

Response Probability Criterion and Subgroup Performance

Canda D. Mueller, Pearson
M. Christina Schneider, CTB/McGraw-Hill
Karla Egan, CTB/McGraw-Hill

Response Probability Criterion and Subgroup Performance

In the standard setting literature, there has been much debate about the most appropriate response probability (RP) to use in an item mapping procedure such as the Bookmark Standard Setting Procedure. Little research has been implemented, however, that investigates the relationship between various RP values and how students are classified into achievement levels. This study examined the effect that cut scores produced using different RP criteria have on the assignment of students to achievement level categories. Data from three grades and two subjects were used. Each examinee's score was categorized into an achievement level based upon the cut score set using each RP criterion. Binary logistic regression analyses were conducted for race/ethnicity subpopulations in each grade and subject.

Response Probability Criterion and Subgroup Performance

Item mapping based standard setting methods such as the Bookmark Standard Setting Procedure (BSSP; Lewis, Mitzel & Green, 1996), the Mapmark method (Schulz & Mitzel, 2005), and the Item-descriptor Matching Method (IDM, Ferrara, Perie, and Johnson, 2002) order items in an ordered item booklet (OIB) based upon a response probability (RP) value. That is, the likelihood that a person answers the item correctly. Though item mapping is based in item response theory, Loricé (2001) aptly noted that item mapping requires the human judgmental process because the RP must be determined in order to align an item with a specific score point.

When conducting a standard setting centered in item mapping, the selection of an RP value is an important policy decision for two reasons. First, the OIB is created by mapping items onto the scale so that the probability of a student answering an item correctly is equal to the selected RP value. Second, in the BSSP and the Mapmark method, judges are asked to place the bookmark for an achievement level after the last item that they deem a student with just enough knowledge, skills, and abilities could answer correctly with the selected RP value. That is, the bookmark is placed after the last item judges feel students can answer correctly with a particular level of mastery, and it is placed on the item that judges believe represents more knowledge, skills, and abilities than expected for the student who has the minimal knowledge to be considered in the achievement level.

The selection of an RP value has been discussed extensively for areas such as item mapping and scale anchoring (Donoghue, 1996, 1997; Zwick, Senturk, Wang, &

Loomis, 2001; Kolstad, 1999) for the purposes of describing student abilities at specific scale locations to help the public understand what students know and can do on an assessment. The selection of an RP value for the purpose of standard setting has not received the same level of attention publically, although it is often discussed with states in technical advisory meetings in preparation for standard setting. Lewis, Mitzel, and Green (1996) used the guess adjusted RP67 to order items in the Bookmark Standard Setting Procedure. Huynh (1998, 2000a, 2000b, 2006) suggested that the most important issue to consider is obtaining the maximum amount of information from the correct response. Huynh points out that this just happens to occur mathematically at an RP value of .67 or $2/3$ for both the Rasch and 2PL models. He goes on to note that for the 3PL model one must take the pseudo-guessing parameter into account so that the information from a correct response is maximized at $(2+c)/3$. Although one may argue psychometrically for different RP values, it should be noted that because standard setting is a judgmental process, it is important to consider what panelists who are content experts believe denotes sufficient mastery.

Zwick, Senturk, Wang, and Loomis (2001) determined that, in general, content experts found RP50 too low a threshold to determine mastery. Rather, content experts generally felt comfortable with values around RP70. Therefore, Zwick, et al. determined that RP65 – RP74 align more closely with content experts' expectations. Williams and Schulz (2005) noted that participants were more comfortable with mastery defined as RP67 than RP50 when conducting field trials of the Mapmark method using RP50 and RP67. The National Academies of Sciences (2005) reported similar findings in their study of using RP50, RP67, and RP80.

In their review of the literature, Karantonis and Sireci (2006) found that the research on panelists' perceptions seems to support the use of RP67; however, they noted that little study has been directed toward determining how the choice of RP value affects student classification into achievement levels. It is possible that different RP values may systematically narrow or widen existing achievement gaps on the reporting metrics used to interpret state assessments, and this is an area in need of study.

It has been documented in the educational research literature and state testing results that an achievement gap between African-American students and white students exists. Cited reasons for achievement differences include larger proportions of minority students living in poverty, high poverty home conditions may lack educational stimulation, and high mobility (Barton, 2002; Ramey, 2000; Slavin & Madden, 2001). Although each of these reasons has historically had an impact on minority achievement, various RP values may exacerbate the gap in student achievement when those comparisons are made using achievement level classifications.

Achievement level classifications are the most common metric used by states in the evaluation of schools for AYP, and states oftentimes have their judges review impact data disaggregated by subgroup. Mueller and Peyton (2007) reported that changing the RP value impacted the achievement level classification of students when disaggregated by SES status. They found that the placement of the cut score based upon RP value widened or narrowed the achievement gap in terms of the percent of student in each achievement level. Their research utilized the Rasch model, in which dictomous items remain in the same order regardless of RP value, and it provided support for using RP67

because the achievement gap was narrowed when the cut score was set using RP67 in comparison to RP80 when the basic category was the analyses focus.

As noted by Beretvas (2004), items in an OIB will change locations as IRT models and RP values are varied. Post hoc analyses allow researchers to determine how classification of students change as the RP value is varied from RP50 to RP67 to RP80. In a BSSP, items modeled with 3PL are located in the OIB by setting the pseudo-guessing parameter to zero (Cizek & Bunch, 2007). That is, the 3PL model is adjusted for guessing. In a post hoc study, cut scores are previously established via a standard setting. Therefore to investigate various RP values, the data must analyzed after calculating cut scores based on different RP values for a known bookmark placement. This requires the assumption that if panelists had received an OIB ordered by a different RP value, they would still have placed the bookmark on the same item. That is, the assumption is that the item, regardless of RP value, is the item that represents more than enough skill for a student to be considered proficient. Therefore, the parameters for this item can be used to calculate the cut score under different RP values.

