AZELLA Reassessment 2018 Stages III-V Mode Comparability Report

June 2018

Produced by Pearson in association with the Arizona Department of Education Assessment Section.

AZELLA Reassessment 2018 Stage III-V Mode Comparability Study

Hirotaka Fukuhara and Lietta Scott

June 2018

Fukuhara, H., & Scott, L. (2018). *AZELLA Reassessment 2018 Stages III-V mode comparability report.* Phoenix, AZ: Arizona Department of Education, Assessment Section.

Executive Summary

A mode comparability study was performed on the Spring 2018 AZELLA Reassessment, Arizona's English learner language proficiency assessment, as this was the first online administration of Stages III through V of the assessment (Grades 3-12). A logistic regression model for differential item functioning (DIF) with a purification procedure was used on the anchor items to investigate any shift in item performance due to the mode change. The DIF analysis revealed that the anchor set was stable for the test administration transition from paper-and-pencil to online. These results were confirmed through DIF analysis using the Mantel-Haenszel method. However, Arizona's standard equating process, displacement by Winsteps (Linacre, 2015), did flag some of the anchor items. Those flagged items were freely calibrated along with all other non-anchor core items by using a final stable anchor set so that the resulting item parameter estimates were on the base scale set in 2013. After equating was complete, an impact analysis was conducted as a check for reasonableness. The impact analysis showed no evidence of shift in student performance due to the mode change.

Background

Beginning Spring 2018, the Arizona Department of Education (ADE) moved from a paper-based administration Arizona's English learner language proficiency reassessment (AZELLA Reassessment) for Stages III through V (Grades 3 through 12) to an online mode. This move for the upper grades was made while maintaining a paper-based administration for Stages I and II (students in Kindergarten through Grade 2). For students with disabilities, whose Individualized Education Program (IEP) indicates they cannot participate in an online assessment, a paper version of the AZELLA Reassessment test was made available for the small number of students in grades 3 and above expected to need this accommodation. The online mode for Stages III through V, with the paper accommodation for students who need it, will be continued for both Placement and Reassessment administrations starting with school year 2018-2019.

One psychometric concern, on transitioning from a paper-and-pencil based assessment to an online assessment is mode comparability. To explore any shift in item and/or scale performance due to this transition, Arizona's Technical Advisory Council (TAC) made several procedural recommendations at their November 2016 meeting. To examine item level changes across modes, they recommended the use of a logistic regression differential item functioning (DIF) method with an iterative purification process by an R package, lordif (Choi, Gibbons, & Crane, 2011) to identify stable anchor items between modes. These stable anchor items would then, in turn, be used to equate the Spring 2018 online forms back to the AZELLA base scale for comparable scale reporting. This report documents the processes and results of both the DIF analyses and the equating analyses performed on the online AZELLA Reassessment data in April 2018. The purpose of this work was to ensure that the student scores and performance levels reported in May 2018 were comparable to those received in prior years when the assessment was administered via paper-and-pencil.

Mode Comparability Method

Since the vast majority of students (as expected, all but 4) took an online form in Spring 2018 the only available comparison group who took a paper test for a mode comparability study were students who took the same items during Spring 2017. Therefore, the study was conducted by combining student responses from Spring 2017 with those from Spring 2018. The items identified as having the highest probability for stability (and therefore chosen as most likely to be stable anchors) were those in a multiple-choice format and that were administered in the same position on both the Spring 2017 and 2018 forms. DIF analyses were performed on these anchor items that were identified, and noted within the 2018 form test-maps, prior to the 2018 administration. The lordif (Choi et al., 2011) program ran the following ordinal logistic regression models:

Model 1 (Impact Model): $logit\ P(u_i \geq k) = \alpha_k + \beta_1\theta_j$, Model 2 (Uniform DIF Model): $logit\ P(u_i \geq k) = \alpha_k + \beta_1\theta_j + \beta_2 G_j$, and

Model 3 (Non-uniform DIF Model): $logit\ P(u_i \geq k) = \alpha_k + \beta_1\theta_j + \beta_2 G_j + \beta_3\theta_j G_j$,

where $P(u_i \geq k)$ expresses the cumulative probabilities of a response on item i falling into k category or higher, θ_j is an ability of student j, and G_j is a group indicator ($G_j = 1$ if a student took a paper-and-pencil assessment, $G_j = 0$ otherwise).

The purpose of this analysis was to evaluate if an item manifested any type of DIF due to a change in mode by comparing a model fit on Model 3 (Non-uniform DIF model) and Model 1 (Impact model). A default method to compare the model fit by lordif (Choi et al., 2011) was the likelihood ratio χ^2 test with *df*=2. However, the likelihood ratio χ^2 test is known to be sensitive to a large sample size. In fact, the smallest sample size for this study was the total of around 18,000 for Stage V. Thus, it was decided to use pseudo-R^2 (i.e., McFadden R^2) for a DIF detection criterion, where any item with McFadden R^2 greater than 0.02 between Model 3 and Model 1, which is considered as a small effect size (Choi et al., 2011, Cohen,1988), was flagged for DIF.

Mode Comparability Results

No anchor items were flagged by the logistic regression DIF method with the pseudo-R^2 criterion. Appendix A presents a *p*-value associated with the likelihood ratio χ^2 test as well as McFadden R^2 between Model 3 and Model 1. McFadden R^2 was quite small relative to the criterion of 0.02 for all items. To validate the results by the logistic regression DIF method, Mantel-Haenszel (MH) DIF method was also applied. Note that MH DIF can only investigate uniform DIF, which examines the difference in difficulty between modes across the ability range. Out of the 58 items investigated across three stages, the MH DIF method detected only one item on Stage IV. This item was flagged as B-DIF by the ETS DIF criteria indicating a negligible shift in difficulty for this item between modes. A *p*-value associated with MH χ^2, MH D-DIF, and the ETS DIF flag are also presented in Appendix A. Overall, the logistic regression DIF and MH DIF produced consistent results. Based on the DIF results, the original anchor set was kept for equating. Further exploration of any change in difficulty would take place during the equating process using the anchor items' displacement values using a fixed anchor method.

Equating

Equating the 2018 forms to the 2017 AZELLA scale was performed using the test's standard fixed anchor, non-equivalent groups anchor item (NEAT) design. This was implemented within Winsteps 3.90.0 (Linacre, 2015) where item difficulty for dichotomously scored items is modeled using the Rasch model (Rasch, 1960) and for polytomously scored items the partial credit model (Masters, 1982) is used. In the literature, a displacement (change in difficulty from the fixed value to the new value, if the item were freely estimated) of greater than 0.5 logits in magnitude is of concern (Linacre, 2018). Arizona, however, flags any item with a value of displacement greater than 0.3 in magnitude, so that only anchor items that have estimated difficulty values within approximately 1/3 logit of their fixed value are maintained as anchors. Those that do not meet this threshold, are released from the anchor set and freely estimated in an iterative process releasing the item with the largest flagged displacement and re-equating the test until no more anchors are flagged for displacement.

Using the original anchor set for equating with the fixed anchor method revealed that there were a few items flagged by displacement with an absolute value > 0.3 for Stages III, IV, and V (2, 4, and 2 items, respectively). The original anchor set for 2018 was relatively small in comparison to that of previous years due to the constraints for anchor item selection criteria, in which

1) items should not be impacted by mode (i.e., multiple-choice items) and
2) items should not have any position shift from Spring 2017 to Spring 2018.

Writing prompts were not considered for equating since the students typed their responses on a computer in Spring 2018 while they hand-wrote the responses in Spring 2017. In addition, open-ended writing responses for Stages III through V were scored by Pearson's automated scoring engine in Spring 2018 as opposed to by human scoring by professionally trained scores as in previous years. Speaking and oral reading items were also not considered for equating since students took these items on computer with a head-set with a microphone in Spring 2018 as opposed to via telephone in Spring 2017.

