Design Goals:
NAEP 2002 and Beyond

Andrew Kolstad

January 7, 2002

NATIONAL CENTER FOR EDUCATION STATISTICS

Four Design Goals

1. Speed up sampling and weighting by administering tests for different subjects (such as mathematics, science, and reading) in the same classroom
2. Speed up scaling by conducting pilot tests of candidate items two years in advance and pre-calibration field tests one year in advance of data collection
3. Link findings deriving from old and new designs by temporarily overlapping old and new methods of data collection
4. Increase NAEP's power to measure performance gaps

Goal 1: Common Block Design

Assessment time	5	10	15	20	25	30	35	40	45	50	55	60
Reading	BQ_1	first block					second block					BQ_2
Mathematics	BQ_1	BQ_2	first block			second block			third block			

BQ_1 is a five-minute questionnaire asking generic background questions; BQ_2 is a five-minute questionnaire asking subject-specific background questions

Test questions are grouped into blocks of about a dozen items; test booklets contain either one 50-minute or two 25-minute blocks (reading) or three 15-minute blocks (mathematics)

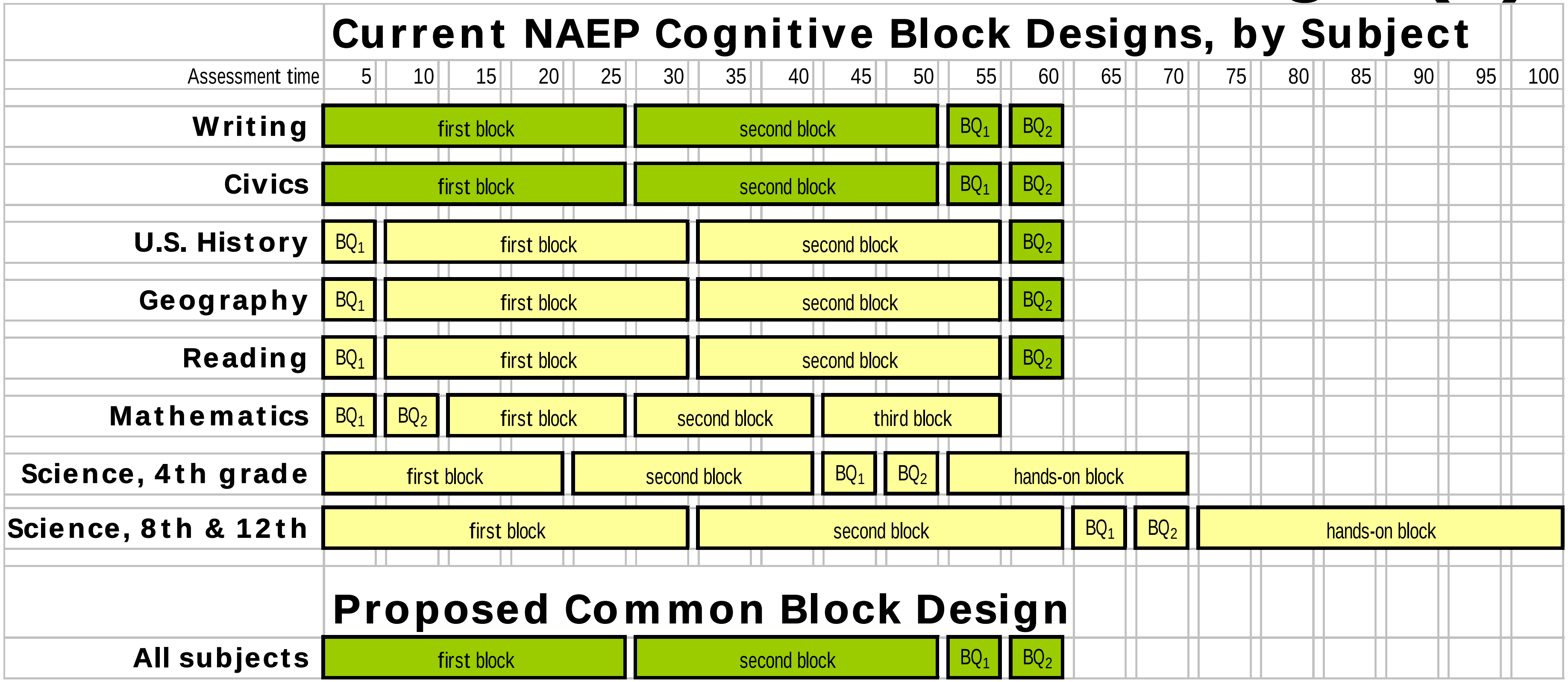
- Problem: timing of instructions requires separate groups for different subjects
- Consequences: more rooms, administrative complexity, different sets of sampling weights take time to produce, small samples difficult to conduct

