Alpine
Testing Solutions

Ask A Psychometrician: Psychometric Analyses & Operations

Lisa S. O’Leary, Ph.D.
September 3, 2014

Agenda

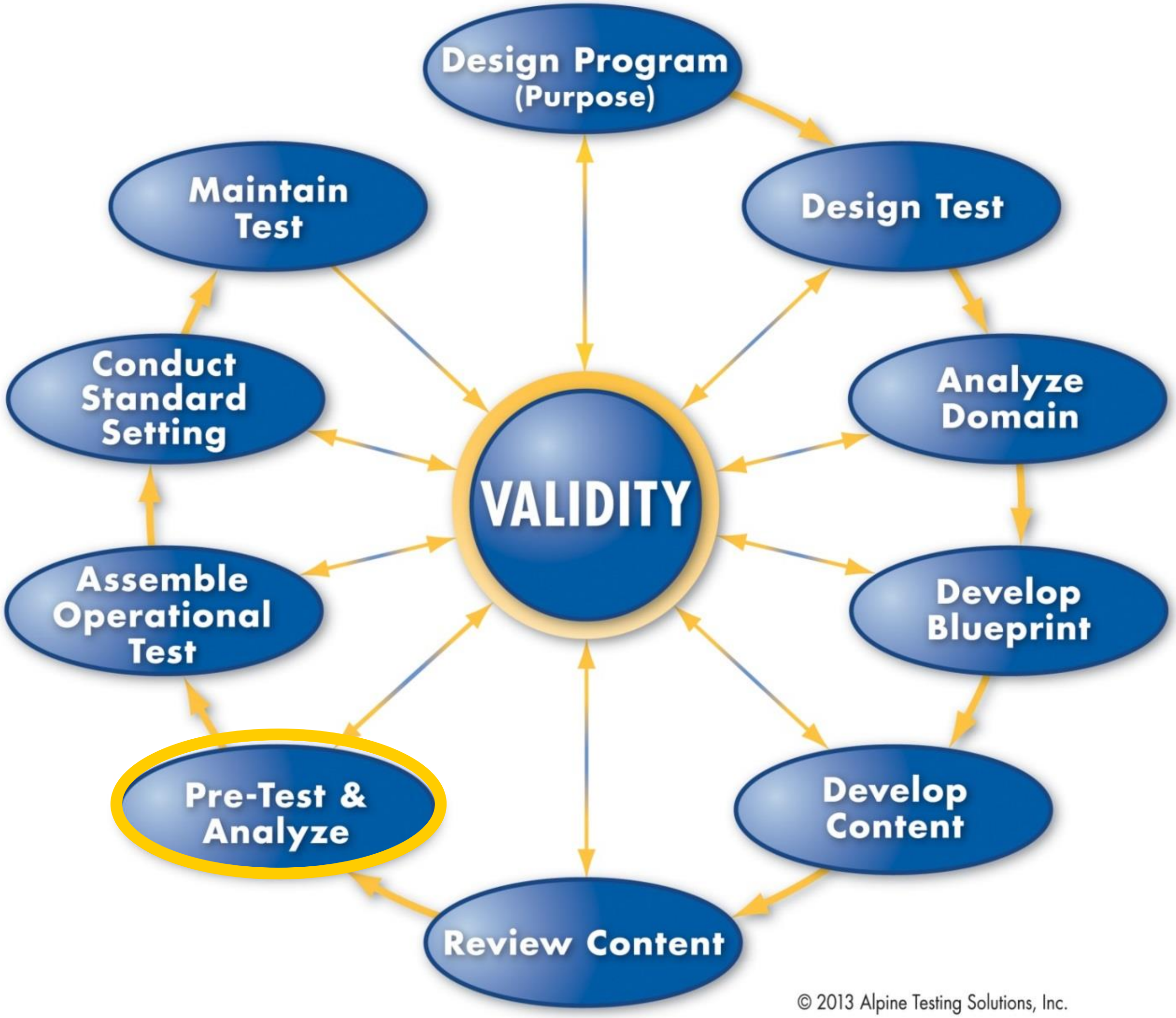
- Discussion of methods and statistical approaches for:
  - Beta Testing Prior to Live Form Administration
  - Item-Level and Form-Level Analyses
  - Forms Assembly & Equating

Alpine
Testing Solutions

Establish Baseline Parameters Through Effective & Advantageous

BETA TESTING

Stage in Process



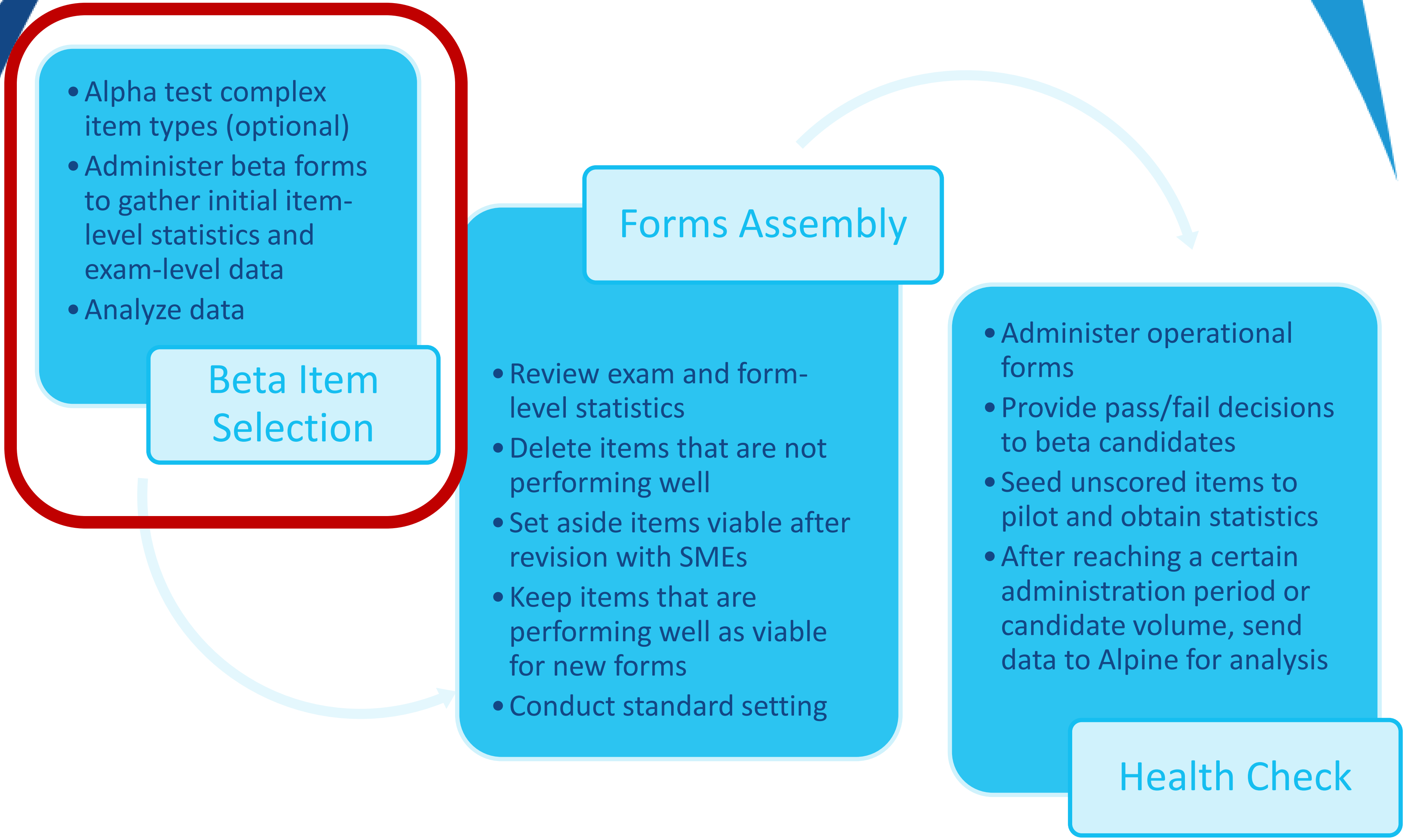
Alpine Testing Solutions

Purpose

- Gather evidence of appropriateness of the items to the content specifications and intended use of test scores
  - Can collect data during a beta exam or concurrent with operational administration

Beta Test	Seed Pilot Items
Purpose of the test program changes	Purpose of the test program remains the same
Significant content domain changes	Minimal content domain changes
Expectations or definition of MQC changes	Similar expectations or definition of the MQC
All new/large set of items are being developed	Small set of new items being developed

Exam Release Cycle



Beta Item Selection

- Alpha test complex item types (optional)
- Administer beta forms to gather initial item-level statistics and exam-level data
- Analyze data

Forms Assembly

- Review exam and form-level statistics
- Delete items that are not performing well
- Set aside items viable after revision with SMEs
- Keep items that are performing well as viable for new forms
- Conduct standard setting

