North Carolina Personalized Assessment Tools (NCPAT)

Draft Technical Report

September 28, 2023

North Carolina Department of PUBLIC INSTRUCTION

TABLE OF CONTENTS

TABLE OF TABLES

TABLE OF FIGURES

TABLE OF APPENDICES

CHAPTER 1. INTRODUCTION

The intent of this technical report is to provide comprehensive and detailed evidence in support of the validity and reliability of the multistage fixed adaptive summative component of the North Carolina Personalized Assessment Tools (NCPAT) system. North Carolina Department of Public Instruction (NCDPI) has partnered with the Office of Assessment, Evaluation, and Research Services (OAERS) at the University of North Carolina at Greensboro to provide psychometric and technical support in the design, development, and pilot of the NCPAT system in mathematics and reading for grades 4 and 7. Specifically, this report provides detailed evidence of the technical processes used to develop multistage adaptive End-of-Grade (EOG) forms with corresponding validity evidence in support of score interpretation.

For a comprehensive detail on purpose and background of the North Carolina state testing program, test design, item development, field-test plan, test administration, scoring and scale development, and standard setting, refer to the NCDPI Technical Reports for mathematics and reading on the NCDPI website (https://www.dpi.nc.gov) and search technical report or access directly via the link provided below:

Technical Report for EOG Mathematics and EOC Math I 2018–19 (Edition 5).pdf
NCDPI_Edition5_Reading_TechnicalReport_2020-21.pdf

1.1. Overview of the Technical Report

Chapter 2 summarizes procedures used to evaluate field-test items that were embedded within operational administrations for the development of NCPAT multistage EOG forms. The field test analysis followed standard NCDPI established evaluation criteria based on Classical Test Theory (CTT), Item Response Theory (IRT), as well as differential item functioning analysis. The main goal was to evaluate and filter out items with less-than-optimal characteristics. The final section briefly described the IRT fixed parameter calibration method used to put field test items onto the operational EOG scale.

Chapter 3 starts with IRT-based form assembly process based on EOG test blueprint with additional constraints used to build multistage fixed adaptive EOG forms that are aligned to grade

level blueprint. The next section documents form characteristics, such as test characteristic curves, test information functions, and conditional standard error of measurement for the multistage fixed adaptive EOG forms developed for the NCPAT system. The final section of this chapter briefly describes the routing procedures used to route students to one of the three multistage fixed adaptive EOG forms based on performance data collected throughout the year from NC Check-Ins 2.0.

Chapter 4 presents the scoring procedure for the NCPAT system and discusses scale comparability between the multistage fixed adaptive EOG administered under the NCPAT system and the traditional EOG.

Chapter 5 presents validity evidence collected in support of the interpretation of the multistage fixed adaptive EOG test scores. The first two sections in this chapter present validity evidence in support of internal structure of EOG assessments. Evidence presented in this section includes reliability, standard error estimates at cut scores, classification consistency and accuracy of the reported achievement levels, and results of principal component analysis with a confirmatory factor analysis in support of the unidimensional interpretation of scores.

CHAPTER 2. ITEM ANALYSIS

This chapter summarizes procedures and criteria that the NCDPI uses to analyze and evaluate the statistical and psychometric characteristic of newly developed test items, which hereinafter will be referred to as field-test items. Item analysis serves as the final quantitative process for item review and establishing grade level operational item pool for form development. Standard 4.10 (AERA, APA, & NCME, 2014) states, *"When a test developer evaluates the psychometric properties of items, the model used for that purpose should be documented. The process by which items are screened and the data used for screening, such as item difficulty, item discrimination, or differential item functioning (DIF) for major examinee groups, should also be documented"* (p.89). Most large-scale assessment programs rely on two measurement theories–CTT and IRT–to screen and evaluate items for calibration, form assembly, and scoring. Another important procedure in traditional item analysis is the statistical evaluation of DIF, which is commonly used to evaluate fairness and potential item bias across subgroups. The NCDPI psychometric specifications for item review use statistical criteria from both CTT and IRT in addition to Mantel-Haenszel (MH) DIF statistics. These procedures and their various criteria used for item screening and analysis are described in the following sections.

2.1. Statistical Item Flagging Criteria

All field-test items are classified into one of three NCDPI item flagging categories (Keep, Reserve, and Weak), with the goal to rank items in the final pool based on overall statistical quality during form assembly. These specifications are routinely updated to continuously ensure that only items that meet the expected technical specifications are selected into the final item pool for operational form development.

- *Keep*: These are items with acceptable statistical properties from CTT, IRT, and DIF statistical procedures used for item analysis. Items flagged as "Keep" are first choice from the item pool during form assembly. Their main statistical properties are within the established NCDPI ranges considered as optimal items.
- *Reserve*: These are items with at least one major statistical parameter that is barely outside the range to be considered as "Keep" items. These items are only included in the final form assembly pool if they are needed to meet content or statistical specifications of the

operational form. When any item flagged as "Reserve" from field tests is placed on a new form, it must undergo additional content review to ensure the content is accurate.

- *Weak*: These are items with at least one major statistical parameter being significantly outside the range to be considered as optional items based on field-test analysis. When complete field-test data are available, these items are generally not included in the item pool used for form assembly. The only exception to this rule is when exceptional circumstances cause field-test data to be incomplete or unreliable. In such situations, thorough vetting is required from the content experts and psychometricians.

2.2. Field-Test Item Calibration Procedure

During each EOG test administration window, multiple alternate forms are administered in every grade level. Using a matrix sampling design, subsets of field-test items are embedded in base forms shown in Figure 2.1 to test a large number of field test items without increasing the test length. All form and flavor combinations are randomly spiraled within schools at the student level across the state. This ensures that base forms with field-test items are randomly administered to a representative sample of students at the grade level. Tables 2.1 and 2.2 show demographic information of the field-test samples that were used to obtain the CTT- and IRT-based item statistics for mathematics and reading grades 4 and 7. It shows that the sample sizes and distributions of sex, ethnicity, and EDS are very similar within each grade.

Figure 2.1. Data collection for embedded field–test design.

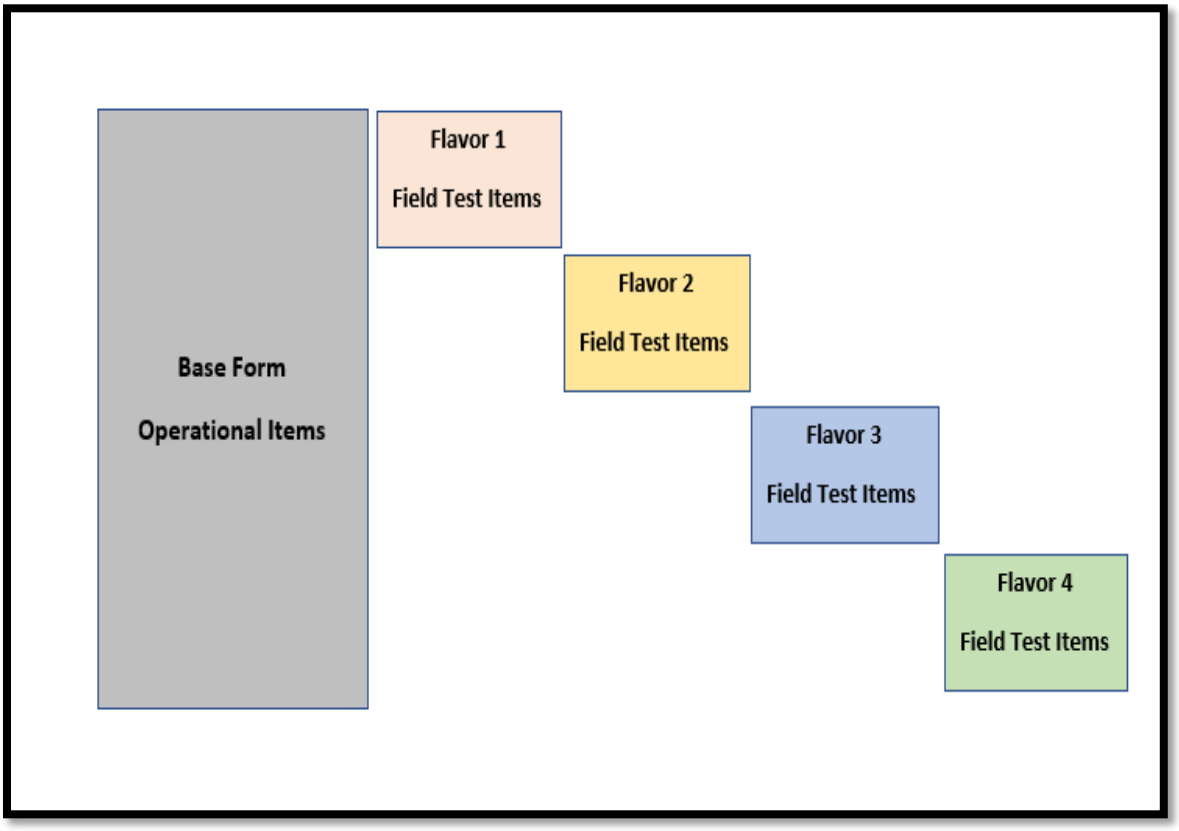


Table 2.1. Demographic information of the field–test sample for EOG mathematics grades 4 and 7, spring 2022.

Grade	Form	Total	Ethnicity				Gender		EDS
			W	B	H	Other	M	F	
4	M	42,797	47.3	25.5	15.2	11.9	48.6	51.3	38.0
	O	41,768	48.0	24.9	15.3	11.7	48.8	51.1	37.5
	All	84,565	47.7	25.2	15.2	11.8	48.7	51.2	37.7
7	M	48,133	45.4	25.2	18.4	11.0	49.0	50.9	36.8
	O	47,358	45.5	25.7	18.3	10.3	48.9	51.0	36.8
	All	95,491	45.5	25.4	18.4	10.7	49.0	50.9	36.8

Note. W = White, B = Black, H = Hispanic, M = Male, F = Female, EDS = Economically disadvantaged student.

Table 2.2. Demographic information of the field–test sample for EOG reading grades 4 and 7, spring 2022.

Grade	Form	Total	Ethnicity				Gender		EDS
			W	B	H	Other	M	F	
4	M	42,797	43.9	23.9	21.1	10.9	50.9	49.0	41.2
	O	41,768	43.9	24.1	20.9	10.9	51.0	48.9	41.1
	All	84,565	43.9	24.0	21.0	10.9	51.0	48.9	41.1
7	M	48,133	43.5	25.3	20.9	10.2	51.2	48.7	39.1
	O	47,358	43.7	25.6	20.7	10.0	51.4	48.5	38.9
	All	95,491	43.6	25.4	20.8	10.1	51.3	48.6	39.0

Note. W = White, B = Black, H = Hispanic, M = Male, F = Female, EDS = Economically disadvantaged student.

2.2. CTT-Based Item Analysis

Item level CTT statistics such as proportion correct (p-value), item-to-total correlation (biserial correlation), and distractor analysis are used as a first step to screen item quality following field tests. The first step involves conducting a series of CTT analysis to determine if these items meet the minimum psychometric requirements to be considered for further evaluation. Brief descriptions of item p-value and biserial correlation are provided below.

- Item *p-value* summarizes the proportion of examinees from a given sample that answers the item correctly and is used as an indicator of preliminary item difficulty. Item p-value for dichotomously scored items ranges from 0 to 1, where values close to 0 indicate more difficult items (few students selected the correct response) and values close to 1 indicate easier items (almost all students answered correctly).
- The *biserial correlation* describes the relationship between a dichotomous variable (item score) and a continuous variable (test score). The biserial correlation provides evidence of item discrimination or more specifically, the strength of the relationship between an item and the unidimensional construct being measured. Theoretical range for biserial correlation is –1 to 1. Negative biserial correlation generally indicates that the item might be measuring a separate unintended construct.

Table 2.3 shows the CTT-based item flagging criteria based on item p-value and biserial correlation. The CTT descriptive statistics for mathematics and reading field-test item pool in the 2021–22 school year are provided in Tables 2.4 and 2.5, respectively. The initial CTT results

indicate that most of the field-test items are classified as meeting the NCDPI optimal standards of "Keep." Moreover, the range of p-values and biserial correlations show that the item pool had items with a wide range of item difficulty and discrimination for high quality operational form assembly.

Table 2.3. CTT-Based item flagging criteria.

CTT Statistics	Flagging Criteria
Item p-value	
$0.150 \leq$ p-value ≤ 0.850	Keep
$0.100 \leq$ p-value ≤ 0.149 or $0.851 \leq$ p-value ≤ 0.900	Reserve
p-value ≤ 0.099 or p-value ≥ 0.901	Weak
Biserial Correlation	
biserial ≥ 0.250	Keep
$0.150 \leq$ biserial ≤ 0.249	Reserve
biserial < 0.150	Weak

Table 2.4. CTT p-values and biserial correlations for mathematics grades 4 and 7 field-test item pool, spring 2022.

Grade	CTT Flag	# of Items	P–value				Biserial Correlation			
			Mean	SD	Min	Max	Mean	SD	Min	Max
4	Keep	186	.57	.16	.17	.84	.61	.10	.27	.78
	Reserve	36	.58	.23	.10	.88	.50	.12	.24	.71
	Weak	18	.63	.30	.02	.97	.36	.22	.04	.79
7	Keep	223	.44	.17	.15	.84	.57	.13	.25	.81
	Reserve	46	.41	.22	.10	.79	.48	.18	.20	.81
	Weak	51	.22	.22	.01	.71	.42	.26	-.07	.92

Table 2.5. CTT p-values and biserial correlations for reading grades 4 and 7 field-test item pool, spring 2022.

Grade	CTT Flag	# of Items	P–value				Biserial Correlation			
			Mean	SD	Min	Max	Mean	SD	Min	Max
4	Keep	366	.63	.13	.22	.85	.57	.10	.26	.78
	Reserve	63	.59	.17	.20	.90	.43	.15	.15	.79
	Weak	23	.41	.12	.17	.58	.19	.09	.05	.32
7	Keep	330	.59	.13	.24	.84	.54	.10	.26	.77
	Reserve	86	.53	.15	.12	.86	.37	.10	.15	.73
	Weak	58	.47	.12	.17	.71	.22	.09	-.10	.38

2.3. IRT-Based Item Analysis

Compared to CTT, which uses relatively weak assumptions based on the relationship between observed score, true score, and measurement error, IRT offers a more robust approach to item analysis. A limitation of CTT is that it focuses on properties of a given test and results are often group dependent (Hambleton, 2000; Yen & Fitzpatrick, 2006). The IRT-based item parameters, on the other hand, are assumed to be sample independent, and item performance is related to a latent trait called "ability" measured by the test (Anastasi & Urbina, 1997). IRT offers many features to the testing program that may be difficult to get with CTT, mostly because IRT defines a scale for the underlying ability on which both student performance and items can be placed. This aspect of IRT means comparable scores may be computed for examinees who did not take the same set of items without intermediate equating steps (Thissen & Orlando, 2001).

In IRT, a series of statistical models are used to describe the relationship between an individual's response to a single item and their ability based on the location of the item on the ability scale. All IRT models assume this relationship to be monotonic, meaning that as the ability level increases, the probability of a correct response also increases. According to Yen and Fitzpatrick (2006, p. 112), all IRT models can be classified by item type (e.g., dichotomous item, polytomous item), the number of abilities that an item measures (e.g., unidimensional, multidimensional), and the relationship between item and each ability.

The current field-test item pools for mathematics and reading were made up of all dichotomously scored (only two possible outcomes: correct or incorrect) items. Based on the current test format

and established NCDPI psychometric procedures, the unidimensional three-parameter logistic (3PL) model for multiple-choice and technology enhanced items and the two-parameter logistic (2PL) model for numeric or gridded response items were used to calibrate all field-test items. These two IRT models make three major assumptions:

- Unidimensionality: There is one dominant ability being measured (e.g., math ability for mathematics and reading ability for reading), and this single ability is the dominant factor accounting for variability in examinee's performance on each item.
- Local independence: Responses to different items on the test are independent given the underlying ability; that is, the correlation of responses to different items can be explained entirely by the ability being measured by the items.
- Measurement invariance: Item parameters are invariant to any group of subjects who have answered the items.

The mathematical function for the 3PL model is:

$$P_i(\theta) = c_i + \frac{1 - c_i}{1 + \exp\left[-Da_i(\theta - b_i)\right]},$$

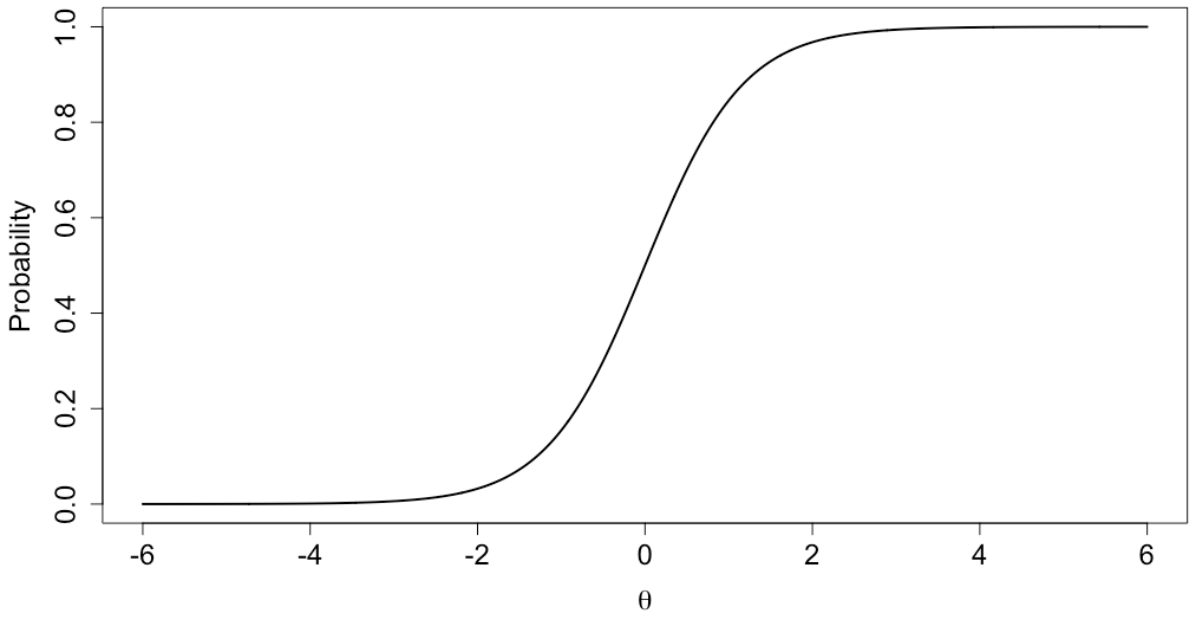
where $P_i(\theta)$ is the probability that an examinee with ability θ answers item i correctly; a_i is the discrimination parameter of item i; b_i is the difficulty parameter of item i; c_i is the guessing parameter for item i; and D is a scaling factor of either 1.0 or 1.7 (NCDPI sets D equal to 1.0). The major difference between the 2PL and 3PL models is that the 2PL model does not account for guessing. The 2PL model can be expressed as a special case of the 3PL model with $c_i = 0$ (see the equation below). For numeric and gridded response items, students are required to provide their answers rather than to select an answer from several choices, and therefore, the chance to get an item correct by guessing is near zero. The mathematical function for the 2PL model is:

$$P_i(\theta) = \frac{1}{1 + \exp[-Da_i(\theta - b_i)]}.$$

Once item parameters are estimated, a probabilistic relationship between each item along the ability continuum of -∞ to +∞ can be represented with a nonlinear monotonically increasing curve $P_i(\theta)$, which is often referred to as the item characteristic curve (ICC; Hambleton & Swaminathan,

1985). The ICC represent a summary figure, which can be used to evaluate the statistical properties for each item. Inferences about difficulty, discrimination, and guessing for each item can be made for examinees at different abilities along the ability continuum. Such inferences are critical during form assembly when items are selected to match a statistical target. An example of an ICC is shown in Figure 2.2. The vertical axis represents the probability of correct response, and the horizontal axis represents the underlying ability scale. In Figure 2.2, the item difficulty parameter is the ability level halfway between the lower and upper asymptotes, the item discrimination parameter is related to the slope of the ICC at the item difficulty parameter, and the guessing parameter is the value of the lower asymptote. Typically, the ability scale is set so that the mean and standard deviation of the abilities for the group at hand are 0 and 1, respectively. The ICC in Figure 2.2 shows an item with moderate difficulty in which an examinee with average ability of 0 will have a 50% chance to answer the item correctly.

