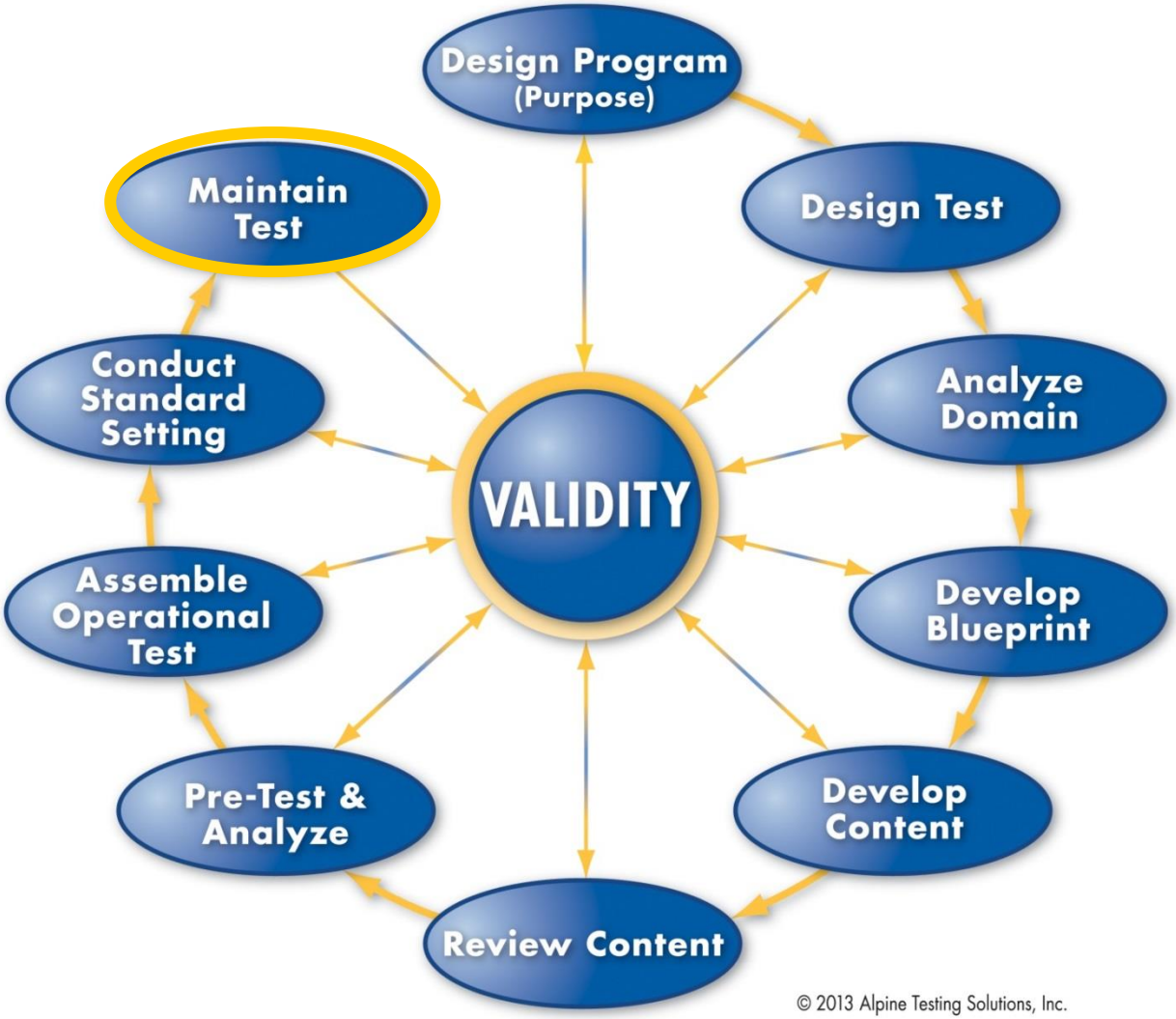
Alpine Testing Solutions

Monitoring Exam Data With Your Credential Management System

Brett P. Foley, Ph.D.

November 5, 2014

Stage in Process



Alpine
Testing Solutions

Why is data monitoring important?

