Using a Performance Test Development & Validation Framework

James B. Olsen

Russell W. Smith

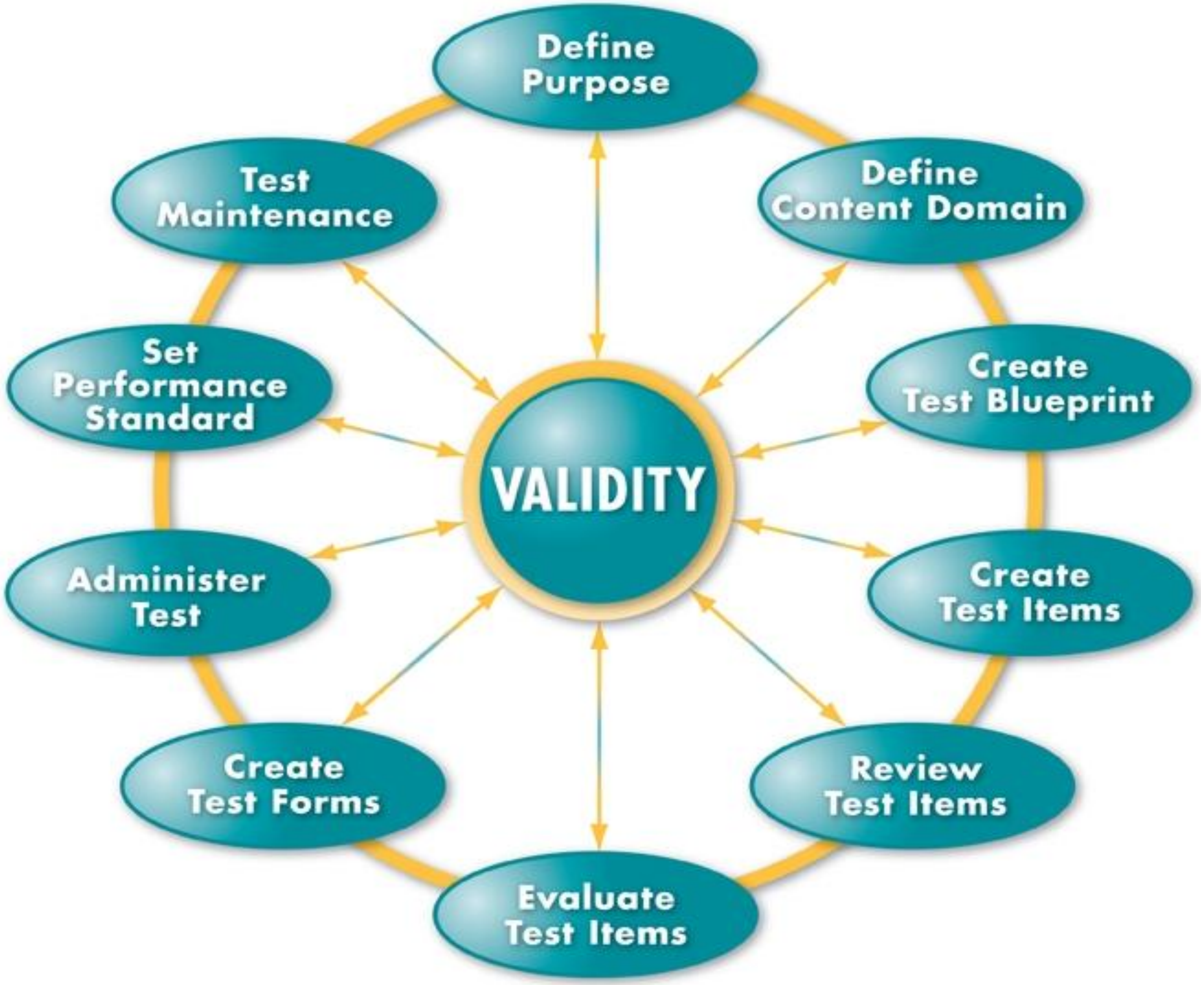
Cristina Goodwin

Alpine Testing Solutions

Presentation Overview

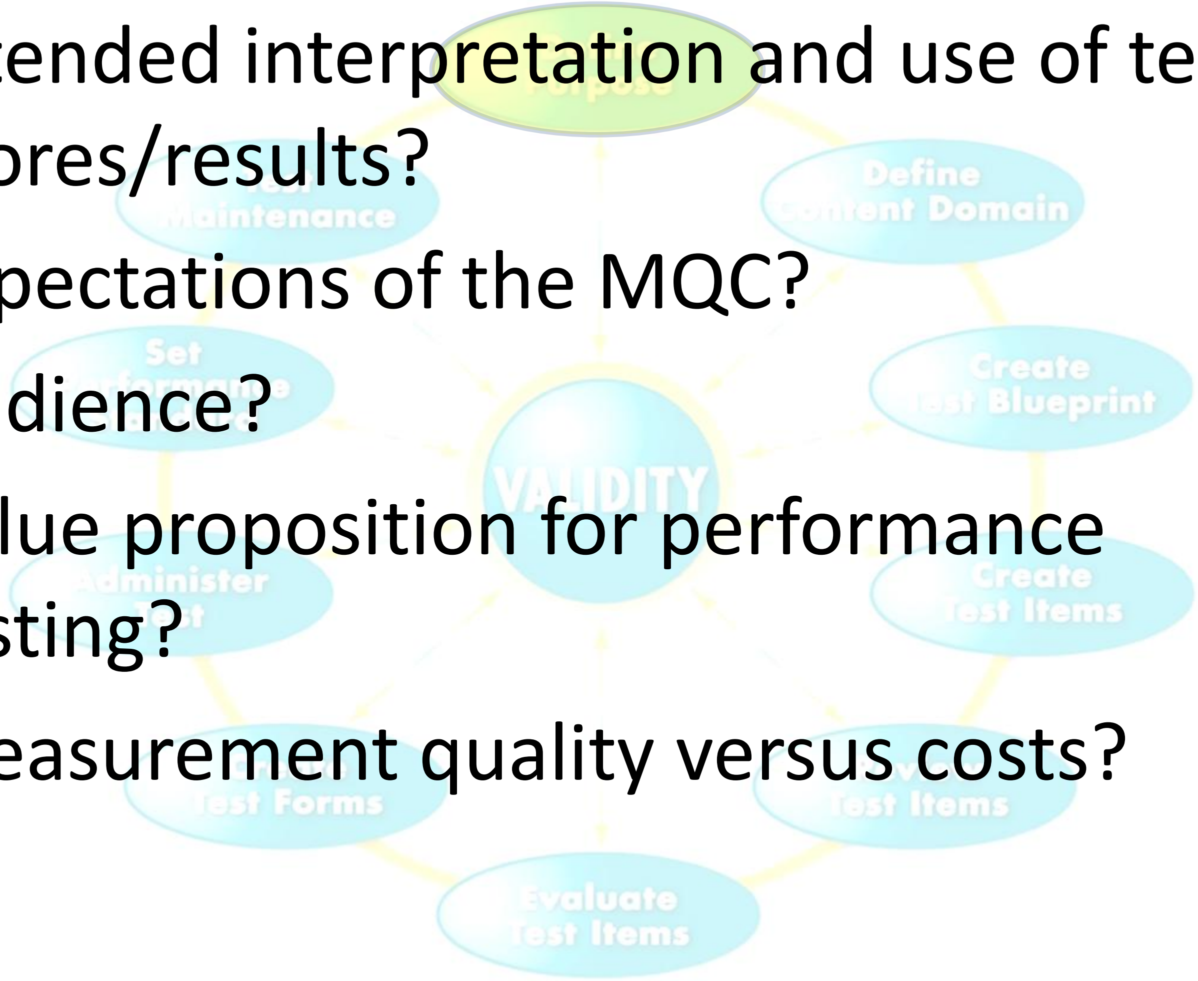
- Present a useful performance test development and validation framework
- Describe, and provide examples of, psychometrics for performance item types

Performance Test Development & Validation Framework



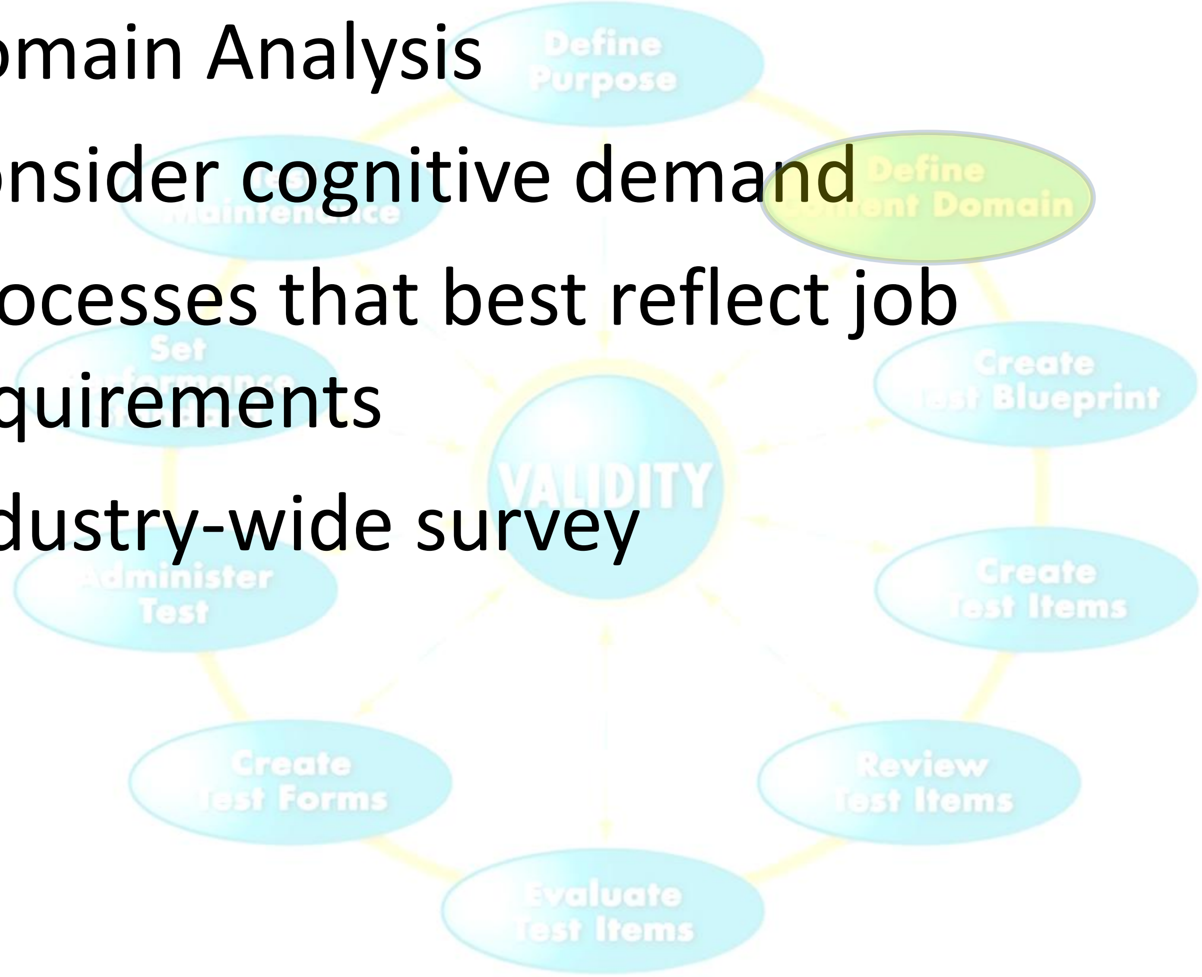
Define Purpose

- Intended interpretation and use of test scores/results?
- Expectations of the MQC?
- Audience?
- Value proposition for performance testing?
- Measurement quality versus costs?