Figure 2.2. Graphical representation of an item characteristic curve for the 2PL model.



To evaluate final item quality, the NCDPI has established several flagging criteria based on IRT item parameters to classify field-test items into one of the three categories (see Table 2.6). As stated earlier, the final item pool for form development is made of items flagged as psychometric "Keep" and "Reserve." During form assembly, however, priority is given to items with a "Keep" status.

Table 2.6. IRT-based items flagging criteria.

IRT Parameters	Flagging Criteria
Item Difficulty (b)	
$-2.500 \leq b \leq 2.500$	Keep
$-3.000 \leq b \leq -2.501$ or $2.501 \leq b \leq 3.000$	Reserve
$b \leq -3.001$ or $b \geq 3.001$	Weak
Item Discrimination (a)	
$1.190 \geq a$	Keep
$0.850 \leq a \leq 1.189$	Reserve
$a \leq 0.849$	Weak
Guessing (c)	
$c \leq 0.350$	Keep
$0.351 \leq c \leq 0.450$	Reserve
$c > 0.451$	Weak

2.4. IRT Parameter Estimation

IRT parameters of the embedded field-test items are estimated by calibrating item responses using the software IRTPRO® (Cai, Thissen, & du Toit, 2011) with a Bayesian prior for the guessing parameters (c) set to a beta distribution with shape parameters 5 and 15. This Bayesian prior ensures that the guessing parameter estimates in the 3PL model does not carry too far away from 0.25 for a four-option multiple-choice item. The IRT calibration phase to serve two main purposes:

- Form Development: The first main purpose of calibration is to develop an item bank with items of known statistical properties that are on the same IRT grade-level ability scale. These calibrated items expressed on the same IRT scale offers the NCDPI the flexibility to build multiple alternate forms without the need for traditional post equating.
- Scaling: The second purpose of calibration is to establish final IRT parameters for field-test items that are later used to create an IRT raw-to-scale score table for new forms before they are operationally administered. This is the essence of the NCDPI decentralized and immediate scoring for EOG assessments.

Prior to 2020, the NCDPI used random groups design to calibrate new field-test items for the purpose of building parallel forms for the same edition. The main assumption of the random groups design was that students' ability distributions were equivalent across years, and therefore,

calibrating IRT parameters independently each year by treating students as randomly equivalent groups will place all IRT parameters on the scale of the base year. However, due to the impact of COVID-19 that affected the normal functioning of schools and had significant negative impact on student performance, the students' ability distributions from EOG tests between pre- and post-COVID-19 were different. As a result, the item parameter estimates for the EOG tests obtained with pre- and post-COVID-19 student populations were no longer on the same scale. To handle this issue, especially for new field-test items, the NCPDI opted to change the item calibration method to fixed parameter calibration (Kim, 2006). Fixed parameter calibration fixes the parameter estimates for operational items to their existing values and calibrates the field-test items along with the ability distribution. This procedure ensures that parameter estimates from newly administered field-test items are placed on the same base scale as the operational items. The rationale is because the estimated ability distribution that is used to estimate the parameters for the field-test items is on the scale of the operational items. Thus, using fixed parameter calibration, the NCDPI was able to create an item bank with new field-test items that are on the base scale.

Tables 2.7 and 2.8 show descriptive statistics for the field-test item IRT parameters for mathematics and reading, respectively. The items flagged as "Keep" and "Reserve" were considered acceptable and made up the final item pool for form assembly.

Table 2.7. Fixed parameter calibration field-test item statistics for mathematics grades 4 and 7, spring 2022.

				Discrimination (*a*)				Difficulty (*b*)				Guessing (*c*)			
Grade	Flags	N	%	Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max
4	Keep	186	78	1.92	0.46	1.19	3.13	-0.19	0.67	-1.52	1.47	0.19	0.07	0.02	0.35
	Reserve	36	15	1.39	0.53	0.86	2.96	-0.15	1.11	-1.92	2.12	0.25	0.11	0.02	0.45
	Weak	18	8	1.66	1.26	0.12	4.47	-0.28	2.21	-3.66	3.60	0.33	0.15	0.17	0.62
7	Keep	223	70	2.23	0.56	1.22	3.70	0.28	0.66	-1.32	1.59	0.20	0.09	0.01	0.35
	Reserve	46	15	2.06	0.94	0.90	3.98	0.54	0.92	-1.48	1.93	0.24	0.13	0.01	0.41
	Weak	48	15	1.87	0.87	0.23	3.38	1.64	1.22	-1.59	3.93	0.24	0.15	0.01	0.52

Note. Items with CTT negative biserial correlations were excluded from IRT calibration; summary statistics for guessing (*c*) are based on only the items calibrated with the 3PL model.

Table 2.8. Fixed parameter calibration field-test items statistics for reading grades 4 and 7, spring 2022.

Grade	Flags	N	%	Discrimination (*a*)				Difficulty (*b*)				Guessing (*c*)			
				Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max
4	Keep	366	81	2.13	0.62	1.19	4.35	-0.22	0.57	-1.44	1.36	0.20	0.05	0.03	0.34
	Reserve	63	14	1.43	0.76	0.85	3.83	-0.06	1.02	-1.70	1.93	0.19	0.07	0.08	0.39
	Weak	23	5	0.99	0.72	0.36	3.05	1.44	1.06	-0.16	4.31	0.19	0.08	0.10	0.47
7	Keep	330	70	1.99	0.54	1.19	3.64	-0.10	0.60	-1.41	1.40	0.20	0.05	0.04	0.33
	Reserve	86	18	1.21	0.43	0.85	2.80	0.20	0.91	-1.49	2.42	0.18	0.06	0.09	0.38
	Weak	56	12	0.76	0.44	0.22	2.91	0.94	1.07	-1.16	3.35	0.19	0.05	0.10	0.30

Note. Items with CTT negative biserial correlations were excluded from IRT calibration; summary statistics for guessing (*c*) are based on only the items calibrated with the 3PL model.

2.5. Bias and Sensitivity DIF Analysis

The final step in item analysis is a statistical evaluation of potential bias. The Standard 3.3 (AERA, APA, & NCME, 2014) states "*Those responsible for test development should include relevant subgroups in validity, reliability/precision, and other preliminary studies used when constructing the test"* (p. 64). Statistical DIF procedure, which is often referred to as bias analysis, examines the degree to which students of various subgroups (e.g., males versus females) perform differently on an item. It is expected that students with the same latent trait level should have similar probability for answering items correctly, regardless of their background characteristics. An item is flagged as exhibiting DIF when students from different socioeconomic or demographic backgrounds with similar estimated knowledge and skill on the overall construct being measured perform substantially different on the same item (AERA, APA, & NCME, 2014). It is important to remember that the presence or absence of true bias is a qualitative decision based on the content of the item and the context within which it appears.

The NCDPI has adopted the Mantel-Haenszel (Holland & Thayer, 1988; Mantel & Haenszel, 1959) DIF detection method with the ETS Delta classification scheme to identify potential DIF items (Camilli & Sheppard, 1994) for further qualitative bias and sensitivity scrutiny by expert panels. The data used by the MH method are in the form of K 2×2 contingency tables, where K is the number of score levels on the matching variable. For the MH method, test score is typically used as the matching variable. Table 2.9 shows a 2×2 contingency table for the kth level of the

matching variable. In Table 2.9, A_k and C_k are the number of examinees in the reference and focal groups, respectively, who answer the studied item correctly, and B_j and D_j are the number of examinees in the reference and focal groups, respectively, who answer the item incorrectly.

Table 2.9. A 2×2 contingency table for the kth level of the matching variable.

Group	Score on Studied Item		Total
	1	0	
Reference (R)	A_k	B_k	n_{Rk}
Focal (F)	C_k	D_k	n_{Fk}
Total	m_{1k}	m_{0k}	T_k

The MH method tests the null hypothesis that the correct response probabilities for the studied item are equal between the reference and focal groups for all test scores; that is,

$$H_0: \frac{A_j}{n_{Rk}} = \frac{C_j}{n_{Fk}}, \quad k = 1, \dots, K.$$

Alternatively, the null hypothesis can be expressed in terms of odds ratio as follows:

$$H_0: \frac{A_k/B_k}{C_k/D_k} = \frac{A_k D_k}{B_k C_k} = 1, \quad k = 1, \dots, K.$$

Mantel and Haenszel (1959) also provided an overall odds ratio, which for a given studied item is defined by:

$$\alpha_{MH} = \frac{\sum_{k=1}^{K} A_k D_k / T_k}{\sum_{k=1}^{K} B_k C_k / T_k}.$$

The overall odds ratio α_{MH} is an estimate of DIF effect size that ranges from 0 to ∞, with a value of 1 indicating that the studied item is DIF free. In general, the natural logarithm of α_{MH} is used instead of α_{MH} because $\ln(\alpha_{MH})$ ranges from $-\infty$ to ∞ and is centered at 0.

As hypothesis tests are sensitive to sample size, NCDPI uses the ETS classification scheme (Holland & Thayer, 1985) instead, which is based on $\Delta_{\text{MH}} = -2.35 \ln(\alpha_{MH})$, to identify DIF items. The ETS classification scheme classifies items into the following three categories:

- A: An item with no substantial difference between the two matched groups. An item is identified as an "A" item if $|\Delta_{MH}| < 1.0$.
- B: An item with small to moderate differences between the two matched groups. An item is identified as a "B" item if $1.0 < |\Delta_{MH}| < 1.5$. An item with positive values of Δ_{MH} favors the focal group, whereas an item with negative values favors the reference group.
- C: An item with substantial differences between the two matched groups. An item is identified as a "C" item if $1.5 < |\Delta_{MH}|$. An item with positive values of Δ_{MH} favors the focal group, whereas an item with negative values favors the reference group.

All field-test items are quantitatively evaluated for DIF based on five main demographic and socioeconomic groupings identified by NCDPI:

- Demographic grouping:
    - Male (reference group) and Female (focal group)
    - White (reference group) and Black (focal group)
    - White (reference group) and Hispanic (focal group)
- Socioeconomic grouping:
    - Urban schools (reference group) and Rural schools (focal group)
    - Not Economically Disadvantaged (reference group) and Economically Disadvantaged (focal group)

Tables 2.10 and 2.11 show EOG mathematics and reading field-test item pool DIF results for the 2021–22 school year, respectively, based on the ETS classification scheme. For both subjects and grades, more than 90% of the items were classified as "A" items, indicating that the items are DIF free. The NCDPI rule is to exclude all items from the final pool that are flagged as "C" items, which indicate significant DIF. These items are either retired or sent back to the item writing process to undergo significant revisions and a new round of field tests and analysis. Items flagged as "B" are kept in the pool but will need to undergo further bias review by a panel if selected to be placed on a form. The panel decides whether the items are free of implied bias.

Table 2.10. Mantel-Haenszel DIF results for the mathematics grades 4 and 7 field-test items, spring 2022.

Grade	Gender			Ethnicity						EDS		
				White/Black			White/Hispanic					
	A	B	C	A	B	C	A	B	C	A	B	C
4	219	19	2	212	23	5	223	15	2	239	1	0
7	288	24	8	270	33	17	301	12	7	310	6	4

Note. EDS = Economically disadvantaged student.

Table 2.11. Mantel-Haenszel DIF results for the reading grades 4 and 7 field-test items, spring 2022.

Grade	Gender			Ethnicity						EDS		
				White/Black			White/Hispanic					
	A	B	C	A	B	C	A	B	C	A	B	C
4	425	24	3	425	26	1	413	30	9	447	5	0
7	450	22	2	462	8	4	448	22	4	473	1	0

Note. EDS = Economically disadvantaged student.

At the conclusion of item analysis based on field-test data, the final item pool for form assembly consisted of items flagged as psychometric “Keep” or “Reserve” and a DIF flagging classification of “A” or “B.” All items with field-test psychometric classification flag of “Weak” or DIF classification of “C” are excluded from consideration during form assembly.

CHAPTER 3. OPERATIONAL FORM ASSEMBLY, ANALYSIS, AND ROUTING PROCEDURE

AERA, APA, & NCME (2014) states, "*The test developer is responsible for documenting that the items selected for the test meet the requirements of the test specifications. In particular, the set of items selected for a new test form or an item pool for an adaptive test must meet both content and psychometric specifications*" (p. 82). To adhere to the Standard, this chapter documents the form assembly procedures that are used to create the multistage fixed adaptive forms for the NCPAT system.

3.1. Form Assembly

The NCPAT system is comprise of two main components: NC Check-Ins 2.0 interims and multistage fixed adaptive EOG. The NC Check-Ins 2.0 interim component consists of three interims with a primary purpose to provide teachers and students with immediate feedback on selected grade level standards so that instruction can be adjusted throughout the year. The multistage fixed adaptive EOG component consists of three levels that are made up of a common item set and a unique subset. Each student is only expected to complete one full set that consists of the common set and one unique subset. All three forms for each grade are built based on the same grade level blueprint used for traditional EOG forms (see the link provided in Appendix A for more details).

Major requirements of IADA are that states provide evidence to show their innovative system is valid, reliable, and comparable. The test specification is technically sound and aligns to the depth and breadth of content standards. In the current design of the NCPAT multistage fixed adaptive EOG forms, the goal was to align each form to the current grade level test specification. To accommodate the current EOG fixed form test specifications to fit an adaptive test design, the main components were divided into primary and secondary test specifications constraints.

- Primary test specification constraints - These are features of the multistage fixed adaptive EOG forms designed to align with current EOG fixed form test specifications. The following components were set as primary constraints during form assembly:

  - Content domain - Content domains for each form are aligned to EOG content blueprint. All three levels of the multistage fixed adaptive EOG form are aligned to grade level content standards adopted by NC State Board of Education. There is no off-grade level content for any of the levels. As with the regular EOG all multistage forms are grade level specific.
  - Test length - For each subject and grade level, test lengths for all three levels of the multistage fixed adaptive EOG form are fixed and matched the length of the current EOG fixed forms.
  - Test format - For mathematics, the number of calculator active items is set equal to current EOG specifications. For reading, the distribution of selection type (Information and Literature) matches current EOG specifications.

- Secondary test specification constraints - These are features of the multistage fixed adaptive EOG forms that were allowed to vary across the three levels during test assembly. The following components of the multistage fixed adaptive EOG form specifications are categorized as secondary constraints:
  - Cognitive complexity - For each level of the multistage fixed adaptive EOG form, the Depth of Knowledge (DOK) ranges are allowed a certain degree of variability. The design goal is for DOK to progress from form levels aligned to the lower end of the scale to form levels aligned to the upper end of the scale while at the same time not exceeding the maximum range of DOK expectation specified in the grade level blueprint.
  - Form difficulty - The current design of the multistage fixed adaptive EOG has three levels currently labeled as A, B and C. Each level is set to a different statistical target. The statistical target for level A is designed to maximize information at the lower range of the ability scale. Level B maximizes information in the middle range of the scale and is statistically equivalent to the current EOG fixed forms. Level C is designed to maximize information at the higher range of the scale. The three levels of the multistage fixed adaptive EOG will hereinafter be referred to as Forms A, B, and C.
  - Item type - Item type can differ from form to form with MC items being the dominant item type.

- Standards within domain – The distribution of items by standard varies across forms.

An additional design constraint for all multistage fixed adaptive EOG forms within each grade is that the three forms share 40-50% of the total items. These items are selected to closely match the average statistical property of the base year EOG fixed form. Complete form summary of the multistage fixed adaptive EOG forms is presented in Appendices B, C, D and E, which show the alignment of each form to the content level content and specification blueprint. The item statistics of the three forms are provided in Tables 3.1 and 3.2 for mathematics and reading, respectively.

Table 3.1. Item statistics of the three multistage fixed adaptive EOG forms for mathematics grades 4 and 7.

		Discrimination (*a*)				Difficulty (*b*)				Guessing (*c*)			
Grade	Form	Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max
4	A	1.84	0.45	1.07	2.97	-0.56	0.58	-1.48	0.81	0.18	0.08	0.01	0.37
	B	1.97	0.42	1.14	2.97	-0.23	0.65	-1.48	1.22	0.16	0.10	0.01	0.46
	C	1.89	0.48	1.10	2.97	0.22	0.74	-1.48	1.90	0.19	0.09	0.01	0.45
7	A	2.07	0.52	0.90	3.31	-0.19	0.51	-1.32	0.95	0.20	0.09	0.03	0.35
	B	2.12	0.51	1.07	3.57	0.11	0.52	-1.13	1.42	0.22	0.08	0.06	0.44
	C	2.18	0.57	1.22	3.26	0.42	0.55	-0.64	1.66	0.22	0.08	0.06	0.37

Note. Item statistics for guessing (*c*) are based on only the items calibrated with the 3PL model.

Table 3.2. Item statistics of the three multistage fixed adaptive EOG forms for reading grades 4 and 7.

		Discrimination (*a*)				Difficulty (*b*)				Guessing (*c*)			
Grade	Form	Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max
4	A	1.93	0.61	1.07	4.11	-0.39	0.72	-1.70	1.21	0.19	0.06	0.03	0.39
	B	1.91	0.63	0.90	4.32	-0.14	0.76	-1.42	1.34	0.18	0.06	0.03	0.31
	C	1.91	0.58	0.99	3.46	0.21	0.73	-1.13	1.93	0.20	0.07	0.03	0.36
7	A	1.83	0.54	0.93	3.05	-0.24	0.68	-1.44	1.75	0.18	0.06	0.07	0.33
	B	1.89	0.61	1.03	3.26	0.02	0.72	-0.89	2.05	0.20	0.06	0.08	0.32
	C	1.80	0.52	0.99	3.14	0.27	0.67	-0.89	1.75	0.20	0.06	0.11	0.32

Note. Item statistics for guessing (*c*) are based on only the items calibrated with the 3PL model.

3.1.1. Test Characteristic Curve

In IRT, test characteristic curves (TCCs) are essential for form assembly and scaling. A TCC is an 'S-shaped' curve with flatter ends that depict the expected summed score as a function of ability (Thissen, Nelson, Rosa, & Mcleod, 2001). Mathematically, a TCC is the sum of ICCs over all items in a given test; that is,

$$\mathrm{TCC}(\theta) = \sum_{i=1}^{n} P_i(\theta),$$

where n is the number of items; θ is a given ability level; and $P_i(\theta)$ is the ICC for item i (see Item Analysis chapter for description of ICC). The TCCs for the three multistage fixed adaptive EOG forms are provided in Figures 3.1 and 3.2 for mathematics and reading, respectively. The location of TCCs for the three forms varies because they are designed to maximize information at different parts of the ability scale. For example, the TCC for Form A is located on the left side of the ability scale to better serve students with abilities located on the lower end of the ability scale, whereas the TCC for Form C is located on the right side of the scale to better serve students with abilities located on the upper end of the scale.

Figure 3.1. TCCs of the three multistage fixed adaptive EOG forms for mathematics grades 4 and 7.

