Pearson

# Spring 2020 Post-Equating Evaluation Plan

## Presentation to the Illinois Technical Advisory Committee

28 February 2020

# Post-Equating Evaluation Rationale

- The purpose of the post-equating evaluation in spring 2020 is to evaluate the stability of the pre-equated item statistics and pre-equated raw score to scale score tables.
- Spring 2019 was the first year for pre-equating the ELA assessments and the first year for administering the Alternative Blueprinting Option, ABO, shortened forms.
- Post-equating evaluation in spring 2019 identified items with pre-equated item statistics that shifted in difficulty and/or discrimination after post-equating which resulted in shifts in the performance level distributions.
- Guidance from the TAC is needed to determine criteria for implementing the post-equated scoring tables over the pre-equated scoring tables.

# Spring 2020 Administration Windows

| Administration Windows | Start | End | Form |
|---|---:|---:|---:|
| Illinois: ELA grades 3–8 | 3/11/20 | 4/24/20 | ABO |
| DoDEA: ELA grades 3–8, 10 | 3/30/20 | 5/8/20 | ABO |
| District of Columbia: ELA grades 3–11 | 4/6/20 | 5/22/20 | Flagship |
| New Jersey: ELA grades 3–10 | 4/20/20 | 6/8/20 | ABO |

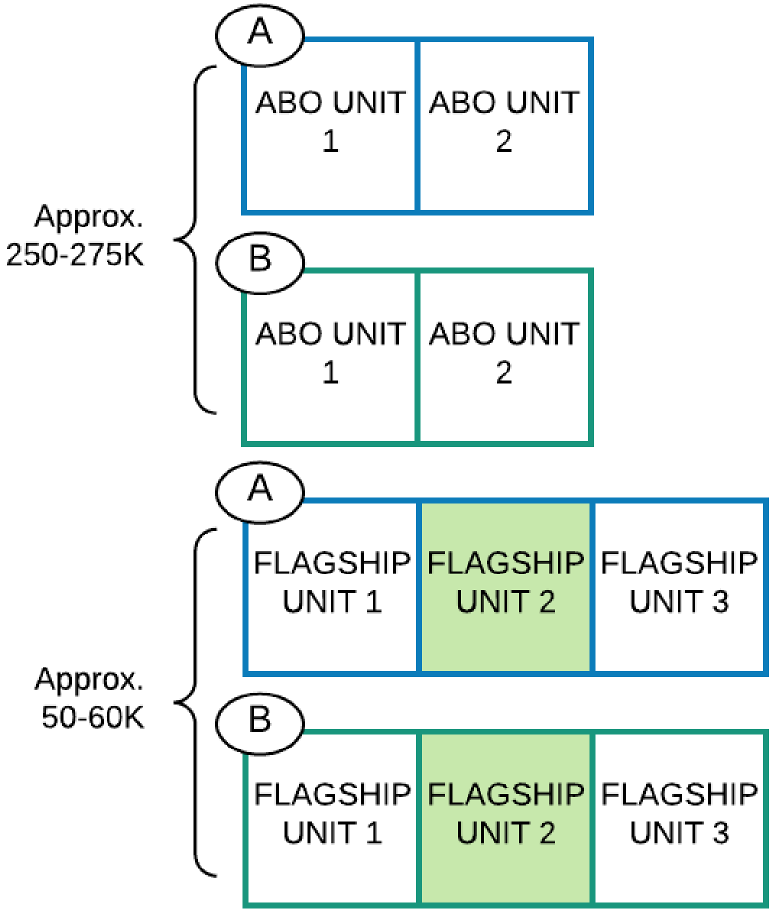
- Majority of students will take the ABO forms (approx. 120,000 per form)
- Fewer students will take the Flagship forms (approx. 3,000 per form)

Due to the number of students testing across forms in 2020, post-equating analyses will need to link through the ABO forms. The common items for linking will need to be placed on the ABO forms rather than the Flagship forms for spring 2020.