Goal 1: Common Block Design (2)

Assessment time	5	10	15	20	25	30	35	40	45	50	55	60
Writing	first block					second block					BQ_1	BQ_2
Civics	first block					second block					BQ_1	BQ_2

- Solution: timing of instructions permits assessing different subjects in the same groups
- Consequences: fewer classrooms needed, less administrative complexity, smaller design effect (fewer cases per school), simpler sampling weights

Goal 1: Common Block Design (3)



Current NAEP Cognitive Block Designs, by Subject

Assessment time	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
Writing	first block					second block					BQ_1	BQ_2								
Civics	first block					second block					BQ_1	BQ_2								
U.S. History	BQ_1	first block					second block					BQ_2								
Geography	BQ_1	first block					second block					BQ_2								
Reading	BQ_1	first block					second block					BQ_2								
Mathematics	BQ_1	BQ_2	first block			second block			third block											
Science, 4th grade	first block				second block				BQ_1	BQ_2	hands-on block									
Science, 8th & 12th	first block						second block						BQ_1	BQ_2	hands-on block					

Proposed Common Block Design

Assessment time	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
All subjects	first block					second block					BQ_1	BQ_2								

- Block order change needed in U.S. history, geography, reading & mathematics
- Test block reconfigurations needed in mathematics & science

Goal 2: Pre-calibrate Items

	Current developmental cycle					
Year	2000	2001	2002	2003	2004	2005
Reading	item dev	field testing	data collection	reporting		
Mathematics			item dev	field testing	data collection	reporting

- Problem: weighting, scoring & scaling occur after data collection
- Consequence: unavoidable delay in reporting

Goal 2: Pre-calibrate Items (2)

	Current developmental cycle					
Year	2000	2001	**2002**	2003	2004	2005
2002 Reading	item dev	field testing	data collection	reporting		
2002 Writing	item dev	field testing	data collection	reporting		
2003 Civics		item dev	field testing	data collection	reporting	
2004 Science			item dev	field testing	data collection	reporting

- This figure shows assessments that are currently scheduled
- Next figure shows the concept for new reading and mathematics assessment cycles

Goal 2: Pre-calibrate Items (3)

Year	one	two	three	four	five	six	seven
Year four	item dev	pilot test	field test/ calibration	data collection & reporting			
Year five		item dev	pilot test	field test/ calibration	data collection & reporting		
Year six			item dev	pilot test	field test/ calibration	data collection & reporting	
Year seven				item dev	pilot test	field test/ calibration	data collection & reporting