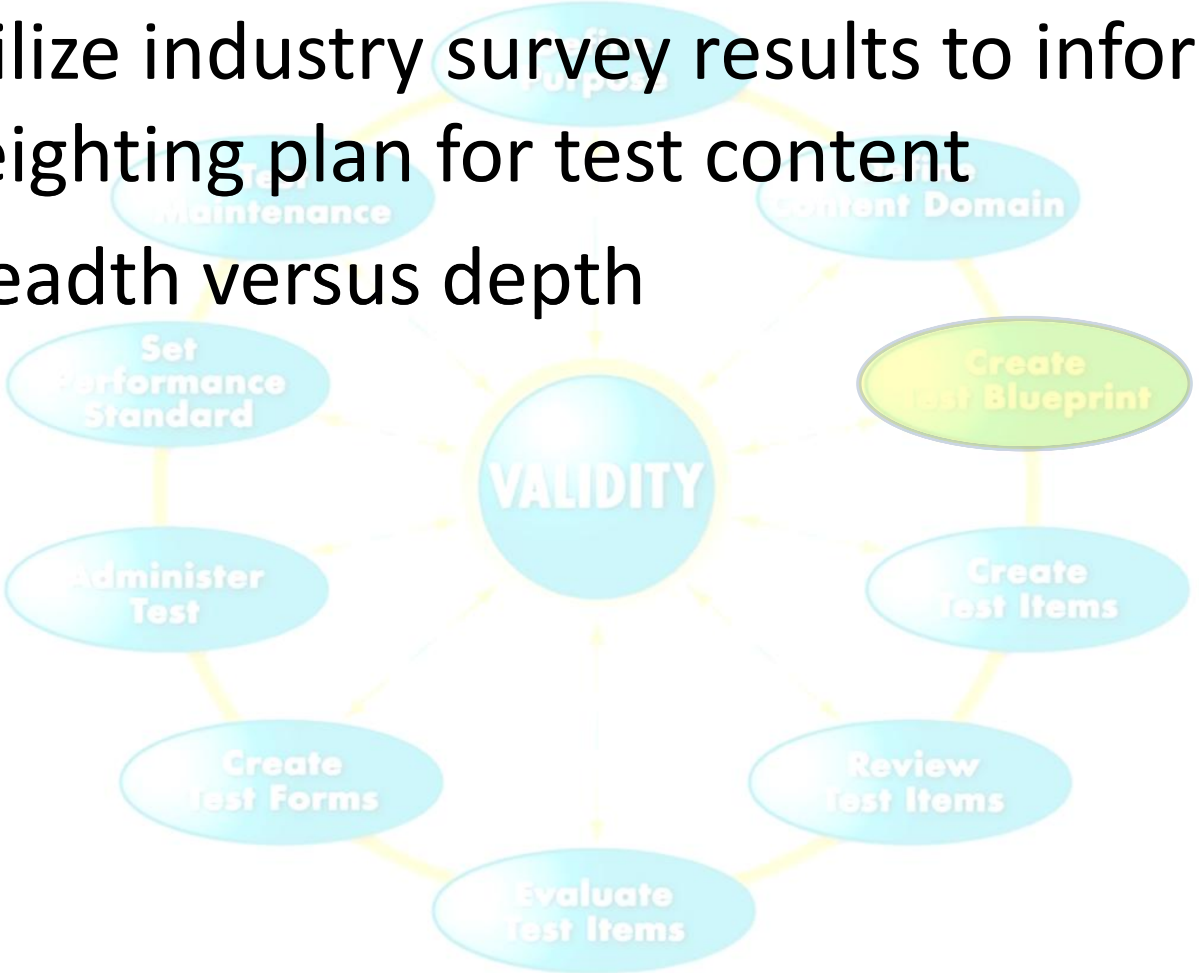
Define Content Domain

- Domain Analysis
- Consider cognitive demand processes that best reflect job requirements
- Industry-wide survey

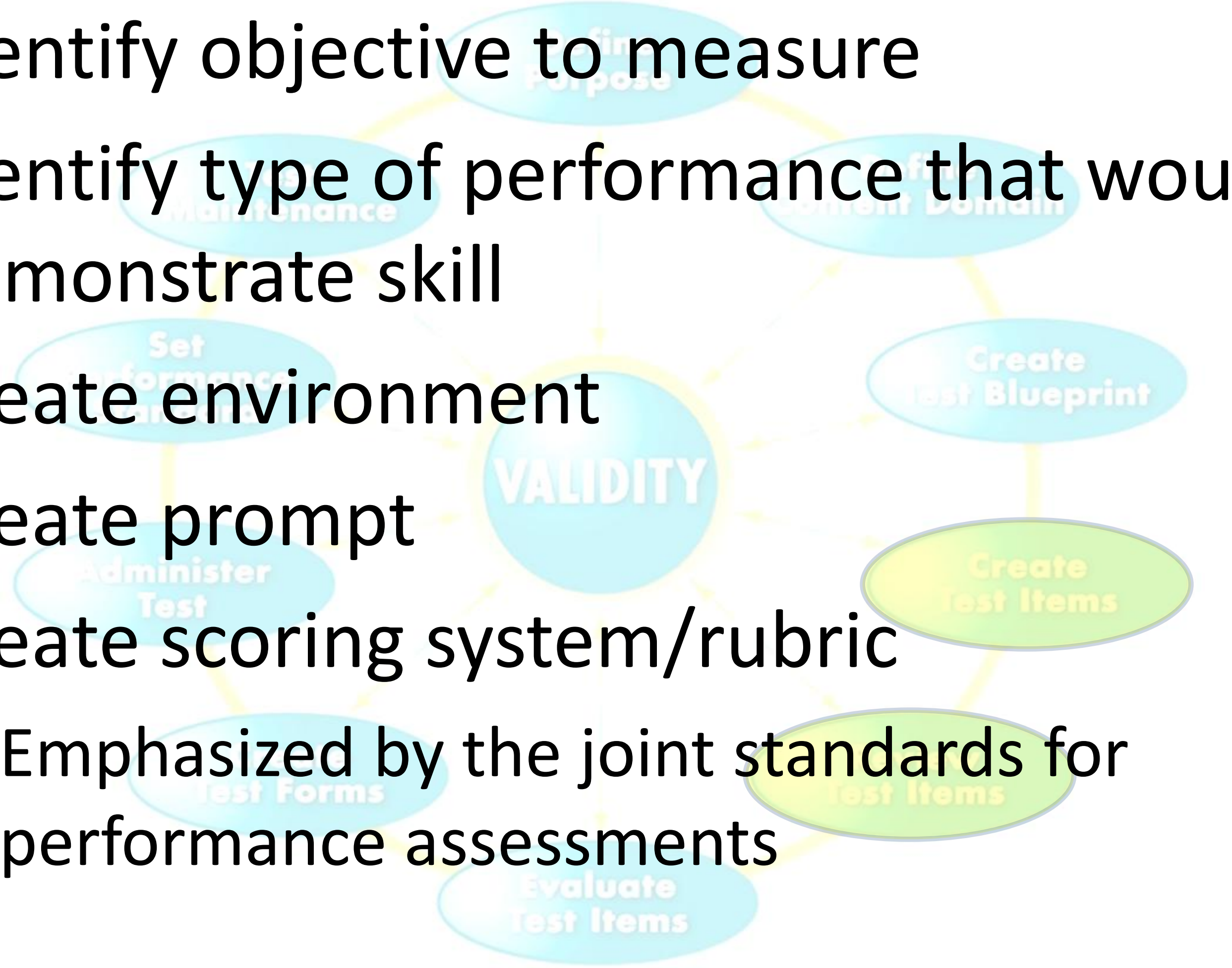


Create Test Blueprint

- Utilize industry survey results to inform weighting plan for test content
- Breadth versus depth



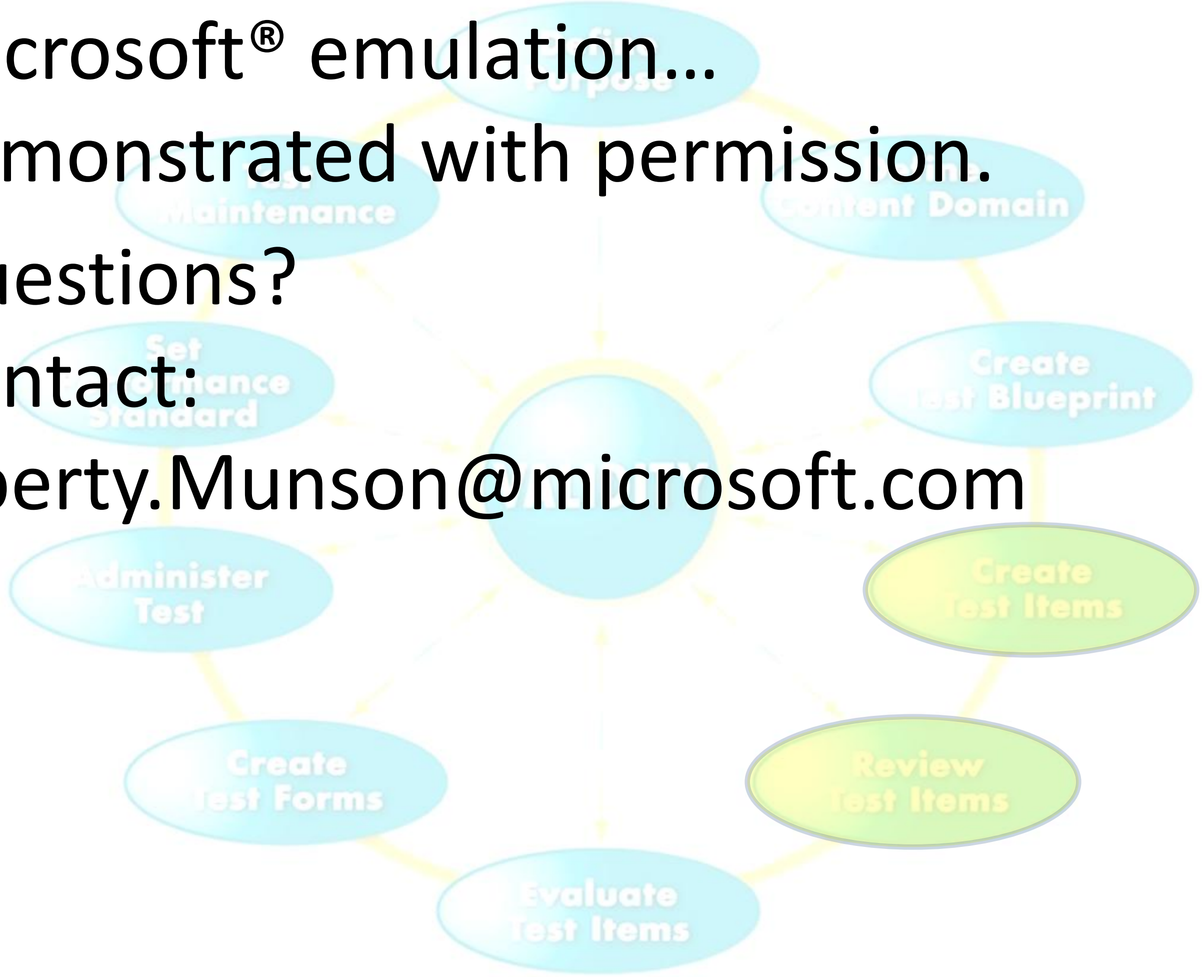
Create Test Items (Tasks)



- Identify objective to measure
- Identify type of performance that would demonstrate skill
- Create environment
- Create prompt
- Create scoring system/rubric
  - – Emphasized by the joint standards for performance assessments

