Computerized Adaptive Testing (and other uses of tests with computers)

Lecture #9

ICPSR Item Response Theory Workshop

Lecture Overview

- Computer-based Tests
- Computerized Adaptive Testing
- Multistage Testing (MST)

COMPUTER BASED TESTING

Importance of Computer Based Testing (CBT)

- Prevalence of operational testing programs currently administering CBTs

- Many organizations are interested in taking advantage of more sophisticated options offered by CBT

Advantages of CBT

- Innovative item formats
- Increased availability
- Faster score reporting
- Increased security
- Adaptive administration of items
- Many others…

Computer Based Tests

- Linear CBT
  - Fixed-form CBT
  - Basically, putting a paper-and-pencil test directly on the computer

Linear CBT

Fixed Form

60 items administered in a traditional way, only computer delivered instead of using a paper-and-pencil format

Computerized Adaptive Tests

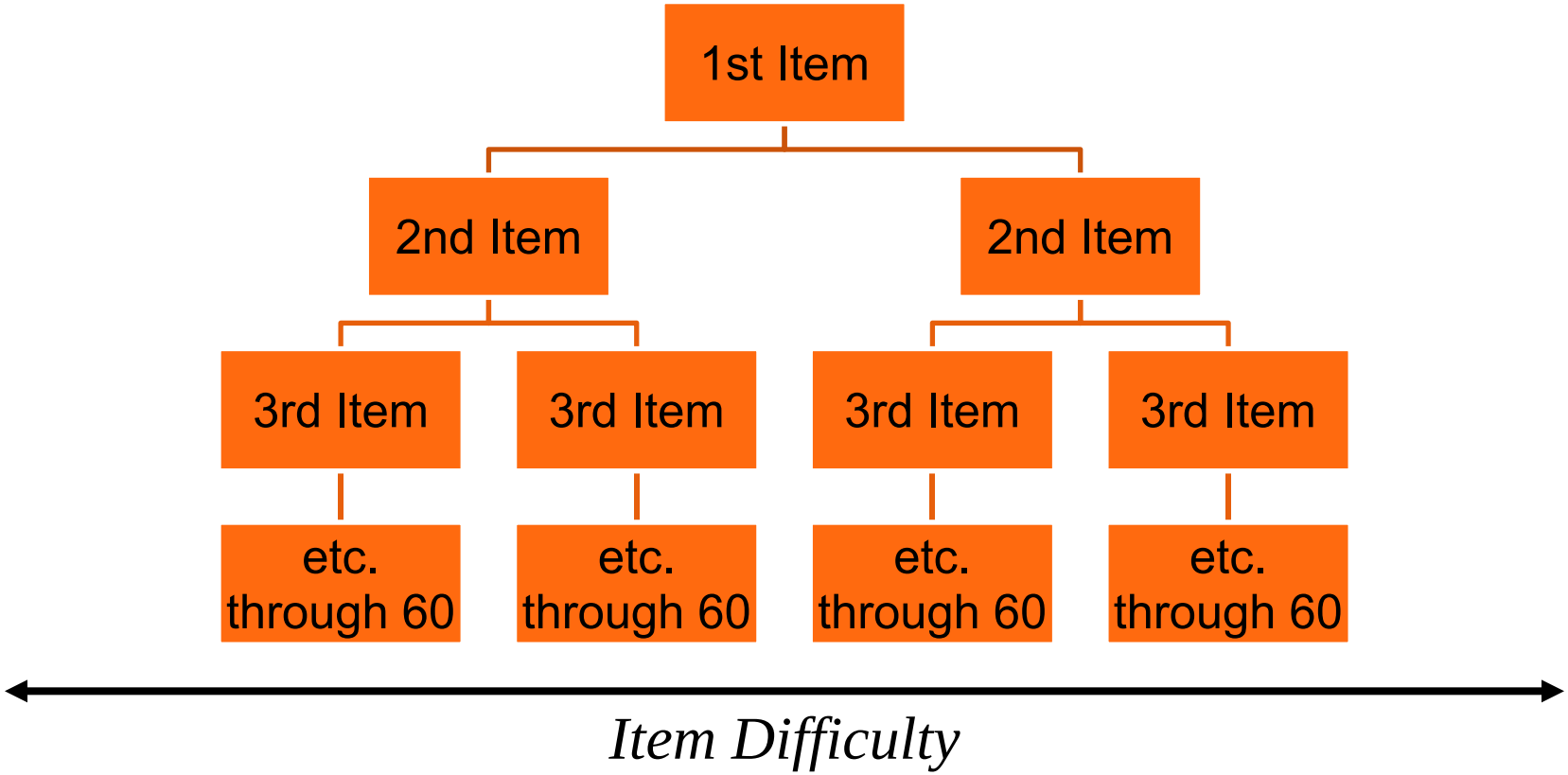
- Administration of each item is adapted to person responses
- Selection of each subsequent item is targeted to person proficiency
- Dramatically increased measurement efficiency

Examples of Computerized Adaptive Tests

- The Graduate Record Examination (GRE)

- The National Council of State Boards of Nursing Exam
  - https://www.ncsbn.org/1216.htm

CAT



CAT could be administered for any item selection procedure

Item Selection

- The “gold standard” of item selection is a procedure that can maximize information while minimizing exposure and maintaining proper content representation

- Why that might be helpful?