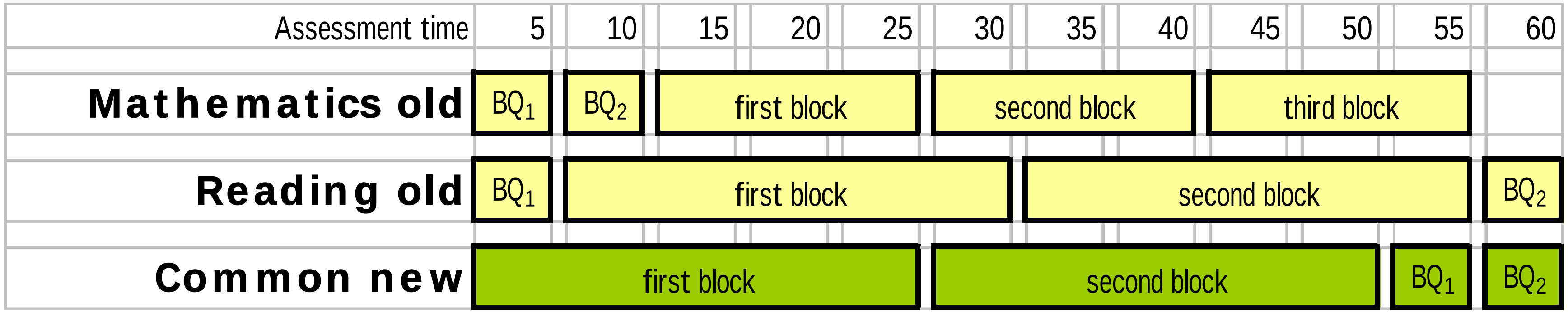
- Solution: calibration of items in advance, integrated samples, and distributed scoring permit more rapid reporting
- Consequence: increased data utility, but longer lead time needed

Goal 2: Pre-calibrate Items (4)

	Proposed developmental cycle						
Year	2000	2001	**2002**	2003	2004	2005	2006
Reading 2003	item dev	pilot test	field test/ calibration	data collection & reporting			
Reading 2005			item dev	pilot test	field test/ calibration	data collection & reporting	
Reading 2007					item dev	pilot test	field test/ calibration
Mathematics 2003		item dev & re-use	field test/ calibration	data collection & reporting			
Mathematics 2005			item dev	pilot test	field test/ calibration	data collection & reporting	
Mathematics 2007					item dev	pilot test	field test/ calibration

