Impact of Item Drift on Candidate Ability Estimation

Sarah Hagge, PhD, Psychometrician
Ada Woo, PhD, Senior Psychometrician
Phil Dickison, PhD, RN, Director, Examinations

NCSBN
National Council of State Boards of Nursing

Background

- Computerized adaptive testing
  - Item response theory (IRT)
  - Item pools
  - Ability estimates
- Drift of item parameters can occur over time
  - Security breaches
  - Shifts in instruction or changes in practice
- Accuracy of candidate ability estimates depends on accurate item parameter estimates

Overview of Relevant Literature

- Fixed Forms
  - Impact of item parameter drift on ability estimates is small, even with unidirectional drift (Wells, Subkoviak, & Serlin, 2002)
  - Ability estimates are robust to drift, even when abilities and item difficulties are not normally distributed (Stahl, Bergstrom, & Shneyderman, 2002; Witt, Stahl, Bergstrom, & Muckle, 2003)
  - Although results were mixed, longitudinally, item parameter drift may negatively impact the linking process and resulting candidate ability estimates (Wollack, Sung, & Kang, 2006)
- A real data and simulation study of a CAT program found minimal impact to score stability, though scale drift was also minimal (Guo & Wang, 2003)

NCSBN
National Council of State Boards of Nursing

Purpose and Research Questions

- To investigate the impact of item difficulty drift on candidate ability estimates for variable-length CAT. Specifically,
  1. How robust are candidate ability estimates when item difficulty drift is present to varying degrees in a CAT item pool?
  2. To what extent are pass/fail decisions impacted when item difficulty drift occurs in a CAT item pool?

Data

- Two large-scale licensure examinations
- Variable-length computerized adaptive tests (CAT) scored using the Rasch model
- Exam 1: 18,004 candidates
- Exam 2: 52,765 candidates

Investigation Conditions

- Only item difficulty parameter drift (Rasch model)
- Conditions
  - Percentage of items with drift
    - 5%, 10%, 20%
  - Magnitude of drift
    - 0.50, 0.75, 1.00 logits
  - Direction of drift
    - All items easier, all items harder, half and half
- Conditions fully crossed resulting in 27 conditions for each exam

Analysis

- Item drift randomly introduced into the operational item pool
- 20% of items in the operational pool were randomly selected to exhibit item drift
  - Items for the 10% condition were randomly selected from the 20%
  - Items for the 5% condition were randomly selected from the 10%

Analysis (cont.)

- The magnitude and direction of drift were applied to all items
  - For example,
    - Percentage: 20%
    - Magnitude: 0.50
    - Direction: All easier
    - Drift of -0.50 was applied to all 20% of the items
- Candidate ability estimates were re-estimated by anchoring items using the drifted item difficulty estimates

