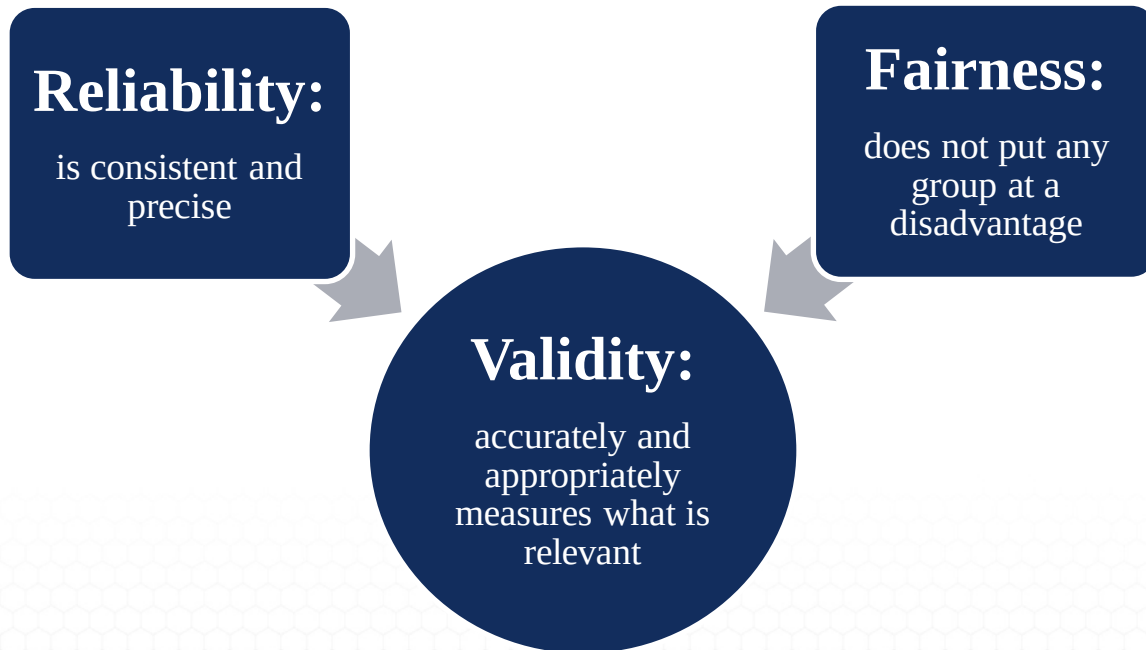


All About Health Checks for Alpine University

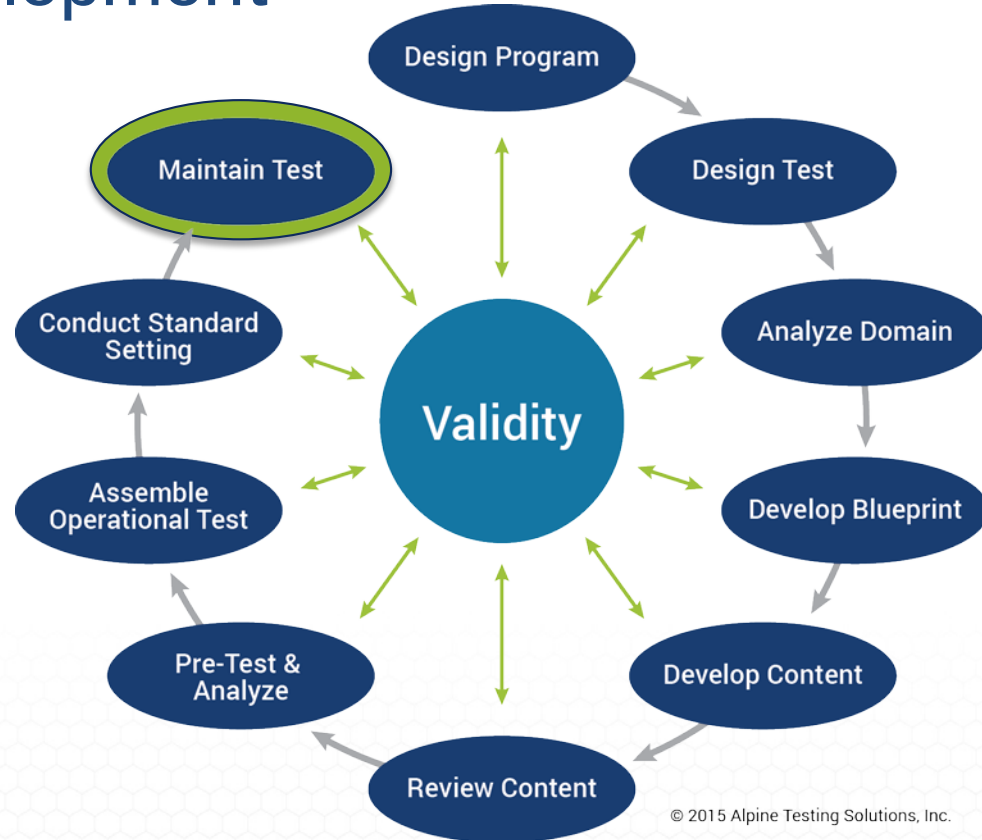
Lisa S. O'Leary, Ph.D.
Senior Psychometrician



A good test...