- Math 2003 development cycle is abbreviated

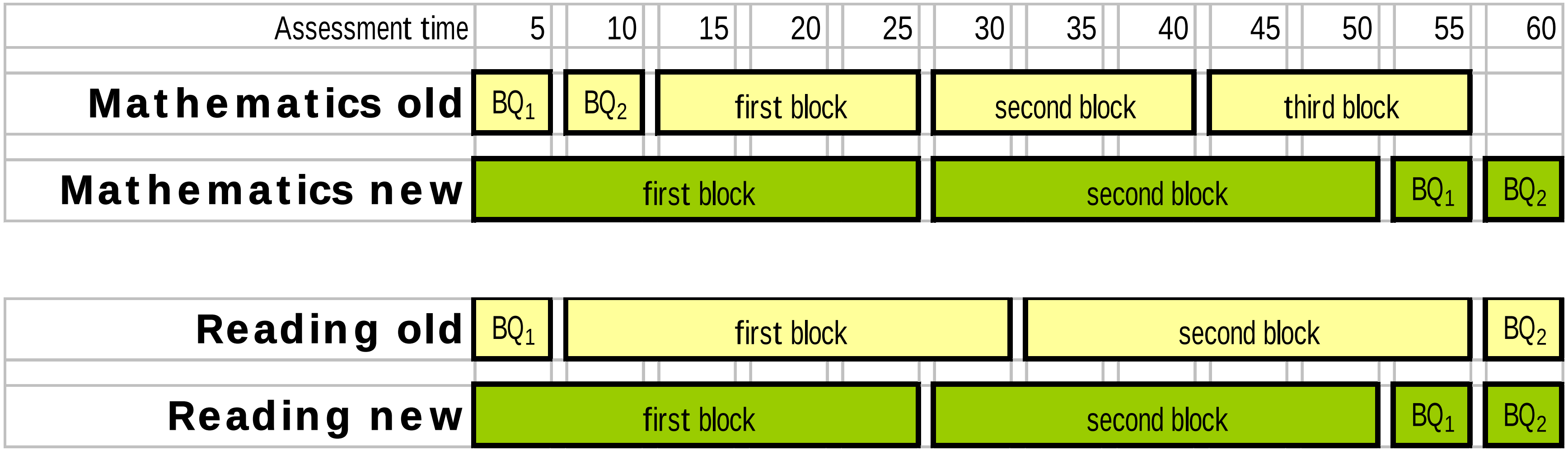
Goal 3: Link Methods



Assessment time	5	10	15	20	25	30	35	40	45	50	55	60
Mathematics old	BQ_1	BQ_2	first block			second block			third block			
Reading old	BQ_1	first block					second block					BQ_2
Common new	first block					second block					BQ_1	BQ_2

- Problem: reconfiguring the test questions into longer blocks (math) and moving BQ_1 (reading) could change scale parameters
- Consequences: potential loss of trend and achievement level cut scores

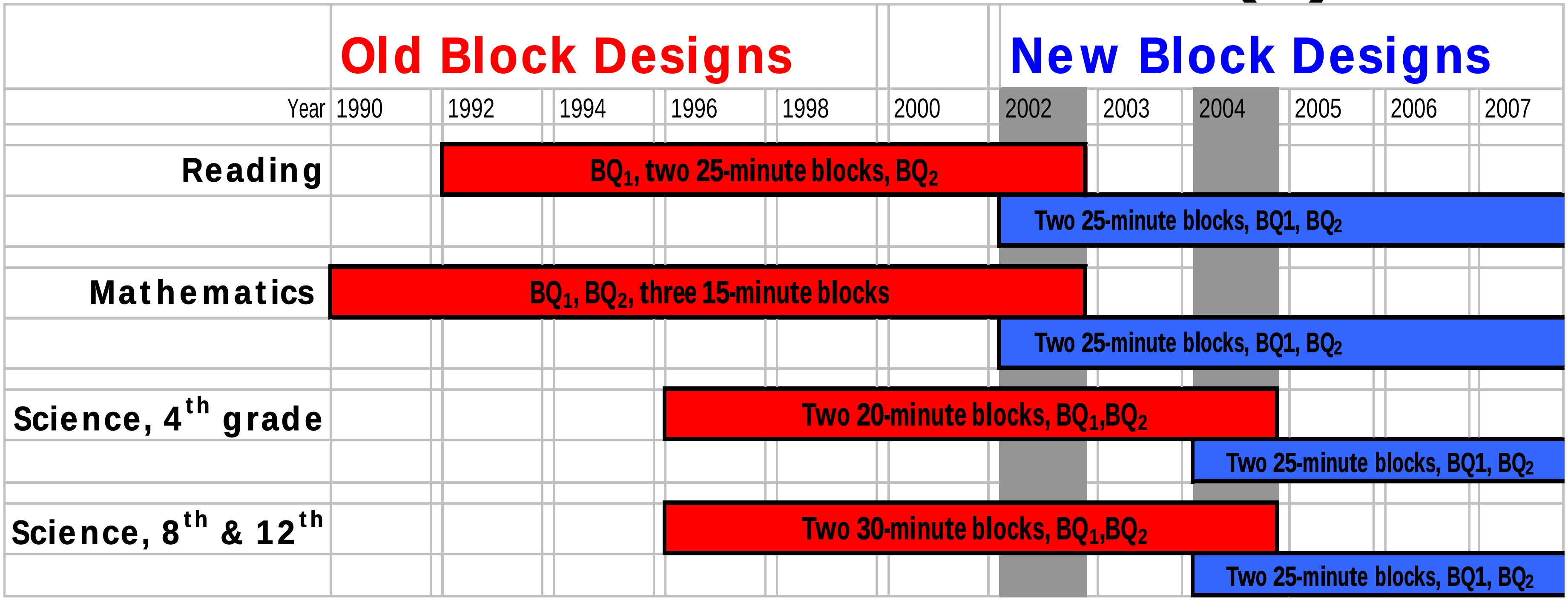
Goal 3: Link Methods (2)