- Critical indicator of the "health" of a program
- Early detection and mitigation of testing issues
- Professional responsibility: validity and fairness

Agenda

- Issues in monitoring exam data
  - Interpreting item and form statistics
  - Monitoring item and form performance over time
  - Understanding security metrics
  - Recognizing the effect of sample size on test maintenance activities
  - Automating regular item and form analyses

Alpine Testing Solutions

Delve Deeper

- Ask A Psychometrician: Psychometric Analyses & Operations
  - Lisa O'Leary, Ph.D., *Psychometrician*

- Enhancing your Program Security with CertMetrics
  - Jill Burroughs, *Director and Senior Security Consultant*

- Visit www.alpinetesting.com to view webinars

Alpine Testing Solutions

Item & Form Statistics

Item- & Form-Level Analyses

- Evaluate statistical data regarding form- and item-level performance during operational administration
- Continually provide evidence of the following:
  - 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

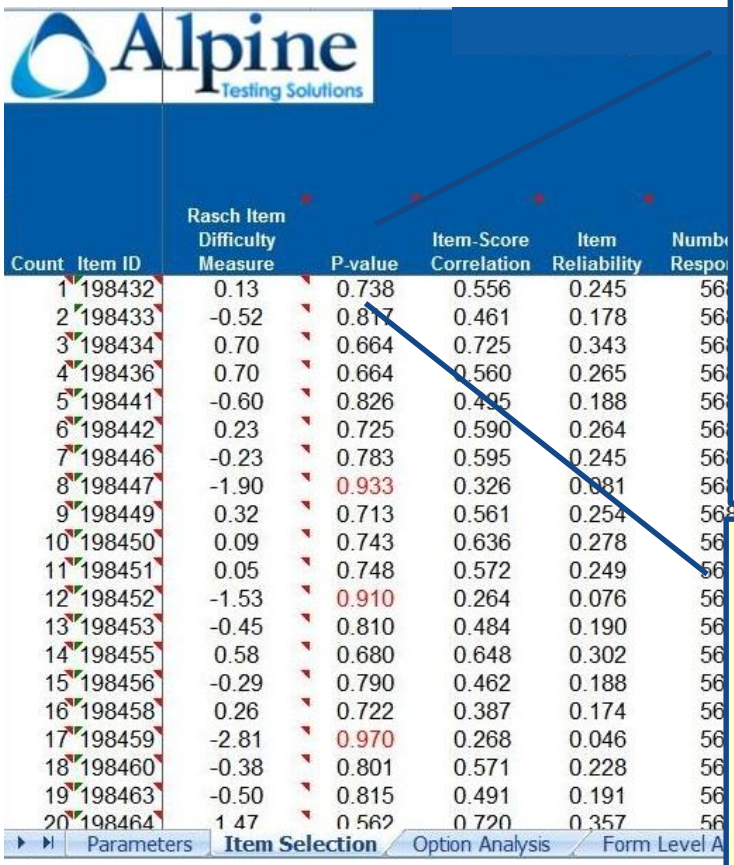
Alpine Testing Solutions

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

Alpine Testing Solutions

Item-Level Statistics: Item Difficulty



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	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	56
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