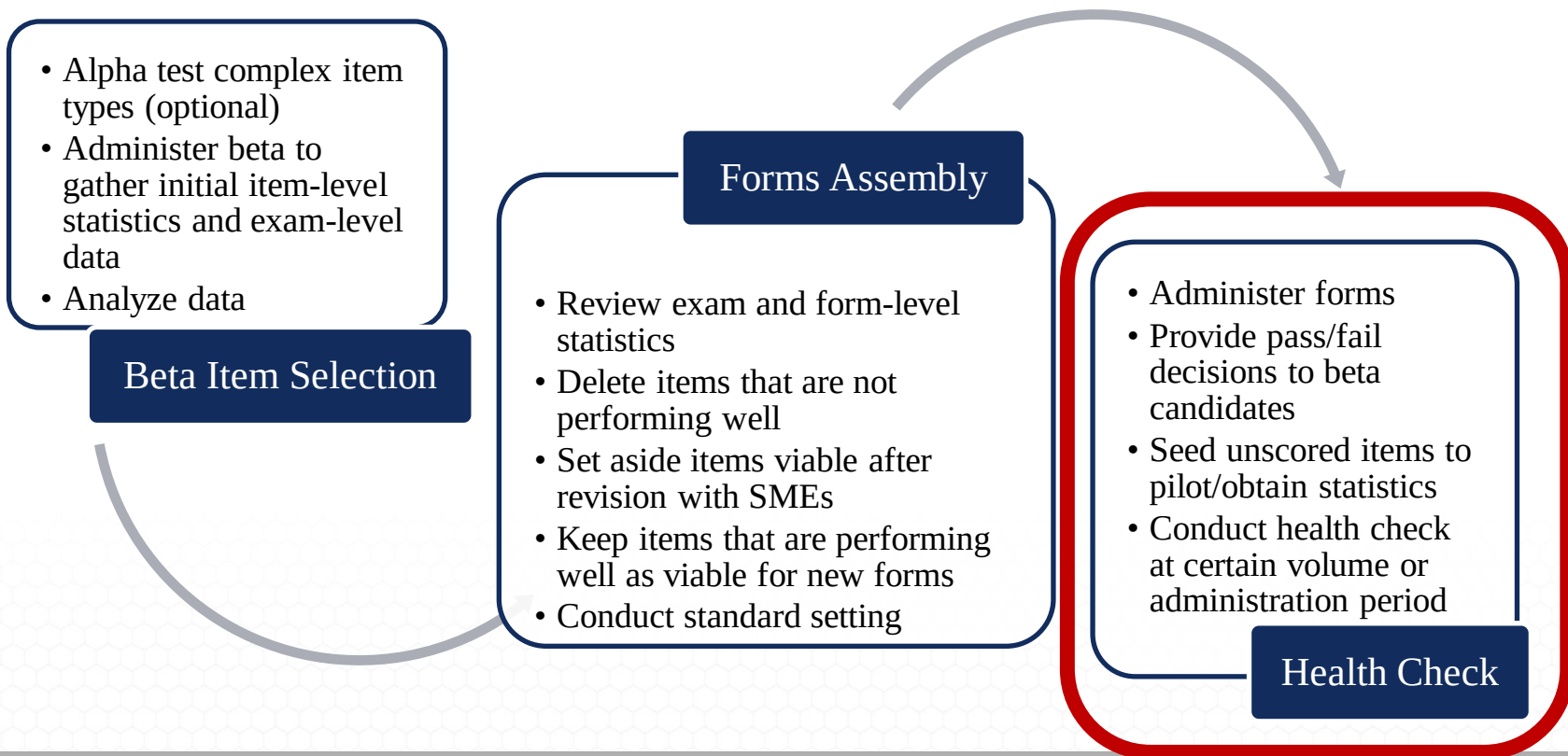


Exam Development



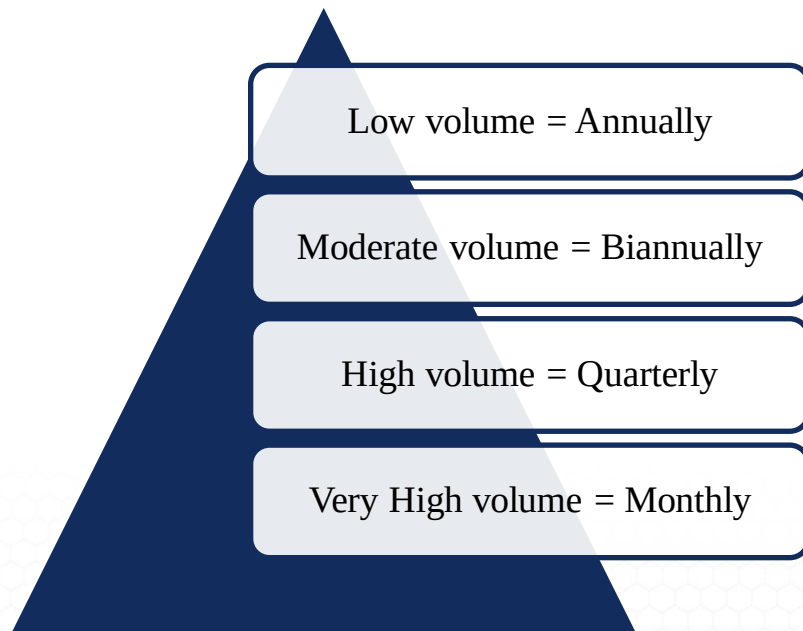
© 2015 Alpine Testing Solutions, Inc.

Exam Release Cycle



When to Conduct Health Checks

- » Need to determine how often to re-visit exam forms, considering the following:
- Candidates volumes, including target audience sample size and representativeness,
 - domain and content relevancy and speed of changes/updates,
 - and exposure or other security concerns.