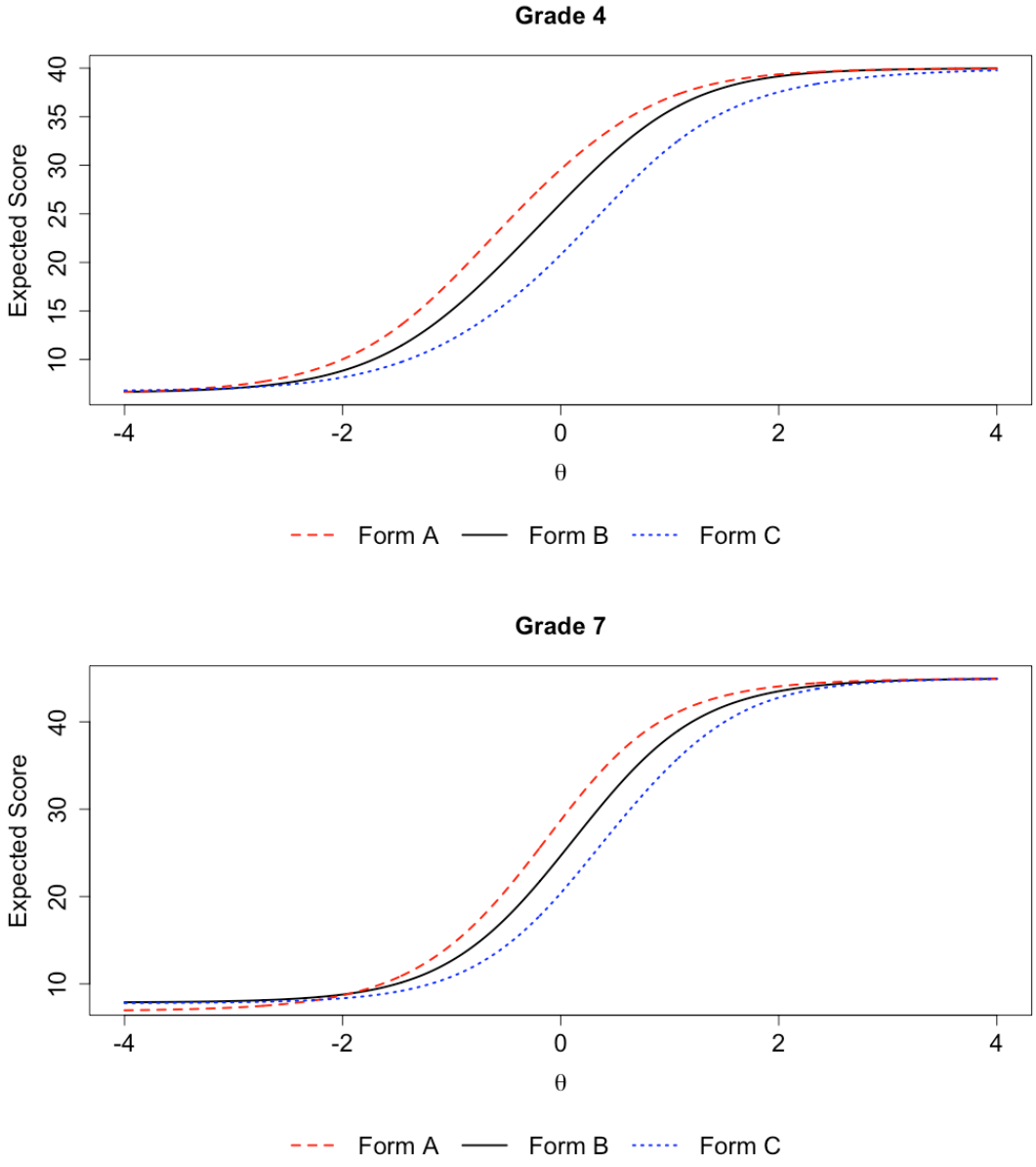
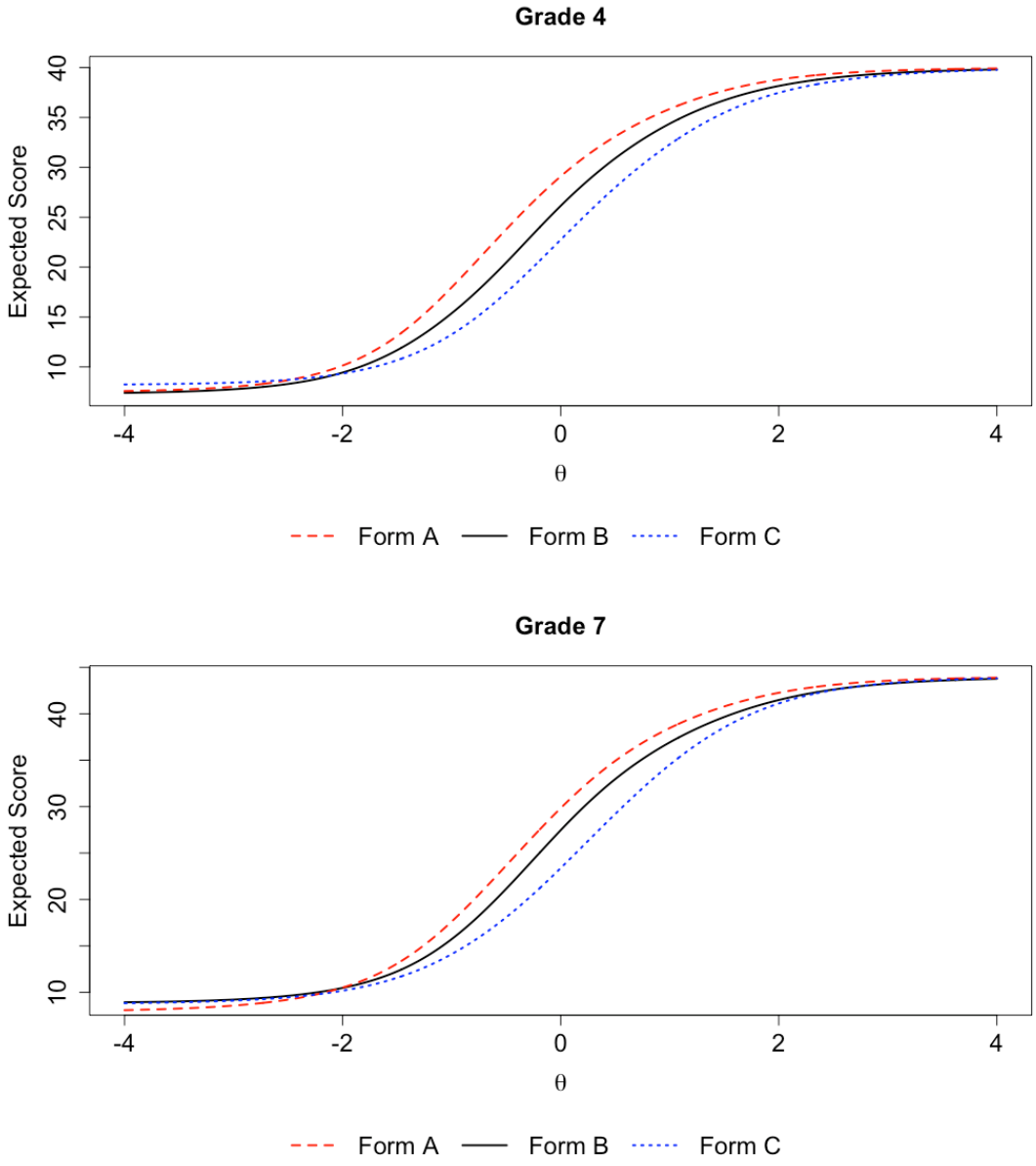


Figure 3.2. TCCs of the three multistage fixed adaptive EOG forms for reading grades 4 and 7.



3.1.1. Test Information Function and Conditional Standard Error of Measurement

The concept of test reliability is central in CTT when evaluating the overall consistency of scores over repeated measurement. A concept related to reliability in CTT is the standard error of measurement (SEM), which is the standard deviation of the observed scores around the true score over repeated measurement. Reliability and SEM in CTT are not a property of a specific test

because they are both test and population dependent. In addition, it is generally assumed that SEM is identical for all examinees regardless of where they are located on the score scale. However, examinees with different response patterns or at different points on the score scale might show variations in the amount of measurement precision. In IRT, the test information function (TIF) is a similar concept to reliability but provides local measures of accuracy. Specifically, the TIF provides how much information the test provides in estimating ability over the entire ability scale and is defined by:

$$I(\theta) = \sum_{i=1}^{n} \frac{[P'_i(\theta)]^2}{P_i(\theta)Q_i(\theta)},$$

where $P_i(\theta)$ is the ICC for item i at a given ability level θ; $Q_i(\theta) = 1 - P_i(\theta)$; and $P_i'(\theta)$ is the first derivative of $P_i(\theta)$. For more information about the TIF, see Hambleton and Swaminathan (1985), and Thissen and Orlando (2001).

Using the TIF, the measurement precision at a given value θ, which is often referred to as the conditional SEM (CSEM), is computed as follows:

$$\text{CSEM}(\theta) = \frac{1}{\sqrt{I(\theta)}}.$$

If a test provides more information in estimating ability at a given θ, then the CSEM at that θ will be small. As depicted in the left panels of Figures 3.3 and 3.4, for both mathematics and reading, Form A provides more information below the mean ability of 0, Form B provides more information at abilities around the mean, and Form C provides higher information at abilities above the mean. To put it differently, Form A, B, and C provide more precise ability estimates (i.e., ability estimates with smaller CSEMs) on the lower end, in the middle, and on the upper end of the ability scale, respectively.

Figure 3.3. TIFs (left panels) and CSEMs (right panels) of the three multistage fixed adaptive EOG forms for mathematics grades 4 and 7.

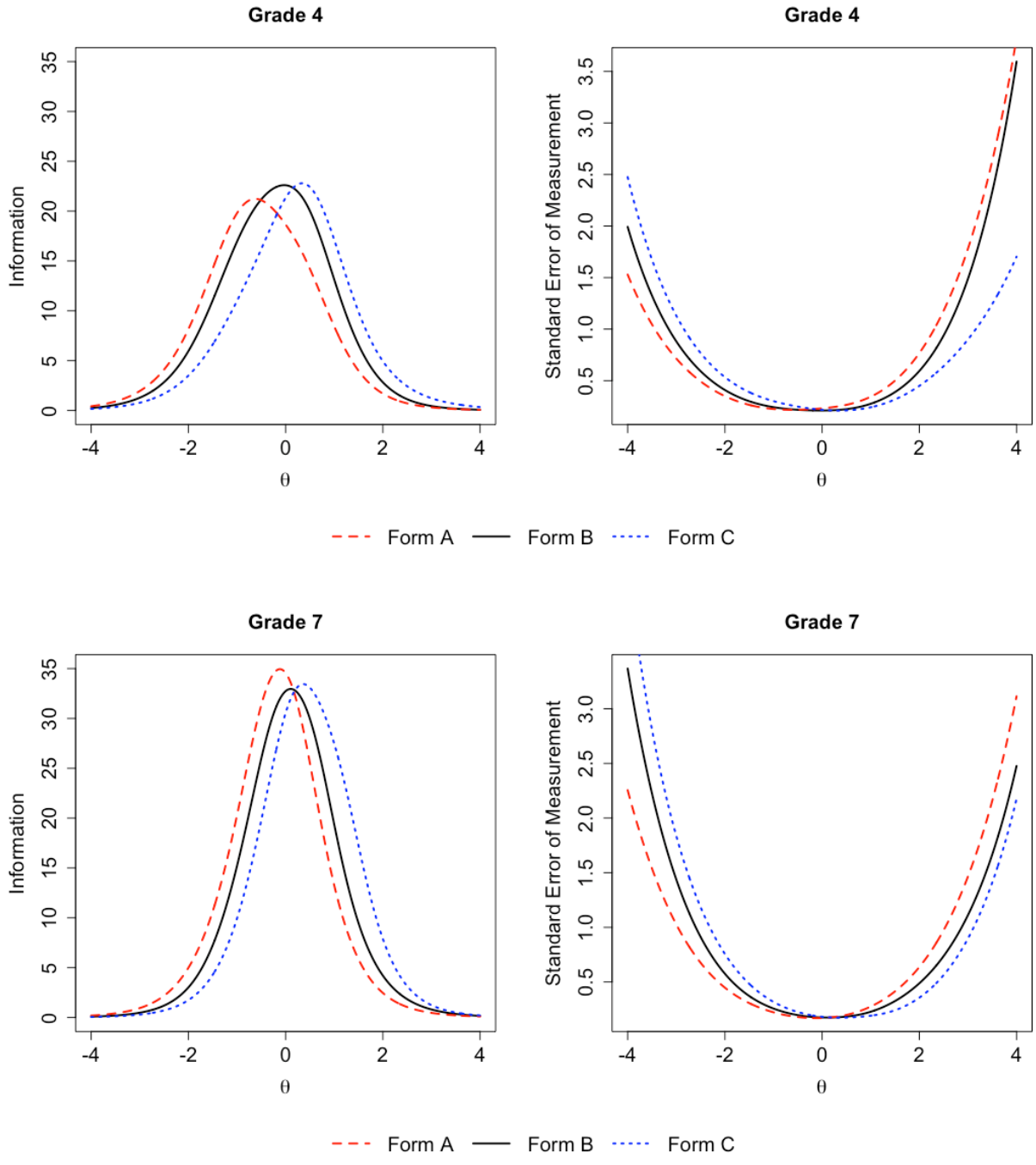
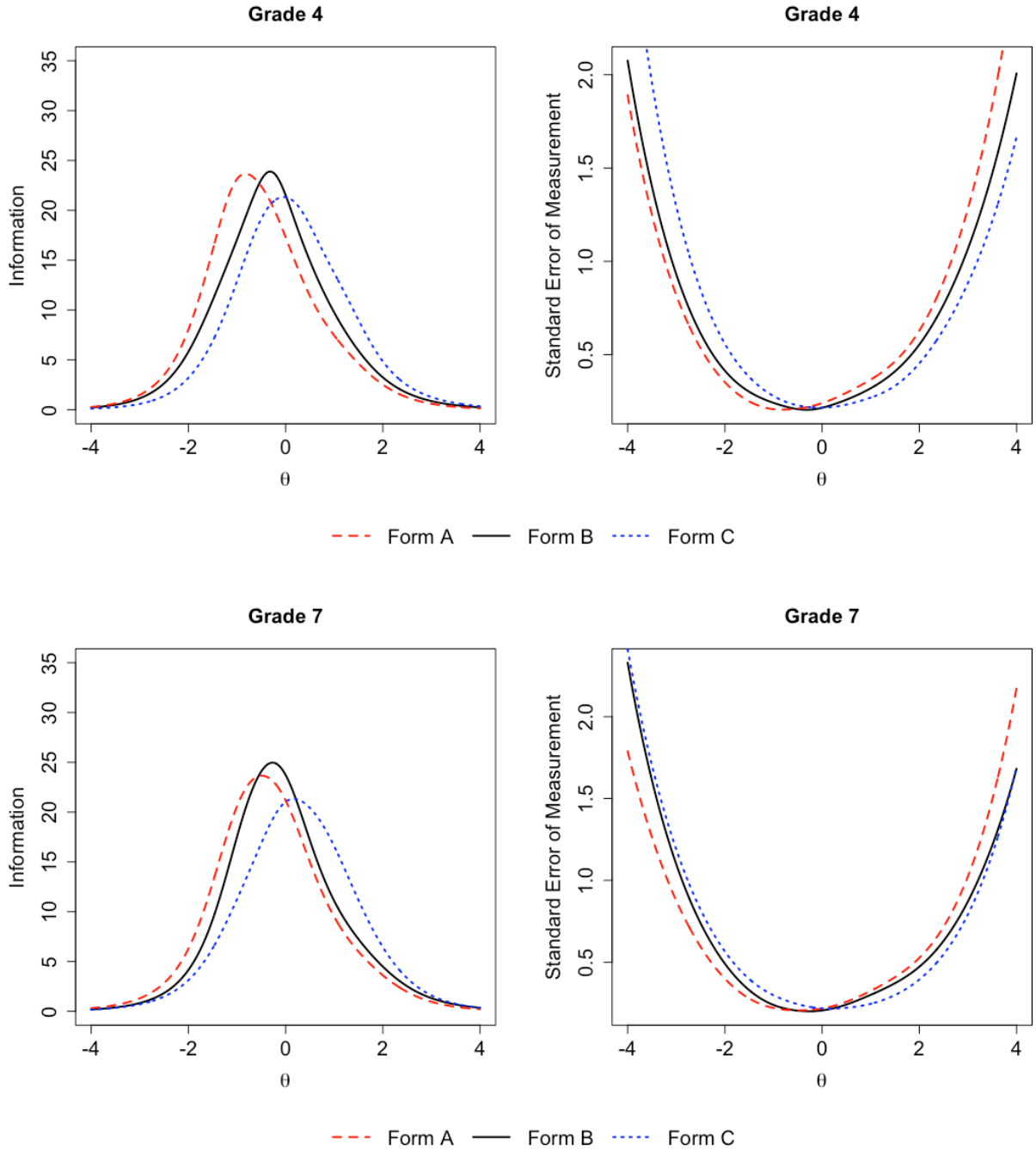


Figure 3.4. TIFs (left panels) and CSEMs (right panels) of the three multistage fixed adaptive EOG forms for reading grades 4 and 7.



3.2. Routing Procedures

This section outlines the procedures used to obtain cut scores and route students to the three multistage fixed adaptive EOG forms for the NCPAT system. Data from NC Check-Ins 2.0 serve as a reliable prior of student ability for the multistage fixed adaptive EOG forms. The goal of routing is to ensure students are assigned to the form that best matches their expected ability and is tailored to enhance testing experience. The plan is to continuously monitor the relationship between the NC Check-Ins 2.0 and multistage fixed adaptive EOG forms and adjust the routing methodology as needed.

3.2.1. Routing Method

The three multistage fixed adaptive EOG forms are designed to enhance students' testing experience and improve measurement precision across the entire ability scale. Students are routed to one of the three forms based on their performance on the NC Check-Ins 2.0. Among a variety of routing methods, the *defined population intervals* method (Luecht, Brumfield, & Breithaupt, 2006) is used for the NCPAT system, which assigns a predefined proportion of students to each of the three forms. The proportion currently used for the NCPAT system is 30%-40%-30%; that is, approximately 30% of the students are routed to Form A, approximately 40% of the students are routed to Form B, and approximately 30% of the students are routed to Form C. The procedures used to obtain students' scores and cut scores for the NCPAT system are described below.

3.2.2. Scores Used for Routing

Scores for the NC Check-Ins 2.0 are used to route students to the three multistage fixed adaptive EOG forms. For each grade, three NC Check-Ins 2.0 are administered throughout the school year; therefore, students can have from one up to three NC Check-In 2.0 scores. When multiple scores are available, the average score is used for routing, which is computed as follows:

- For students with all three NC Check-In 2.0 scores, the average of the two highest scores is used for routing.
- For students with two NC Check-In 2.0 scores, the average of the two scores is used for routing.

- Students with no or only one NC Check-In 2.0 score are assigned to form B as it is determined that they do not have enough information to be reliably routed to either Form A or C.

3.2.3. Cut Scores for Routing

Two cut scores are generated to route students to one of the three multistage fixed adaptive EOG forms. An IRT based methodology is used in the base years to create a relationship based on students estimated ability (θ) scores on the NC Check-Ins 2.0. Student response data from all valid NC Check-Ins 2.0 are concurrently calibrated and scored to obtain a θ estimate for each student using the maximum likelihood (ML) estimation method. As the ML estimation method cannot estimate θ scores for students with certain response patterns, such as all incorrect or correct responses, the minimum and maximum θ scores are set to -6 and 6, respectively. Then, the 30^{th} and 70^{th} quantiles of the θ scores are set as the two θ cuts. These cuts are then converted to (expected) raw scores by first combining the three NC Check-Ins 2.0 item parameter estimates to obtain a single TCC, and then finding the raw scores that correspond to the two θ cuts from the ICC. The final raw cuts are converted back to the scale of a single NC Check-In 2.0 by dividing the raw cuts by the total number of NC Check-Ins 2.0, which is three. This is because students are routed based on their average scores on two NC Check-Ins 2.0, not their total scores.

3.2.4. Routing to the Multistage Fixed Adaptive Forms

Comparing students' average scores on the NC Check-Ins 2.0 and the two cut scores, say C_1 and C_2, students are routed to one of the three multistage fixed adaptive forms as follows:

- Students whose average NC Check-In 2.0 scores are equal to or less than C_1 are routed to Form A.
- Students whose average NC Check-In 2.0 scores are higher than C_1 and equal to or less than C_2 are routed to Form B.
- Students whose average NC Check-In 2.0 scores are higher than C_2 are routed to Form C.
- Students with a single or missing NC Check-In 2.0 score are assigned to Form B

CHAPTER 4. SCORING AND SCALE COMPARABILITY

This chapter summarizes procedures used to generate IRT summed to scale scores for the multistage fixed adaptive forms that are on the same scale as traditional EOG. The evidence documented in this chapter serves as validity evidence to show scores from the multistage adaptive forms are comparable and thus reported on the same EOG grade level scale. In other words, the summed-to-scale score tables for each multistage form were created using IRT item parameters that were placed on the base year EOG scale through the fixed parameter calibration procedure. The scores generated from the NCPAT multistage forms are statistically equivalent to those from the traditional EOG forms, and both scores are reliable and can be used and interpreted as valid measure of student performance on grade level content expectations.