Assessment time	5	10	15	20	25	30	35	40	45	50	55	60
Mathematics old	BQ$_1$	BQ$_2$	first block			second block			third block			
Mathematics new	first block					second block					BQ$_1$	BQ$_2$
Reading old	BQ$_1$	first block						second block				BQ$_2$
Reading new	first block					second block					BQ$_1$	BQ$_2$

- Solution: administer old and new versions to different samples
- Consequence: measurable impact of changing block design

Goal 3: Link Methods (3)



Old Block Designs

New Block Designs

Year 1990 1992 1994 1996 1998 2000 2002 2003 2004 2005 2006 2007

Reading

BQ_1, two 25-minute blocks, BQ_2

Two 25-minute blocks, BQ1, BQ_2

Mathematics

BQ_1, BQ_2, three 15-minute blocks

Two 25-minute blocks, BQ1, BQ_2

Science, 4th grade

Two 20-minute blocks, BQ_1,BQ_2

Two 25-minute blocks, BQ1, BQ_2

Science, 8th & 12th

Two 30-minute blocks, BQ_1,BQ_2

Two 25-minute blocks, BQ1, BQ_2

- Both old and new instruments administered to different samples
  - in 2002 (reading & mathematics)
  - in 2004 (science)

Summary: Four NAEP Session Types in 2002

Session Type A

Reading and writing assessments	
National component ◆ 42,500 students ◆ grades 4, 8 & 12 ◆ private as well as public schools	State component ◆ fewer than 50 jurisdictions ◆ public schools only ◆ 525,000 students in grades 4 & 8 only ◆ also counted in national sample
	Reading BQ blocks moved to end of booklet.
	Reading and writing booklets spiraled together and administered in the same groups

Session Type B

Reading and mathematics field tests for 2003 and pilot tests for 2004	
National only ◆ 111,000 students ◆ grades 4, 8	Math items reconfigured into 25-minute blocks. BQ blocks moved to end.
	Reading BQ blocks moved to end of booklet.
	Math and reading pilot tests large enough to identify improvements in items to be used in 2004
	Math and reading field tests large enough to estimate item parameters to be used in 2003
	Reading and math booklets spiraled together and administered in the same groups, along with Session Type A

