

Automated Scoring to Detect and Control Reader Drift

**CCSSO Conference
June 19-23, 2010**

Susan Lottridge (Presenter)

Howard Mitzel (Presenter) – presenting for Matt Schulz

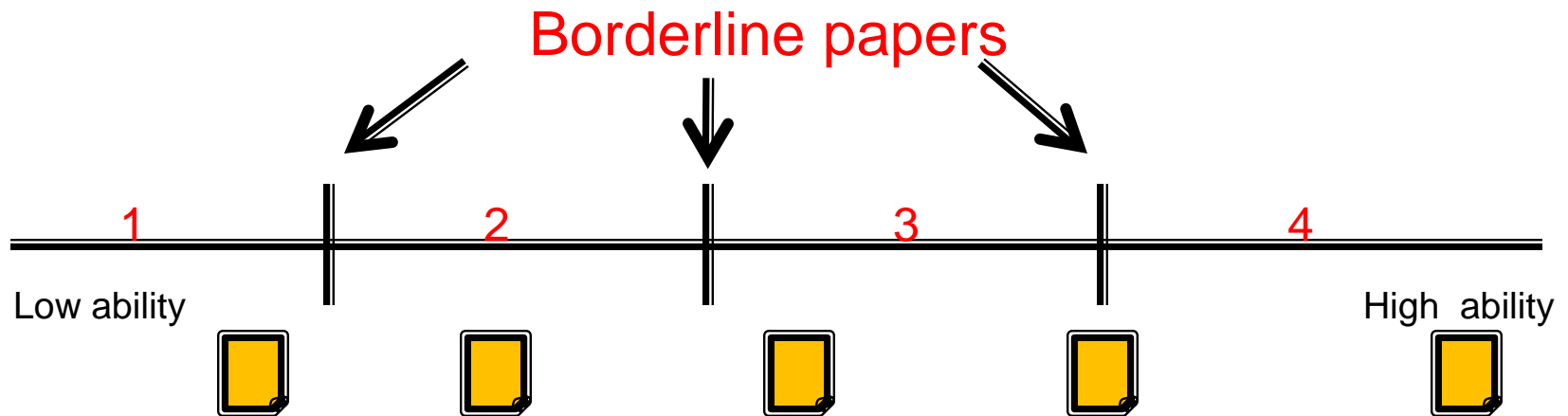
Wayne Camara (Discussant)