4.1. IRT Scoring and Scale Scores

The NCDPI uses an IRT summed-to-scale score procedure for form level scoring and transforming student number correct responses into reportable scale scores. The scoring tables for converting number correct scores to scale scores are generally established after form development and review are complete and before test forms are operationally administered to students. This process of establishing scoring tables for test forms before the forms are administered operationally to students is referred to as a pre-equated scoring model. The use of pre-equated scoring model in North Carolina dates to the early 1990s and remains an important feature in the NCDPI grades 3–8 and high school state assessment program. The use of this model allows the NCDPI to take full advantage of test design properties offered through IRT while also allowing for decentralized scoring system based on number correct. Another practical consequence is that the NCDPI can use a short administration window for EOG that is usually the last 5–10 days of the school year and is still able to provide and use scores for end of year reporting.

4.1.1. IRT Summed Score Procedure

For the multistage forms, field-test items were calibrated using the fixed parameter calibration procedure, which allowed field test items administered in 2021-22 to be placed on the base scale used for EOG form building and scoring. With both sets of item parameters on the same scale, scoring was done using the IRT summed-to-scale score procedure to produce final raw-to-scale

conversion tables. Two main advantages of using IRT-based scale scores over scale scores based on raw scores are that:

- IRT-based scale scores provide a standard metric to report scores from multiple test forms, particularly in the case of adaptive test design where students are by design administered different levels of items based on their routed outcome. Fixed parameter calibration allows for the development of independent summed score tables using item parameters that are on the same established grade level scale that are comparable and offer the same meaning. IRT enables the continuous development, calibration and, scoring of new forms on the same existing IRT scale.
- By reporting on a common scale, performance from students who took the multistage and traditional EOG can be fairly compared without the need for any additional score adjustments. Separate raw-to-scale tables for each form accounts for statistical differences across NCPAT multistage forms. Students are neither penalized nor gain unfair advantage based on final form that was assigned to them.

Estimates of students' ability from NCPAT multistage adaptive forms were derived from number correct scores using IRT summed to scale score procedure based on expected a posteriori (EAP) ability estimate. These EAP estimates were then transformed and reported using an NCDPI custom scale metric. Following Standard 5.2 (AERA, APA, & NCME, 2014), "*the procedures for constructing scales used for reporting scores and the rationale for these procedures should be described clearly*" (p.102), the IRT summed-to-scale score procedures used to derive student ability estimates from number correct scores are outlined below. For reference of full description of the IRT summed to scale score procedure, see Thissen and Orlando (2001, p.119). For any IRT model for items indexed by i with item scores $u_i = 0$ or 1, the likelihood function for summed score $x = \sum u_i$ is:

$$L(\theta \mid x) = \sum_{(u_i)=x} L(\boldsymbol{u} \mid \theta)$$

where $(u_i) = x$ is all response patterns that produce summed score x; and

$$L(\boldsymbol{u} \mid \theta) = \prod_i P_i(\theta)^{u_i}\big(1 - P_i(\theta)\big)^{1-u_i}.$$

Here $P_i(\theta)$ is the correct response probability for item i at ability level θ. The EAP estimate based on summed score is:

$$EAP(\theta \mid x) = \int_{-\infty}^{\infty} \theta \frac{L(\theta \mid x)\phi(\theta)}{\int_{-\infty}^{\infty}(\theta \mid x)\phi(\theta)\, d\theta} d\theta$$

where $\phi(\theta)$ is the probability density function of θ, which is often set a standard normal distribution. The corresponding standard deviation is:

$$SD(\theta \mid x) = \sqrt{\int_{-\infty}^{\infty} [\theta - \mathrm{EAP}(\theta \mid x)]^2 \frac{L(\theta \mid x)\phi(\theta)}{\int_{-\infty}^{\infty}(\theta \mid x)\phi(\theta)\, d\theta} d\theta}\,.$$

The values of $EAP(\theta \mid x)$ were used to transform summed scores to IRT θ scores, and the values of $SD(\theta \mid x)$ were used as a measure of the uncertainty associated with those θ scores.

Scoring was done in IRTPRO® using banked item parameters from the base year scale obtained using the fixed parameter calibration procedure to estimate EAP scores. To ensure students ability estimates from new parallel forms are placed on the same common IRT scale, the population density distribution (mean and standard deviation) used to generate final scale scores were set to the base year default values used for each EOG.

4.2. Scale Comparability

Scale comparability assures users that students with the same scale possessed the same level of proficiency with respect to the domain of knowledge and skills that a test was intended to measure (Perie, 2020). Some basic principles regarding the degree of comparability pertains to similarity of the assessments content, administration conditions, and the psychometric properties of the assessments. As documented in previous chapters of this report, the multistage forms are based on the same EOG grade level content blueprint. From a content perspective, all forms of the NCPAT system and traditional EOG are equivalent. Statistically, the emphasis of measurement precision and information provided by each form varies across the level of the multistage forms as would be expected under an adaptive form design. The use of IRT parameters that are calibrated onto the base scale to separately score each form to create summed-to-scale scores tables allow for reported

scale scores to be equivalent and on the same scale. The content similarities together with the IRT scoring procedure used are sufficient evidence to support the claim that scale scores for students who participated in the NCPAT multistage pilot are comparable to scale scores reported for students who participated in the traditional EOG. Additional NCDPI also conducted a propensity score matching study at the student level to look at investigate comparability of scale scores and achievement levels between the multistage fixed adaptive EOG administered as a pilot under the NCPAT system and the general EOG administered to students across the state. The goal was to document evidence to show scores from students who took the multistage forms are generally consistent with scores from an equivalent subset of students who took the EOG. The study with results is provided in Appendix F.

CHAPTER 5. VALIDITY EVIDENCE

This chapter presents additional validity evidence collected in support of the interpretation of test scores for the multistage fixed adaptive forms. Evidence presented in these sections include reliability, standard error of measurement at the achievement level cuts, classification consistency and accuracy of reported achievement levels, principal component analysis (PCA), and confirmatory factor analysis (CFA) to support the unidimensional interpretation of test scores.

5.1. Marginal Reliability

Reliability provides a sample-based summary statistic that describes the proportion of the reported score variability that is attributed to true score variance. To justify valid use of test results in large-scale standardized assessments, evidence must be documented that shows test results are stable, consistent, and dependable for the intended population. A reliable assessment produces scores that are expected to be relatively stable if the test is administered repeatedly under similar conditions to the same students. Scores from a reliable test reflect examinees' estimated ability on the construct being measured with very little measurement error. Reliability coefficients range from 0 to 1, where a coefficient of 1 refers to a perfectly reliable measure with no measurement error. In IRT, the concept of measurement precision/reliability is conditional on location of the underlying ability scale. Instead of a single value used to summarize measurement precision as is the case with CTT, IRT allows for varying degree of precision along the scale. In the context of adaptive test design, the definition of a single value to summarize precision is further complicated by the fact that homogenous subset of students is administered separate subset of items. In the current context of multistage fixed forms, homogenous subset of students is routed to specific form levels with the intent to maximize measurement precision. The disadvantage of classical reliability is that it is dependent on the heterogeneity of samples; the same test will have a higher value of reliability with a heterogenous examinee sample than a homogenous sample (Thissen, 2000).

For the multistage adaptive design where the assumption of constant error variance has been violated, marginal reliability is estimated by averaging over conditional standard error of measurement across all levels. This generates comparable reliability estimates that can be

interpreted in the traditional sense. The marginal reliability ($\bar{\rho}$) for the multistage fixed adaptive EOG forms is defined by:

$$\bar{\rho} = 1 - \frac{(\sum_{i=1}^{N} CSEM_i^2/N)}{\sigma^2},$$

where σ^2 is the scale score variance, $CSEM_i^2$ is the CSEM for examinee i, and N is the total number of examinees. For high-stakes assessments, reliability estimates of 0.85 or higher are generally deemed desirable.

Tables 5.1 to 5.4 show the marginal reliability estimates for the three multistage fixed adaptive EOG forms for mathematics and reading. For both subjects and grades, overall marginal reliability estimates ranged from 0.86 to 0.91. With a few exceptions, subgroup marginal reliabilities were also consistently higher than 0.85. Some subgroups showed reliability estimates lower than 0.85 because of their small variability of the observed scores (denominator of the reliability formula), not due to large measurement error. The CSEMs for those subgroups were similar to the subgroups with reliability estimates higher than 0.85.

Table 5.1. Marginal Reliability and standard error of measurement by subgroup for mathematics grade 4.

Grade	Group		N	Mean	SD	Min	Max	Reliability	SEM
4	All		9,090	545.9	10.4	522	570	0.91	3.08
	Sex	Female	4,430	545.6	10.1	522	570	0.91	3.08
		Male	4,545	546.5	10.6	525	570	0.91	3.07
	Ethnicity	Black	2,119	541.6	9.5	522	570	0.88	3.33
		Hispanic	1,734	545.3	9.8	525	570	0.90	3.04
		Other	902	544.6	10.5	525	570	0.91	3.19
		White	4,335	548.6	10.1	522	570	0.92	2.94
	EDS	Yes	5,678	543.6	9.8	522	570	0.89	3.18
	ELS	Yes	872	542.6	9.4	525	570	0.88	3.21
	SWD	Yes	1,856	541.0	10.1	522	570	0.88	3.43

Note. Reliability = Marginal reliability estimates for all multistage forms; reliability estimates are displayed only for major ethnic groups and accommodations investigated in DIF analysis with acceptable sample size; SEM is the average CSEM across all examinees.

Table 5.2. Marginal reliability and standard error of measurement by subgroup for mathematics Grade 7.

Grade	Group		N	Mean	SD	Min	Max	Reliability	SEM
7	All		9,233	544.5	9.3	526	573	0.86	3.43
	Sex	Female	4,442	544.4	9.1	526	573	0.86	3.42
		Male	4,685	544.8	9.5	526	573	0.87	3.44
	Ethnicity	Black	2,247	540.5	7.2	527	573	0.72	3.84
		Hispanic	1,852	543.6	8.9	526	571	0.85	3.48
		Other	866	543.5	9.3	528	573	0.85	3.56
		White	4,268	547.3	9.5	526	573	0.89	3.15
	EDS	Yes	5,584	542.1	8.2	526	571	0.80	3.65
	ELS	Yes	706	538.8	6.7	526	567	0.65	3.99
	SWD	Yes	1,571	539.2	7.9	526	573	0.74	4.07

Note. Reliability = Marginal reliability estimates for all multistage forms; reliability estimates are displayed only for major ethnic groups and accommodations investigated in DIF analysis with acceptable sample size; SEM is the average CSEM across all examinees.

Table 5.3. Marginal reliability and standard error of measurement by subgroup for reading Grade 4.

Grade	Group		N	Mean	SD	Min	Max	Reliability	SEM
4	All		9,099	541.4	10.2	515	567	0.89	3.41
	Sex	Female	4433	542.0	9.9	516	567	0.88	3.38
		Male	4551	541.1	10.5	515	567	0.89	3.44
	Ethnicity	Black	2121	538.1	9.7	515	567	0.87	3.52
		Hispanic	1737	539.7	9.9	518	565	0.88	3.46
		Other	902	540.7	10.3	518	565	0.89	3.46
		White	4339	543.9	9.9	516	567	0.89	3.33
	EDS	Yes	5685	539.2	9.8	515	567	0.87	3.46
	ELS	Yes	873	535.7	9.2	518	562	0.84	3.65
	SWD	Yes	1861	535.4	10.2	517	567	0.87	3.72

Note. Reliability = Marginal reliability estimates for all multistage forms; reliability estimates are displayed only for major ethnic groups and accommodations investigated in DIF analysis with acceptable sample size; SEM is the average CSEM across all examinees.

Table 5.4. Marginal reliability and standard error of measurement by subgroup for reading Grade 7

Grade	Group		N	Mean	SD	Min	Max	Reliability	SEM
7	All		9,273	550.8	9.8	528	580	0.87	3.50
	Sex	Female	4,465	551.3	9.4	528	579	0.87	3.44
		Male	4,701	550.4	10.2	528	580	0.88	3.56
	Ethnicity	Black	2,258	547.0	8.7	528	579	0.82	3.68
		Hispanic	1,864	549.2	9.4	528	580	0.86	3.56
		Other	868	549.5	9.7	529	576	0.86	3.56
		White	4,283	553.7	9.8	528	579	0.88	3.37
	EDS	Yes	5,618	548.5	9.2	528	579	0.85	3.60
	ELS	Yes	710	542.3	6.7	528	566	0.64	4.03
	SWD	Yes	1,582	544.2	9.4	528	579	0.82	4.00

Note. Reliability = Marginal reliability estimates for all multistage forms; reliability estimates are displayed only for major ethnic groups and accommodations investigated in DIF analysis with acceptable sample size; SEM is the average CSEM across all examinees.

5.2. Conditional Standard Error of Measurement at Scale Score Cuts

The information provided by the CSEM at a given cut score is important because it helps determine the accuracy of examinees' classifications to the four achievement levels, which are Not Proficient, Level 3, Level 4, and Level 5. The CSEMs at the lowest obtainable scale score (LOSS), highest obtainable scale score (HOSS), and the three scale score cuts that divide the four achievement levels are provided in Tables 5.5 and 5.6 for mathematics and reading, respectively. Among the three multistage fixed adaptive EOG forms, Form A, which is designed to maximize information at the lower end of the score scale, tends to provide smaller CSEMs at the Level 3 cut. This indicates that more accurate decisions can be made for students whose scale scores are near the Level 3 cut than the Level 4 and 5 cuts. On the other hand, Form C is designed to maximize information at the higher end of the scale and tends to provide smaller CSEMs at the Level 4 and Level 5 cuts. Although CSEMs for Form B are comparable to those for Form A at the Level 3 cut and to those for Form C at the Level 4 and Level 5 cuts, Forms A and C provide smaller CSEMs on the lower and upper ends of the ability scale than Form B, respectively (see Figures 3.3 and 3.4). CSEMs at the LOSS and HOSS are much larger than those at the three cut scores. The higher CSEMs at the LOSS and HOSS are typical because extreme scores have less measurement precision due to the lack of informative items at those score ranges.

Table 5.5. CSEMs at achievement level cuts for mathematics grades 4 and 7.

Grade	Form	Min		Level 3		Level 4		Level 5		Max	
		LOSS	SE	Cut	SE	Cut	SE	Cut	SE	HOSS	SE
4	A	522	5	547	2	552	3	560	4	570	5
	B	525	5	547	2	552	2	560	3	570	5
	C	527	5	547	3	552	2	560	3	570	5
7	A	526	5	546	2	550	2	560	3	573	5
	B	529	5	546	2	550	2	560	2	573	5
	C	530	5	546	3	550	2	560	2	573	5

Table 5.6. CSEMs at achievement level cuts for reading grades 4 and 7.

Grade	Form	Min		Level 3		Level 4		Level 5		Max	
		LOSS	SE	Cut	SE	Cut	SE	Cut	SE	HOSS	SE
4	A	515	5	544	3	548	3	557	4	567	6
	B	516	5	544	3	548	3	557	4	568	6
	C	520	6	544	3	548	3	557	3	568	6
7	A	526	5	554	3	559	3	567	4	580	6
	B	528	5	554	3	559	3	566	4	580	6
	C	529	6	554	3	559	3	566	3	580	5

5.3. Classification Consistency and Accuracy

The No Child Left Behind Act of 2001 (USED, 2002) and subsequent Race to the Top Act of 2009 (2009) emphasized the measurement of adequate yearly progress (AYP) with respect to the percentage of students at or above performance standards set by states. With this emphasis on the achievement level classification, it is very important to provide evidence that shows all students are consistently and accurately classified into one of the four achievement levels. The importance of classification consistency as a measure of the categorical decisions when the test is used repeatedly has been recognized in Standard 2.16 (AERA, APA, & NCME, 2014), which states, "*When a test or combination of measures is used to make categorical decisions, estimates should be provided of the percentage of examinees who would be classified in the same way on two applications of the procedure*" (p. 46).

Classification consistency refers to "the agreement between classifications based on two non-overlapping, equally difficult forms of the test," and classification accuracy refers to "the extent to which the actual classifications of test takers (on the basis of a single test score) agree with those

that would be made on the basis of their true scores, if their true scores could somehow be known" (Livingston & Lewis, 1995, p. 178). That is, classification consistency refers to the agreement between two observed scores, while classification accuracy refers to the agreement between observed and true scores. The classification indices for the three multistage fixed adaptive EOG forms were computed using an IRT-based method presented by Lee et al. (2002). This method provides coherent formulas for both consistency and accuracy for multiple classifications—classifying individuals into one of three or more categories, such as achievement levels based on a set of cut scores.

Tables 5.7 and 5.8 show the classification indices for mathematics and reading, respectively. The classification indices for both subjects and grades are fairly high, with a few exceptions at the Level 5 cut for reading. Note that the values in Tables 5.7 and 5.8 are the classification consistency and accuracy that examinees will be classified into the same category below or above the cut score for Level 3, 4, or 5. For example, Grade 7 Mathematics for Flex B has a consistency value of 0.90 (bolded). This means that if a student takes Form B twice, there is a 90% probability that the student will be classified into the same category, either not proficient or proficient (Levels 3, 4, and 5), both times. To put it differently, the probability of misclassifying a student into either the not proficient or proficient category is about 10%.

Table 5.7. Classification consistency and accuracy for mathematics grades 4 and 7.

Grade	Form	Level 3		Level 4		Level 5	
		CC	CA	CC	CA	CC	CA
4	A	0.96	0.97	0.98	0.99	0.99	0.99
	B	0.90	0.93	0.90	0.93	0.92	0.94
	C	0.99	0.99	0.98	0.98	0.88	0.90
7	A	0.92	0.94	0.96	0.97	0.99	0.99
	B	**0.90**	0.93	0.91	0.93	0.93	0.95
	C	0.99	0.99	0.99	0.99	0.93	0.94

Note. CC = Classification consistency; CA = Classification accuracy.

Table 5.8. Classification consistency and accuracy for reading grades 4 and 7.

Grade	Form	Level 3		Level 4		Level 5	
		CC	CA	CC	CA	CC	CA
4	A	0.99	0.99	0.99	0.99	0.99	0.99

	B	0.89	0.92	0.89	0.92	0.92	0.94
	C	0.95	0.96	0.91	0.94	0.82	0.88
7	A	0.94	0.96	0.97	0.98	0.99	0.99
	B	0.89	0.93	0.90	0.93	0.94	0.95
	C	0.97	0.98	0.93	0.95	0.84	0.88

Note. CC = Classification consistency; CA = Classification accuracy.

5.4. Dimensionality

The three multistage fixed adaptive EOG forms for the NCPAT system are designed based on the assumption that scores on the three forms represent an estimate of students' ability based on grade level content standards. It is therefore important that the NCDPI test design show relevant validity evidence to support the unidimensional use and interpretation of test scores on the multistage fixed adaptive forms. Empirical evidence of overall dimensionality for the multistage fixed adaptive EOG forms was explored using PCA and CFA. Between the two methods, PCA is an exploratory technique that seeks to summarize observed variables using a small number of new variables that are linear combinations of the original observed variables (the new variables are referred to as principal components). The primary hypothesis in PCA is to determine the fewest reasonable components that can explain most of the observed variance in the data. Two commonly used criteria to determine the number of meaningful dimensions for a set of observed variables are:

1. Retaining components whose eigenvalues are greater than the average of all the eigenvalues, which is 1 when the observed variables are standardized; and
2. Plotting eigenvalues against principal components (scree plot) and counting the number of components above where the bend occurs.

It is common to rely on both criteria simultaneously when evaluating the number of possible components for a given set of observed variables. Principal components are generally extracted from the tetrachoric correlation matrix for a test with dichotomously scored items to determine the number of meaningful principal components.

In contrast to PCA, CFA is a confirmatory technique that is used to verify the factor structure (i.e., construct of interest) of a set of observed variables. Several fit indices are typically used to evaluate the fit of a CFA model, such as the model chi-square, root mean square error of approximation (RMSEA), comparative fit index (CFI), and standardized root mean squared residual (SRMR).

Model chi-square is used to test the difference between the observed and model generated covariance matrices of the observed variables. A chi-square value close to zero indicates that the difference between the two covariance matrices is small. Unlike many other hypothesis tests, a non-significant result (i.e., p-value greater than the user-specified significance level) implies that the model fit the observed data well. The other fit indices do not have statistical tests, and therefore, threshold values proposed by Hu and Bentler (1999) are widely used to evaluate model fit. The RMSEA measures closeness of the model generated covariance matrix to the observed covariance matrix, with a value close to 0.06 indicating a good model fit. The CFI compares the user-specified model to a model that assumes independence among the observed variables, and a value close to 0.95 indicates good model fit. Finally, the SRMR is the average residual value between each element of the observed and model generated covariance matrices. Typically, a good model fit is indicated by an SRMR value close to 0.08.