Item Information Function

- "Item Information" indicates an item's usefulness for assessing ability
- By "usefulness" we basically mean how good an item is at distinguishing people with lower ability levels from those with higher ability levels
- Information :: Precision

Item Information Function

- Items are basically more informative where the slope of the ICC is steepest, which happens when…

  b_i is relatively close to θ_s,
  a_i is relatively high, and
  c_i is relatively low

- If $c_i = 0$, an item provides its maximum information when $\theta_s = b_i$

Test Information Function

- Just like we add up ICCs to get a TCC, we add up IIFs to get a TIF
- Information will continue to increase as we add test items, therefore increasing precision
- All things equal, longer tests provide increased measurement precision

Conditional Standard Error

- The imprecision of ability estimation is therefore inversely related to the amount of ***Information*** with respect to ability that is available

- Since Information increases with the ***quality*** and number of items, the SE conversely decreases…

CAT Item Selection

- Maximum Information
- a-stratified with b-blocking
- Specific Information Item Selection

Maximum Information

- Proposed by Fisher, developed by Lord
- Each item is administered to provide maximum information, given the provisional estimate of examinee ability
- Results in the most efficient test possible, given what's in the item pool
- Generally results in selecting items with the highest discrimination…
  - can anyone see what problems may arise with such a routine?

a-stratified with b-blocking

- Proposed by Chang
- Stratify the discrimination parameters across difficulty of items
- Administer items with low ***a*** early, then more discriminating items
- Rationale: initial θ estimates aren't that great, so why waste our most informative items before we know the approximate "neighborhood" for examinee θ?

a-stratified with b-blocking

- Benefits:
  - Allows for more equitable use of the item pool, thereby reducing exposure
  - Samples items more evenly
  - Administers the most informative items when they can do the most

Specific Information Item Selection

- Proposed by Davey & Fan
- Administer items to achieve pre-specified information targets
- We don’t necessarily need a flat information function for all possible administrations…
- Relatively large SE at the tails, or far from a cut-score, might be just fine for our purposes
- Allows for flexibility and more evenly distributed use of item pool

Issues in Adaptive Testing

- Score comparability
- Exposure control
- Content representation
- Item pools must be large

Score Comparability

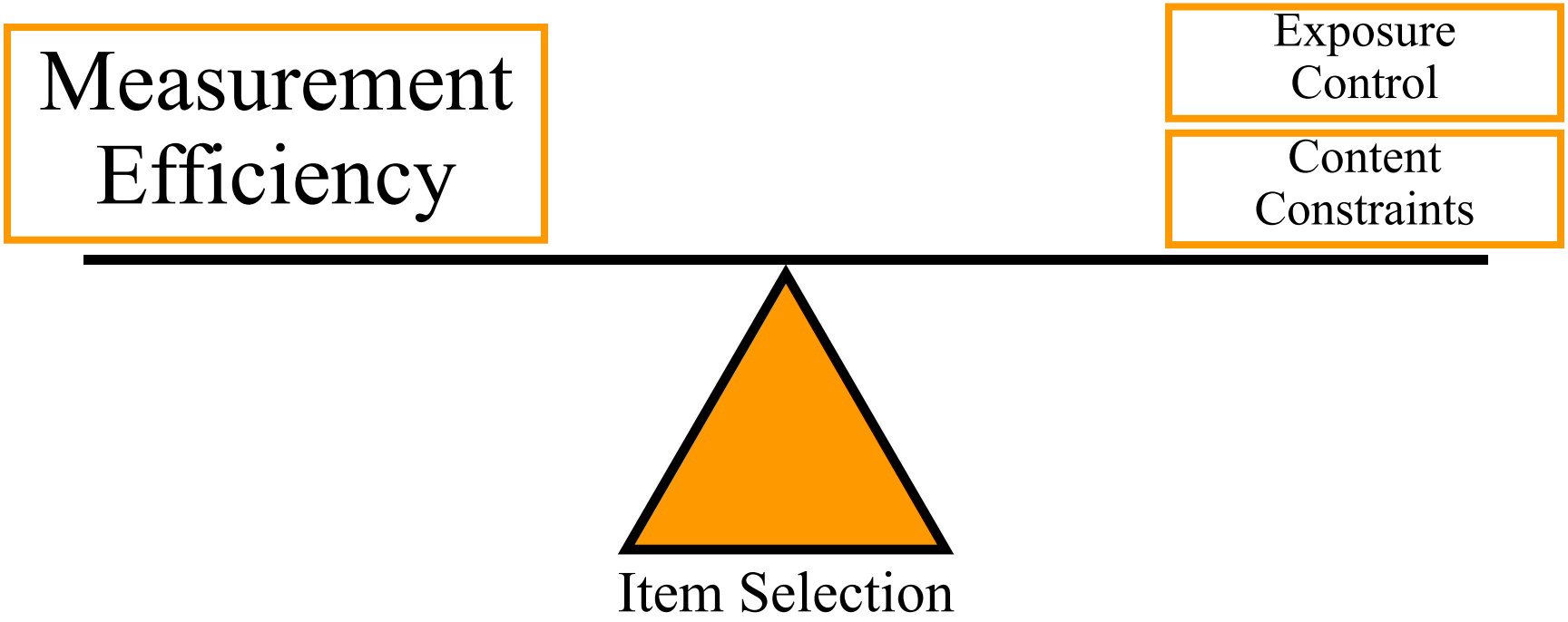
- Measurement precision, efficiency are the benefits, but they come at a price…
- In terms of the models features, scores are comparable, in that they are all based on a pool of items whose parameters are on the same scale
- However, most examinees see completely different test forms, and there is no answer review/change