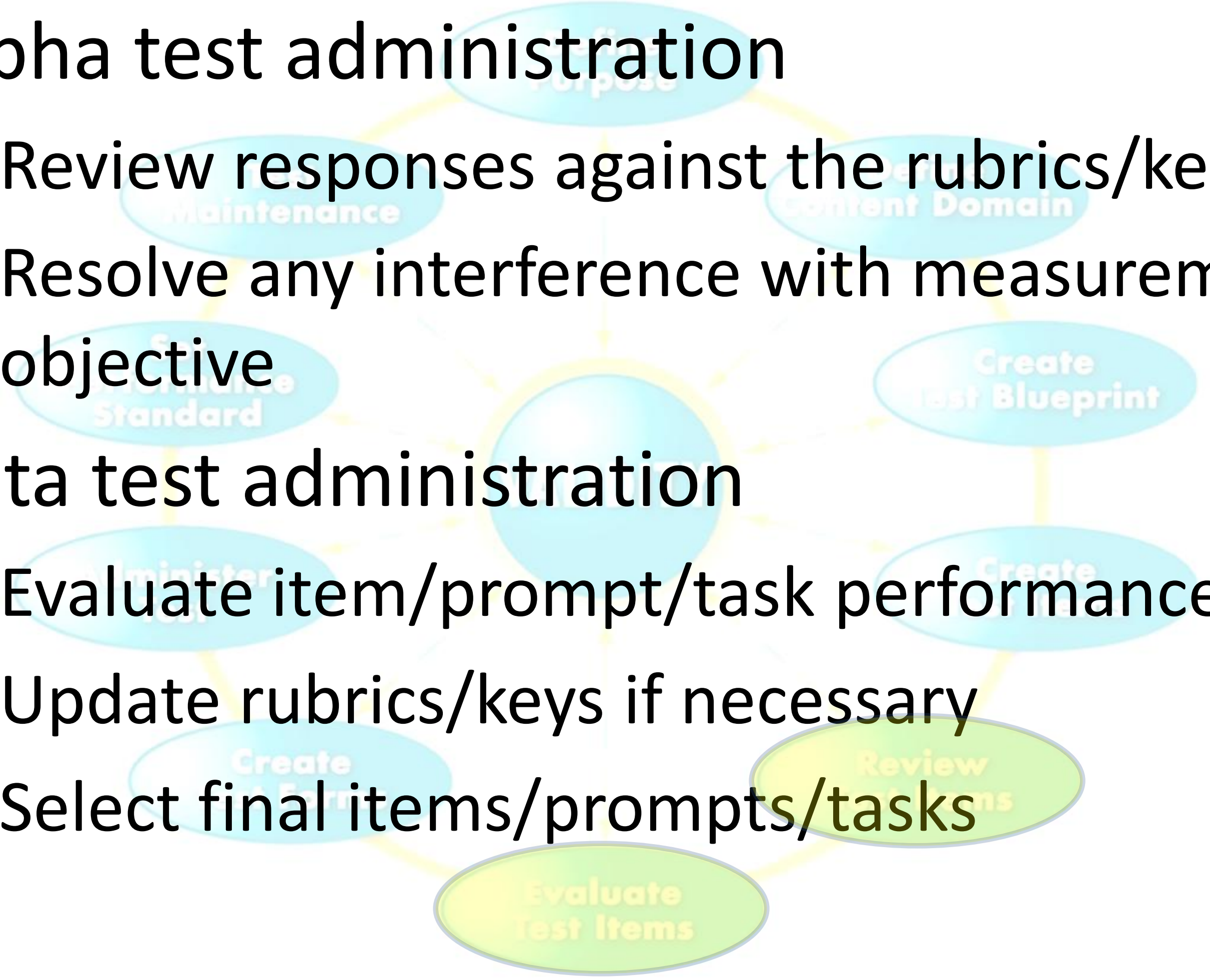
Create Test Items (Tasks)

- Microsoft® emulation… demonstrated with permission.
- Questions?
  Contact:
  Liberty.Munson@microsoft.com



Evaluate Test Items (Tasks)

- Alpha test administration
  - – Review responses against the rubrics/keys
  - – Resolve any interference with measurement objective
- Beta test administration
  - – Evaluate item/prompt/task performance
  - – Update rubrics/keys if necessary
  - – Select final items/prompts/tasks



Evaluate Test Items (Tasks)

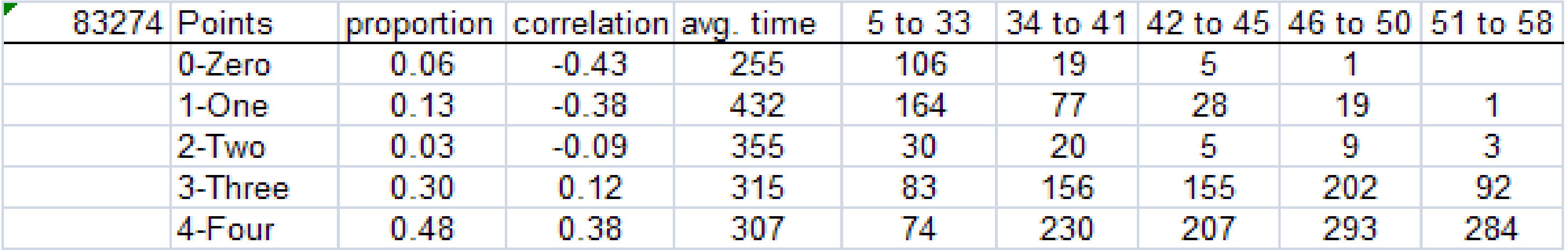
- Example: 4pt Task

Average Score	Proportion	Item-Score Correlation	Median Response Time
3.022	.756	.363	289 seconds

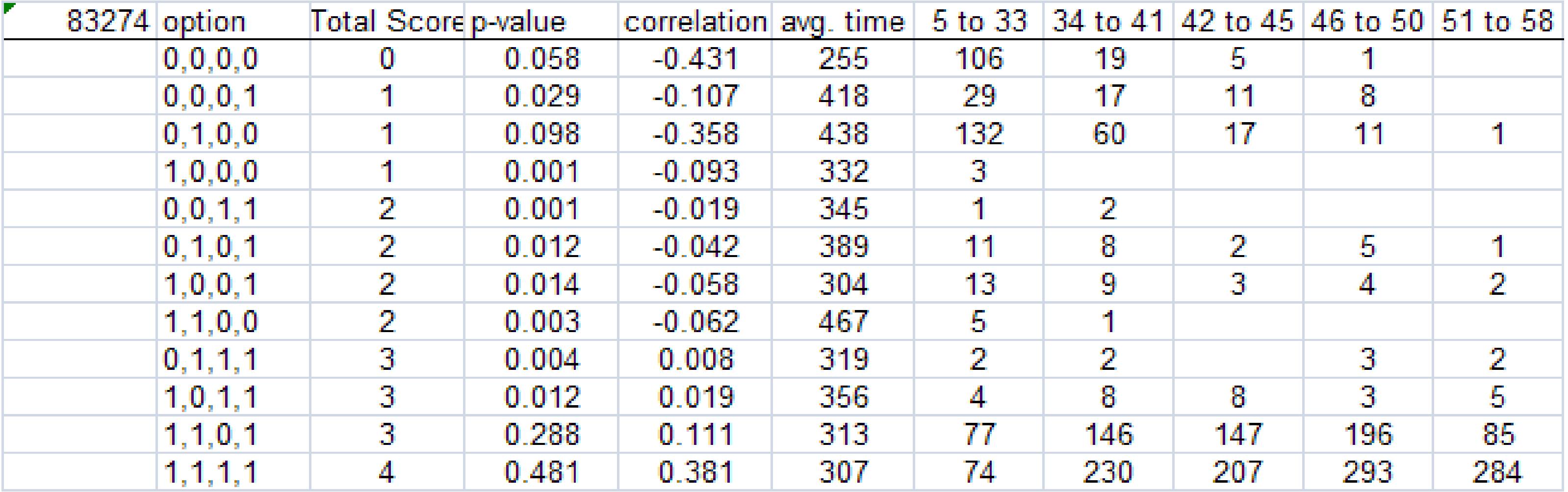
	Point 1	Point 2	Point 3	Point 4
P-value	.80	.88	.50	.84
Point biserial	.57	.40	.39	.58

Evaluate Test Items (Tasks)

- Example: 4pt Task



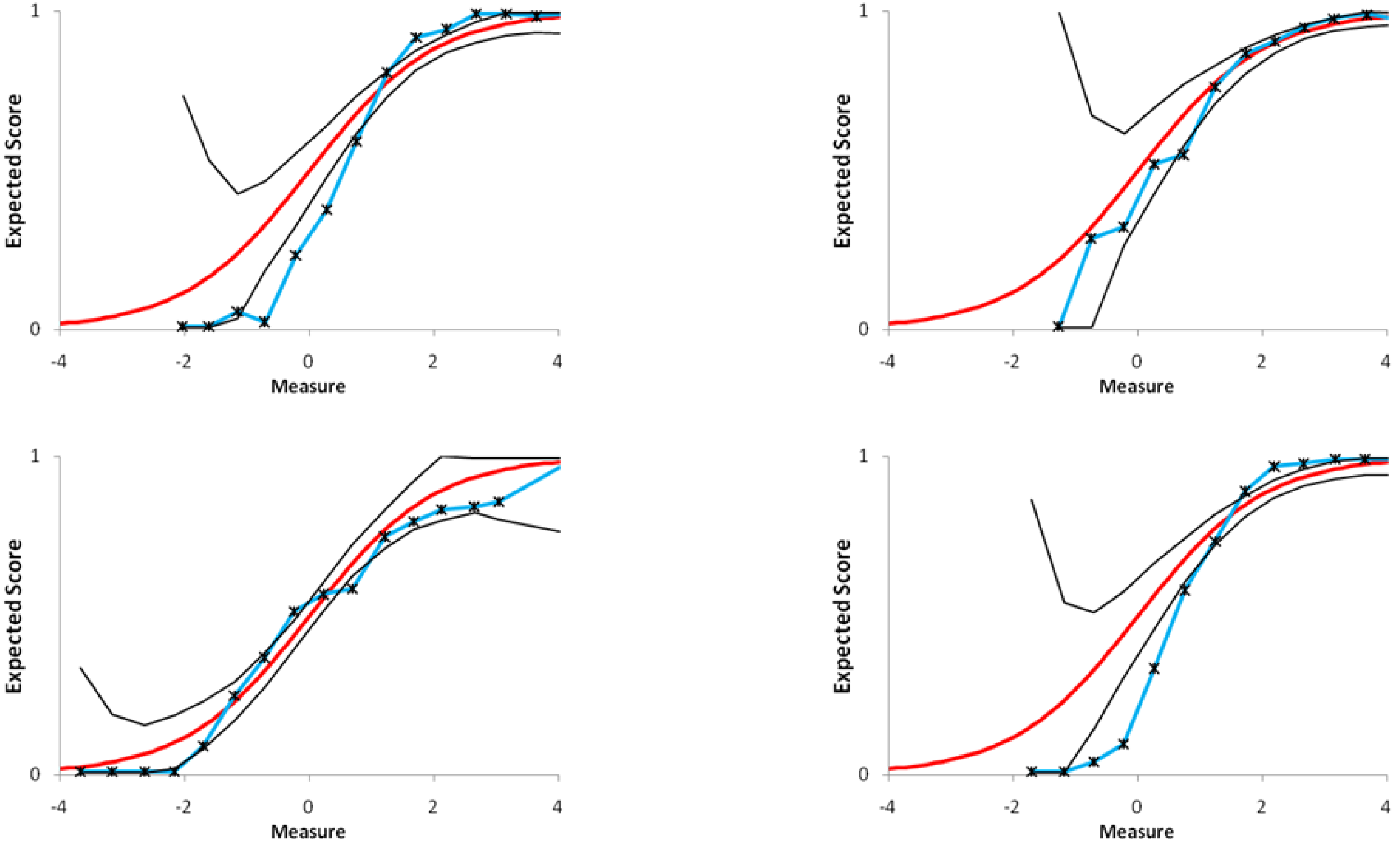
83274	Points	proportion	correlation	avg. time	5 to 33	34 to 41	42 to 45	46 to 50	51 to 58
	0-Zero	0.06	-0.43	255	106	19	5	1	
	1-One	0.13	-0.38	432	164	77	28	19	1
	2-Two	0.03	-0.09	355	30	20	5	9	3
	3-Three	0.30	0.12	315	83	156	155	202	92
	4-Four	0.48	0.38	307	74	230	207	293	284



83274	option	Total Score	p-value	correlation	avg. time	5 to 33	34 to 41	42 to 45	46 to 50	51 to 58
	0,0,0,0	0	0.058	-0.431	255	106	19	5	1	
	0,0,0,1	1	0.029	-0.107	418	29	17	11	8	
	0,1,0,0	1	0.098	-0.358	438	132	60	17	11	1
	1,0,0,0	1	0.001	-0.093	332	3				
	0,0,1,1	2	0.001	-0.019	345	1	2			
	0,1,0,1	2	0.012	-0.042	389	11	8	2	5	1
	1,0,0,1	2	0.014	-0.058	304	13	9	3	4	2
	1,1,0,0	2	0.003	-0.062	467	5	1			
	0,1,1,1	3	0.004	0.008	319	2	2		3	2
	1,0,1,1	3	0.012	0.019	356	4	8	8	3	5
	1,1,0,1	3	0.288	0.111	313	77	146	147	196	85
	1,1,1,1	4	0.481	0.381	307	74	230	207	293	284