Agenda

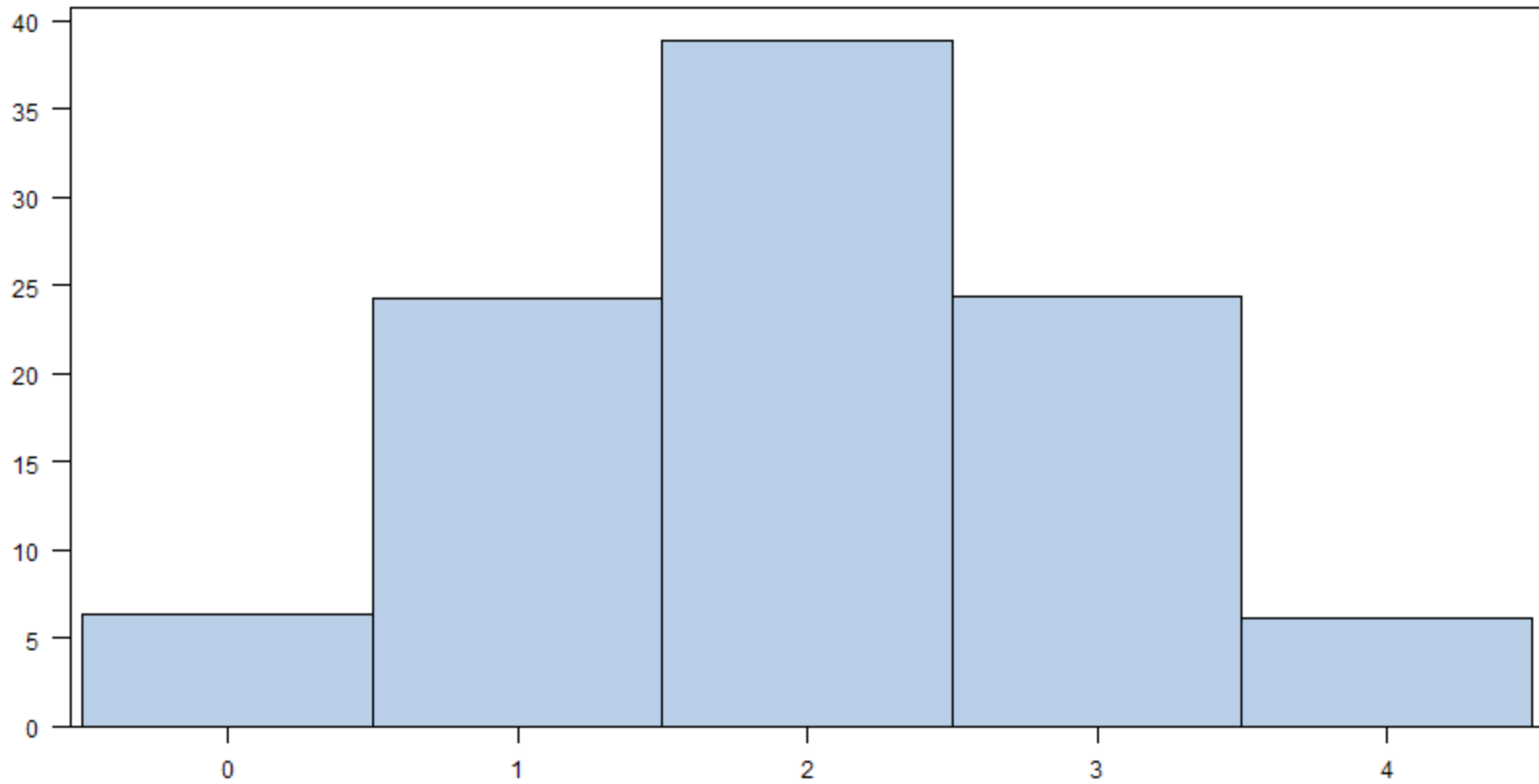
- What is drift and why does it occur?
- Evidence of drift from empirical data
- Using automated scoring to reduce drift
- Models for using AS to monitor and reduce drift