With the loss of the anchors through displacement, it was decided to relax the anchor selection constraint in terms of the position shift to up to 3 positions away from the item location in the Spring 2017 to investigate whether additional items could serve as anchors this year. This relaxed criterion enabled us to recruit up to 5 more possible anchors per stage.

With the revised anchor set, the fixed anchor method was re-run and again flagged items for displacement in each stage. The final anchor set had 23 items for Stage III (35% of test length, 28% of total score), 18 items for Stage IV (26% of test length, 21% of total score), and 17 items for Stage V (24% of test length, 20% of total score). The final anchor set, founded to be stable between modes, was used to equate Spring 2018 forms to the AZELLA base scale so that the reported scale scores of online forms in Spring 2018 were comparable to the scores of paper forms in the past. Note that since Proficiency level cuts were updated after Spring 2016, the reported scale scores for domains and subdomains, except for Total Combined, are comparable only between Spring 2017 and Spring 2018. A summary of the anchor investigation using the displacement statistic is presented in Table 1.

Table 1. Summary of Anchor Set

Stage	Number of Original Anchors	Number of Additional Anchors Studied	*Number of Studied Anchors by Domain	**Studied Anchors Dropped (Domain/ Displacement)	*Number of Final Anchors by Domain	% of Final Anchors	
						Item	Point
III	20	5	LS: 5 RD: 13 WR: 7	Item 7 (LS/ -0.5818) Item 35 (RD/ 0.3029)	LS: 4 RD: 12 WR: 7	35%	28%
IV	21	1	LS: 5 RD: 8 WR: 9	Item 17 (LS/ -0.5284) Item 8 (LS/ -0.4121) Item 31 (RD/ 0.3751) Item 19 (RD/ 0.3212)	LS: 3 RD: 6 WR: 8	26%	21%
V	17	5	LS: 6 RD: 10 WR: 6	Item 44 (RD/ 0.7646) Item 8 (LS/ -0.7017) Item 10 (LS/ -0.3582) Item 15 (LS/ -0.3604) Item 2 (LS/ -0.3381)	LS: 2 RD: 9 WR: 6	24%	20%

*LS: Listening, RD: Reading, WR: Writing
**Items are presented in an order of being dropped

After equating with the stable anchor items was complete, an impact analysis was conducted on both test characteristics and student performance for the 2018 online forms, by comparing the results for the 2018 forms against historical trends as a reasonableness check. The historical trends are summarized by stage in Appendix B. Except for raw score cuts on Total Combined scale score for Stage V, the analysis within each stage was further broken down by grade. (For Stage V, the raw score cuts are the same across all grades assessed.)

In terms of test characteristics, the average p-value for each test form remained relatively consistent with that of prior years (varying a maximum of .03, in Stage III forms A and C, from that of 2017). The Rasch difficulty values for the 2018 online forms, however were slightly higher than the previous year except for Stage V Forms A and C, where they were approximately 0.04 logits lower. An increase in Rasch difficulty for forms purposeful and expected as effort to increase difficulty of test in order to increase the precision around the

Proficient cut was made through item selection during Arizona's test construction process. Consequently, raw score cuts for the 2018 online forms were slightly higher than in 2017. Student performance on the 2018 forms were comparable to 2017's for all grades with respect to the average scale score and percentage of passing (i.e., Proficient). The cumulative frequency distribution of scale score was plotted for Spring 2018 forms and Spring 2017 form by grade to visually present the comparability of scale scores across years and forms across the ability distribution (see Figures B.1.a through B.1.c for grades 3-5, B.2.a through B.2.c for grades 6-8, and B.3.a through B.3.d for High School grades). The distributions were similar among the Spring 2018 forms and Spring 2017 form.

Conclusions

In Spring 2018, a mode comparability was conducted on the AZELLA Stages III through V Reassessment forms to investigate, and moderate if evident, any effects caused by the transition from paper-and-pencil to online administration. Online administration was instituted during the Spring 2018 AZELLA Reassessment window, for grades 3 and above, and will remain in effect for Placement and Reassessment administrations for these grades for the foreseeable future.

The logistic regression DIF analysis results revealed that no anchor items showed item drift due to the change in mode. The MH DIF results were consistent with the logistic regression DIF, flagging only one item which had moderate DIF. There were, however, some anchor items that were flagged by displacement within the fixed anchor calibration. It appeared that the flagging criteria for the logistic regression and MH DIF methods were less sensitive to changes in item difficulty than the criterion for the displacement. Nonetheless, the Spring 2018 forms were equated with stable anchor items to the AZELLA base scale so that the reported scale scores were comparable to the scale scores based on the paper-and-pencil forms administered in the past. Based on equated forms, the impact data, such as average scale score, percent of students at the Proficiency level, and cumulative frequency distribution of scale score across administrations, indicated no evidence of shift due to the mode change.

References

Choi, S. W., Gibbons, L. E., Crane, P. K. (2001). lordif: An R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and Monte Carlo simulations. Journal of Statistical Software, 39(8), 1-30.

Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences. 2nd edition. Lawrence Earlbaum Associates, Hillsdale, NJ.

Linacre, J. M. (2015) Winsteps® (Version 3.90.0) [Computer Software]. Beaverton, Oregon: Winsteps.com. Retrieved January 1, 2015. Available from http://www.winsteps.com/.

Linacre, J. M. (2018). A user's guide to WINSTEPS® MINISTEP Rasch-model computer programs 4.1.0. Beaverton, OR: Winsteps.

Masters, G.N. (1982). A Rasch model for partial credit scoring. Psychometrika, 47,149-174.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Copenhagen, Denmark: Danmarks Paedogogiske Institut.

Appendix A. Logistic Regression DIF and MH DIF Results

Table A.1. Stage III DIF Results

	N-count		Logistic Regression DIF (Model 3 vs Model 1)		MH DIF		
Item	**Paper (2017)**	**Online (2018)**	**χ^2 *p*-value**	**McFadden R^2**	**MH-χ^2 *p*-value**	**MH D-DIF**	**ETS Flag**
1	22847	23060	< 0.01	0.001	< 0.01	-0.20	A
*2	22847	23060	< 0.01	0.005	< 0.01	-0.78	A
3	22847	23060	< 0.01	0.002	< 0.01	-0.26	A
4	22847	23060	< 0.01	0.004	< 0.01	0.76	A
5	22847	23060	< 0.01	0.003	< 0.01	0.50	A
6	22847	23060	< 0.01	0.001	< 0.01	0.22	A
7	22847	23060	< 0.01	0.000	0.07	-0.09	A
8	22847	23060	< 0.01	0.001	< 0.01	0.29	A
9	22847	23060	< 0.01	0.001	< 0.01	0.25	A
10	22847	23060	< 0.01	0.000	< 0.01	0.16	A
11	22847	23060	< 0.01	0.002	< 0.01	0.49	A
12	22847	23060	< 0.01	0.000	< 0.01	-0.15	A
13	22847	23060	0.02	0.000	0.46	0.04	A
14	22847	23060	< 0.01	0.003	< 0.01	0.44	A
15	22847	23060	0.05	0.000	0.05	-0.10	A
16	22847	23060	< 0.01	0.001	< 0.01	-0.44	A
*17	22847	23060	< 0.01	0.007	< 0.01	-0.97	A
18	22847	23060	0.19	0.000	1.00	0.00	A
19	22847	23060	< 0.01	0.004	< 0.01	-0.68	A
20	22847	23060	0.32	0.000	0.09	0.08	A