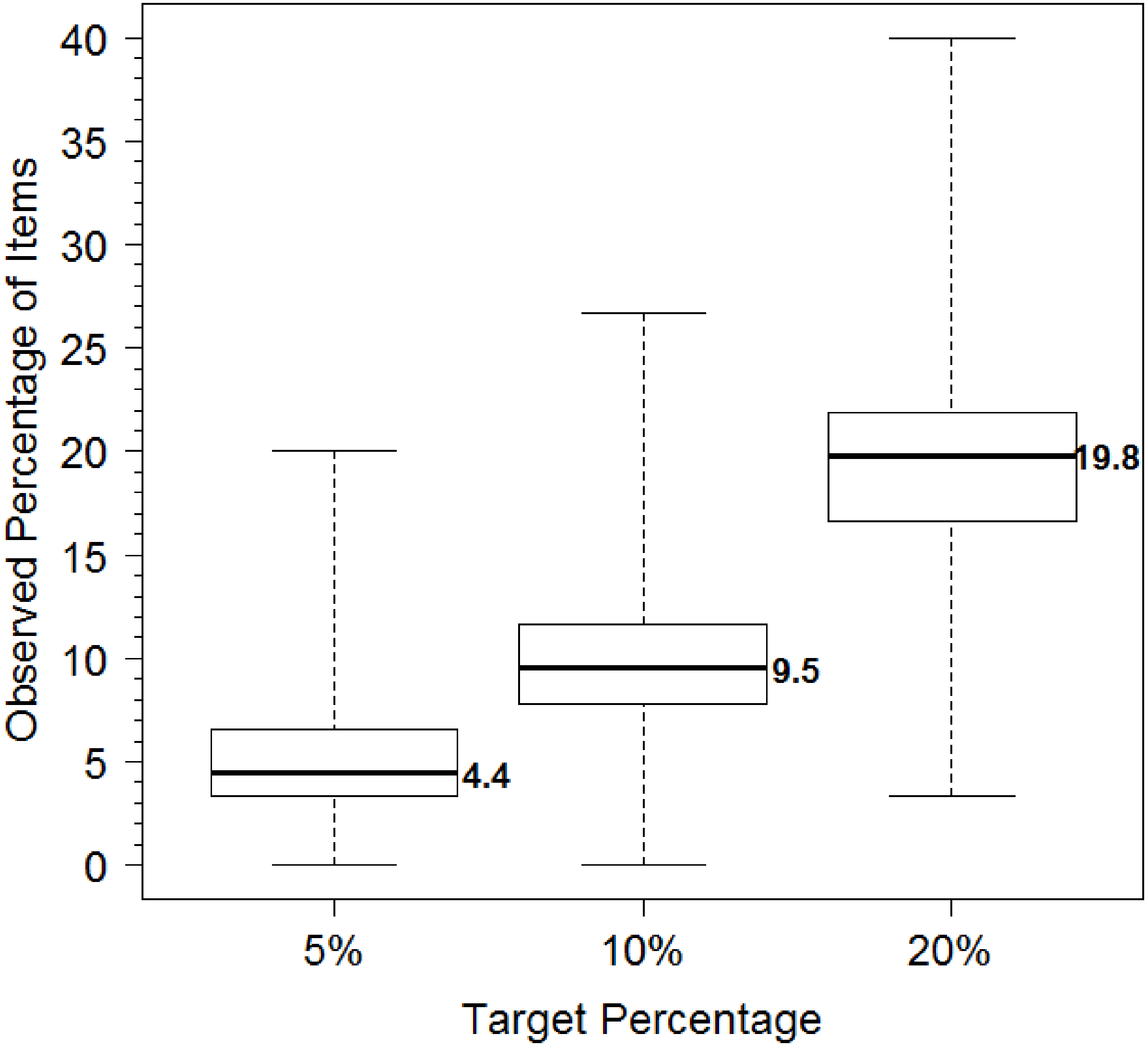
Evaluation

- Difference between re-calibrated candidate ability estimates and original candidate ability estimates
  - Re-calibrated candidate ability estimate minus original candidate ability estimate
  - Minimum, maximum, mean and standard deviation of differences
- Pass/fail decision consistency