5.4.1. Principal Component Analysis

A scree plot of eigenvalue against component number, which is often used to show the graphical result from PCA to determine the dimensionality of a test based on the aforementioned two criteria, is provided in Figures 5.1 and 5.2 for mathematics and reading, respectively. The left vertical axis shows the eigenvalues, and the right vertical axis displays the explained variance. The same information for the first three components is summarized in Tables 5.9 and 5.10 for mathematics and reading, respectively. Based on the PCA results, the eigenvalue of the first component is much larger than the eigenvalue of the second component. As the eigenvalue for each component is related to the total variance explained by the component, the first component explains a large amount of total variance compared to the other components. Furthermore, evaluation of the scree plots show that the bend of the plots occurs after the first component. Such results provide exploratory evidence in support of the assumption of unidimensionality—a single dominant component explains a significant amount of the total variance of the test scores on the multistage fixed adaptive forms.

Figure 5.1. PCA scree plot and explained variance for mathematics grades 4 and 7.

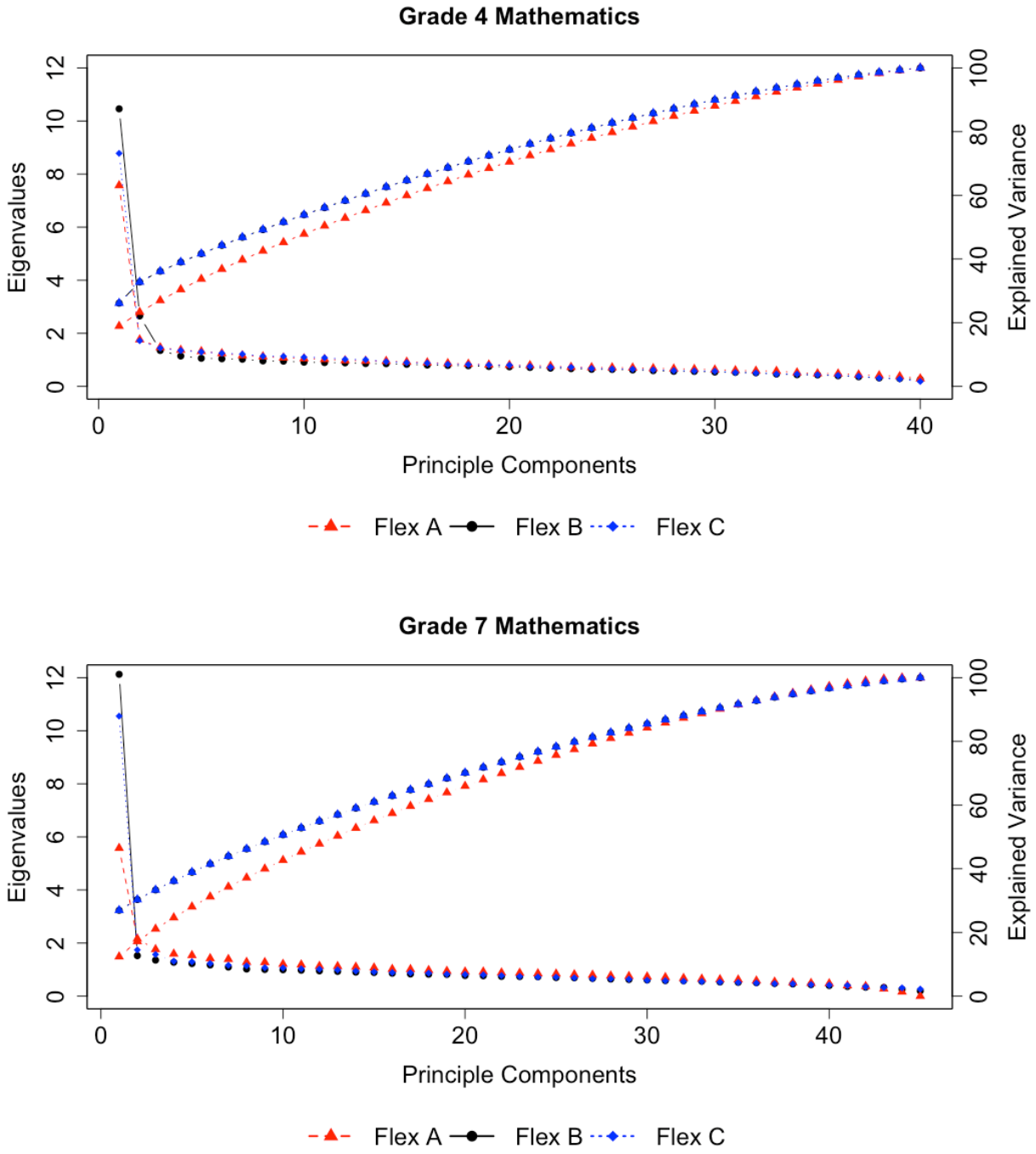


Figure 5.2. CA scree plot and explained variance for reading grades 4 and 7.

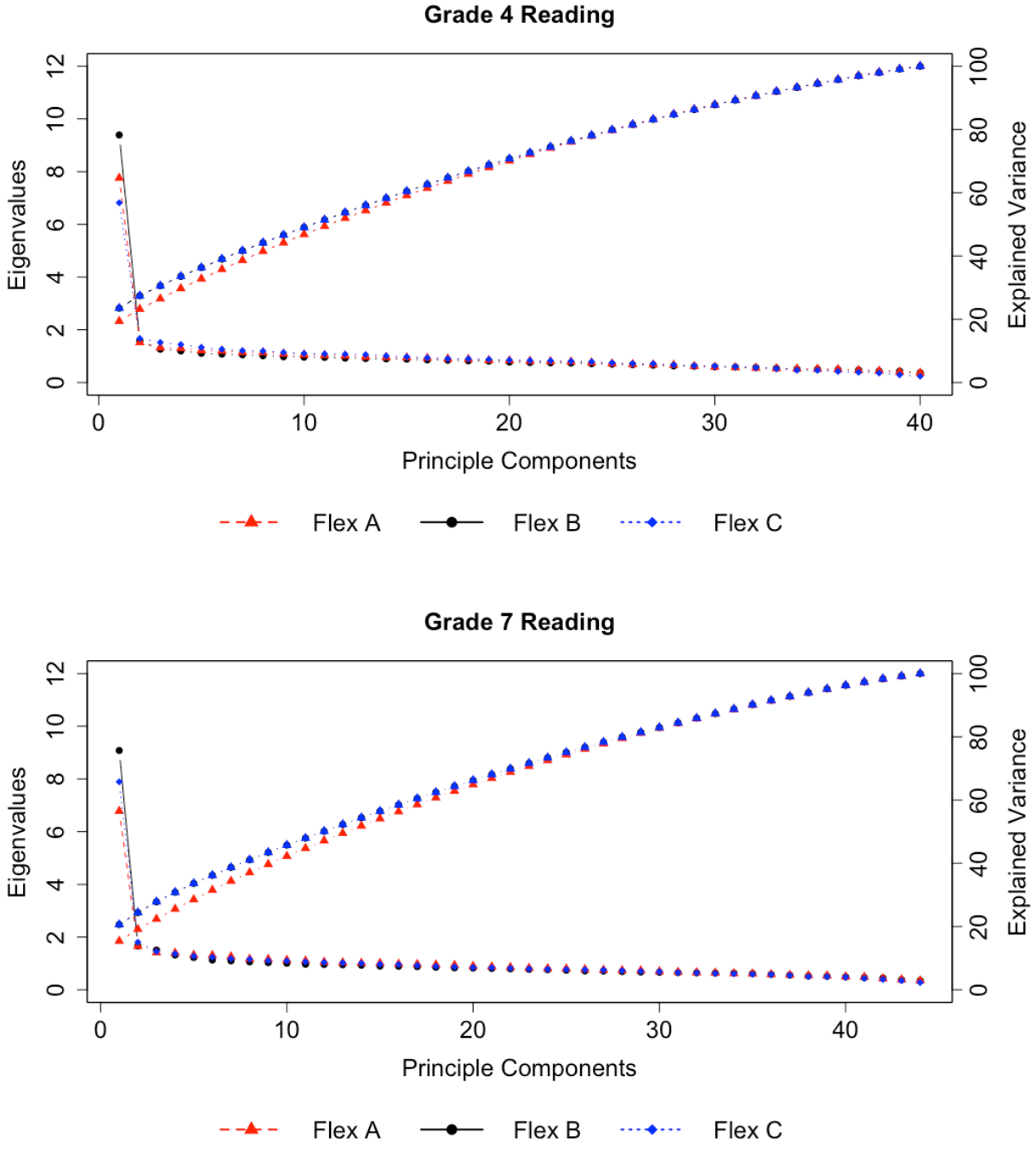


Table 5.9. Eigenvalues and explained variance for the first three components of mathematics grades 4 and 7.

Grade	Form	Principle Component	Eigenvalue	Explained Variance	Cumulative Variance
4	A	1	7.57	18.93%	18.93%
		2	1.76	4.40%	23.33%
		3	1.47	3.67%	27.00%
	B	1	10.46	26.15%	26.15%
		2	2.65	6.63%	32.78%
		3	1.35	3.38%	36.17%
	C	1	8.78	21.95%	21.95%
		2	1.72	4.30%	26.25%
		3	1.44	3.61%	29.87%
7	A	1	5.58	12.39%	12.39%
		2	2.16	4.80%	17.20%
		3	1.76	3.92%	21.12%
	B	1	12.12	26.95%	26.95%
		2	1.52	3.39%	30.34%
		3	1.36	3.02%	33.35%
	C	1	10.55	23.45%	23.45%
		2	1.74	3.87%	27.32%
		3	1.57	3.49%	30.81%

Table 5.10. Eigenvalues and explained variance for the first three components of reading grades 4 and 7.

Grade	Form	Principle Component	Eigenvalue	Explained Variance	Cumulative Variance
4	A	1	7.76	19.40%	19.40%
		2	1.52	3.80%	23.19%
		3	1.32	3.30%	26.50%
	B	1	9.39	23.48%	23.48%
		2	1.58	3.95%	27.43%
		3	1.26	3.15%	30.59%
	C	1	6.82	17.04%	17.04%
		2	1.67	4.18%	21.22%
		3	1.52	3.80%	25.02%
7	A	1	6.78	15.41%	15.41%
		2	1.66	3.76%	19.18%
		3	1.41	3.21%	22.39%
	B	1	9.08	20.64%	20.64%
		2	1.66	3.78%	24.42%
		3	1.51	3.43%	27.84%
	C	1	7.89	17.94%	17.94%
		2	1.81	4.10%	22.04%
		3	1.45	3.30%	25.34%

5.4.2. Confirmatory Factor Analysis

A single-factor CFA model was fit to the response data for each of the three multistage fixed adaptive EOG forms. The values of the model chi-square, RMSEA, CFI, and SRMR are provided in Tables 5.11 and 5.12 for mathematics and reading, respectively. For both subjects and grades, the p-values for the model chi-square were smaller than 0.05, indicating poor model fit. However, the fit of the single-factor CFA model cannot be determined solely based on the model chi-square because, similar to other statistical tests, model chi-square is sensitive to sample size. In contrast to the model chi-square, other fit indices indicate that the single-factor CFA model fits the data well. One exception was for mathematics grade 7 Form A, for which the value of CFI was smaller than the threshold 0.90; however, unlike CFI, RMSEA and SRMR indicated good model fit.

Based on the two evaluation criteria described above, eigenvalues and scree plots, a compelling argument can be made that the three multistage fixed adaptive forms for Grades 4 and 7

Mathematics and Reading are unidimensional. In other words, PCA results with one dominant component support interpreting scores for the three multistage fixed adaptive forms on Grades 4 and 7 Mathematics and Reading using a unidimensional scale.

Table 5.11. Values of several fit statistics for the CFA model of mathematics grades 4 and 7.

Grade	Form	Model Chi-Square	RMSEA	CFI	SRMR
4	A	0.000	0.020	0.947	0.045
	B	0.000	0.039	0.904	0.058
	C	0.000	0.020	0.947	0.053
7	A	0.000	0.017	0.832	0.061
	B	0.000	0.023	0.957	0.038
	C	0.000	0.022	0.952	0.047

Table 5.12. Values of several fit statistics for the CFA model of reading grades 4 and 7.

Grade	Form	Model Chi-Square	RMSEA	CFI	SRMR
4	A	0.000	0.015	0.975	0.038
	B	0.000	0.018	0.972	0.036
	C	0.000	0.015	0.938	0.051
7	A	0.000	0.016	0.955	0.041
	B	0.000	0.020	0.957	0.038
	C	0.000	0.014	0.960	0.045

Appendix A

The North Carolina EOG Test Specifications

https://www.dpi.nc.gov/districts-schools/testing-and-school-accountability/technical-information-state-tests#standard-setting-resources-and-reports

Appendix B

Mathematics Grade 4 Multistage Fixed Adaptive EOG Forms Summary

Table 1b. Content blueprint.

Domain	Blueprint By Domain		
	Min	Max	Items
Measurement and Data, Geometry	23.00%	27.00%	10
Number and Operations - Fractions	30.00%	34.00%	13
Number and Operations in Base Ten	25.00%	29.00%	11
Operations and Algebraic Thinking	14.00%	18.00%	6

Table 2b. DOK specification.

DOK Range		
DOK1	**DOK2**	**DOK3**
35%-45%	50%-60%	5%

Table 3b. Content blueprint for the three multistage fixed adaptive EOG forms.

Domains	Form						Blueprint	
	A		B		C			
	N	%	N	%	N	%	N	%
Measurement and Data, Geometry	10	25.00	10	25.00	9	22.50	29	24.17
Number and Operations - Fractions	13	32.50	13	32.50	13	32.50	39	32.50
Number and Operations in Base Ten	11	27.50	11	27.50	11	27.50	33	27.50
Operations and Algebraic Thinking	6	15.00	6	15.00	7	17.50	19	15.83
All	40	100.00	40	100.00	40	100.00	120	100.00

Table 4b. Test Format for the three multistage fixed adaptive EOG forms.

Form	Calculator Use						Total
	Active		Inactive				
	GR	MC	DD	GR	MC	TI	N
A	2	18	.	2	17	1	40
B	.	20	.	.	20	.	40
C	2	18	1	2	17	.	40

Table 5b. DOK range specification for the three multistage fixed adaptive EOG forms.

DOK	Form						Blueprint	
	A		B		C			
	N	%	N	%	N	%	N	%
Recall	19	47.50	16	40.00	14	35.00	49	40.83
Skill / Concept	21	52.50	22	55.00	24	60.00	67	55.83
Strategic Thinking	.	.	2	5.00	2	5.00	4	3.33
All	40	100.00	40	100.00	40	100.00	120	100.00

Table 6b. Summary of CTT and IRT item statistics for the three multistage fixed adaptive EOG forms.

Form	N	Mean			SD		
		p-value	IRT a	IRT b	p-value	IRT a	IRT b
A	40	0.66	1.84	-0.56	0.13	0.45	0.58
B	40	0.59	1.97	-0.23	0.13	0.42	0.65
C	40	0.50	1.89	0.22	0.15	0.48	0.74
All	120	0.59	1.90	-0.19	0.15	0.45	0.73

Table 7b. Number of common and unique items for the three multistage fixed adaptive EOG forms by content domain.

Domain	Type				Total Items
	Common	Form A	Form B	Form C	
Measurement and Data, Geometry	4	6	6	5	21
Number and Operations - Fractions	6	7	7	7	27
Number and Operations in Base Ten	6	5	5	5	21
Operations and Algebraic Thinking	3	3	3	4	13
All	19	21	21	21	82

Table 8b. Item Statistics of common and unique items for the three multistage fixed adaptive EOG forms.

Type	N	Mean			SD		
		p-value	IRT a	IRT b	p-value	IRT a	IRT b
Common	19	0.59	1.96	-0.24	0.13	0.47	0.63
Form A	21	0.72	1.72	-0.85	0.09	0.40	0.32
Form B	21	0.60	1.97	-0.22	0.13	0.38	0.68
Form C	21	0.43	1.82	0.64	0.13	0.48	0.56
All	82	0.58	1.87	-0.17	0.16	0.44	0.77

Table 9b. DIF summary for the three multistage fixed adaptive EOG forms.

Form	DIFF	
	AAAAA	DIF(B)
A	31	9
B	39	1
C	39	1
All	109	11

Table 10b. Item quality summary for the three multistage fixed adaptive EOG forms.

Form	Item Quality		
	Keep	Reserve	Weak
A	36	4	.
B	38	1	1
C	35	5	.
All	109	10	1

Figure 1b: Percent correct distribution for the three multistage fixed adaptive EOG forms.

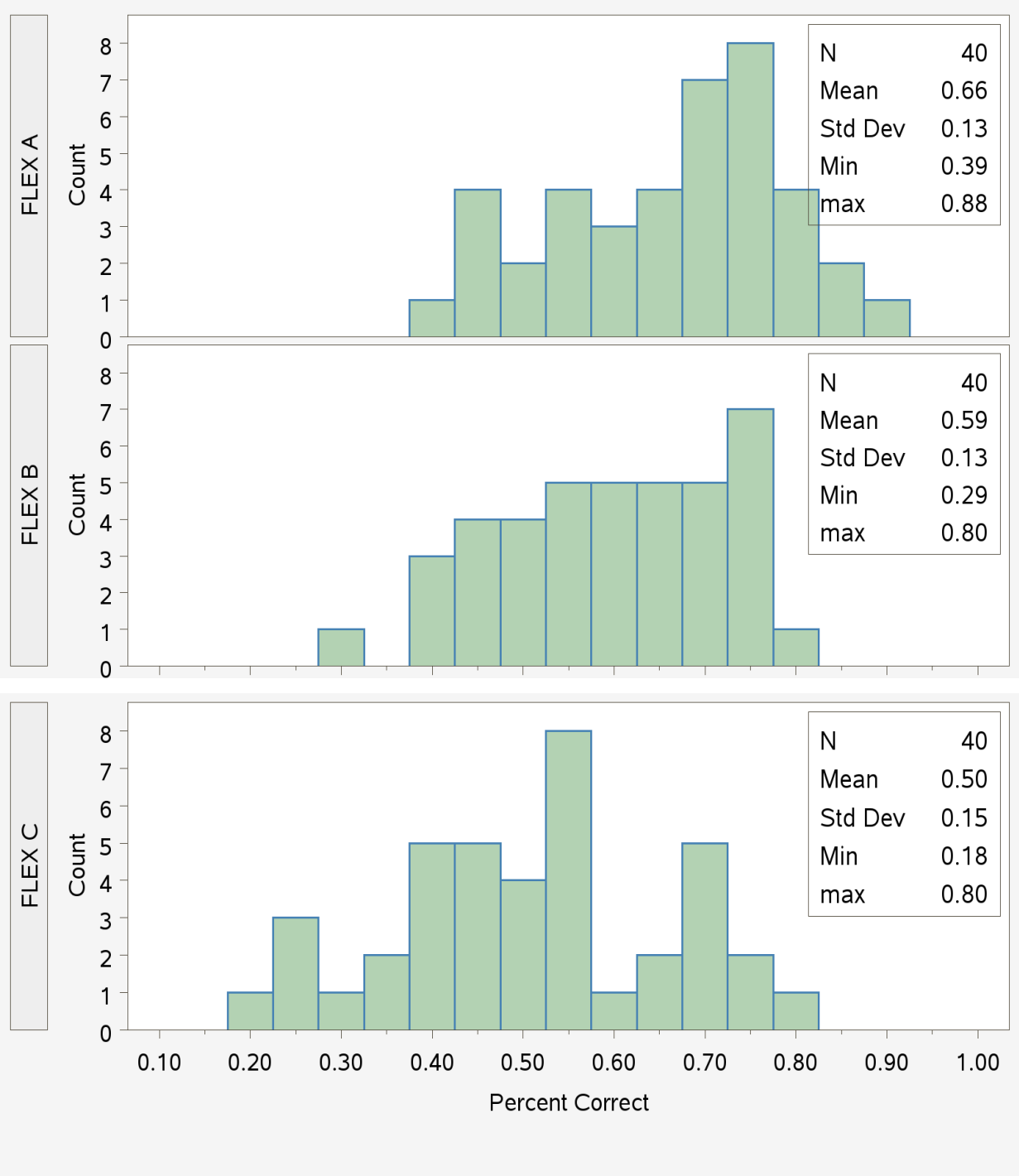


Figure 2b. Test characteristic curves for the three multistage fixed adaptive EOG forms.