Parameters | Item Selection | Option Analysis | Form Level A

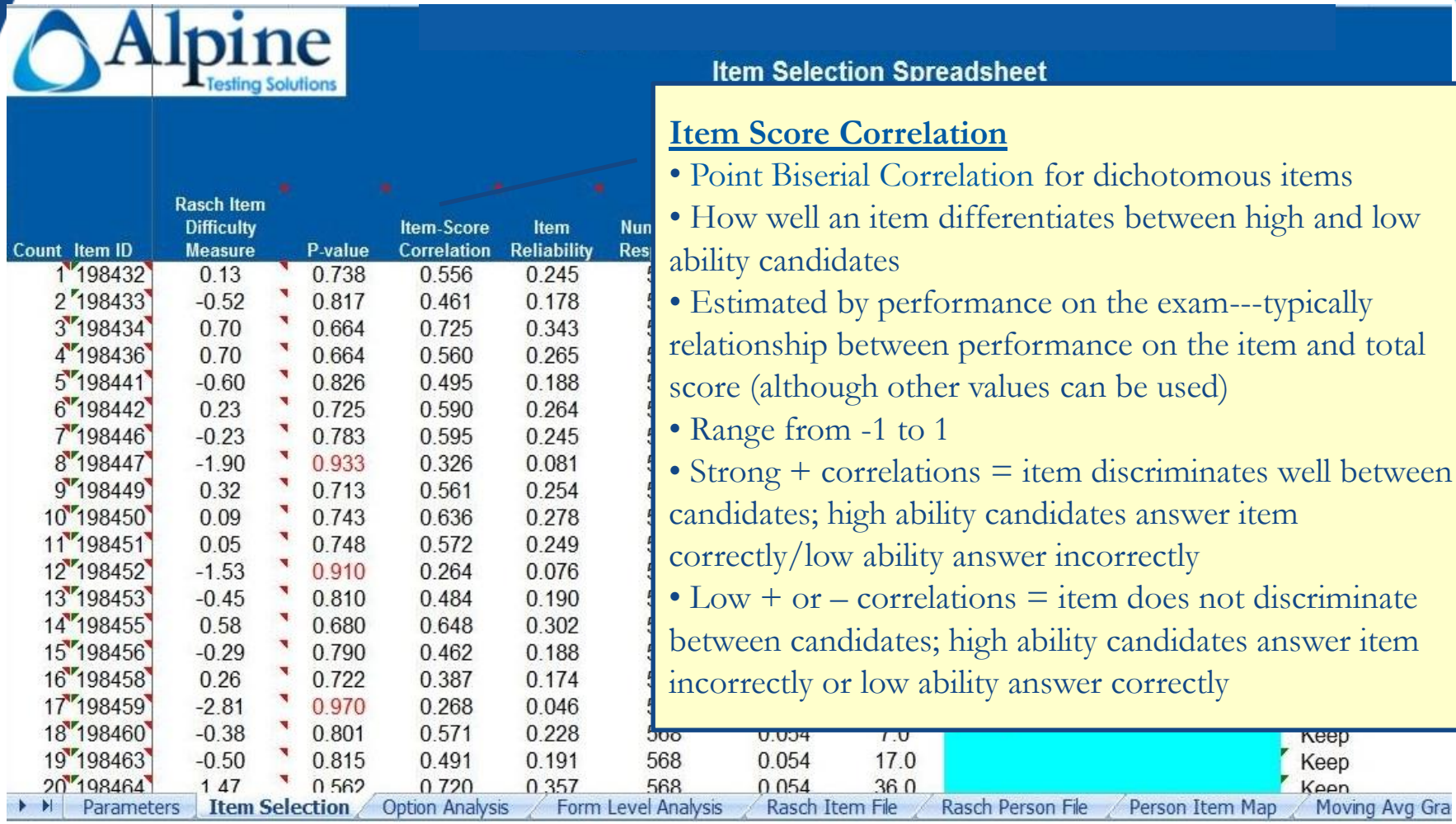
P-Value

- 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: Correlation

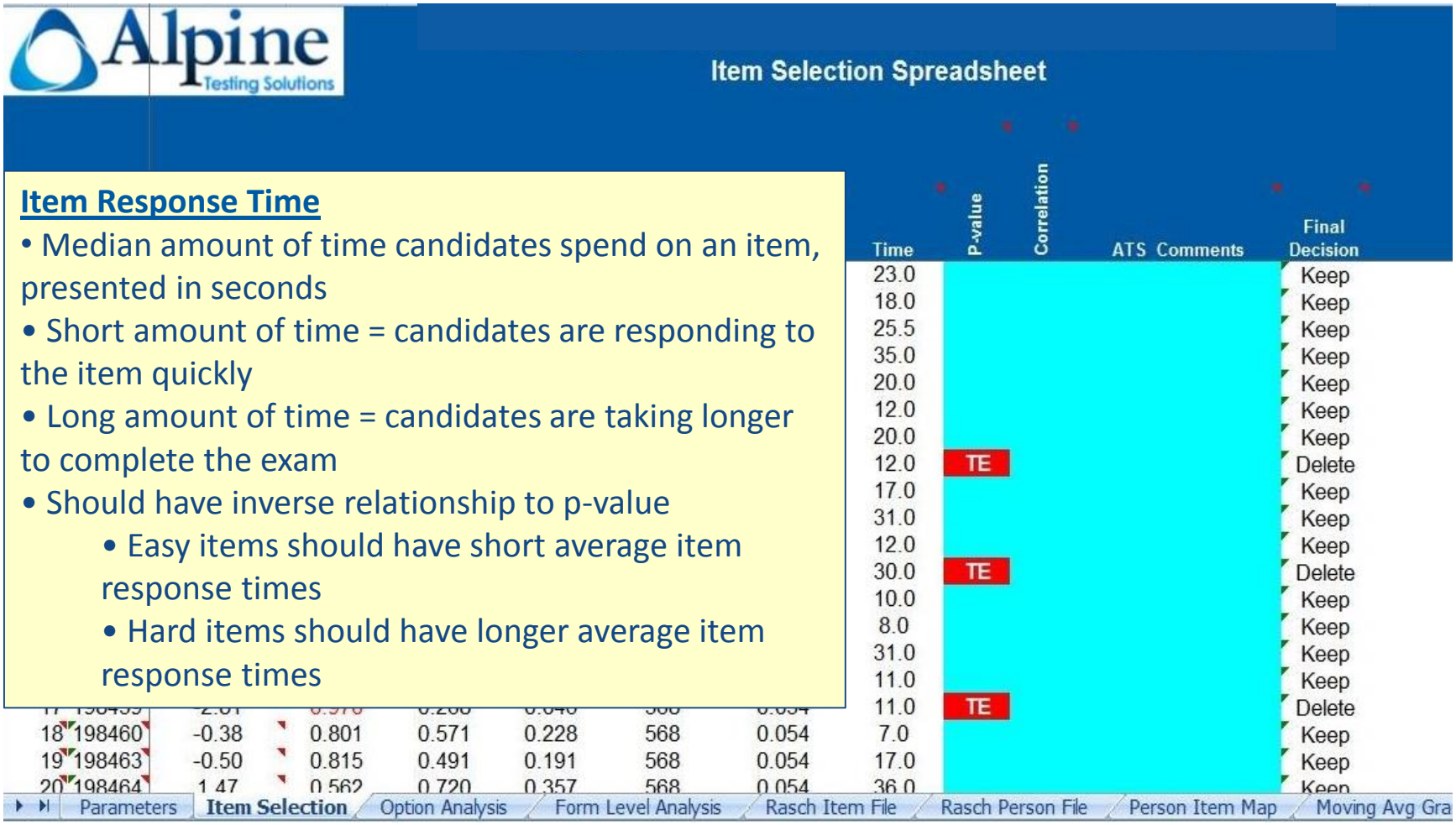


Item Score Correlation

- 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

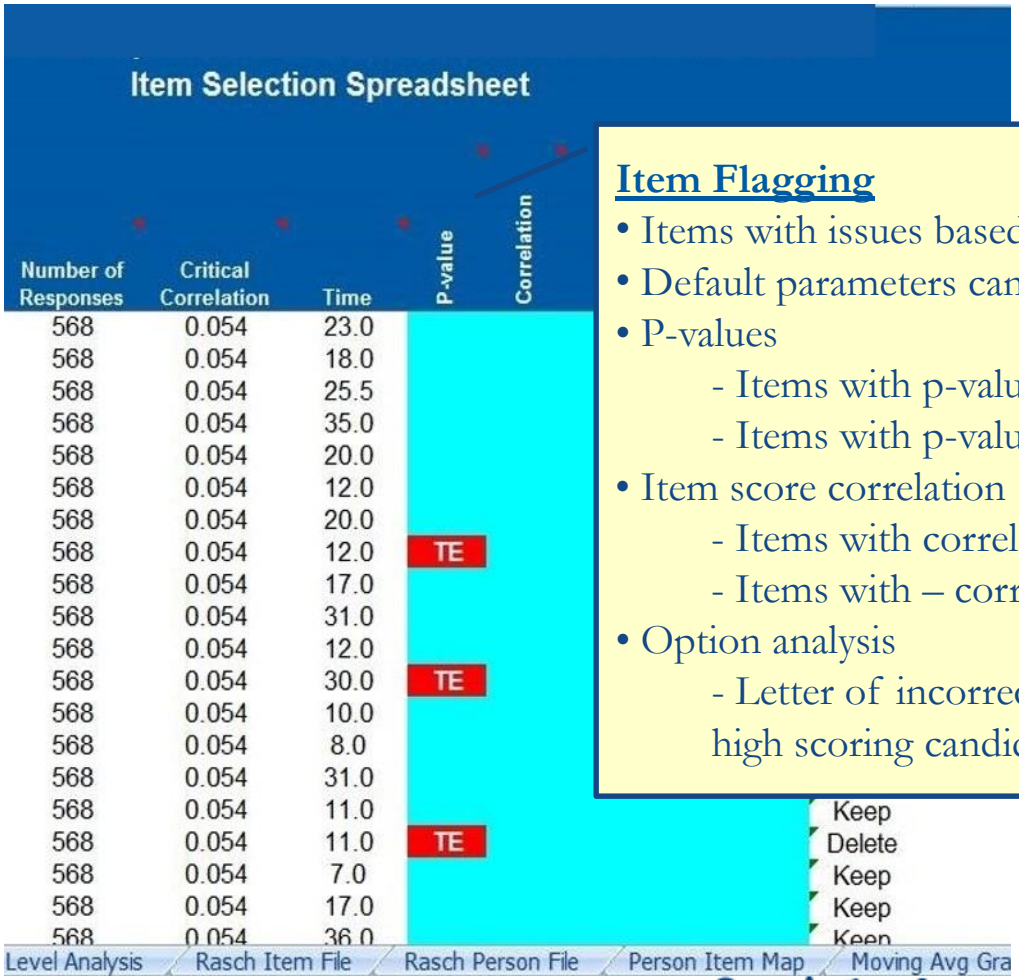
Alpine Testing Solutions

Item-Level Statistics: Response Time



Alpine Testing Solutions

Item-Level Statistics: Identification of Poorly Performing Items



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

Alpine Testing Solutions

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

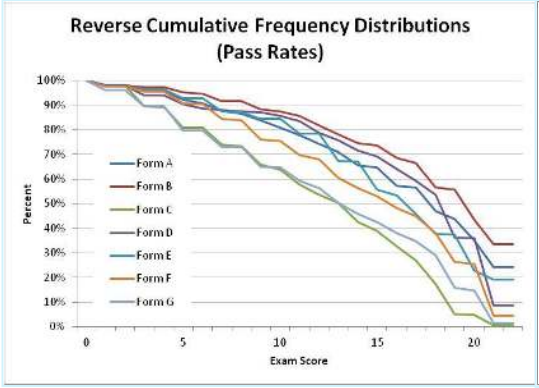
Reliability: Consistency of items as an entire exam, how well the items as a test seem to be measuring the same knowledge, should be > 0.85 for certification exams

