Section 6. English Test

There are five parts to this section. The first part describes how the operational (scoring) items were selected for the new English test. The second part describes how the items were scaled, and the third part describes the factor analyses of the English forms. Summary statistics of student achievement and measures of classification consistency comprise the fourth and fifth parts.

Operational Item Selection

This section summarizes the procedures used for selecting ETS recommended English operational items. It reflects changes MSDE made to Form G, adding one extra SR item for subscore 2 in addition to the ETS recommendation. ETS recommended Form K to be the standard setting form and Form E or H be the secondary form depending on whether easier or harder items are needed for standard setting. Summary statistics (i.e., p-values, item-total correlations) for each of the recommended forms are presented in this document. Data files associated with item selection are posted on the MSDE DocuShare site. Three kinds of separate data files are posted.

1. Augmented form planners with item statistics and operational item designation are in Excel data files named: FP_English_0505_V5_Form***X***.D042005.xls where ***X*** stands for form code. Please note the operational item designation in the column heading "anchor_status." The value "O" indicates an operational item.
2. Item analyses by student ability group summary. This file, called "IA by category offload.xls," contains the item difficulties and item-total correlations for the high (H), medium (M) and low (L) ability groups.
3. Distractor analyses by student ability group. This is a flat text file called "IA by category Offload Distracter analyses.txt." This file contains the distractor analyses for all three ability groups. Please note that this is an extract file of the ETS item analyses output so there was no Maryland ID associated with each item. Instead, a form code and sequence number are added to the first column of the output for the reader to understand the statistics.

The processes that were used to select the operational items for the English forms are as follows:

1. Research conducted Item Analyses and DIF analyses and flagged items unavailable as operational items.
   Flagging criteria:
   P > 0.9 or P < 0.2
   Item-total correlation < 0.2
   Distractor item-total correlation > 0
   Omit rate (conditional code A or B) > 15%[3]
   Item with C-DIF

[3] Omit rates were considered not as crucial as the test blueprints according to correspondence with MSDE. After matching the test blueprint, CR items with omit rates as high as 23.32% were used.

2. Research conducted preliminary IRT calibration for all items except for items that were flagged for poor quality. Items that had poor fit were also excluded from item selection.
3. Research provided TD with augmented form planners with IA, DIF, and item fit flags.
4. Form K was thought to be the best form for standard setting. TD first selected operational items for the standard setting form – Form K.
5. TD used the test blueprint and reporting categories to select the operational items. One BCR item in form G was mislabeled as an ECR item so this item was not scored. TD was able to replace this BCR item with 2 SR items.
6. Research reviewed average item difficulties by form and suggested changes to item selection when necessary.

Table 6.1 shows the frequency distribution of item difficulty (P value) and item-total correlation (R_ITT) by SR and CR items for the proposed target form – Form K. Tables 6.2 to 6.10 show the same frequency distributions for the other forms. Table 6.11 presents the mean and standard deviation of the item difficulty and item-total correlation for each proposed operational form. Table 6.12 presents the number of items per subscore by form. Table 6.13 presents the number of items excluded from item selection by reason.

Table 6.1. Form K (Target)

	SR items		CR items	
P value/ R_ITT value interval	P value	R_ITT	P value	R_ITT
0.1-0.3	0	0	0	0
0.3-0.4	1	11	0	0
0.4-0.5	5	21	2	0
0.5-0.6	4	13	1	1
0.6-0.7	15	1	1	1
0.7-0.8	17	0	0	2
0.8-0.9	4	0	0	0
0.9-1	0	0	0	0

Table 6.2. Form E

	SR items		CR items	
P value/ R_ITT value interval	P value	R_ITT	P value	R_ITT
0.1-0.3	0	2	0	0
0.3-0.4	4	7	0	0
0.4-0.5	7	19	3	0
0.5-0.6	5	15	1	1
0.6-0.7	17	3	0	3
0.7-0.8	12	0	0	0
0.8-0.9	1	0	0	0
0.9-1	0	0	0	0

Table 6.3. Form F

	SR items		CR items	
P value/ R_ITT value interval	P value	R_ITT	P value	R_ITT
0.1-0.3	1	3	0	0
0.3-0.4	0	9	2	0
0.4-0.5	3	19	0	0
0.5-0.6	11	15	1	1
0.6-0.7	18	0	1	3
0.7-0.8	12	0	0	0
0.8-0.9	1	0	0	0
0.9-1	0	0	0	0

Table 6.4. Form G

	SR items		CR items	
P value/ R_ITT value interval	P value	R_ITT	P value	R_ITT
0.1-0.3	0	2	0	0
0.3-0.4	2	12	0	0
0.4-0.5	6	25	2	0
0.5-0.6	9	9	1	1
0.6-0.7	20	1	0	1
0.7-0.8	10	0	0	1
0.8-0.9	2	0	0	0
0.9-1	0	0	0	0