Why Conduct Health Checks

- » Evaluate statistical data regarding form and item-level performance during operational administration
- » Continually provide evidence of the following:
 - Adherence to the defined purpose of the exam
 - Quality of psychometric and statistical attributes
 - Appropriateness of standard setting results
 - Exposure and security review
 - Evaluation of fairness
 - Alignment with policy and administrative goals
- » Inform future decisions regarding exam, forms, and items

Why Conduct Health Checks

- » Provide evidence of the health of an exam and its items
 - **Use:** Track exam volumes and pass rates over time
 - **Performance:** Ensure forms and items are functioning as intended in operational environment
 - **Exposure:** Track both item- and form-level exposure to address security concerns
- » Provide support that the interpretation of exam scores remains appropriate over time

Item-Level Analyses

- » Diagnoses potential item issues with:
 - Item wording (stems and distractors),
 - Scoring (keys, options, logic, points possible), and
 - Relationship between particular items and the entire test.
- » Enables a selection of the final items viable for operational forms re-assembly

Item-Level Analyses Interpretation

Classical Item Difficulty



		Rasch Item Difficulty		Item-Score Correlation	Item Reliability
Count	Item ID	Measure	P-value		
19	1.2.081	0.11	0.799	0.252	0.101
20	1.2.098	-4.80	1.000	0.000	0.000
21	1.2.107	0.80	0.667	0.232	0.110
22	1.2.111	-0.40	0.863	0.348	0.120
23	1.2.133	-0.21	0.826	0.434	0.166
24	1.2.144	0.02	0.829	0.483	0.183
25	1.2.146	-1.93	0.964	0.276	0.052
26	1.2.170	-1.56	0.950	0.380	0.083
27	1.2.181	-0.73	0.884	0.457	0.147
28	1.3.010	-2.47	0.978	0.439	0.064
29	1.3.019	-0.21	0.826	0.285	0.109
30	1.3.035	-0.49	0.886	0.291	0.093

p-value

- Item difficulty for dichotomous items (0,1)
- Proportion of candidates who answered the item correctly
- Ranges from 0 to 1, or 0% to 100%
- High values indicate easier items; low values indicate hard items
- Lower values indicate easier items; higher values indicate more difficult items

Item-Level Analyses Interpretation

Classical Item Difficulty



Count	Item ID	Rasch Item Difficulty Measure	Avg. Item Score	Item-Score Correlation	Item
1	1.1.1.3	-0.05	5.833	0.245	
2	1.2.1.3	0.09	3.050	0.067	
3	1.2.2.3	-0.31	4.233	0.349	
4	1.3.1.3	0.06	2.000	0.217	
5	1.4.1.3	0.01	5.583	0.345	
6	2.1.1.3	-1.10	7.800	0.245	
7	2.2.1.3	0.10	4.817	0.270	
8	2.3.1.3	-0.26	2.900	0.557	
9	2.4.1.3	0.14	2.900	0.407	
10	2.5.1.3	0.03	2.250	0.219	
11	2.5.2.3	0.03	9.867	0.529	

Average Item Score

- Item difficulty for polytomous items (0 through maximum points value)
- Average number of score points earned by candidates
- Ranges from 0 to maximum number of points
- Interpret on the scale of the maximum number of points

0.818	328	0.071	245.0
0.682	328	0.071	288.5
0.272	328	0.071	269.5
3.018	328	0.071	701.0

Item-Level Analyses Interpretation

IRT Item Difficulty



Rasch Item Difficulty Measure

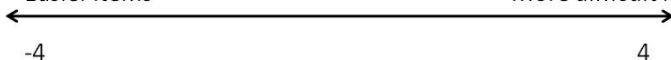
- Item difficulty for dichotomous and polytomous items on the Rasch scale
- Approximate scale of -4 to +4, with item difficulty typically centered at 0
- Lower values indicate easier items; higher values indicate more difficult items
- Estimate of item difficulty is equal to the ability level of the candidate who has a 50% probability of answering the item correctly

Alpine Testing Software				
Rasch Item Difficulty Measure				
Count	Item ID	Measure	Avg. Item Score	Correlation
1	1.1.2.3	0.00	2.419	0.125
2	1.1.4.3	0.80	0.291	0.012
3	1.2.1.3	0.27	1.791	0.221
4	1.2.3.3	0.00	2.860	0.319
5	1.3.1.3	-0.10	3.070	0.372
6	1.3.2.3	-0.75	4.779	0.091
7	2.1.2.3	0.31	1.570	0.103
8	2.2.2.3	-0.21	19.733	0.506
9	2.3.1.3	0.14	3.518	0.175
10	2.3.4.3	-0.43	86	411.5
11	2.4.1.3	0.13		
12	2.4.2.3	-0.11		

Rasch scale puts candidate ability and item difficulty on the same scale

Lower ability candidates
Easier items

Higher ability candidates
More difficult items



NO

Item-Level Analyses Interpretation

Item Score Correlation



		Rasch Item Difficulty		Item-Score	
Count	Item ID	Measure	Avg. Item Score	Correlation	Item Relia
1	1.1.2.3	0.00	2.419	0.125	0.258
2	1.1.4.3	0.80	0.291	0.012	0.009
3	1.2.1.3	0.27	1.791	0.221	0.351
4	1.2.3.3	0.00	2.860	0.319	0.656
5	1.3.1.3	-0.10	3.070	0.372	0.493
6	1.3.2.3	-0.75	4.779	0.091	0.131
7	2.1.2.3	0.31	1.570	0.103	0.239
8	2.2.2.3	-0.21	19.733	0.506	3.518
9	2.3.1.3	0.14	2.500	0.216	0.269
10	2.3.4.3	-0.43	5.314	0.273	0.529
11	2.4.1.3	0.13	3.151	0.142	0.240
12	2.4.2.3	-0.11	17.058	0.262	1.003