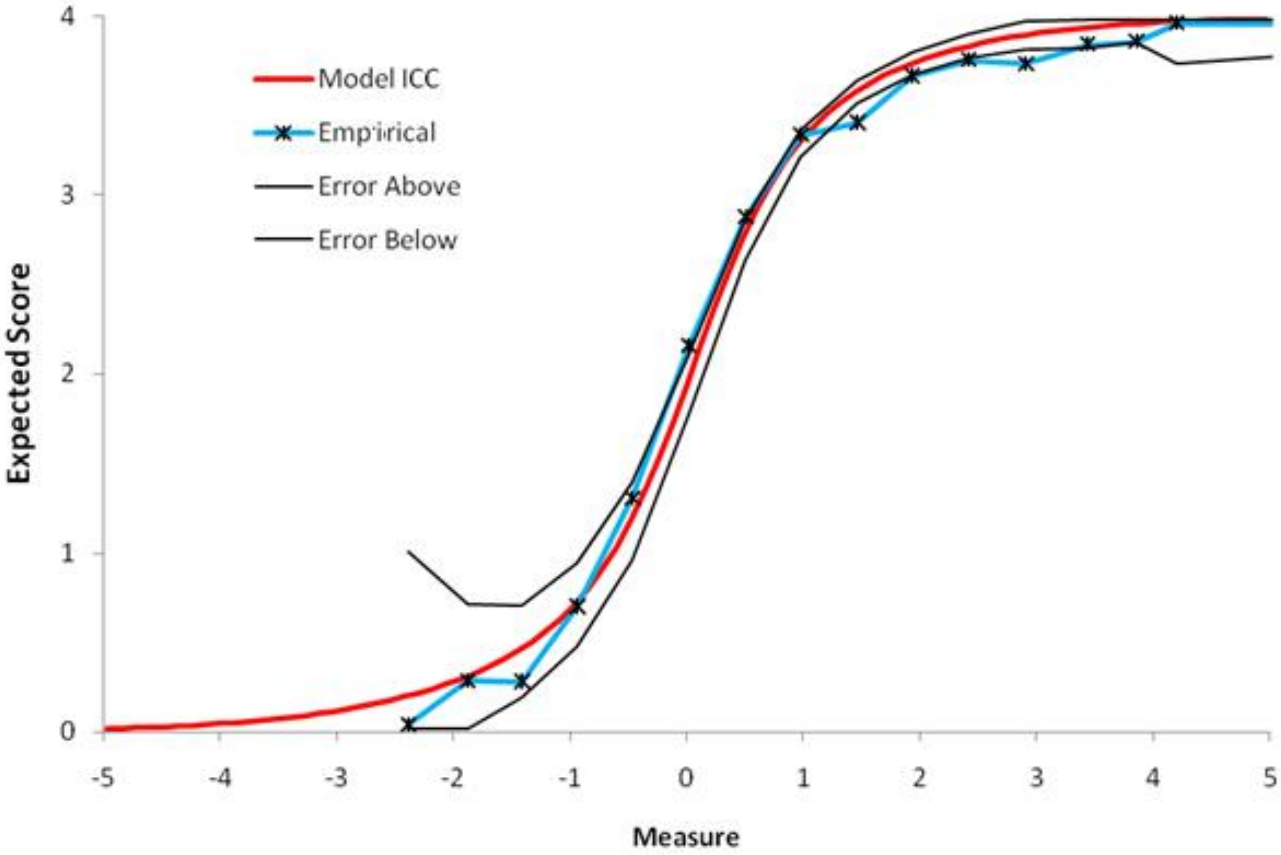
Evaluate Test Items (Tasks)

- Example: 4pt Task



Evaluate Test Items (Tasks)

- Example: 4pt Task



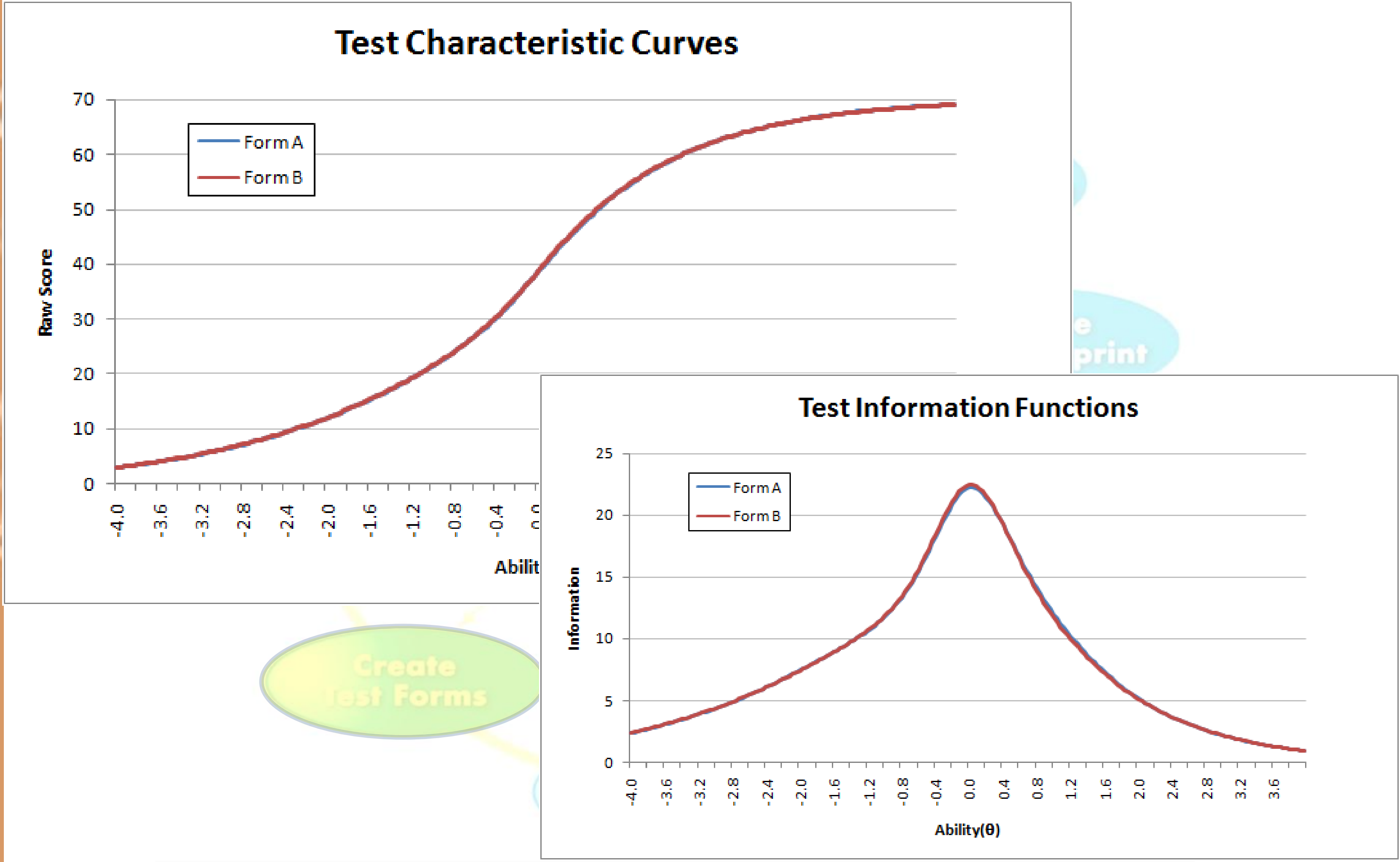
Evaluate Test Items (Tasks)

Item type	Count	Est. p-value	Point measure**	Item time***
Multiple choice	639	.76	.36	58.8
Multiple select	286	.67	.41	69.2
Drag & drop	42	.62	.38	99.1
Graphical drag & drop	5	.53	.40	142.5
Flash dynamic hybrid item	2	.62	.61	629.2
Simlet	7	.52	.52	628.2
Simulation	27	.58	.60	594.3
Testlet	3	.71	.54	256.4

