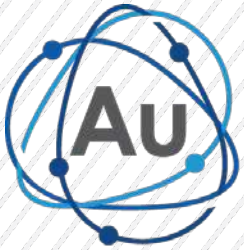


Alpine University

Visual Displays: Design Principles for Testing Programs

Brett P. Foley, PhD

Psychometrician

December 22, 2015

AlpineTesting.com
au@alpinetesting.com
(844) 625-7463

Workshop Objectives

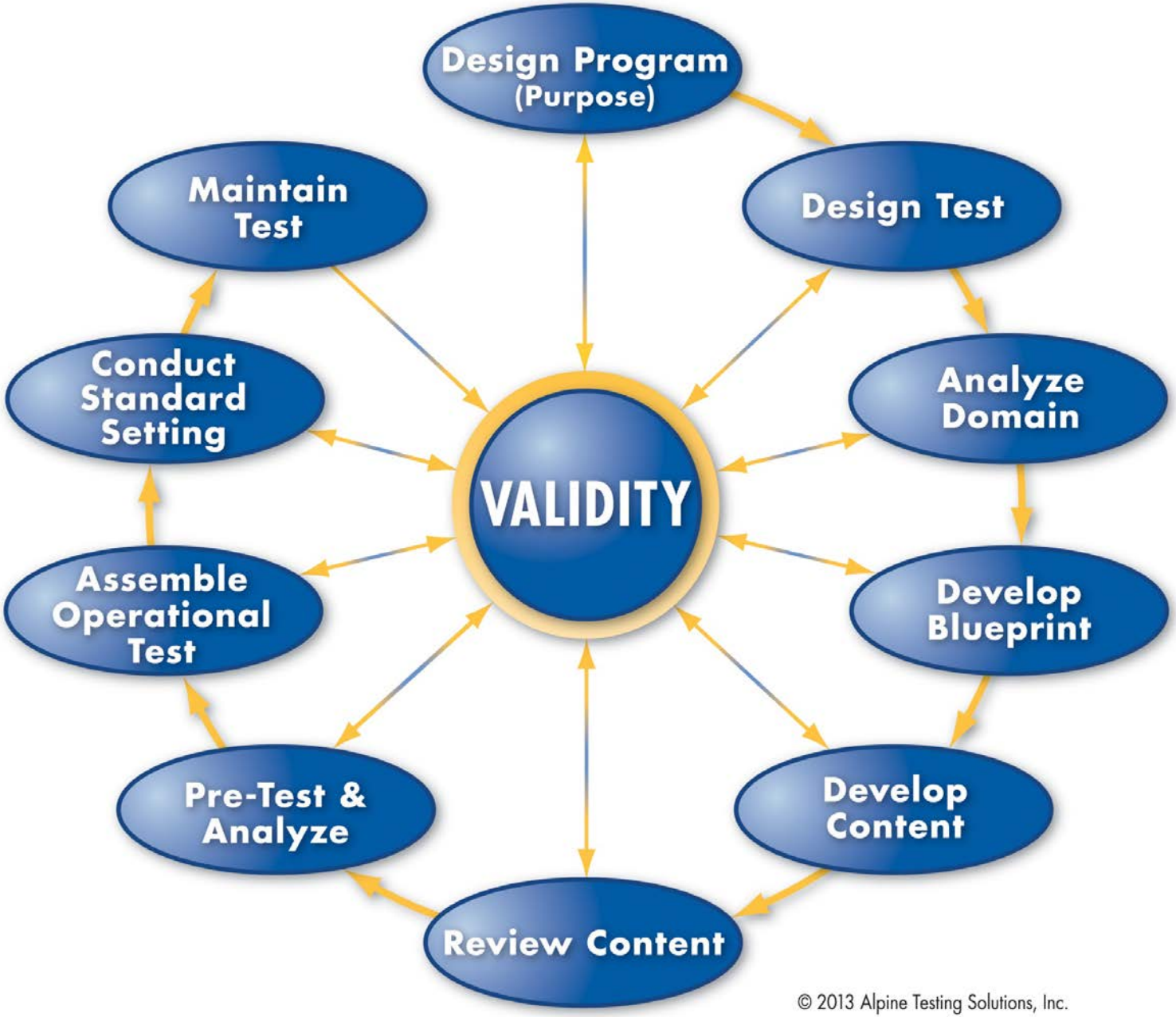
To provide…

- relevant examples of visual data displays designed to facilitate the test development processes
- Design principles for effective displays

Participants in the development process

- Organization administrators
- Subject matter experts
- Test users/consumers

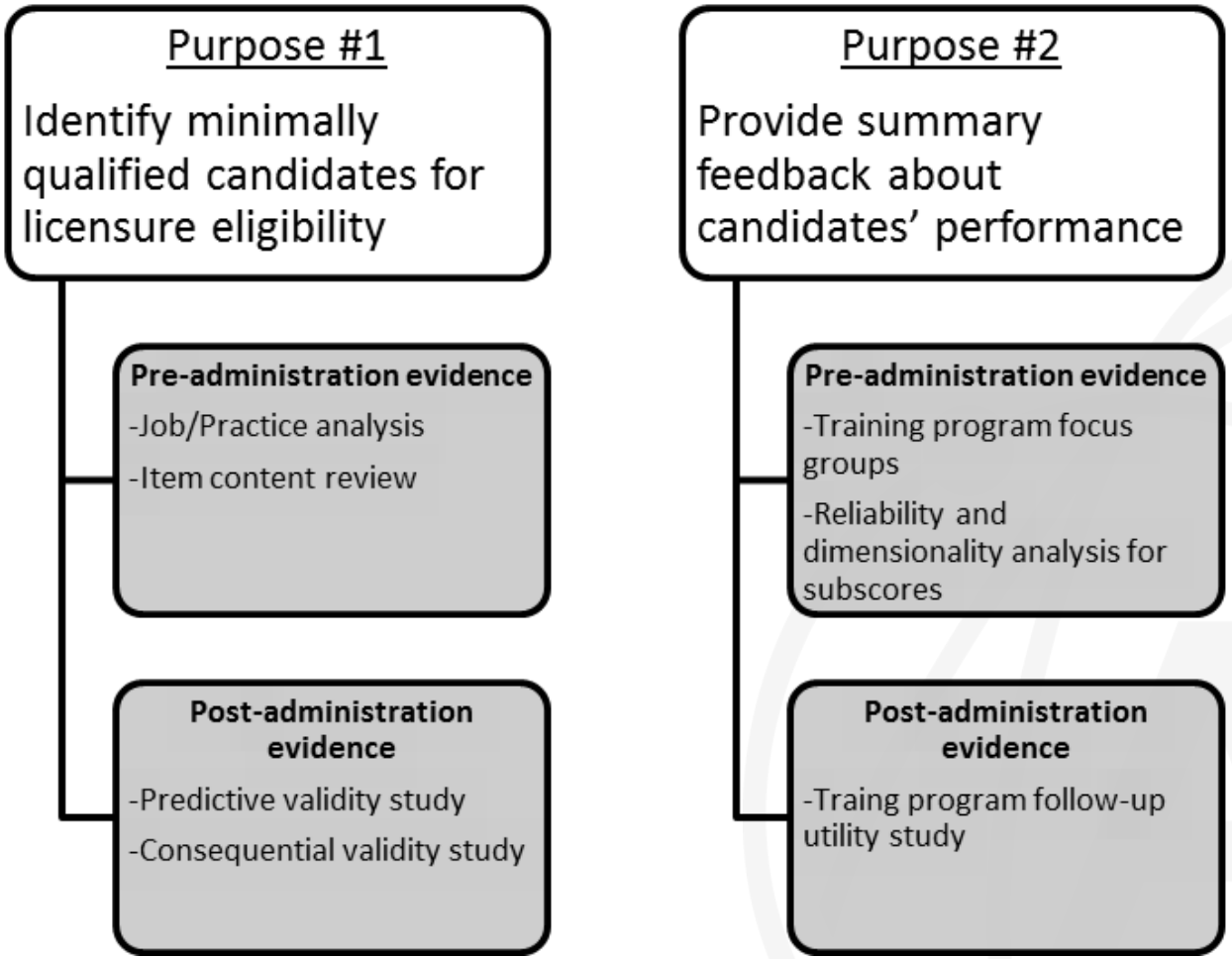
- *Needed for test development*
- *May have little or no psychometric background*



Design Program

- What is the purpose of the test?
- What question(s) is the test designed to answer?
- Who will take the test?

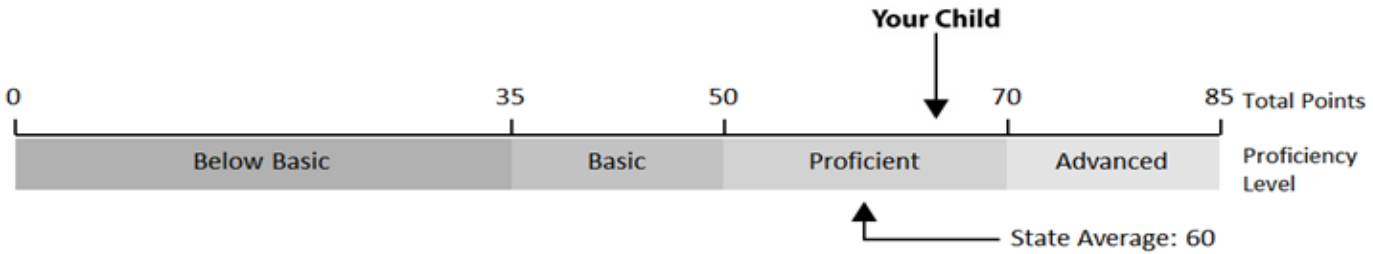
Summarizing two purposes within a validity framework



Excerpt from a sample score report

Jane Doe
Your Child's Total Score is 65
Overall Proficiency Level is Proficient*

A student's test score can vary if the test were taken several times. If your child were tested again, it is likely that your child would receive a score between **62** and **68**.



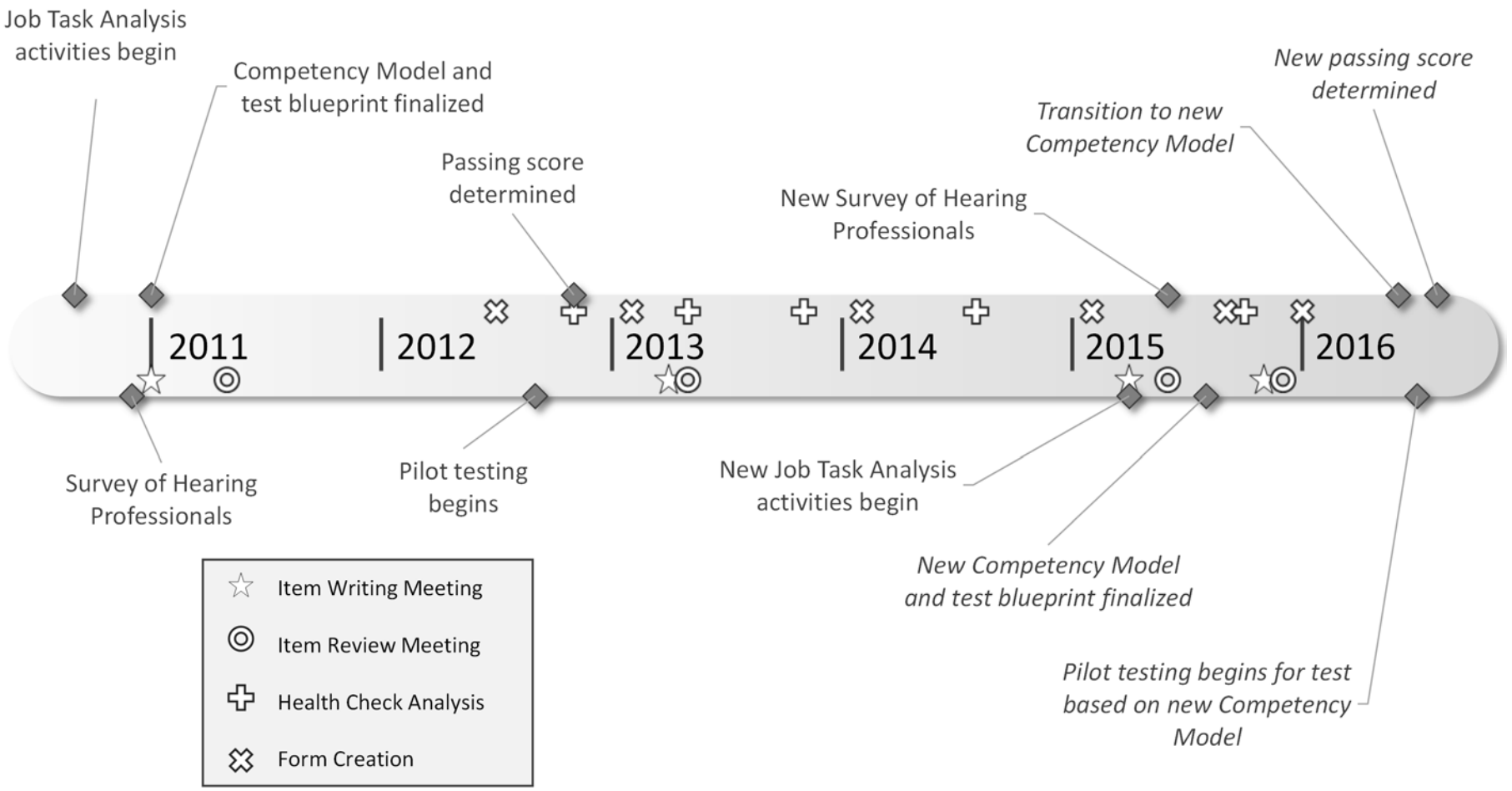
Mathematics Scores by Strand

Strand*	Possible Points	Your Child's Score**	Average Score*** School District	Average Score*** State
Number Sense	25	19	18	20
Measurement	10	7	6	7
Geometry	15	10	12	10
Algebra	20	18	15	13
Statistics and Probability	15	11	11	10

* *Please refer to page X for some of the knowledge and skills a student at the Proficient level has learned in Math and for definitions of the mathematics strands.*

** *These strand scores are added together to determine your child's total score.*

*** *Due to rounding, the sum of the strand score averages may not equal the average total score in the above graph.*

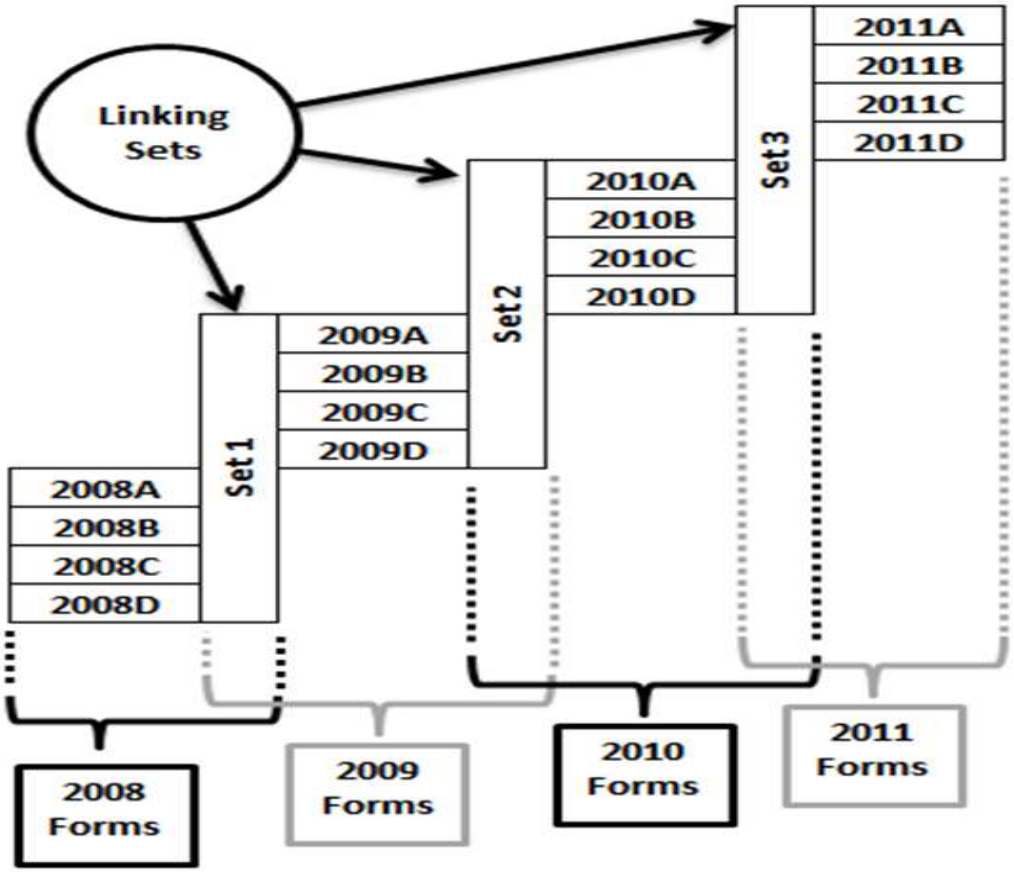


Note: Events in *italics* are in-progress/forthcoming

Design Test

- What format should the test be?
- How will the test be administered?
- How many test forms are necessary?