*Flagged by displacement based on the original anchor set

Table A.2. Stage IV DIF Results

	N-count		Logistic Regression DIF (Model 3 vs Model 1)		MH DIF		
Item	**Paper (2017)**	**Online (2018)**	**χ^2 *p*-value**	**McFadden R^2**	**MH-χ^2 *p*-value**	**MH D-DIF**	**ETS Flag**
1	11899	14956	< 0.01	0.005	< 0.01	-0.88	A
2	11899	14956	0.20	0.000	0.10	-0.10	A
*3	11899	14956	< 0.01	0.009	< 0.01	-1.14	B
4	11899	14956	< 0.01	0.001	< 0.01	-0.41	A
*5	11899	14956	< 0.01	0.002	< 0.01	-0.60	A
*6	11899	14956	< 0.01	0.000	< 0.01	0.19	A
7	11899	14956	< 0.01	0.008	< 0.01	0.99	A
8	11899	14956	< 0.01	0.002	< 0.01	0.44	A
9	11899	14956	< 0.01	0.001	< 0.01	0.26	A
*10	11899	14956	< 0.01	0.004	< 0.01	0.69	A
11	11899	14956	< 0.01	0.001	< 0.01	-0.21	A
12	11899	14956	< 0.01	0.003	< 0.01	0.33	A
13	11899	14956	< 0.01	0.001	< 0.01	-0.43	A
14	11899	14956	0.03	0.000	< 0.01	-0.19	A
15	11899	14956	< 0.01	0.001	0.02	0.15	A
16	11899	14956	0.23	0.000	0.18	-0.09	A
17	11899	14956	< 0.01	0.000	< 0.01	0.19	A
18	11899	14956	0.42	0.000	0.12	-0.10	A
19	11899	14956	< 0.01	0.001	< 0.01	-0.36	A
20	11899	14956	< 0.01	0.001	0.01	0.17	A
21	11899	14956	< 0.01	0.003	< 0.01	0.22	A

*Flagged by displacement based on the original anchor set

Table A.3. Stage V DIF Results

	N-count		Logistic Regression DIF (Model 3 vs Model 1)		MH DIF		
Item	**Paper (2017)**	**Online (2018)**	**χ^2 *p*-value**	**McFadden R^2**	**MH-χ^2 *p*-value**	**MH D-DIF**	**ETS Flag**
1	7668	10588	< 0.01	0.002	< 0.01	-0.49	A
*2	7668	10588	< 0.01	0.003	< 0.01	-0.59	A
3	7668	10588	< 0.01	0.001	< 0.01	-0.33	A
4	7668	10588	0.38	0.000	0.02	-0.21	A
5	7668	10588	< 0.01	0.002	< 0.01	0.46	A
6	7668	10588	< 0.01	0.001	< 0.01	0.33	A
7	7668	10588	< 0.01	0.001	< 0.01	0.38	A
8	7668	10588	0.01	0.000	0.49	0.05	A
9	7668	10588	< 0.01	0.001	< 0.01	-0.54	A
*10	7668	10588	< 0.01	0.006	< 0.01	0.90	A
11	7668	10588	< 0.01	0.001	0.62	-0.04	A
12	7668	10588	< 0.01	0.001	< 0.01	-0.42	A
13	7668	10588	0.73	0.000	0.91	-0.01	A
14	7668	10588	< 0.01	0.001	< 0.01	-0.47	A
15	7668	10588	< 0.01	0.002	< 0.01	-0.68	A
16	7668	10588	0.01	0.000	0.01	0.20	A
17	7668	10588	< 0.01	0.003	< 0.01	0.60	A

*Flagged by displacement based on the original anchor set

Appendix B. Historical Trend in Test Characteristics and Student Performance

Table B.1.a. Historical Trend in Test Characteristics for Stage III

Year	*Form	Average P-value	Average Rasch	Weighted Raw Score Cuts on Total			
				Grade	Basic	Intermediate	Proficient
2018	A	0.53	0.1858	3	46	74	107
				4	59	86	118
				5	65	91	127
	C	0.53	0.1856	3	46	73	107
				4	59	86	118
				5	65	90	127
	D	0.55	0.0888	3	49	78	110
				4	63	90	121
				5	70	95	130
2017		0.56	-0.0146	3	57	83	114
				4	69	94	124
				5	76	99	132

*In Spring 2018, there were 3 core forms with embedded field test items to make 6 forms. Form A and C were the same form except that they had different Writing prompt questions.

Table B.1.b. Historical Trend in Student Performance by Grade for Stage III

Grade	*Year	N	Average Scale Score	SD Scale Score	Percent at Overall Proficiency Level			
					Pre-/Emergent	Basic	Intermediate	Proficient
3	2018	7150	2417.10	40.35	9	40	44	6
	2017	8243	2417.96	46.34	11	37	45	7
4	2018	8114	2444.27	47.49	12	30	51	7
	2017	8245	2448.23	53.57	14	25	52	10
5	2018	7796	2468.29	54.72	11	21	56	12
	2017	6418	2462.52	56.81	14	19	56	11

*After Spring 2016, Proficient cut was increased. Thus, data prior to Spring 2017 is not comparable.

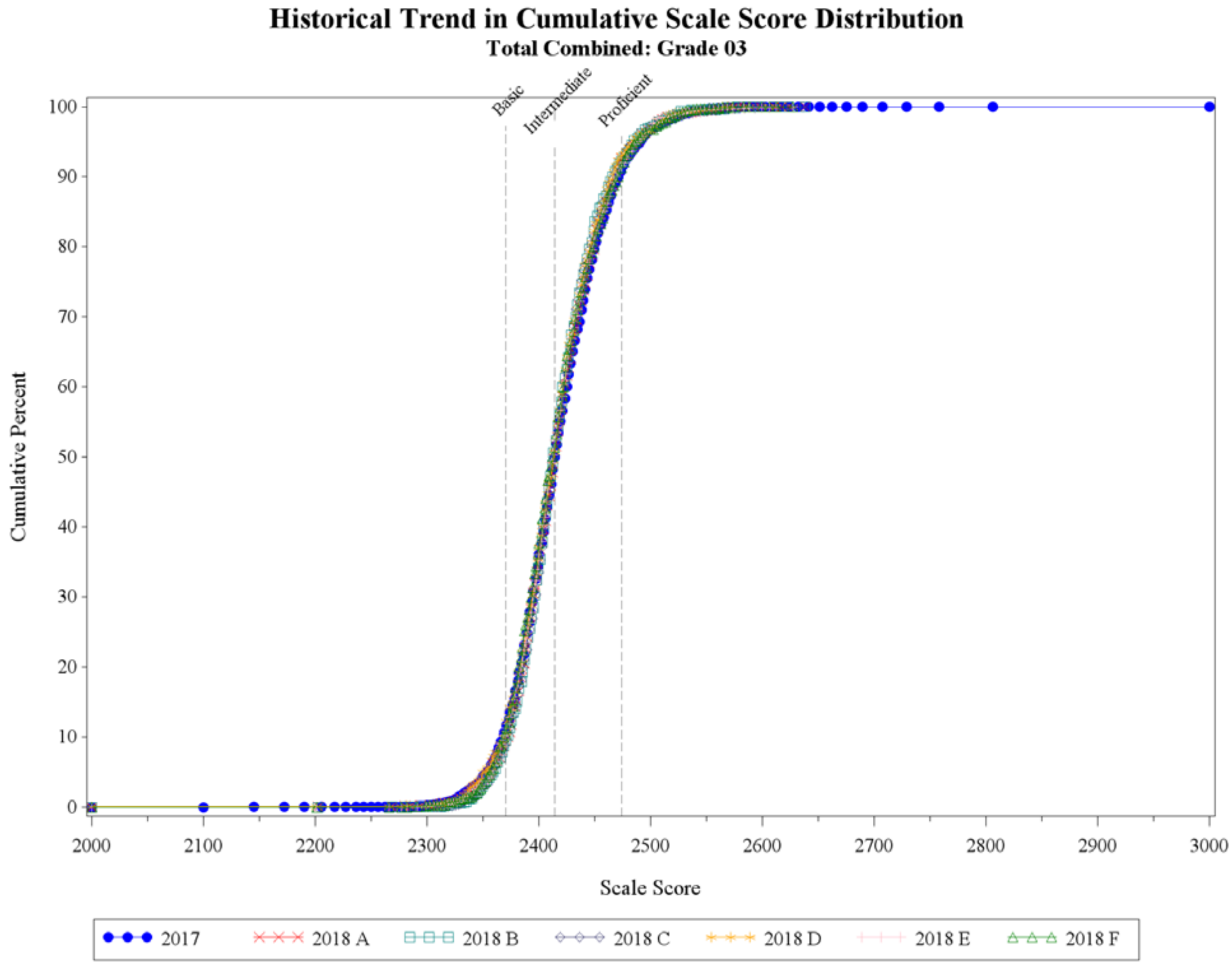

Figure B.1.a. Historical Trend in Cumulative Distribution of Total Combined Scale Score at Grade 3