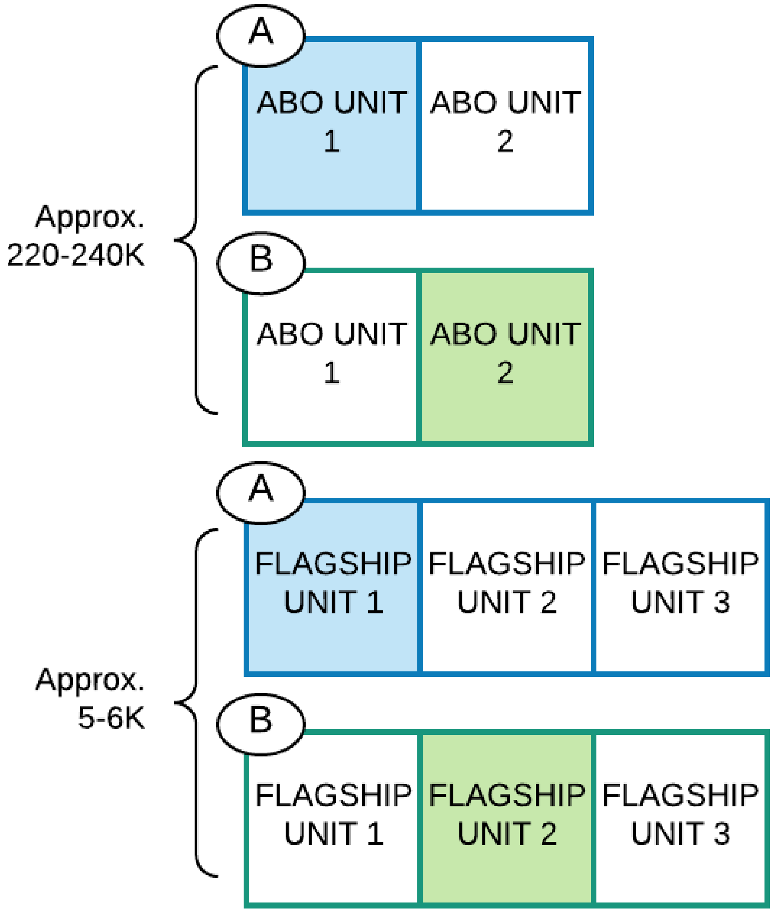
# ELA/L Common Items for Spring 2019 and Spring 2020 Equating Design

| Common Items 2019 | Common Items 2020 |
|---|---|
| • Flagship forms had common items across forms | • Flagship forms may not have items in common across forms |
| • ABO forms were a subset of the Flagship forms | • ABO forms may not be a subset of the Flagship forms |
| • ABO forms may not have items in common across forms | • ABO forms may not have items in common across forms |
| • Flagship forms had prior operational items for linking to the item bank scale | • ABO forms have prior operational items for linking to the item bank scale |
| • Number of student sufficient to post-equated ABO and Flagship | • Number of students may not be sufficient to post-equate Flagship |

# Spring 2019 Linking Design



# Spring 2020 Linking Design



# Spring 2020 Test Construction

- Online 1 and Online 2 ABO forms require one unit that has prior operational use in order to have stable common items for linking
- All Flagship forms and Accommodated ABO forms require prior operational sets or field test sets in common with the Online 1 and Online 2 ABO forms due to low student populations.

# Spring 2020 Post-Equating Evaluation Considerations

- Pearson will perform the post administration evaluation analyses.
- Hand-scoring and training by the Intelligent Essay Assessor (IEA) for scoring open-ended items need to be completed for the post-equating sample.
- HumRRO is recommended as the third-party replicator, if needed.
- Clearly defined criteria for selecting post-equating is recommended prior to the administration.

# Sampling Considerations

- The samples will consist of ABO and Flagship data. Only items on the Flagship in common with the ABO will have item parameters updated during post-equating.
- For Flagship only items, the pre-equated item statistics, based on prior operational data or field test data, are based on larger number of students.
- Due to differences in administration windows, the majority of Illinois student data will be available when New Jersey student data begins processing.
- It’s recommended that multiple states be included in the sample to reflect prior consortium-based analyses.
- Illinois student data will be reduced to be consistent with prior representation of Illinois data in the post-equating samples.

# Sampling by Grade Level

- An early sample of 25-30% of the ABO administrations for grades 3-8 are recommended.
- For ELA grades 3 – 8, New Jersey DOE recommends pulling New Jersey student data once 30% of New Jersey students have completed testing.

# Post-Equating Evaluation Process

- Sampling
  - Pull an early sample of spring 2020 student data. The majority of student data will consist of Illinois and New Jersey students. A small portion of students from the Department of Defense Education Activity (DoDEA) and District of Columbia will be included.
- Item-Level Analysis
  - Conduct classical item analysis and IRT calibrations.
- Test-Level Analysis
  - Generate scoring tables based on the post-equating sample IRT calibrations.
- Evaluation
  - Compare the pre-equated and post-equated item-level analysis and test-level analysis.

# Sampling Process

- Using spring 2019 data, sampling targets will be estimated based on state representation, demographic groups, and prior year performance level distributions.
  - Only states or agencies participating in spring 2020 are included.
- Due to administration windows, the majority of student data available from DoDEA, District of Columbia, and New Jersey will be included in the post-equating evaluation samples.
  - Comparison to the sampling targets may result in a small reduction in the student data.
- For Illinois, samples will be pulled from the student data that represent the demographic and performance data for Illinois in spring 2019.
  - The sample size will be proportional to the data available from the other states such that Illinois representation reflects the spring 2019 proportion.