Diagram illustrating an equating design with linking sets of items



Analyze Domain

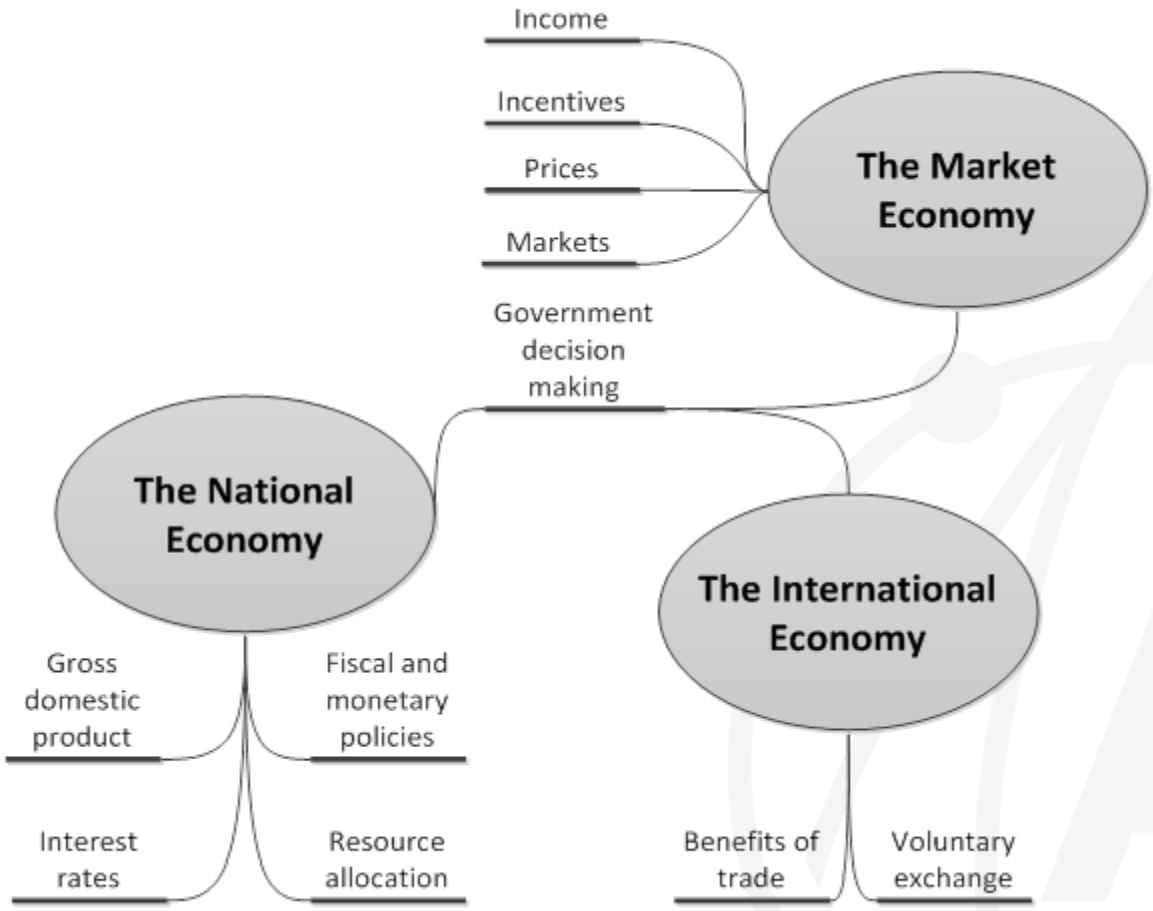
- What content should appear on the test?

- What knowledge/skills/abilities should the test address and at what level of cognitive complexity?

Develop Blueprint

- How many test items should be devoted to each content area?
- What item format is most appropriate for each content area?
- How many items should be developed for each cognitive complexity level?

Cluster diagrams representing the content domain of grade 12 economics



Three-dimensional test blueprint (objective, cognitive complexity, item type)

	Item type					
	Multiple choice			Constructed response		
	Cognitive complexity level			Cognitive complexity level		
Objective	1	2	3	1	2	3
1	8	1	1			
2	2	7	1			1
3	1	6	3			1
4	8	1	1			
5	2	3	5			2
6	5	2	3			
7	5	2	3			
8	2	8	0			
9	1	2	7			1
10	8	1	1			

Develop Content

- How will quality items be created?
- Who will create them?
- How many items should be written?

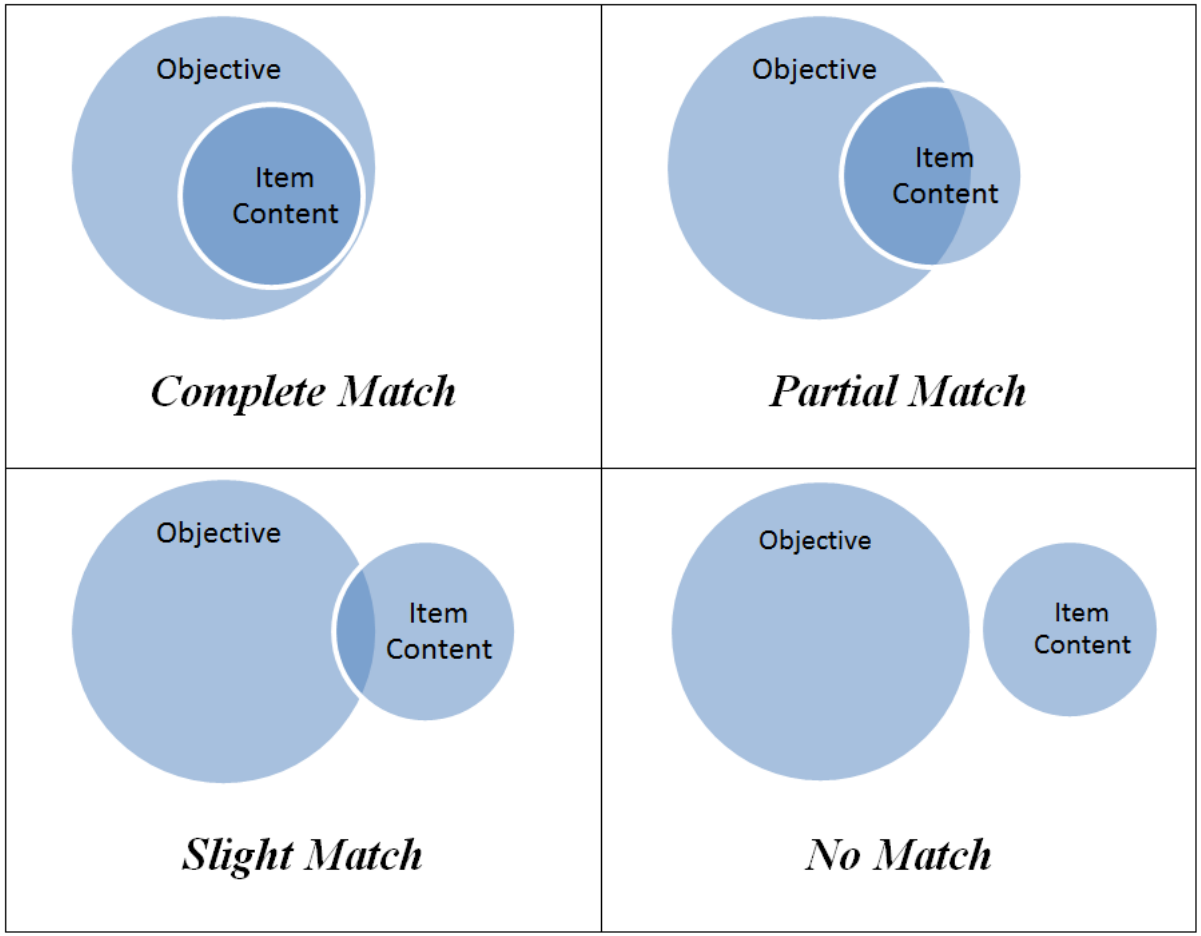
Conditional formatting for illustrating real-time item development progress

Section	Objective	Needed	Written	Progress
Section 1	1.1	9	4	
	1.2	7	0	
	1.3	7	5	
Section 2	2.1	7	1	
	2.2	6	1	
	2.3	7	0	
Section 3	3.1	10	10	
	3.2	7	1	
	3.3	10	10	
Section4	4.1	9	5	
	4.2	7	6	
	4.3	8	4	
	Total:	94	47	

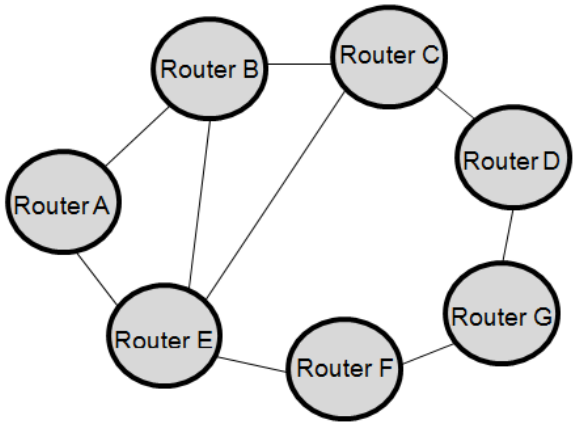
Review Content

- What steps should be taken to make sure items are not biased against certain subgroups?
- How can we ensure the items are technically accurate?
- How can we ensure that items meet published criteria for quality?
- Are there sufficient items to create the forms required by our test blueprint and program design?

Venn Diagrams to illustrate degrees of relationship between the content of a test item and the content of an objective



Graphic inclusion review as part of the item review process

Item without diagram	*Item with diagram*
Router A connects to Routers B and E. Router B connects to Routers A and E. Router C connects to Routers B and D. Router D Routers C and G. Router G connects to Routers D and F. Router F connects to Routers G and E. Router E connects to Routers A, B, C, and F. Rip is the only routing protocol configured on the routers. What is the shortest path from Router A to Router G? a. Through Routers B and C b. Through Routers E and F c. Through Routers B and E d. Through Routers E and C	Refer to the Exhibit.  What is the shortest path from Router A to Router G? a. Through Routers B and C. b. Through Routers E and F. c. Through Routers B and E. d. Through Routers E and C.

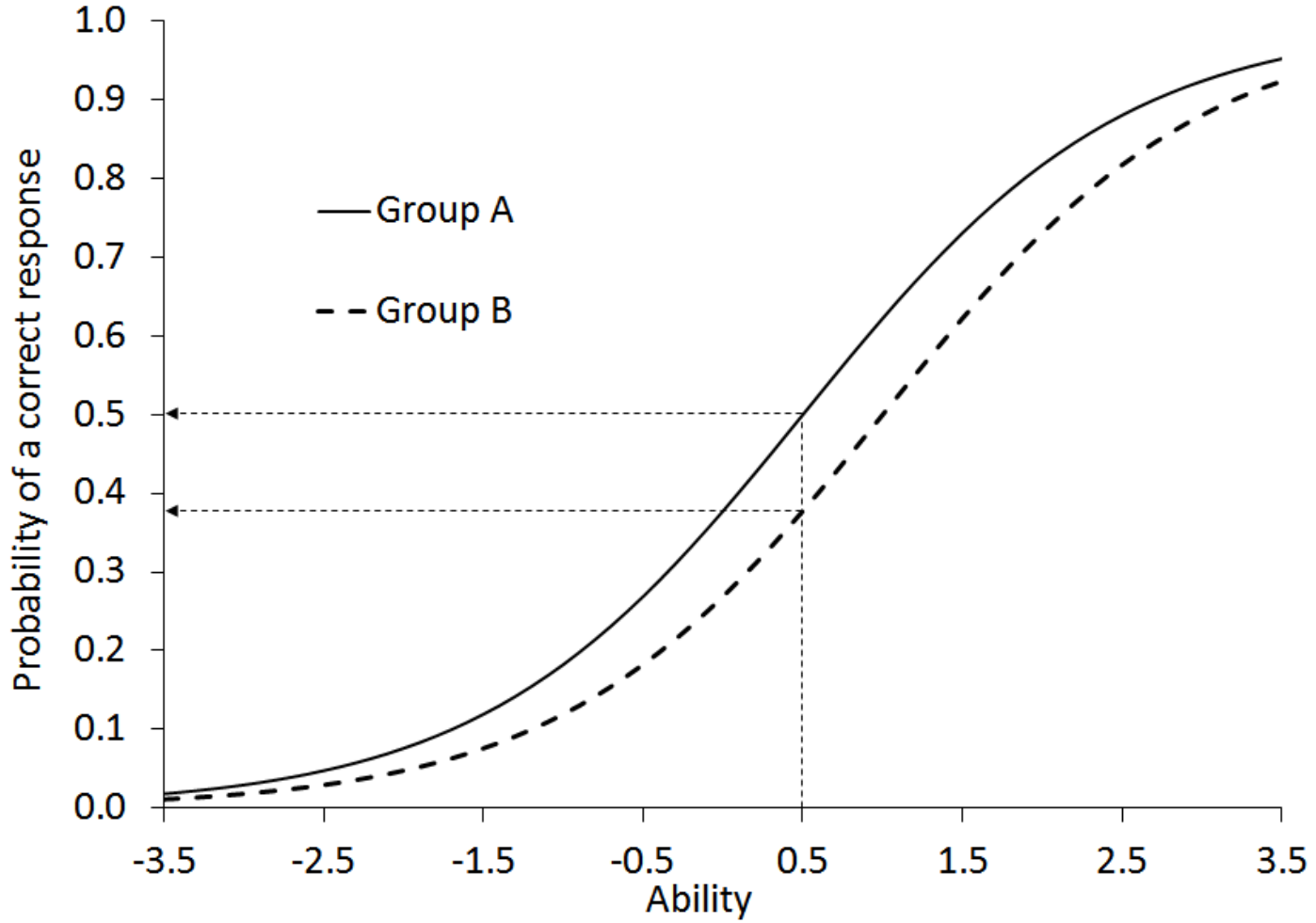
Pre-test and Analyze

- Why is it important to field test items before administering them operationally?
- What kind of analyses should be performed?
- How large of a field test sample should we have?