**Average point measure correlation

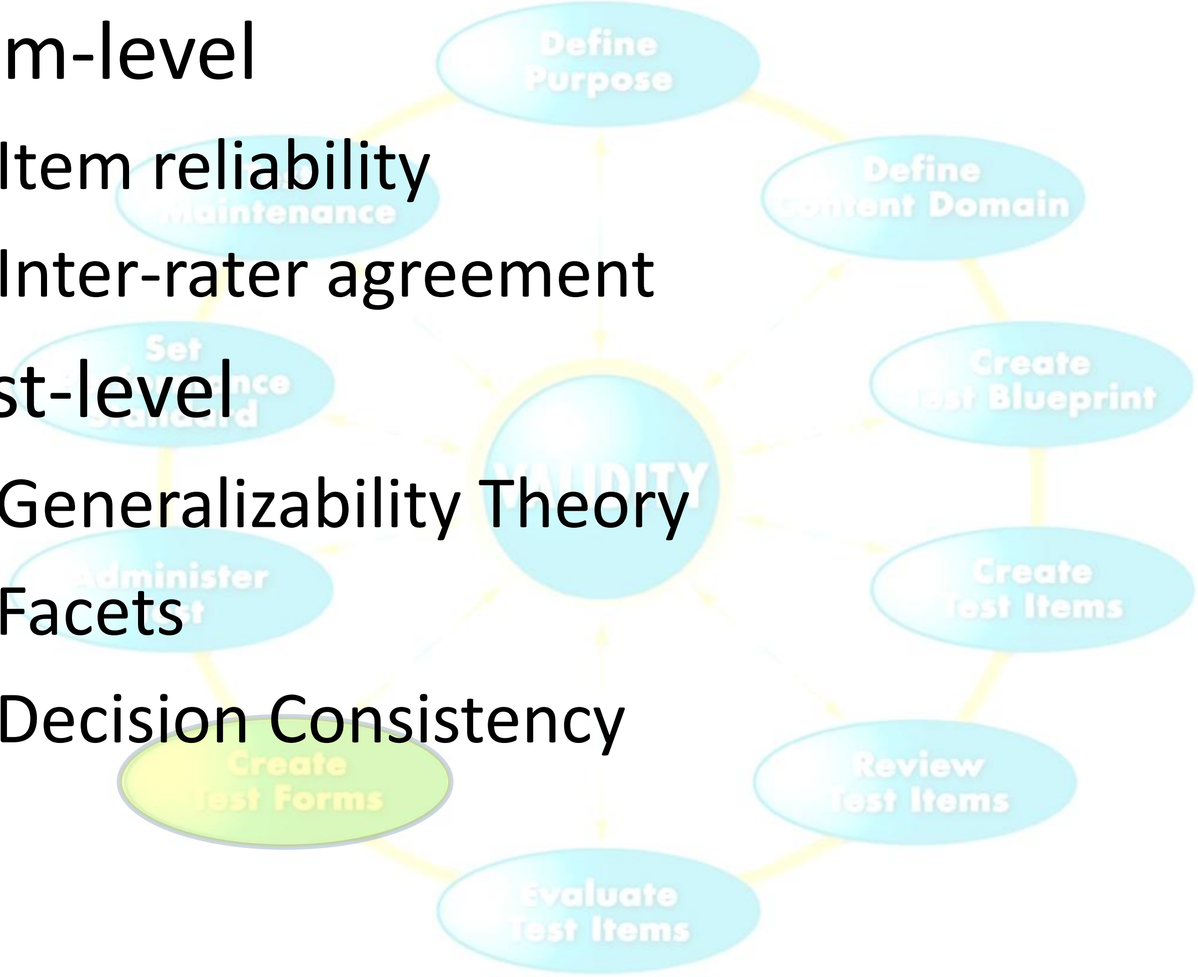
*** Time in seconds

Create Test Forms



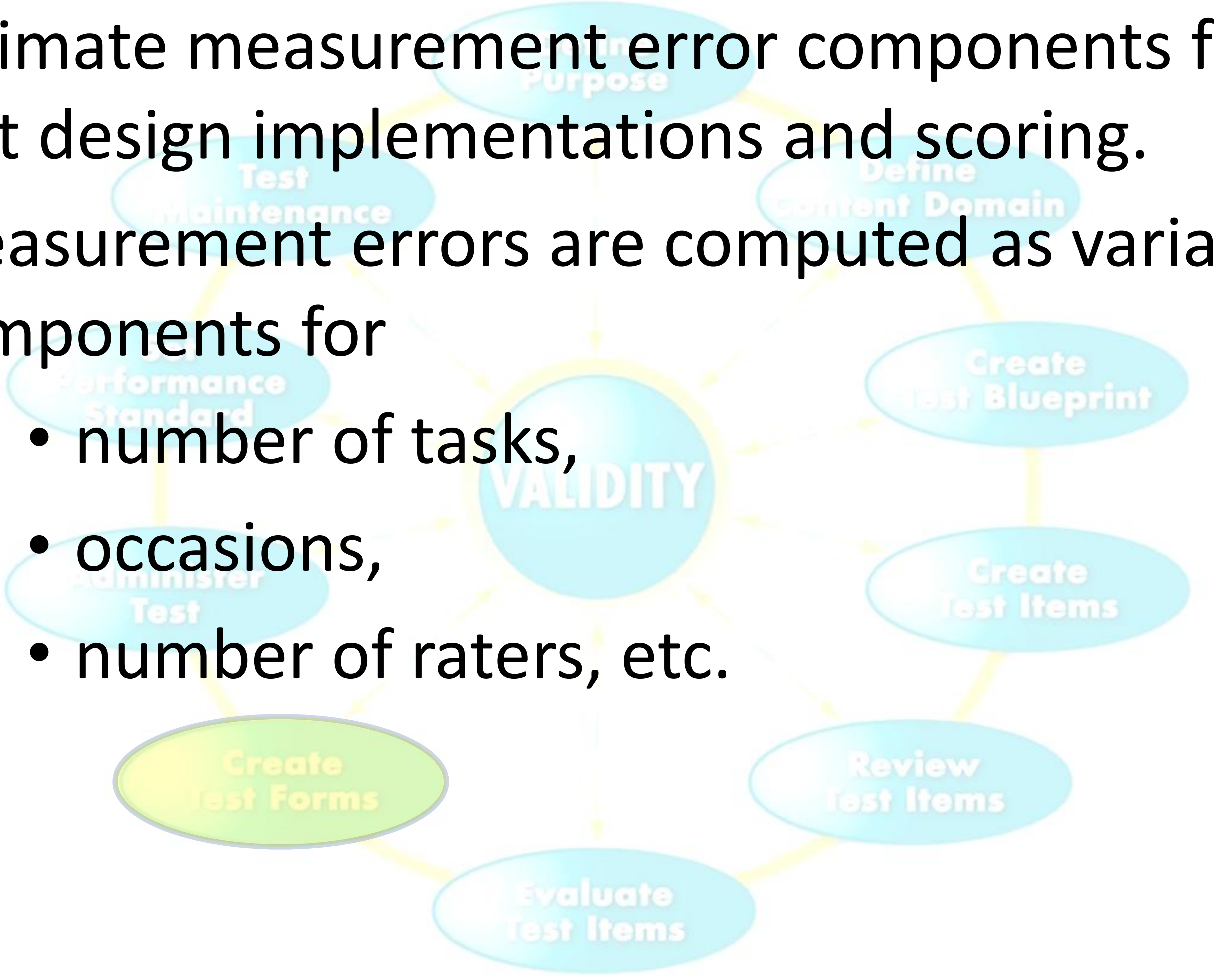
Create Test Forms

- Item-level
  - Item reliability
  - Inter-rater agreement
- Test-level
  - Generalizability Theory
  - Facets
  - Decision Consistency



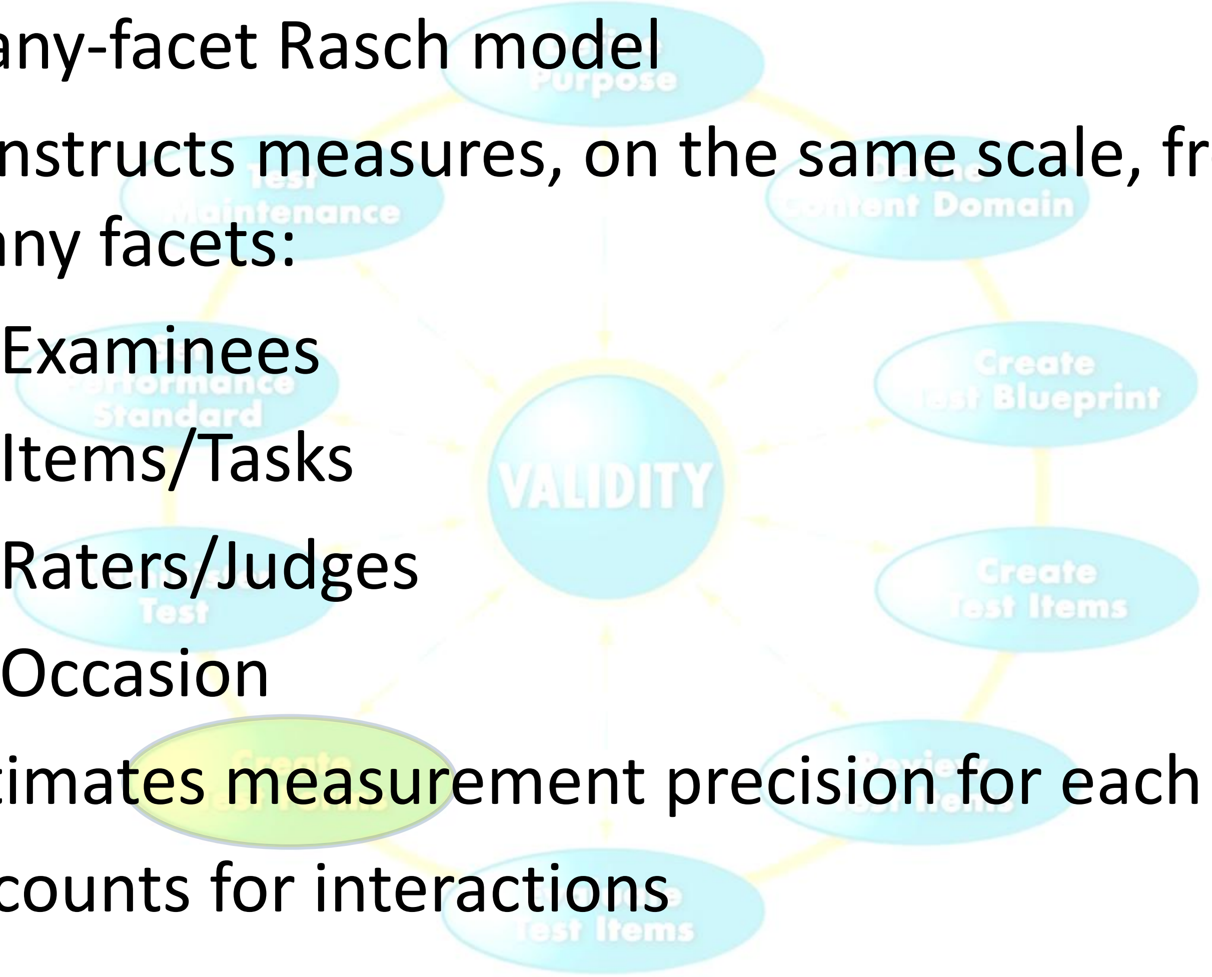
Generalizability Theory

- Estimate measurement error components for test design implementations and scoring.
- Measurement errors are computed as variance components for
    - number of tasks,
    - occasions,
    - number of raters, etc.

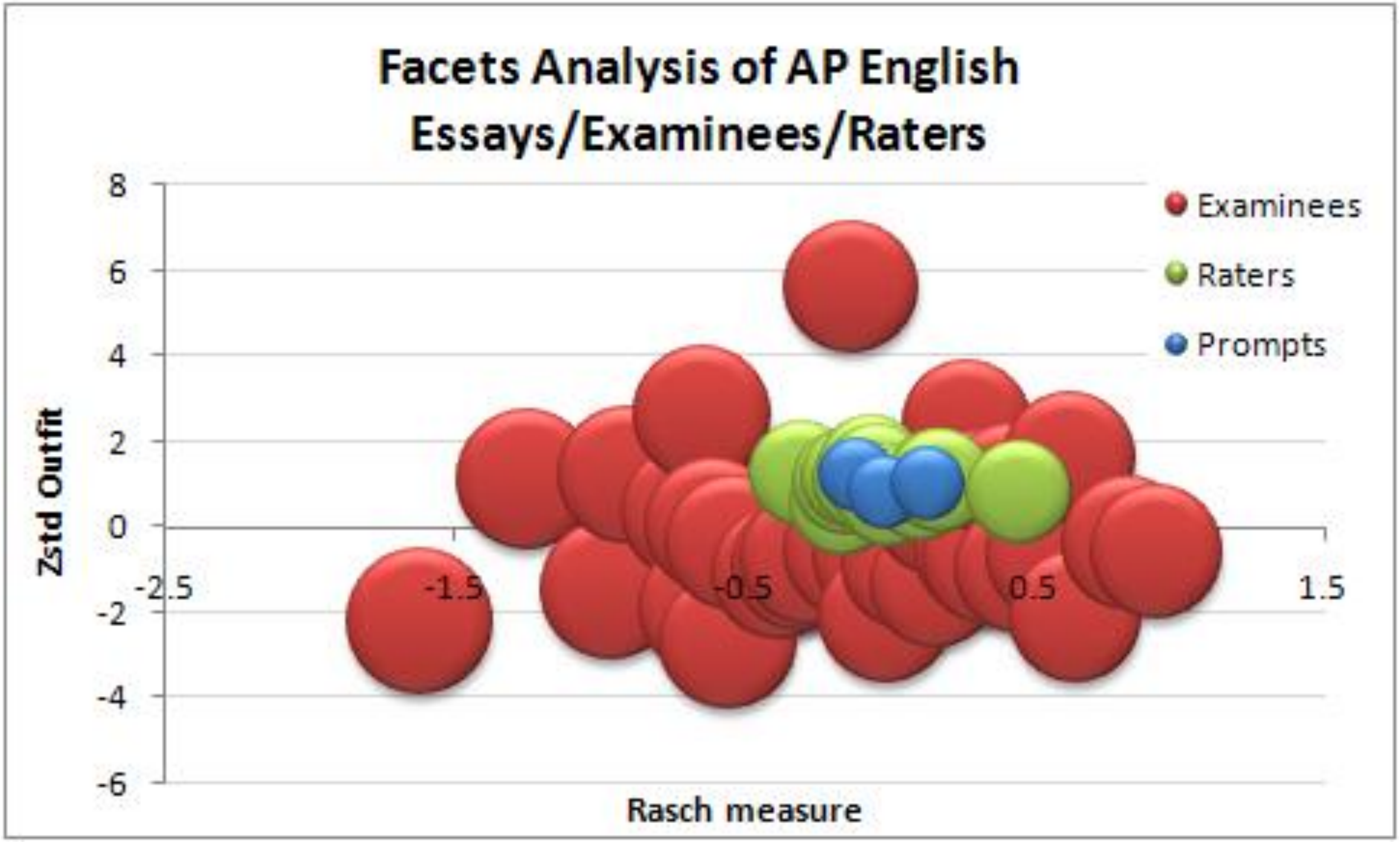


Facets Analysis

- Many-facet Rasch model
- Constructs measures, on the same scale, from many facets:
  - Examinees
  - Items/Tasks
  - Raters/Judges
  - Occasion
- Estimates measurement precision for each
- Accounts for interactions