Exposure Control

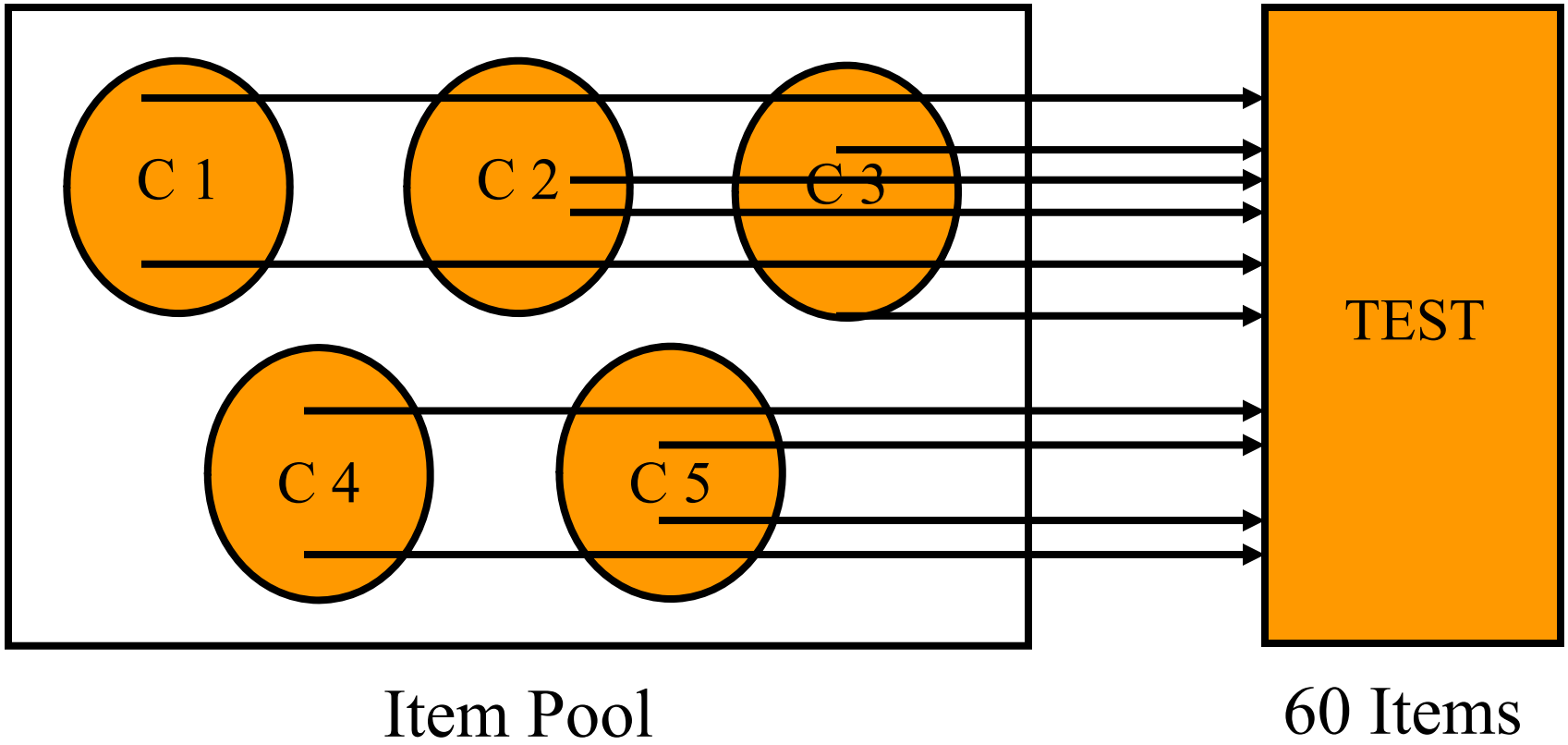
- Through repeated administrations, many items can become over-exposed, thus possibly changing the item parameters
- Different item selection procedures “prefer” certain types of items
- Generally, exposure controls must be written into the CAT algorithm

Content Representation

- There is a challenge to representing test specifications in an adaptive environment
- Algorithms must also include instructions by which items are sampled from within content areas to ensure proper composition
- In CAT, we must balance measurement efficiency and content constraints through the process of item selection