Conditional formatting and spreadsheet-embedded sparklines for evaluation of item level pretest data

Item	Number of responses at scorepoint...				Percent of responses at scorepoint...				Distribution of scored responses	Average score
	0	*1*	*2*	*3*	*0*	*1*	*2*	*3*		
1	12	46	33	59	8%	31%	22%	39%		2.9
2	71	38	35	6	47%	25%	23%	4%		1.3
3	0	59	6	85	0%	39%	4%	57%		3.3
4	18	6	67	59	12%	4%	45%	39%		3.2
5	1	41	23	85	1%	27%	15%	57%		3.4
6	52	12	54	32	35%	8%	36%	21%		2.2
7	45	57	15	33	30%	38%	10%	22%		1.9
8	25	48	45	32	17%	32%	30%	21%		2.3
9	3	49	49	49	2%	33%	33%	33%		2.9
10	44	1	77	28	29%	1%	51%	19%		2.4

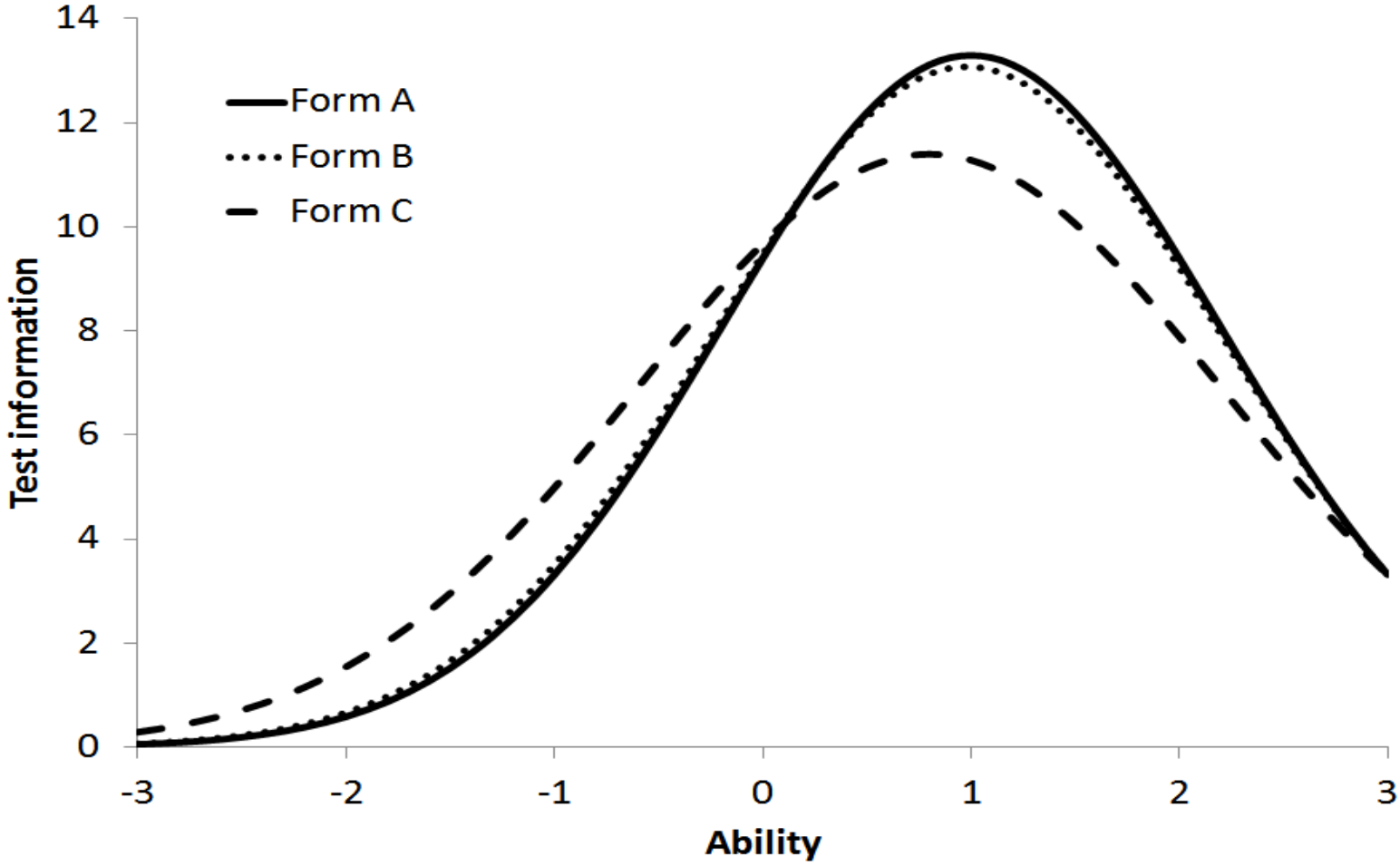
Using item characteristic curves to show the effect of DIF



Assemble Operational Test

- How is pre-test data used to assemble operational tests?
- How do we make sure results from different forms are comparable?

Example of using test information functions to evaluate form comparability

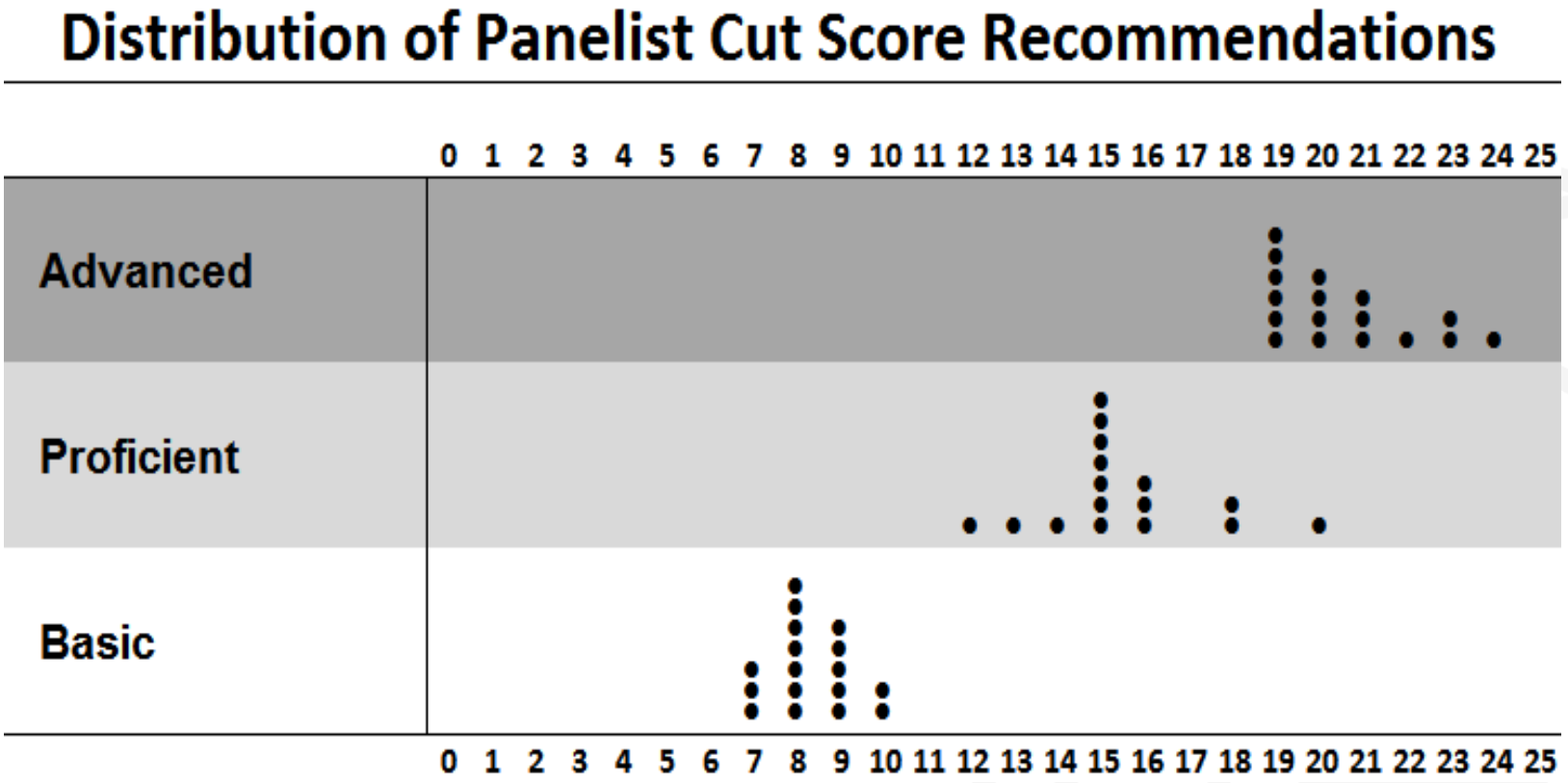


Conduct Standard Setting

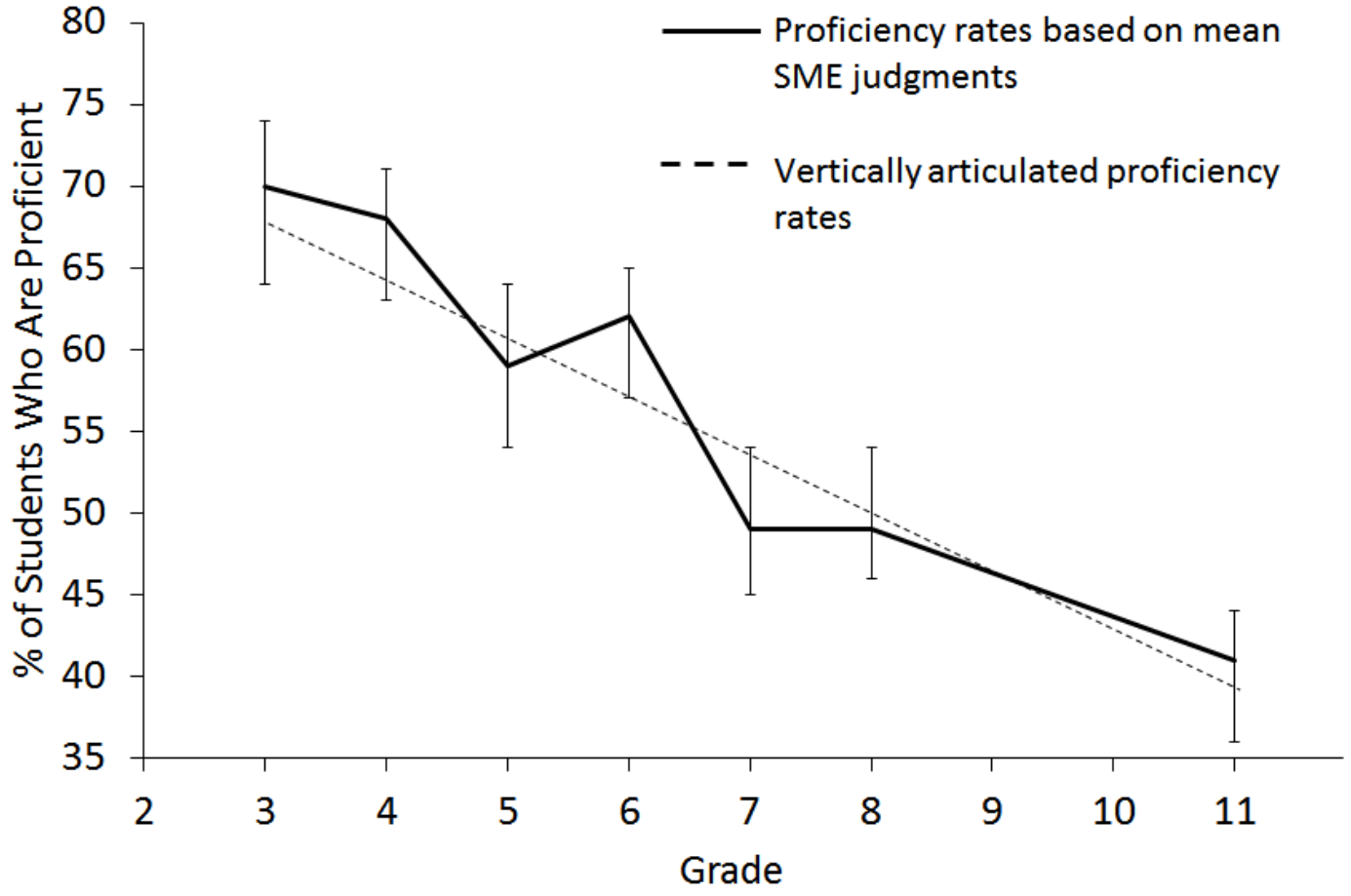
- If criterion-referenced decisions are going to be made based on the assessment results, how should we determine the cut-point(s) on the test score scale?

- What about norm-referenced tests?

Example of stacked dot plots for standard setting panelist feedback



Example of a line graph with confidence intervals for vertical articulation in standard setting