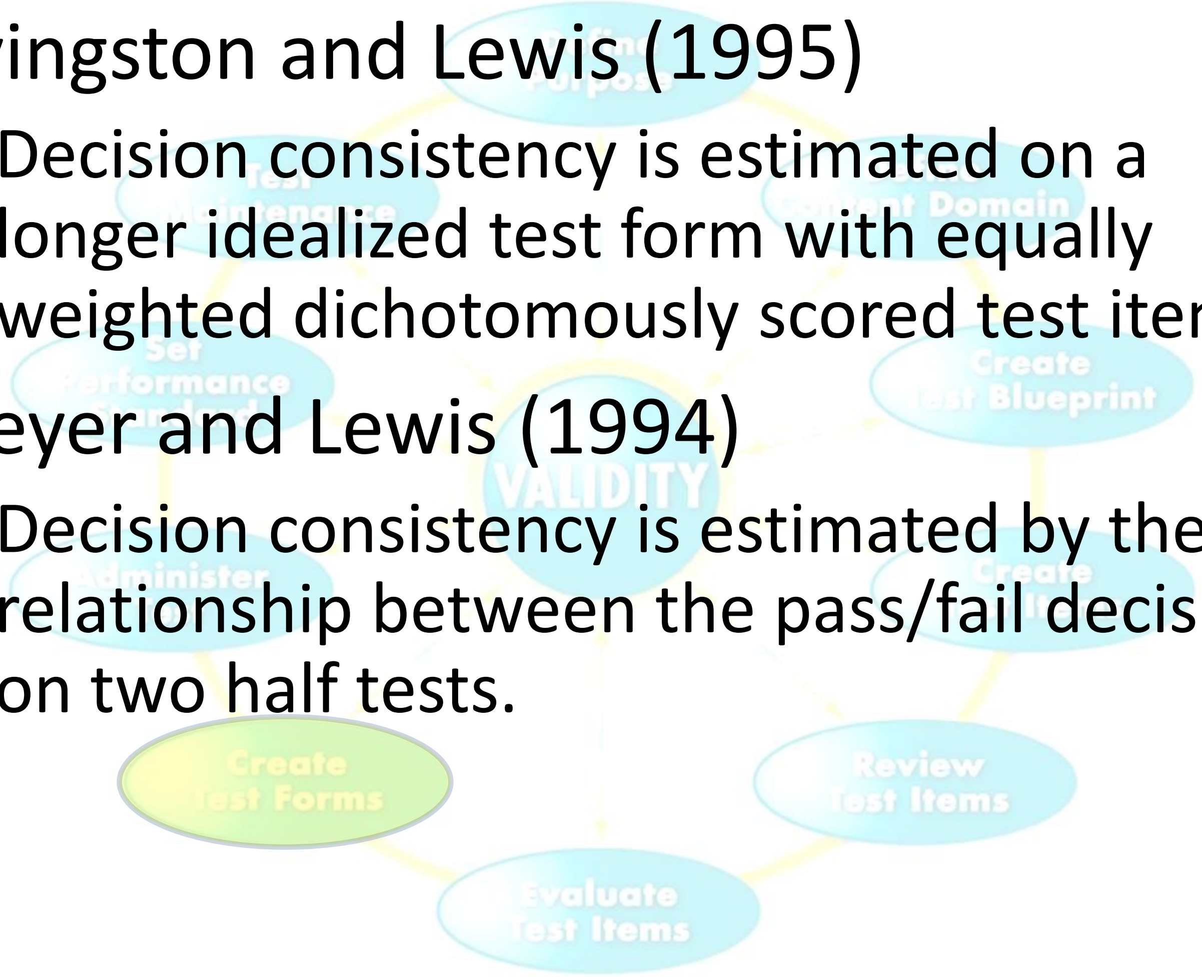


Facets Analysis



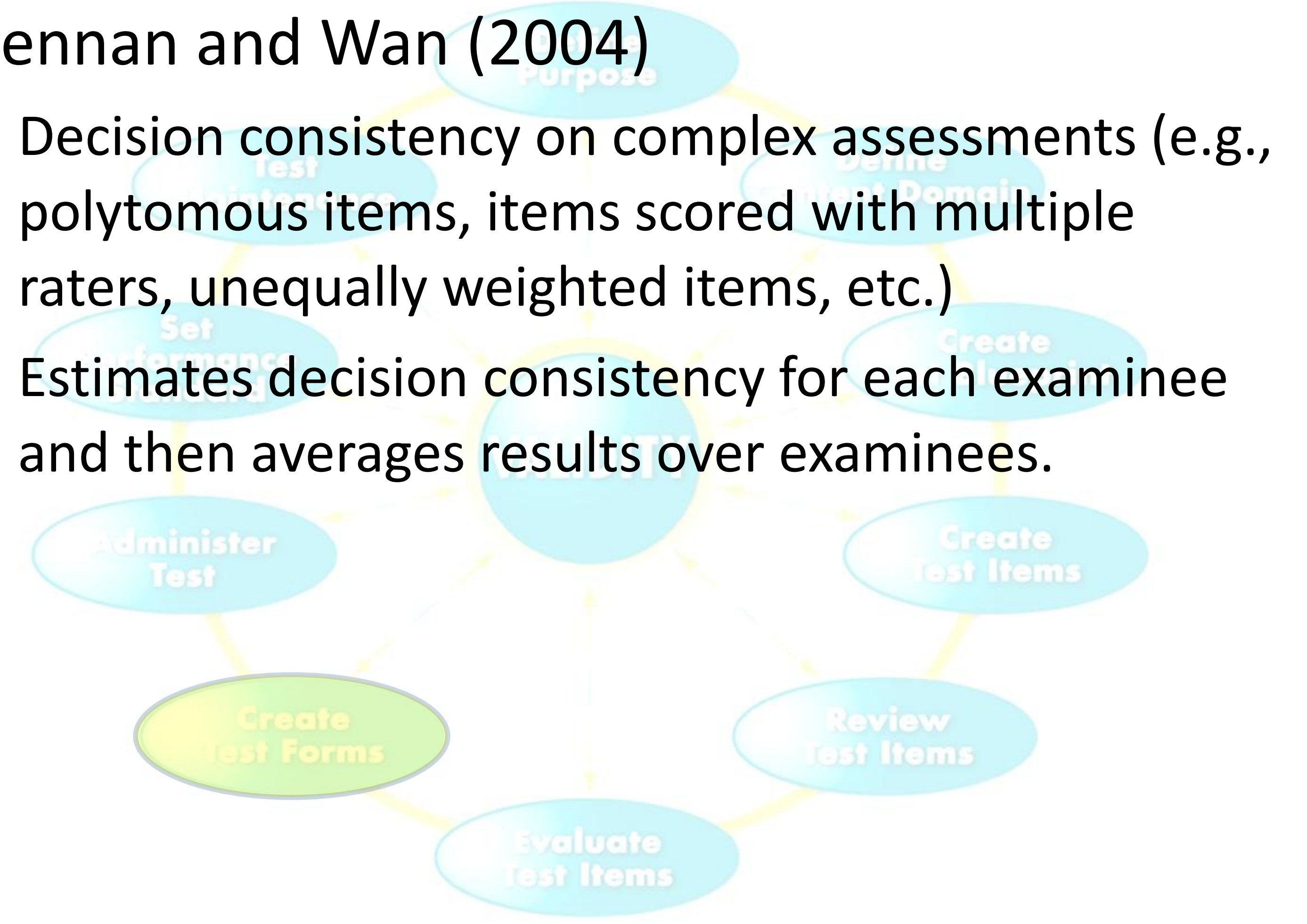
Decision Consistency

- Livingston and Lewis (1995)
  - Decision consistency is estimated on a longer idealized test form with equally weighted dichotomously scored test items.
- Breyer and Lewis (1994)
  - Decision consistency is estimated by the relationship between the pass/fail decision on two half tests.

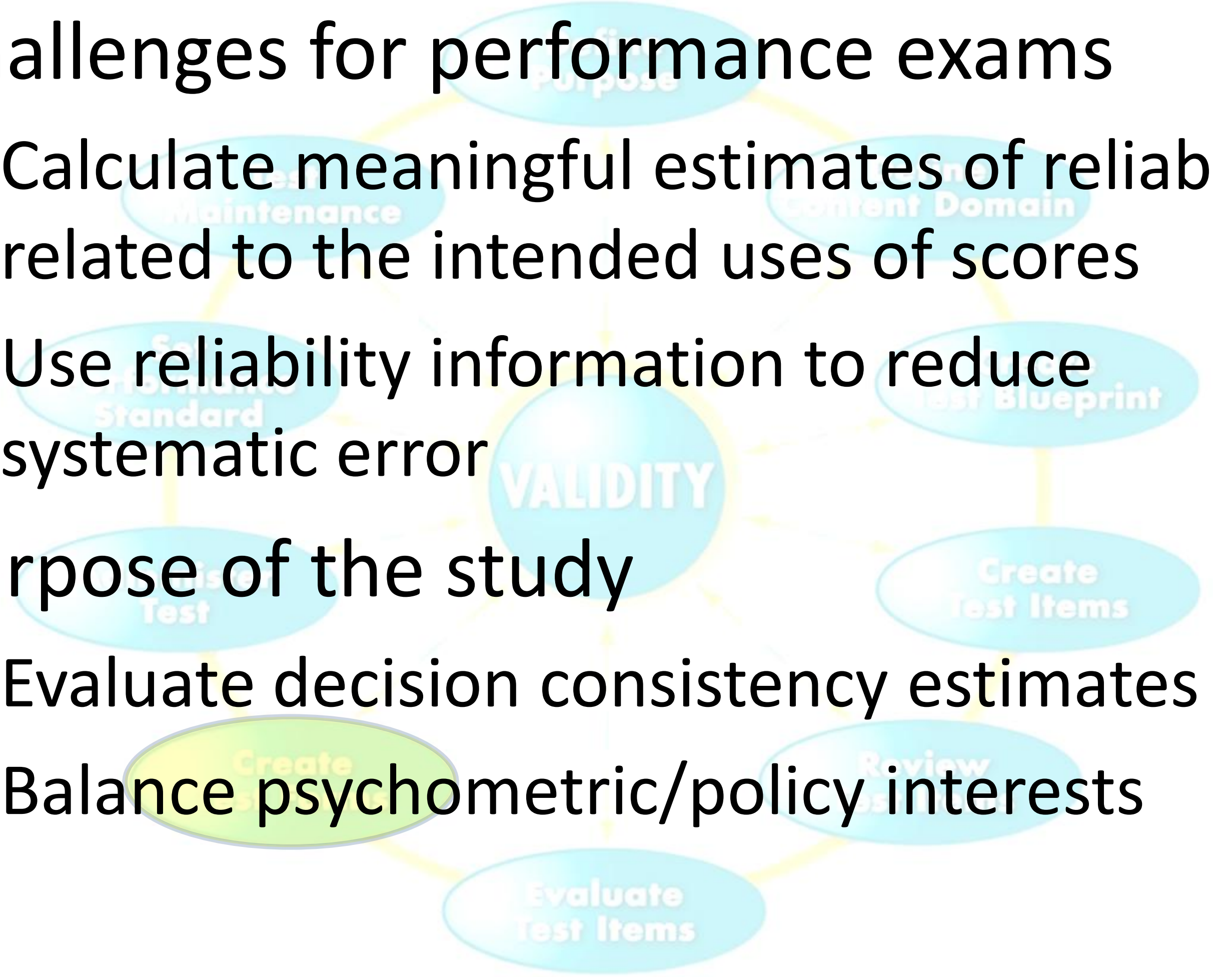


Decision Consistency

- Brennan and Wan (2004)
  - Decision consistency on complex assessments (e.g., polytomous items, items scored with multiple raters, unequally weighted items, etc.)
  - Estimates decision consistency for each examinee and then averages results over examinees.