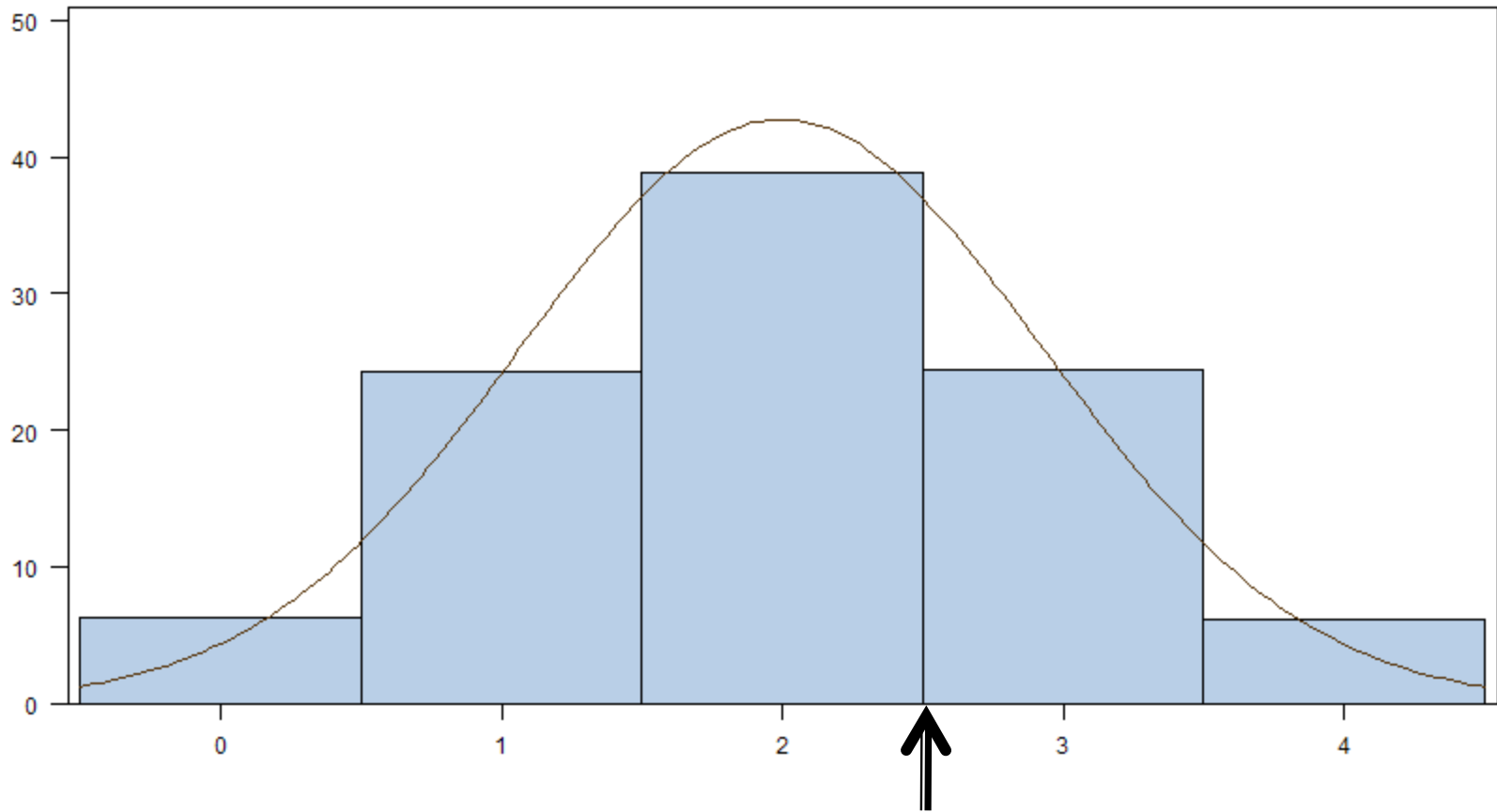
The scoring problem

- Students writing ability is continuous
- Readers are asked to categorize writing scores into scores