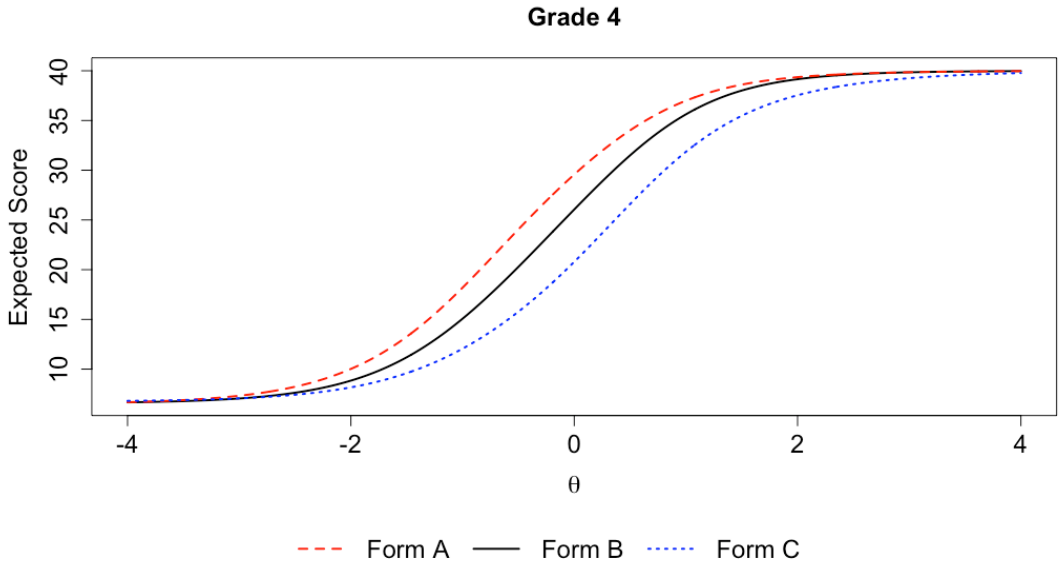
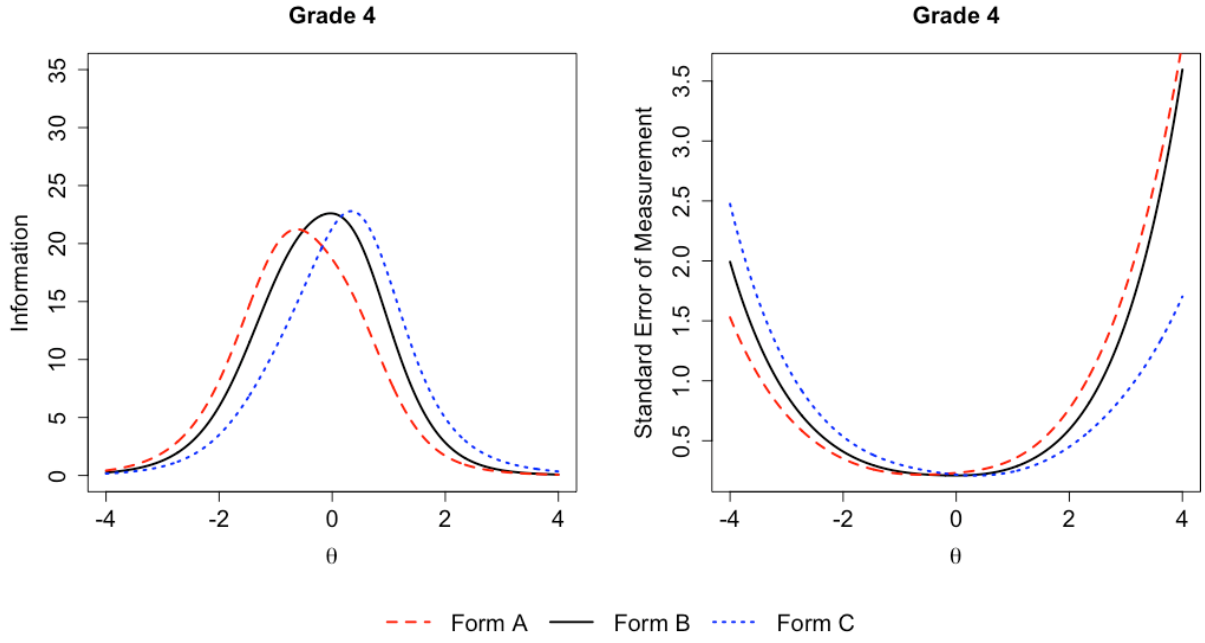


Figure 3b: Test information functions for the three multistage fixed adaptive EOG forms.



Appendix C

Mathematics 7 Multistage Fixed Adaptive Forms Summary

Table 1c. Content blueprint.

Domain	Blueprint By Domain		
	Min	Max	Items
Expressions and Equations	20.00%	24.00%	10
Geometry	16.00%	20.00%	7
Ratio and Proportional Relationships	24.00%	28.00%	12
Statistics and Probability	22.00%	26.00%	11
The Number System	8.00%	12.00%	5

Table 2c. DOK specification.

DOK Range		
DOK1	**DOK2**	**DOK3**
25-35%	50-60%	8-15%

Table 3c. Content blueprint for the three multistage fixed adaptive EOG forms.

Domains	Form						Blueprint	
	A		B		C			
	N	%	N	%	N	%	N	%
Expressions and Equations	11	24.44	10	22.22	10	22.22	31	22.96
Geometry	7	15.56	7	15.56	7	15.56	21	15.56
Ratio and Proportional Relationships	13	28.89	12	26.67	12	26.67	37	27.41
Statistics and Probability	11	24.44	11	24.44	11	24.44	33	24.44
The Number System	3	6.67	5	11.11	5	11.11	13	9.63
All	45	100.00	45	100.00	45	100.00	135	100.00

Table 4c. Test Format for the three multistage fixed adaptive EOG forms.

Form	Calculator Use						Total
	C-Active			C-Inactive			
	DD	GR	MC	DD	GR	MC	DD
A	1	5	23	1	5	23	1
B	.	5	25	.	5	25	.
C	2	4	24	2	4	24	2

Table 5c. DOK range specification for the three multistage fixed adaptive EOG forms.

DOK	Form					
	A		B		C	
	N	%	N	%	N	%
Recall	13	28.89	14	31.11	11	24.44
Skill / Concept	31	68.89	27	60.00	29	64.44
Strategic Thinking	1	2.22	4	8.89	5	11.11
All	45	100	45	100	45	100

Table 6c. Summary of CTT and IRT item statistics for the three multistage fixed adaptive EOG forms.

Form	N	Mean			SD		
		p-value	IRT a	IRT b	p-value	IRT a	IRT b
A	45	0.54	2.07	45	0.54	2.07	45
B	45	0.49	2.12	45	0.49	2.12	45
C	45	0.43	2.18	45	0.43	2.18	45
All	135	0.48	2.13	135	0.48	2.13	135

Table 7c. Number of common and unique items for the three multistage fixed adaptive EOG forms by content domain.

Domain	Type				
	Common	Form A	Form B	Form C	Total Items
Expressions and Equations	4	7	6	6	23
Geometry	3	4	4	4	15
Ratio and Proportional Relationships	6	7	6	6	25
Statistics and Probability	4	7	7	7	25
The Number System	1	2	4	4	11
All	18	27	27	27	99

Table 8c. Item Statistics of common and unique items for the three multistage fixed adaptive EOG forms.

Type	N	Mean			SD		
		p-value	IRT a	p-value	IRT a	p-value	IRT a
Common	18	0.49	2.01	0.08	0.12	0.40	0.44
Form A	27	0.57	2.12	-0.37	0.13	0.59	0.48
Form B	27	0.48	2.20	0.13	0.16	0.56	0.58
Form C	27	0.38	2.30	0.64	0.11	0.64	0.50
All	99	0.48	2.17	0.12	0.15	0.57	0.63

Table 9c. DIF summary for the three multistage fixed adaptive EOG forms.

Form	DIFF	
	AAAAA	DIF(B)
A	36	9
B	42	3
C	39	6
All	117	18

Table 10c. Item quality summary for the three multistage fixed adaptive EOG forms.

Form	Item Quality		
	Keep	Reserve	Weak
A	41	4	.
B	42	3	.
C	43	2	.
All	126	9	.

Figure 2c. Test characteristic curves for the three multistage fixed adaptive EOG forms.

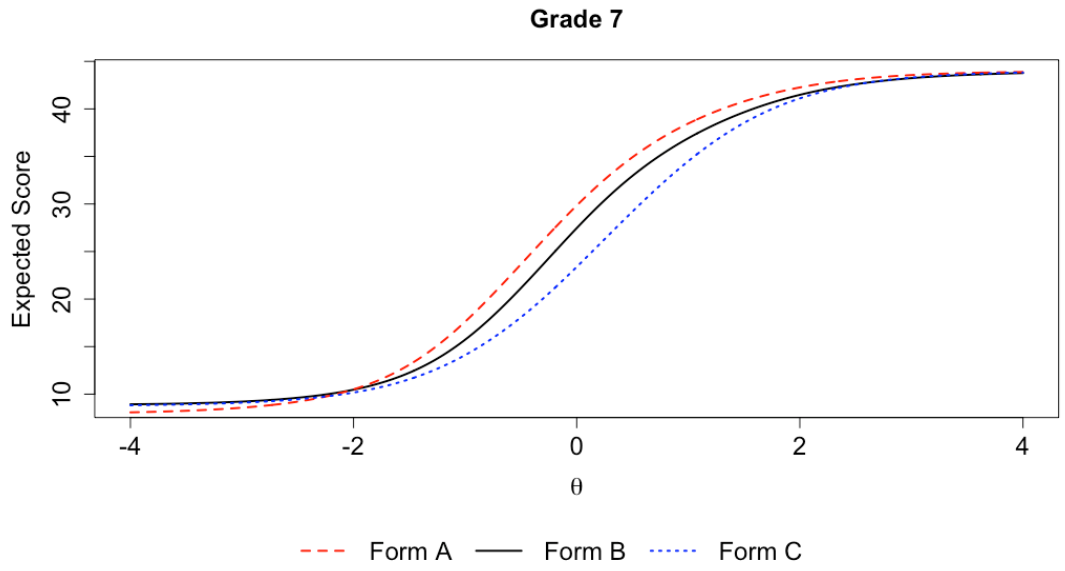
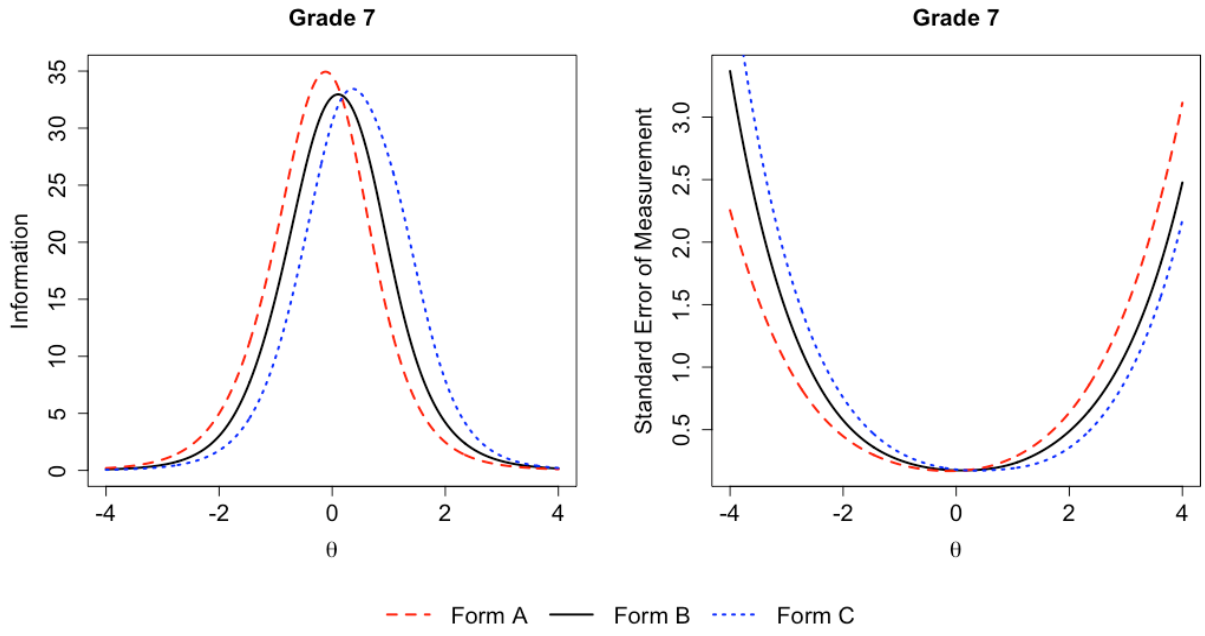


Figure 3c. Test information functions for the three multistage fixed adaptive EOG forms.



Appendix D

Reading Grade 4 Multistage Fixed Adaptive EOG Forms Summary

Table 1d. Content blueprint.

Domain	Weight	Blueprint By Domain	
		%	Items
Language	13% - 15%	14	6
Reading for Informational Text	46% - 50%	46	18
Reading for Literature	38% - 42%	40	16

Table 2d. DOK specification.

DOK Range		
DOK1	**DOK2**	**DOK3**
12%-25%	50%-75%	5%-10%

Table 3d. Content blueprint for the three multistage fixed adaptive EOG forms.

Domains	Form						Blueprint	
	A		B		C			
	N	%	N	%	N	%	N	%
Language	5	12.50	5	12.50	5	12.50	15	12.50
Reading for Informational Text	19	47.50	20	50.00	19	47.50	58	48.33
Reading for Literature	16	40.00	15	37.50	16	40.00	47	39.17
All	40	100.00	40	100.00	40	100.00	120	100.00

Table 4d. Test format selection type for the three multistage fixed adaptive EOG forms.

Form	Informational	Literature	Total
A	24	16	40
B	24	16	40
C	24	16	40

Table 5d. Test format item type for the three multistage fixed adaptive EOG forms.

Form	N	Item Type			
		DD	MC	SR	TD
A	40	2	35	2	1
B	40	1	34	2	3
C	40	2	35	1	2
All	120	5	104	5	6

Table 6d. DOK range specification for the three multistage fixed adaptive EOG forms.

DOK	Form					
	A		B		C	
	N	%	N	%	N	%
Recall	9	22.50	6	15.00	6	15.00
Skill / Concept	28	70.00	30	75.00	27	67.50
Strategic Thinking	3	7.50	4	10.00	7	17.50
All	40	100	40	100	40	100

Table 7d. Summary of CTT and IRT item statistics for the three multistage fixed adaptive EOG forms.

Form	N	Mean			SD		
		p-value	IRT a	IRT b	p-value	IRT a	IRT b
A	40	0.65	1.93	-0.39	0.14	0.61	0.72
B	40	0.60	1.91	-0.14	0.15	0.63	0.76
C	40	0.53	1.91	0.21	0.14	0.58	0.73
All	120	0.60	1.92	-0.11	0.15	0.60	0.77

Table 8d. DIF summary for the three multistage fixed adaptive EOG forms.

Form	DIFF	
	AAAAA	DIF(B)
A	35	5
B	37	3
C	34	6
All	106	14

Table 9d. Item quality summary for the three multistage fixed adaptive EOG forms.

Form	Item Quality		
	Keep	Reserve	Weak
A	35	5	.
B	35	5	.
C	33	7	.
All	103	17	.

Figure 1d. Percent correct distribution for the three multistage fixed adaptive EOG forms.

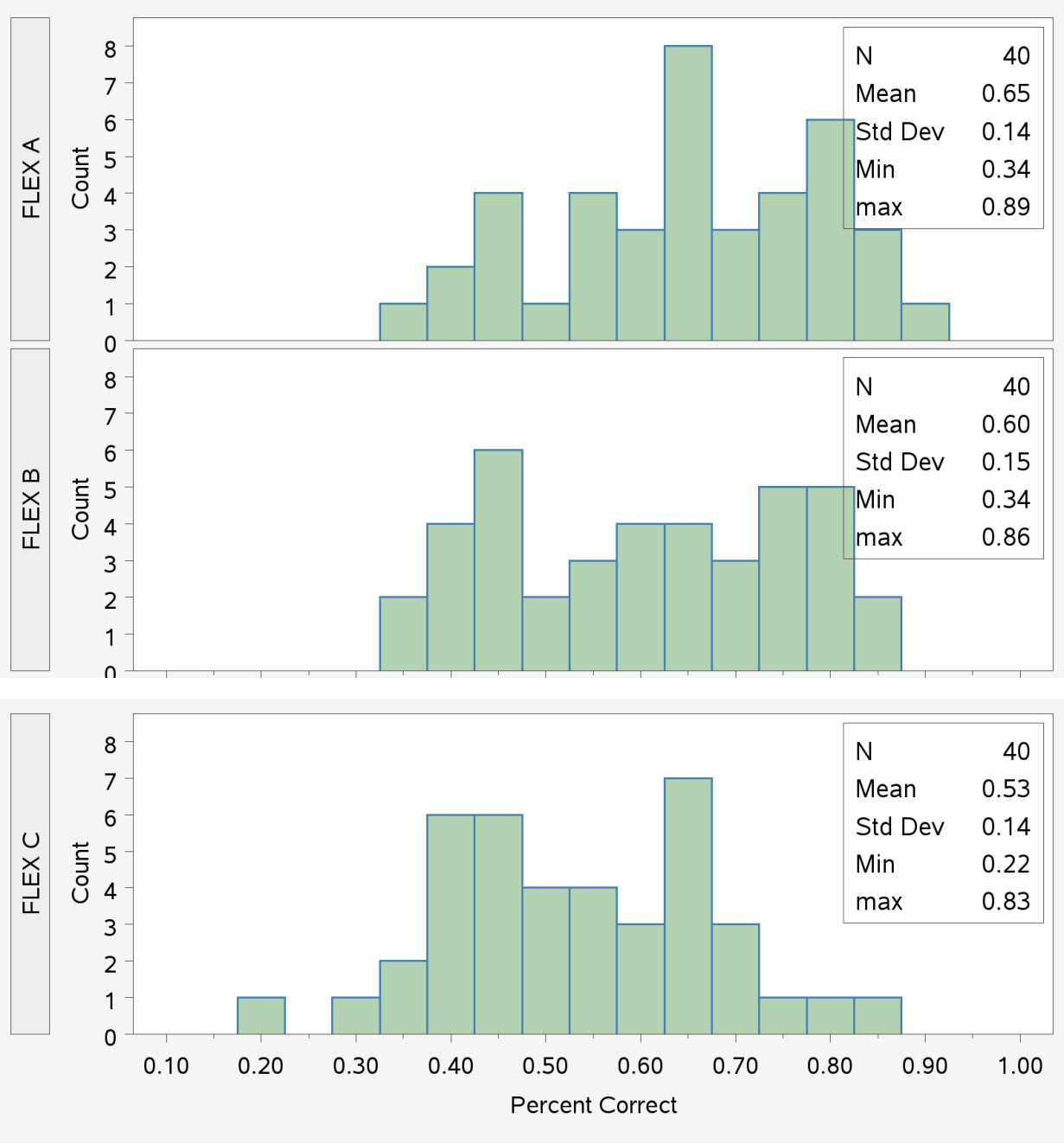


Figure 2d. Test characteristic curves for the three multistage fixed adaptive EOG forms.