Rater agreement and decision consistency analysis example (Buckendahl, 2009)



- Challenges for performance exams
  - Calculate meaningful estimates of reliability related to the intended uses of scores
  - Use reliability information to reduce systematic error
- Purpose of the study
  - Evaluate decision consistency estimates
  - Balance psychometric/policy interests

Typodont and Manikins

Define Purpose

Create
est Items

w
ms

Create
Test Forms

Results – Embedded Performances

	Exact Agreement (D.C.)
By section	
Endodontics (n=30)	68% (72%)
Fixed Pros (n=30)	13% (76%)
By site	
Site 1 (n=10 models)	47% (70%)
Site 2 (n=6)	47% (72%)
Site 3 (n=10)	45% (83%)
Site 4 (n=4)	17% (63%)

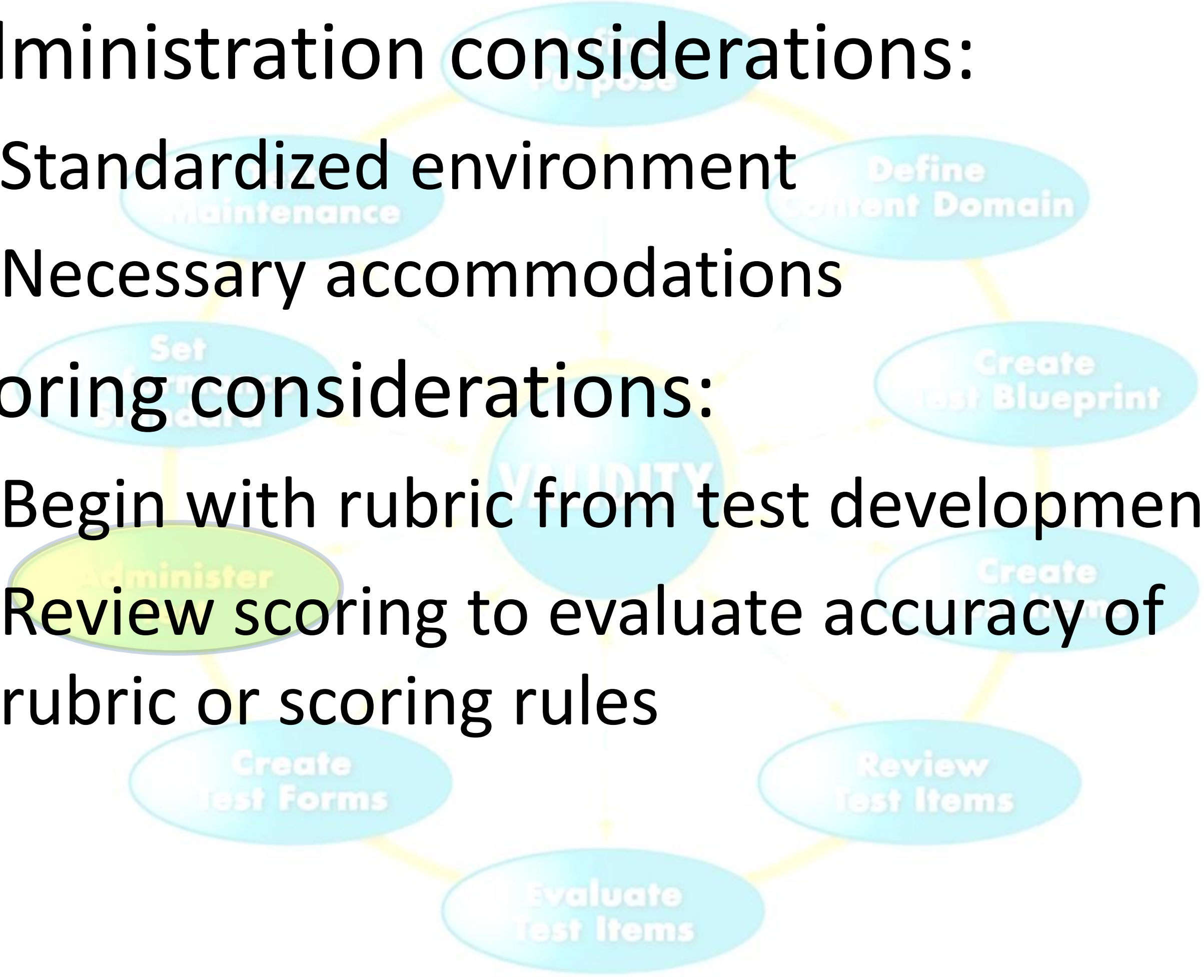
Results – Decision Consistency (2007)

	Dec. Cons. (Flag. Exam.)
Amalgam	95% (4)
Composite	97% (3)
Endodontics	98% (3)
Fixed Pros	94% (8)

- % of instances where examiners individually would have agreed with the actual decision across ~300 candidates

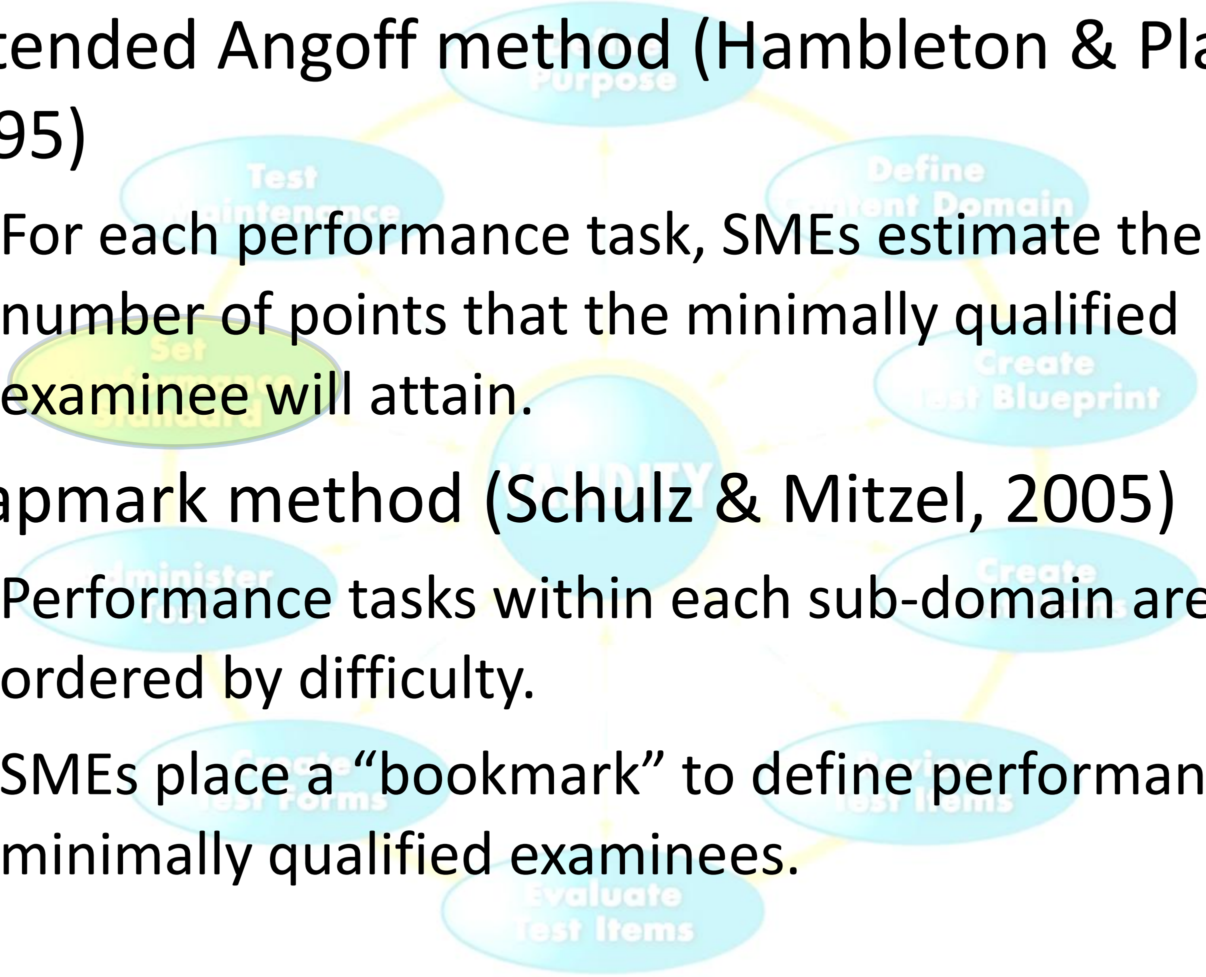
Administer Test

- Administration considerations:
  - Standardized environment
  - Necessary accommodations
- Scoring considerations:
  - Begin with rubric from test development
  - Review scoring to evaluate accuracy of rubric or scoring rules

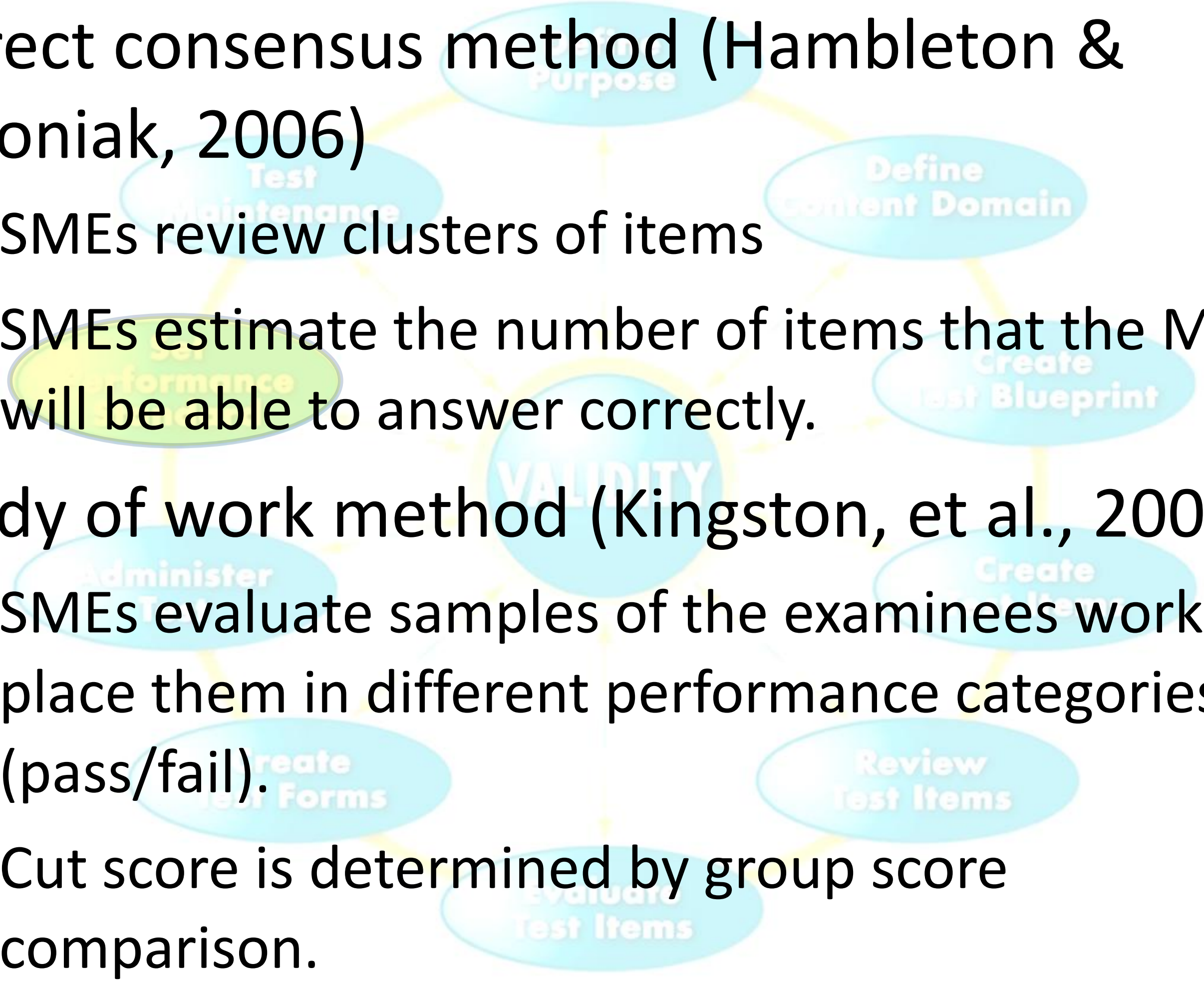


Set Performance Standard

- Extended Angoff method (Hambleton & Plake, 1995)
  - For each performance task, SMEs estimate the number of points that the minimally qualified examinee will attain.
- Mapmark method (Schulz & Mitzel, 2005)
  - Performance tasks within each sub-domain are ordered by difficulty.
  - SMEs place a “bookmark” to define performance of minimally qualified examinees.