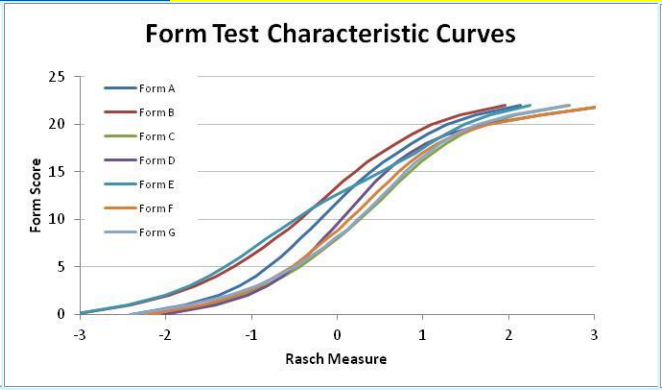
Alpine Testing Solutions

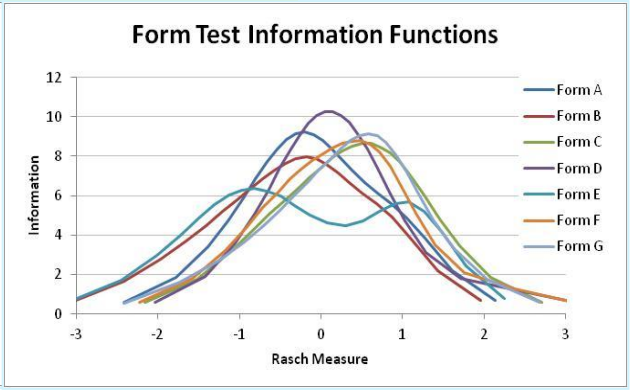
Form-Level Analysis

- 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%





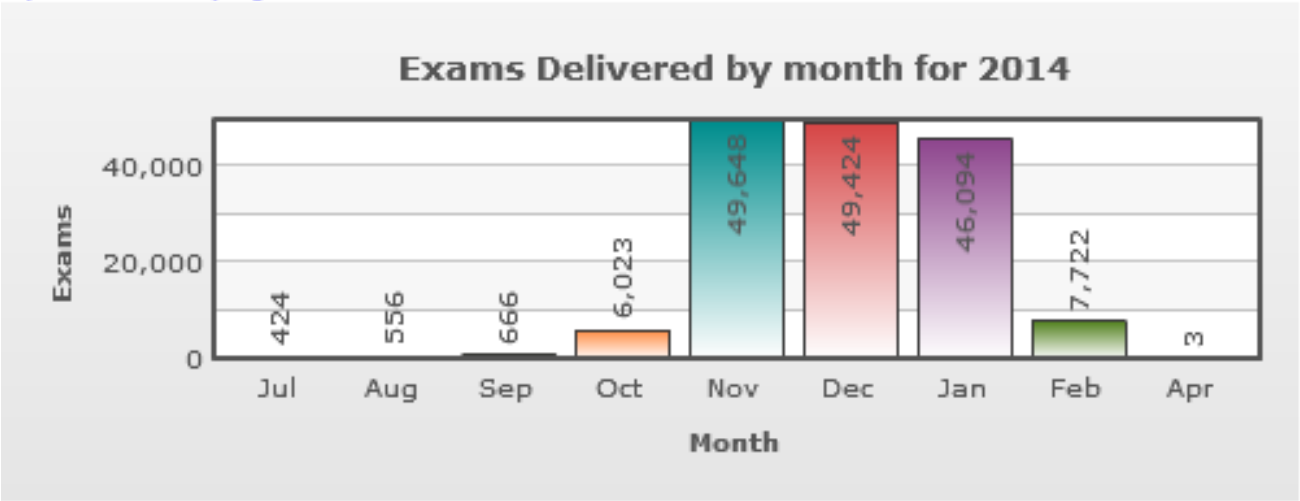


Alpine Testing Solutions