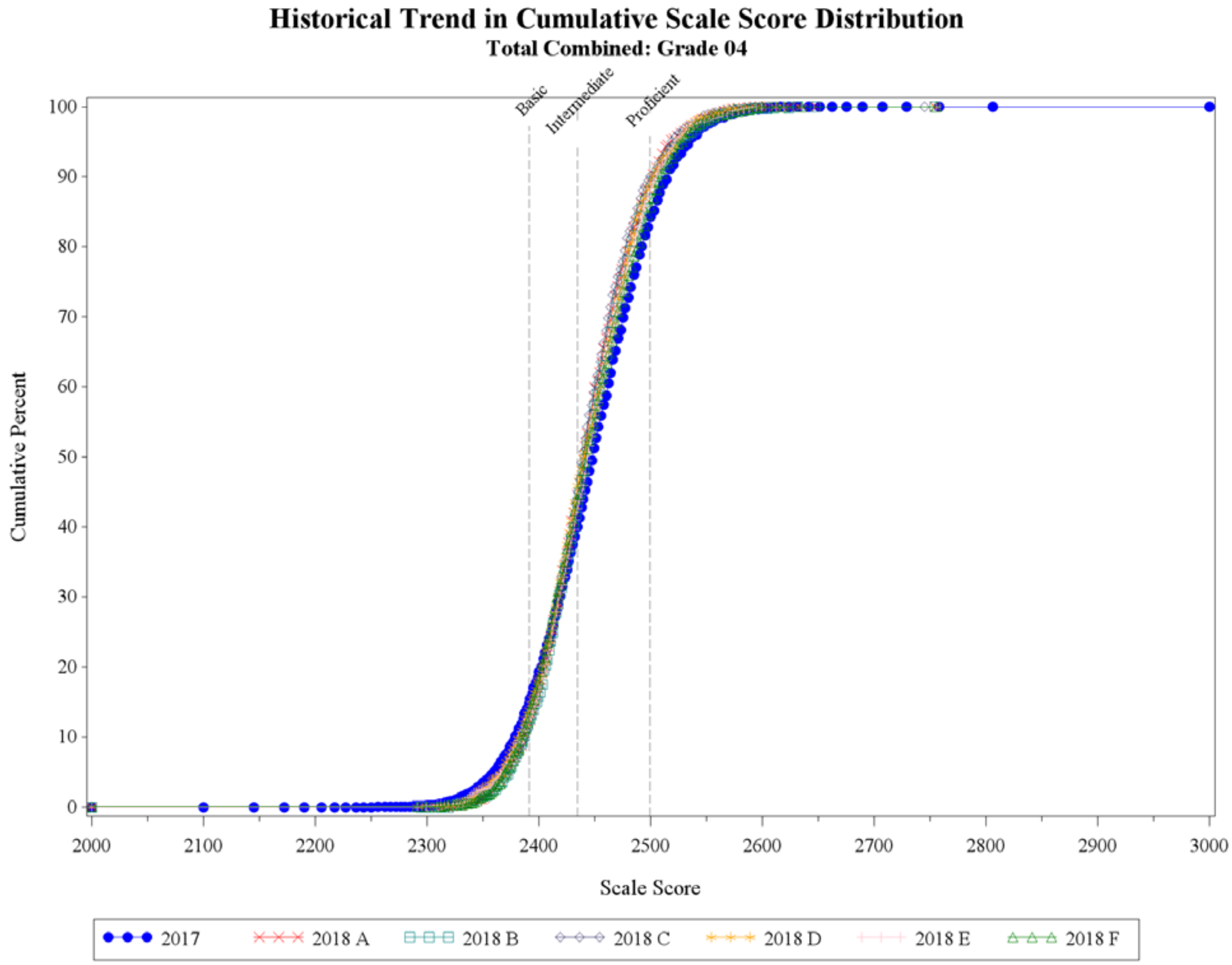

Figure B.1.b. Historical Trend in Cumulative Distribution of Total Combined Scale Score at Grade 4

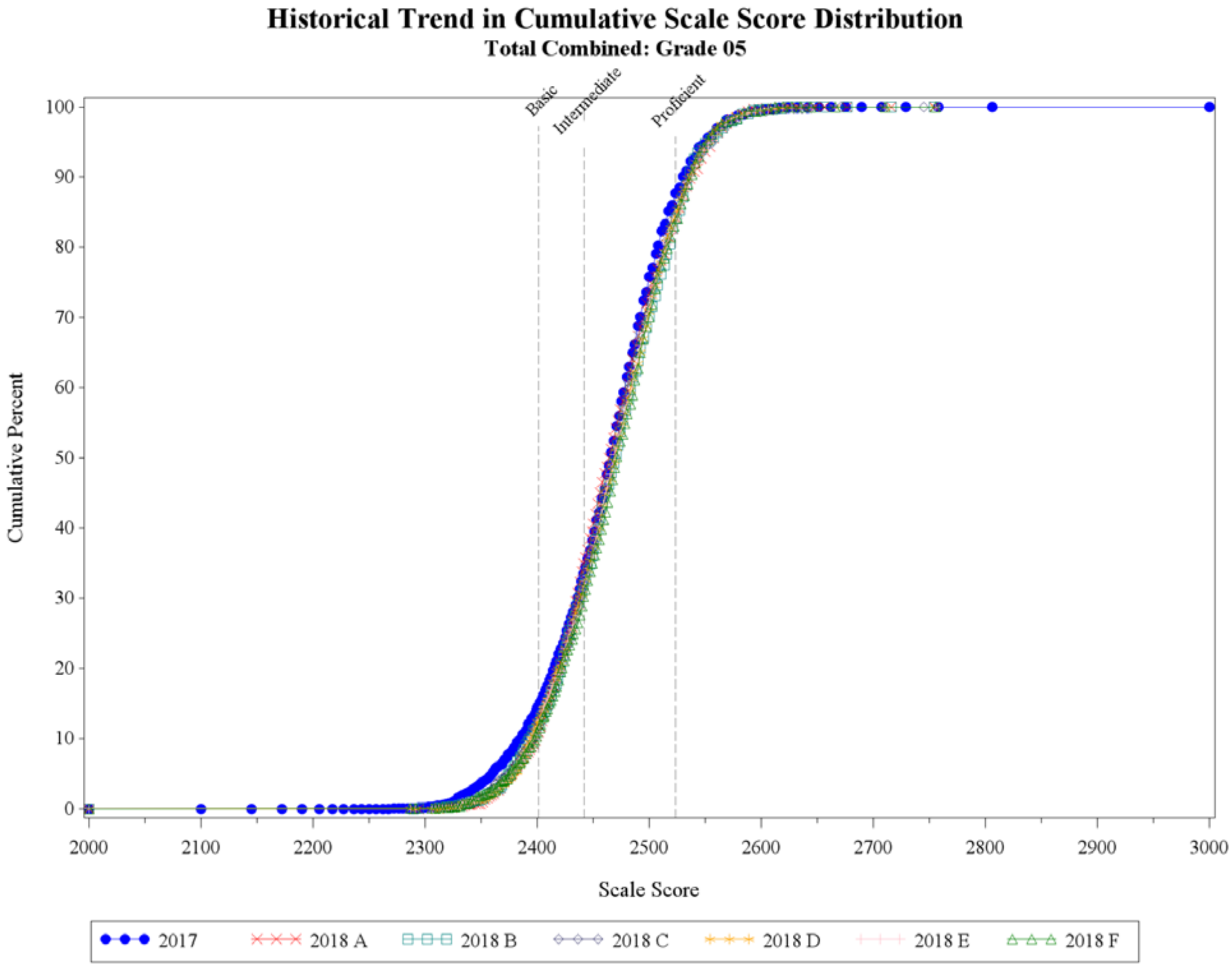

Figure B.1.c. Historical Trend in Cumulative Distribution of Total Combined Scale Score at Grade 5