To determine the cut score values for different RP values, the BSSP guessing adjusted RP value can used for three RP values: 50, 67, and 80. In the typical BSSP, items are located in the OIB using an adjustment to RP67. An item's guess-adjusted RP67 means that an examinee with a scale score at the item's location would have a 67 percent chance of answering the item correctly, after guessing is factored out when the three-parameter logistic model is used. If item characteristic curve (ICC) for item i with ICC $P_i(\theta)$ and guessing parameter c_i is adjusted for guessing using the following: $P_i^*(\theta) = [P_i(\theta) - c_i] / (1 - c_i)$, then the RP Guess Adjusted 67 (RP) of item i is given by θ such that

$P_i^*(\theta) = 0.67$ (Lorie, Egan, Mercado, Brandstrom, and Tele'a, 2004). For the remainder of this paper, RP values assume an adjustment for guessing.

Research Question, Data, and Methods

For this study, the locations of cut scores for Grade 3, 5, and 8 English Language Arts and mathematics were compared for RP50, RP67, and RP80. Scale score frequency distributions were applied to the item locations to understand how impact data varied by RP value. Next, using logistic regression, the odds of being classified proficient or above for RP67 and RP80 were compared against RP50 for the total group and for the race/ethnicity subgroups to explore the affect varying RP values have on the classifications of students on statewide assessments in English Language Arts and Mathematics.

This research extends the earlier work by Mueller (2005). In that research, both SES and race/ethnicity gaps were examined using the Rasch model on existing statewide data. At RP50, more students were classified as proficient or above regardless of race or SES. The research used multinomial logistic regression to investigate the categorization of students at four achievement levels. This study extends Mueller's work in two ways using logistic regression. First, rather than the Rasch model, the 3PL model will be used. Second, the affect of varying RP values on race/ethnicity subpopulations will be compared in terms of classifying students as proficient or above and below proficient.

Data Source

English Language Arts (ELA) and mathematics data for grades 3, 5, and 8 were used in this research. For confidentiality purposes, the state is not named and only selected grades are reported. The initial cut scores for this state data set were determined

using the BSSP with RP67. Table 1 shows the number of students categorized as proficient or above on the standard setting form used in this study for overall, African-American, and white subgroups. Multiple race/ethnicity categories are included in the overall calculation, however only African-American and white students were included for the subpopulation analysis in this study.

Table 1: Frequency of Students Meeting Initial Proficiency Categorization

Subject	Grade	Overall	African-American	White
ELA	3	69,729	17,999	48,266
	5	71,802	18,512	50,059
	8	69,324	16,977	49,226
Math	3	69,905	18,019	48,400
	5	71,905	18,529	50,130
	8	69,853	17,193	49,491

The percentage of students meeting the requirements of the proficiency or above category in each grade and subject are shown in Table 2 for the overall population and the African-American and white subgroups. As can be noted from the table, the percentage of students passing the exam was fairly high in both subjects and in all three grade levels. The lowest percentage passing the exam appears in 8th grade mathematics for African-American students. The remaining percentages indicate that more than three-fifths of students meet the proficiency requirements for every grade and subject.

Table 2: Percentage of Students Meeting Initial Proficiency Categorization (RP67)

Subject	Grade	Overall	African-American	White
ELA	3	80.83%	68.22%	85.38%
	5	81.93%	71.11%	86.33%
	8	81.24%	67.74%	86.40%
Math	3	79.07%	63.52%	85.13%
	5	82.48%	66.03%	88.71%
	8	67.79%	43.26%	76.59%

In this study, the categorical dependent variable of achievement level was used. The achievement level variable was coded as below proficient and as proficient or above. Because proficient or above for the purpose of NCLB serves as a pass status, it served as the reference category used to study the significance of the affect of changing the RP value. The No Child Left Behind Act of 2002 requires that all students be at the level of proficiency by 2014. Using pass as the reference category in each analysis allows for comparison of the levels below proficiency to the level prescribed by the legislation. RP50 was chosen to serve as the reference category for the independent variable to ease the interpretation. Race/ethnicity status was included as an additional independent variable with the values of African-American and white. In this case, classification in the white category served as the reference group. The use of logistic regression allows for prediction of achievement level categorization given a particular RP value for the student population disaggregated by race/ethnicity status.

The logistic regression model compared a model containing only an intercept to a model containing the RP factor, the race/ethnicity factor, and the interaction of the two factors. This model is shown as Equation 1. The main focus was the interaction of RP and race/ethnicity status. A significant change in the χ^2 value between the intercept only

model and the full model (including the main effects of RP, race/ethnicity, and the interaction of the two) indicates that a relationship exists between the set of factors and the dependent variable of achievement level. That is, changing the response probability value impacts the categorization of students into achievement level categories differently based on subgroup.

$$\hat{Y}_i = \frac{\exp(A + B_1 X_1 + B_2 X_2)}{1 + \exp(A + B_1 X_1 + B_2 X_2)} \quad (1)$$

where \hat{Y}_i = probability of an outcome (*fail, pass*),

A = intercept value,

B_1 = parameter estimate for the RP variable,

X_1 = RP variable,

B_2 = parameter estimate for the race/ethnicity variable,

X_2 = race/ethnicity variable.

Significant interactions of RP and race/ethnicity using a Bonferroni correction to adjust for inflated Type I error were further analyzed.