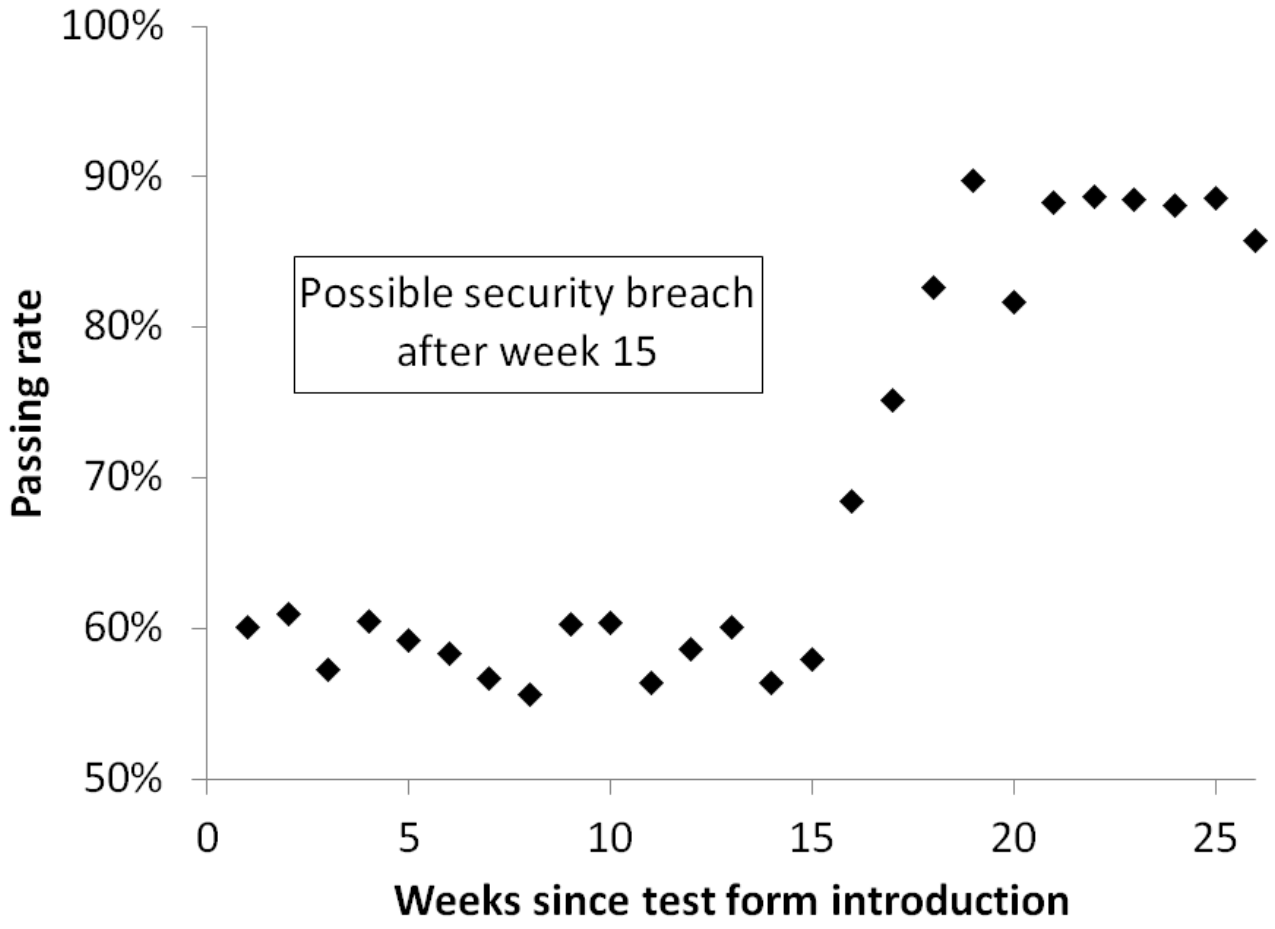


Maintain Test

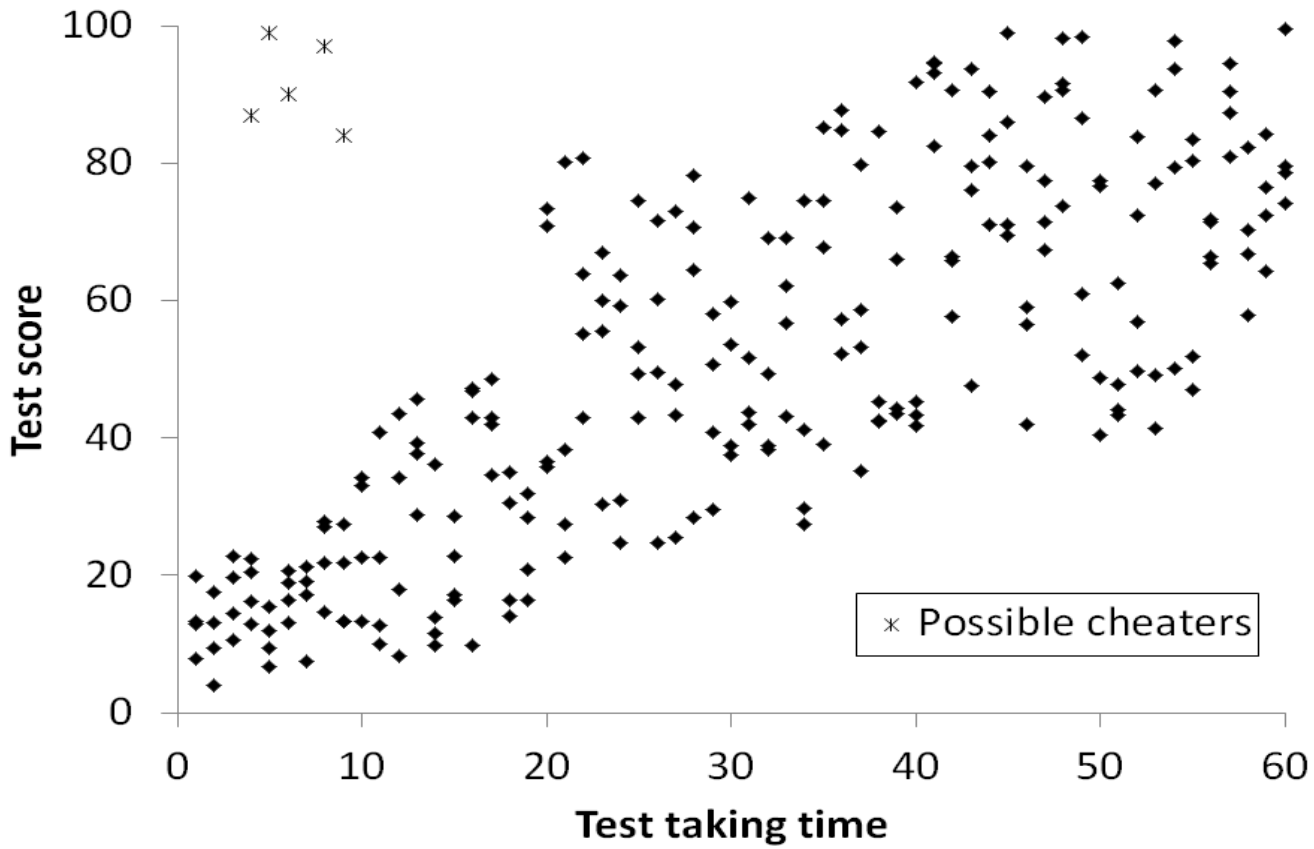
- Once a test is created, what are the test creators' ongoing responsibilities?

- What additional validity studies are necessary to support test use?

Scatter plot showing passing rates over time



Example of a scatter plot comparing test score for different test-taking times



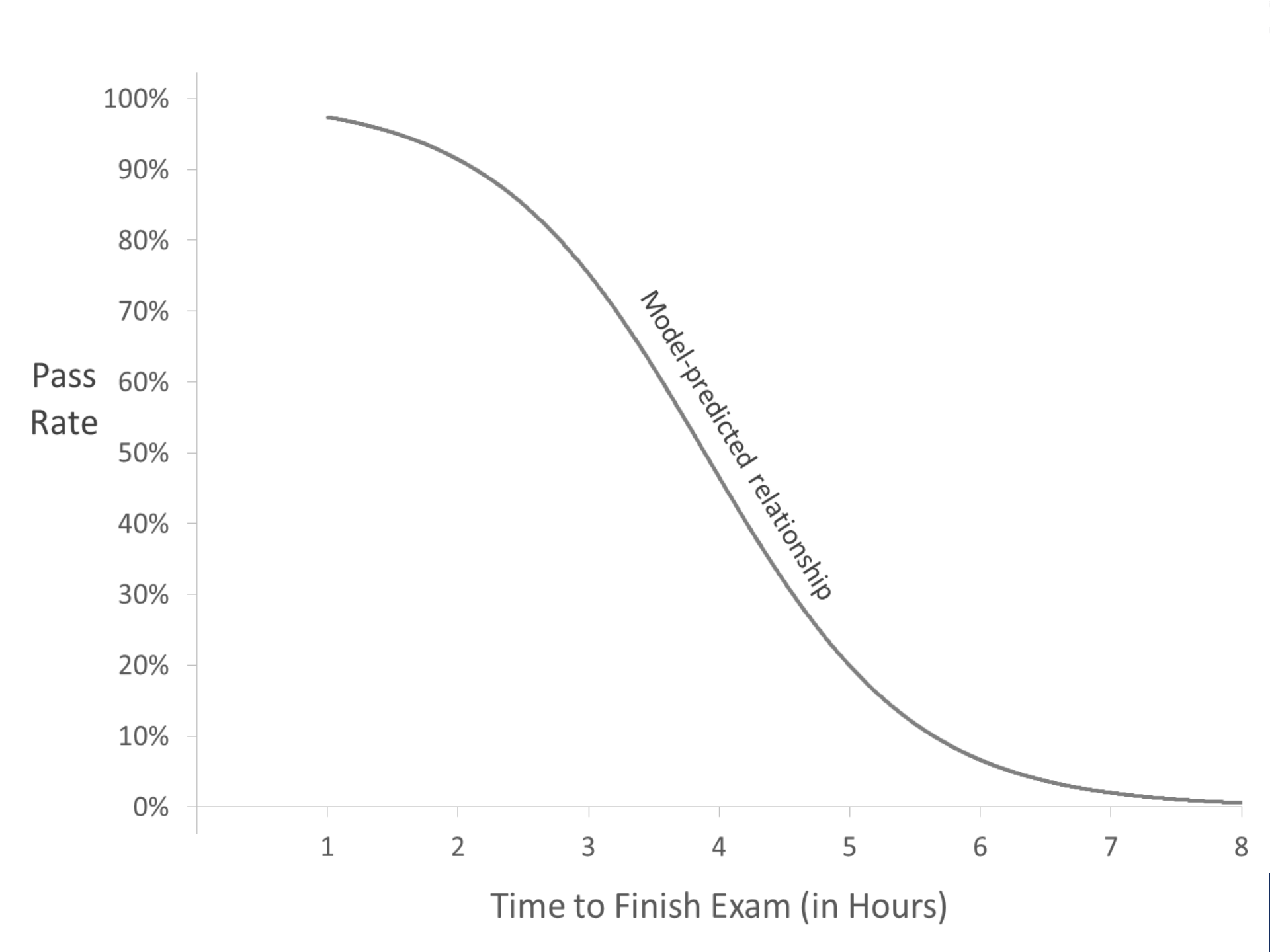
Characteristics of Effective Visual Displays

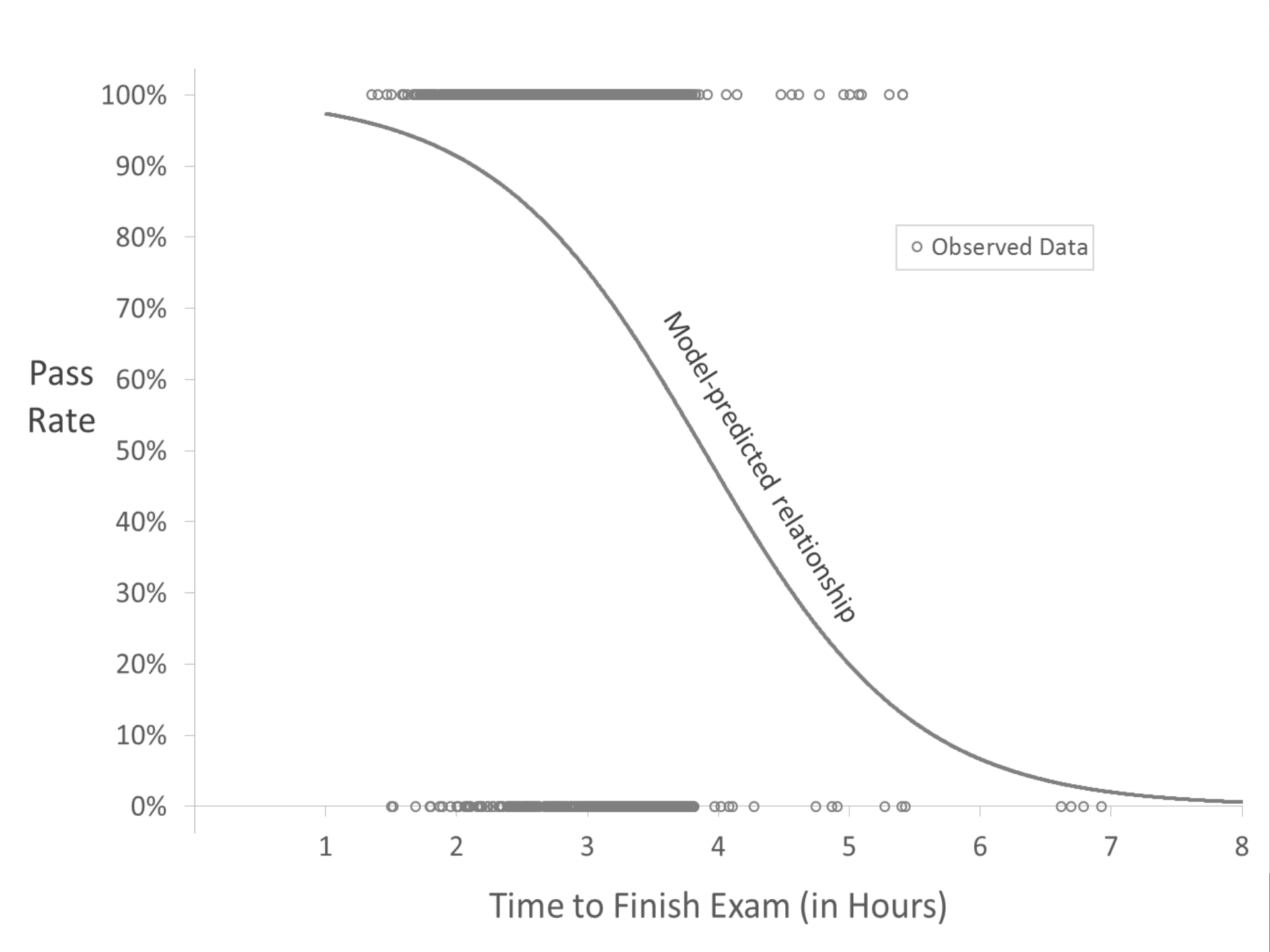
- Choose displays that are appropriate for the data
- Aim for simplicity
- Integrate text, numbers, and figures
- Highlight what is important
- Do not intentionally mislead
- Pair the design with the audience

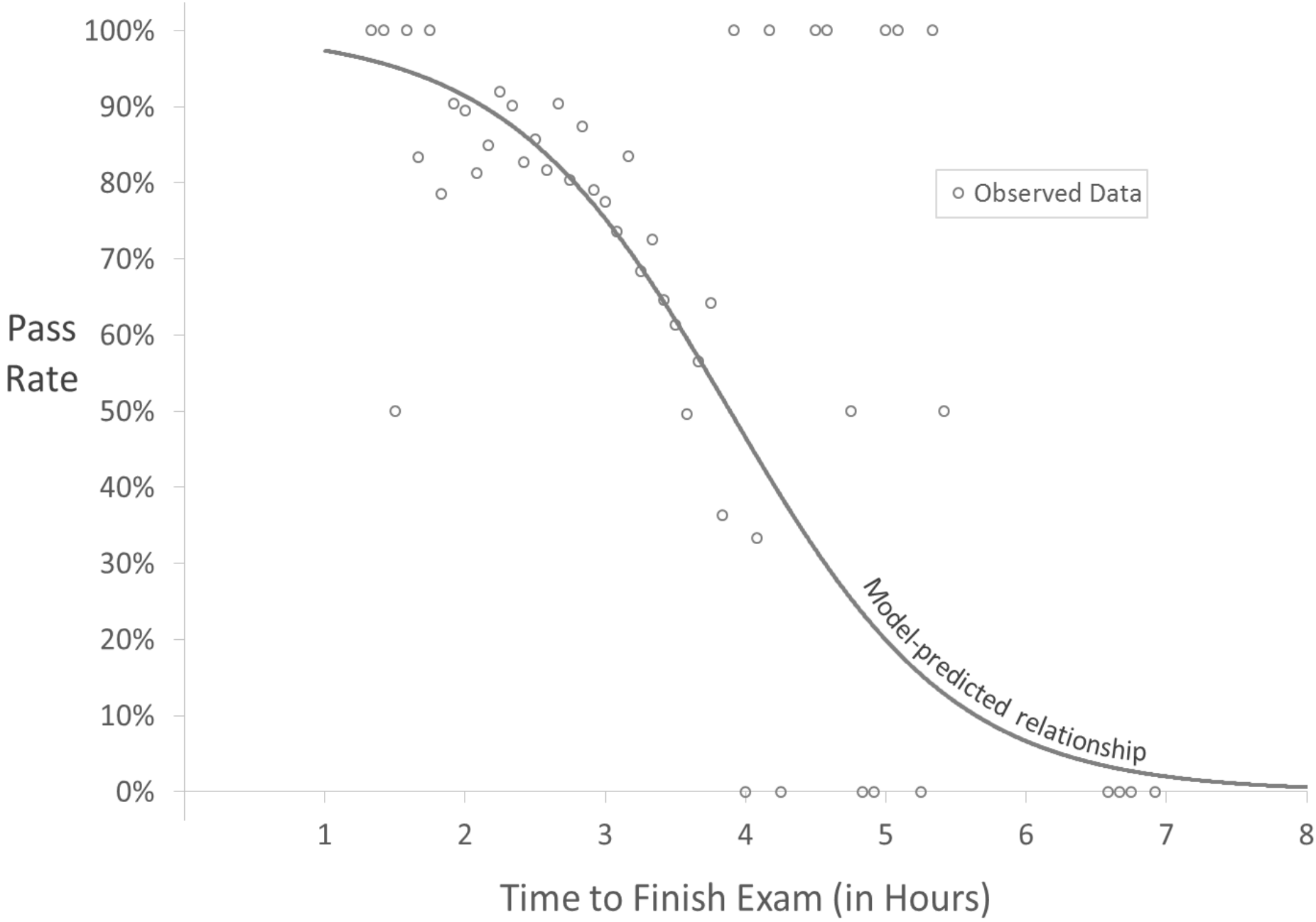
Choose displays that are appropriate for the data