Table B.2.a. Historical Trend in Test Characteristics for Stage IV

Year	*Form	Average P-value	Average Rasch	Weighted Raw Score Cuts on Total			
				Grade	Basic	Intermediate	Proficient
2018	A	0.55	0.5116	6	56	80	124
				7	56	80	126
				8	56	80	128
	C	0.55	0.5021	6	57	81	126
				7	57	81	128
				8	57	81	130
	D	0.55	0.5293	6	56	79	123
				7	56	79	125
				8	56	79	127
2017		0.56	0.4011	6	61	85	129
				7	61	85	131
				8	61	85	132

*In Spring 2018, there were 3 core forms with embedded field test items to make 6 forms. Form A and C were the same form except that they had different Writing prompt questions.

Table B.2.b. Historical Trend in Student Performance by Grade for Stage IV

Grade	*Year	N	Average Scale Score	SD Scale Score	Percent at Overall Proficiency Level			
					Pre-/Emergent	Basic	Intermediate	Proficient
6	2018	6032	2463.73	47.34	10	22	62	5
	2017	4312	2457.77	50.39	15	23	57	6
7	2018	4662	2472.38	55.28	12	17	64	7
	2017	4199	2471.19	58.62	14	17	61	9
8	2018	4262	2487.60	57.83	9	13	67	11
	2017	3392	2478.79	60.68	13	15	62	10

*After Spring 2016, Proficient cut was increased. Thus, data prior to Spring 2017 is not comparable.

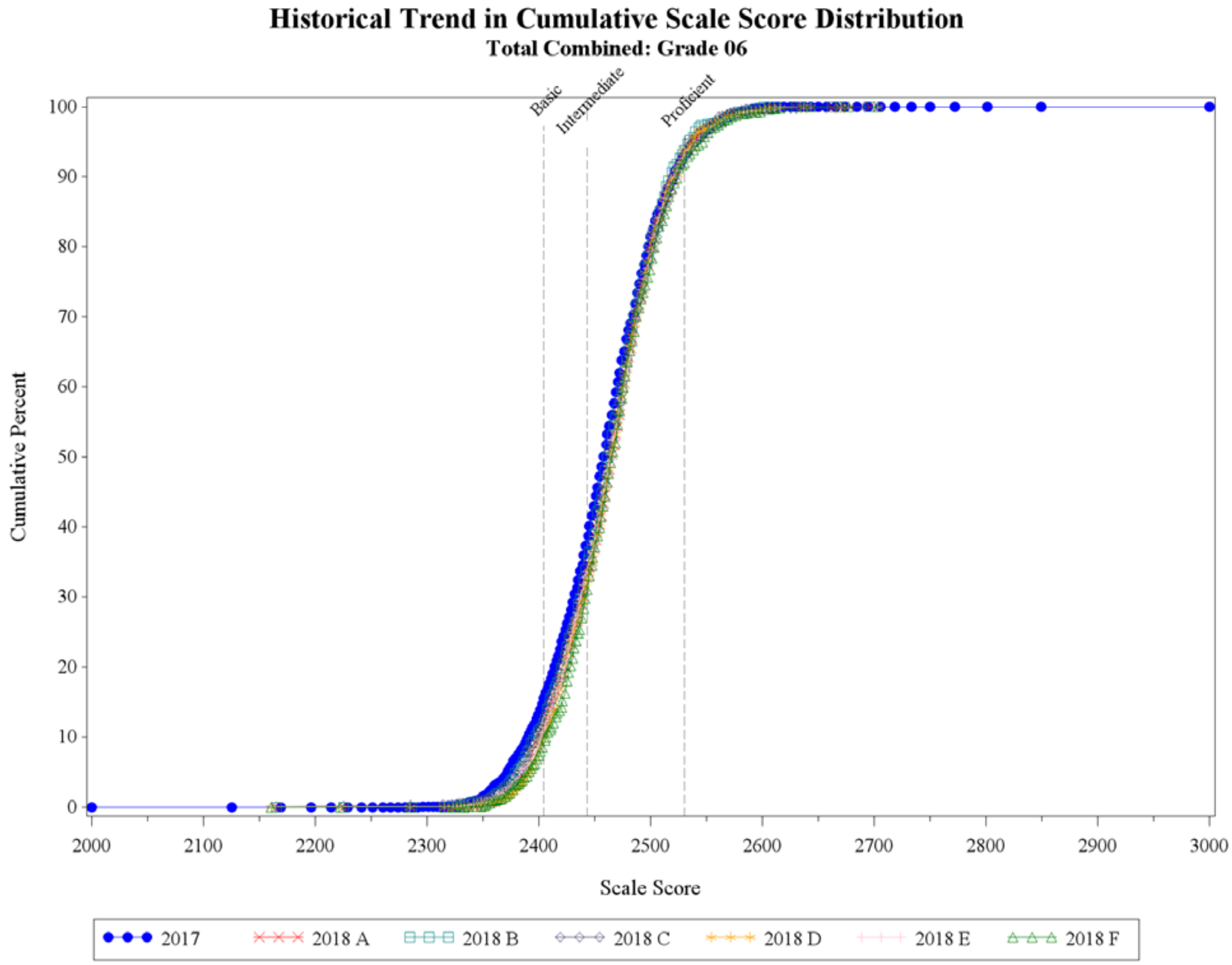

Figure B.2.a. Historical Trend in Cumulative Distribution of Total Combined Scale Score at Grade 6