Results

Table 3 shows the post-hoc calculated cut scores based on different RP values using the item parameters for the item used to calculate the cut score for RP67. This required the assumption that if panelists had received an OIB ordered by a different RP value, they would still have placed their bookmark on the same item. As is expected, in

each case the cut score is lowest for RP50. Table 4 shows the percentage of students as proficient or above for each RP level.

Table 3: Scale Score Cut Score Values for Each RP

Subject	Grade	RP50	RP67	RP80
ELA	3	604	615	624
	5	617	631	645
	8	641	651	661
Math	3	579	591	603
	5	596	619	640
	8	642	672	702

Table 4: Percentage of Students Meeting Proficiency for each RP

RP	Subject	Grade	Overall	African-American	White	
50	ELA	3	86.65%	77.64%	89.75%	
		5	89.36%	83.03%	92.02%	
		8	86.67%	76.86%	90.58%	
	Math	3	86.16%	75.31%	90.41%	
		5	92.78%	85.75%	95.52%	
		8	85.94%	72.74%	90.75%	
	67	ELA	3	80.83%	68.22%	85.38%
			5	81.93%	71.11%	86.33%
			8	81.24%	67.74%	86.40%
Math		3	79.07%	63.52%	85.13%	
		5	82.48%	66.03%	88.71%	
		8	67.79%	43.26%	76.59%	
80		ELA	3	70.57%	53.75%	76.78%
			5	71.03%	54.65%	77.44%
			8	74.36%	56.83%	80.87%
	Math	3	69.33%	49.26%	77.12%	
		5	65.72%	41.07%	75.04%	
		8	40.10%	15.48%	48.70%	

In reporting the results of the binary logistic regression, it is important to correct for the multiple comparisons that were involved in the analyses. A Bonferroni-type correction was used to control for inflated Type I error. For the analyses presented here, $\alpha = .05$ was divided by one to account for the main effects of race/ethnicity so that these parameter estimates were compared to $\alpha = .05$. To account for the main effects of differing RP values, $\alpha = .05$ was divided by 2 so that these parameter estimates were compared to $\alpha = .025$. Finally, $\alpha = .05$ was divided by 2 so that the parameter estimates for the interaction effects were compared to $\alpha = .025$. If an RP by race interaction was found to be significant at this level, the simple main effects were compared to $\alpha = .025/3 = .008$ to account for the three possible simple main effect comparisons.

ELA Results

In each grade, the results showed that adding RP and race/ethnicity to the intercept-only model was justified based on the results of the chi-square goodness-of-fit tests. For each grade, the goodness-of-fit tests indicated that it was necessary to evaluate the full model containing the main effects of RP and race/ethnicity along with the interaction of these two variables.

The third grade data indicated that the model accounted for 9.3% of the total variance in achievement level categorization. Table 5 shows the parameter estimates for the model including the interaction of RP and race/ethnicity. The main interest was in the interaction effects since these effects represent achievement level categorization gaps between African-American and white students. Significant interactions were further analyzed through follow-up tests using a Bonferroni correction as described previously.

Table 5: Parameter Estimates Showing the Likelihood of Achievement Categorization for Grade 3 ELA

		Fail		
Type of Effect	I.V.	B	S.E.	Exp(B)
Main				
Race/Ethnicity	A-A	-0.92	.02	0.40*
RP	67	-0.41	.02	0.67**
	80	-0.97	.02	0.38**
Interaction				
Race/Ethn and RP	67A-A	-0.08	0.03	.93**
	80A-A	-0.12	0.03	.89**
Constant		2.17	0.02	

Note. I.V.= independent variable, RP= response probability, B= regression coefficient, S.E.= standard error, A-A=African-American
 Pass, RP50, and white are reference categories
 * $p < .05$, ** $p < .025$

In comparing students classified below proficient (fail) to those classified as proficient and above (pass), there was a significant difference at both RP levels compared to RP50 for African-American students versus white students. Follow-up analyses were conducted on each significant difference, and the results are in Table 6. The odds of being classified as below proficient for African-American students are approximately one-third as high as white students at RP50, but at RP67 the odds of this classification are two-thirds as high. At RP80, the odds of African-American students being classified as below proficient are again approximately one-third as high.

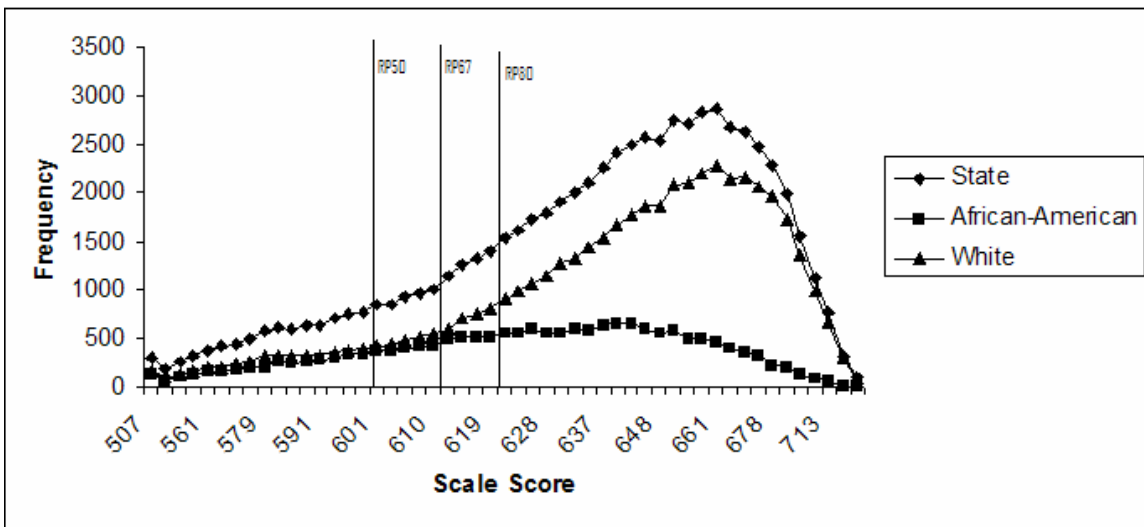
Table 6: Simple Main Effects Split by RP for Grade 3 ELA