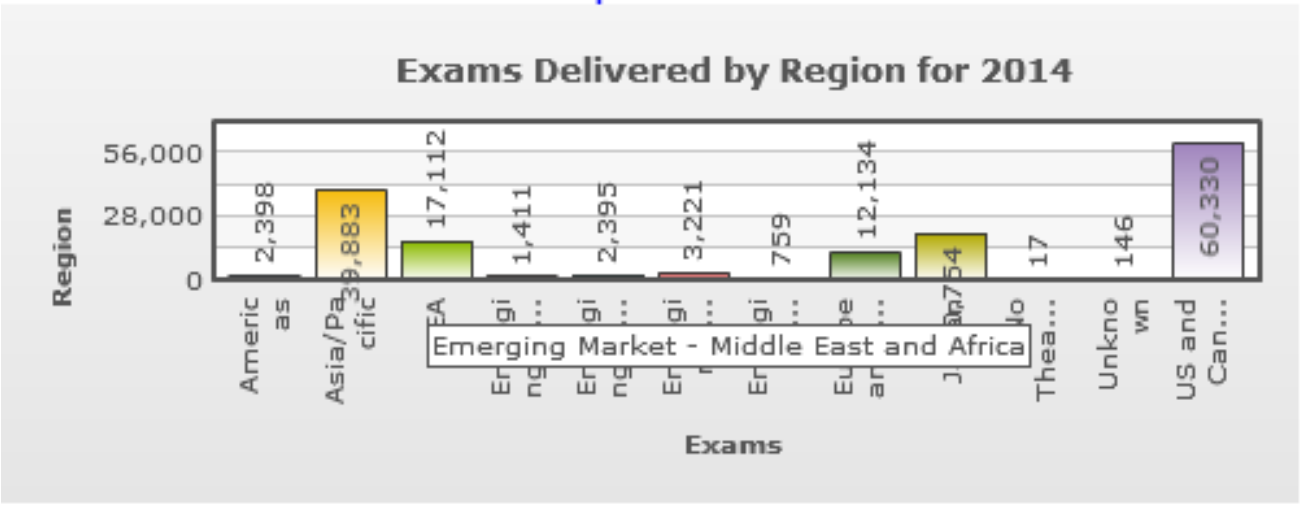
STATISTICS OVER TIME

Exam Volume

By month By Quarter



Export Data

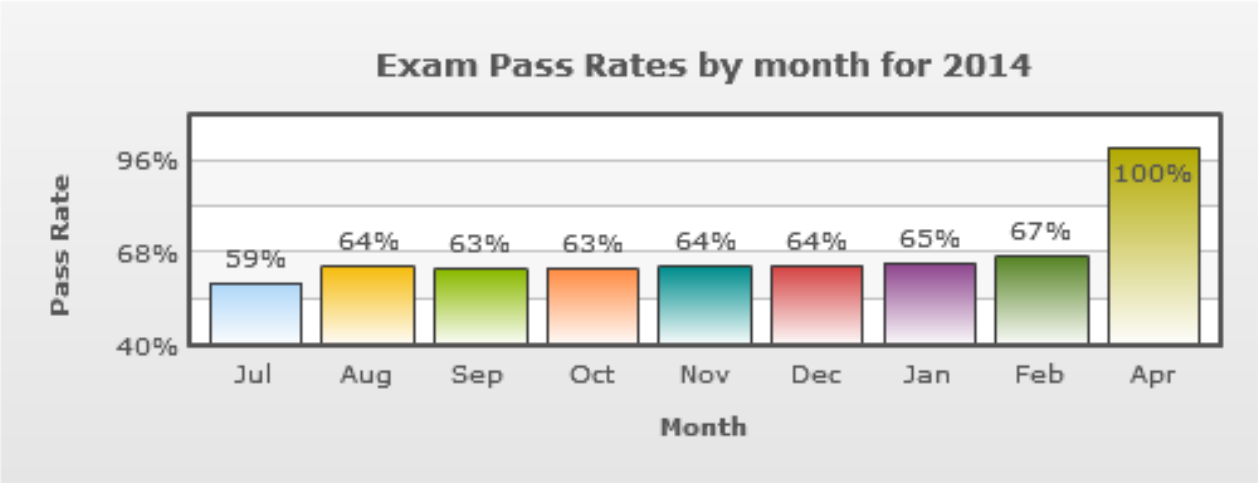


Note: Region totals are based on where the candidate lives, not where the Exam was taken.

