76	76
Mean	41.47	33.35
SD	13.49	13.01
Median	40.5	33.5
Standard Error of the Mean	2.25	2.23
95% confidence interval +/-	4.41	4.37
Minimum	18	2
Maximum	66	60
Skewness	0.16	0.00
Kurtosis	-1.14	-0.24
Alpha Reliability	0.927	0.917
SEM	3.64	3.74
95% confidence interval +/-	7.14	7.34

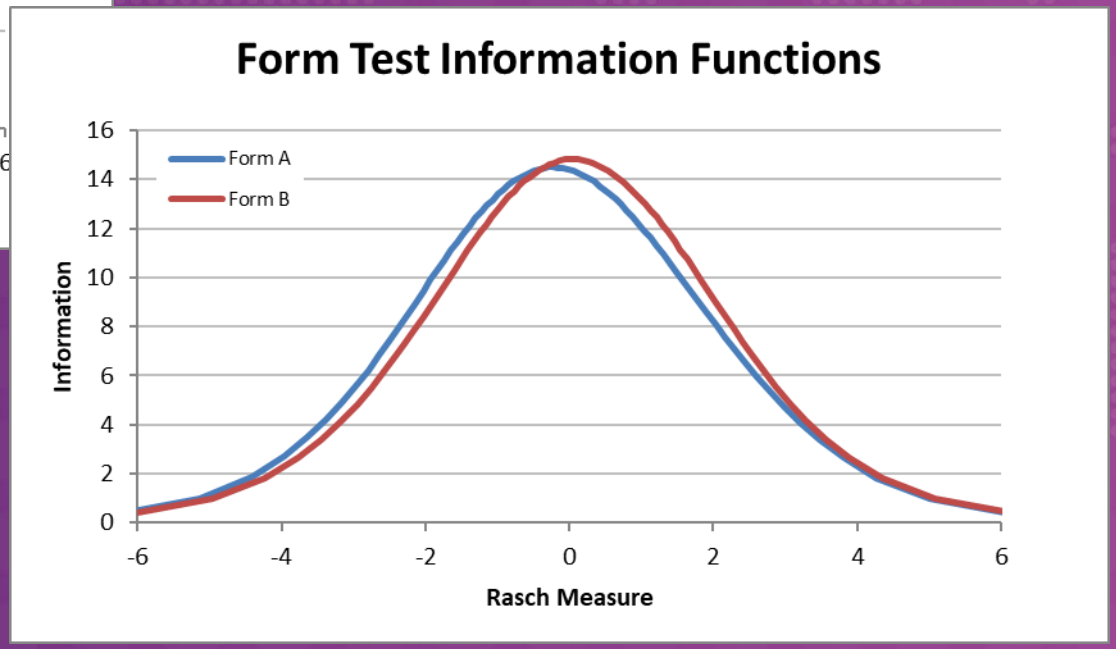


Beta Form Performance



Rasch Item Measures	Mean	Std. Dev.
Overall	0.0002	1.2375
Form A unique	-0.0836	1.2648
Form B unique	0.1659	1.1612
Common	-0.2295	1.3666

Rasch Person Measures	Mean	Std. Dev.
Overall	-0.0819	1.0817
Form A	0.1411	1.0347
Form B	-0.3179	1.0952



Candidate Fairness

- It should be irrelevant to a candidate which of multiple forms they are administered. Their success/failure should be based on their knowledge without an interaction with the items administered.
- In this case fairness was clearly a concern.

Candidate Fairness

- Standard setting included all Keep items
- IRT was used to determine the theta cut at the selected raw cut
- The theta cut was applied to the live forms and the beta candidate rescore forms
- Beta candidates were rescored against the keep items from their beta form

Now What?

- **The two forms perform differently**
- **The candidates on the two forms are differentially able**

- **We must assemble two live forms**
  - Policy decision
  - Re-take needs (when candidates fail)

Balance Live Forms

- Content (meet the blueprint)
- Statistics (next slide)
- Item variables

Item Type	Form A	Form B
Check-All-That-Apply	12	12
Standard Multiple Choice	48	48

Prior Form	Form A	Form B
Beta Form A	25	24
Beta Forms A&B	11	11
Beta Form B	24	25

Balance Statistics (CTT & IRT)