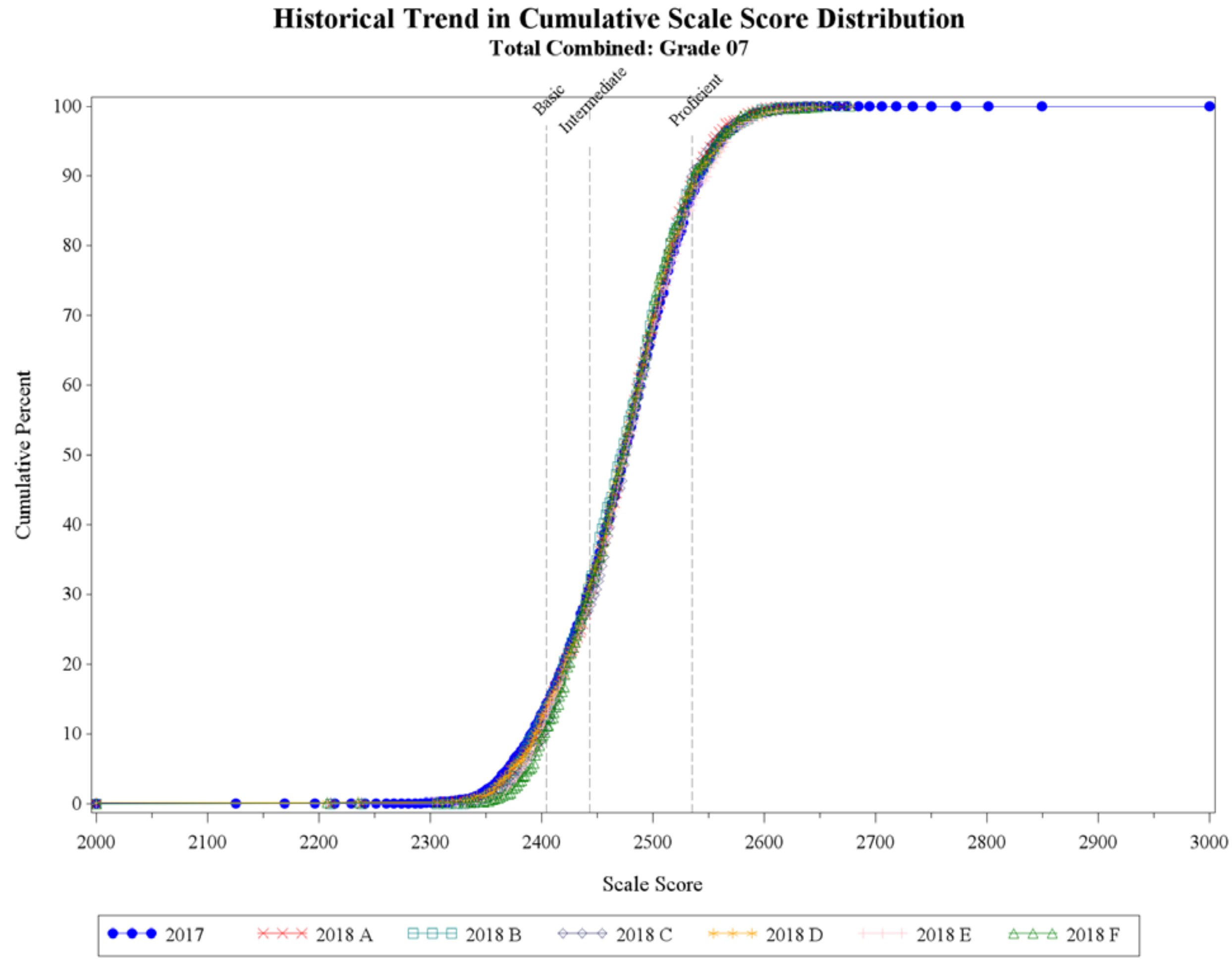

Figure B.2.b. Historical Trend in Cumulative Distribution of Total Combined Scale Score at Grade 7

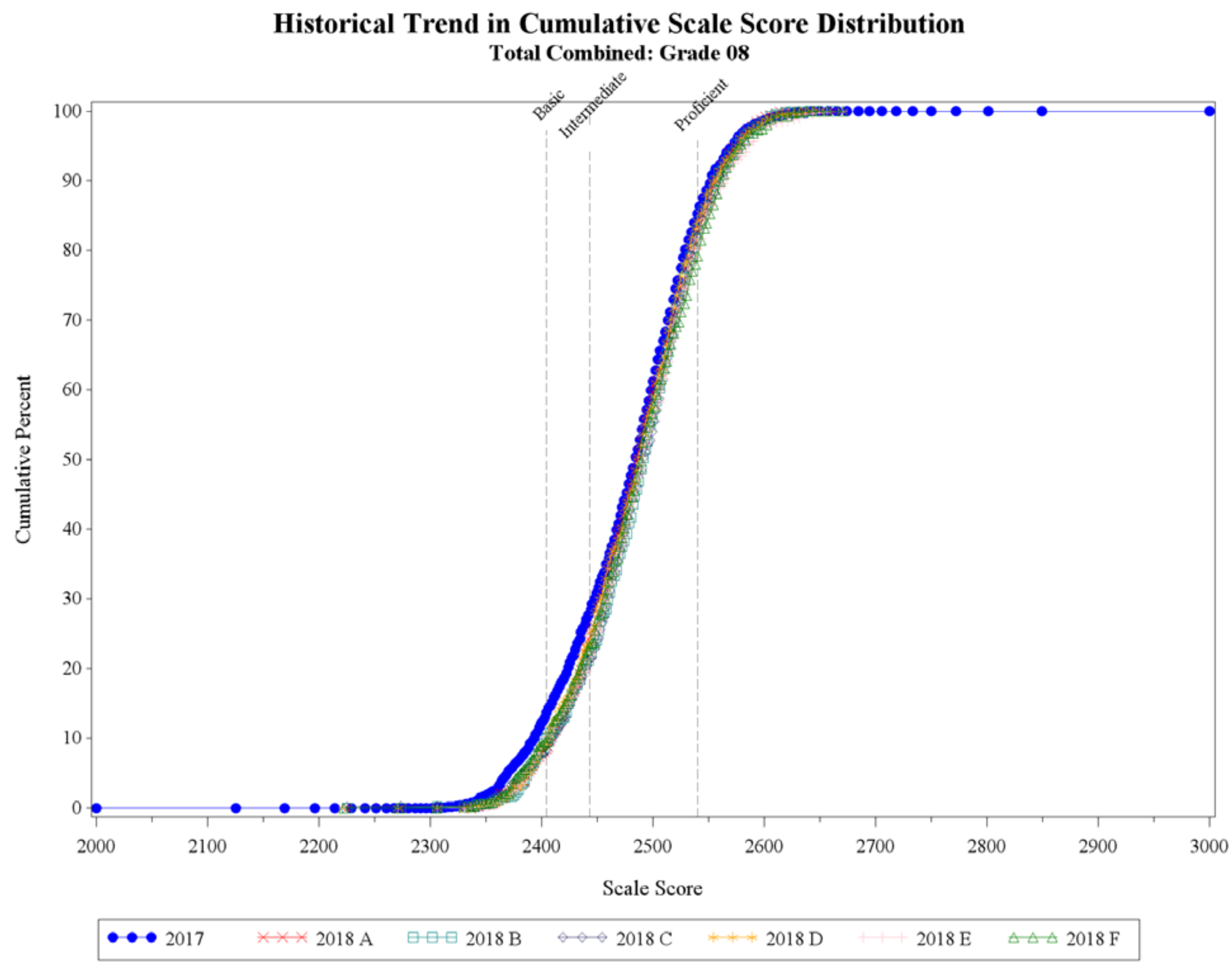

Figure B.2.c. Historical Trend in Cumulative Distribution of Total Combined Scale Score at Grade 8

Table B.3.a. Historical Trend in Test Characteristics for Stage V

Year	*Form	Average P-value	Average Rasch	Weighted Raw Score Cuts on Total		
				Basic	Intermediate	Proficient
2018	A	0.56	0.7447	60	85	126
	C	0.56	0.7427	61	86	127
	D	0.56	0.8068	59	83	125
2017		0.54	0.7830	62	85	125

*In Spring 2018, there were 3 core forms with embedded field test items to make 5 forms. Form A and C were the same form except that they had different Writing prompt questions.

Table B.3.b. Historical Trend in Student Performance by Grade for Stage V

Grade	*Year	N	Average Scale Score	SD Scale Score	Percent at Overall Proficiency Level			
					Pre-/Emergent	Basic	Intermediate	Proficient
9	2018	3869	2487.28	54.63	14	23	53	9
	2017	2902	2471.13	57.16	24	24	47	6
10	2018	3049	2490.61	56.63	13	22	53	12
	2017	2298	2486.92	53.84	13	22	57	8
11	2018	2074	2502.91	52.53	8	18	61	14
	2017	1503	2496.07	51.33	10	18	63	9
12	2018	1596	2500.49	51.34	7	18	62	13
	2017	965	2498.48	52.51	8	21	60	11

*After Spring 2016, Proficient cut was increased. Thus, data prior to Spring 2017 is not comparable.