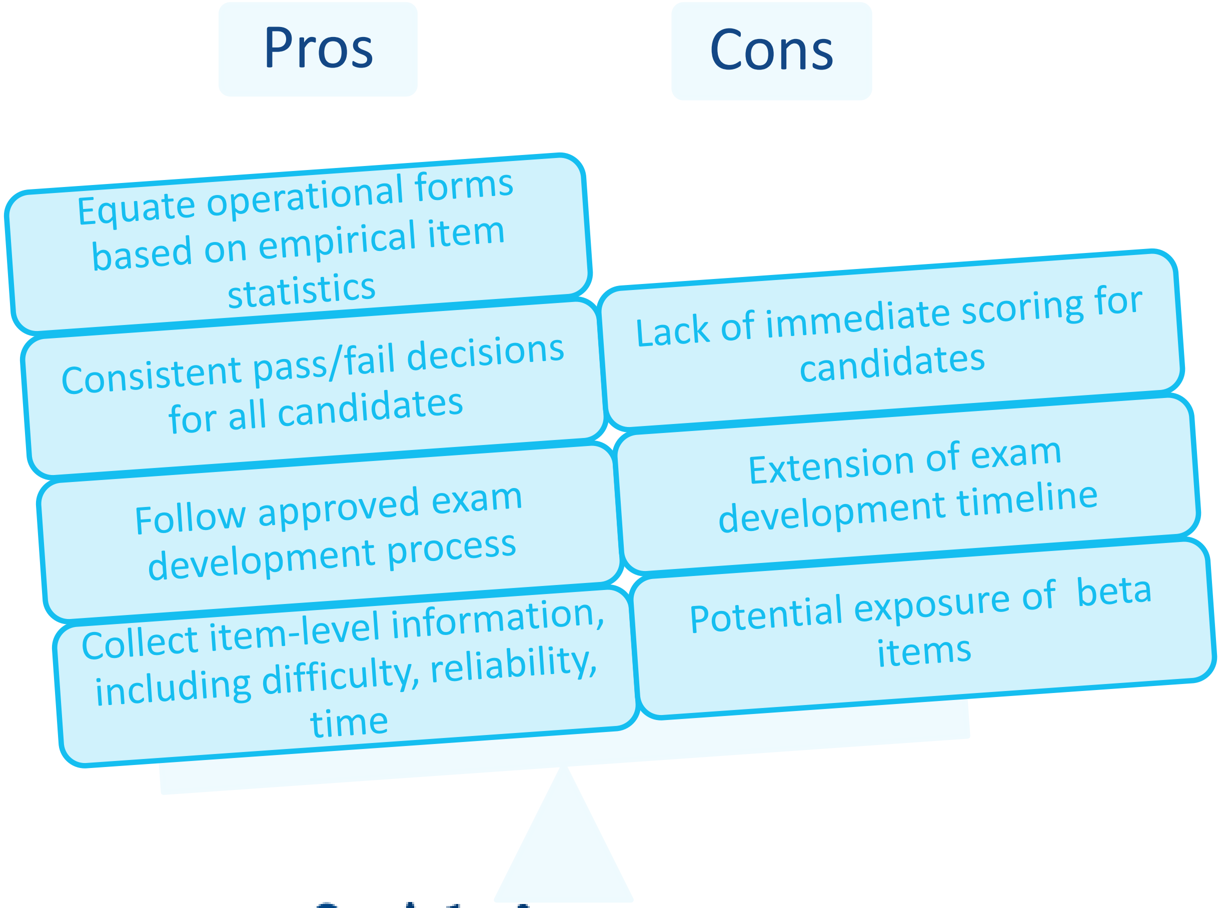
Health Check

- Administer operational forms
- Provide pass/fail decisions to beta candidates
- Seed unscored items to pilot and obtain statistics
- After reaching a certain administration period or candidate volume, send data to Alpine for analysis

Alpine
Testing Solutions

Beta Testing

- Decision on whether or not to beta test is based on a set of competing factors



Pros

- Equate operational forms based on empirical item statistics
- Consistent pass/fail decisions for all candidates
- Follow approved exam development process
- Collect item-level information, including difficulty, reliability, time

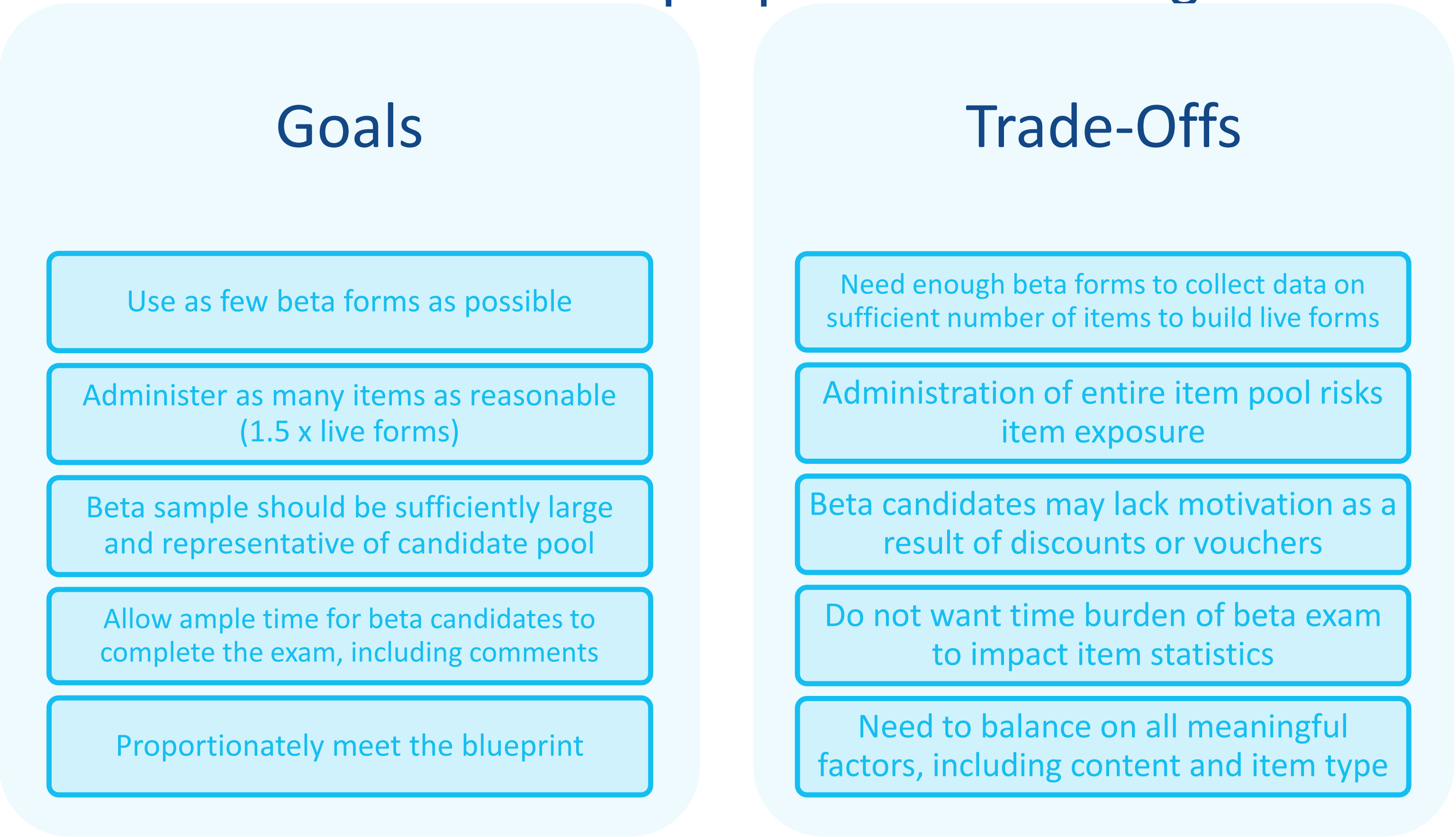
Cons

- Lack of immediate scoring for candidates
- Extension of exam development timeline
- Potential exposure of beta items

Alpine Testing Solutions

Beta Testing

- Need to determine the appropriate number of beta forms based on exam purpose and design



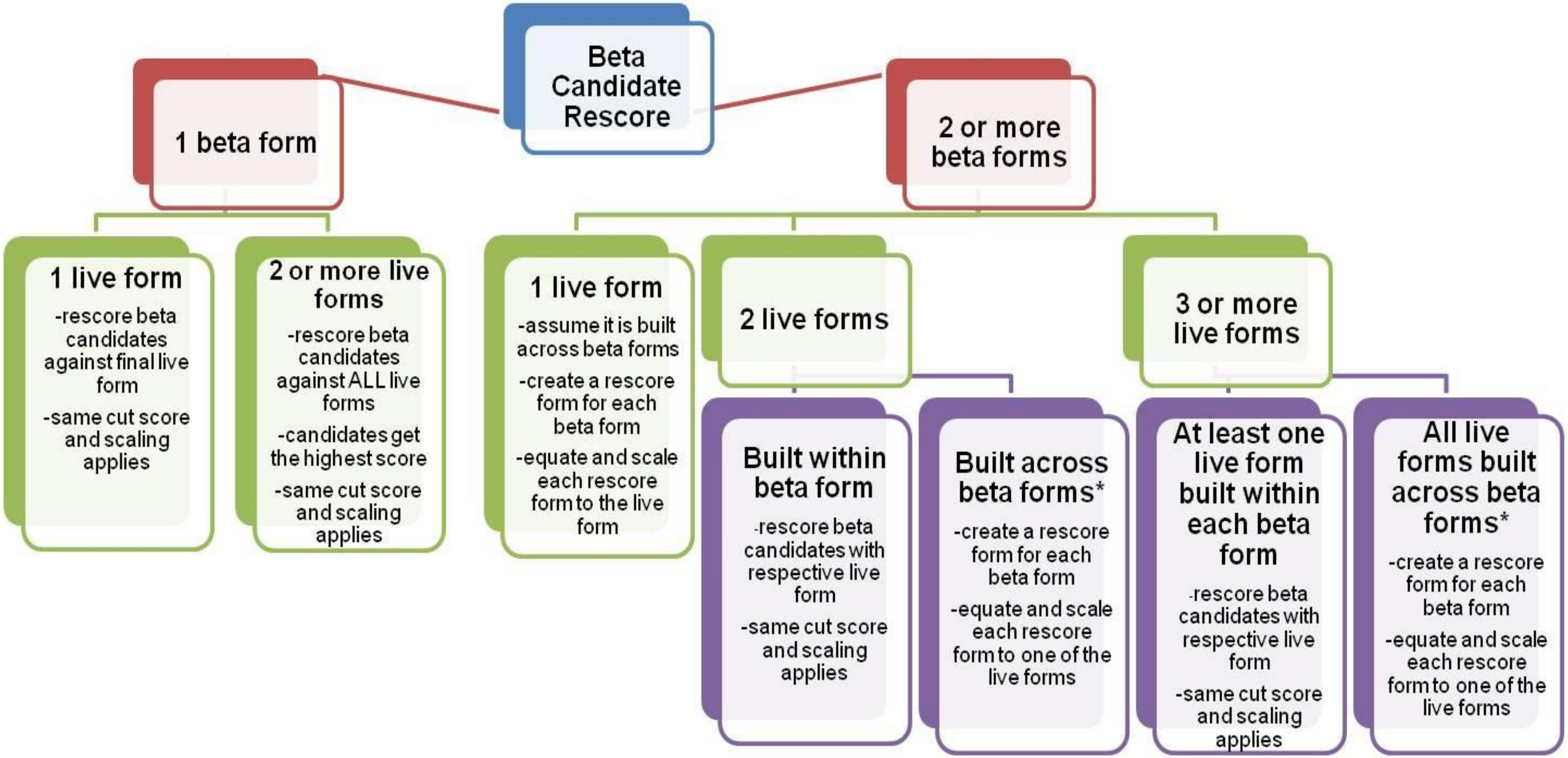
Goals

- Use as few beta forms as possible
- Administer as many items as reasonable (1.5 x live forms)
- Beta sample should be sufficiently large and representative of candidate pool
- Allow ample time for beta candidates to complete the exam, including comments
- Proportionately meet the blueprint

Trade-Offs

- Need enough beta forms to collect data on sufficient number of items to build live forms
- Administration of entire item pool risks item exposure
- Beta candidates may lack motivation as a result of discounts or vouchers
- Do not want time burden of beta exam to impact item statistics
- Need to balance on all meaningful factors, including content and item type