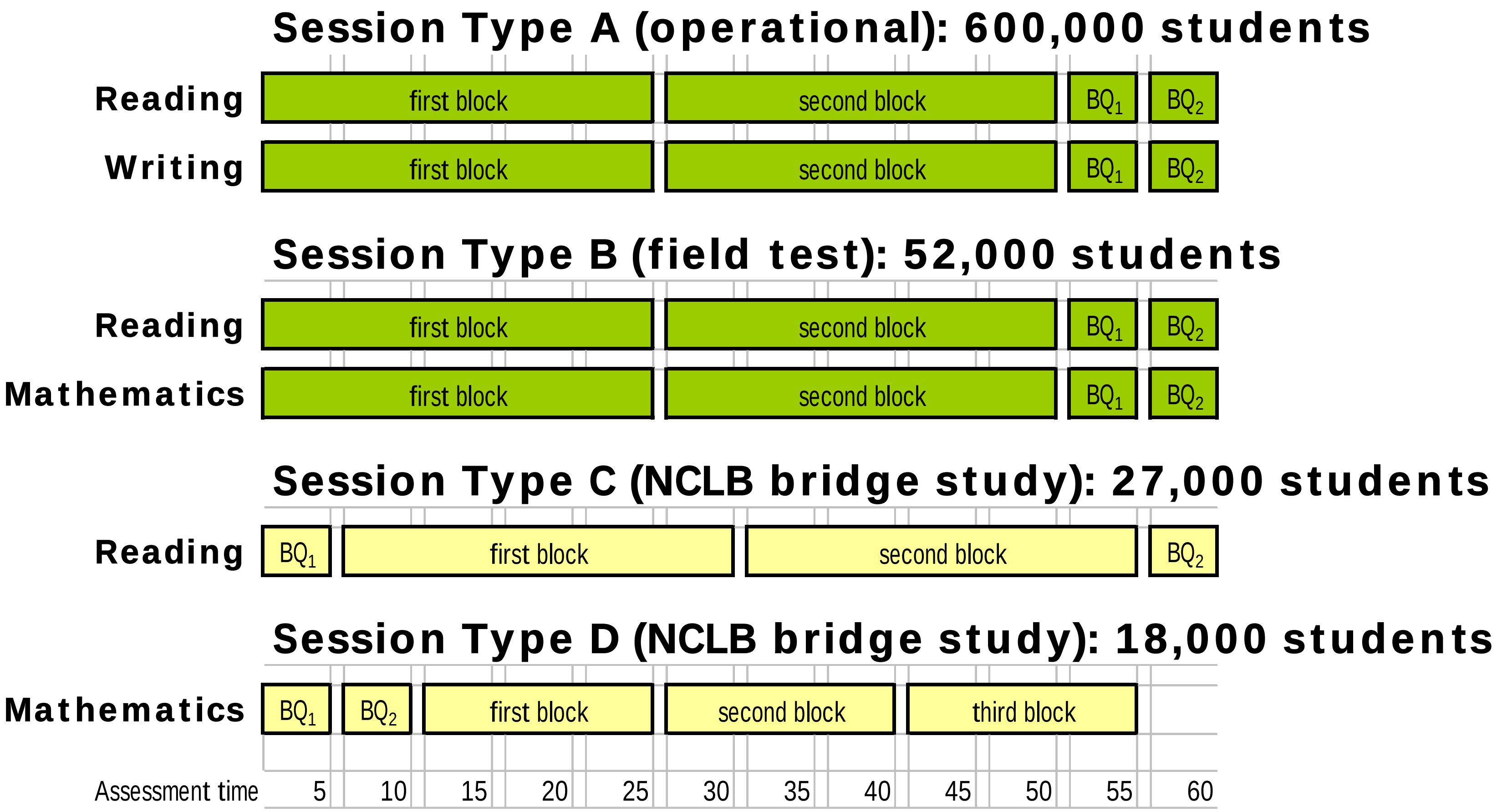
Session Type C

Reading bridge assessment	
National only ◆ 32,000 students	Uses the existing 25-minute block configuration and BQ placement
	Administered in separate groups

Session Type D

Math equating assessment	
National only ◆ 18,000 students	Uses the existing 15-minute block configuration and BQ placement
	Administered in separate groups