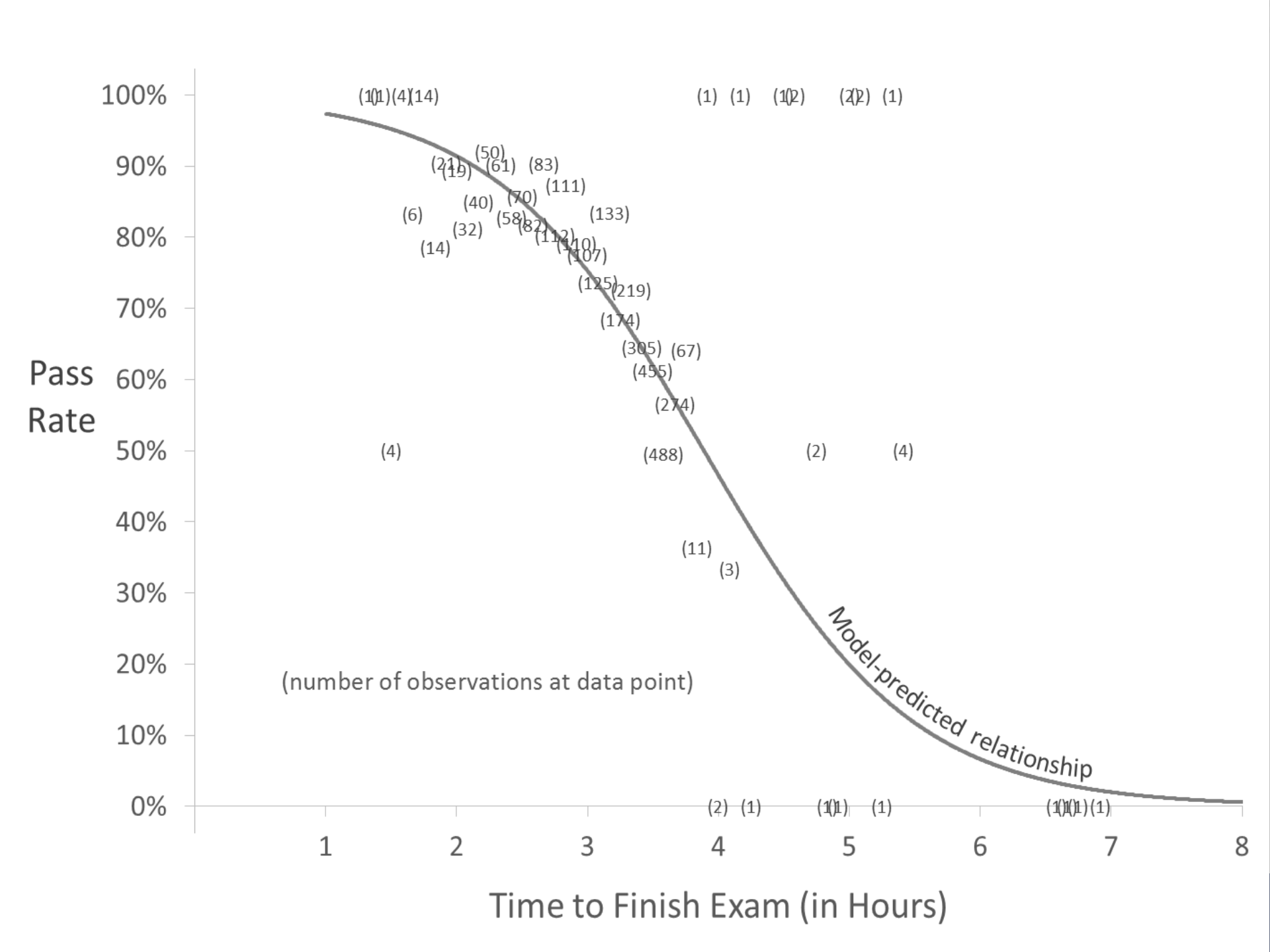
- Tables:
  - Present specific values
  - Useful when a significant amount of detail is necessary
- Graphs
  - Show trends and relationships

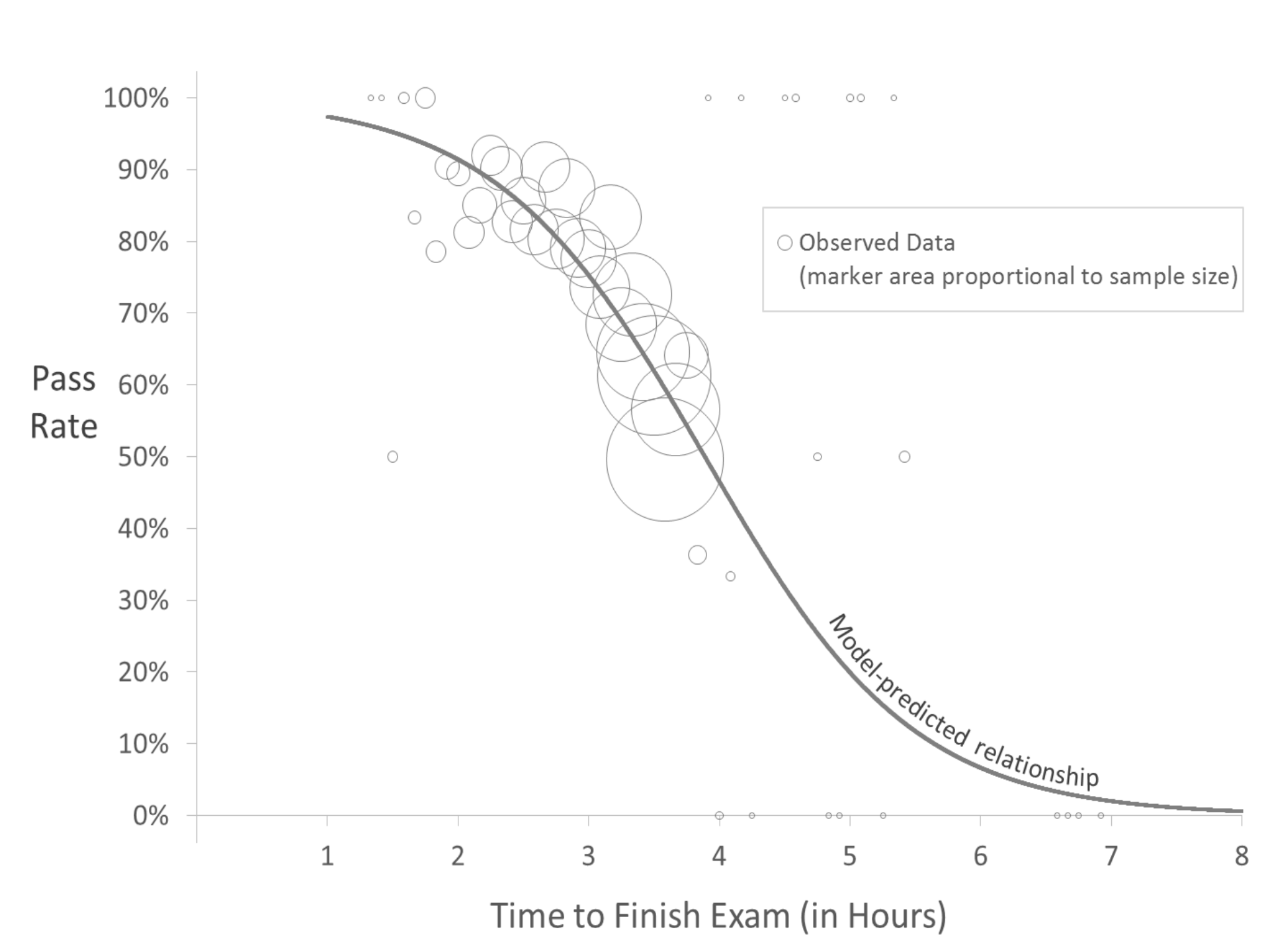
- Corollary:
  - Choose the appropriate data for the graph


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Pass Rate
Model-predicted relationship
1
2
3
4
5
6
7
8
Time to Finish Exam (in Hours)


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Pass Rate
Observed Data
Model-predicted relationship
1
2
3
4
5
6
7
8
Time to Finish Exam (in Hours)

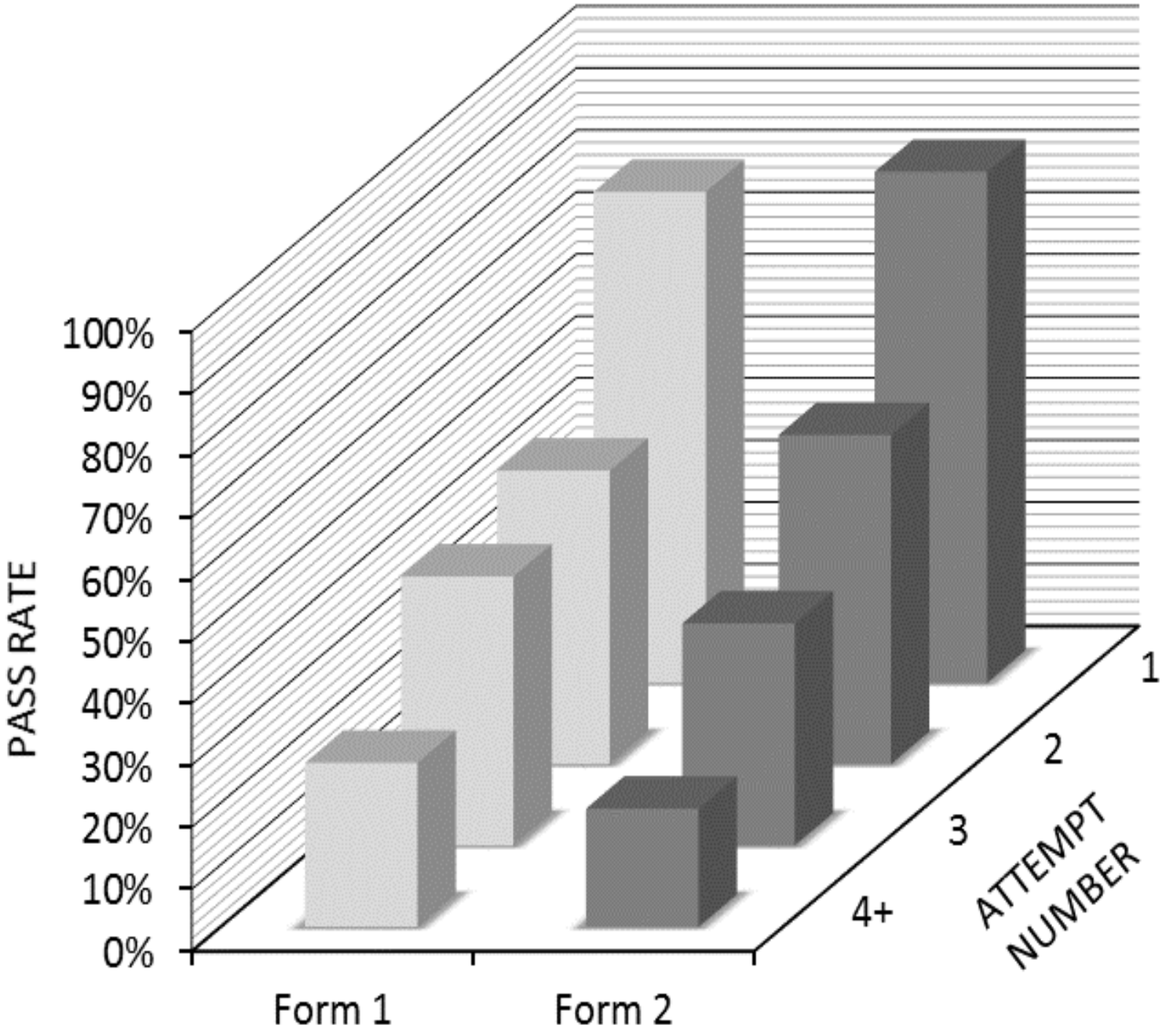

100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Pass Rate
1
2
3
4
5
6
7
8
Time to Finish Exam (in Hours)
Observed Data
Model-predicted relationship

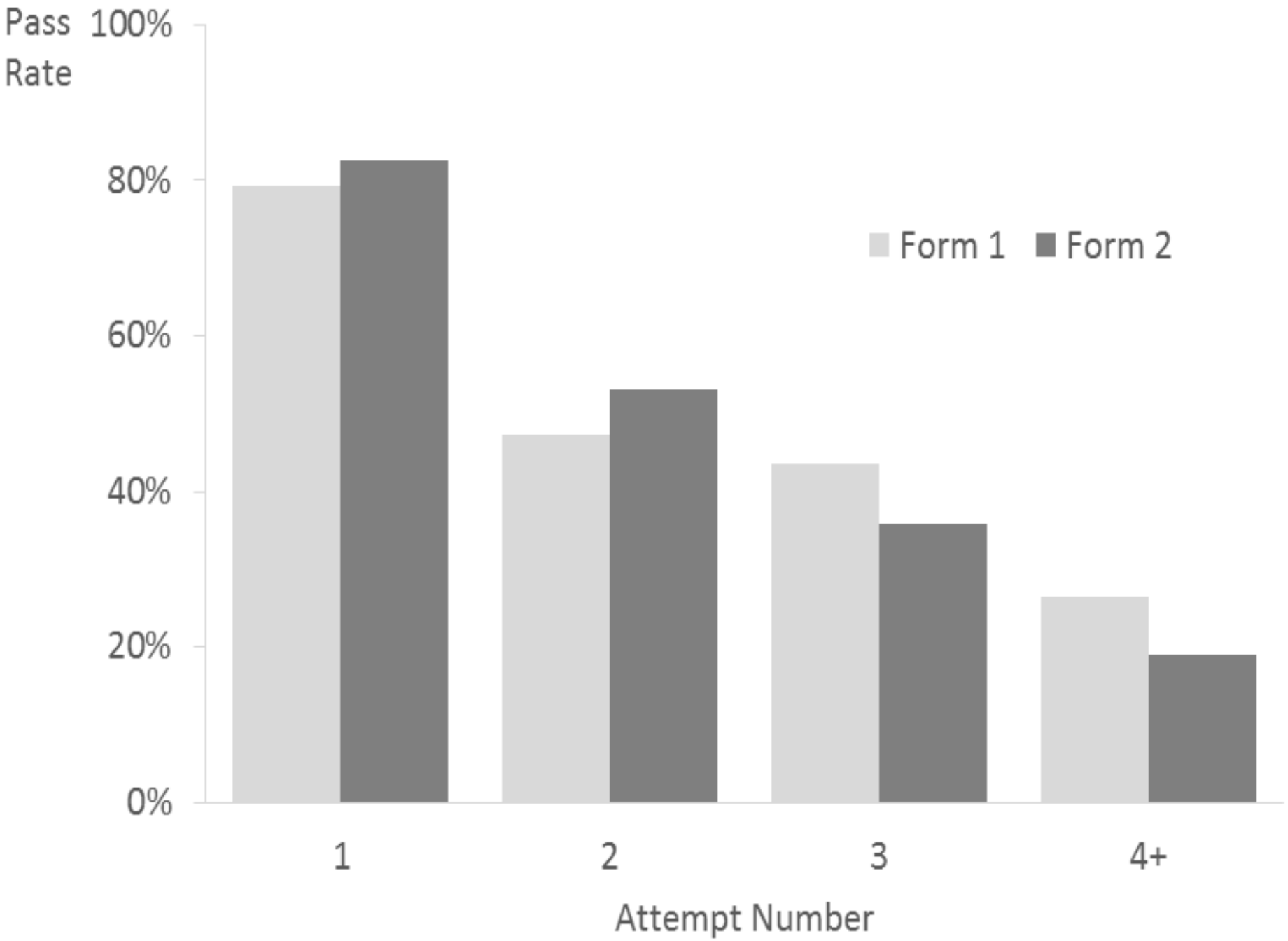

Pass Rate
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
1
2
3
4
5
6
7
8
Time to Finish Exam (in Hours)
(number of observations at data point)
Model-predicted relationship


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Pass Rate
1
2
3
4
5
6
7
8
Time to Finish Exam (in Hours)
Observed Data
(marker area proportional to sample size)
Model-predicted relationship

Aim for simplicity

- Should be simple, but complete.
- Graphical elements that do not directly convey information should be removed or minimized.
- Avoid the use of 3D rendering for 2D or 1D information.
- Limit the number and type of fonts and symbols.


PASS RATE
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Form 1
Form 2
1
2
3
4+
ATTEMPT NUMBER


Pass Rate
100%
80%
60%
40%
20%
0%
Form 1
Form 2
1
2
3
4+
Attempt Number

Integrate text, numbers, and figures

- Information necessary for interpretation should be proximate to the display.
- Place labels close to the elements they are describing, when possible
- Place relevant annotations near the data

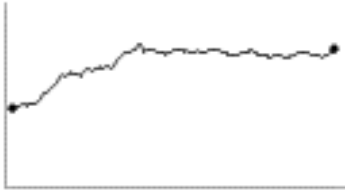
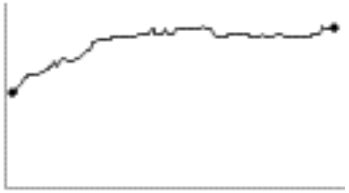
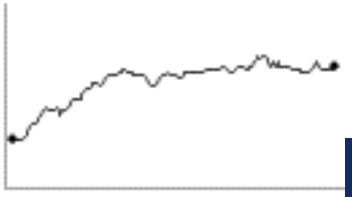
Pass rates by attempt for several countries. This table integrates graphs with numbers through the addition of sparklines.

County	Attempt 1	2	3	4+	Trend
All Candidates	**81%**	50%	40%	23%	
Country*					
USA	**88%**	63%	48%	25%	
India	37%	37%	**38%**	24%	
Philippines	**37%**	36%	28%	17%	
Other	**42%**	25%	27%	27%	

Note: Highest pass rate for each country is shown in **bold**.