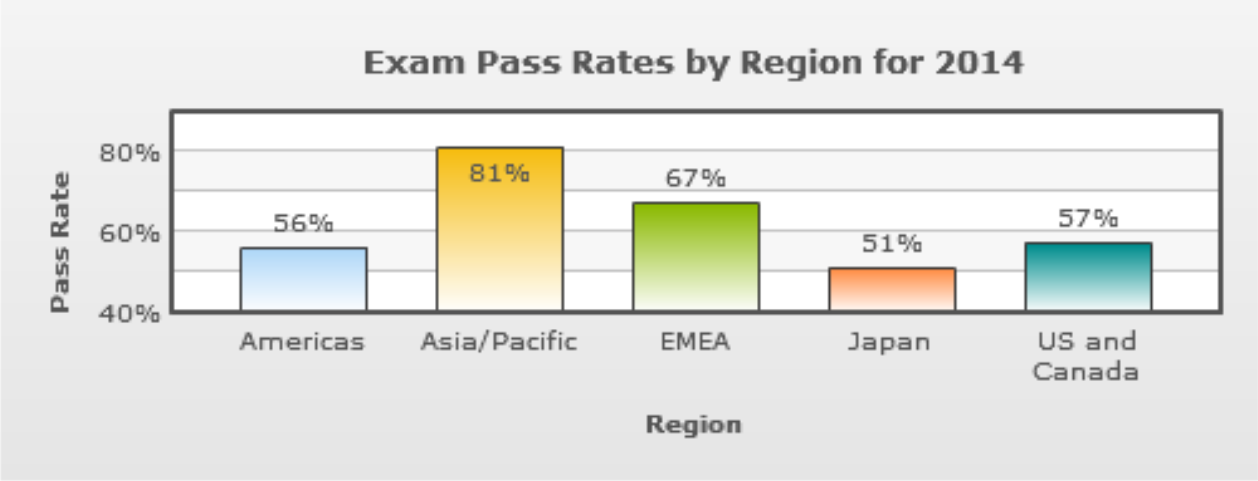
Export Data

Pass Rate

By month By Quarter



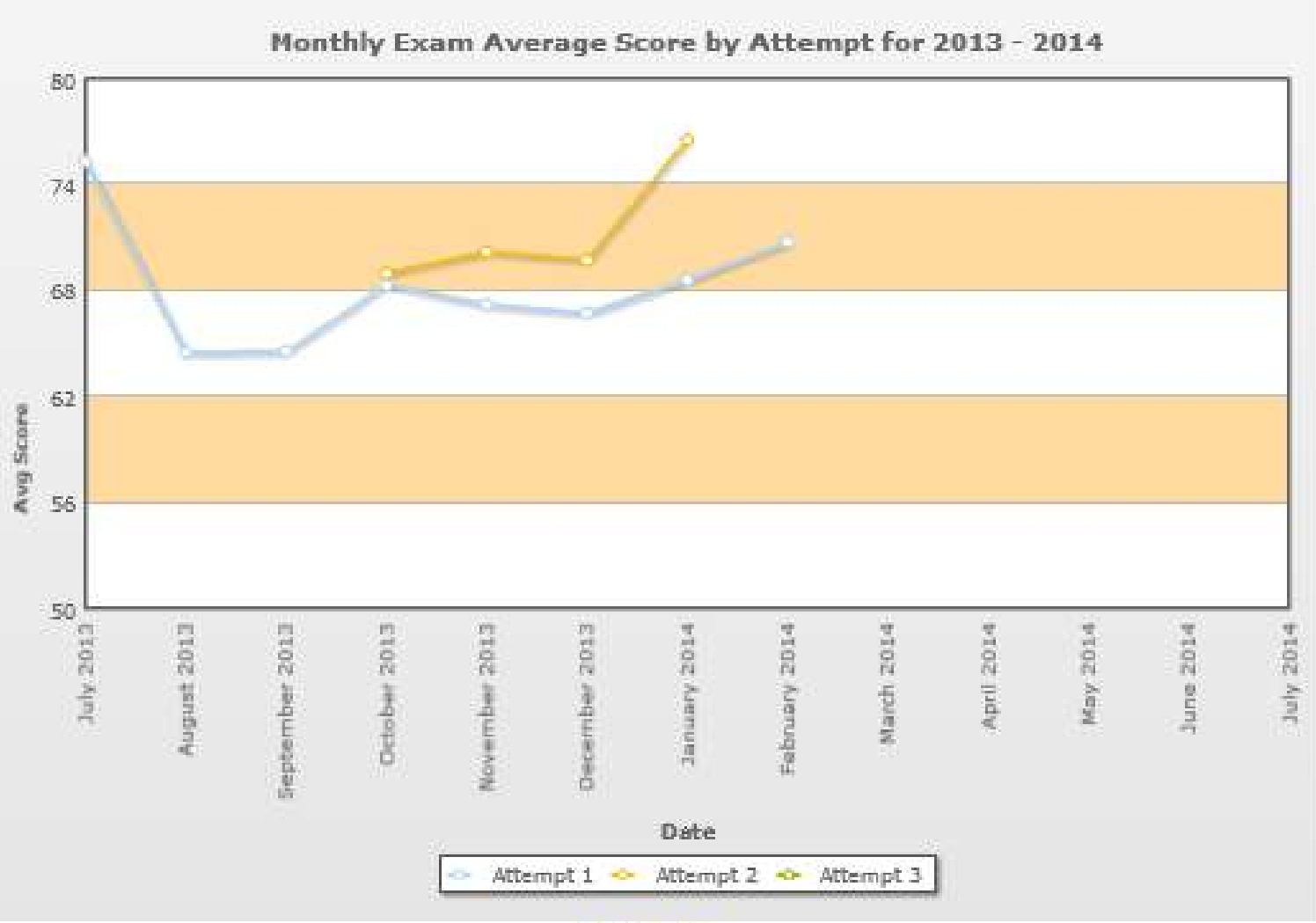
Export Data



Note: Region totals are based on where the candidate lives, not where the Exam was taken.