Measurement Efficiency
Exposure Control
Content Constraints
Item Selection

Content Constraints



Large Item Pools

- A CAT basically guarantees a geometric progression of administered test forms
- With “on demand” testing becoming the standard for CAT programs, very large item pools must developed and maintained
- Adaptive administration also requires many items across the span of possible item difficulties to make it worthwhile

Cheating on Computerized Adaptive Tests

- Because CATs depend on the item pool, many successful cheating attempts have been made to CATs
  - Predominantly based on ways to capture items
- Some countries sponsored cheating programs
  - Examinees reportedly were asked to report all items they could remember post-test
  - Chang's research showed only a handful of examinees remembering only a handful of items could compromise the entire test bank
- The GRE was subject to massive cheating during the early part of the 2000s (and maybe before)
  - Therefore…the GRE is now about to change

Solution? Maybe Multi-Stage Tests (MST)…

- Blocks of items, representing partial tests, are administered adaptively
- Groups instead of individual items
- Allows test makers to review all possible forms of a test prior to administration
- Assurance of content representation
- Less efficient than CAT, but more than a linear fixed form administration
- Doesn’t require as large an item pool

Computer-based Testing

Example MST Designs

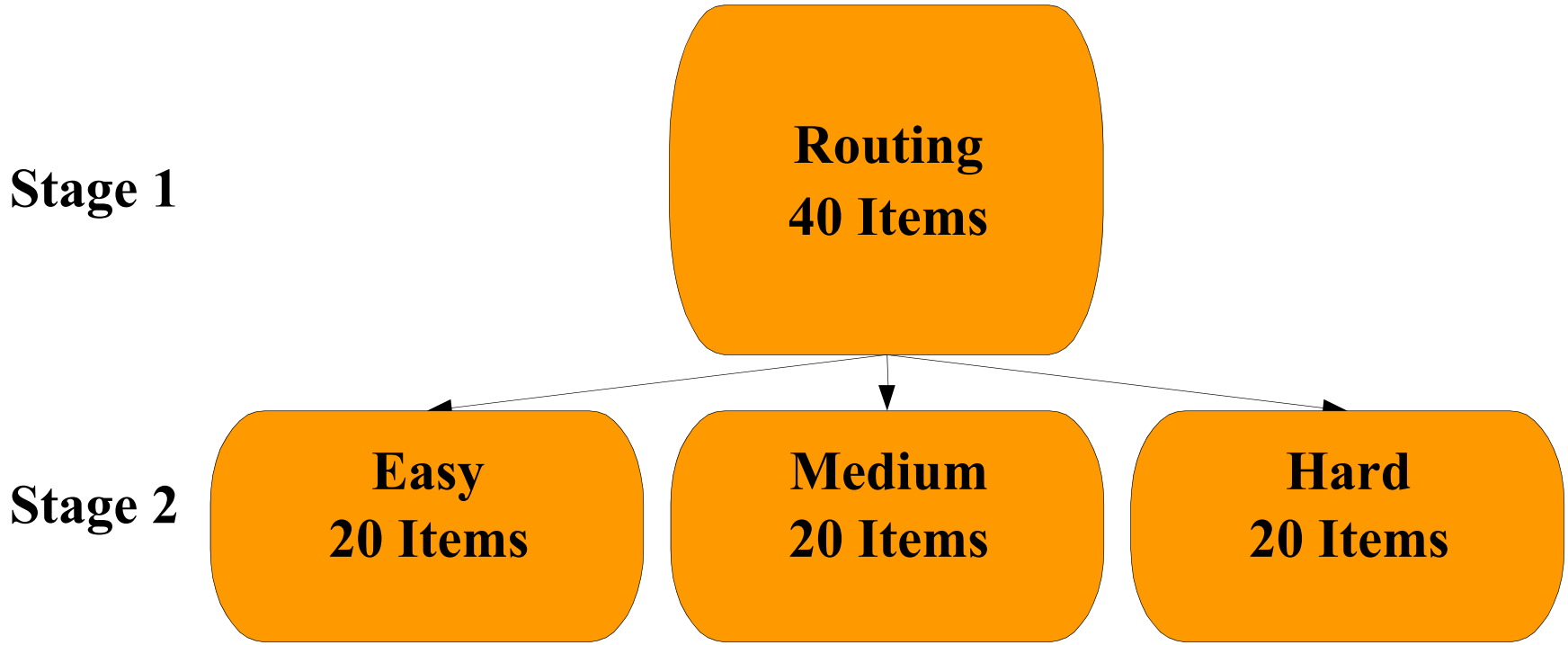
- 2-Stage
- 3-Stage
- 4-Stage
- 5-Stage
- 6-Stage

MST Designs

- 6 Designs → test length = 60
  - 1-3 (40,20)
  - 1-3-3 (40,10,10)
  - 1-3-3 (20,20,20)
  - 1-3-3-3 (30,10,10,10)
  - 1-3-3-3-3 (20,10,10,10,10)
  - 1-3-3-3-3-3 (10,10,10,10,10,10)