Writing Prompt Score Distribution



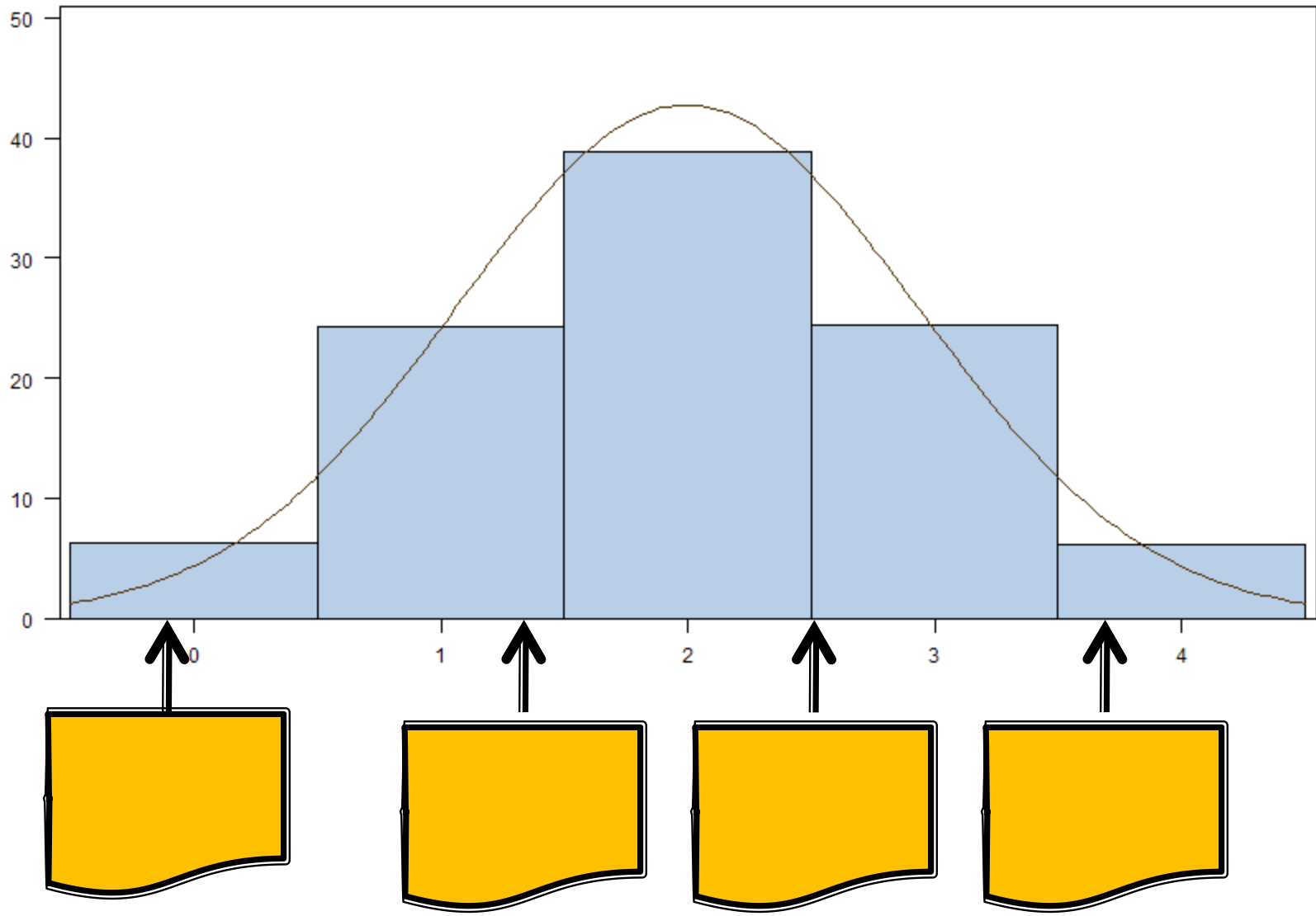


**Borderline
Paper**

Maximum Agreement Rate

| Scenario | Rater 1 | Rater 2 |
|----------|---------|---------|
| Agree | 2 | 2 |
| Adjacent | 2 | 3 |
| Adjacent | 3 | 2 |
| Agree | 3 | 3 |

Assuming two raters are independent, well-calibrated and observant, our expectation is (conservatively) a 50% exact agreement rate in this situation



- We should not expect agreement rates of 100%
- Agreement rates are influenced by the score distribution
- AS agreement rates with humans are often lower (3-5%) than human-human
- Hand-scoring agreement rates may be inflated due to drift

What is drift?

- Drift is the systematic bias in reader scoring
 - Severity/leniency
 - Avoidance/overuse of extreme categories
- We know from psychophysics that accurate judgments tend to also be relative (i.e., there is a context)
- It is dynamic, changing according to raters' internal schema and context factors

Drift and Scoring Writing

- The scoring of writing is a good example of where drift can occur:
 - Readers are expected to make score judgments based upon multiple criteria
 - Judgments are often made over long periods, and across administrations (year-to-year)
 - Substantial pressure exists to achieve high levels of reader agreement
 - There is pressure to score many responses quickly

Monitoring Drift

- Monitoring drift usually conducted with:
 - Two readers / ~10% read-behind
 - Seeded papers
 - Comparison to all readers
- Drift is very difficult to detect:
 - Lack of statistical power
 - Non-random assignment of papers to readers
 - Non-random testing (i.e., across window)
 - Social interaction between readers
 - Nature of drift is dynamic

Empirical Example of Drift

English Writing Assessment (Spring, 2009)

- Part of a state's testing program
- Students responded to a single prompt (6 prompts used)
- Forms spiraled
- Response was scored on two dimensions
- 4 point rubric (1 to 4) in each dimension
- Student score was the sum of two independent readers
 - 8 points on each dimension
 - 16 points total

Data

- Three-week online testing window
- About 31,000 responses
- Responses were divided into three time periods

Response Volume by Date

| | date | Cumulative Frequency | Cumulative Percent | Frequency | Percent |
|---------------|------|-------------------------|-----------------------|-----------|---------|
| Time 1 | 0429 | 65 | 0.21 | 65 | 0.21 |
| | 0430 | 60 | 0.20 | 125 | 0.41 |
| | 0501 | 1994 | 6.52 | 2119 | 6.93 |
| | 0504 | 3533 | 11.56 | 5652 | 18.49 |
| | 0505 | 3627 | 11.86 | 9279 | 30.35 |
| | 0506 | 2664 | 8.71 | 11943 | 39.07 |
| Time 2 | 0507 | 2248 | 7.35 | 14191 | 46.42 |
| | 0508 | 1587 | 5.19 | 15778 | 51.61 |
| | 0511 | 3374 | 11.04 | 19152 | 62.65 |
| | 0512 | 4120 | 13.48 | 23272 | 76.13 |
| | 0513 | 3460 | 11.32 | 26732 | 87.45 |
| Time 3 | 0514 | 1699 | 5.56 | 28431 | 93.01 |
| | 0515 | 472 | 1.54 | 28903 | 94.55 |
| | 0518 | 319 | 1.04 | 29222 | 95.59 |
| | 0519 | 479 | 1.57 | 29701 | 97.16 |
| | 0520 | 435 | 1.42 | 30136 | 98.58 |
| | 0521 | 149 | 0.49 | 30285 | 99.07 |
| | 0522 | 284 | 0.93 | 30569 | 100.00 |