Beta Rescore Considerations



Beta Candidate Rescore

1 beta form

1 live form

-rescore beta candidates against final live form

-same cut score and scaling applies

2 or more live forms

-rescore beta candidates against ALL live forms

-candidates get the highest score

-same cut score and scaling applies

2 or more beta forms

1 live form

-assume it is built across beta forms

-create a rescore form for each beta form

-equate and scale each rescore form to the live form

2 live forms

Built within beta form

-rescore beta candidates with respective live form

-same cut score and scaling applies

Built across beta forms*

-create a rescore form for each beta form

-equate and scale each rescore form to one of the live forms

3 or more live forms

At least one live form built within each beta form

-rescore beta candidates with respective live form

-same cut score and scaling applies

All live forms built across beta forms*

-create a rescore form for each beta form

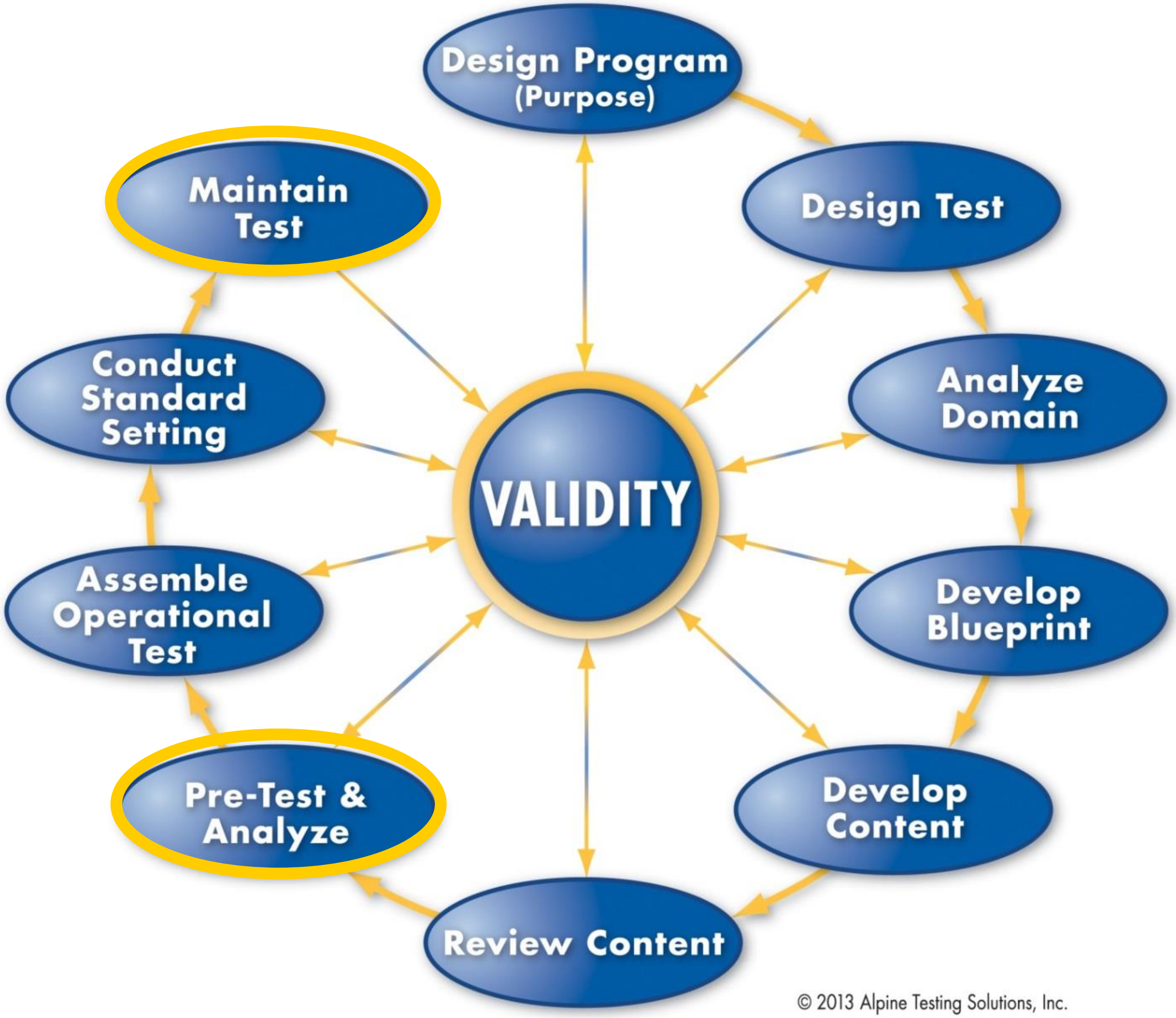
-equate and scale each rescore form to one of the live forms

Establish Item & Form Performance Through

ITEM & FORM ANALYSIS & STATISTICS

Alpine Testing Solutions

Stage in Process



Alpine Testing Solutions

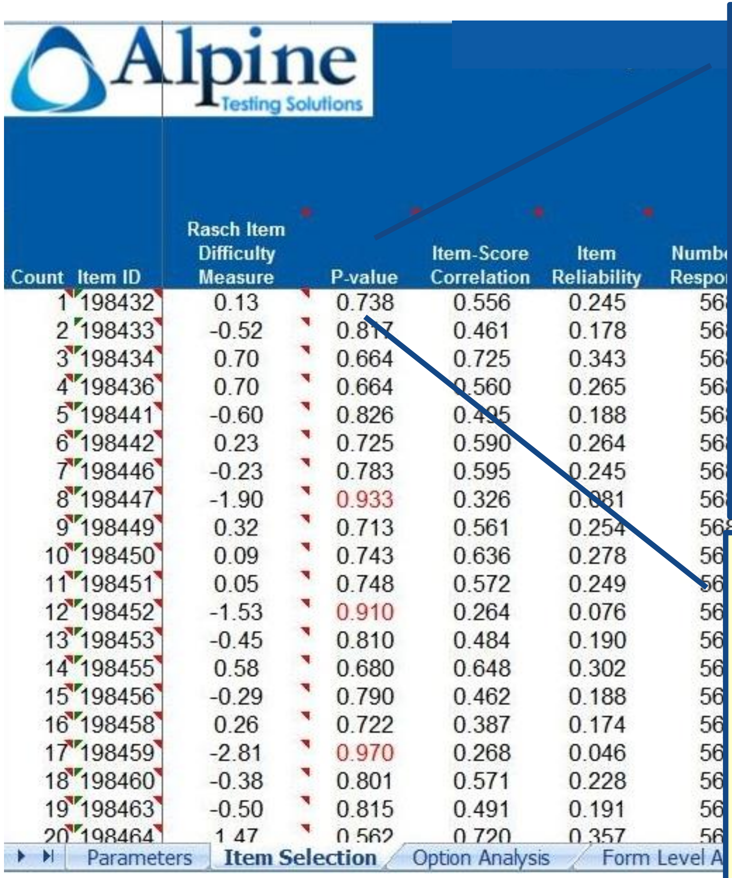
Item- & Form-Level Analyses

- 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

Item- & Form-Level Analyses

- 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 Statistics: Item Difficulty



Count	Item ID	Rasch Item Difficulty Measure	P-value	Item-Score Correlation	Item Reliability	Numb Respo
1	198432	0.13	0.738	0.556	0.245	56
2	198433	-0.52	0.817	0.461	0.178	56
3	198434	0.70	0.664	0.725	0.343	56
4	198436	0.70	0.664	0.560	0.265	56
5	198441	-0.60	0.826	0.495	0.188	56
6	198442	0.23	0.725	0.590	0.264	56
7	198446	-0.23	0.783	0.595	0.245	56
8	198447	-1.90	0.933	0.326	0.081	56
9	198449	0.32	0.713	0.561	0.254	568
10	198450	0.09	0.743	0.636	0.278	56
11	198451	0.05	0.748	0.572	0.249	56
12	198452	-1.53	0.910	0.264	0.076	56
13	198453	-0.45	0.810	0.484	0.190	56
14	198455	0.58	0.680	0.648	0.302	56
15	198456	-0.29	0.790	0.462	0.188	56
16	198458	0.26	0.722	0.387	0.174	56
17	198459	-2.81	0.970	0.268	0.046	56
18	198460	-0.38	0.801	0.571	0.228	56
19	198463	-0.50	0.815	0.491	0.191	56
20	198464	1.47	0.562	0.720	0.357	56

0.054 17.0 Keep

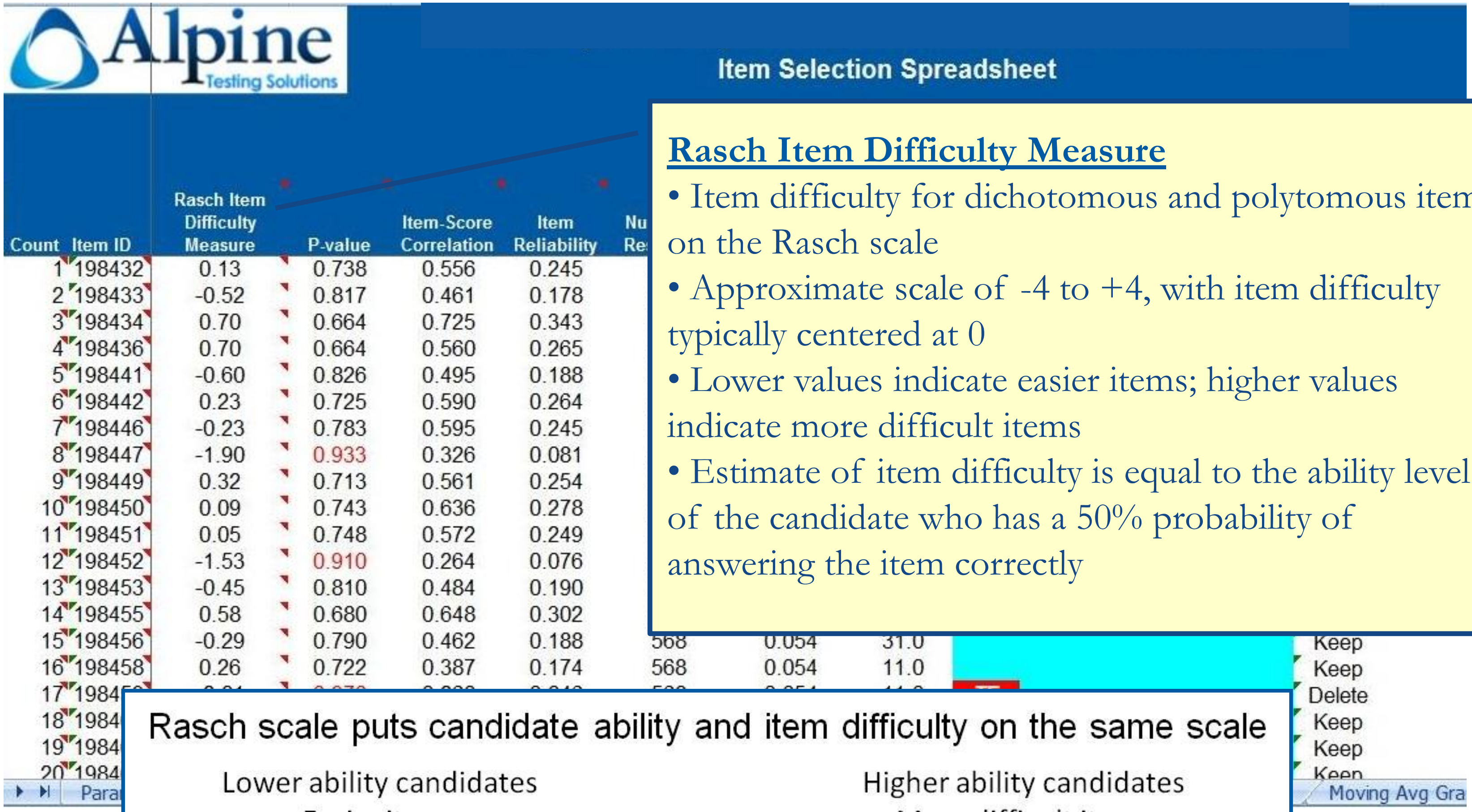
P-Value

- Item difficulty for dichotomous items (0,1) in CTT
- 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