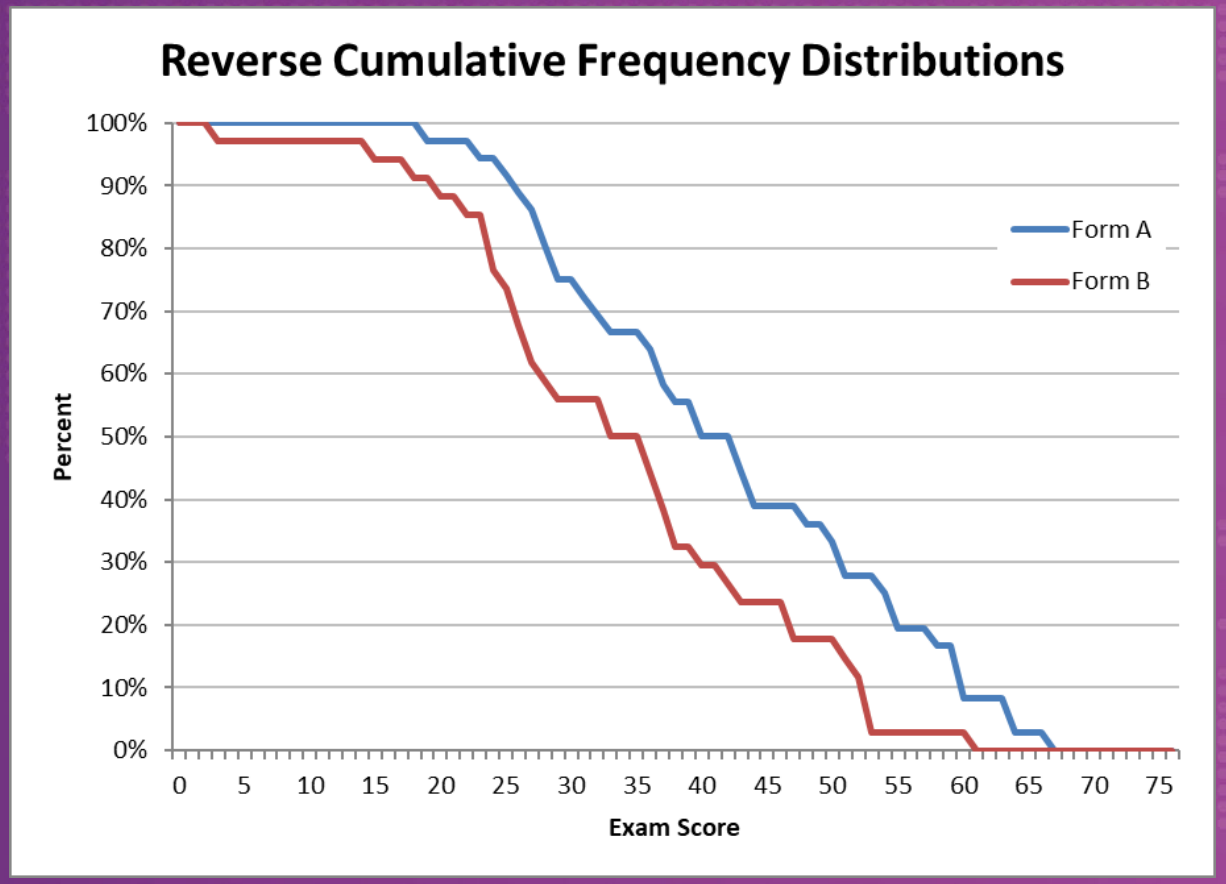
Stats	Form	
	A	B
Number of items	60	60
Number of points	60	60
Items\Percent shared with Form A		33%
Items\Percent shared with Form B	20	
Mean	31.73	31.69
Standard deviation	12.33	12.31
SE of mean*	1.95	1.95
95% confidence interval for mean (plus or minus)*	3.82	3.81
Alpha Reliability	0.93	0.93
Standard Error of Measurement	3.22	3.22
95% confidence interval for score (plus or minus)*	6.31	6.32
Average Test Time (minutes)	48.72	48.84
Average Item Measure	-0.24	-0.25
Standard deviation of item measures	1.04	1.04
Information at target cut score	12.17	12.18
SE at target cut score	0.29	0.29

The Waiting

- Two live forms
- Get at least 100 candidates before pulling data to see what happened
- Forms were administered beginning in early November 2019
- Data pull in January 2021 with 110 records