Score Trends over Time

All Prompts, Both Raters, Both Dimensions Combined

| Score | Percent by Time | | |
|---------------------------------|-----------------|------|------|
| | 1 | 2 | 3 |
| 0 | 2% | 3% | 4% |
| 1 | 5% | 4% | 6% |
| 2 | 29% | 26% | 25% |
| 3 | 49% | 57% | 55% |
| 4 | 15% | 11% | 10% |
| P-value | 0.68 | 0.68 | 0.65 |
| InterRater Exact Agreement: | 67 | 74 | 79 |
| IntraRater Dimension Agreement: | 95 | 98 | 99 |
| MC total: | 22.2 | 22.5 | 21.6 |

Use of Automated Scoring (AS)

- Because AS cannot drift, it can be used to examine the extent of human reader drift
- Score distributions and results that follow are for one of the six prompts
- AS engine was trained using a sample of data from Time 1
- Results are representative

Results for Dimension 1

R1 = Reader 1 AS = Automated Scoring

| Score | Percent by Time and Source | | | | |
|----------------------------|----------------------------|------|--|--------|------|
| | Time 1 | | | Time 3 | |
| | R1 | AS | | R1 | AS |
| 1 | 4 | 4 | | 5 | 8 |
| 2 | 28 | 28 | | 24 | 30 |
| 3 | 53 | 53 | | 60 | 49 |
| 4 | 15 | 15 | | 11 | 13 |
| Mean Essay Score | 2.78 | 2.78 | | 2.76 | 2.65 |
| Mean MC Score: | 22.83 | | | 21.85 | |
| Correlation with MC Score: | 0.43 | 0.43 | | 0.46 | 0.49 |
| Agreement with Reader 2: | 72% | 68% | | 82% | 67% |

- Results were similar
 - for other dimensions
 - other prompts
 - in comparisons across administrations
(December to May)

Summary

- AS provides consistent results over time
 - Results are consistent with MC data
- Reader 1 is more likely to agree with Reader 2 than with AS but this is associated with
 - drift to middle of the rubric (3's)
 - lower correlation with MC scores

Hypotheses

- Drift may be indicative of rater behavior problems:
 - raters tend to avoid highest/lowest scores in order to meet unrealistically high expectations for exact agreement
 - scoring becomes overly focused on features that are easily identified at the expense of valid, but more complex features of writing

Recommendations

- Automated and human scoring may complement one another:
 - Humans provide judgment to a complex task
 - Machines provide consistency in scoring within and across administrations

Using Automated Scoring to Monitor and Reduce Drift

Automated Scoring as Monitor

- Perceived Benefits in Scoring
 - Consistency
 - Per-response labor costs removed, allowing for 100% read-behind
 - Uses a continuous variable to represent writing skill
- Perceived Drawbacks in Scoring
 - Outlier responses will not be well-scored
 - Computer cannot provide judgment

Results of Drift Analysis

- The State agreed to remove one human reader and replace with the automated scorer
- The State was resistant to allowing computer score to be used in score of record
- The AS score was used only for monitoring/read-behind purposes

Steps Involved

- Identify training sample
- Train engine
- Determine processing rules
- Work with hand-scoring vendor
- Monitor during administration
- Analyze data (post-administration)

Training sample

- The choice of training sample is critical
- Consider:
 - Data on which items already calibrated
 - Representativeness
 - Any issues with human reader bias/error
- Training sample choice will shape scoring
 - Expert-only reads vs. ‘typical’ reads
 - Time period