Average Item Score

- Item difficulty for polytomous items (0 through maximum points value) in CTT
- 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

Item-Level Statistics: Item Difficulty



Item Selection Spreadsheet

Count	Item ID	Rasch Item Difficulty Measure	P-value	Item-Score Correlation	Item Reliability	Nu Re			
1	198432	0.13	0.738	0.556	0.245				
2	198433	-0.52	0.817	0.461	0.178				
3	198434	0.70	0.664	0.725	0.343				
4	198436	0.70	0.664	0.560	0.265				
5	198441	-0.60	0.826	0.495	0.188				
6	198442	0.23	0.725	0.590	0.264				
7	198446	-0.23	0.783	0.595	0.245				
8	198447	-1.90	0.933	0.326	0.081				
9	198449	0.32	0.713	0.561	0.254				
10	198450	0.09	0.743	0.636	0.278				
11	198451	0.05	0.748	0.572	0.249				
12	198452	-1.53	0.910	0.264	0.076				
13	198453	-0.45	0.810	0.484	0.190				
14	198455	0.58	0.680	0.648	0.302				
15	198456	-0.29	0.790	0.462	0.188	568	0.054	31.0	Keep
16	198458	0.26	0.722	0.387	0.174	568	0.054	11.0	Keep
17	1984								Delete
18	1984								Keep
19	1984								Keep
20	1984								Keep

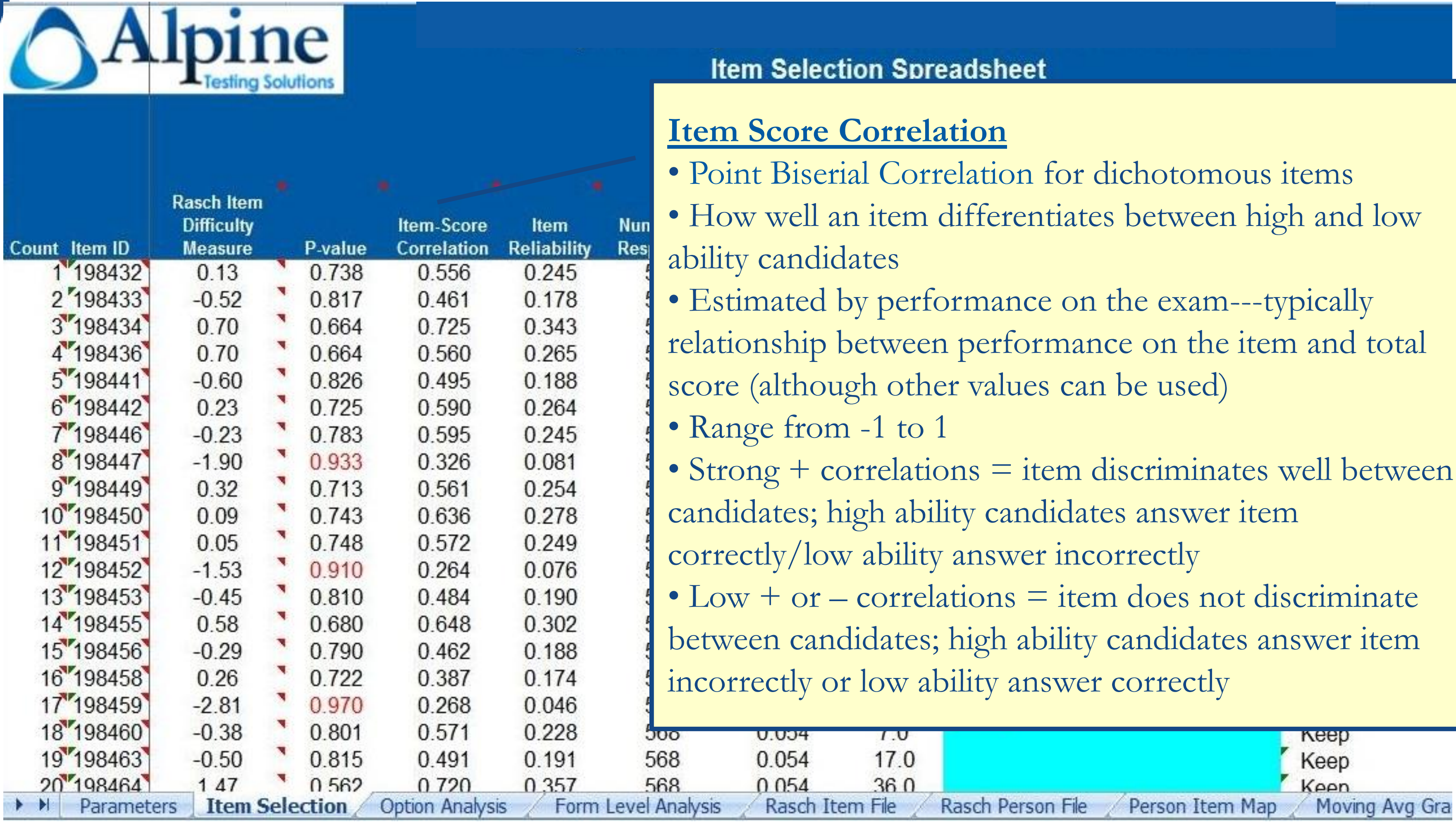
Rasch Item Difficulty Measure

- 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

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

Lower ability candidates Easier items	Higher ability candidates More difficult items
-4	4

Item-Level Statistics: Correlation



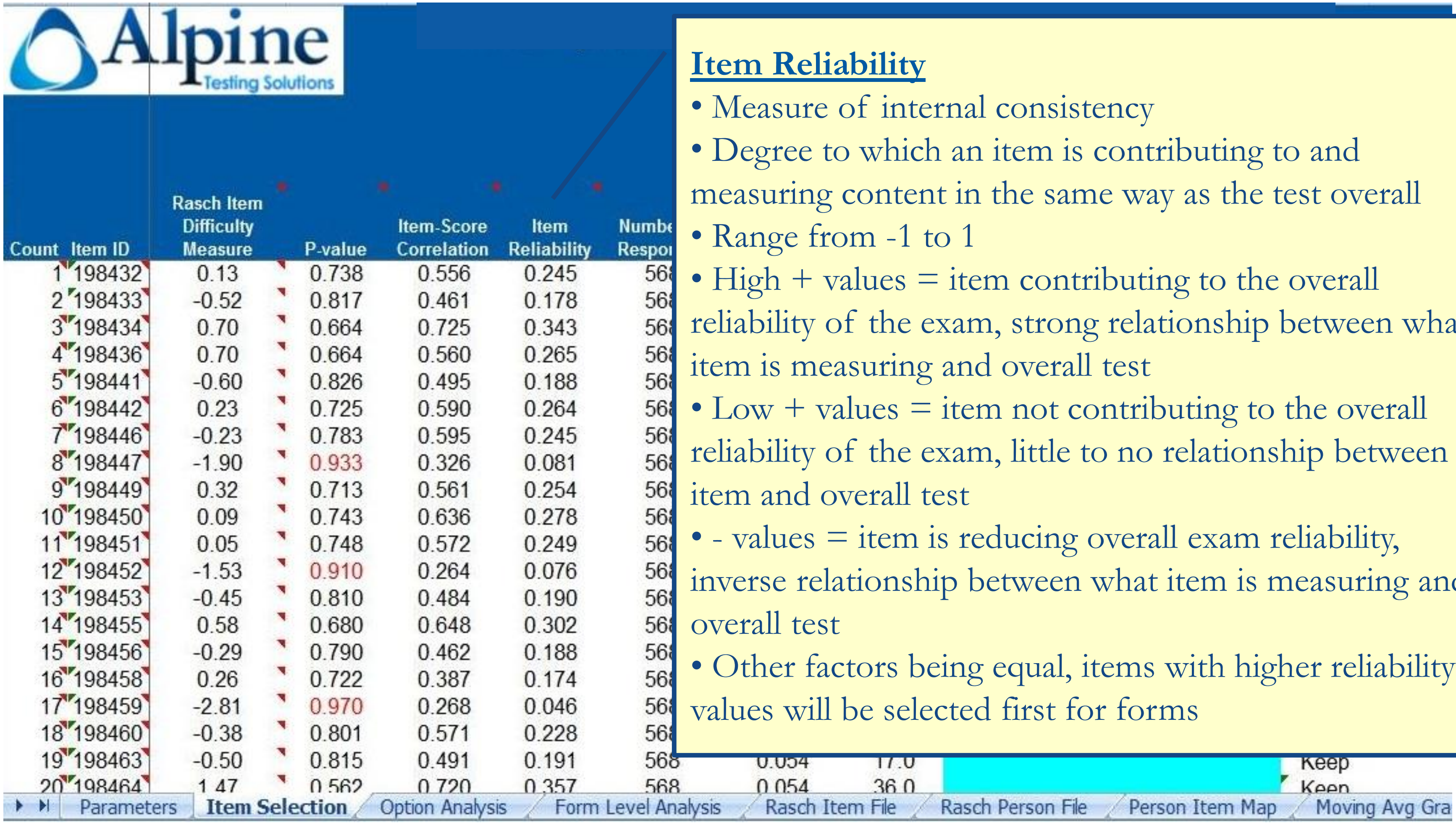
Item Selection Spreadsheet

Count	Item ID	Rasch Item Difficulty Measure	P-value	Item-Score Correlation	Item Reliability	Num Res				
1	198432	0.13	0.738	0.556	0.245					
2	198433	-0.52	0.817	0.461	0.178					
3	198434	0.70	0.664	0.725	0.343					
4	198436	0.70	0.664	0.560	0.265					
5	198441	-0.60	0.826	0.495	0.188					
6	198442	0.23	0.725	0.590	0.264					
7	198446	-0.23	0.783	0.595	0.245					
8	198447	-1.90	0.933	0.326	0.081					
9	198449	0.32	0.713	0.561	0.254					
10	198450	0.09	0.743	0.636	0.278					
11	198451	0.05	0.748	0.572	0.249					
12	198452	-1.53	0.910	0.264	0.076					
13	198453	-0.45	0.810	0.484	0.190					
14	198455	0.58	0.680	0.648	0.302					
15	198456	-0.29	0.790	0.462	0.188					
16	198458	0.26	0.722	0.387	0.174					
17	198459	-2.81	0.970	0.268	0.046					
18	198460	-0.38	0.801	0.571	0.228	568	0.054	7.0		Keep
19	198463	-0.50	0.815	0.491	0.191	568	0.054	17.0		Keep
20	198464	1.47	0.562	0.720	0.357	568	0.054	36.0		Keep