		Fail		
RP	Race/Ethnicity	B	S.E.	Exp(B)
50	A-A	-1.00	.01	0.37*
67	A-A	-0.44	.02	0.65*
80	A-A	-1.02	.01	0.36*

Note. RP= response probability, B= regression coefficient, S.E.= standard error
 Pass and African-American are reference categories
 * $p < .008$

The graphical representation of these results is shown as Figure 1. In this figure, the vertical lines represent the cut scores associated with RP50, RP67, and RP80. Based on the figure, there is indication that the proportion of students classified as below proficient are similar for the subgroup populations for the three RP values.

Figure 1: 3rd Grade Race/Ethnicity Scale Scores per RP in ELA



The fifth grade data set provided similar results. The full model accounted for 10.6% of the total variance in achievement level categorization. Table 7 shows the parameter estimates for the model including the interaction of RP and race/ethnicity.

Table 7: Parameter Estimates Showing the Likelihood of Achievement Categorization for Grade 5 ELA

Fail				
Type of Effect	I.V.	B	S.E.	Exp(B)
Main				
Race/Ethnicity	A-A	-0.86	.03	0.42*
RP	67	-0.60	.02	0.55**
	80	-1.21	.02	0.30**
Interaction				
Race/Ethn and RP	67A-A	-0.09	0.03	.92**
	80A-A	-0.19	0.03	.83**
Constant		2.45	0.02	

Note. I.V.= independent variable, RP= response probability, B= regression coefficient, S.E.= standard error, A-A=African-American

Pass, RP50, and white are reference categories

* $p < .05$, ** $p < .025$

As was the case with the third grade data set, there was a significant difference at both RP levels compared to RP50 for African-American students versus white students. Follow-up analyses were conducted on each significant difference, and the results are in Table 8. The odds of being classified as below proficient are again approximately one-third as high for African-American students as for white students at RP50, but at RP67 the odds of this classification are one-half as high. At RP80, the odds of African-American students being classified as below proficient were slightly less than one-third as high as for white students.

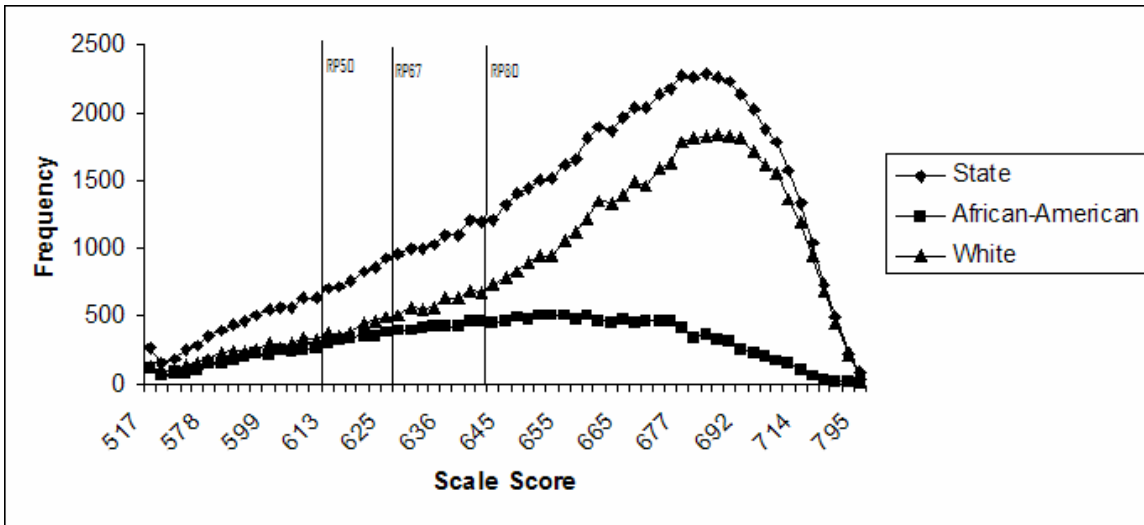
Table 8: Simple Main Effects Split by RP for Grade 5 ELA

RP	Race/Ethnicity	Fail		
		B	S.E.	Exp(B)
50	A-A	-0.97	.01	0.38*
67	A-A	-0.64	.02	0.53*
80	A-A	-1.29	.02	0.28*

Note. RP= response probability, B= regression coefficient, S.E.= standard error
 Pass and African-American are reference categories
 * $p < .008$

The graphical representation of these results is shown as Figure 2. In this figure, the vertical lines represent the cut scores associated with RP50, RP67, and RP80. Based on the figure, there is indication that the proportion of students classified as below proficient are similar for the subgroup populations for the three RP values.

Figure 2: 5th Grade Race/Ethnicity Scale Scores per RP in ELA



The eighth grade ELA data set indicated results very similar to those from the third grade. In this case, the model accounted for 9.1% of the total variance in achievement level categorization. Table 9 shows the parameter estimates for the model including the interaction of RP and race/ethnicity. The main interest was in the interaction effects since these effects represent achievement level categorization gaps

between African-American and white students. Significant interactions were further analyzed through follow-up tests using a Bonferroni correction as described previously.