Item Score Correlation

- **Point Biserial Correlation** for dichotomous items
- How well an item differentiates between high and low ability candidates
- Estimated by performance on the exam---typically relationship between performance on the item and total score (although other values can be used)
- Range from -1 to 1
- Strong + correlations = item discriminates well between candidates; high ability candidates answer item correctly/low ability answer incorrectly
- Low + or - correlations = item does not discriminate between candidates; high ability candidates answer item incorrectly or low ability answer correctly

Item-Level Analyses Interpretation

Item Reliability



Rasch Item Difficulty		Item-Score		Item Reliability	Number
Item ID	Measure	P-value	Correlation		
1.2.081	0.11	0.799	0.252	0.101	
1.2.098	-4.80	1.000	0.000	0.000	
1.2.107	0.80	0.667	0.232	0.110	
1.2.111	-0.40	0.863	0.348	0.120	
1.2.133	-0.21	0.826	0.434	0.166	
1.2.144	0.02	0.829	0.483	0.183	
1.2.146	-1.93	0.964	0.276	0.052	
1.2.170	-1.56	0.950	0.380	0.083	
1.2.181	-0.73	0.884	0.457	0.147	
1.3.010	-2.47	0.978	0.439	0.064	
1.3.019	-0.21	0.826	0.285	0.109	
1.3.035	-0.49	0.886	0.291	0.093	

Item Reliability

- Measure of internal consistency
- Degree to which an item is contributing to and measuring content in the same way as the test overall
- Range from -1 to 1
- High + values = item contributing to the overall reliability of the exam, strong relationship between what item is measuring and overall test
- Low + values = item not contributing to the overall reliability of the exam
- - values = item is reducing overall exam reliability, inverse relationship between what item is measuring and overall test

Item-Level Analyses Interpretation

Response Time

Item Response Time

- Median amount of time candidates spend on an item, presented in seconds
- Short amount of time = candidates are responding to the item quickly
- Long amount of time = candidates are taking longer to complete the exam
- Should have inverse relationship to p-value
 - Easy items should have short average item response times
 - Hard items should have longer average item response times

Count		Item Selection Spreadsheet				
		Responses	Critical Correlation	Time	P-Value	Correlation
19			0.027	59.0		
20			0.027	29.0	TE	NO
21			0.038	52.0		
22			0.027	73.0		
23			0.038	35.0		
24			0.038	16.0		
25			0.027	11.0	TE	
26			0.027	33.0	TE	
27			0.038	54.0		
28			0.027	24.0	TE	
29			0.038	19.0		
30			0.038	15.0		

Item-Level Analyses Interpretation

Flagging of Poorly Performing Items

Testing Solutions						
P-value	Item-Score Correlation	Critical Correlation	P-Value	Correlation	ATS Comments	Final Decision
0.928	0.202	0.038	TE	4		Delete
0.957	0.301	0.038	TE			Delete
0.884	0.481	0.038				Keep
0.614	0.084	0.038				Keep
0.725	0.284	0.038				Keep
0.771	0.249	0.038				Keep
0.377	0.530	0.038				Keep
0.914	0.381	0.038	TE			Delete
0.884	0.360	0.038				Keep
0.629	-0.027	0.038			NO	Delete
0.435	0.424	0.038				Keep
0.899	0.312	0.027				Keep
0.671	0.446	0.038				Keep
0.892	0.240	0.027				Keep
0.971	0.165	0.038	TE			Delete

Item Flagging

- Items with issues based on their statistical performance
- Default parameters can be set depending on exam situation (CUSTOMIZABLE)
 - p -values
 - Items with p -values > 0.9 = “too easy”
 - Items with p -values < 0.1 = “too hard”
 - Item score correlation
 - Items with correlation $<$ critical = “no”
 - Items with – correlation = “neg”
 - Option analysis
 - Letter of incorrect response with higher correlation, p -value, or high scoring candidates than correct option

Item-Level Analyses Interpretation

Recommendations for Item Status

Item Decisions

- Recommended keep/delete decisions for each item based on item statistics
- Recommendations should be reviewed in conjunction with SMEs
 - Review candidate comments
 - Evaluate too easy/too hard items
 - Review items with correlation issues
 - Look at response options for possible miskeys and problematic distractors

103	0.128	26.0							
105	0.127	20.0	TE	NO					
103	0.128	29.0							
105	0.127	46.0							
109	0.124	37.0							

Comments	Final Decision	Client Comments	Form(s)
	Delete		A
	Keep		A
	Keep		A
	Keep		B
	Keep		B
	Delete	recall	B
	Keep		A, B
	Keep		A, B
	Keep		A, B
	Delete	recall	B
	Keep		B
	Keep		A
	Keep	Anchor	A
	Keep		B
	Delete		A
	Keep		B
	Keep		A
	Keep		A, B

Item-Level Analyses Interpretation

Option Analyses

- » Provides breakdown of how well each response is performing as a correct (key) or incorrect (distractor) answer
- p-value: Distractors with p -values higher than the key
 - Item-Score Correlation: Distractors with high positive correlations
 - Frequency count: Distractors selected frequently by high scorers