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 Statistics: Reliability

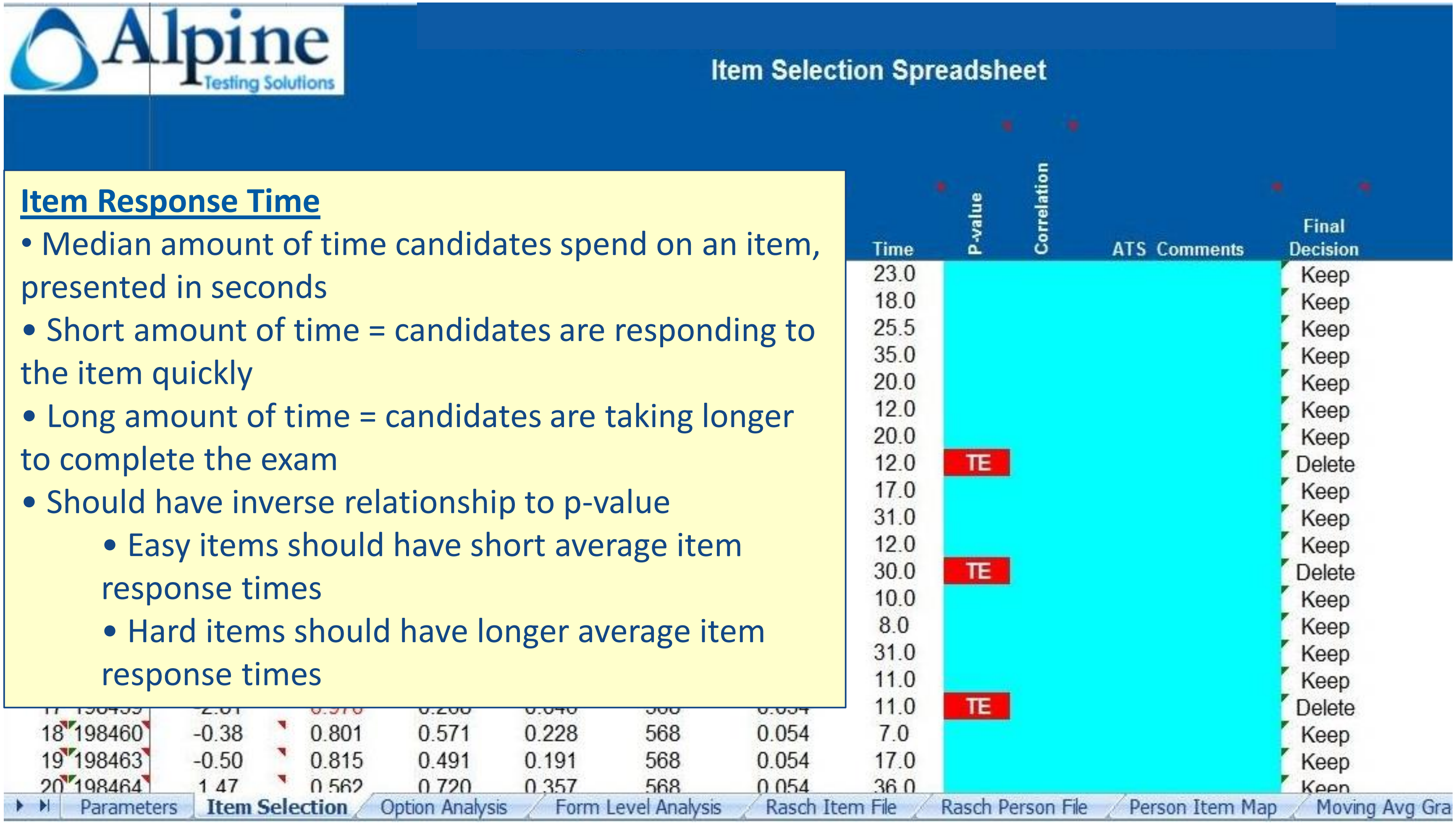


Count	Item ID	Rasch Item Difficulty Measure	P-value	Item-Score Correlation	Item Reliability	Numbe Respo			
1	198432	0.13	0.738	0.556	0.245	568			
2	198433	-0.52	0.817	0.461	0.178	568			
3	198434	0.70	0.664	0.725	0.343	568			
4	198436	0.70	0.664	0.560	0.265	568			
5	198441	-0.60	0.826	0.495	0.188	568			
6	198442	0.23	0.725	0.590	0.264	568			
7	198446	-0.23	0.783	0.595	0.245	568			
8	198447	-1.90	0.933	0.326	0.081	568			
9	198449	0.32	0.713	0.561	0.254	568			
10	198450	0.09	0.743	0.636	0.278	568			
11	198451	0.05	0.748	0.572	0.249	568			
12	198452	-1.53	0.910	0.264	0.076	568			
13	198453	-0.45	0.810	0.484	0.190	568			
14	198455	0.58	0.680	0.648	0.302	568			
15	198456	-0.29	0.790	0.462	0.188	568			
16	198458	0.26	0.722	0.387	0.174	568			
17	198459	-2.81	0.970	0.268	0.046	568			
18	198460	-0.38	0.801	0.571	0.228	568			
19	198463	-0.50	0.815	0.491	0.191	568	0.054	17.0	Keep
20	198464	1.47	0.562	0.720	0.357	568	0.054	36.0	Keep

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, little to no relationship between item and overall test
- - values = item is reducing overall exam reliability, inverse relationship between what item is measuring and overall test
- Other factors being equal, items with higher reliability values will be selected first for forms

Item-Level Statistics: Response Time



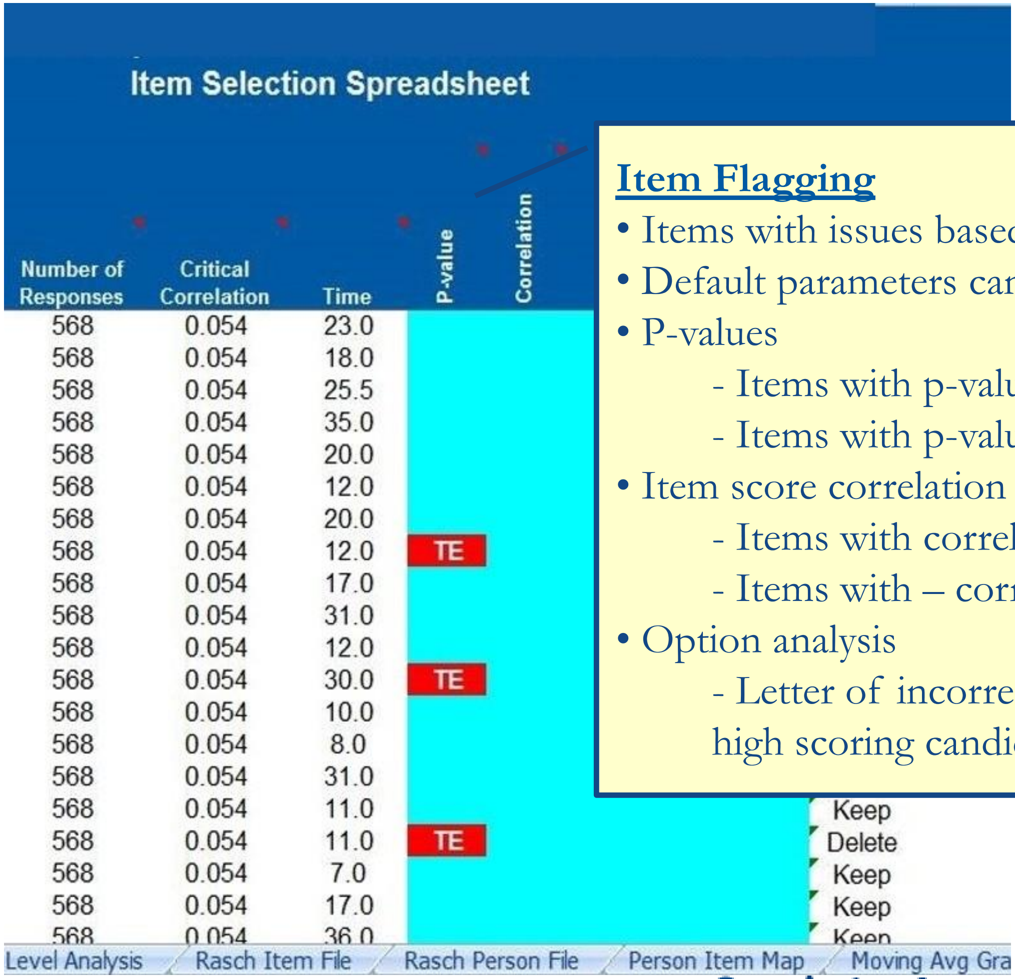
Item Selection Spreadsheet

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

								Time	P-value	Correlation	ATS Comments	Final Decision
								23.0				Keep
								18.0				Keep
								25.5				Keep
								35.0				Keep
								20.0				Keep
								12.0				Keep
								20.0				Keep
								12.0	TE			Delete
								17.0				Keep
								31.0				Keep
								12.0				Keep
								30.0	TE			Delete
								10.0				Keep
								8.0				Keep
								31.0				Keep
								11.0				Keep
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	11.0	TE			Delete
18	198460	-0.38	0.801	0.571	0.228	568	0.054	7.0				Keep
19	198463	-0.50	0.815	0.491	0.191	568	0.054	17.0				Keep
20	198464	1.47	0.562	0.720	0.357	568	0.054	36.0				Keep

Item-Level Statistics: Identification of Poorly Performing Items



Item Selection Spreadsheet

Number of Responses	Critical Correlation	Time	P-value	Correlation	
568	0.054	23.0			
568	0.054	18.0			
568	0.054	25.5			
568	0.054	35.0			
568	0.054	20.0			
568	0.054	12.0			
568	0.054	20.0			
568	0.054	12.0	TE		
568	0.054	17.0			
568	0.054	31.0			
568	0.054	12.0			
568	0.054	30.0	TE		
568	0.054	10.0			
568	0.054	8.0			
568	0.054	31.0			
568	0.054	11.0			Keep
568	0.054	11.0	TE		Delete
568	0.054	7.0			Keep
568	0.054	17.0			Keep
568	0.054	36.0			Keep