Table 9: Parameter Estimates Showing the Likelihood of Achievement Categorization for Grade 8 ELA

Fail				
Type of Effect	I.V.	B	S.E.	Exp(B)
Main				
Race/Ethnicity	A-A	-1.06	.02	0.35*
RP	67	-0.42	.02	0.66**
	80	-0.82	.02	0.44**
Interaction				
Race/Ethn and RP	67A-A	-0.04	0.03	.96**
	80A-A	-0.10	0.03	.90**
Constant		2.26	0.02	

Note. I.V.= independent variable, RP= response probability, B= regression coefficient, S.E.= standard error, A-A=African-American
 Pass, RP50, and white are reference categories
 * $p < .05$, ** $p < .025$

As shown in Table 10, the results are very similar to those found for the third grade data set. Specifically, the odds of being classified as below proficient are approximately one-third as high for African-American students as for white students at RP50, but at RP67 the odds of this classification are two-thirds as high. At RP80, the odds of African-American students being classified as below proficient are slightly more than one-third as high.

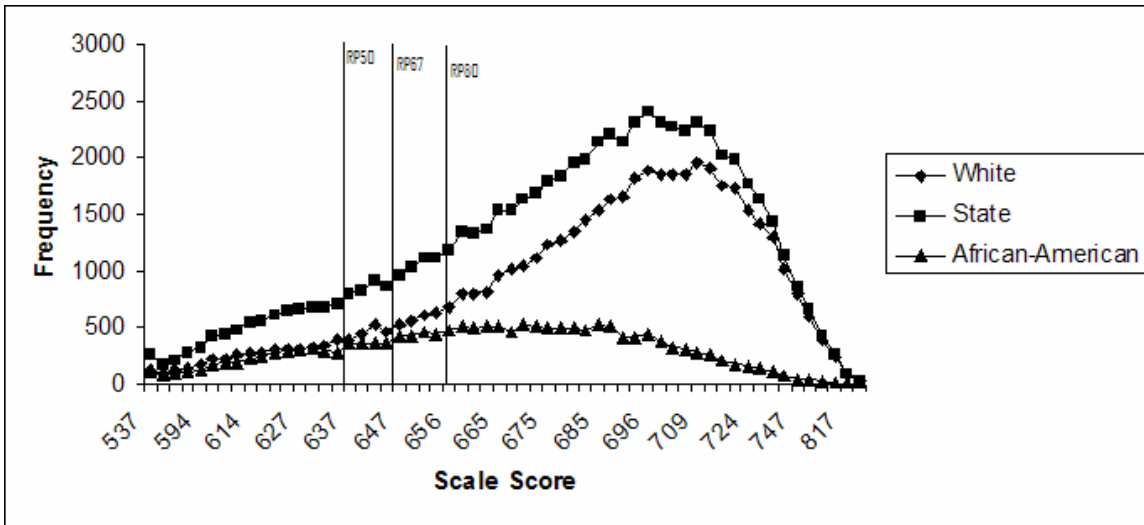
Table 10: Simple Main Effects Split by RP for Grade 8 ELA

RP	Race/Ethnicity	Fail		
		B	S.E.	Exp(B)
50	A-A	-1.12	.01	0.33*
67	A-A	-0.43	.02	0.65*
80	A-A	-0.86	.02	0.42*

Note. RP= response probability, B= regression coefficient, S.E.= standard error
 Pass and African-American are reference categories
 * $p < .008$

The graphical representation of these results is shown as Figure 3 below. In this figure, the vertical lines represent the cut scores associated with RP50, RP67, and RP80.

Figure 3: 8th Grade Race/Ethnicity Scale Scores per RP in ELA



Mathematics Results

The mathematics data sets yielded results that closely mirrored those presented for the English Language Arts data sets. All three will be presented, but discussion will be very brief.

The grade three data indicated that the model accounted for 11.9% of the total variance in achievement level categorization. Table 11 shows the parameter estimates for the model including the interaction of RP and race/ethnicity. The main interest was in the

interaction effects since these effects represent achievement level categorization gaps between African-American and white students. Significant interactions were further analyzed through follow-up tests using a Bonferroni correction as described previously.

Table 11: Parameter Estimates Showing the Likelihood of Achievement Categorization for Grade 3

Mathematics

Fail				
Type of Effect	I.V.	B	S.E.	Exp(B)
Main				
Race/Ethnicity	A-A	-1.13	.02	0.32*
RP	67	-0.50	.02	0.61**
	80	-1.03	.02	0.36**
Interaction				
Race/Ethn and RP	67A-A	-0.06	0.03	.94**
	80A-A	-0.12	0.03	.89**
Constant		2.24	0.02	

Note. I.V.= independent variable, RP= response probability, B= regression coefficient, S.E.= standard error, A-A=African-American

Pass, RP50, and white are reference categories

* $p < .05$, ** $p < .025$

Follow-up analyses were conducted for significant differences, and the results are in Table 12. The odds of being classified as below proficient are approximately one-third as high for African-American students as for white students at RP50 and again at RP80; however, at RP67 the odds of this classification are slightly more than one-half as high.