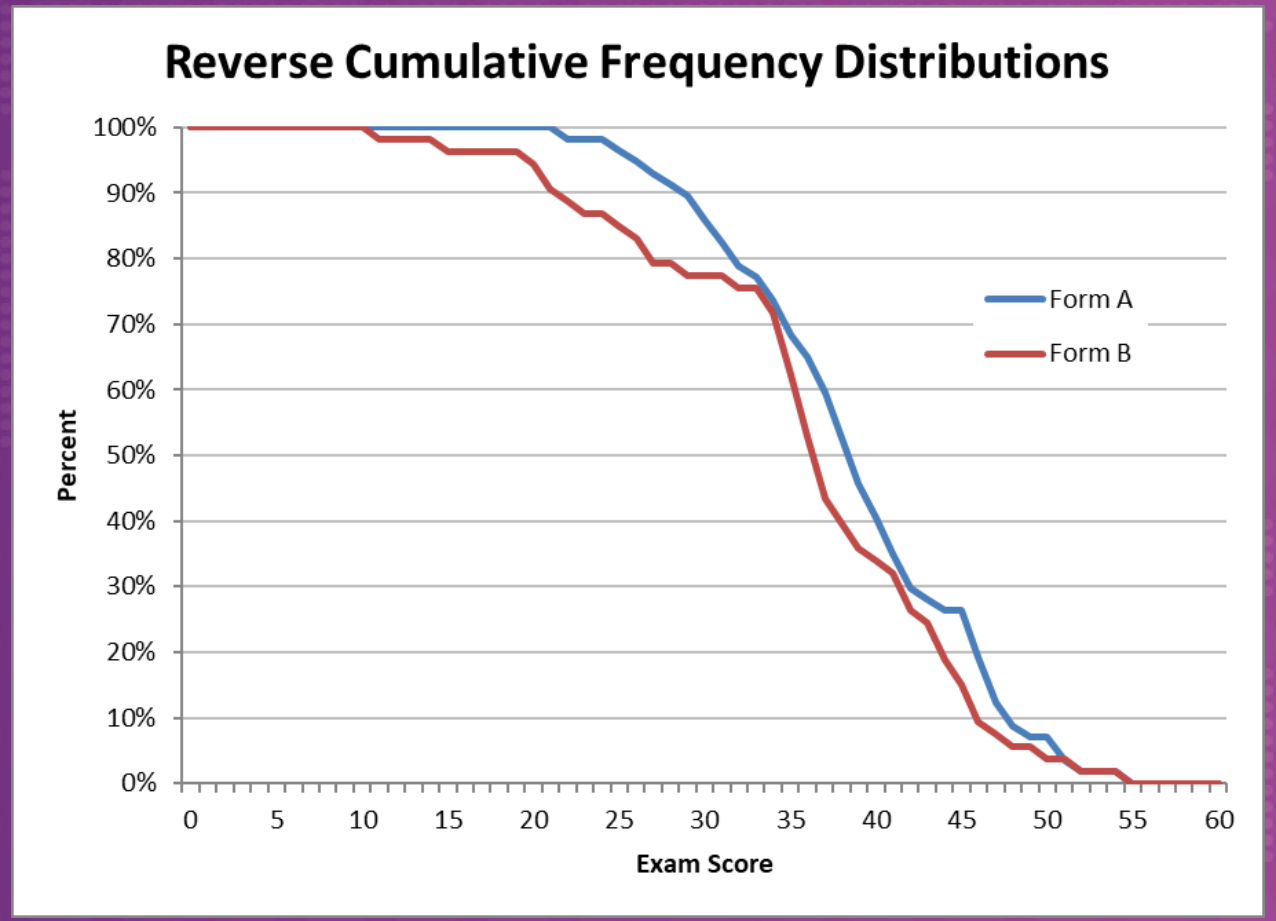
Beta Form Reminder

Health Check	Form A	Form B
Candidate Count	36	34
Number of Items	76	76
Mean	41.47	33.35
SD	13.49	13.01
Median	40.5	33.5
Standard Error of the Mean	2.25	2.23
95% confidence interval +/-	4.41	4.37
Minimum	18	2
Maximum	66	60
Skewness	0.16	0.00
Kurtosis	-1.14	-0.24
Alpha Reliability	0.927	0.917
SEM	3.64	3.74
95% confidence interval +/-	7.14	7.34