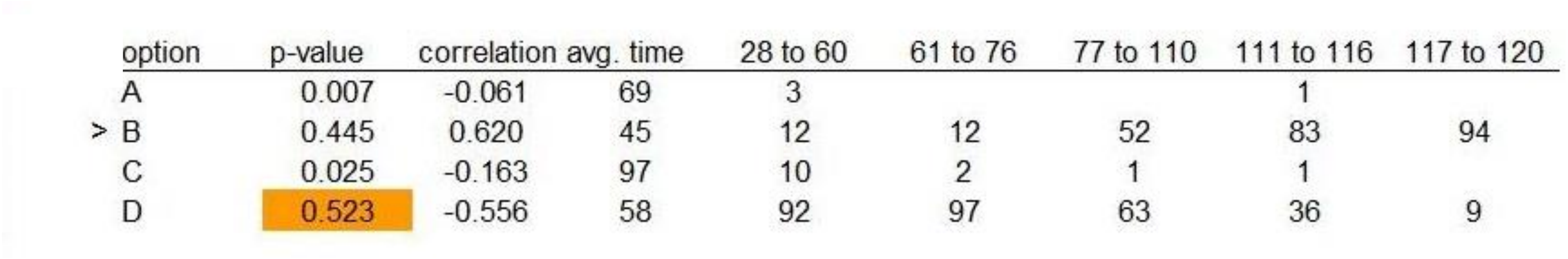
Item Flagging

- Items with issues based on their statistical performance
- Default parameters can be set depending on exam situation
- P-values
  - Items with p-values > 0.90 = "too easy"
  - Items with p-values < 0.10 = "too hard"
- Item score correlation
  - Items with correlation < critical correlation = "no"
  - Items with – correlation < critical correlation = "neg"
- Option analysis
  - Letter of incorrect response with higher correlation, p-value, or high scoring candidates than correct option

Alpine Testing Solutions

Item-Level Statistics: Option Analysis

- 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 or correlations higher than the key
  - Frequency count: Distractors with frequent selection by high performing examinees



option	p-value	correlation	avg. time	28 to 60	61 to 76	77 to 110	111 to 116	117 to 120
A	0.007	-0.061	69	3			1	
> B	0.445	0.620	45	12	12	52	83	94
C	0.025	-0.163	97	10	2	1	1	
D	0.523	-0.556	58	92	97	63	36	9

Form-Level Analysis

- Provides the overall test statistics by form

Health Check	Form A
Candidate Count	568
Exam Length	120
Mean	88.93
SD	27.65
Median	99.5
Mode	118
Avg. Time on Test	66.2
SD of Time on Test	27.6
Standard Error of the Mean	1.16
95% confidence interval +/-	2.27
Minimum	28
Maximum	120
Skewness	-0.38
Kurtosis	-1.36
Alpha Reliability	0.981
SEM	3.84
95% confidence interval +/-	7.53
# Items in Test Pool	120

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

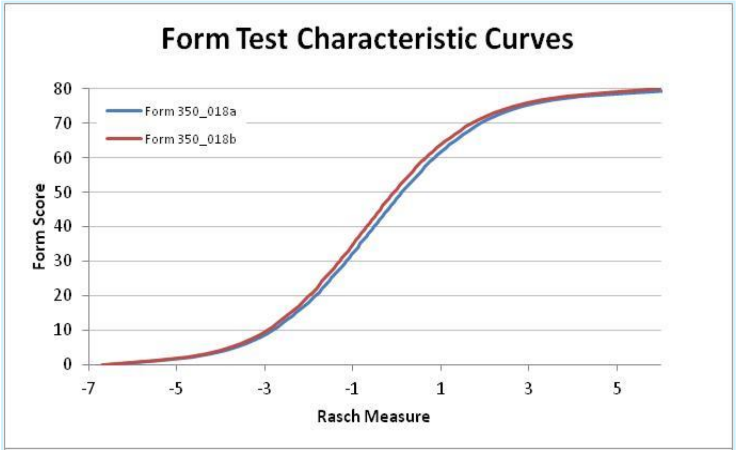
Reliability: Consistency of items as an entire exam, how well the items as a test seem to be measuring the same knowledge, should be > 0.85 for certification exams

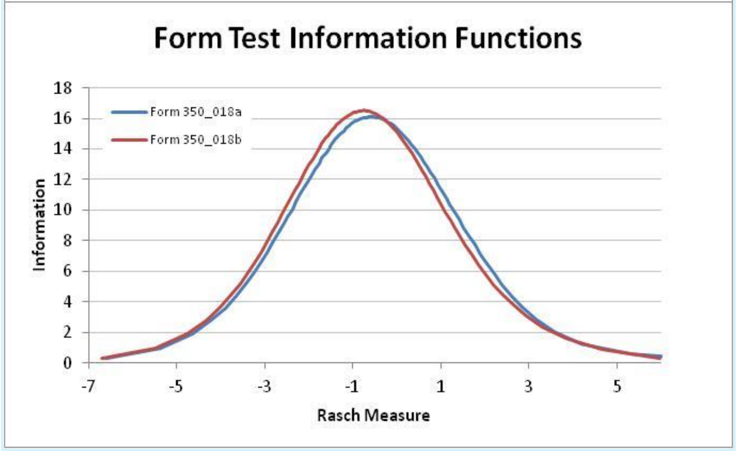
Alpine Testing Solutions

Form-Level Analysis

- Form-level analysis shows test-level statistics by form
  - Metrics such as average difficulty, time, and pass rates
  - Imbalanced statistics and differing item difficulties along the ability continuum indicate current misalignment

Health Check	Form A	Form B
Exam Length	80	80
Mean	72.00	72.29
SD	10.25	10.84
Rasch Measure at Cut Score	1.90	1.70
Avg. Time on Test	35.4	37.3
Standard Error of the Mean	0.46	0.44
95% confidence interval +/-	0.90	0.85
Minimum	11	14
Maximum	80	80
Alpha Reliability	0.953	0.959
SEM	2.23	2.19
95% confidence interval +/-	4.38	4.29
Pass Rate	80.2%	81.4%


Form Test Characteristic Curves

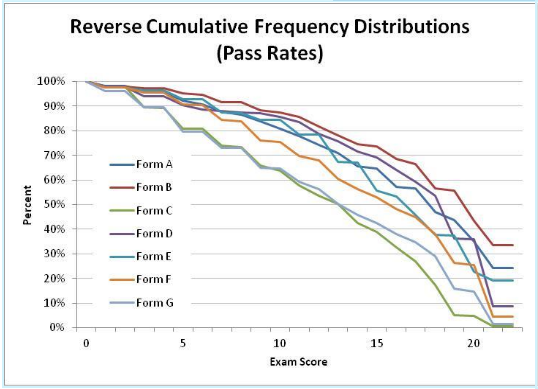

Form Test Information Functions

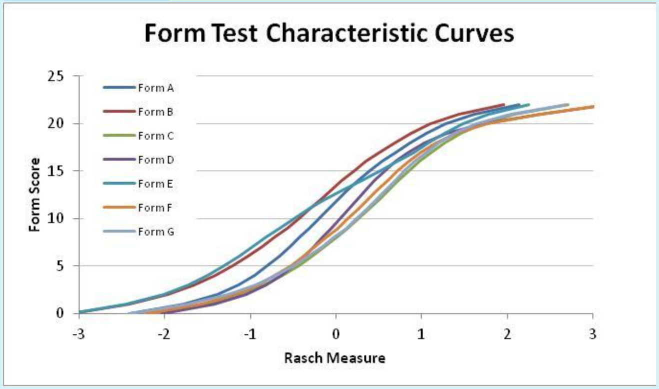
Alpine Testing Solutions

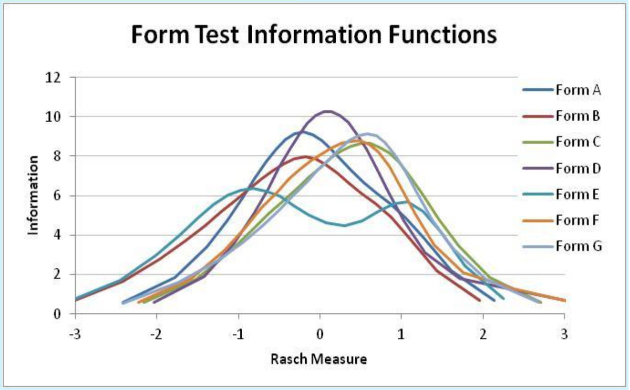
Form-Level Analysis

- Differences in difficulty indicative of non-equivalent exam forms
  - Need to have overlapping content to enable either pre- or post-equating of the forms to ensure fair scoring

Health Check	Form A	Form B	Form C	Form D	Form E	Form F	Form G
Candidate Count	321	223	275	231	249	495	313
Exam Length	10	10	10	10	10	10	10
Mean	15.51	16.90	11.44	15.56	15.00	13.87	11.96
SD	6.16	5.73	5.76	5.80	5.79	5.89	6.36
Rasch Measure at Cut Score	0.69	0.51	1.11	0.84	0.91	0.99	1.05
Standard Error of the Mean	0.34	0.38	0.35	0.38	0.37	0.26	0.36
95% confidence interval +/-	0.67	0.75	0.68	0.75	0.72	0.52	0.70
Alpha Reliability	0.831	0.828	0.757	0.813	0.814	0.788	0.811
SEM	2.53	2.38	2.84	2.50	2.50	2.71	2.76
95% confidence interval +/-	4.96	4.66	5.57	4.91	4.90	5.31	5.41
Pass Rate	56.7%	66.4%	26.9%	59.3%	45.8%	45.1%	34.8%

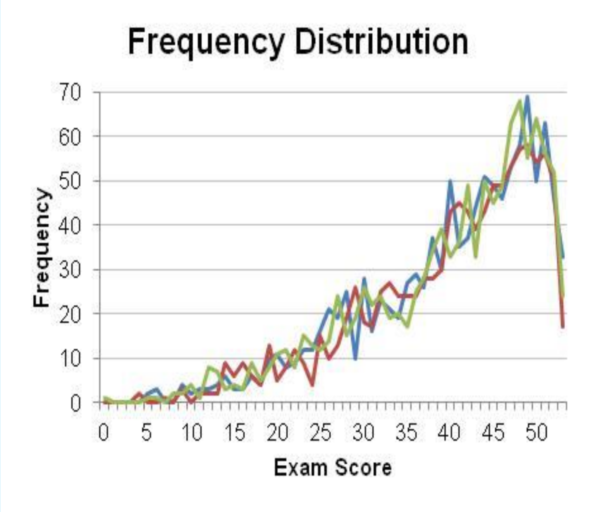

Reverse Cumulative Frequency Distributions (Pass Rates)