# Prior Sampling Research

- Using spring 2016 data from eight states,
  - sampling analyses were conducted for all grades and assessments. A baseline data set (all students) and four sample data sets representing the first 25%, 30%, 40%, and 50% of the baseline data based on administration date were evaluated.
  - post-equating analyses were conducted on a selection of assessments that represented the grade bands. The sample item parameter estimates were compared to the baseline item parameter estimates. The raw score to scale score conversion files were compared for meaningful differences on the reported scale score and the performance level categories.
- The early post-equating sampling research report is provided for supplemental information.

  *Early Post-Equating Sampling Research Report Final_03242016_Approved.pdf*

# Prior Sampling Research

- An ANOVA identified the demographic variables tending to explain more of the variability in the summative scale scores than other variables.
    - Based on the minimum post-equating sample identified as sufficient for score reporting, criteria associated with the ANOVA were established for each demographic variable.
- Of the eight states scheduled to participate in the spring 2016 administrations, five states were found to be consistently represented in each of the early equating samples (25%, 30%, 40%, and 50% samples). The proportions tended to be within 11% of the baseline.
    - Based on the minimum post-equating sample identified as sufficient for score reporting, criteria associated with difference between the baseline and equating sample were established.
- The following slides summarize the established criteria and are proposed for the spring 2020 early post-equating sampling for the ABO assessments, where relevant.

# Sampling Criteria ELA/L Grades 3 – 8

- Post-equating sampling criteria based on prior criteria implemented in spring 2017 and spring 2018.

| ELA/L Grades 3 – 8 | |
|---|---|
| **Demographic Variables** | **Online Criteria** |
| Student with Disabilities (Yes) | Proportions within 3% of the 2019 distributions |
| Economically disadvantaged (Yes) | Proportions within 6% of the 2019 distributions |
| Ethnicity | Effect size less than .15 of the 2019 distributions for each group |
| **Sample Size** | |
| Overall N Count | Minimum 50,000-75,000 (25-30%) |
| Per Form | Average 22,000 |
| Per Item | Minimum 5,000 |
| **State** | |
| At least four | Proportions within 12% of the 2019 distributions |
| **Prior Scores** | |
| Summative Scale Score | Percent difference in CDF within 3% of 2019 |
| Performance Level | Proportions within 3% of the 2019 distribution |

# Item-Level Analysis

- Compare pre-equated and post-equated classical item level statistics including item mean scores, item-to-total-score correlations, and score point distributions.
- Compare pre-equated and post-equated IRT parameters for the common items and for all items. Determine the correlation between the IRT parameters and evaluate scatter plots. Evaluate item fit plots.

# Test-Level Analysis

- Compare pre-equated and post-equated raw score to scale score tables including any differences in the associated scale scores.
- Compare the test characteristic curves.
- Compare performance level scale score cuts for pre-equated and post-equated score tables.
- Evaluate impact data for the pre-equated and post-equated score tables. Determine the percent of students in each performance level and percent of students with different performance levels based on the equating tables.

# Evaluation Criteria

- Differences are expected between pre-equating and post-equating scoring tables due to larger sample sizes being available for post-equating.
- Dorans and Feigenbaum (1994) described a “difference that matters” criteria as scale score differences greater than half a scale score point (0.5) which would round to a different scale score value.
- In spring 2019, differences in the percent of students at the college and career readiness cut (Level 4) were evaluated.
  - Evaluating the raw score, DTM identified values of 0 to 1 raw score points for the majority of the ELA/L grades 3 – 11 performance level cuts. ELA/L grades 4, 9, and 10 had at least one performance level cut that shifted by 2 raw scores.

Dorans, N. J., & Feigenbaum, M. D. (1994). Equating issues engendered by changes to the SAT and PSAT/NMSQT (ETS Research Memorandum No. RM-94-10). Princeton, NJ: ETS.

# TAC Questions

- Does the TAC have feedback regarding the sampling process?
- Does the TAC have guidance for evaluating the item-level and test-level comparisons between the pre-equated and post-equated analyses?
- Does the TAC have a recommendation for the criteria to use to determine if the post-equating tables should be implemented?
- What evaluation criteria or guidelines does the TAC suggest using to inform the decision about whether to implement post-equating tables by individual grade levels vs. for all grade levels?

# Spring 2019 Evaluation

- The following tables provide Test-Level results from the spring 2019 evaluations to provide context as criteria for spring 2020 are discussed.