Impact of Training Sample Scoring of December 2009 Data

| | R1 | AS-T1 | AS-W |
|-----------------------------------|------|-------|------|
| Percent in Score Category | | | |
| 1 | 6% | 6% | 4% |
| 2 | 32% | 29% | 26% |
| 3 | 52% | 51% | 57% |
| 4 | 10% | 14% | 14% |
| Mean | 2.66 | 2.74 | 2.81 |
| SD | .74 | .77 | .71 |
| Agreement with second human rater | | | |
| Exact Agreement | 75% | 67% | 67% |
| Adj. Agreement | 25% | 33% | 33% |
| Correlation with MC | .50 | .49 | .46 |

R1: Human Reader

AS- T1: Automated Scoring system trained on time period 1

AS-W: Automated Scoring system trained on entire May 2009 window

Revise Processing Rules

- Second human reader replaced by automated scorer
- Resolution conducted on non-adjacent scores
- Only single human score used in score of record

Hand-Scoring Vendor

- Change in processing rules required changes for sub-contractor:
 - Reader training
 - Modifications to reports
 - Buy-in
 - Resolution rule changes

Monitoring During Administration

- Table leaders reviewed score distribution reports and rater agreement reports against the automated scoring
 - Reader agreement was computed between human reader and AS
 - Reader score distribution compared to all AS results (all readers)

Post-Administration Analysis

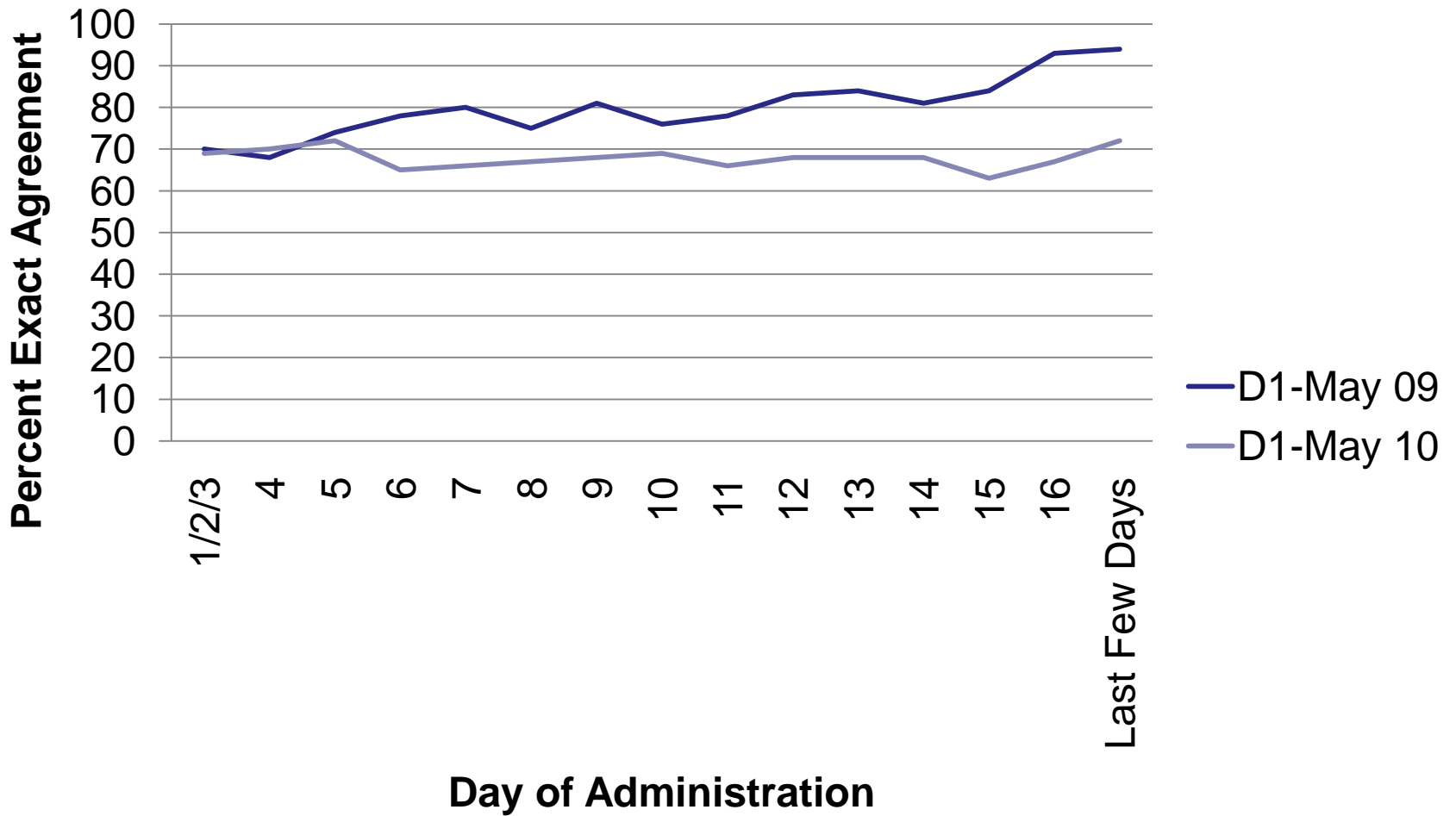
- Did computer scoring change the human scoring?
- Was the observed drift in previous spring administration reduced?