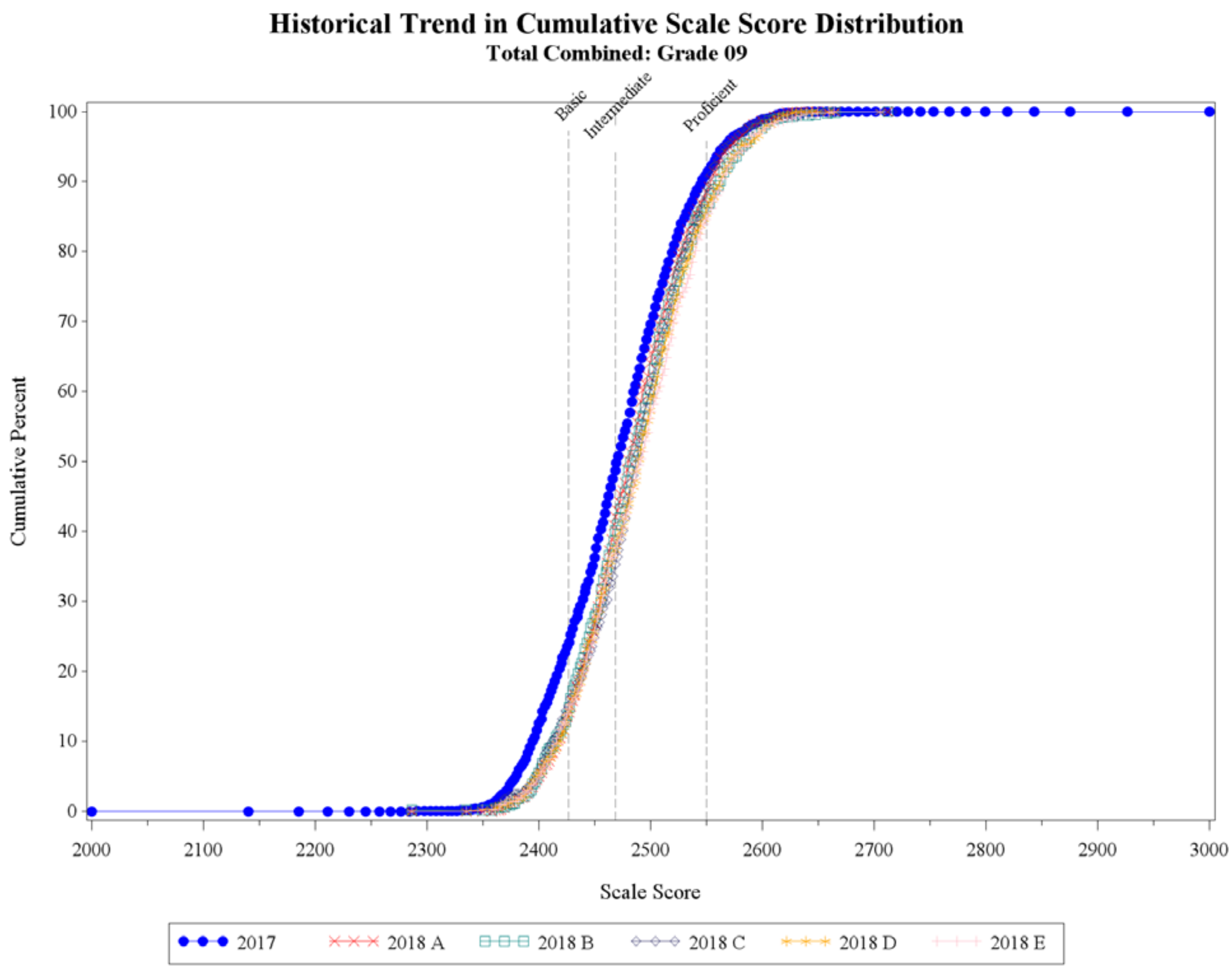

Figure B.3.a. Historical Trend in Cumulative Distribution of Total Combined Scale Score at Grade 9

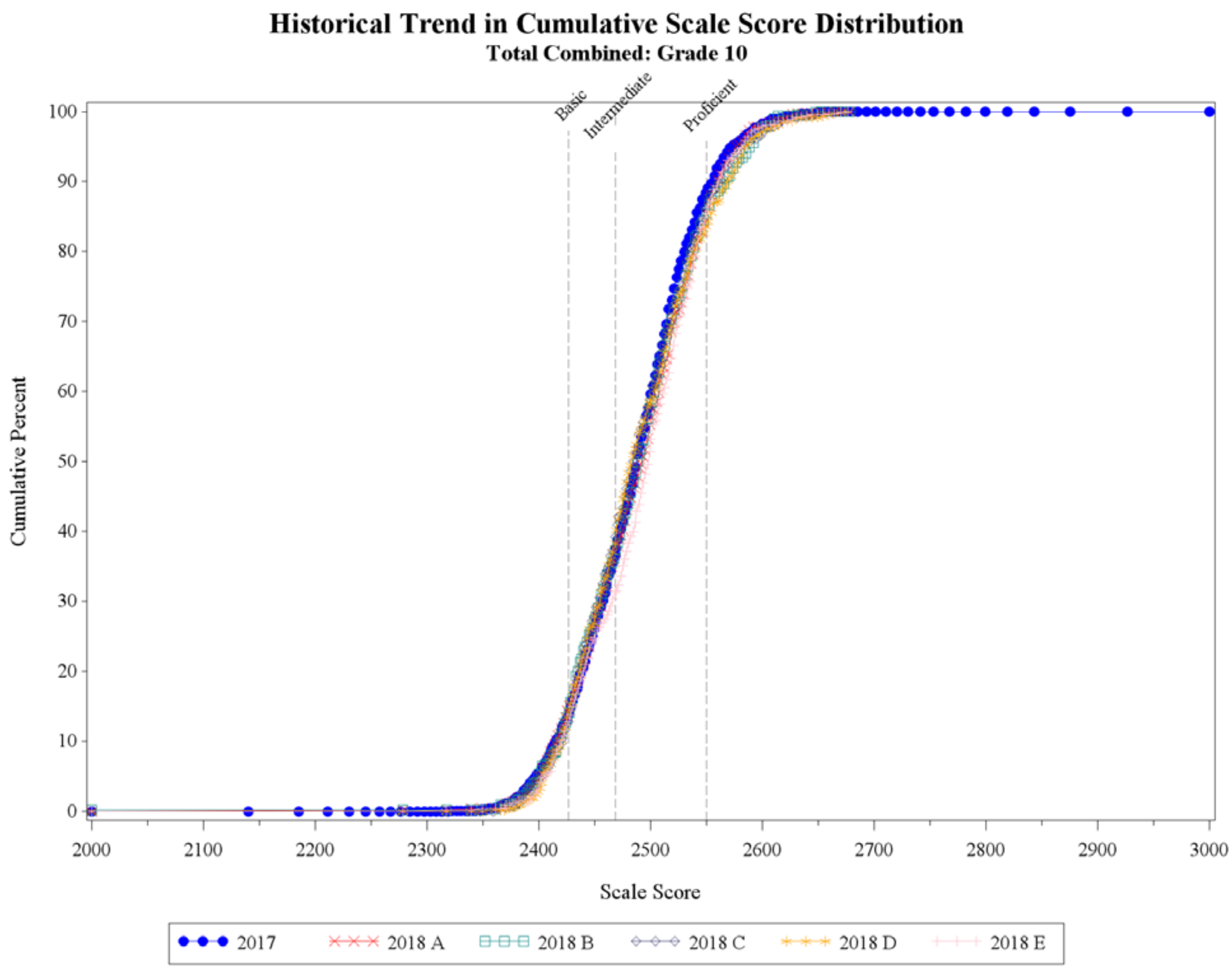

Figure B.3.b. Historical Trend in Cumulative Distribution of Total Combined Scale Score at Grade 10