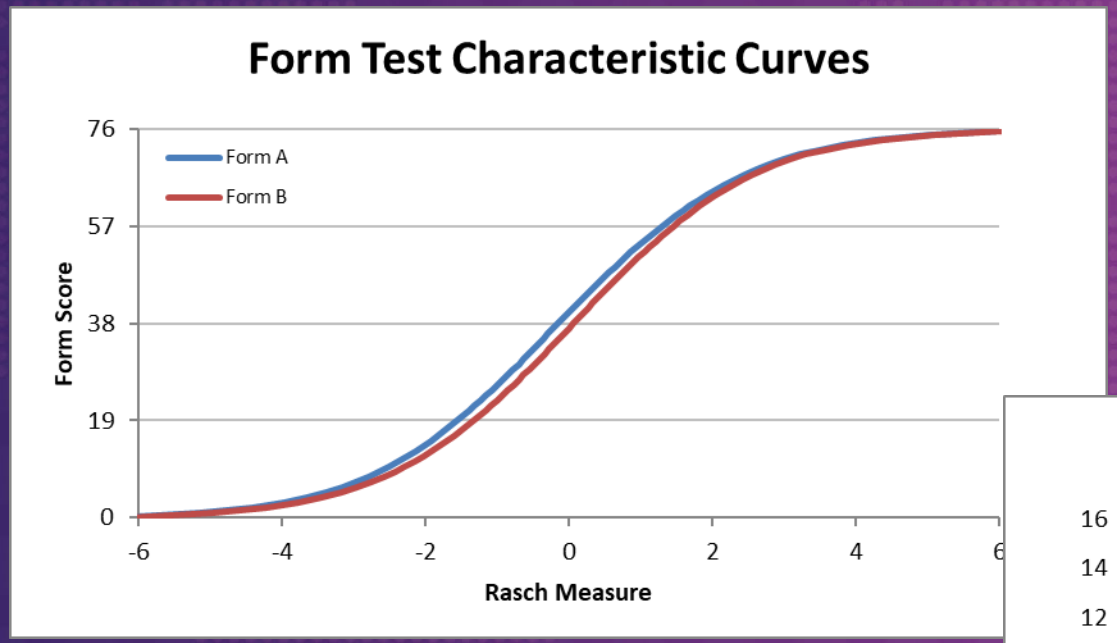


Live Form Performance

Health Check	Form A	Form B
Candidate Count	57	53
Number of Items	60	60
Mean	37.98	35.28
SD	7.42	9.36
Median	38.0	36.0
Standard Error of the Mean	0.98	1.29
95% confidence interval +/-	1.93	2.52
Minimum	21	10
Maximum	54	54
Skewness	-0.08	-0.60
Kurtosis	-0.47	0.24
Alpha Reliability	0.806	0.877
SEM	3.27	3.28
95% confidence interval +/-	6.40	6.44
Pass Rate	73.7%	71.7%
CSEM at Passing Score	3.475	3.487
Decision Consistency (Livingston-Lewis)	0.838	0.876
Accuracy (Livingston-Lewis)	0.886	0.914