Summary: Timing of Blocks by 2002 Session Type



Session Type A (operational): 600,000 students

Reading: first block | second block | BQ_1 | BQ_2

Writing: first block | second block | BQ_1 | BQ_2

Session Type B (field test): 52,000 students

Reading: first block | second block | BQ_1 | BQ_2

Mathematics: first block | second block | BQ_1 | BQ_2

Session Type C (NCLB bridge study): 27,000 students

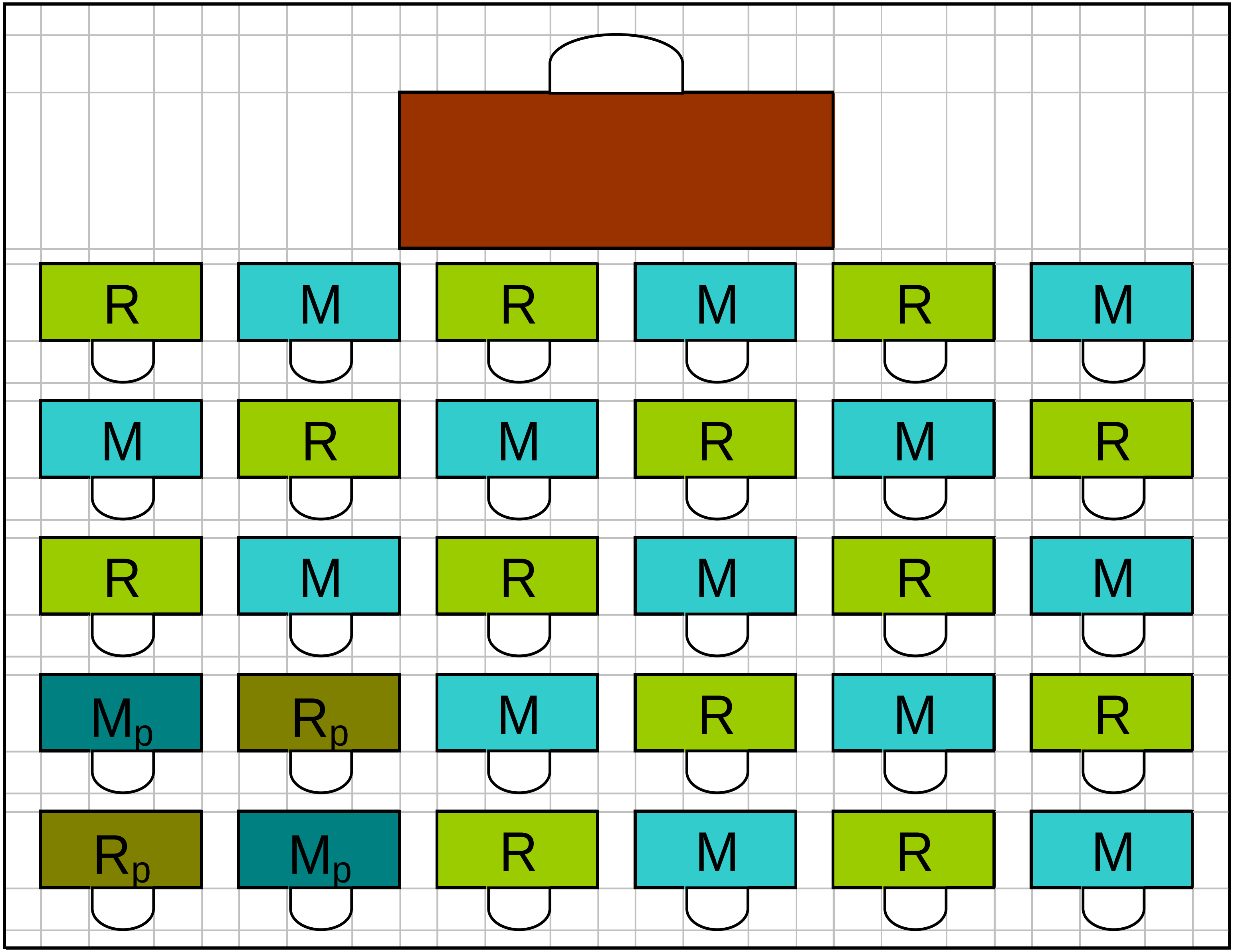
Reading: BQ_1 | first block | second block | BQ_2

Session Type D (NCLB bridge study): 18,000 students

Mathematics: BQ_1 | BQ_2 | first block | second block | third block

Assessment time 5 10 15 20 25 30 35 40 45 50 55 60

2003 Booklet Spiraling



R	M	R	M	R	M
M	R	M	R	M	R
R	M	R	M	R	M
M_p	R_p	M	R	M	R
R_p	M_p	R	M	R	M

Goal 4: Measure Gaps (1)