Set Performance Standard

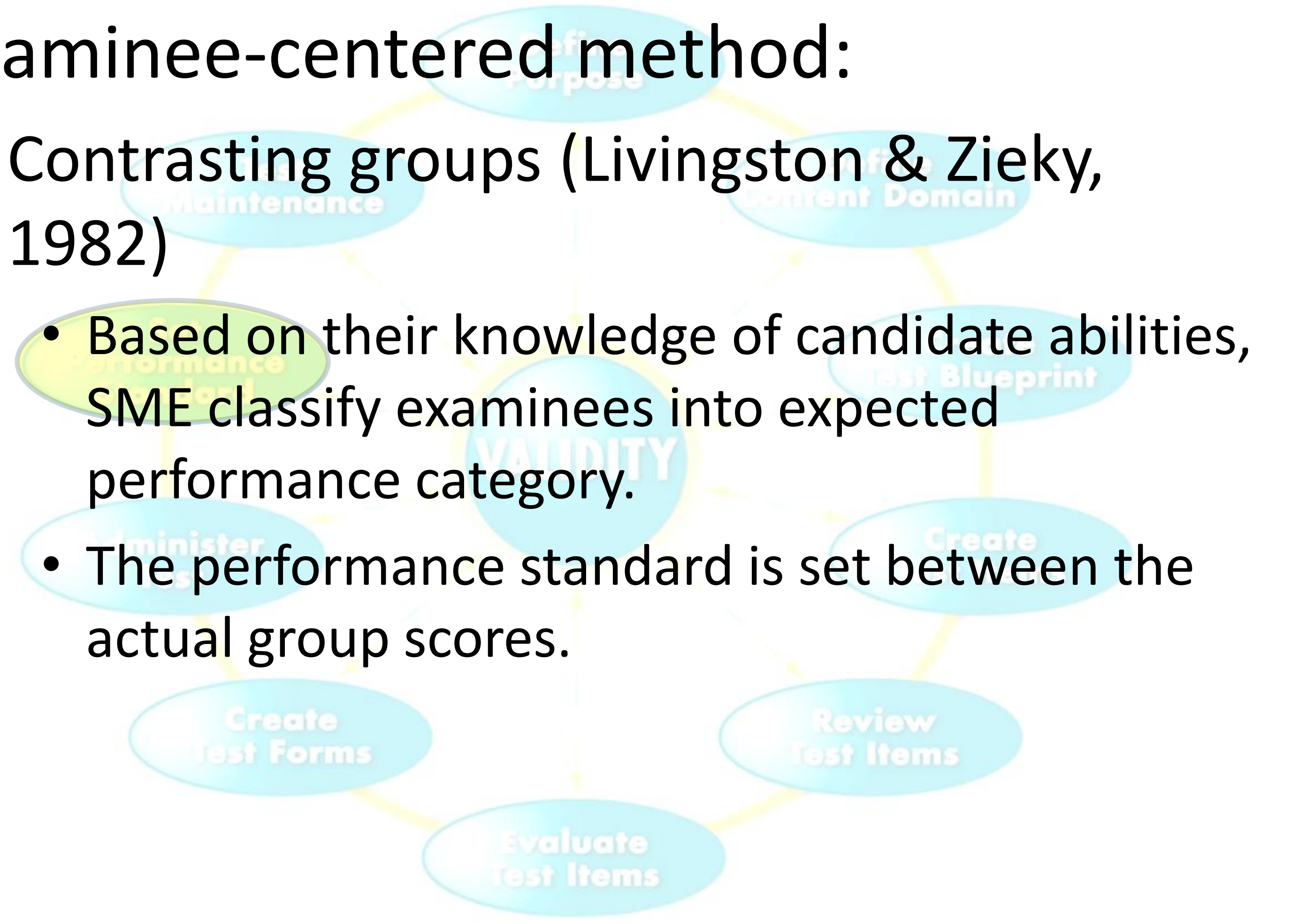


- Direct consensus method (Hambleton & Pitoniak, 2006)
  - SMEs review clusters of items
  - SMEs estimate the number of items that the MQC will be able to answer correctly.
- Body of work method (Kingston, et al., 2001)
  - SMEs evaluate samples of the examinees work and place them in different performance categories (pass/fail).
  - Cut score is determined by group score comparison.

Set Performance Standard

- Dominant profile method (Plake, et al., 1997)
  - SMEs review candidate score profiles across different performance tasks
  - Create a policy and/or combination of decision rules to represent a performance standard.
- Judgment policy capturing (Jaeger, 1995)
  - SMEs review score candidate profiles across performance tasks and classify each score profile to a proficiency category
  - Candidate scores are analyzed to determine each panelist’s standard setting policy.

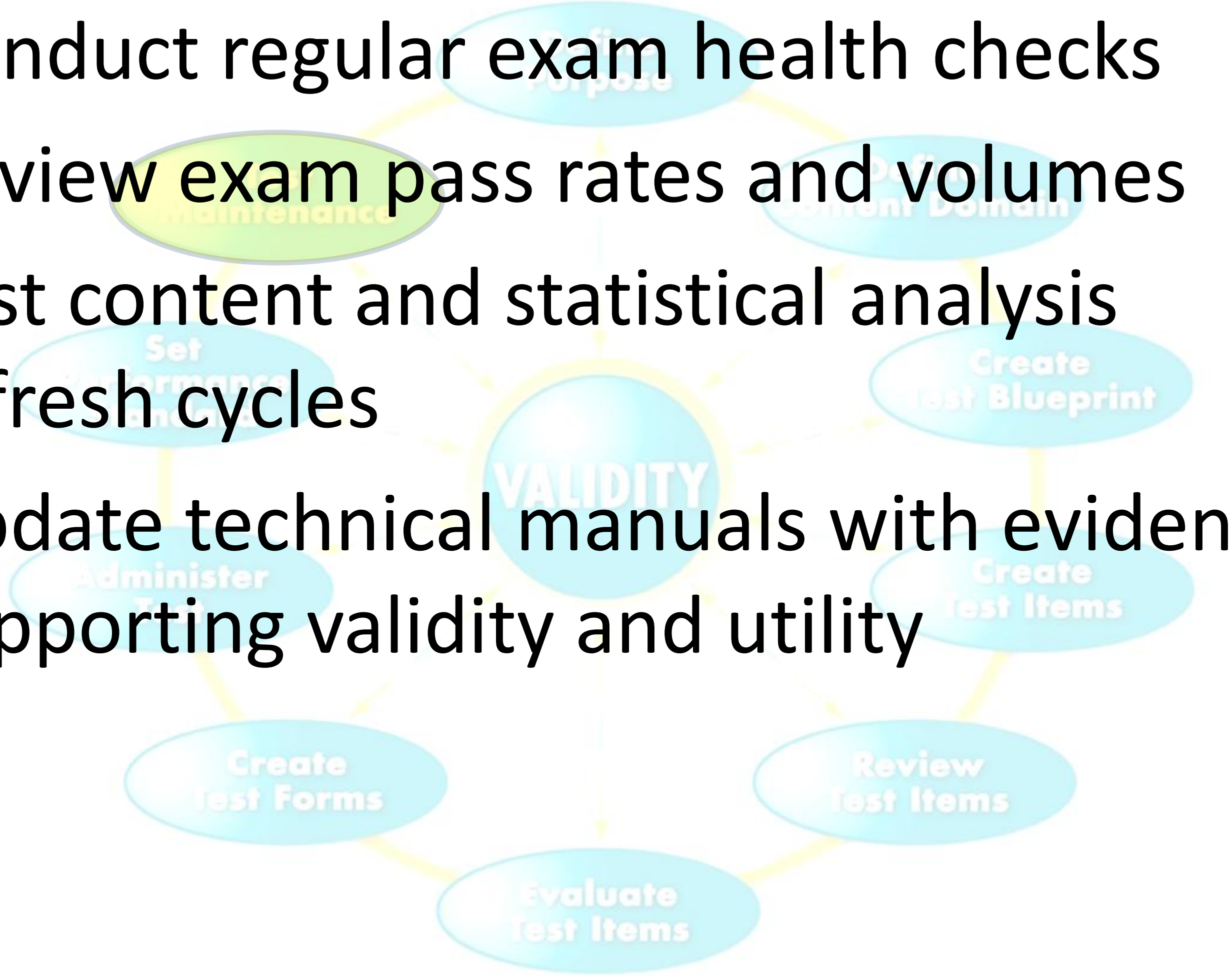
Set Performance Standard



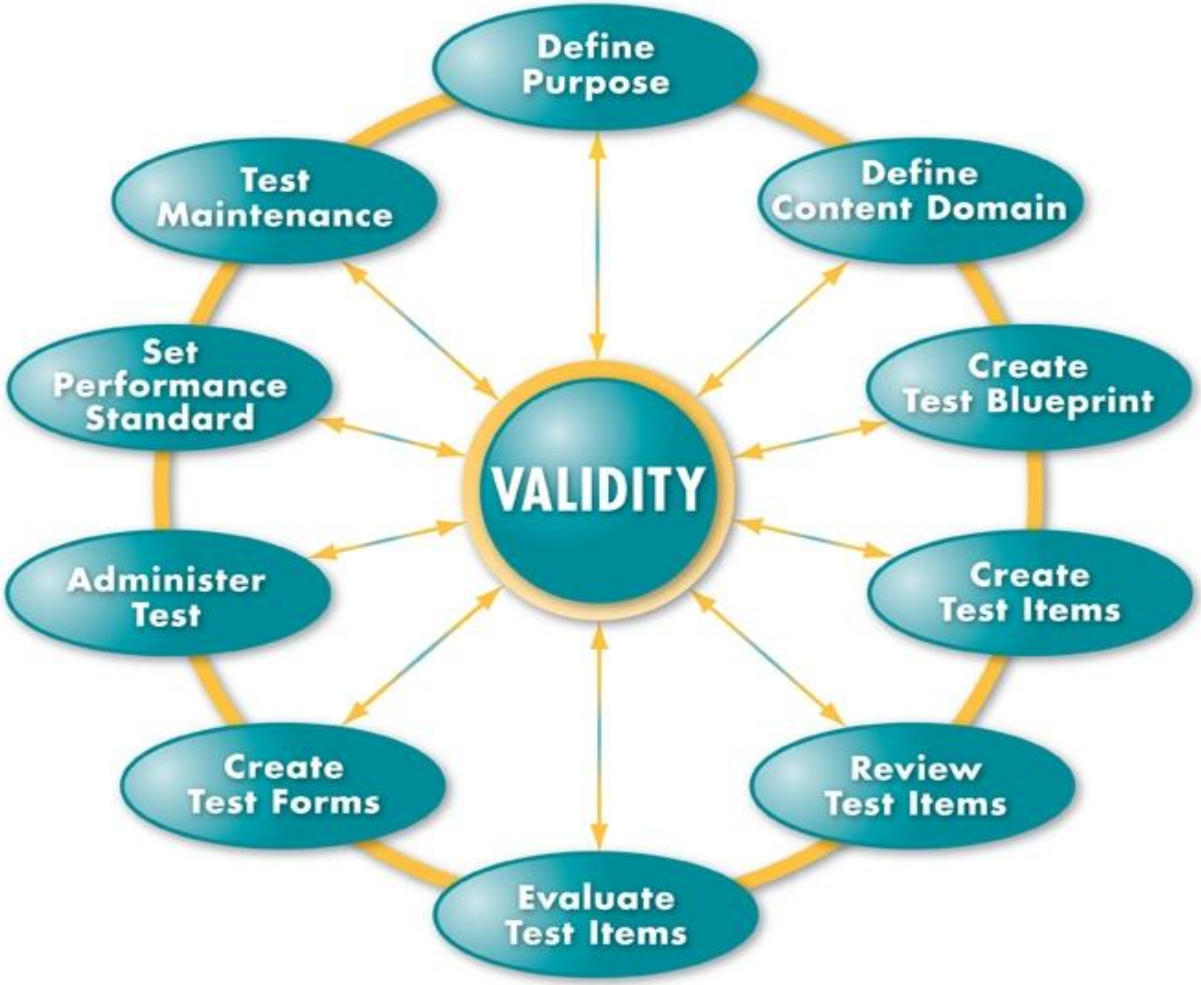
- Examinee-centered method:
  - Contrasting groups (Livingston & Zieky, 1982)
    - Based on their knowledge of candidate abilities, SME classify examinees into expected performance category.
    - The performance standard is set between the actual group scores.

Test Maintenance

- Conduct regular exam health checks
- Review exam pass rates and volumes
- Test content and statistical analysis refresh cycles
- Update technical manuals with evidence supporting validity and utility



Performance Test Development & Validation Framework



Conclusions and recommendations

- Center validity framework on the *interpretation and use of test results.*
- Plan for test enhancement and revision.
- Consider cost-benefit of design, development, delivery and scoring.
- Demonstrate that scores/decisions are reliable.
- Determine the unique measurement capabilities of various item types.

FINALLY….

- *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) apply to performance testing and assessment environments.
- However, still opportunities for research on compiling and documenting evidence for validity, reliability, fairness and legal defensibility.

Contact Information

- James B. Olsen, PhD
  jim.olsen@alpinetesting.com

- Russell W. Smith, PhD
  russell.smith@alpineteting.com

- Cristina Goodwin
  cristina.goodwin@alpineteting.com

Thank You

Alpine Testing Solutions