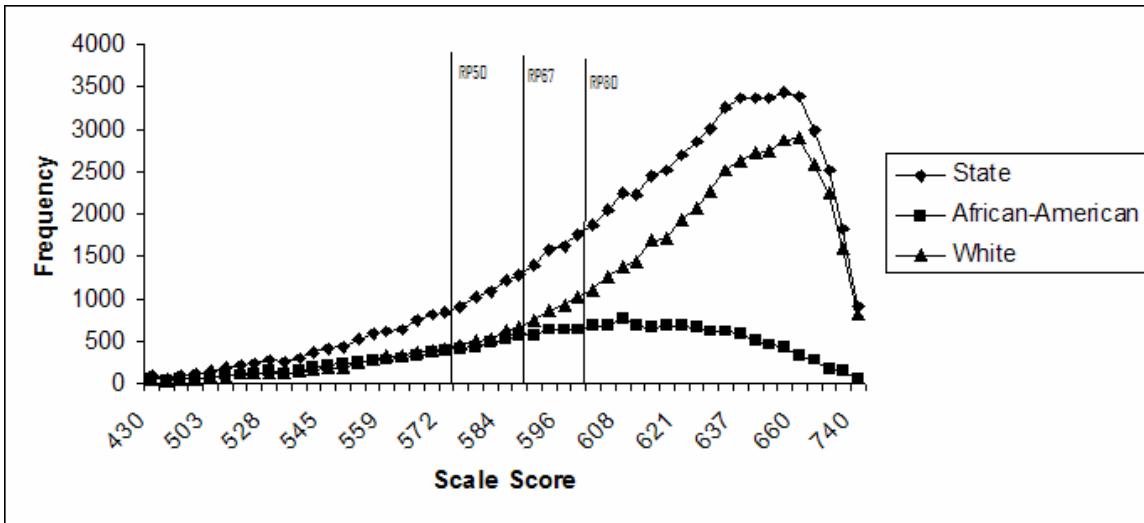
Table 12: Simple Main Effects Split by RP for Grade 3 Mathematics

RP	Race/Ethnicity	Fail		
		B	S.E.	Exp(B)
50	A-A	-1.20	.01	0.30*
67	A-A	-0.53	.02	0.59*
80	A-A	-1.08	.02	0.34*

Note. RP= response probability, B= regression coefficient, S.E.= standard error
 Pass and African-American are reference categories
 * $p < .008$

The graphical representation of these results is shown as Figure 4 below. In this figure, the vertical lines represent the cut scores associated with RP50, RP67, and RP80. Based on the figure, there is indication that the proportion of students classified as below proficient are similar for the subgroup populations for the three RP values.

Figure 4: 3rd Grade Race/Ethnicity Scale Scores per RP in Mathematics



In the fifth grade mathematics data set, the model accounted for 21.3% of the total variance in achievement level categorization. Table 13 shows the parameter estimates for the model including the interaction of RP and race/ethnicity. Significant interactions were further analyzed through follow-up tests using a Bonferroni correction as described previously.

Table 13: Parameter Estimates Showing the Likelihood of Achievement Categorization for Grade 5

Mathematics

Fail				
Type of Effect	I.V.	B	S.E.	Exp(B)
Main				
Race/Ethnicity	A-A	-1.27	.03	0.28*
RP	67	-1.00	.03	0.37**
	80	-1.96	.02	0.14**
Interaction				
Race/Ethn and RP	67A-A	-0.13	0.04	.88**
	80A-A	-0.20	0.04	.82**
Constant		3.06	0.02	

Note. I.V.= independent variable, RP= response probability, B= regression coefficient, S.E.= standard error, A-A=African-American

Pass, RP50, and white are reference categories

* $p < .05$, ** $p < .025$

Follow-up analyses were conducted on each significant difference, and the results are in Table 14. The odds of being classified as below proficient are approximately one-fourth as high for African-American students as for white students at RP50. At RP67, the odds of this classification for African-American students are one-third as high. At RP80, the odds of African-American students being classified as below proficient were slightly more than one-tenth as high as for white students.

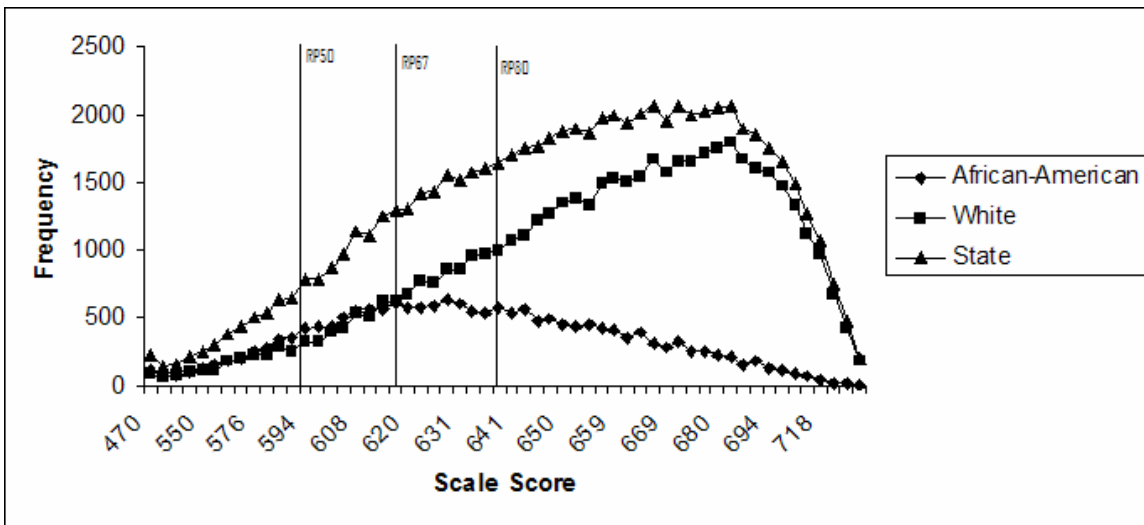
Table 14: Simple Main Effects Split by RP for Grade 5 Mathematics

RP	Race/Ethnicity	Fail		
		B	S.E.	Exp(B)
50	A-A	-1.41	.01	0.25*
67	A-A	-1.07	.02	0.34*
80	A-A	-2.05	.02	0.13*

Note. RP= response probability, B= regression coefficient, S.E.= standard error
 Pass and African-American are reference categories
 * $p < .008$

The graphical representation of these results is shown as Figure 5 below. In this figure, the vertical lines represent the cut scores associated with RP50, RP67, and RP80.