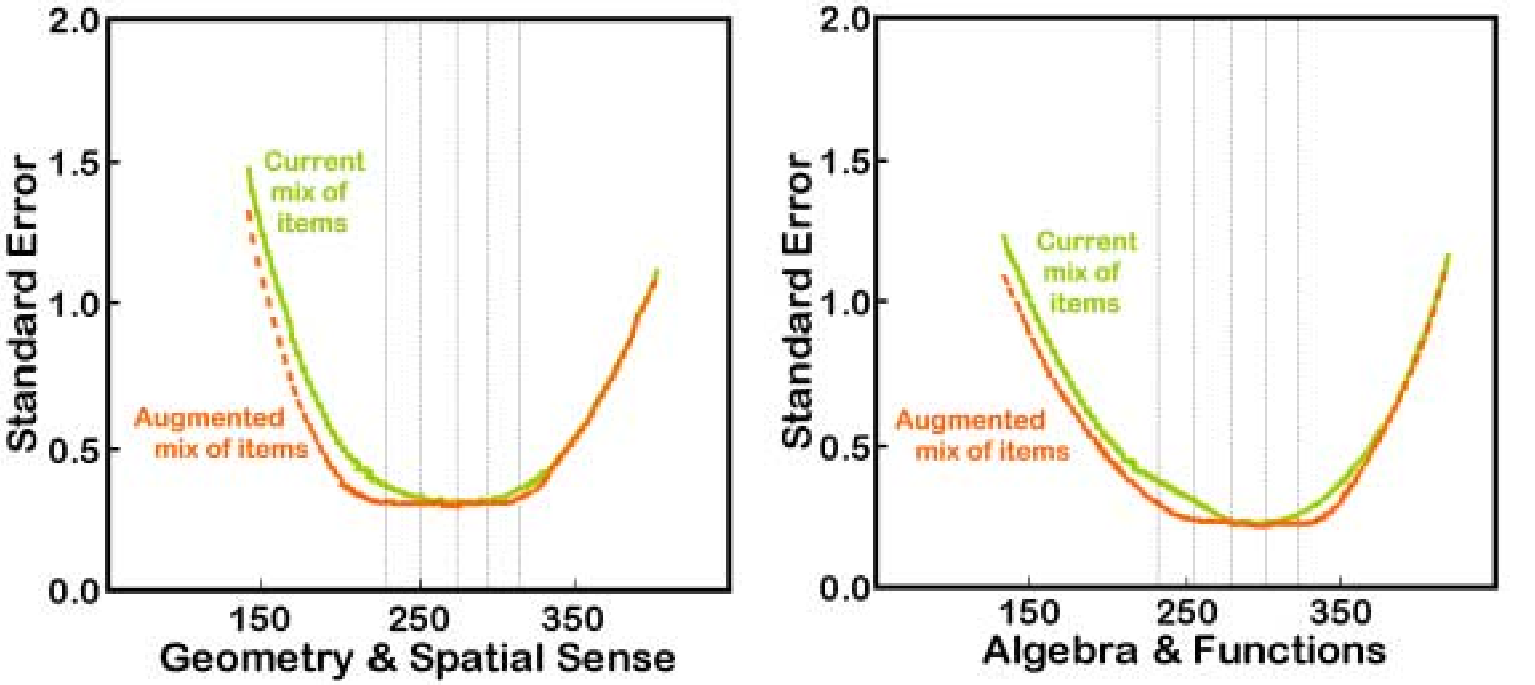
- Two factors will affect NAEP's capacity to measure performance gaps:

  1. *Sample sizes*: accuracy of NAEP score for a subgroup depends on the size of the subgroup sample
  2. *Assessment scales*: accuracy of NAEP score for a subgroup is better at the middle of the scales than at the ends of the scales

Goal 4: Measure Gaps (2)

- *Solution 1:* Ensure adequate sample sizes of targeted groups, or states as a whole:
  - Racial/ethnic gaps: Black, Hispanic, and White children
  - Socioeconomic gaps: Children eligible for free and reduced-price school lunch program
  - Performance gaps: Children at the 10-25th percentiles and those at the 75th-90th percentiles

Goal 4: Measure Gaps (3)



- *Solution 2:* Add test questions at both ends of difficulty range (but more at the lower end)