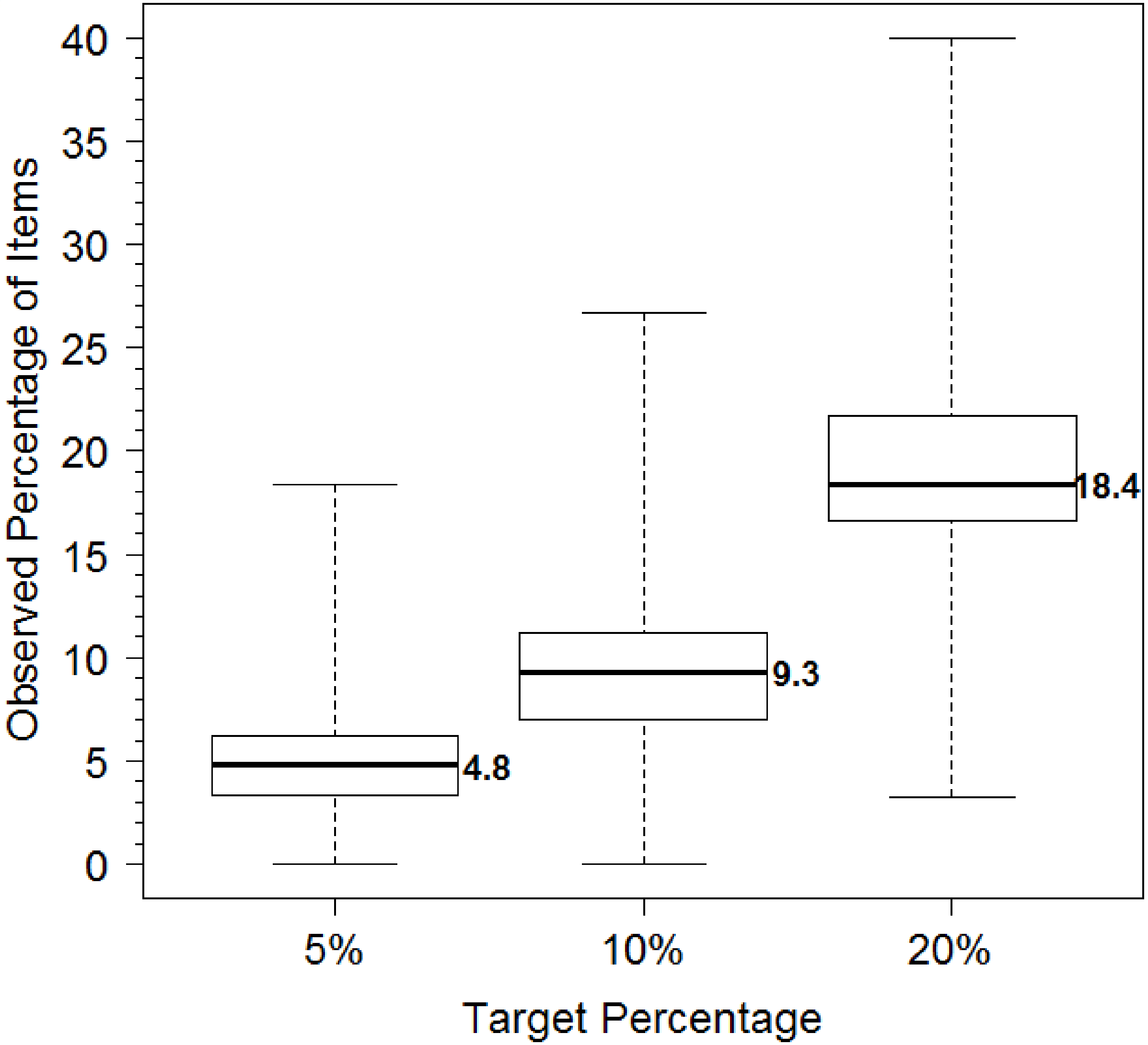
Results

- Percentage of drifted items on individual exams
- Theta differences
- Pass/fail decision consistency