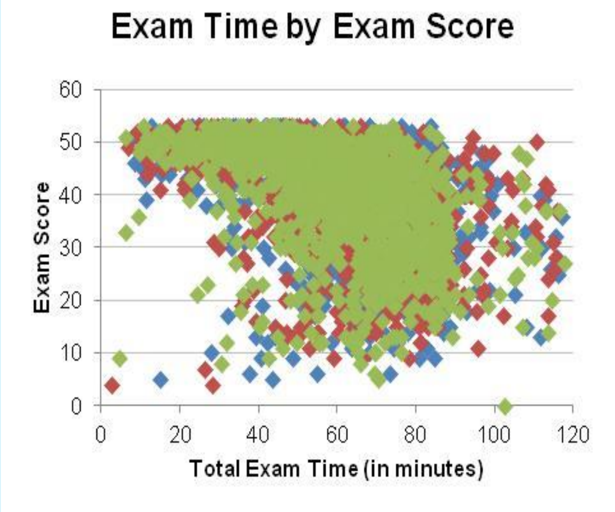

Form Test Characteristic Curves

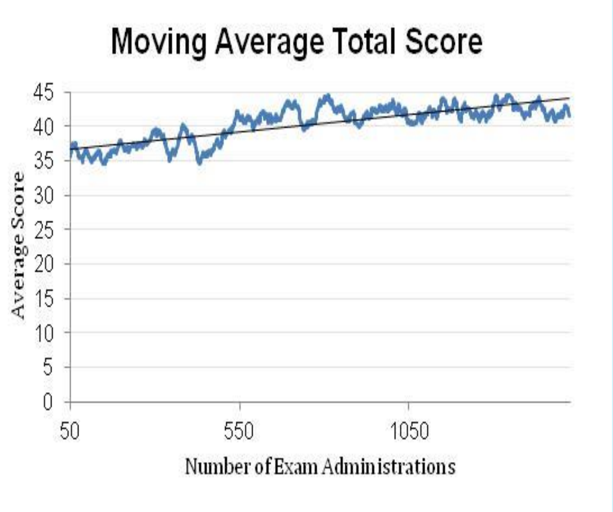

Form Test Information Functions

Form-Level Analysis

- Can provide evidence of exam security or potential compromise



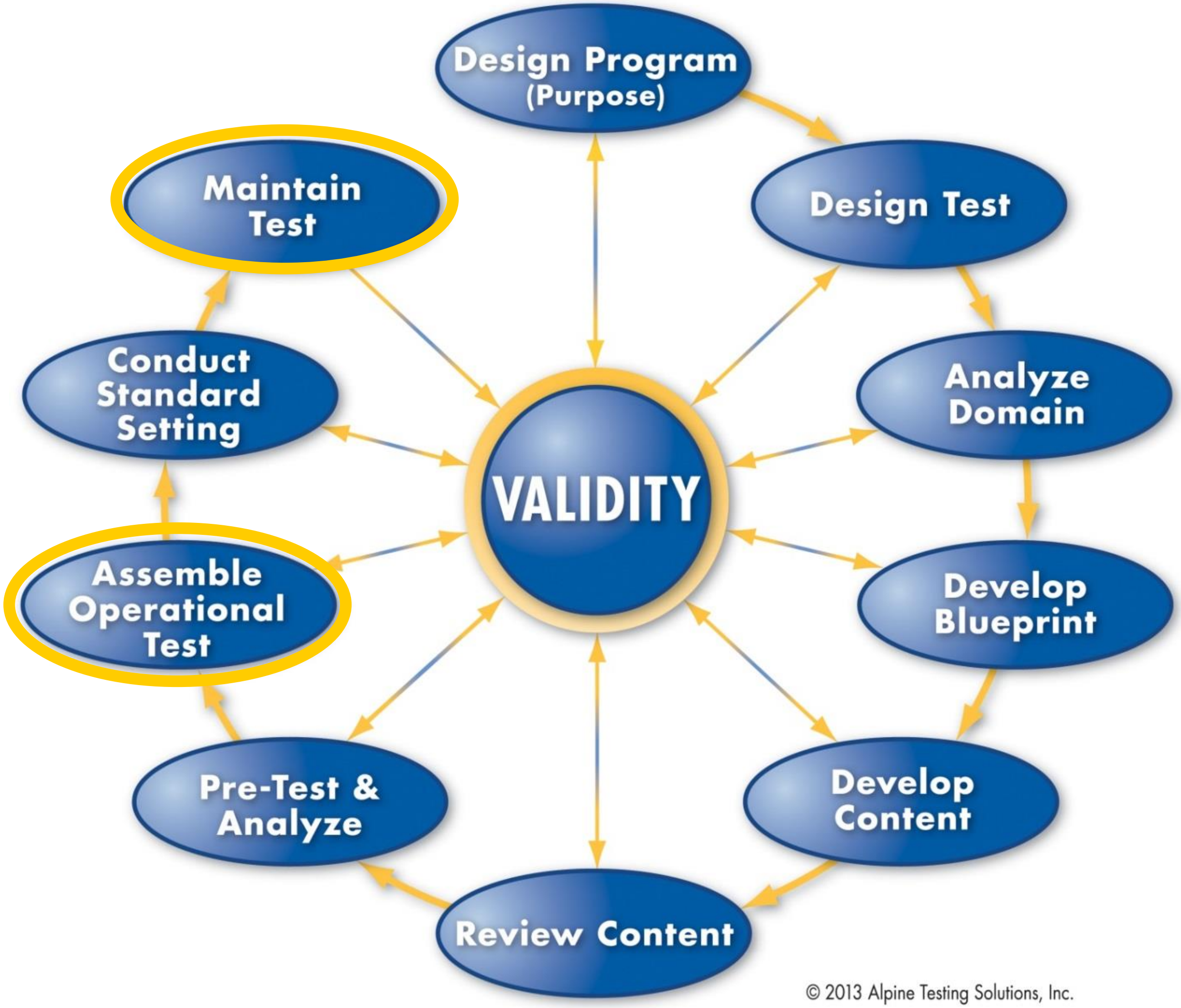




Publish Parallel and Balanced Forms Through

FORMS ASSEMBLY & EQUATING

Stage in Process

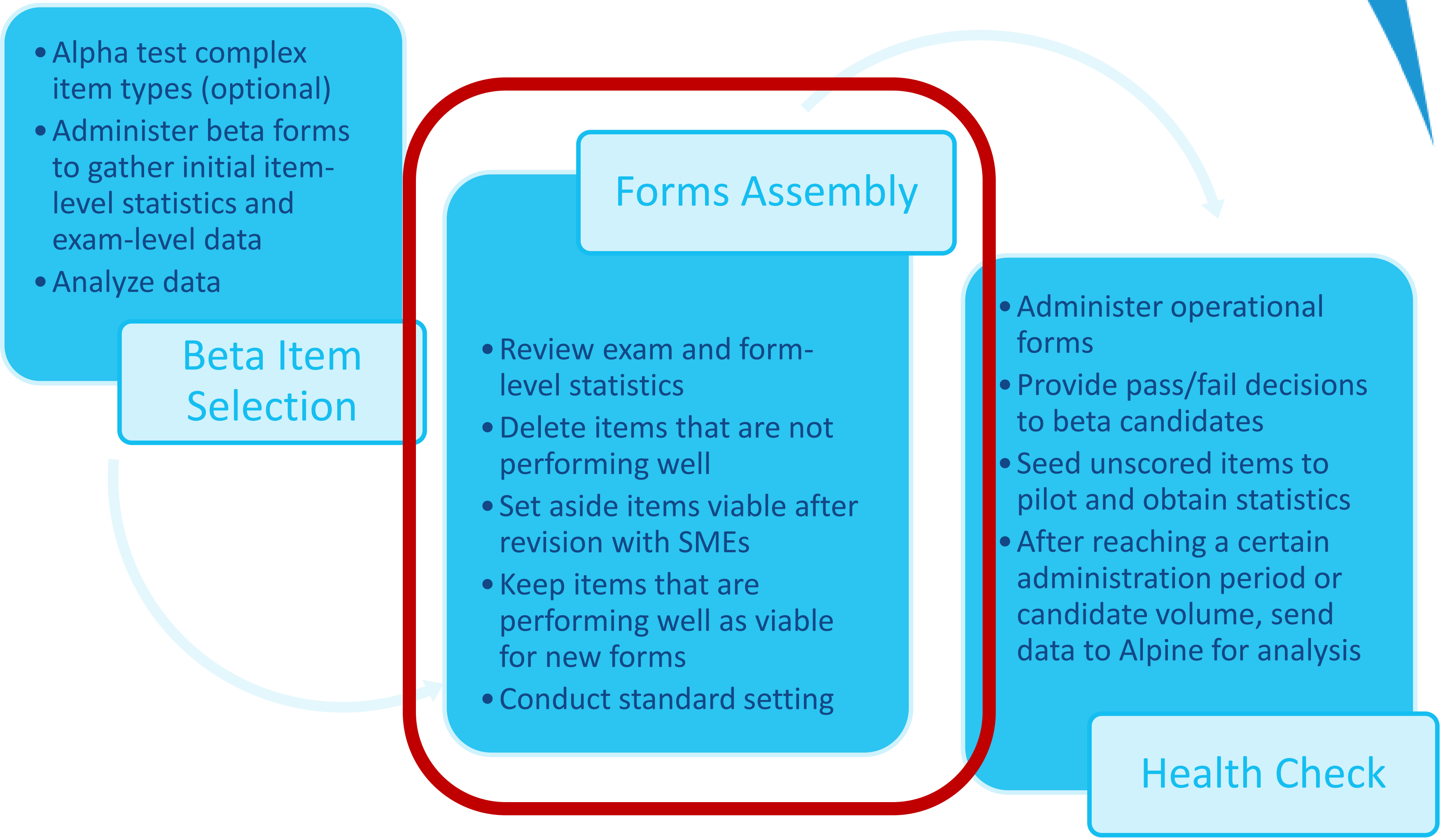


Alpine Testing Solutions

Purpose

- Determine specifications for live exam including number of test forms, items and/or points per form, and administration time
- Assemble one or more parallel operational forms
  - Parallel test forms should have equivalent statistical characteristics and proper blueprint representation
- **Provide fair, equated scores resulting in similar score interpretation for all candidates *regardless of test form taken***

Exam Release Cycle



Beta Item Selection

- Alpha test complex item types (optional)
- Administer beta forms to gather initial item-level statistics and exam-level data
- Analyze data