Figure 5: 5th Grade Race/Ethnicity Scale Scores per RP in Mathematics



In the eighth grade mathematics data set, the model accounted for 29.5% of the total variance in achievement level categorization. Table 15 shows the parameter estimates for the model including the interaction of RP and race/ethnicity. Significant interactions were further analyzed through follow-up tests using a Bonferroni correction as described previously.

Table 15: Parameter Estimates Showing the Likelihood of Achievement Categorization for Grade 8

Mathematics

Fail				
Type of Effect	I.V.	B	S.E.	Exp(B)
Main				
Race/Ethnicity	A-A	-1.30	.02	0.27*
RP	67	-1.10	.02	0.33**
	80	-2.34	.02	0.10**
Interaction				
Race/Ethn and RP	67A-A	-0.16	0.03	.86**
	80A-A	-0.34	0.03	.71**
Constant		2.28	0.02	

Note. I.V.= independent variable, RP= response probability, B= regression coefficient, S.E.= standard error, A-A=African-American
 Pass, RP50, and white are reference categories
 * $p < .05$, ** $p < .025$

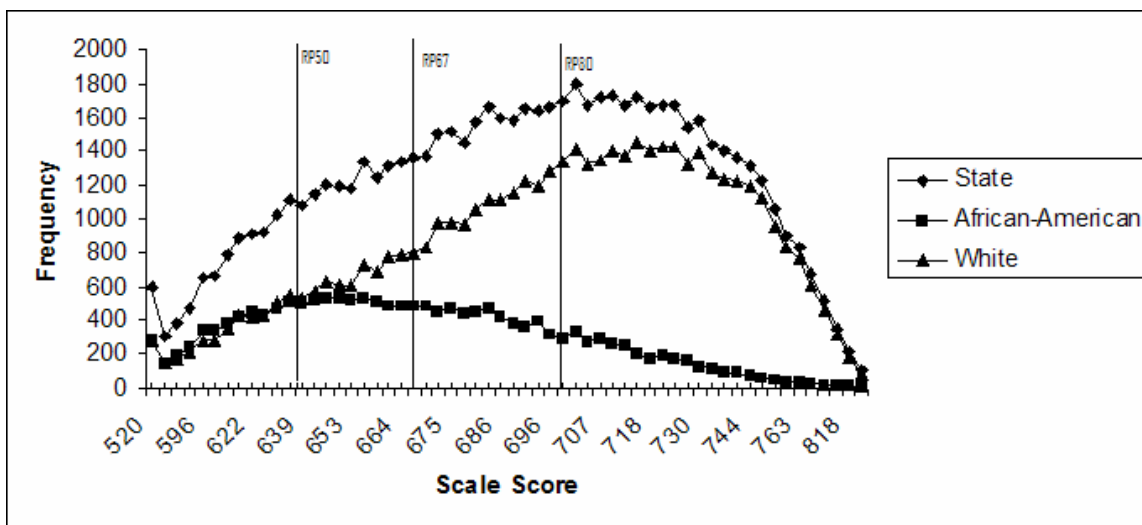
Table 16 provides the results for the follow-up analyses. The odds of being classified as below proficient are approximately one-fourth as high for African-American students as for white students at RP50. The odds of this classification are slightly less than one-third as high at RP67 and about one-tenth as high at RP80.

Table 16: Simple Main Effects Split by RP for Grade 8 Mathematics

		Fail		
RP	Race/Ethnicity	B	S.E.	Exp(B)
50	A-A	-1.47	.01	0.23*
67	A-A	-1.17	.02	0.31*
80	A-A	-2.44	.02	0.09*

Note. RP= response probability, B= regression coefficient, S.E.= standard error
 Pass and African-American are reference categories
 * $p < .008$

The graphical representation of these results is shown as Figure 9. In this figure, the vertical lines represent the cut scores associated with RP50, RP67, and RP80.

Figure 6: 8th Grade Race/Ethnicity Scale Scores per RP in Mathematics

Discussion

The results from this research suggest that there is an interaction between student race and the RP value used to order item for an item-mapping procedure. It appears that the odds of African American students being categorized as below proficient tend to be slightly higher under RP67 and RP80 than when they are compared to RP50; however, on this particular assessment the differences were not dramatic.

These results are not surprising given the distribution of student performance on the assessments that were studied. All students tended to be highly successful on these assessments. In fact, student abilities were generally higher than the items that comprise the test. In general, the RP50 cut score tended to be so low that the majority of students were classified as proficient regardless of race/ethnicity. These tests did not have enough variability between subgroups at the cut scores to adequately study the research question. From the figures one can see that the more spread distributions at the varying cut scores,

the more variance in the achievement level varied RP levels accounted for. This should be noted for future research.

Future research can address the deficiency in the current study. The assessments in this study tended to be very easy for the students being tested as is demonstrated by Figures 1 through 6. Researchers should examine assessments that differentiate between low and high ability students (i.e., the ability distributions should optimally be symmetrical rather than skewed as are the current distributions). Future research should also look at the interaction of RP value and other subgroups in addition to the two race/ethnicity subgroups presented in this study. For example, are similar trends observed for students receiving free/reduced lunch or students who are Hispanic? The equity question implied by this research applies to many different subgroups. If equity is an issue when setting an RP value, then it is important that all relevant underserved groups are studied

Finally, researchers may want to examine the impact of retaining the guessing parameter when looking at the interaction of the RP value with subgroup performance. The current research factored the guessing parameter out before calculating the item location for different RP values. In the future, the presence of the guessing parameter may also be varied.