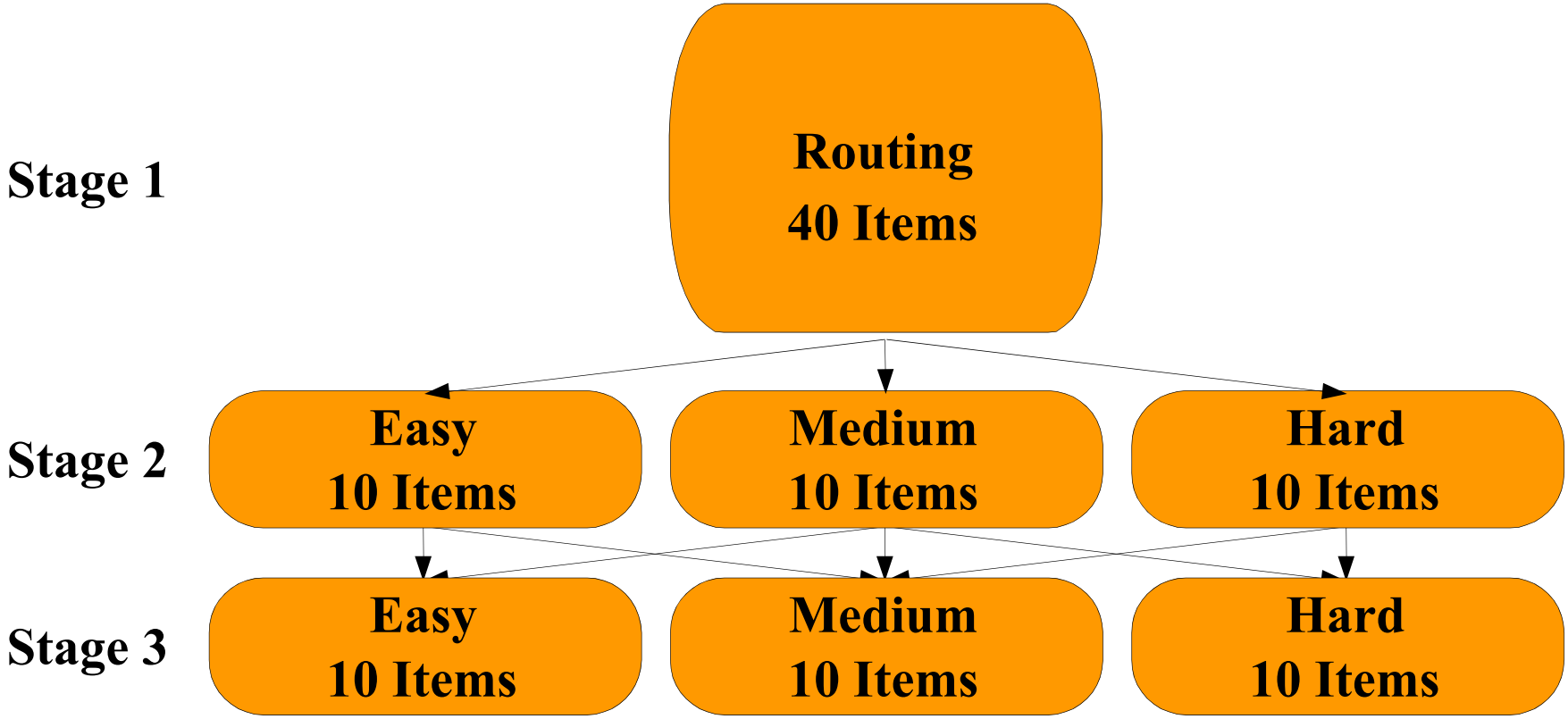
Requires 100 Items

2-Stage Design



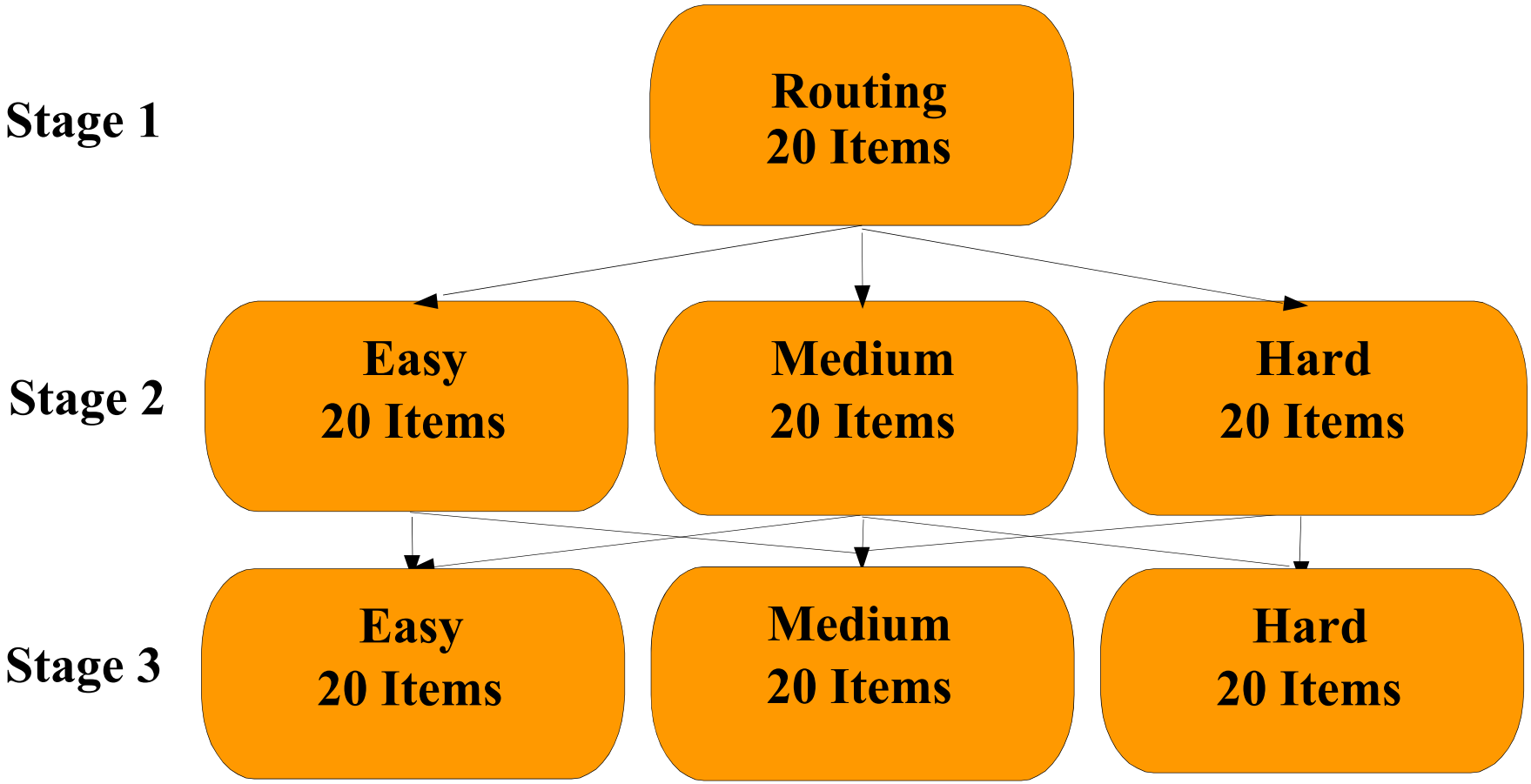
Requires 100 Items

3-Stage Design



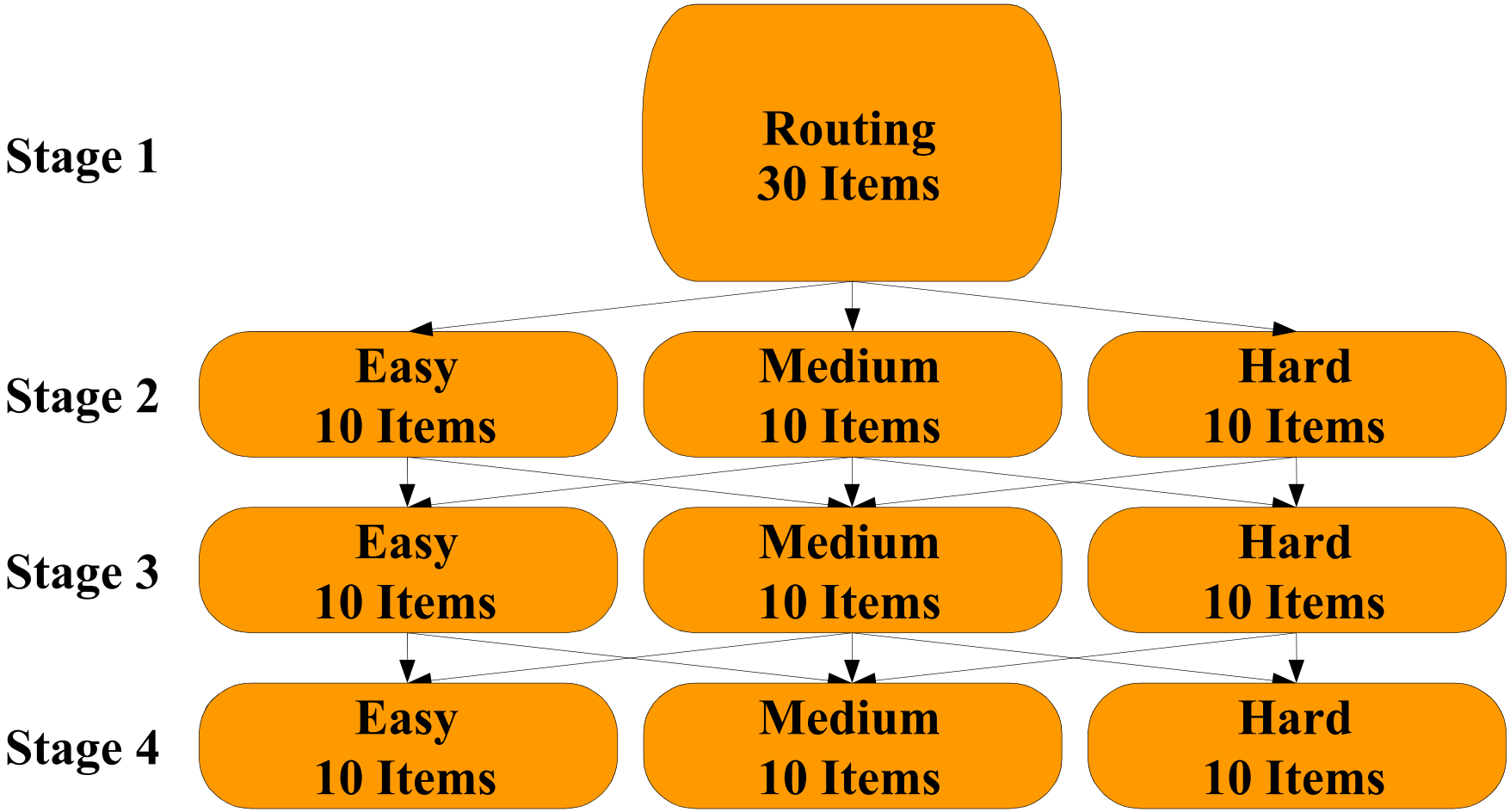
Requires 140 Items

3-Stage Design



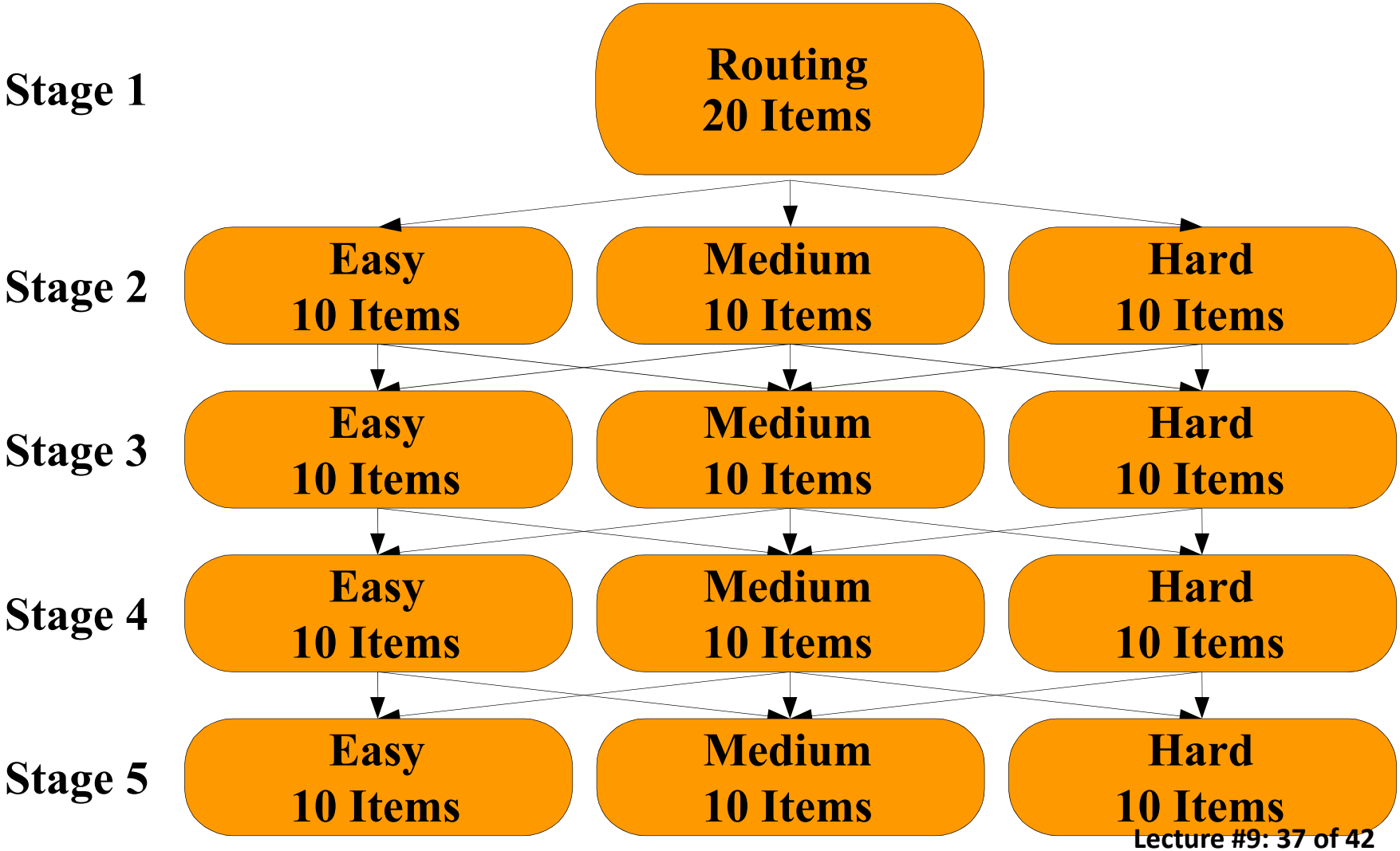
Requires 120 Items