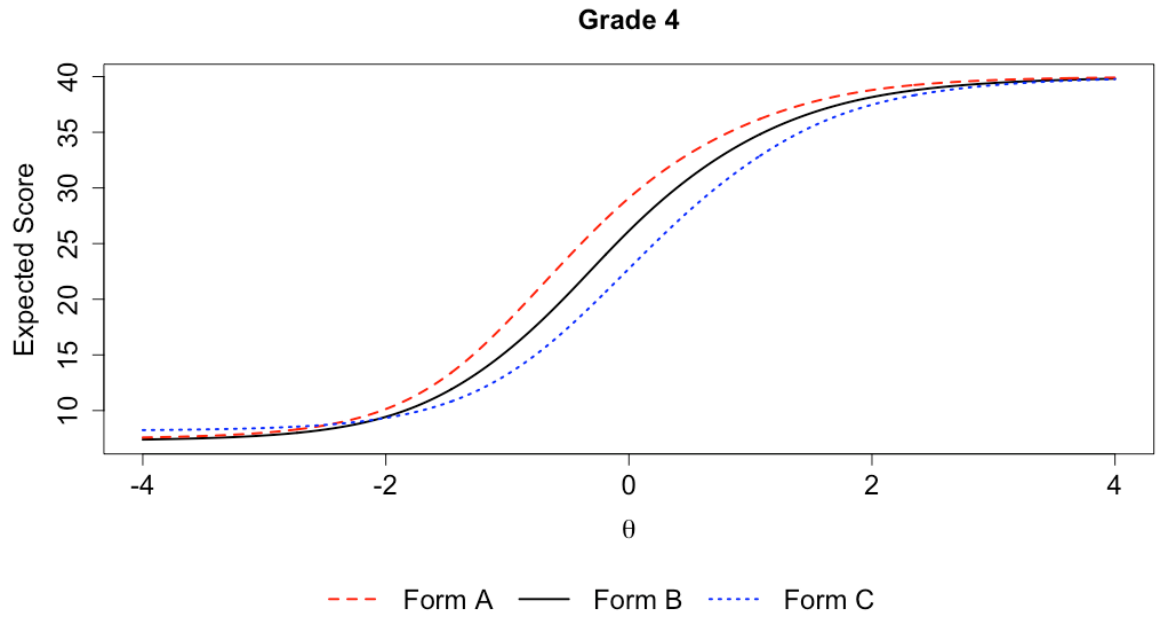
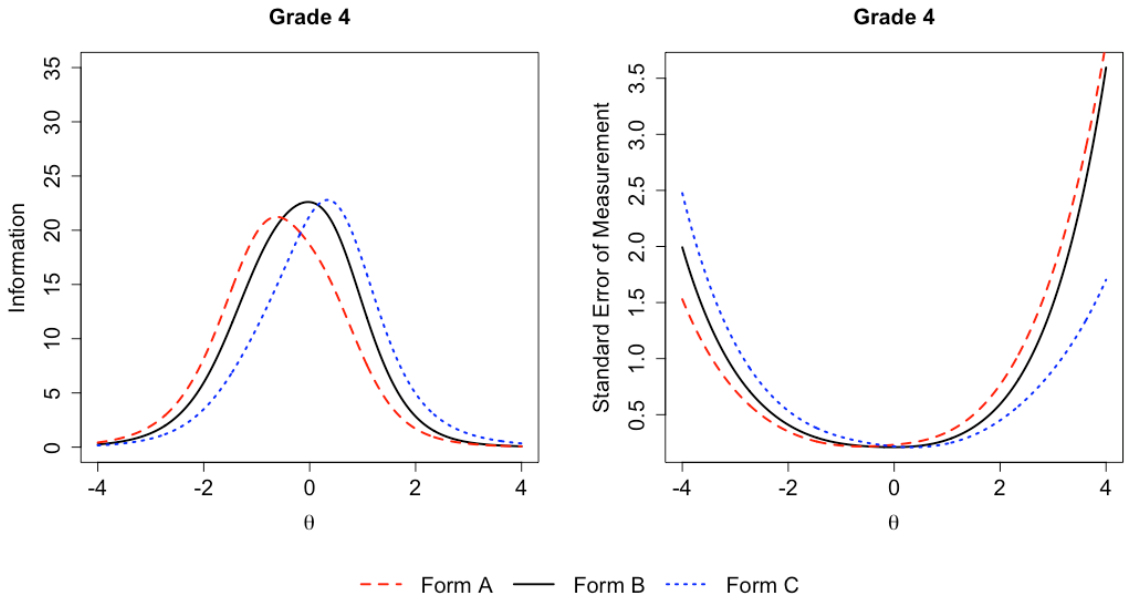


Figure 3d. Test information functions for the three multistage fixed adaptive EOG forms.



Appendix E

Reading Grade 7 Multistage Fixed Adaptive EOG Forms Summary

Table 1e. Content blueprint.

Domain	Weight	Blueprint By Domain	
		%	Items
Language	11% - 16%	15	7
Reading for Informational Text	43% - 47%	46	20
Reading for Literature	36% - 41%	38	17

Table 2e. DOK specification.

DOK Range		
DOK1	**DOK2**	**DOK3**
	60%-82%	18%-40%

Table 3e. Content blueprint for the three multistage fixed adaptive EOG forms.

Domains	Form						Blueprint	
	A		B		C			
	N	%	N	%	N	%	N	%
Language	7	15.91	6	13.64	7	15.91	20	15.15
Reading for Informational Text	19	43.18	20	45.45	21	47.73	60	45.45
Reading for Literature	18	40.91	18	40.91	16	36.36	52	39.39
All	44	100.00	44	100.00	44	100.00	132	100.00

Table 4e. Test format selection type for the three multistage fixed adaptive EOG forms.

Form	Informational	Literature	Total
A	23	16	5
B	23	16	5
C	23	16	5

Table 5e. Test format item type for the three multistage fixed adaptive EOG forms.

Form	N	Item Type				
		DD	MC	SR	TD	TI
A	44	1	38	2	1	2
B	44	3	39	.	1	1
C	44	1	40	1	1	1
All	132	5	117	3	3	4

Table 6e. DOK range specification for the three multistage fixed adaptive EOG forms.

DOK	Form					
	A		B		C	
	N	%	N	%	N	%
Skill / Concept	34	77.27	30	68.18	31	70.45
Strategic Thinking	10	22.73	14	31.82	13	29.55
Total	44	100	44	100	44	100

Table 7e. Summary of CTT and IRT item statistics for the three multistage fixed adaptive EOG forms.

Form	N	Mean			SD		
		p-value	IRT a	IRT b	p-value	IRT a	IRT b
A	44	0.61	1.83	-0.24	0.14	0.54	0.68
B	44	0.58	1.89	0.02	0.14	0.61	0.72
C	44	0.52	1.80	0.27	0.13	0.52	0.67
All	132	0.57	1.84	0.02	0.14	0.55	0.71

Table 8e. DIF summary for the three multistage fixed adaptive EOG forms.

Form	DIFF	
	AAAAA	DIF(B)
A	37	7
B	43	1
C	43	1
All	123	9

Table 9e. Item quality summary for the three multistage fixed adaptive EOG forms.

Form	Item Quality		
	Keep	Reserve	Weak
A	37	7	.
B	38	6	.
C	37	7	.
All	112	20	.

Figure 1e. Percent correct distribution for the three multistage fixed adaptive EOG forms.

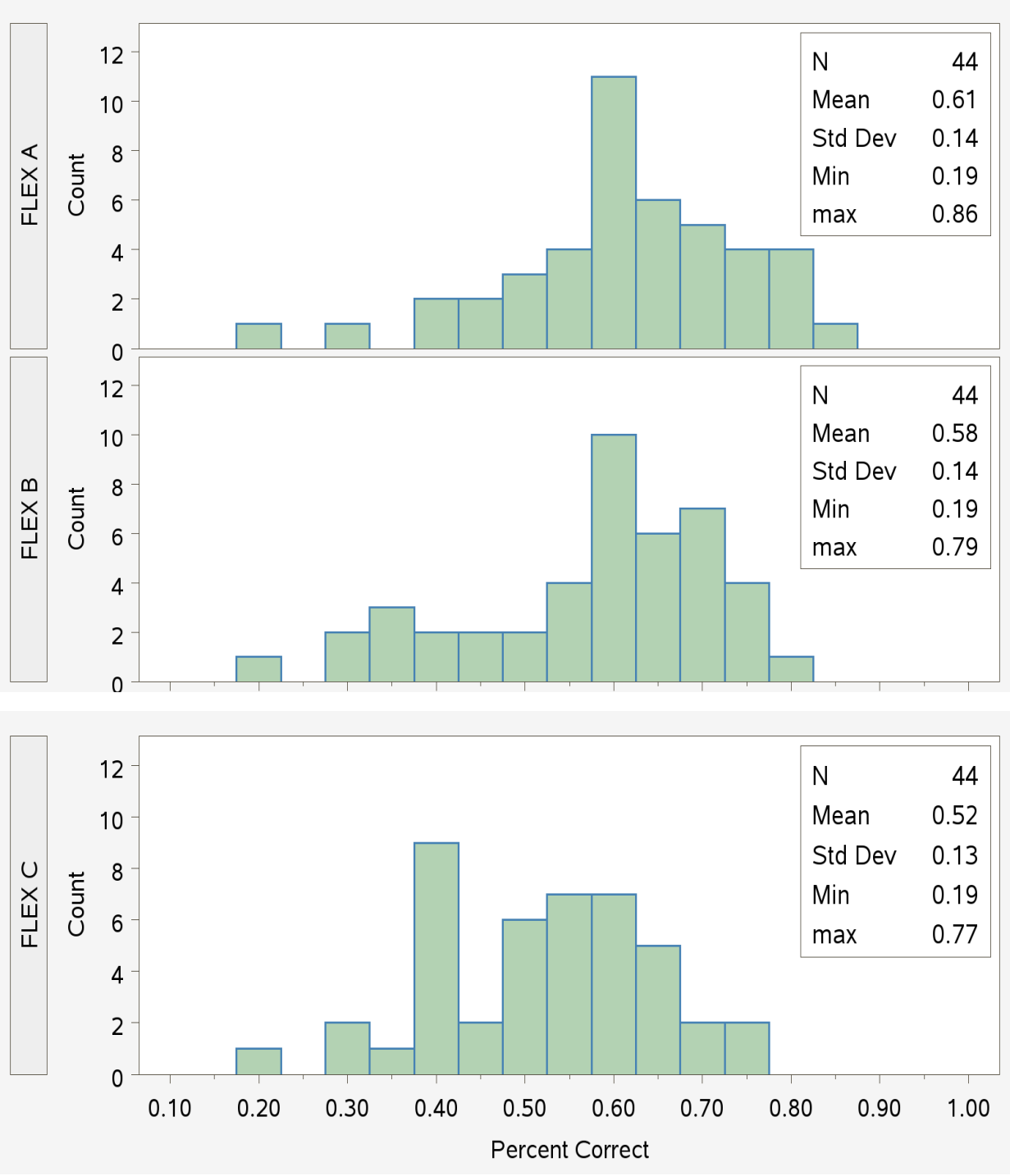


Figure 2e. Test characteristic curves for the three multistage fixed adaptive EOG forms.

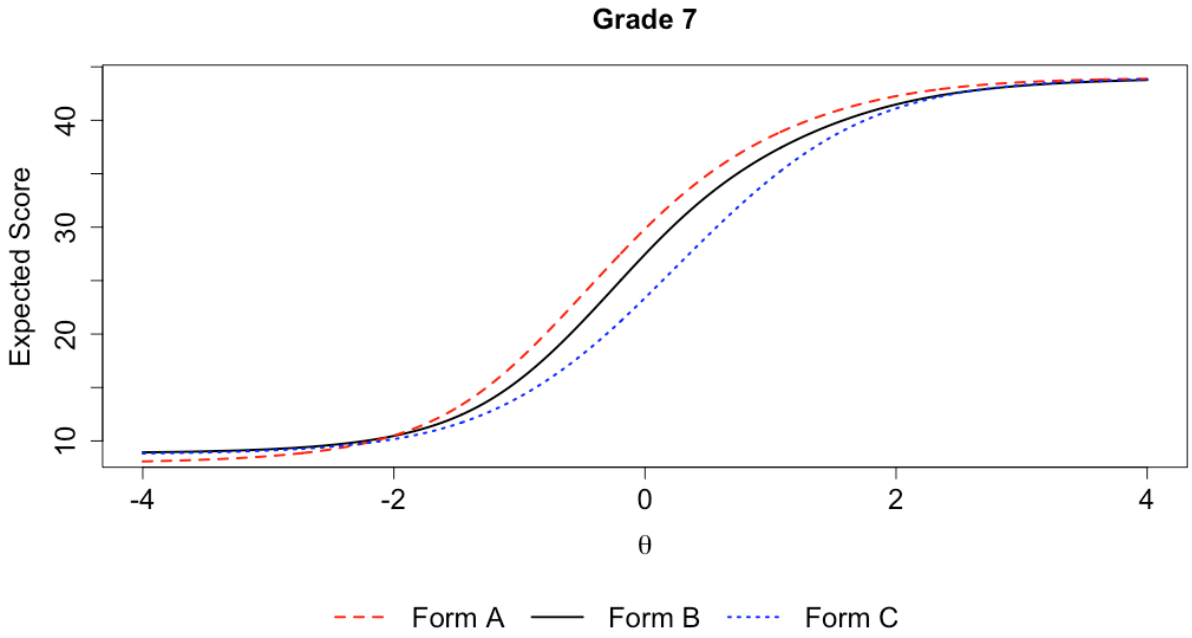
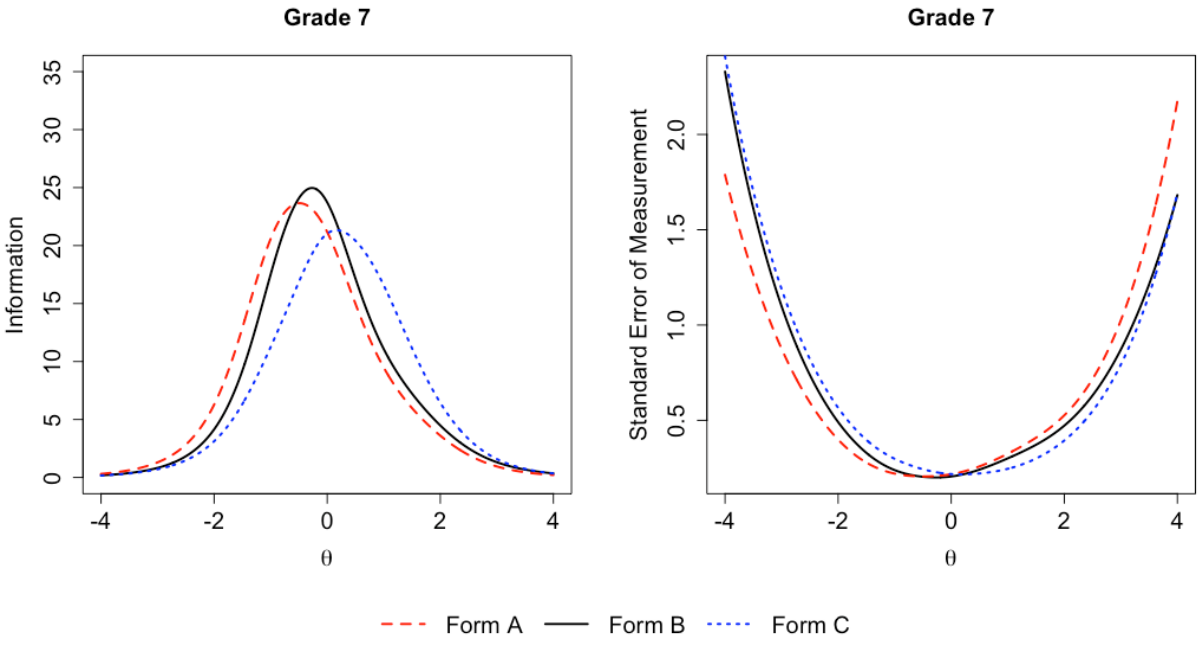


Figure 3e. Test information functions for the three multistage fixed adaptive EOG forms.



Appendix F
Comparability Study

The results of scale comparability analysis between the multistage fixed adaptive EOG administered as a pilot under the NCPAT system and the general EOG administered to students across the state are presented here. The multistage fixed adaptive EOG forms for each grade were designed with three different levels, and students were assigned to a specific level based on a routing methodology that used performance on the NC Check-Ins 2.0. In 2022-23 school year, both the NC Check-Ins 2.0 and multistage fixed adaptive EOG components of the NCPAT were administered to students in mathematics and reading grades 4 and 7 as a pilot program in selected volunteer schools across the state. Specifically, students enrolled in schools that participated in the pilot study took the multistage fixed adaptive EOG, whereas students enrolled in non-pilot schools took the traditional EOG assessment. To examine whether students' performance on the multistage fixed adaptive EOG is comparable to performance of students who took the traditional EOG, EOG scores between pilot and non-pilot schools were compared in terms of their mean scale scores and distribution of achievement levels.

As pilot schools were selected voluntarily and not randomly, comparison of the pilot and non-pilot schools may provide biased results. To mimic random assignment and reduce bias in the treatment effect (pilot vs. non-pilot schools) estimates, students in the pilot and non-pilot schools were matched on several covariates at the student and school levels. Previous year scale score was also included in the propensity score model as an outcome predictor because it was strongly correlated with current year scale score, one of the two outcome measures included in the current study (correlation between previous and current year scale scores was larger than 0.82 for both subjects and grades). A statistical approach for matching is to match students on *propensity scores* (Rosenbaum & Rubin, 1983), which are the predicted probabilities of assignment to the pilot school given a set of observed covariates. Propensity score matching facilitates the matching process by reducing several covariates down to a single number. Detailed description of the covariates used for propensity score matching is provided in Table 1f.

Table 1f. Descriptions of the covariates used for propensity score matching.

	Covariates	Type of Data	# of Categories	Description of Categories
Students	Sex	Categorical	2	Male, Female
	Ethnicity	Categorical	4	White, Black, Hispanic, Others
	EDS	Categorical	2	No, Yes
	SWD	Categorical	2	No, Yes
	ELS	Categorical	2	No, Yes
	Previous Year Scale Score	Continuous		
School	Region	Categorical	8	North Central, Northeast, Northwest, Piedmont-Triad, Sandhills, Southeast, Southwest, Western
	School Type	Categorical	2	Public, Charter

F.1. Description of Data

The data used for the comparability study were collected during the 2022-23 school year from North Carolina students enrolled in public and charter schools in mathematics and reading grades 4 and 7. Sample sizes of the total student population are summarized in Table 2f. For both subjects and grades, roughly 8% of students were from pilot schools and 92% were from non-pilot schools.

F.1.1. Pre-Matched Data

In the current study, the number of students with missing data on at least one of the covariates was small. Therefore, they were removed from the analysis instead of imputing missing data (the resulting data set will hereinafter be referred to as "pre-matched" data). As shown in Table 2f, the pre-matched data consisted of approximately 93% of the original data.

Table 2f. Sample sizes for the original, pre-matched, and matched data sets.

Grade	Dataset	Mathematics			Reading		
		All	Pilot	Non-Pilot	All	Pilot	Non-Pilot
4	Original	110,561	8,775	101,786	111,238	8,783	102,455
	Pre-Matched	102,577	8,148	94,429	103,299	8,161	95,138
	Matched	16,296	8,148	8,148	16,322	8,161	8,161
7	Original	113,693	8,840	104,853	113,759	8,880	104,879
	Pre-Matched	105,540	8,124	97,416	105,655	8,143	97,512

Matched	16,248	8,124	8,124	16,286	8,143	8,143

Prior to conducting propensity score matching to match each student in the pilot schools with a student from a non-pilot school with respect to the covariates, evaluation of covariate balance was performed for the pre-matched data using graphical, descriptive, and inferential measures. Inferential measures used in the current study were the two-proportion Z-test for categorical variables with two categories, a chi-square test for categorical variables with more than two categories, and a two-sample t-test for continuous variables. Because statistical tests tend to be highly sensitive with larger sample sizes, in addition to the inferential measures, effect sizes were computed using Cohen's *h* (Cohen, 1988) for categorical variables with two categories, Cramér's *V* (Cramér, 1946) for categorical variables with more than two categories, and Cohen's *d* (Cohen, 1988) for continuous variables. Note that the general rule of thumb is to interpret effect sizes that are smaller than 0.2 as negligible (Cohen, 1988; Cramér, 1946); in other words, the observed differences are not practically significant.

Tables 3f and 4f provide the test results and effect sizes of the pre-matched data for EOG mathematics and reading, respectively. For most of the covariates, differences between pilot and non-pilot schools were statistically significant at the 0.05 level, with the exception of the sex for the two subjects and grades. Despite the significant results, effect sizes for most covariates were less than 0.2, implying that the differences were not practically significant. One exception was the covariate region, for which the effect sizes were larger than 0.2 for both subjects and grades. This is probably because schools' participation in the pilot study was voluntary and not proportional to the number of schools in each region. Covariate balance evaluation based on graphical measures, shown in Figures 5f and 6f, provides consistent results with the covariate balance evaluation based on effect size measures.