2.3.5.0	option	p-value	correlation	avg. time	25 to 48	49 to 60	61 to 79	80 to 91	92 to 100
	AB	0.053	-0.203	26	3	1	1		
	AC	0.011	-0.159	15	1				
	AD	0.021	-0.059	45		1	1		
	BC	0.266	0.394	11	2		6	8	9
>	BD	0.638	-0.218	23	18	11	14	8	9
	DC	0.011	0.006	24			1		

Item Selection Interpretation



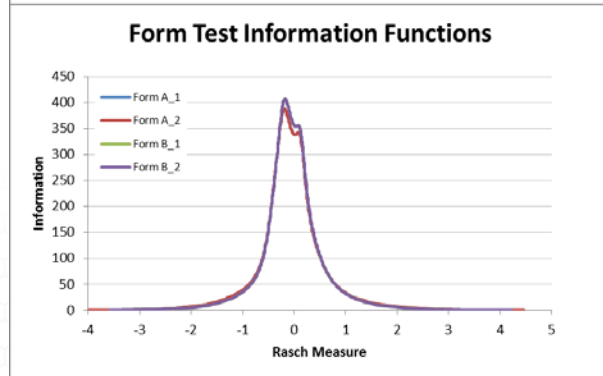
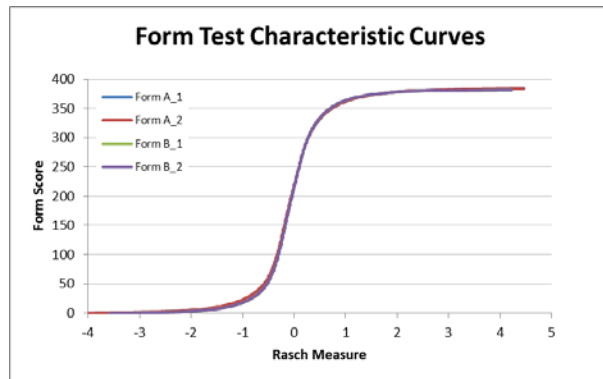
Situation	Indicative item statistics	Recommendation
Very easy item	High p -value, low Rasch measure, low correlation or reliability	Item concept is easy = delete the item Item wording is easy = revise and re-pilot
Very difficult item	Low p -value, high Rasch measure, low correlation or reliability	If content is irrelevant = delete the item If important relevant content = review item wording to check if it is over-complicated
Item miskeyed	Low p -value, low correlation, low reliability, high correlation on option	Fix item key and re-pilot
Multiple correct answers	Low reliability, aberrant option analysis results, longer item response time, high frequency counts in upper quintile(s) of option analysis	Identify truly correct answer, revise or remove problematic distractors(s) and re-pilot
Item is targeting different content than the exam	Low reliability, low correlation	If content is irrelevant = delete the item If important relevant content = refocus item towards intended content
Item compromise	High p -values, low Rasch measures, short item response time	Temporarily or permanently remove item from operational forms

Form-Level Analyses Interpretation

Health Checks

- » Form-level analysis shows test-level statistics by form
- » Differences in difficulty indicative of non-equivalence

Health Check	Form A	Form B
Candidate Count	125	126
Exam Length	60	60
Mean	53.91	53.70
SD	7.98	7.32
Median	58.0	57.0
Mode	60	60
Avg. Time on Test	33.3	36.0
SD of Time on Test	22.4	22.7
Standard Error of the Mean	0.71	0.65
95% confidence interval +/-	1.40	1.28
Minimum	23	29
Maximum	60	60
Skewness	-1.63	-1.52
Kurtosis	2.02	1.73
Alpha Reliability	0.932	0.914
SEM	2.09	2.15
95% confidence interval +/-	4.09	4.22
Pass Rate	92.8%	93.7%
Decision Consistency (Livingston-Lewis)	0.966	0.968
Accuracy (Livingston-Lewis)	0.976	0.978
# Items in Test Pool	60	60



Form-Level Analyses Interpretation

Overall Exam Performance



Health Check	Form A
Candidate Count	579
Exam Length	70
Mean	44.40
SD	8.75
Median	41.0
Mode	44
Avg. Time on Test	69.8
SD of Time on Test	18.2
Standard Error of the Mean	0.76
95% confidence interval +/-	1.49
Minimum	18
Maximum	67
Skewness	-0.32
Kurtosis	-0.48
Alpha Reliability	0.898
SEM	3.39
95% confidence interval +/-	6.65
Pass Rate	69.9%
Decision Consistency (Livingston-Lewis)	0.842
Accuracy (Livingston-Lewis)	0.887
# Items in Test Pool	70

Mean: Average exam score of all examinees, difficulty of exam for candidates

Standard deviation: Variability in exam scores; higher values indicate scores vary greatly from the mean while lower values indicate scores are more closely clustered about the mean