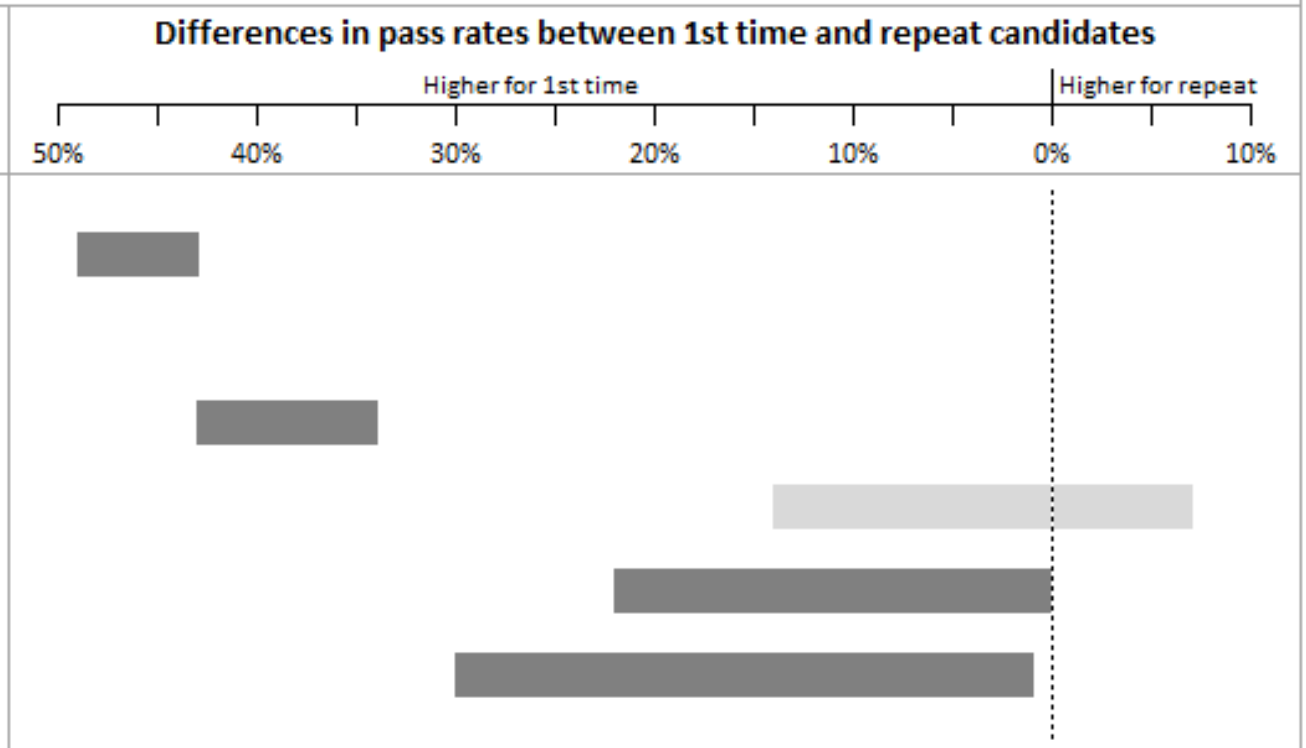
*Country where candidate was educated

Item performance for a subset of items across the first 200 administrations of Form 1. This table integrates graphs with numbers through the addition of sparklines to show how item p-values change with the number of test adminstrations.

	P-value			
Item	After 50 admins	After 200 admins	Change	Moving Average
7	0.42	0.80	0.38 ▲	
12	0.42	0.36	-0.06 ▼	
19	0.52	0.92	0.40 ▲	
38	0.24	0.70	0.46 ▲	

This figure shows comparisons of pass rates for first time and repeat test takers for several different countries The graphic on the right illustrates ranges that, with 95% certainty, capture the true difference in the pass rates between the two types of candidates within each country. If the range is completely on the "Higher for 1st time" or "Higher for repeat" side of the dashed line, the difference between the two pass rates is significant. Therefore, it is correct to say that the candidates of one type performed better or worse than the other type on the assessment. However, if the range crosses the dashed line (representing no difference), the pass rates for first time and repeat test takers are not significantly different. There was no significant difference between first time and repeat test takers who were educated in India (light gray shading).

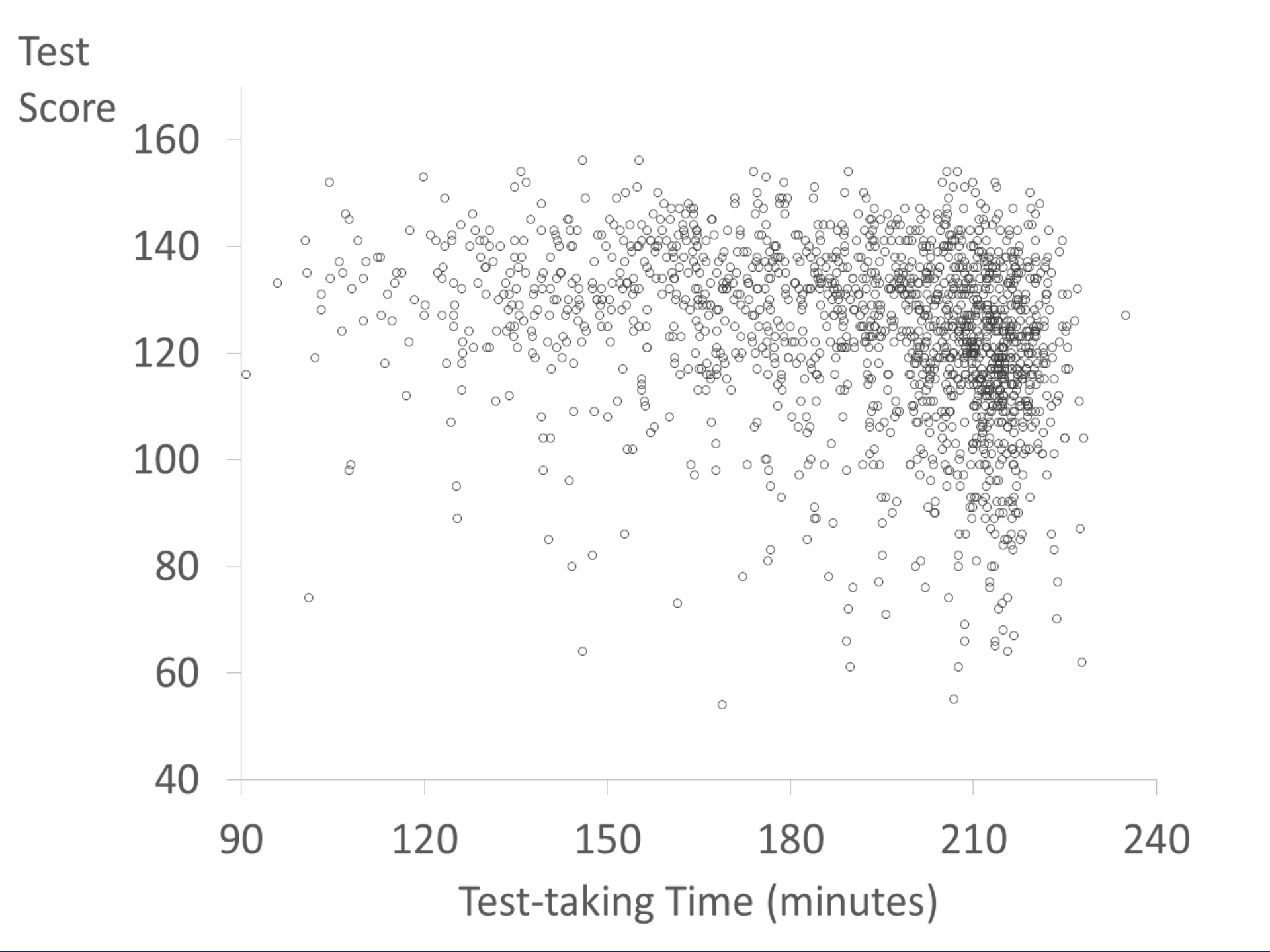
	Pass Rate	
	1st time	Repeat
All Candidates	81%	39%
Country*		
USA	88%	49%
India	37%	33%
Philippines	37%	25%
Other	42%	26%