Table 3f. Pre-matched data covariate balance evaluation for mathematics grades 4 and 7.

Covariates	Test	Grade 4			Grade 7		
		Statistic	p-value	Effect Size	Statistic	p-value	Effect Size
Sex	2propZ	0.82	0.4146	0.0094	-0.57	0.5717	0.0065
Ethnicity	ChiSq	126.78	0.0000*	0.0352	86.13	0.0000*	0.0286
EDS	2propZ	15.97	0.0000*	0.1864	14.53	0.0000*	0.1690
SWD	2propZ	6.42	0.0000*	0.0724	4.42	0.0000*	0.0501
ELS	2propZ	-5.58	0.0000*	0.0667	-4.77	0.0000*	0.0571
Previous Year Scale Score	WlcT	5.50	0.0000*	0.0609	6.35	0.0000*	0.0683
Region	ChiSq	6332.26	0.0000*	0.2485	6675.09	0.0000*	0.2511
School Type	2propZ	-2.17	0.0302*	0.0254	-2.59	0.0096*	0.0304

Note. WlcT = Welch's t-test; * symbol indicates that the test is significant at a significance level of 0.05.

Table 4f. Pre-matched data covariate balance evaluation for reading grades 4 and 7.

Covariates	Test	Grade 4			Grade 7		
		Statistic	p-value	Effect Size	Statistic	p-value	Effect Size
Sex	2propZ	0.77	0.4441	0.0088	-0.4616	0.6444	0.0053
Ethnicity	ChiSq	126.43	0.0000*	0.0350	87.24	0.0000*	0.0287
EDS	2propZ	16.01	0.0000*	0.1868	14.67	0.0000*	0.1705
SWD	2propZ	6.53	0.0000*	0.0735	4.45	0.0000*	0.0503
ELS	2propZ	-5.61	0.0000*	0.0670	-4.83	0.0000*	0.0578
Previous Year Scale Score	WlcT	8.05	0.0000*	0.0898	8.00	0.0000*	0.0891
Region	ChiSq	6382.13	0.0000*	0.2486	6676.14	0.0000*	0.2514
School Type	2propZ	-1.98	0.0472*	0.0232	-2.64	0.0084*	0.0309

Note. WlcT = Welch's t-test; * symbol indicates that the test is significant at significance level of 0.05.

F.1.2. Matching procedure

Propensity score matching was performed by regressing the binary variable, pilot school versus non-pilot school (pilot school = 1; non-pilot school = 0), on the covariates listed in Table 1f. The propensity scores were calculated using logistic regression with the categorical variables dummy coded. Matching on the propensity scores was performed using the R package "MatchIt" (Ho et al., 2018) with the nearest neighbor (or greedy) matching method. For each student in the pilot

schools, greedy matching searches for the "best" available student in a non-pilot school without considering the quality of the match of the entire sample. The matching result data set will hereinafter be referred to as "matched" data.

F.1.3. Matched Data

Sample sizes for the matched data sets are provided in Table 2f. Sample sizes for the pilot and non-pilot schools are equal because greedy matching was used for propensity score matching. In the matched data, differences observed between the pilot and non-pilot schools in terms of the covariates included in the propensity score model were not statistically significant for both subjects and grades (see Tables 5f and 6f). Consistent with the statistical test results, the covariates were well balanced in both groups after matching as shown in Figures 7f and 8f.

Table 5f. Matched data covariate balance tests for mathematics grades 4 and 7.

Covariates	Test	Grade 4			Grade 7		
		Statistic	p-value	Effect Size	Statistic	p-value	Effect Size
Sex	2propZ	-0.14	0.8879	0.0022	-0.06	0.9499	0.0010
Ethnicity	ChiSq	0.31	0.9573	0.0044	0.73	0.8856	0.0067
EDS	2propZ	-0.02	0.9871	0.0003	-0.58	0.5633	0.0091
SWD	2propZ	-0.06	0.9835	0.0009	-0.02	0.9833	0.0003
ELS	2propZ	0.24	0.8104	0.0038	1.12	0.2624	0.0176
Previous Year Scale Score	WlcT	-0.76	0.4458	0.0119	-0.51	0.6118	0.0080
Region	ChiSq	0.87	0.9967	0.0073	0.27	0.9999	0.0041
School Type	2propZ	0.29	0.7707	0.0046	0.64	0.5191	0.0101

Note. WlcT = Welch's t-test; * symbol indicates that the test is significant at significance level of 0.05.

Table 6f. Matched data covariate balance tests for reading grades 4 and 7.

Covariates	Test	Grade 4	Grade 7

		Statistic	p-value	Effect Size	Statistic	p-value	Effect Size
Sex	2propZ	-0.17	0.8633	0.0027	-0.50	0.6159	0.0079
Ethnicity	ChiSq	2.94	0.4004	0.0134	6.55	0.0877	0.0200
EDS	2propZ	0.26	0.7960	0.0040	-0.14	0.8853	0.0023
SWD	2propZ	0.58	0.5613	0.0091	0.82	0.4128	0.0128
ELS	2propZ	1.46	0.1446	0.0228	1.87	0.0622	0.0292
Previous Year Scale Score	WlcT	-0.49	0.6236	0.0077	0.37	0.7119	0.0058
Region	ChiSq	4.52	0.7186	0.0166	0.44	0.9996	0.0052
School Type	2propZ	0.11	0.9156	0.0017	-0.19	0.8523	0.0029

Note. WlcT = Welch's t-test; * symbol indicates that the test is significant at significance level of 0.05.

F.2. Results

The mean scale scores and distribution of achievement levels between the pilot and non-pilot schools were compared using the matched data. Scale scores were compared using a two-sample t-test, and Cohen's *d* was computed as a measure of effect size. Instead of using the original four levels, achievement levels were dichotomized as "not proficient" and "proficient (level 3 or above)" for statistical testing. Therefore, achievement level differences were tested using a two-proportion Z-test and practical significance was examined using Cohen's *h*.

F.2.1. Mathematics

The test results and effect sizes for both grade levels in mathematics are provided in Table 7f. The mean scale scores between the pilot and non-pilot schools for grade 4 were not statistically different for the matched sample. As can be seen from Table 8f, the mean scale scores for the pilot and non-pilot schools were identical as 546.3. Consistent with the overall results, the distribution of scale scores for the pilot and non-pilot schools were comparable across the entire score scale (see Figure 1f). For achievement levels, the frequency distributions were very similar between the pilot and non-pilot schools, and results from the two-proportion Z test was not statistically significant (see Figure 2f).

Grade 7, on the other hand, showed statistically significant difference in the mean scale scores. However, the mean scale scores for the pilot and non-pilot schools were 545.1 and 545.5, respectively, showing a difference of 0.4 in the scale score metric. As shown in Table 7f, a 0.4

difference in scale score translates to a 0.04 difference in standard deviation (Cohen's $d = 0.0450$), which is a negligible difference. Unlike scale scores, no significant difference was observed in achievement levels.

Table 7f. Comparison of scale scores and achievement levels for mathematics grades 4 and 7.

Outcome variable	Test	Grade 4			Grade 7		
		Statistic	p-value	Effect Size	Statistic	p-value	Effect Size
Scale score	WlcT	-0.54	0.5926	0.0084	2.87	0.0042*	0.0450
AL	2propZ	-1.22	0.2216	-0.0192	1.48	0.1391	0.0232

Note. AL = Achievement level; WlcT = Welch's t-test; * symbol indicates that the test is significant at significance level of 0.05.

Table 8f. Descriptive statistics for scale scores of mathematics grades 4 and 7.

Statistics	Grade 4		Grade7	
	Pilot	Non-Pilot	Pilot	Non-Pilot
Mean	546.3	546.3	545.1	545.5
SD	10.33	9.69	9.34	9.03
Median	547.0	547.0	544.0	544.0
Min	522.0	525.0	526.0	529.0
Max	570.0	570.0	573.0	573.0

Figure 1f. Scale score distributions for mathematics grades 4 and 7.

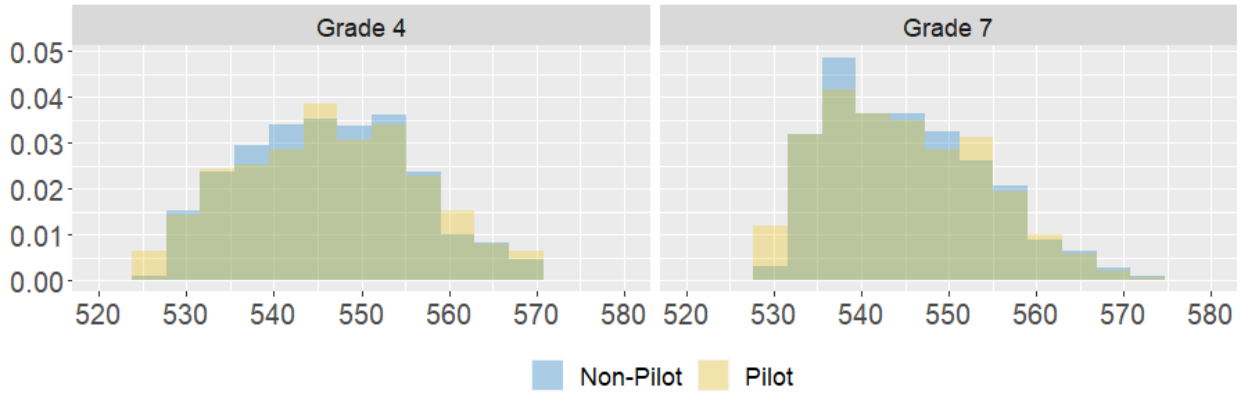
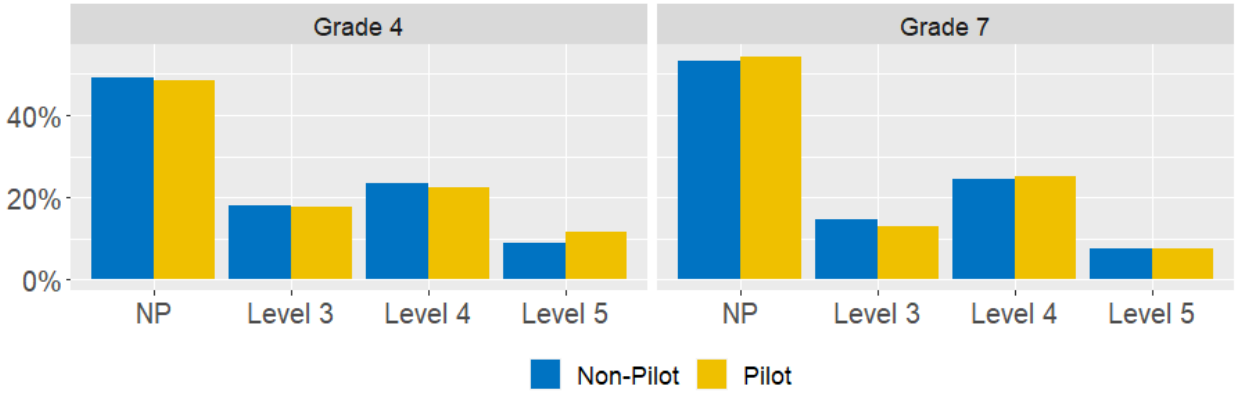


Figure 2f. Achievement level distribution for mathematics grades 4 and 7.



Note. NP = Not Proficient.

F.2.2. Reading

The test results and effect sizes for both grades in reading are provided in Table 9f. The mean scale scores for both grades 4 and 7 were statistically different between the pilot and non-pilot schools for the matched data. As shown in Table 10f, however, the difference in mean scale scores was 0.8 and 0.7 for grades 4 and 7, respectively, which translates to a 0.08 and 0.07 difference in standard deviation. Note that Cohen's *d* is 0.078 and 0.073 for grades 4 and 7, respectively. Such small difference suggests that scale scores were not practically different between the pilot and non-pilot schools. The scale-score distributions for the pilot and non-pilot students were also very similar (see Figure 3f).

Similar to scale scores, the proportion of achievement levels between the pilot and non-pilot schools was also statistically different for both grades. However, effect sizes were negligible, which were 0.047 and 0.071 for grades 4 and 7, respectively. The achievement level distributions for the two groups, which are provided in Figure 4f, also displayed nearly identical patterns for all four achievement levels.

Table 9f. Comparison of scale scores and achievement levels for reading grades 4 and 7.

Outcome Variable	Test	Grade 4			Grade 7		
		Statistic	p-value	Effect Size	Statistic	p-value	Effect Size
Scale score	WlcT	4.97	0.0000*	0.0778	4.66	0.0000*	0.0731
AL	2propZ	2.98	0.0029*	0.0466	4.50	0.0000*	0.0705

Note. AL = Achievement level; WlcT = Welch's t-test; * symbol indicates that the test is significant at significance level of 0.05.

Table 10.f. Descriptive statistics of scale scores for reading grades 4 and 7.

Statistics	Grade 4		Grade7	
	Pilot	Non-Pilot	Pilot	Non-Pilot
Mean	541.8	542.6	551.2	551.9
SD	10.13	10.18	9.77	9.69
Median	543.0	543.0	551.0	552.0
Min	515.0	519.0	528.0	528.0
Max	567.0	568.0	579.0	580.0

Figure 3f. Scale score distributions for reading grades 4 and 7.

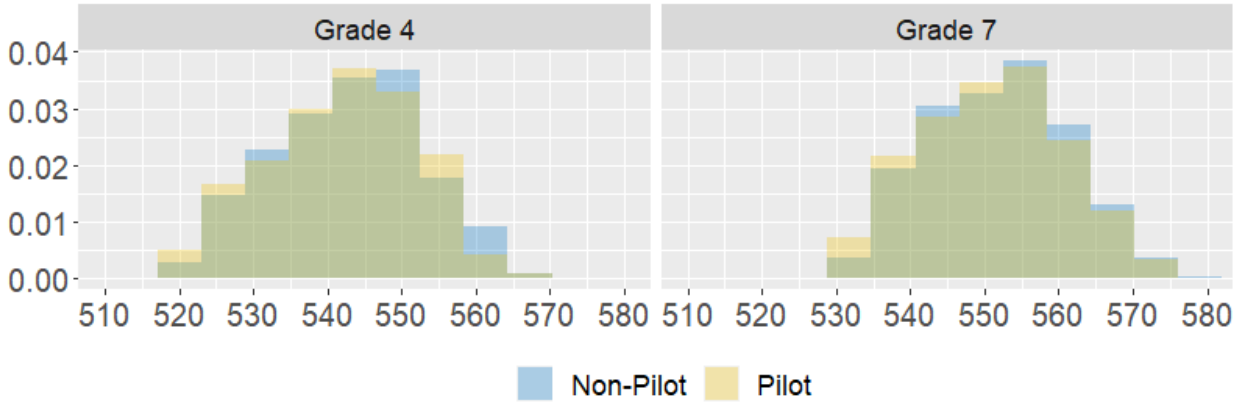
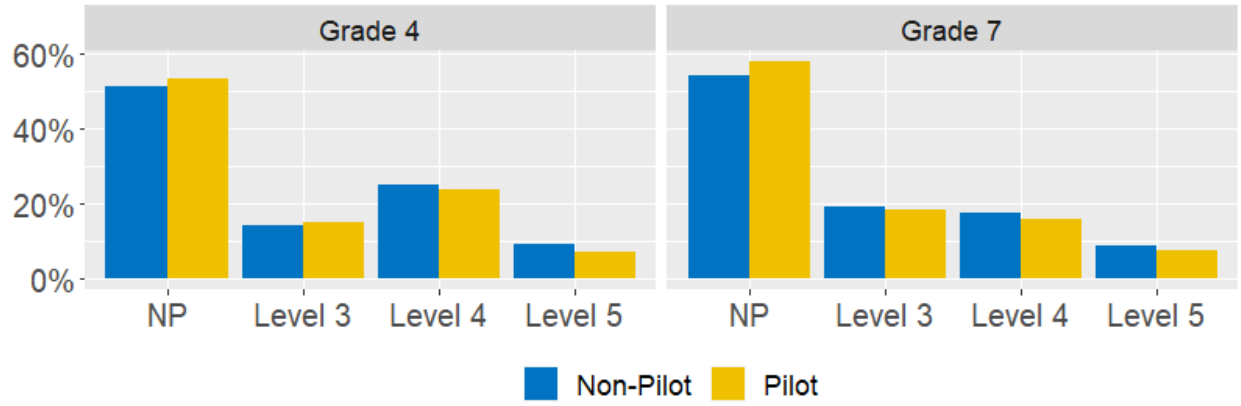


Figure 4f. Achievement level distributions for reading grades 4 and 7.



Note. NP = Not Proficient.

4.3. Conclusion

In summary, after controlling for six covariates at the student level and two covariates at the school level, there were no systematic differences in students' performances as reported by scale scores and achievement levels between students who took the multistage fixed adaptive EOG as part of the NCPAT pilot and students who took the traditional EOG. Although test results were statistically significant for mathematics grade 7 and reading grades 4 and 7, the observed differences were not practically meaningful evidenced by small effect size measures (less than 0.2). Therefore, the results presented in this memo support that scale scores generated from students who took the multistage fixed adaptive EOG are comparable to scale scores from a randomly equivalent subset of students who took the traditional EOG after controlling for sampling differences.

Figure 5f. Graphical balance evaluation of pre-matched data for mathematics grades 4 and 7.

Figure 6f. Graphical balance evaluation of pre-matched data for reading grades 4 and 7.

Figure 7f. Graphical balance evaluation of matched data for mathematics grades 4 and 7.

Figure 8f. Graphical balance evaluation of matched data for reading grades 4 and 7.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (Eds.). (2014). *Standards for educational and psychological testing*. Washington DC: American Educational Research Association.

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Upper Saddle River, NJ: Prentice Hall/Pearson Education.

Cai, L., Thissen, D., & du Toit, S. H. C. (2011). IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling [Computer software]. Lincolnwood, IL: Scientific Software International.

Camilli, G. & Shepard, L.A. (1994). *Methods for Identifying Biased Test Items*. Thousand Oaks, CA: Sage Publications, Inc.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Taylor and Francis.

Cramér, H. (1946). *Mathematical methods of statistics*. Princeton University Press.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika, 22, 297–334.

Hambleton, R. K. (2000). Advances in Performance Assessment Methodology. *Applied Psychological Measurement*, *24*, 291-293.

Hambleton, R. & Swaminathan, H. (1985). *Item response theory*. Boston: Kluwer Nijhoff.

Ho, D., Imai, K., King, G., & Stuart, E. (2018). Matchit: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*, *42*, 1-28.

Holland, P. W., & Thayer, D. T. (1985). *An alternative definition of the ETS delta scale of item difficulty* (Research Report No. 85-64). ETS.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Lawrence Erlbaum Associates, Inc.

Hu, L.-t., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1–55.

Kim. S. (2006). A comparability study of IRT fixed parameter calibration methods. *Journal of Educational Measurement*, *43*, 355-381.

Lee, W., Hanson, B. A., & Brennan, R. L. (2002). Estimating consistency and accuracy indices for multiple classifications. *Applied Psychological Measurement*, *26*, 412-432.

Livingston, S. A. & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement*, *32*, 179–197.

Luecht, R., Brumfield, T., & Breithaupt, K. (2006). A testlet assembly design for adaptive multistage tests. *Applied Measurement in Education*, *19*, 189-202.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, *22*, 719-748.

Perie, M (2020). Comparability Across Different Assessment Systems. In A. I. Berman, E. H. Haertel, & J. W. Pellegrino, *Comparability of Large-Scale Educational Assessments: Issues and Recommendations* (pp. 123-148). Washington, DC: National Academy of Education.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*, 41–45.

Thissen, D. (2000). Reliability and measurement precision. In *Computerized Adaptive Testing* (pp. 159-184). New York, NY: Routledge.

Thissen, D., & Orlando, M. (2001). Item response theory for items scored in two categories. In D. Thissen & H. Wainer (Eds), *Test Scoring* (pp. 73-140). Mahwah, NJ: Lawrence Erlbaum Associates.

Yen, W. M., & Fitzpatrick, A. R. (2006). Item Response Theory. In R. L. Brennan (Ed.), *Educational Measurement* (4th Ed., pp. 111-153). Westport, CT: Praeger.