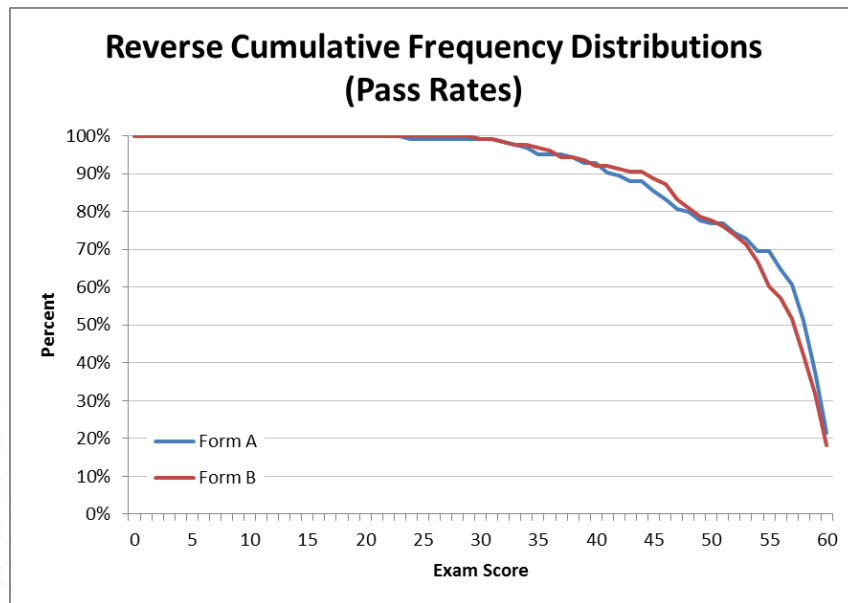
Total Test Time: Median amount of time candidates took on the entire exam; exams with short average time and high performance should be reviewed

Reliability: Consistency of items as an entire exam, how well the items as a test seem to be measuring the same knowledge, should be > 0.85 for certification exams

Form-Level Analyses Interpretation

Pass Rates

- » Contextualize how the cut score is functioning and relative equivalence of the forms across the ability spectrum



Form-Level Analyses Interpretation

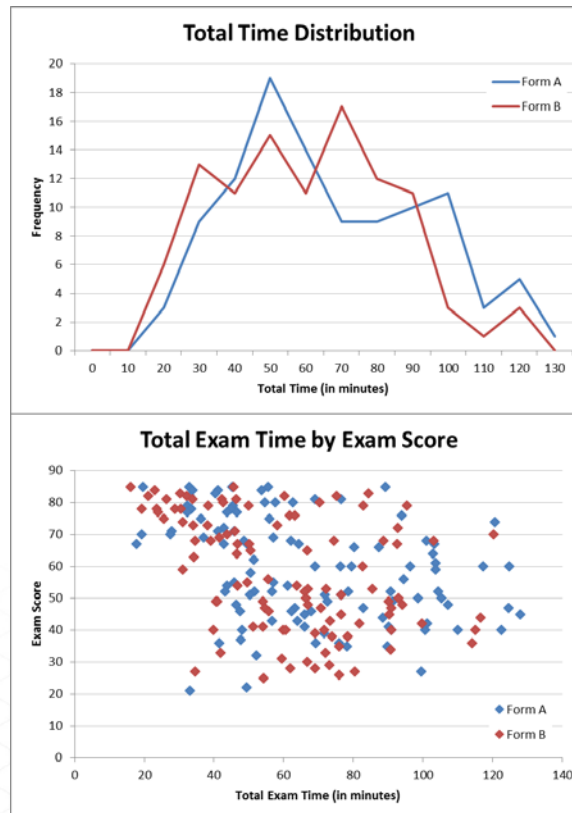
Overall Test Time

» Fairness

- Is a high proportion of candidates using the maximum time?
- What is the time at which 95% of candidates would complete the exam without speediness issues?

» Security

- Are candidates achieving high scores “too” quickly?
- Are candidates seeming to linger on the exam?



Form-Level Analyses Interpretation

Candidate Comments

Reg No	First Name	Last Name	Item Name	Answer	Response	Test Score	Comment
RegNum0011	FirstName11	LastName11	1.01.c.1	ABC	ABC	60	Comment 6
RegNum0010	FirstName10	LastName10	1.01.c.1	ABC	ABC	60	Comment 5
RegNum0018	FirstName18	LastName18	1.01.c.1	ABC	ABC	58	Comment 10
RegNum0017	FirstName17	LastName17	1.01.c.1	ABC	ABC	54	Comment 9
RegNum0009	FirstName9	LastName9	1.01.c.1	ABC	ABC	53	Comment 4
RegNum0006	FirstName6	LastName6	1.01.c.1	ABC	ABC	53	Comment 3
RegNum0014	FirstName14	LastName14	1.01.c.1	ABC	ABC	51	Comment 8
RegNum0013	FirstName13	LastName13	1.01.c.1	ABC	ABC	51	Comment 7
RegNum0001	FirstName1	LastName1	1.01.c.1	ABC	ABC	44	Comment 1
RegNum0004	FirstName4	LastName4	1.01.c.1	ABC	ABC	36	Comment 2

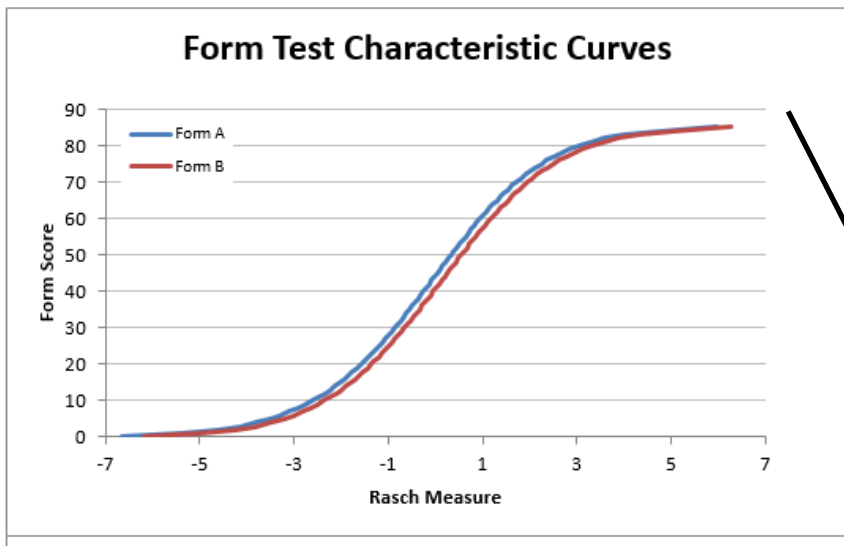
Candidate Comments

- Open-ended comments from candidates regarding their reactions to specific items on item content, stem, distractors
- Useful if reviewed in conjunction with item-level statistics, particularly for flagged items