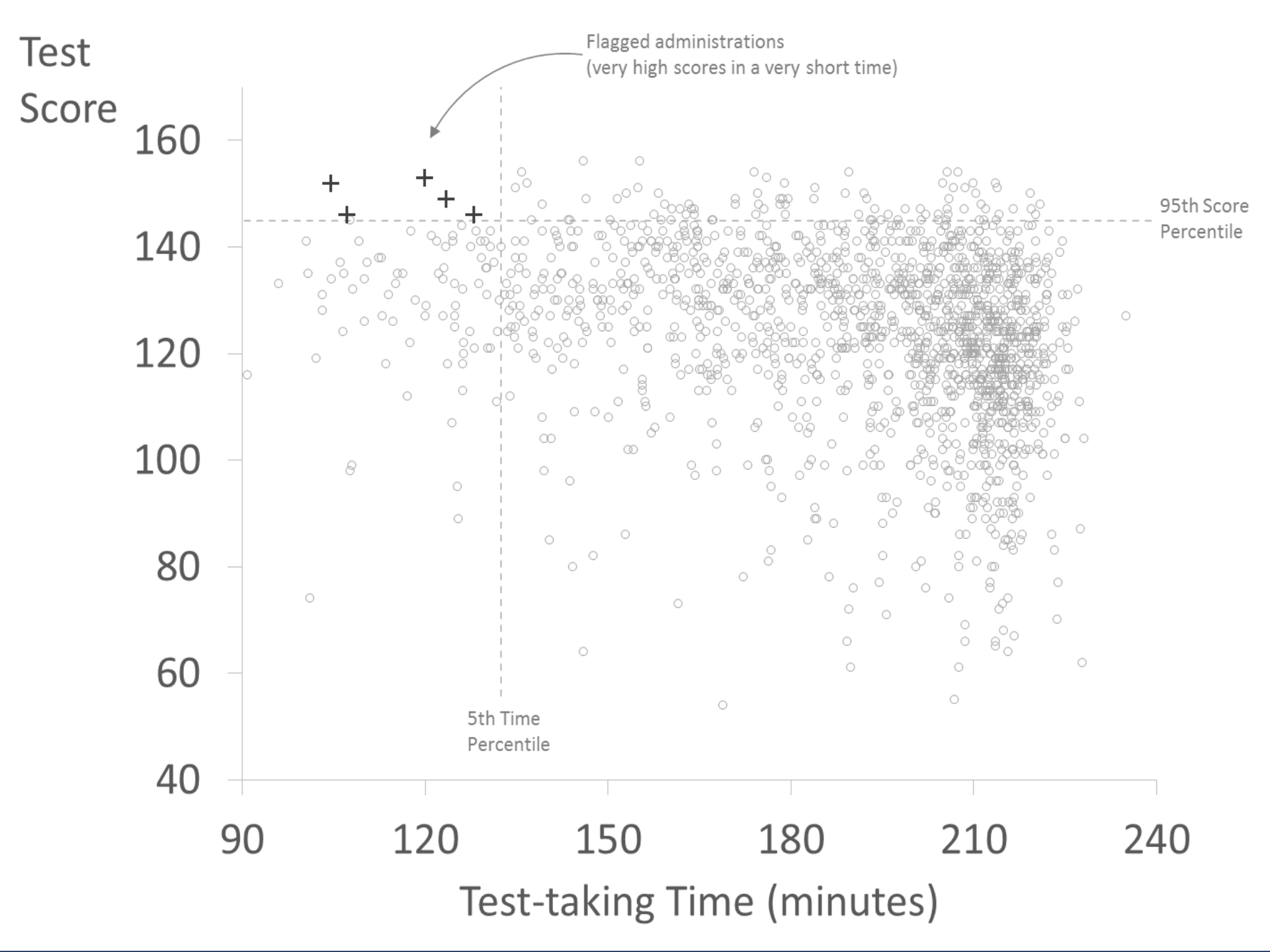


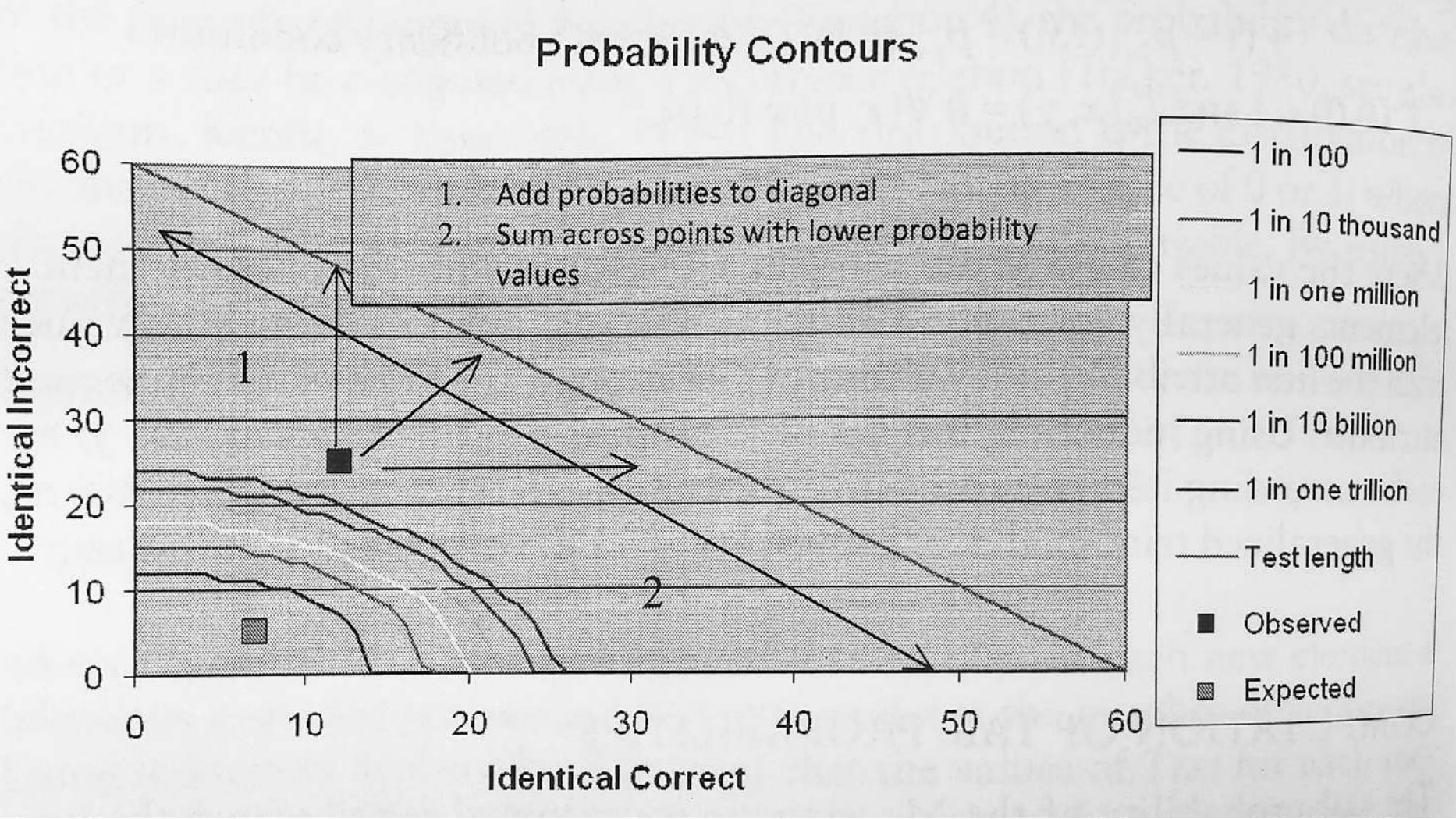
*Country where candidate was educated

Highlight What Is Important

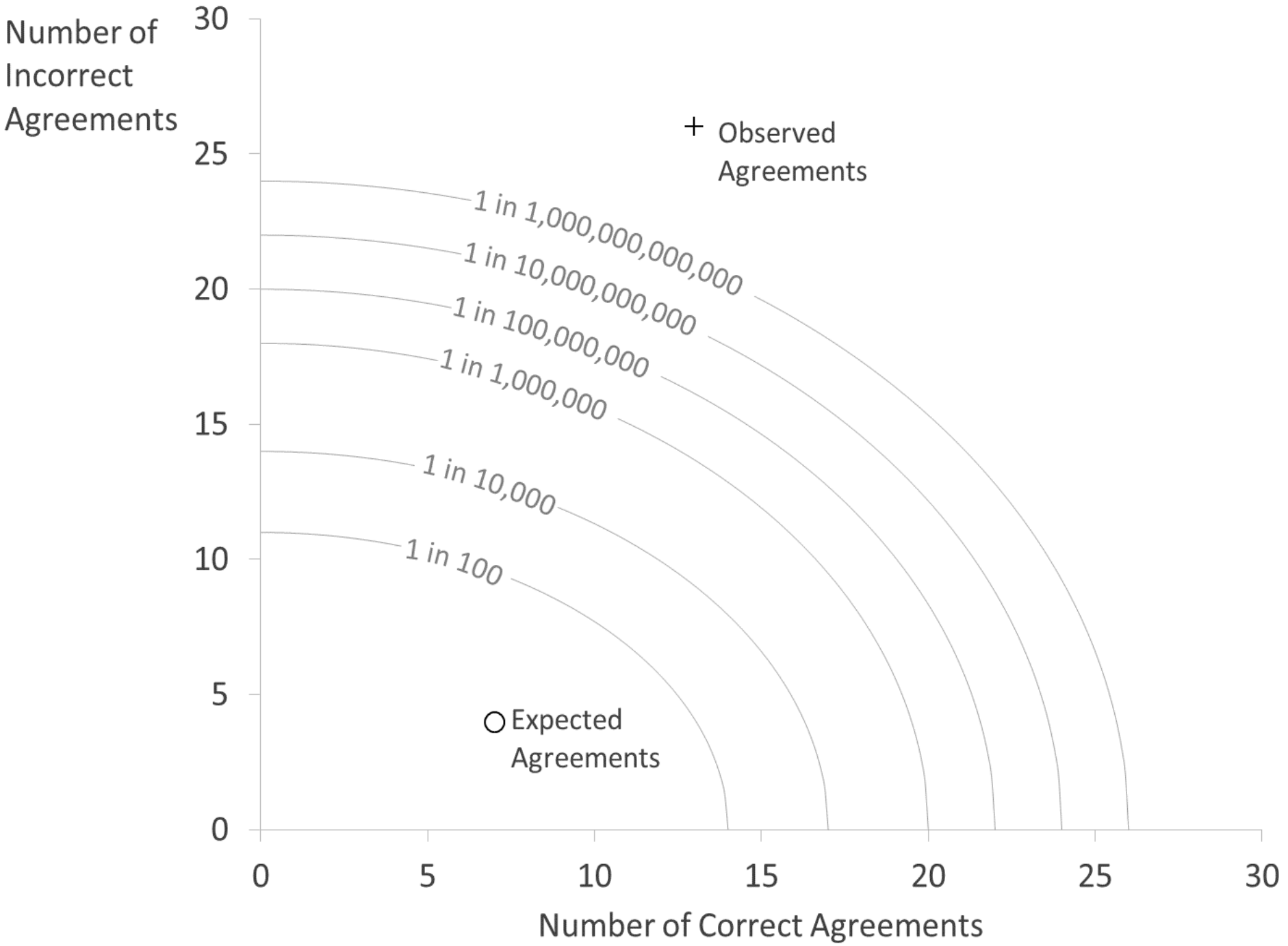
- Non-data elements (e.g., gridlines, scales) should support interpretations, but should not be obtrusive.
- Intense colors should be only be used to highlight important features.
- Use annotations to identify important data features.
- Items that are intended to be compared should be placed close to one another.


Test Score
160
140
120
100
80
60
40
90
120
150
180
210
240
Test-taking Time (minutes)


Test Score
Flagged administrations
(very high scores in a very short time)
160
140
120
100
80
60
40
95th Score Percentile
5th Time Percentile
90
120
150
180
210
240
Test-taking Time (minutes)

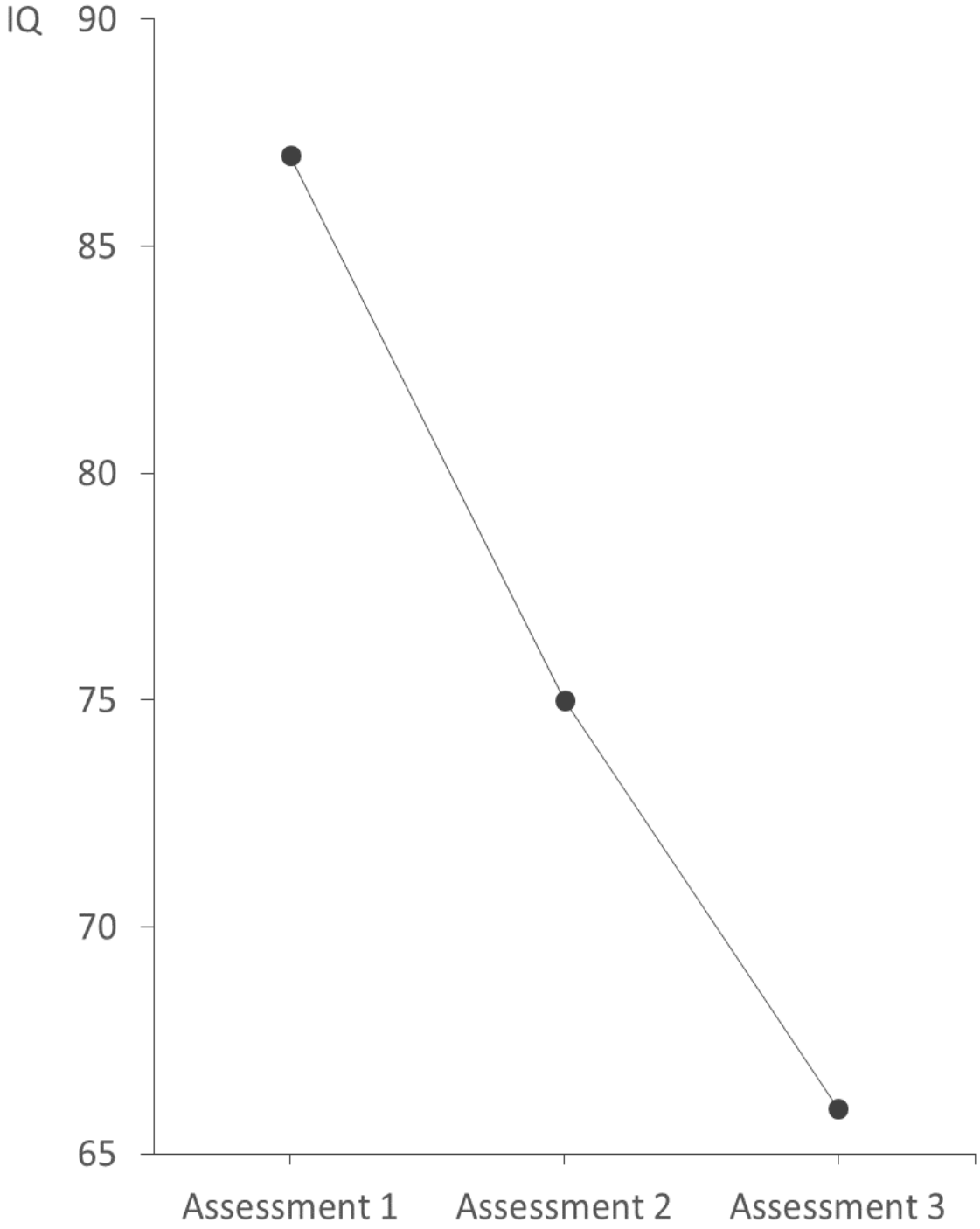


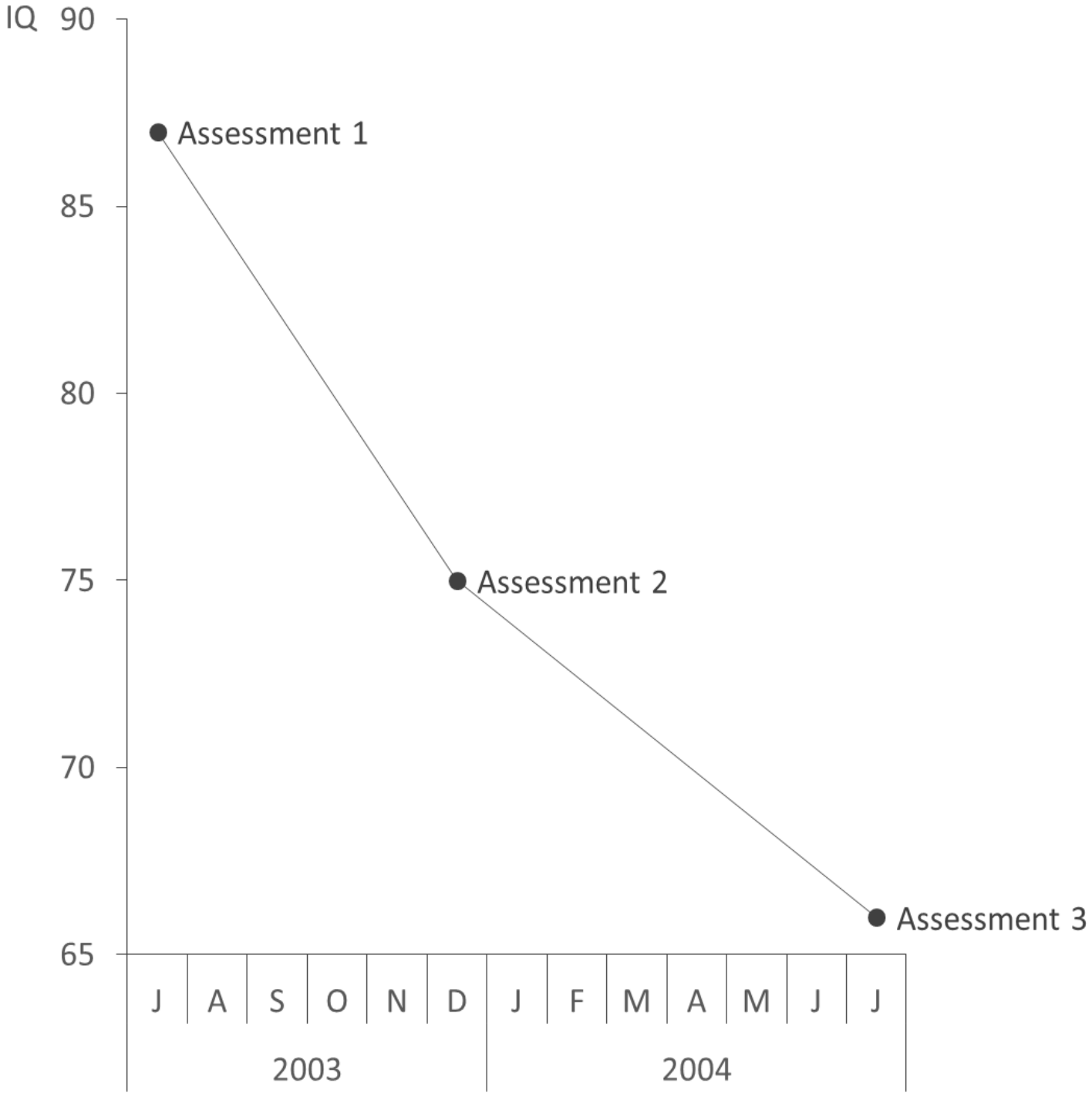
Maynes, D. (2014). Detection of non-independent test taking by similarity analysis. In N. M. Kingston & A. K. Clark (Eds.), *Test fraud: Statistical detection and methodology* (pp. 53-80). New York, NY: Routledge.

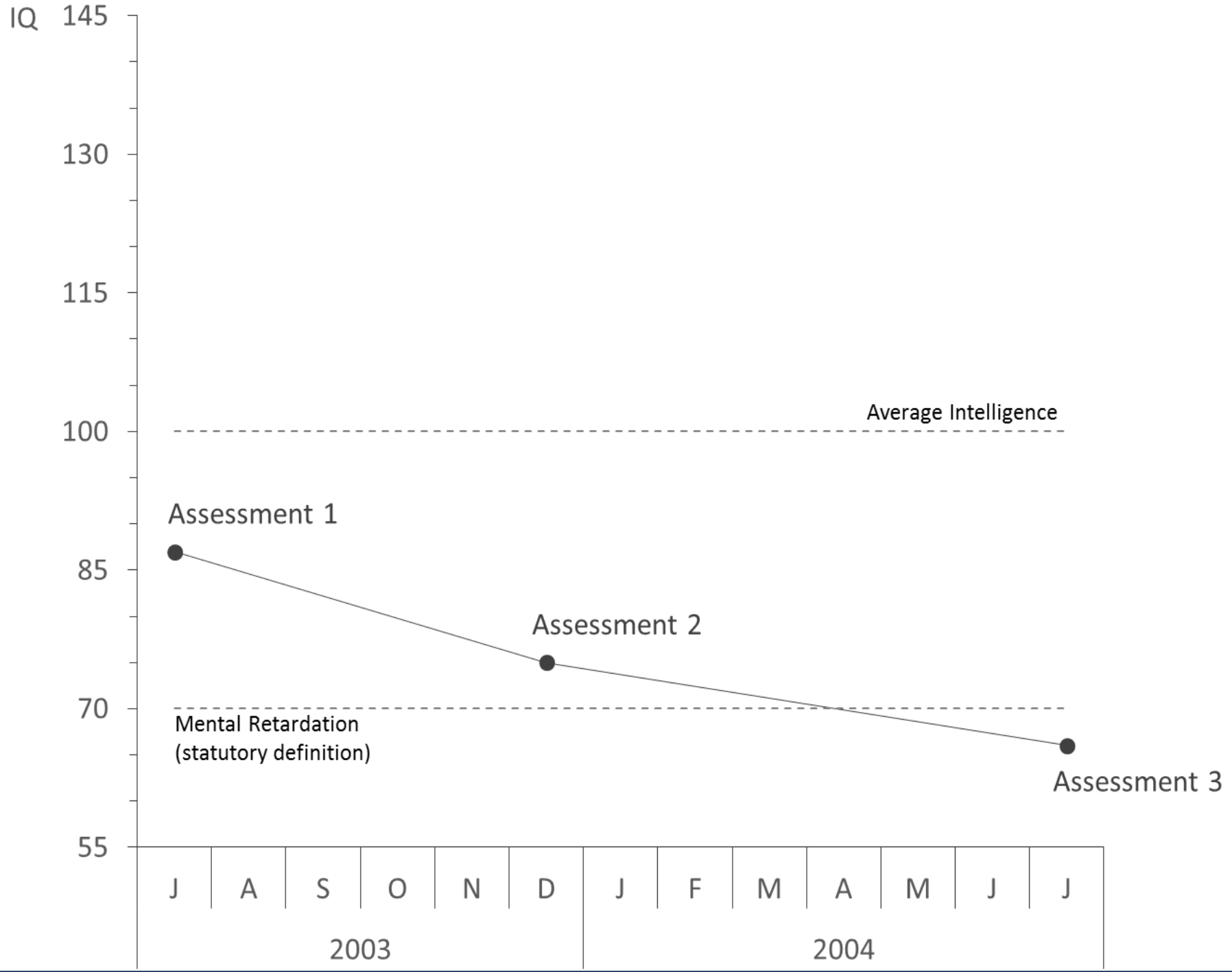

Number of Incorrect Agreements
30
25
20
15
10
5
0
Observed Agreements
1 in 1,000,000,000
1 in 10,000,000
1 in 100,000,000
1 in 1,000,000
1 in 10,000
1 in 100
Expected Agreements
0
5
10
15
20
25
30
Number of Correct Agreements