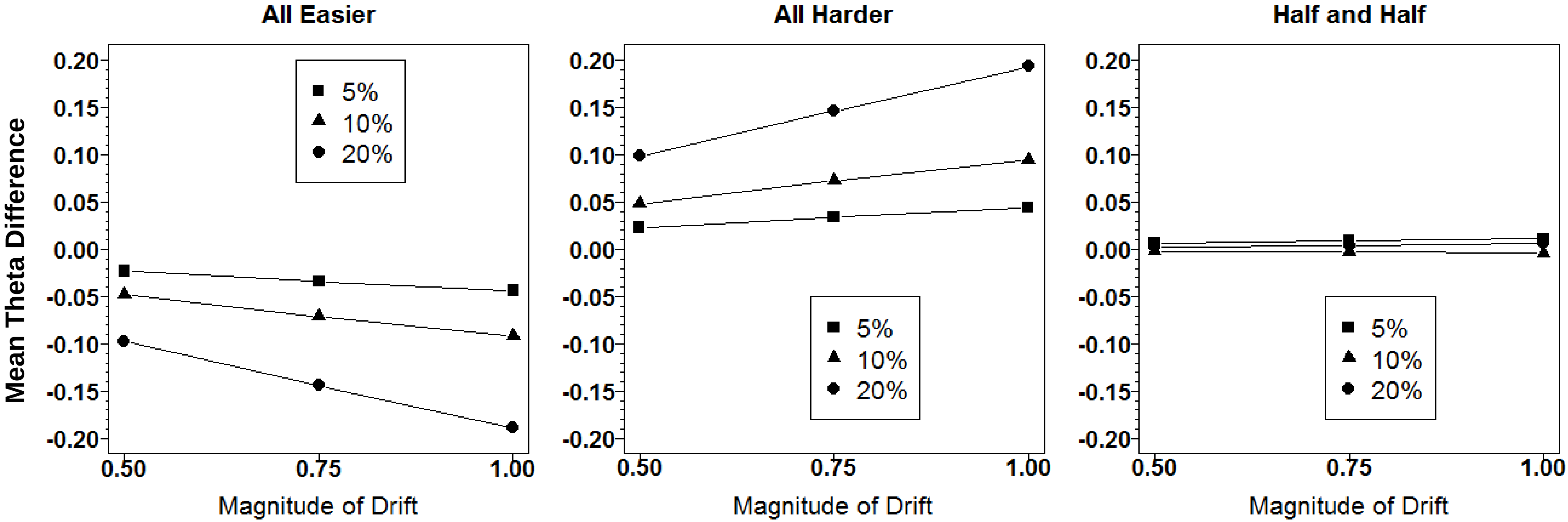
Percentage of Drifted Items on Individual Exams for Exam 1