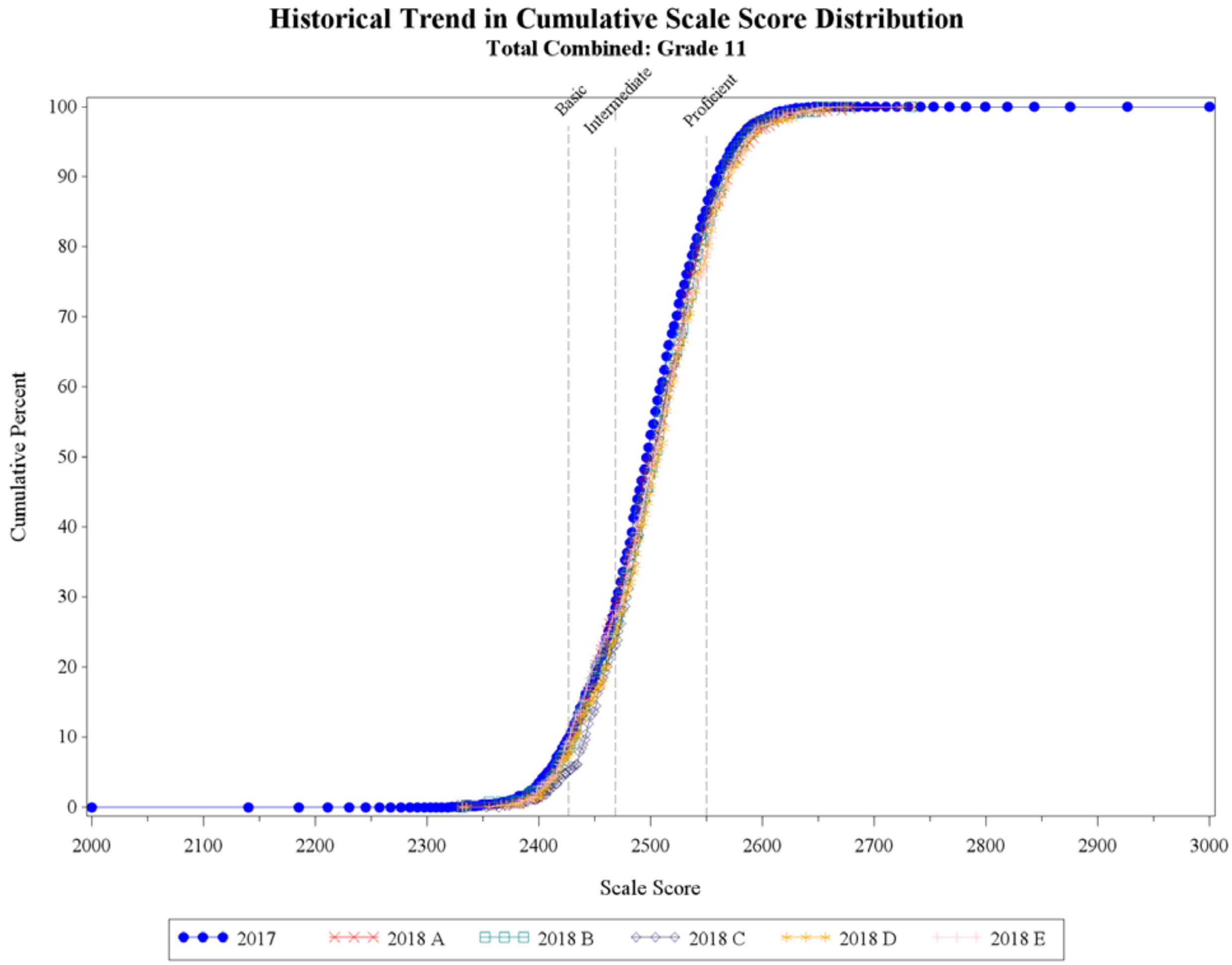

Figure B.3.c. Historical Trend in Cumulative Distribution of Total Combined Scale Score at Grade 11

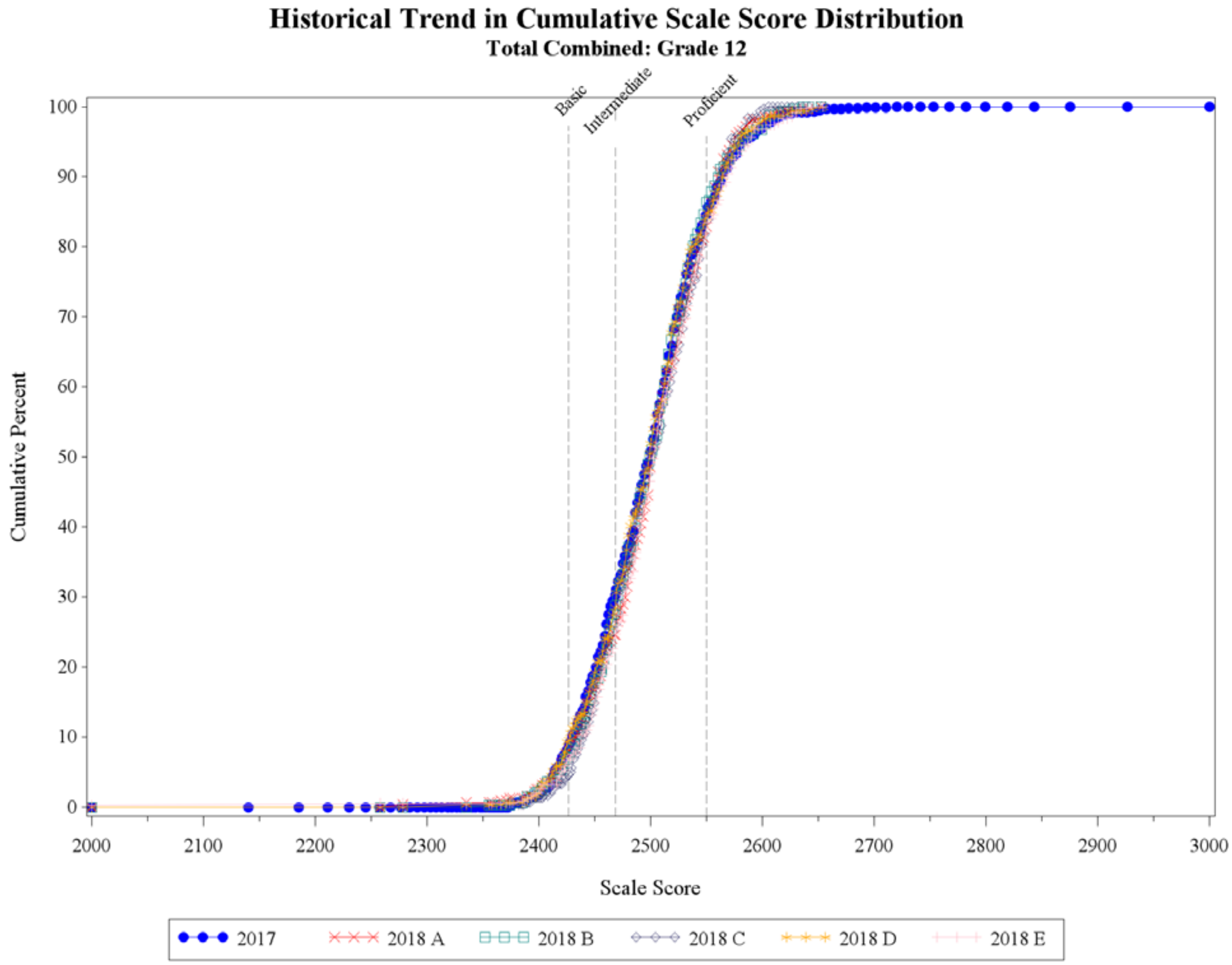

Figure B.3.d. Historical Trend in Cumulative Distribution of Total Combined Scale Score at Grade 12