4-Stage Design

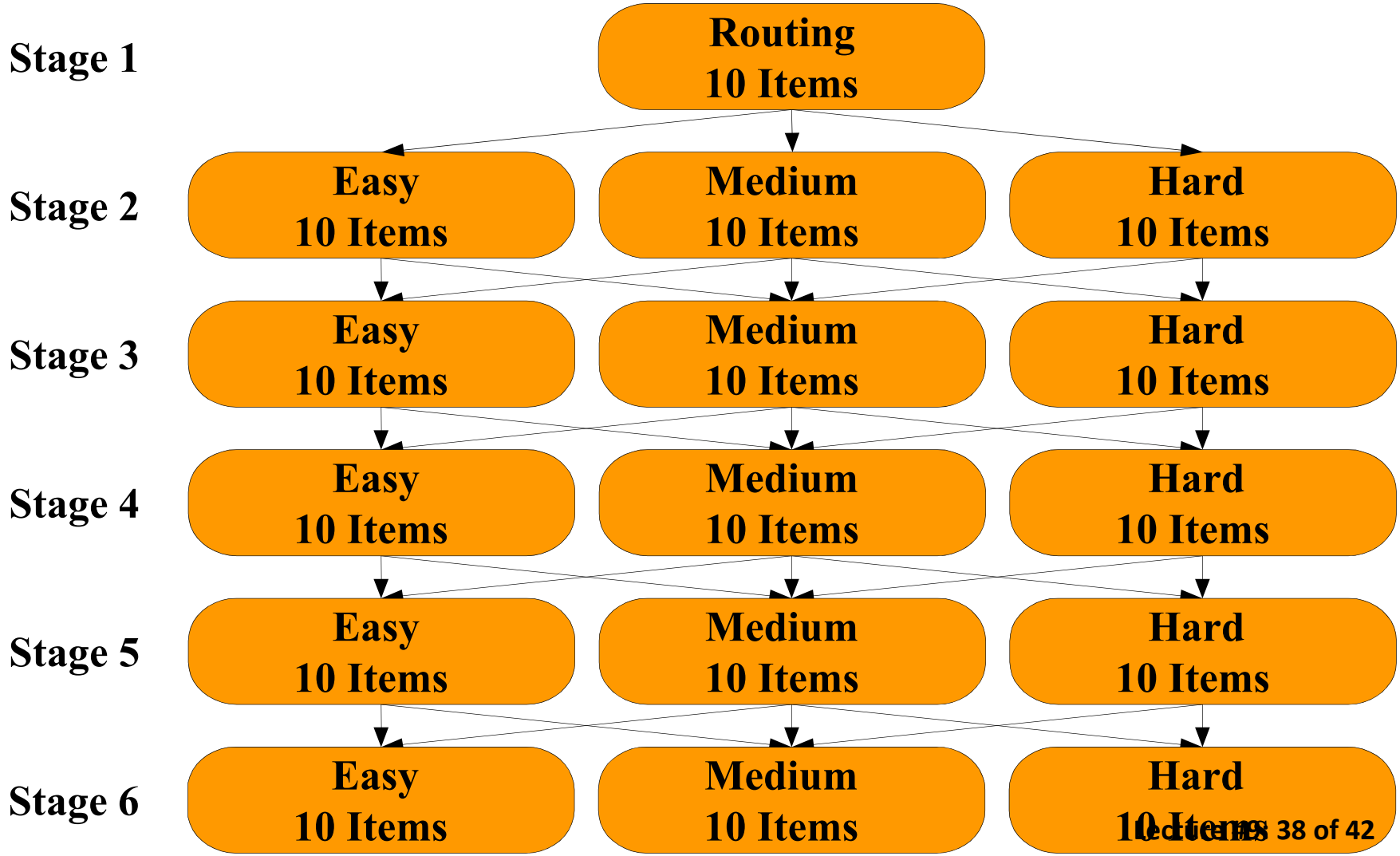


Requires 140 Items

5-Stage Design



Requires 6-Stage Design 160 Items



MST Design Question…

- As each design moves closer to a fully adaptive test, can we identify an "ideal" situation that balances the desired precision and efficiency of θ estimates?

- Good research question…depends, in part, on the nature of the item pool

WRAPPING UP

Concluding Remarks

- Pencil-and-paper tests lack efficiency
  - And therefore may produce lower reliability
- CAT provides maximum efficiency and precision, but at a price:
  - Complicated to implement
  - Harder to defend
  - Difficult to secure
- MST provides a "middle ground" of adaptively administered blocks of items
  - Can be selected a priori
    - Assuring proper test specifications (and doesn't require as large an item pool to maintain)

Up Next

- Computer Activity: Building a test with IRT