May 2009 and 2010 Scores

- No large changes in scoring

| Score | May, 2009 (n=5,188) | | May, 2010 (n=35,064) | |
|------------------------|------------------------|------|-------------------------|------|
| | R1 | AS | R1 | AS |
| 1 | 4% | 6% | 4% | 5% |
| 2 | 26% | 30% | 25% | 28% |
| 3 | 57% | 51% | 55% | 51% |
| 4 | 13% | 14% | 16% | 15% |
| Mean | 2.79 | 2.73 | 2.83 | 2.76 |
| SD | 0.71 | 0.77 | 0.74 | 0.77 |
| Agreement | | | | |
| Exact | 77% | 67% | n.a. | 67% |
| Adj | 23% | 32% | | 33% |
| Correlation with MC | 0.45 | 0.47 | 0.50 | 0.52 |

Daily Exact Agreement Rates



May 2010 Scores, by Time

| | Time 1 (n=12,895) | | Time 2 (n=12,830) | | Time 3 (n=9,339) | |
|--|----------------------|------|----------------------|------|---------------------|------|
| | R1 | AS | R1 | AS | R1 | AS |
| Percent in Score Category | | | | | | |
| 1 | 4 | 4 | 4 | 5 | 6 | 8 |
| 2 | 24 | 27 | 23 | 28 | 28 | 31 |
| 3 | 55 | 53 | 57 | 51 | 54 | 49 |
| 4 | 18 | 16 | 17 | 16 | 12 | 11 |
| Mean Essay Score | | | | | | |
| Mean | 2.87 | 2.82 | 2.87 | 2.78 | 2.72 | 2.66 |
| Agreement between R1 and AS | | | | | | |
| Exact | 66% | | 67% | | 68% | |
| Relationship to Multiple Choice Score | | | | | | |
| r_{MC} | .51 | .52 | .50 | .52 | .49 | .51 |
| Mean | 24.11 | | 23.31 | | 22.27 | |