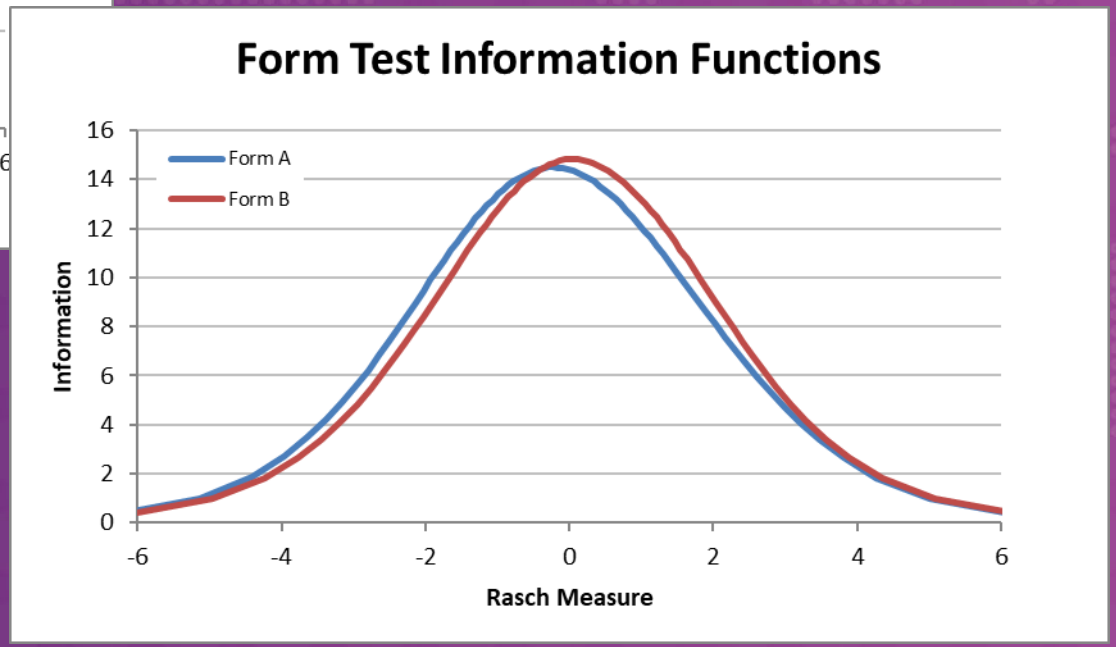


Beta Form Reminder

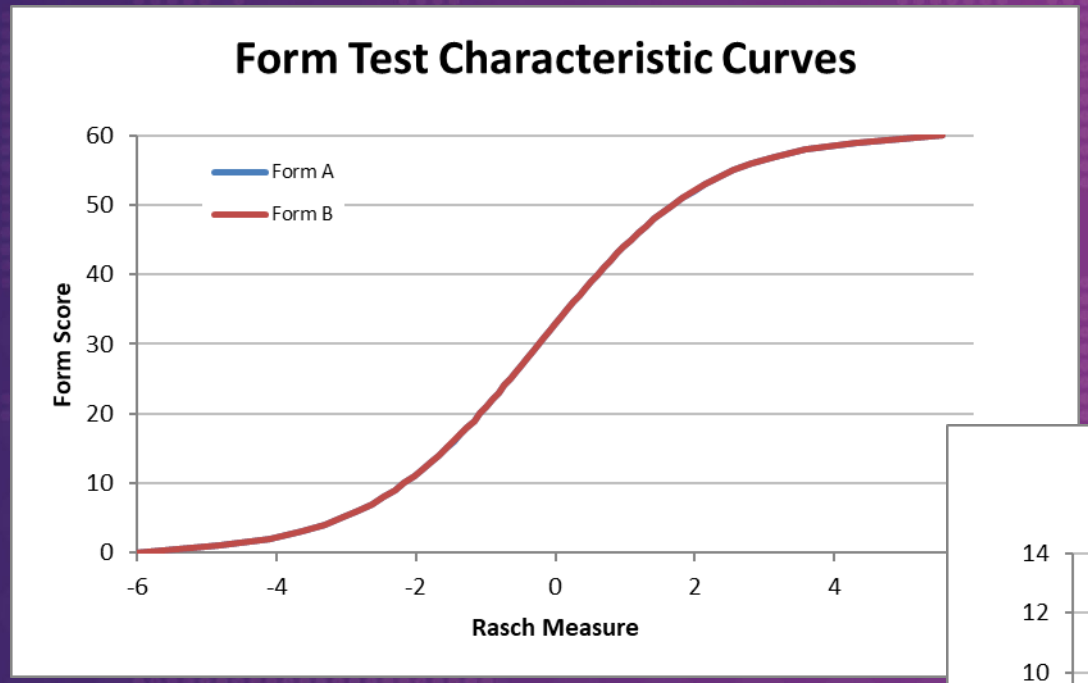


Rasch Item Measures	Mean	Std. Dev.
Overall	0.0002	1.2375
Form A unique	-0.0836	1.2648
Form B unique	0.1659	1.1612
Common	-0.2295	1.3666

Rasch Person Measures	Mean	Std. Dev.
Overall	-0.0819	1.0817
Form A	0.1411	1.0347
Form B	-0.3179	1.0952