Export Data

Results by attempt



Export Data

Additional checks over time

- Item performance
- Median time to complete test
- Frequency distributions

Disaggregate by…

- Test form
- Test center
- Language
- Country
- First time/repeaters
- Other demographics

Security Metrics

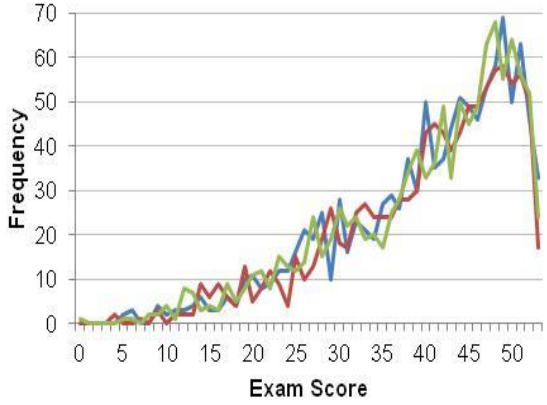
CertMetrics Security Scripts

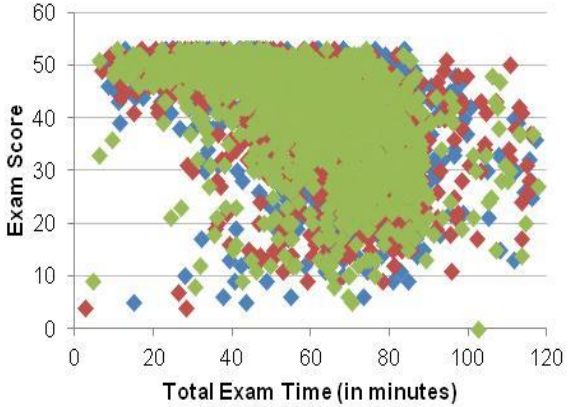
Security Flag	Suspect Action
Exam Retakes	Candidate takes same exam x times within y period
Rapid Exam Completion	x% percent of items completed in less than y seconds
Retake After Pass	Candidate takes the same exam after already passing
Large Score Differential	Score increase by more than x% within 2 attempts within y days
High score/Low time	Exam score above x%, time spent on exam less than y min
Too Little Exam Time	Candidate spent less than x minutes taking an exam
Possible Collusion	Candidates at the same test center on same date and scored within y% of each other on same exam
Security Items	Candidate correctly answered x security items out of y total security items
Differential Item Performance	x% or above on 1st item type and y% or under on 2nd item type
Watch List	Candidate is on the watch list at the time of taking a test
Banned list	Candidate is on the ban list at the time of taking a test

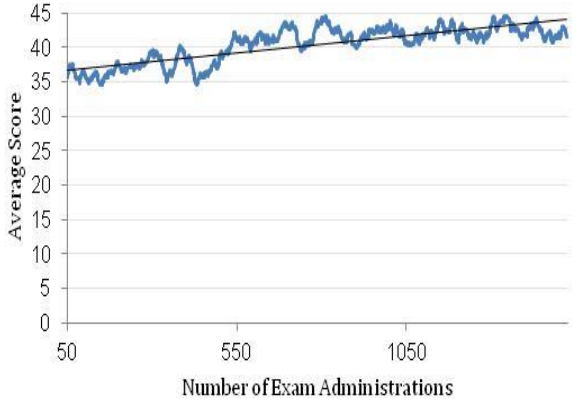
Alpine

Form-Level Analysis

- Can provide evidence of exam security or potential compromise







Alpine Testing Solutions

THE EFFECT OF SAMPLE SIZE

How much information is available?

- When monitoring, amount of data will depend on
  - Volume of program
  - Number of forms
  - Frequency of monitoring
  - Level of disaggregation
  - Size of subgroups

Smaller Sample = Less Certainty

- Use caution when interpreting data based on small samples
  - Don't panic when statistics change
  - Small differences expected
  - Large differences warrant investigation
  - Watch for long term trends

Quick Rules of Thumb

- For a proportion (e.g., pass rate, p-value, % first timers)…

$$\text{Margin of Error} \approx \frac{1}{\sqrt{sample\ size}}$$

- To halve margin of error, quadruple sample size

Choose a monitoring interval

- By time
  - Weekly
  - Once a month
  - Every 6 months
- By number of administrations
  - Every 200 administrations

- Appropriate choice will vary by program and volume
- Sample size should be large enough that statistics are meaningful/useful

AUTOMATING REGULAR ITEM AND FORM ANALYSES

Why Automate

- Data should me monitored consistently and regularly
- Unusual results easier to identify with increased data familiarity
- Reduces administrative burden
- Regular reports serve as historical record, may help with accreditation

QUESTIONS?

BRETT.FOLEY@ALPINETESTING.COM
SHANE.FREEMAN@ALPINETESTING.COM

Alpine
Testing Solutions