May 2009/2010 Comparisons

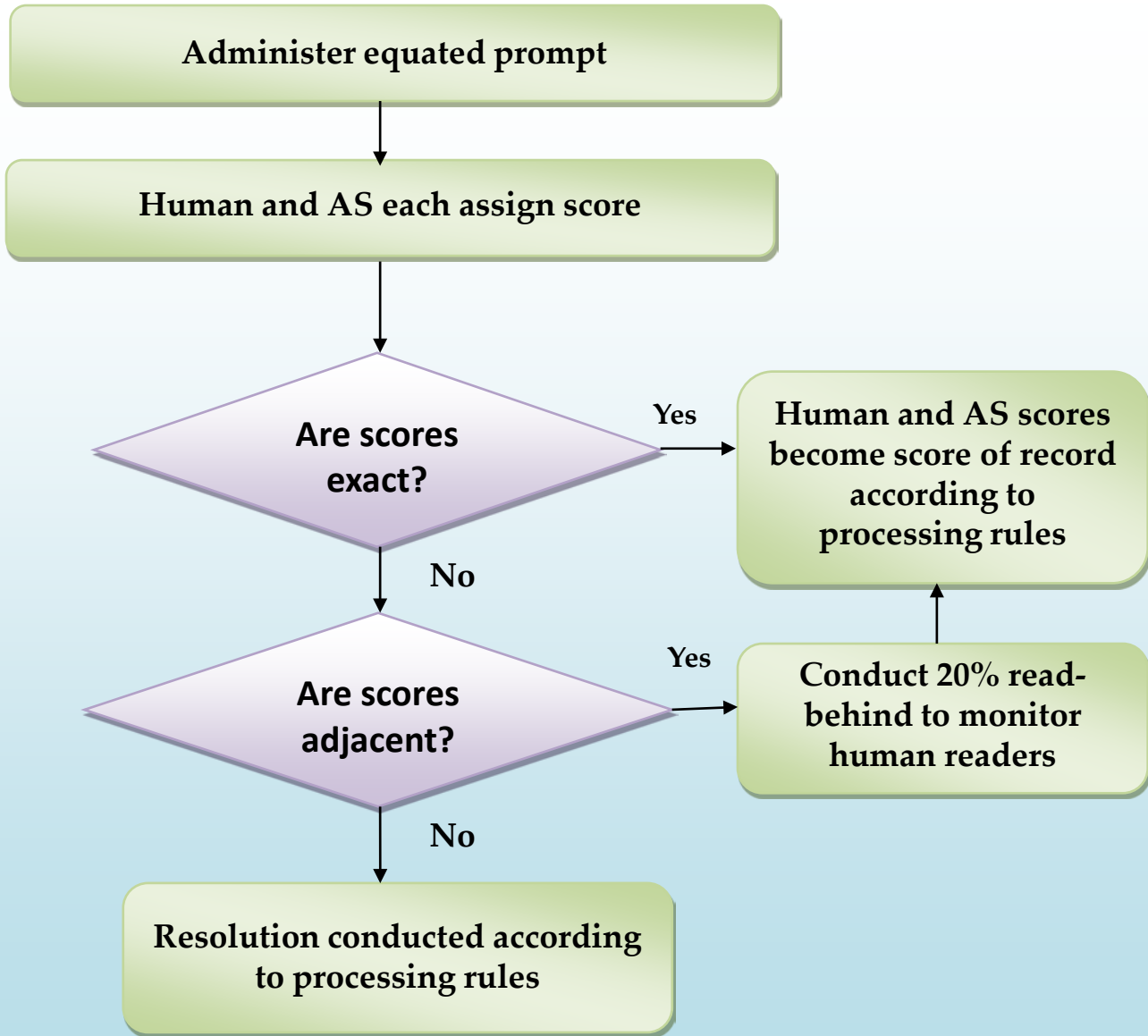
| Essay Score | May 2009 | | | May 2010 | | |
|-------------------------|----------|--------|--------|----------|--------|--------|
| | Time 1 | Time 2 | Time 3 | Time 1 | Time 2 | Time 3 |
| 1 | 4 | 3 | 4 | 4 | 4 | 6 |
| 2 | 28 | 23 | 24 | 24 | 23 | 28 |
| 3 | 53 | 62 | 60 | 55 | 57 | 54 |
| 4 | 15 | 12 | 11 | 18 | 17 | 12 |
| Mean Essay Score | 2.78 | 2.83 | 2.76 | 2.87 | 2.87 | 2.72 |
| MC Total Mean | 22.83 | 23.19 | 21.85 | 24.11 | 23.31 | 22.27 |
| MC Correlation | .43 | .46 | .46 | .51 | .50 | .49 |
| Agreement Rate | 68% | 78% | 82% | 66% | 67% | 68% |

Summary

- **Benefits:**
 - Reduced reader drift across administration
 - Takes two to drift?
 - Reduced number of reads in half
 - No significant changes to score distributions, and improved validity
- **Issues:**
 - Reactions of human readers
 - Identifying problematic raters still difficult due to available reports

Models for using AS in Scoring

- 100% read-behind/second read
 - AS score is NOT included in score of record
 - AS score is included in score of record
 - Resolution rules
- AS score presented to human reader
 - Used as basis for scoring
 - Human reader can retain score or change score



Monitoring using AS

- May require shift in how hand-scoring vendors operate
 - Comparisons made to single scorer (AS) rather than group of raters
 - Requires trust in computer scoring
 - Possible increase in table leader monitoring
- Can compute reader agreements and score distributions on each reader
 - Differences isolated to AS/reader scoring behavior

Thank you!