Percentage of Drifted Items on Individual Exams for Exam 2

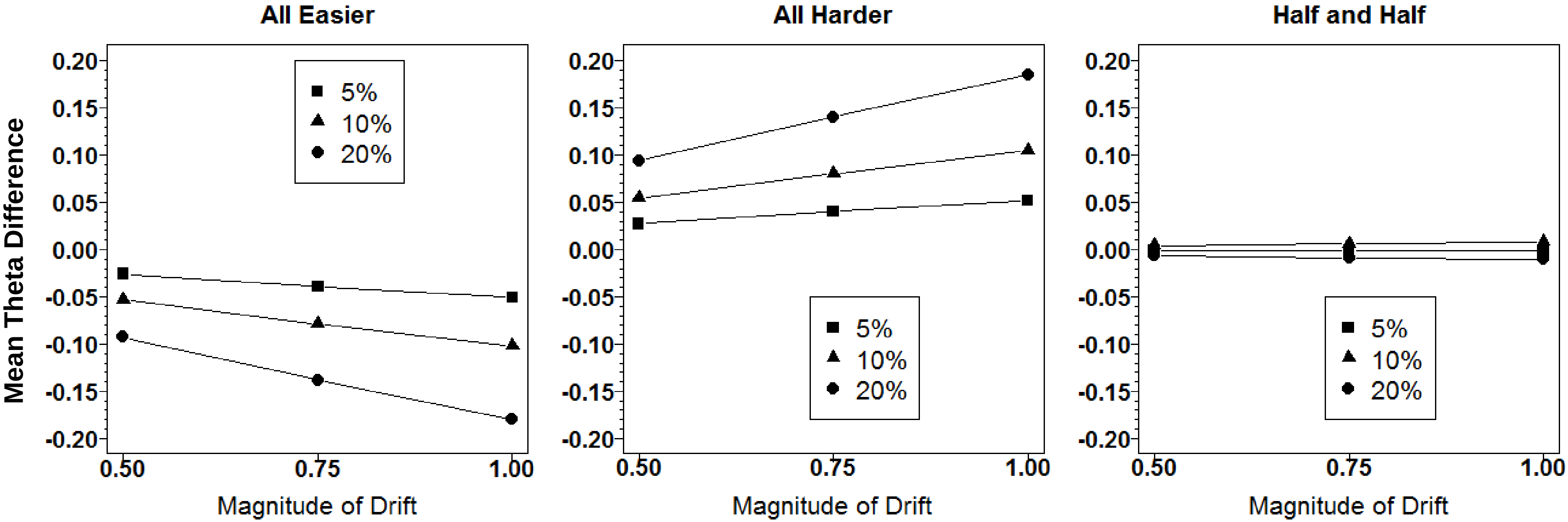


Mean Theta Differences: Exam 1



NCSBN
National Council of State Boards of Nursing

Mean Theta Differences: Exam 2

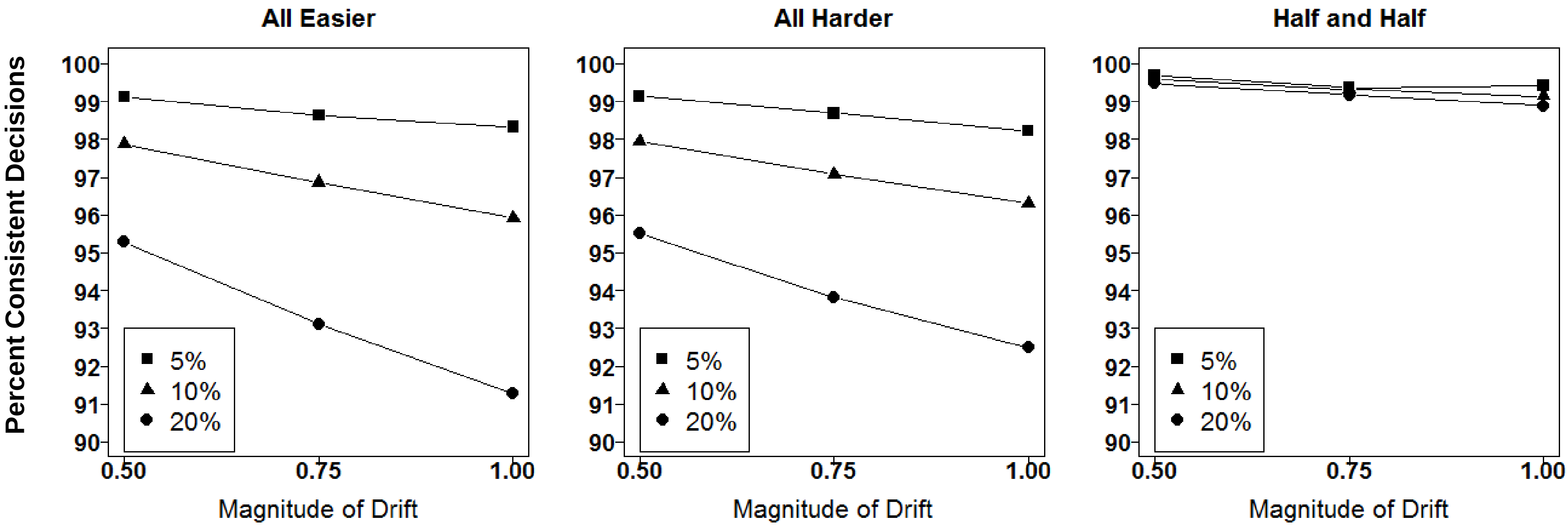


Theta Differences: Exam 1

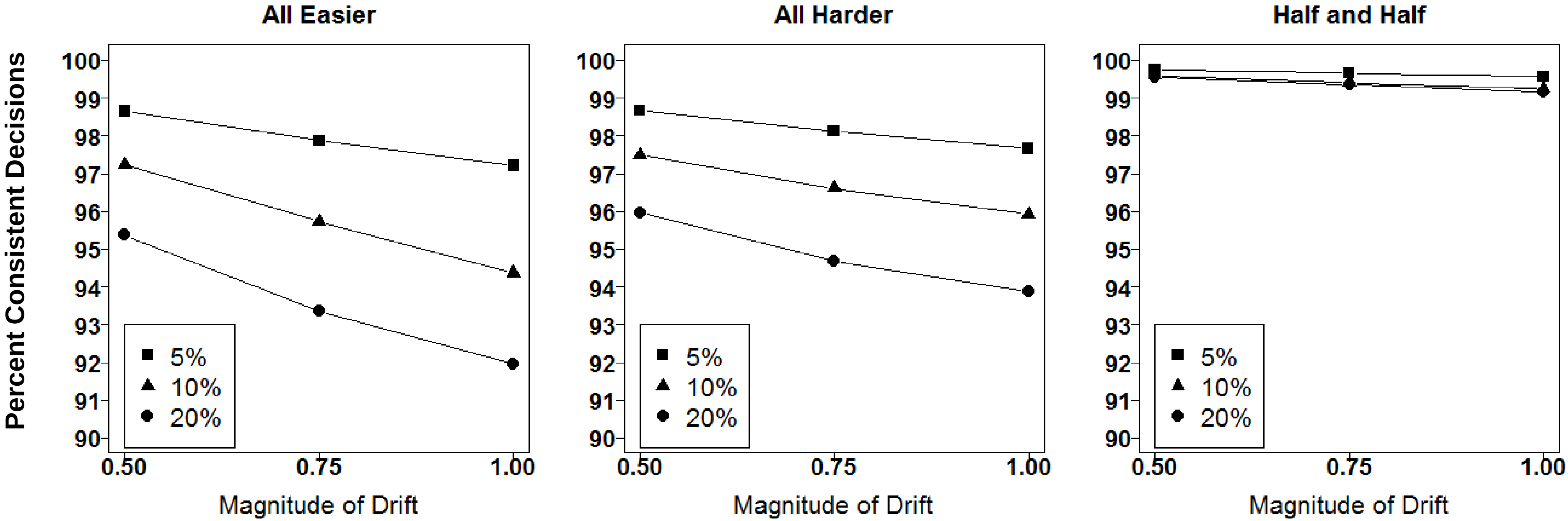
Direction of Drift	Magnitude (Logits)	All Easier		All Harder		Half and Half	
		Min	Max	Min	Max	Min	Max
5%	**0.50**	-0.10	0.00	0.00	0.09	-0.08	0.06
	0.75	-0.15	0.00	0.00	0.14	-0.13	0.15
	1.00	-0.20	0.00	0.00	0.18	-0.16	0.12
10%	**0.50**	-0.13	0.00	0.00	0.14	-0.07	0.08
	0.75	-0.20	0.00	0.00	0.21	-0.10	0.11
	1.00	-0.26	0.00	0.00	0.28	-0.14	0.15
20%	**0.50**	-0.20	-0.02	0.02	0.20	-0.10	0.12
	0.75	-0.29	-0.02	0.03	0.30	-0.16	0.18
	1.00	-0.39	-0.03	0.03	0.40	-0.22	0.24

NCSBN
National Council of State Boards of Nursing

Decision Consistency: Exam 1



Decision Consistency: Exam 2



Summary

- As the percentage of items increased or the magnitude of drift increased, differences in theta estimates also increased
- The largest difference in theta estimates was 0.40 logits for 20% with drift of 1.00 logits
- Decision consistency was greater than 95% for all conditions except 20% with drift of 0.75 or 1.00 logits

Discussion

- For large operational pools, candidate ability estimates appear robust to item drift, especially under conditions that may represent ‘normal’ amounts of drift
- Even with ‘extreme’ conditions of drift (e.g., 20% of items drifting 1.00 logits), decision consistency was still high

Limitations and Future Research

- Limitations
  - Recalibration study
  - Current study conducted on only variable-length CAT exams
- Future Research
  - Comparison with paper-and-pencil based tests
  - Simulation study
    - Replicate simulations of candidate response strings based on various drift conditions
    - Vary size of operational CAT item pool