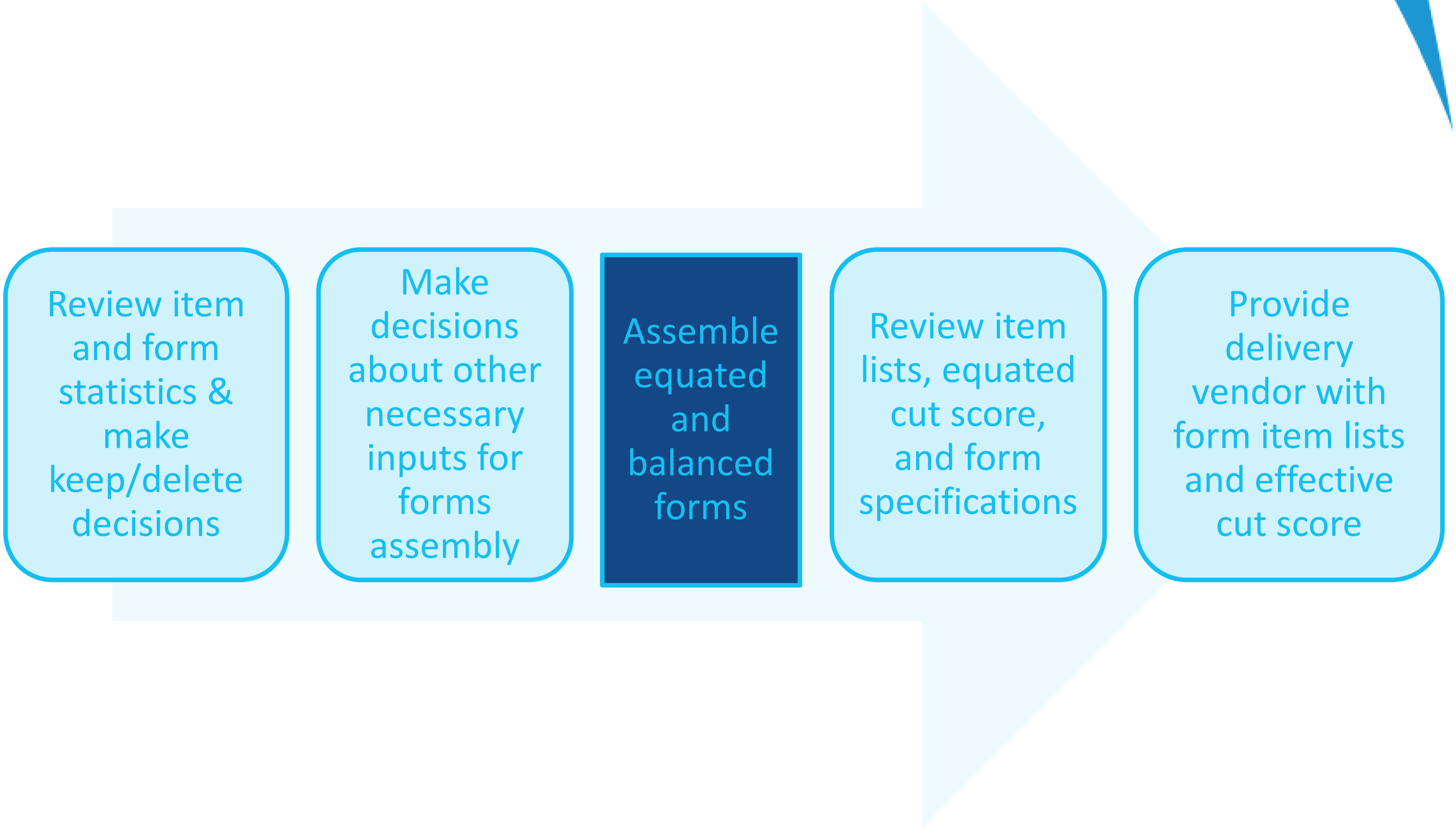
Forms Assembly

- Review exam and form-level statistics
- Delete items that are not performing well
- Set aside items viable after revision with SMEs
- Keep items that are performing well as viable for new forms
- Conduct standard setting

Health Check

- Administer operational forms
- Provide pass/fail decisions to beta candidates
- Seed unscored items to pilot and obtain statistics
- After reaching a certain administration period or candidate volume, send data to Alpine for analysis

Form Assembly Process



Review item and form statistics & make keep/delete decisions

Make decisions about other necessary inputs for forms assembly

Assemble equated and balanced forms

Review item lists, equated cut score, and form specifications

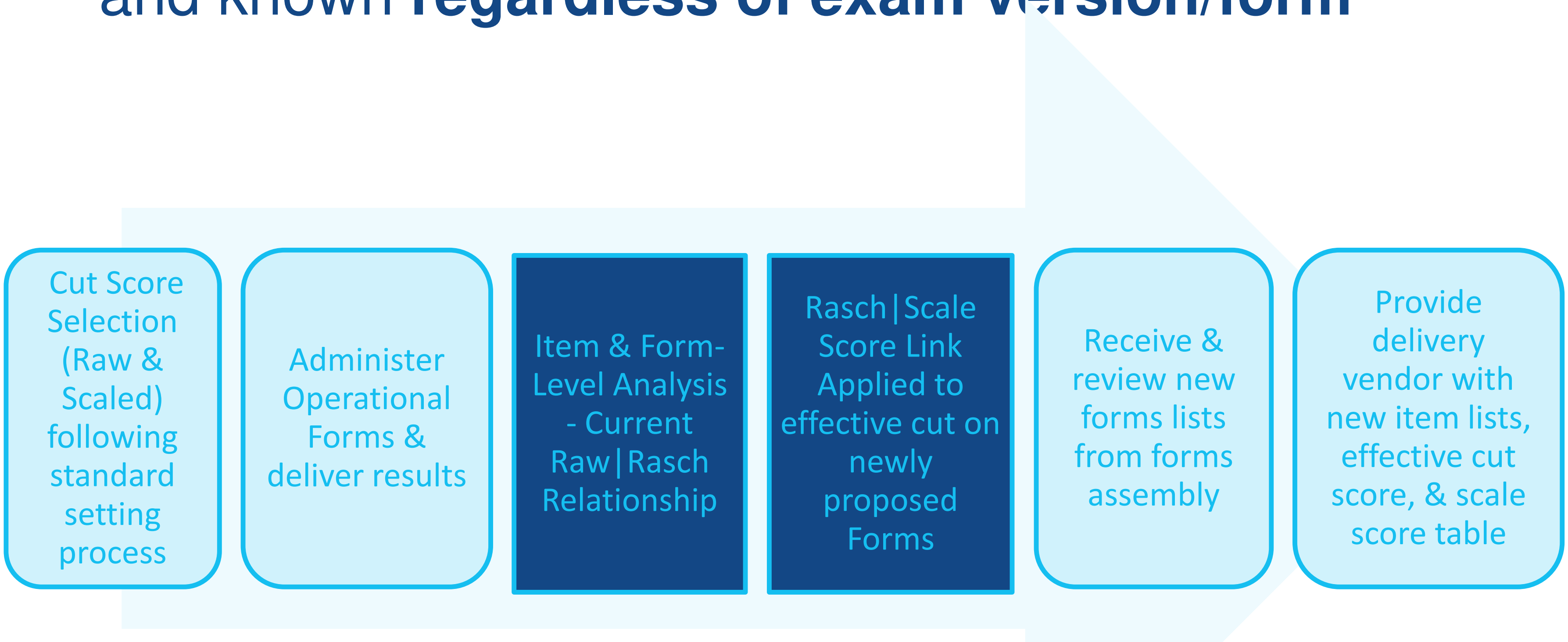
Provide delivery vendor with form item lists and effective cut score

Forms Assembly Considerations

- **Equate** to the raw 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** to the scaled cut score to 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

Equating & Scaling

- Expectations for the ability-level needed to achieve a particular performance level remains consistent, fair, and known **regardless of exam version/form**



Cut Score Selection (Raw & Scaled) following standard setting process

Administer Operational Forms & deliver results

Item & Form-Level Analysis - Current Raw|Rasch Relationship

Rasch|Scale Score Link Applied to effective cut on newly proposed Forms

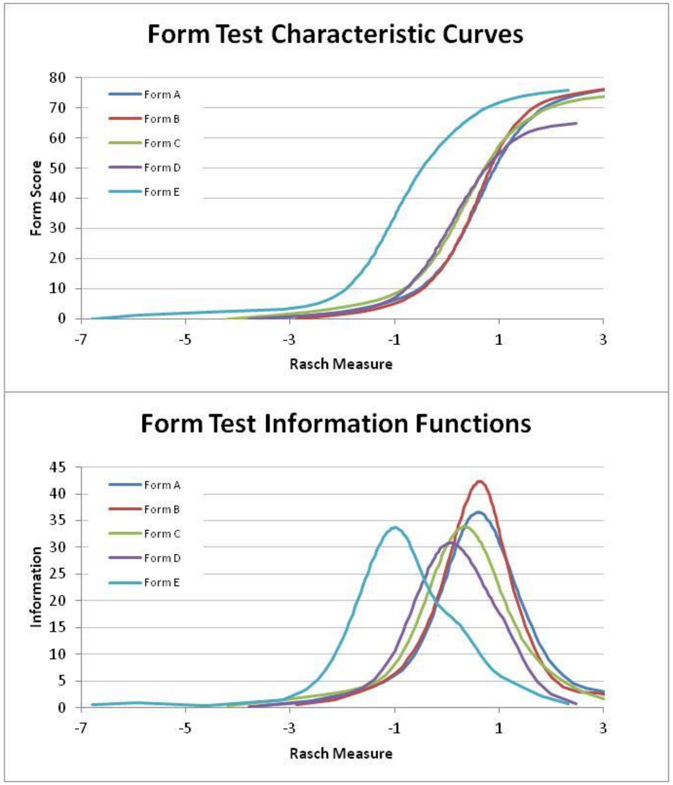
Receive & review new forms lists from forms assembly

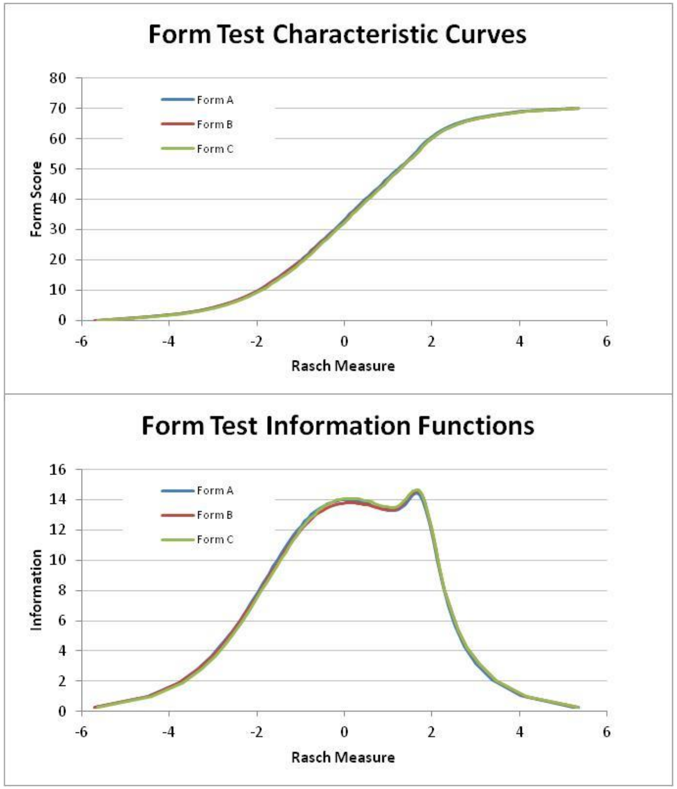
Provide delivery vendor with new item lists, effective cut score, & scale score table

Alpine Testing Solutions

Equating

- Differences in difficulty across forms can be accommodated for through either pre- or post-equating, resulting in aligned scoring decisions





Scaling

- Eases the interpretability of exam scores and pass/fail decisions
- Important to the valid interpretation of exam scores as it assigns meaningful links between raw scores, underlying ability measures, and scale scores
- Ensures consistency of the scale score meaning regardless of which administration/version/form of an exam a candidate receives

References

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (1999). *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.

Alpine
Testing Solutions

QUESTIONS?

Alpine
Testing Solutions