# Pre and Post- Equated Raw Score Change

| Test | Form | Level 2 Change in Raw Score | Level 3 Change in Raw Score | Level 4 Change in Raw Score | Level 5 Change in Raw Score |
|---|---|---|---|---|---|
| ELA03 | ABO1 | 0 | 0 | **0** | -1 |
| | ABO2 | 0 | 0 | **0** | 0 |
| ELA04 | ABO1 | 0 | -1 | **-2** | -2 |
| | ABO2 | 0 | 0 | **1** | 0 |
| ELA05 | ABO1 | 0 | -1 | **-1** | -1 |
| | ABO2 | -1 | 0 | **-1** | 0 |
| ELA06 | ABO1 | -1 | -1 | **-1** | -1 |
| | ABO2 | 0 | -1 | **0** | 0 |
| ELA07 | ABO1 | -1 | -1 | **-1** | -1 |
| | ABO2 | -1 | -1 | **-1** | -1 |
| ELA08 | ABO1 | 0 | 0 | **1** | 1 |
| | ABO2 | -1 | -1 | **-1** | 0 |

Raw Score Change = Pre Raw Score – Post Raw Score

Pearson

# Spring 2019 ABO ELA/L 3 – 5: Pre- and Post-Equating Test Level Summary

| Test | Pre. Vs Post Equating | Form | Total | Scale Score Mean | Scale Score Standard Deviation | CCR* Count | CCR Percent | ABO 1 - ABO 2 CCR Percent | Pre-Equated - Post-Equated CCR Percent* |
|---|---|---|---|---|---|---|---|---|---|
| ELA03 | PRE | ABO1 | 51,858 | 734.03 | 40.76 | 51858 | 35.9 | | |
| | | ABO2 | 51,494 | 734.03 | 38.87 | 51494 | 34.5 | 1.4 | |
| | POST | ABO1 | 51,858 | 732.34 | 39.33 | 51858 | 35.9 | | 0.0 |
| | | ABO2 | 51,494 | 733.52 | 39.22 | 51494 | 34.5 | 1.4 | 0.0 |
| ELA04 | PRE | ABO1 | 59,892 | 738.60 | 37.37 | 59892 | 40.7 | | |
| | | ABO2 | 71,968 | 735.82 | 35.50 | 71968 | 34.5 | 6.2 | |
| | POST | ABO1 | 59,892 | 735.97 | 35.59 | 59892 | 35.5 | | 5.2 |
| | | ABO2 | 71,968 | 736.41 | 35.62 | 71968 | 37.2 | -1.7 | -2.7 |
| ELA05 | PRE | ABO1 | 79,929 | 739.04 | 33.86 | 79929 | 39.3 | | |
| | | ABO2 | 59,330 | 738.16 | 34.30 | 59330 | 37.5 | 1.8 | |
| | POST | ABO1 | 79,929 | 736.59 | 33.09 | 79929 | 36.7 | | 2.6 |
| | | ABO2 | 59,330 | 737.26 | 33.84 | 59330 | 34.9 | 1.8 | 2.6 |

*CCR: College and Career Readiness Cut

Pearson

# Spring 2019 ABO ELA/L 6 - 8: Pre- and Post-Equating Test Level Summary

| Test | Pre. Vs Post Equating | Form | Total | Scale Score Mean | Scale Score Standard Deviation | CCR* Count | CCR Percent | ABO 1 - ABO 2 CCR Percent | Pre-Equated - Post-Equated CCR Percent* |
|---|---|---|---|---|---|---|---|---|---|
| ELA06 | PRE | ABO1 | 69,577 | 738.65 | 31.75 | 69577 | 36.8 | | |
| | | ABO2 | 68,320 | 735.93 | 31.46 | 68320 | 33.8 | 3.0 | |
| | POST | ABO1 | 69,577 | 736.21 | 32.57 | 69577 | 34.8 | | 2.0 |
| | | ABO2 | 68,320 | 735.80 | 32.14 | 68320 | 33.8 | 1.0 | 0.0 |
| ELA07 | PRE | ABO1 | 68,286 | 738.99 | 37.44 | 68286 | 40.6 | | |
| | | ABO2 | 69,607 | 741.65 | 38.84 | 69607 | 43.0 | -2.4 | |
| | POST | ABO1 | 68,286 | 735.86 | 37.96 | 68286 | 38.3 | | 2.3 |
| | | ABO2 | 69,607 | 739.76 | 38.92 | 69607 | 40.7 | -2.4 | 2.3 |
| ELA08 | PRE | ABO1 | 69,443 | 738.61 | 38.10 | 69443 | 39.4 | | |
| | | ABO2 | 69,638 | 739.33 | 38.51 | 69638 | 41.2 | -1.8 | |
| | POST | ABO1 | 69,443 | 739.13 | 39.36 | 69443 | 41.6 | | -2.2 |
| | | ABO2 | 69,638 | 738.27 | 39.45 | 69638 | 39.0 | 2.6 | 2.2 |

CCR: College and Career Readiness Cut

Pearson

Pearson