B	60	Comment 8
B	59	Comment 10
B	59	Comment 6
B	58	Comment 9
B	58	Comment 7
B	57	Comment 5
B	56	Comment 1
B	55	Comment 2
A	46	Comment 3

Form-Level Analyses Interpretation

Test Characteristic Curves



Form A

Score	Measure	S.E.	Information
0	-6.65	1.85	0.29
1	-5.38	1.04	0.92
2	-4.61	0.76	1.75
3	-4.14	0.63	2.53
4	-3.80	0.55	3.28

Form B

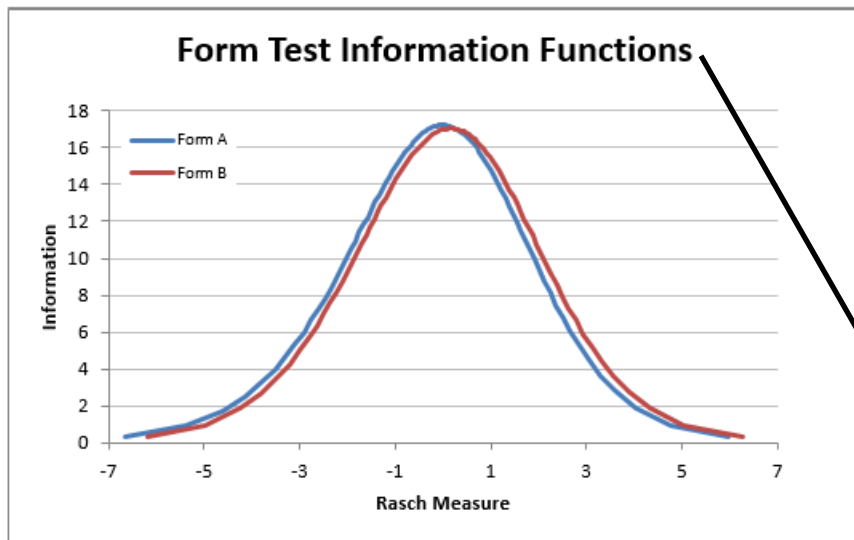
Score	Measure	S.E.	Information
0	-6.20	1.84	0.30
1	-4.97	1.02	0.96
2	-4.24	0.73	1.86
3	-3.80	0.61	2.70
4	-3.47	0.53	3.50

Test Characteristic Curves

- Relates examinee ability and raw exam score along the ability continuum
- Show the estimated number of items correct for a candidate of a given ability
- Coincidence of the TCCs is evidence of the comparability of the forms

Form-Level Analyses Interpretation

Test Information Functions



Form A

Score	Measure	S.E.	Information
0	-6.65	1.85	0.29
1	-5.38	1.04	0.92
2	-4.61	0.76	1.75
3	-4.14	0.63	2.53
4	-3.80	0.55	3.28
5	-3.52	0.50	3.99

Form B

Score	Measure	S.E.	Information
0	-6.20	1.84	0.30
1	-4.97	1.02	0.96
2	-4.24	0.73	1.86
3	-3.80	0.61	2.70
4	-3.47	0.53	3.50
5	-3.21	0.48	4.26

Test Information Functions

- Shows how much information is provided by each form at the target cut score
- Shows the distribution of information provided by the exam across the ability scale
- Coincidence of the TIFs is evidence of the comparability of the forms

Form-Level Analyses Interpretation

Blueprint Tracker



- » Compares blueprint specifications to recommended kept items and shows shortages (if applicable)
 - Depending on blueprint structure, results presented either in items or points and at section or objective-level