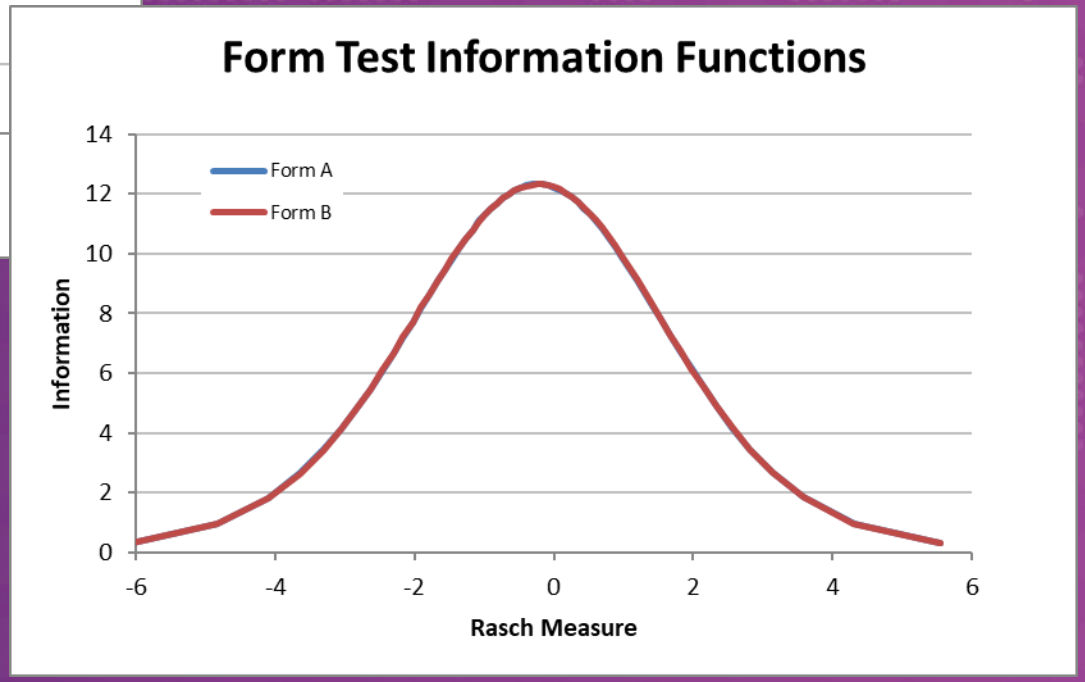


Live Form Performance



Item Measure Displacement	
Minimum	-1.8
Maximum	1.66
Mean	-0.01
Std. Dev.	0.61

Rasch Person Measures	Mean	Std. Dev.
Overall	0.3501	0.7791
Form A	0.4686	0.6915
Form B	0.2226	0.8518



Meaning?

- What are the live forms telling us?
- Did we assemble equivalent forms, even in the face of beta forms and beta form populations that were different?

- Maybe?

All Forms Performance

Stats	Beta Form Performance		Pre-Equated Estimates		Live Form Performance	
	Form A	Form B	Form A	Form B	Form A	Form B
Number of items	76	76	60	60	60	60
Number of candidates	36	34			57	53
Items\Percent shared with Form A		26%		33%		
Items\Percent shared with Form B	20		20			
Mean	41.47	33.35	31.73	31.69	37.98	35.28
Standard deviation	13.49	13.01	12.33	12.31	7.42	9.36
SE of mean*	2.25	2.23	1.95	1.95	0.98	1.29
95% confidence interval for mean (plus or minus)*	4.41	4.37	3.82	3.81	1.93	2.52
Alpha Reliability	0.93	0.92	0.93	0.93	0.81	0.88
Standard Error of Measurement	3.64	3.74	3.22	3.22	3.27	3.28
95% confidence interval for score (plus or minus)*	7.14	7.34	6.31	6.32	6.40	6.44
Average Test Time (minutes)	72.88	75.93			71.44	71.55
Average Item Measure	-0.12	0.06	-0.24	-0.25	-0.24	-0.25
Standard deviation of item measures	1.28	1.22	1.04	1.04	1.04	1.04
Target cut score (theta)			0.06	0.06	0.09	0.08
Information at target cut score	-	-	12.17	12.18	12.14	12.15
SE at target cut score	-	-	0.29	0.29	0.29	0.29

Item Performance Deltas

But what about differences in item performance from appearance on the beta to the live forms

Delta (Beta minus Live)		
100 items	*p* -value	Item-Score Correlation
Increase	19	80
Decrease	81	20
Minimum	-0.36	-0.33
Maximum	0.23	0.52
Mean	-0.08	0.12
Std. Dev.	0.12	0.16

Item Measure Displacement	
Minimum	-1.8
Maximum	1.66
Mean	-0.01
Std. Dev.	0.61

Next Steps

- **Reviewing items for continued content accuracy and relevancy**
  - IT content is everchanging
- **Developing new items due to content updates and to replace items with poor statistical performance**
- **Re-Assemble forms and add unscored items**

ATP Innovations in Testing 2021 Virtual

Questions?

Ping Identity.

Alpine Testing Solutions

ASSESSMENT for a Better World