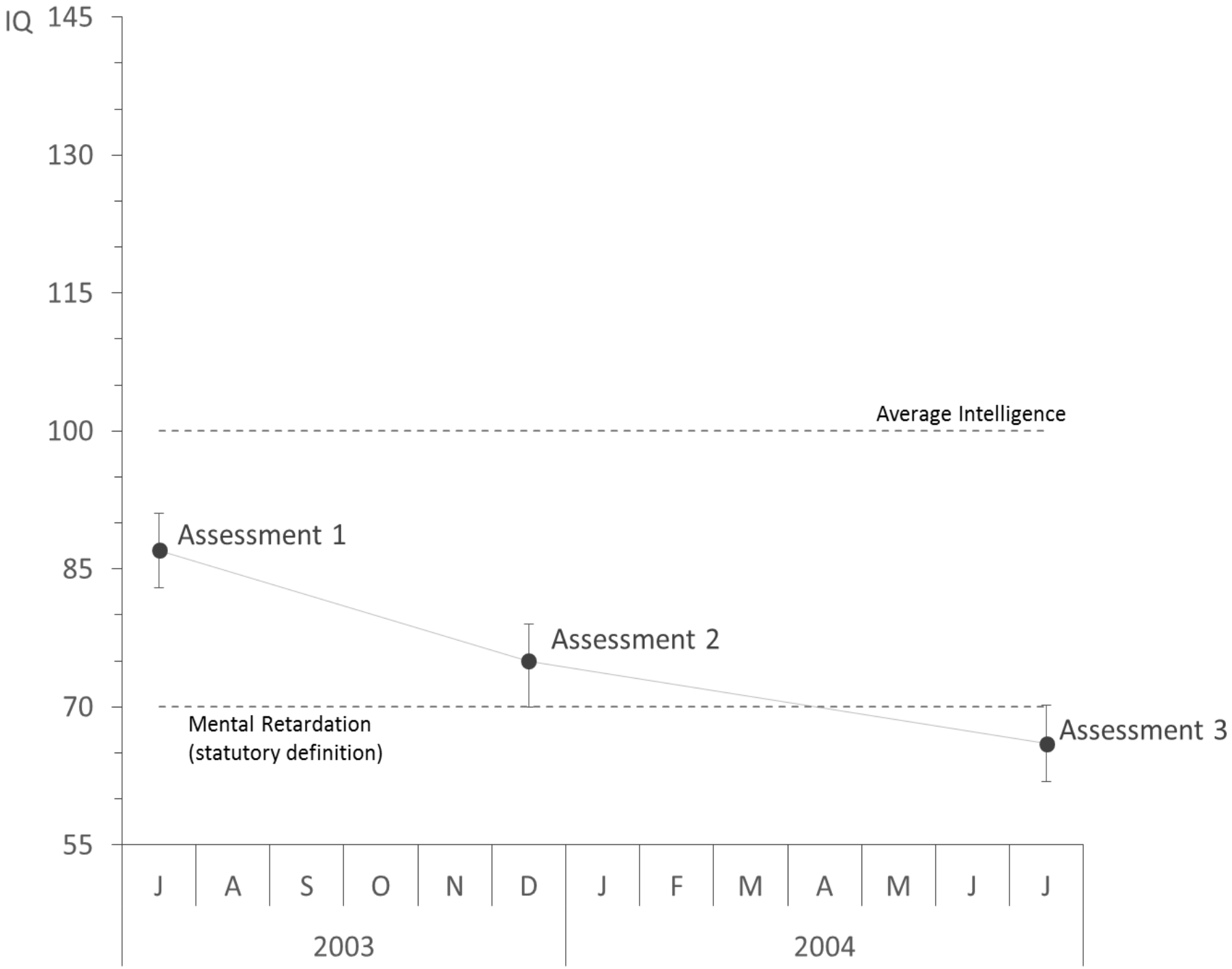
Do Not Intentionally Mislead

- Labels should not be ambiguous.
- Choices of scales and aspect ratios should be defensible and appropriate
- Clearly indicate projections or estimates.
- Units should be appropriate (e.g., actual values vs. percent change for stocks; inflation adjusted vs. unadjusted incomes; age-adjusted vs. unadjusted death rates)


IQ
90
85
80
75
70
65
Assessment 1
Assessment 2
Assessment 3

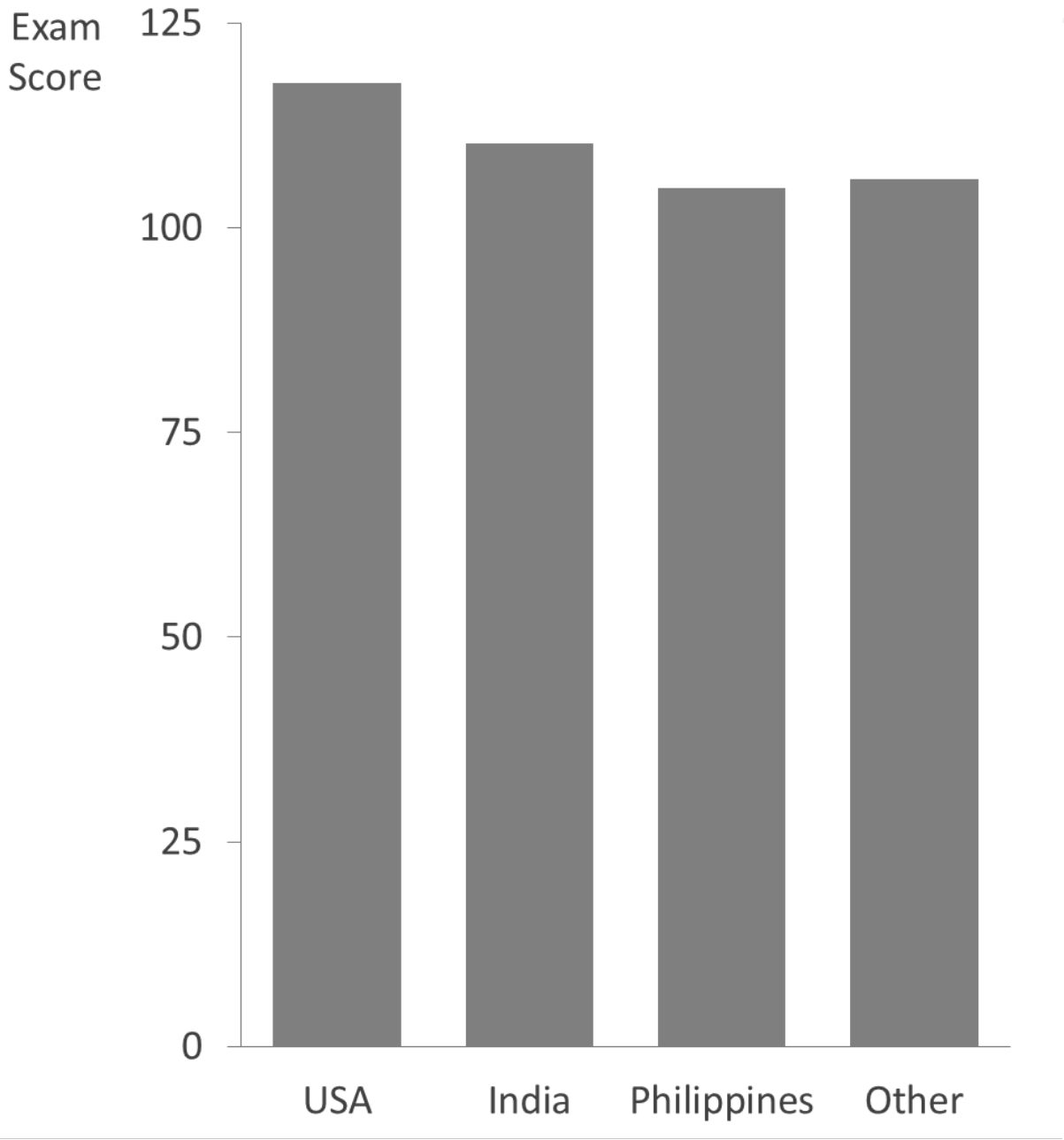

IQ
90
85
80
75
70
65
Assessment 1
Assessment 2
Assessment 3
J
A
S
O
N
D
J
F
M
A
M
J
J
2003
2004

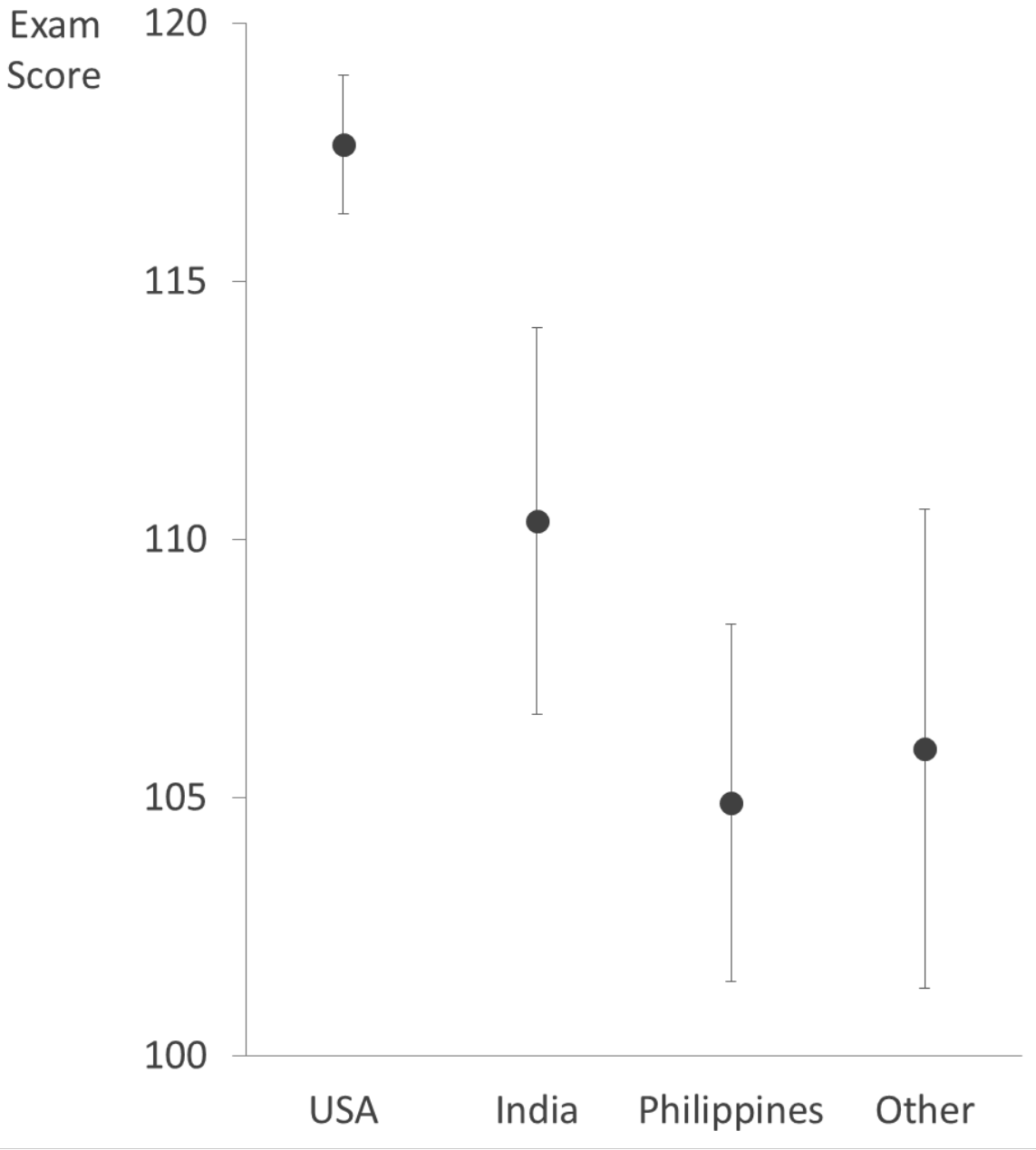

IQ
145
130
115
100
85
70
55
Average Intelligence
Assessment 1
Assessment 2
Assessment 3
Mental Retardation
(statutory definition)
J
A
S
O
N
D
J
F
M
A
M
J
J
2003
2004

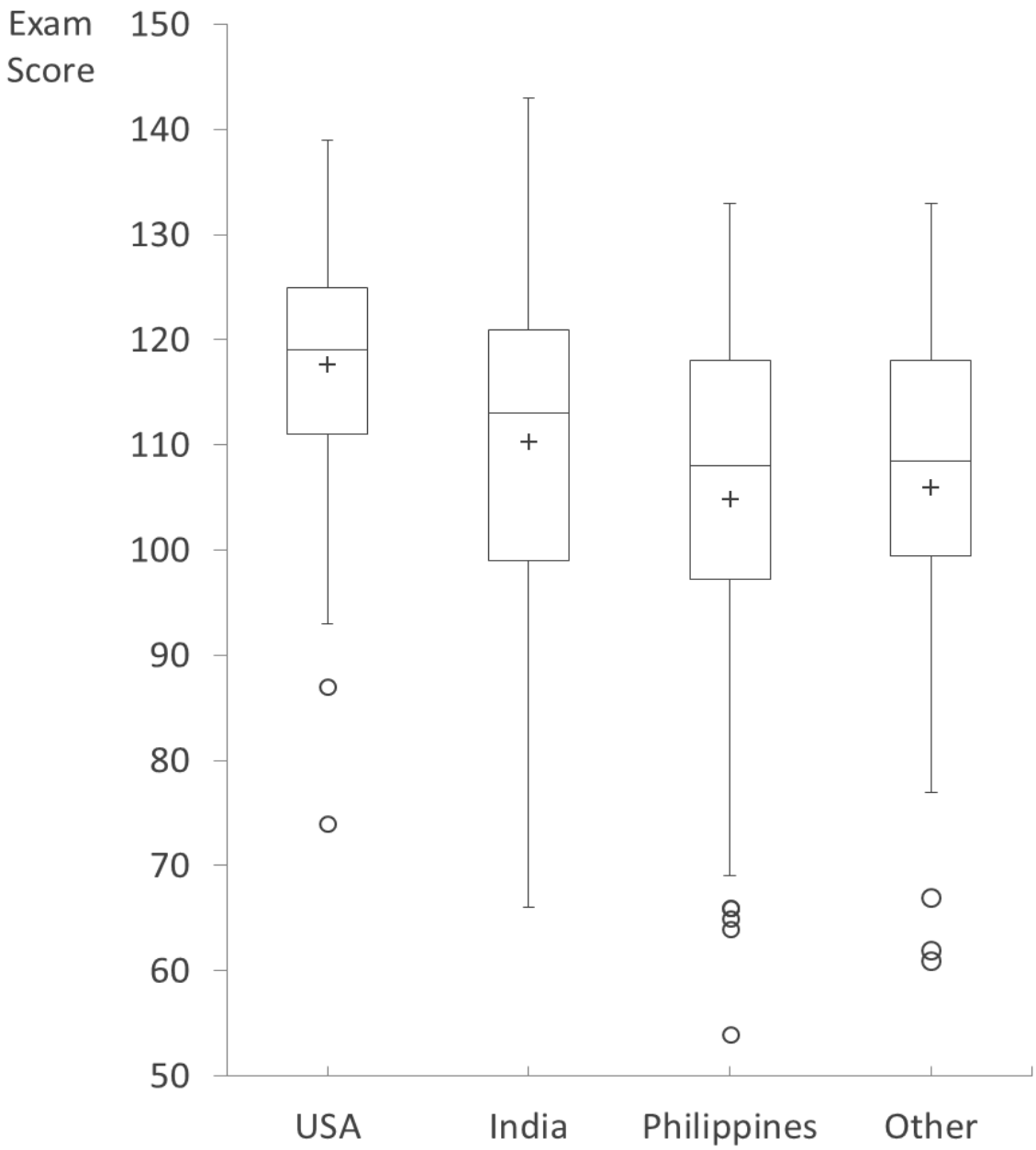

IQ
145
130
115
100
85
70
55
Average Intelligence
Assessment 1
Assessment 2
Assessment 3
Mental Retardation
(statutory definition)
J
A
S
O
N
D
J
F
M
A
M
J
J
2003
2004

Pair the Design with the Audience

- Not all visual displays are equally effective for all audiences.
- Work with test users/consumers to identify the formats and variables that most effectively convey the intended information


Exam Score
125
100
75
50
25
0
USA
India
Philippines
Other


Exam Score
120
115
110
105
100
USA
India
Philippines
Other


Exam Score
150
140
130
120
110
100
90
80
70
60
50
USA
India
Philippines
Other

Take away messages

- Psychometrics is (are?) hard. Visuals can make it easier.
- One size does not fit all
- Be creative and evaluate your creations (see Zenisky & Hambleton, 2009; Zwick, Zapata-Rivera, & Hegarty, 2014)
- Share your results

Contact information

Brett.Foley@alpinetesting.com