Section Count	Section	BluePrint	Kept Item Count
1	1	7	5
2	2	33	34
3	3	20	16

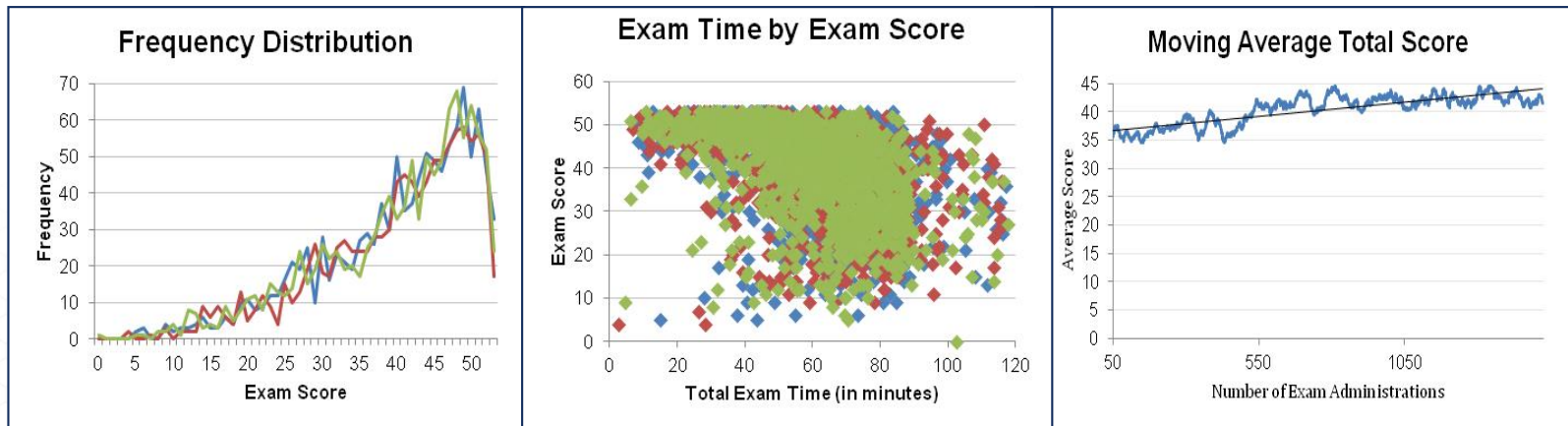
Objective Count	Objective	BluePrint	Kept Item Count	Comments
1	1.01	1	0	Not enough items to meet BP
2	1.02	1	0	Not enough items to meet BP
3	1.03	2	2	
4	1.04	3	3	
5	2.01	2	2	
6	2.02	1	0	Not enough items to meet BP
7	2.03	2	2	
8	2.04	2	2	
9	2.05	3	4	

Form-Level Analyses Interpretation

Security Checks

» Several indicators of possible form exposure and candidate pre-knowledge:

- Left-skewed score distribution with high frequency of high-scoring candidates
- High incidence of high scoring candidates in low amounts of time
- Upward trends/increases of average scores over time



Form-Level Analyses Interpretation

Anomalous Records

- » Lists those candidates that were identified for potentially suspect candidate behavior or exam performance based on default flags
 - Spending less than 5 seconds on 20% or more of the items
 - Scoring greater than 90% in less than 10 minutes
 - Data administration anomalies

Reg No	Decision	First Name	Last Name	Company	Test Date	Testing Center	Reason Flagged
RegNum0012	Kept	FirstName12	LastName12	Company18	1/18/2006	TestCenter4	Scored 98.33% in 7.5 minutes
RegNum0021	Kept	FirstName21	LastName21	Company20	1/19/2006	TestCenter3	Spent <5 Secs on 20% or more of items / Scored 100% in 8.9 minutes
RegNum0053	Kept	FirstName53	LastName53	Company5	2/16/2006	TestCenter2	Spent <5 Secs on 20% or more of items / Scored 98.33% in 8.7 minutes
RegNum0071	Kept	FirstName71	LastName71	Company14	3/2/2006	TestCenter5	Spent <5 Secs on 20% or more of items / Scored 100% in 5.3 minutes
RegNum0083	Kept	FirstName83	LastName83	Company3	3/11/2006	TestCenter6	Spent <5 Secs on 20% or more of items / Scored 100% in 6.9 minutes
RegNum0087	Kept	FirstName87	LastName87	Company14	3/10/2006	TestCenter2	Spent <5 Secs on 20% or more of items
RegNum0124	Kept	FirstName124	LastName124	Company6	3/26/2006	TestCenter6	Spent <5 Secs on 20% or more of items / Scored 95% in 7.3 minutes
RegNum0129	Kept	FirstName129	LastName129	Company15	3/27/2006	TestCenter1	Scored 100% in 8.9 minutes

Health Check Interpretation

- » Advocate for the involvement of at least 2-3 SMEs
 - Provide feedback on item-level results
 - Review candidate comments
 - Provide necessary changes/edits to items with option analysis flags and correlational issues for re-seeding
 - Gauge relevancy of content to current domain
 - Assist with final keep/delete decisions
 - Provide insight into form-level concerns
 - Candidates deviating largely from time or performance expectations
 - Differential pass rates by form or other demographic variable

What Next?

Re-Assemble Forms

- » **Equate** to the existing cut score to ensure fair scoring and equivalent score interpretation across versions
- » **Balance** content, item and form difficulty, reliability, variance and test time across forms
- » **Scale** the equated cut score to the reporting score scale increase interpretability and meaning of candidates' raw scores
- » **Maximize** content relevancy and item quality by replacing older items with previously unscored items
- » **Minimize** item exposure by keeping item overlap low and retiring items with known performance issues

What Next?

Exam Refresh Plan

- » Keep exam, content, and blueprint up-to-date and aligned with exam purpose
- » Be flexible to domain and content shifts while still maintaining the practical, psychometric, and statistical integrity of the exam
- » Extent of changes will drive necessary exam maintenance activities
 - Developments in the purpose of the exam? Changes in the definition of the domain of interest? Shifts in the definition or expectations of the MQC?
 - Consider the necessary schedule, scope, budget, level of effort, and available resources to support continual exam maintenance
 - Constraints may affect capacity/frequency of exam maintenance activities
though these are critical to valid use of the exam scores

References

- American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.
- Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*. Mahwah, NJ; Lawrence Erlbaum Associates, Publishers.
- Crocker, L., & Algina, J. (1986). *Introduction of classical and modern test theory*. New York, NY: Harcourt Brace Jovanovich College Publishers.
- Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking* (3rd ed.). New York, NY: Springer.
- Lord F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison Wesley.
- Rasch, G. (1960) *Probabilistic Models for Some Intelligence and Attainment Tests*. (Copenhagen: 1960) Chicago, IL: MESA Press, 1992
- Wright, B.D., & Stone, M.H. (1979). *Best Test Design*. Chicago, IL: MESA Press.



QUESTIONS?