While it was the hope of the current research project to find an RP value that was fair to both majority and minority students, the results did not clearly favor one RP value over another in terms of increasing equity for African-American students. Until this answer is found, the choice of RP value should reflect the way that a state conceptualizes mastery.

REFERENCES

- Barton, P.E. (2003). *Parsing the achievement gap: Baselines for tracking progress*. Retrieved on October 21, 2004 from <http://www.ets.org/research/pic/parsing.pdf>. Princeton, NJ: Educational Testing Service.
- Cizek, G.J. & Bunch, M.B. (2007). *Standard Setting: A Guide to Establishing and Evaluating Performance Standards on Tests*. Thousand Oaks, CA: Sage Publications.
- Donghue, J.R. (1996). *Issues in item mapping: The maximum category information criterion and item mapping procedures for a composite scale*. Paper presented at the annual meeting of the American Educational Research Association, New York City, NY.
- Donoghue, J.R. (1997). *Item mapping to a weighted composite scale*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.
- Embretson, S. E. & Reise, S.P. (2000). *Item Response Theory for Psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Ferrara, S., Perie, M., & Johnson, E. (2002, April). *Setting performance standards: The item-descriptor (ID) matching procedure*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.
- Huynh, H. (1998). On score locations for binary and partial credit items and their applications to item mapping and criterion-referenced interpretation. *Journal of Educational and Behavioral Statistics*, 23, 35-56.

Huynh, H. (2000a). *On item mappings and statistical rules for selecting binary items for criterion-referenced interpretation and Bookmark standard settings*. Paper presented at the annual meeting of National Council on Measurement in Education, New Orleans, LA.

Huynh, H. (2000b). *On Bayesian rules for selecting 3PL binary items for criterion-referenced interpretations and creating booklets for Bookmark standard setting*. Paper presented at the annual meeting of American Educational Research Association, New Orleans, LA.

Karantonis, A, & Sireci, S.G. (2006). The Bookmark standard setting method: A literature review. *Educational Measurement: Issues and Practice*, 25(1), 4-12

Kolstad, A., Cohen, J., Baldi, S., Chan, T., DeFur, E., & Angeles, J. (1998). *The response probability convention used in reporting data from IRT assessment scales: Should NCES adopt a standard?* Washington, DC: U.S. Department of Education, NCES.

Lorié, W.A. (2001). Setting defensible cut scores: Canonical pseudoresponses, item types, and performance standards. *Dissertation Abstracts International*. (Univeristy Microfilms No. 3038119).

Lorié, W.A. Egan, K. L., Mercado, R.L., Brandstrom, A.J., & Tele'a, D. (2004, April). *A comparison of response probability and canonical pseudoresponse Item locations: Implications for the Bookmark Standard Setting procedure*. Paper presented at the annual meeting of National Council on Measurement in Education, San Diego, CA.

- Lewis, D. M., Mitzel, H. C., & Green, D. R. (1996). *Standard setting: A bookmark approach*. Paper presented at the 26th Annual CCSSO National Conference on Large Scale Assessment in Phoenix, AZ.
- Mueller, C.D. (2005). Assessing the consequences of response probability choice on student performance across subpopulations. *Dissertation Abstracts International*.
- Mueller, C.D. & Peyton, V. (2007). *Assessing consequences of varying response probability values on achievement via multinomial logistic regression*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.
- National Academies of Sciences (2005). *Measuring literacy: Performance levels for adults*. Retrieved March 5, 2008, from <http://www.nap.edu/openbook.php?isbn=0309096529&page=36>.
- No Child Left Behind Act of 2001*, Pub. L. No. 107-110, 115 Stat.1425 (2002).
- Ramey, M. (2000). Reducing the white- nonwhite achievement gap. Paper presented at annual meeting of American Educational Research Association, New Orleans, LA.
- Schulz, E.M., & Mitzel, H.C. (April, 2005). *The Mapmark standard setting method*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Quebec, Canada. ERIC Document ED490463.
- Slavin, R.E. & Madden, N.A. (2001). *Reducing the gap: Success for all and the achievement of African- American and Latino students*. Paper presented at annual meeting of American Educational Research Association, Seattle, WA.

Spellings, M. (Jan. 24, 2007). Secretary launches priorities for NCLB reauthorization.

USDOE press release retrieved January 8, 2008, from

<http://www.ed.gov/news/pressreleases/2007/01/01242007.html> .

United States Department of Agriculture. (2002). *National School Lunch Program*

income guidelines. Retrieved April 12, 2005, from

<http://www.fns.usda.gov/cnd/governance/notices/iegs/IEGs02-03.pdf>

United States Department of Agriculture. (2004). *National School Lunch Program fact*

sheet (2004). Retrieved April 12, 2005, from

<http://www.fns.usda.gov/cnd/Lunch/AboutLunch/NSLPFactSheet.pdf>

Williams, N.J., & Schulz, E.M. (April, 2005). *An investigation of response probability*

values used in standard setting. Paper presented at the annual meeting of the

National Council on Measurement in Education, Montreal, Quebec, Canada.

Zwick, R., Senturk, D., Wang, J., & Loomis, S.C. (2001). An investigation of alternative

methods for item mapping in the National Assessment of Educational Progress.

Educational Measurement: Issues and Practice, 20(2), 15-25.