Table 6.5. Form H

	SR items		CR items	
P value/ R_ITT value interval	P value	R_ITT	P value	R_ITT
0.1-0.3	0	4	0	0
0.3-0.4	2	10	0	0
0.4-0.5	7	22	2	0
0.5-0.6	5	10	1	1
0.6-0.7	12	0	1	3
0.7-0.8	14	0	0	0
0.8-0.9	6	0	0	0
0.9-1	0	0	0	0

Table 6.6. Form J

	SR items		CR items	
P value/ R_ITT value interval	P value	R_ITT	P value	R_ITT
0.1-0.3	0	1	0	0
0.3-0.4	2	12	1	0
0.4-0.5	6	18	1	0
0.5-0.6	6	15	2	0
0.6-0.7	13	0	0	3
0.7-0.8	15	0	0	1
0.8-0.9	4	0	0	0
0.9-1	0	0	0	0

Table 6.7. Form L

	SR items		CR items	
P value/ R_ITT value interval	P value	R_ITT	P value	R_ITT
0.1-0.3	0	1	0	0
0.3-0.4	1	9	0	0
0.4-0.5	5	24	3	0
0.5-0.6	12	10	1	1
0.6-0.7	15	2	0	1
0.7-0.8	12	0	0	2
0.8-0.9	1	0	0	0
0.9-1	0	0	0	0

Table 6.8. Form M

	SR items		CR items	
P value/ R_ITT value interval	P value	R_ITT	P value	R_ITT
0.1-0.3	0	1	0	0
0.3-0.4	0	6	0	0
0.4-0.5	2	17	2	0
0.5-0.6	10	22	1	1
0.6-0.7	13	0	1	2
0.7-0.8	19	0	0	1
0.8-0.9	2	0	0	0
0.9-1	0	0	0	0

Table 6.9. Form N

	SR items		CR items	
P value/ R_ITT value interval	P value	R_ITT	P value	R_ITT
0.1-0.3	0	0	0	0
0.3-0.4	0	10	1	0
0.4-0.5	5	26	1	0
0.5-0.6	11	10	1	1
0.6-0.7	16	0	1	3
0.7-0.8	12	0	0	0
0.8-0.9	2	0	0	0
0.9-1	0	0	0	0

Table 6.10. Form P

	SR items		CR items	
P value/ R_ITT value interval	P value	R_ITT	P value	R_ITT
0.1-0.3	0	1	1	0
0.3-0.4	1	6	1	0
0.4-0.5	4	21	0	0
0.5-0.6	8	18	2	1
0.6-0.7	18	0	0	1
0.7-0.8	14	0	0	2
0.8-0.9	1	0	0	0
0.9-1	0	0	0	0

Table 6.11. Classical Item Statistics Summary by Form

Form code	K (Target)	E	F	G	H	J	L	M	N	P
P value										
mean	0.67	0.61	0.63	0.62	0.65	0.64	0.62	0.66	0.63	0.64
SD	0.12	0.14	0.12	0.11	0.13	0.14	0.11	0.11	0.11	0.12
R_ITT										
mean	0.48	0.48	0.47	0.45	0.45	0.47	0.48	0.49	0.46	0.48
SD	0.09	0.10	0.10	0.10	0.10	0.10	0.09	0.09	0.08	0.09

Table 6.12. Number of Items per Subscore Category by Form

Form code	K (Target)	E	F	G	H	J	L	M	N	P
Subscore 1	14	14	14	14	14	14	14	14	14	14
Subscore 2	12	12	12	14	12	12	12	12	12	12
Subscore 3	10	10	10	10	10	10	10	10	10	10
Subscore 4	14	14	14	14	14	14	14	14	14	14

Table 6.13. Number of Items Excluded from Selection by Reason

Analysis	Number of Items Flagged
Poor Content	4
High Omit Rate (SR)	8
High Omit Rate (CR)	4
DIF	16
Poor Fit	3

Calibration and Scaling

Items identified as operational items by both ETS and MSDE were calibrated using the Three Parameter Logistic (3PL) model for SR items and Generalized Partial Credit Model (GPCM) for CR items. There were 11 linking items shared by all 10 forms and additional linking items shared by adjacent forms. A concurrent calibration allowed us to put all item parameters on the same scale. The concurrent calibration converged successfully and item parameter estimates were obtained. Item fit statistics were examined and no item displayed poor fit. The maximum likelihood ability estimates (MLE) were obtained for all students in the calibration. For students with all correct or all incorrect responses, ability estimates were set to 4 and -4, respectively, on theta-scale. The mean and standard deviation of ability estimates were calculated and a set of transformation constants were derived such that the mean scale score was approximately[4] 400 and the standard deviation was 40. This set of transformation constants was applied to the item parameter estimates of the operational items in order to place the operational item parameters on the reporting scale.

A second calibration was conducted to include all items (both operational and field test items) accepted from the MSDE review. Two items on Form P were considered to have poor fit. MSDE approved the removal of the two misfit items from calibration so a third calibration was conducted removing the two items. In a Stocking-Lord linking procedure, the operational items were used as linking items to bring the field test items on to the reporting scale.

Test Characteristic Curves and the Conditional Standard Error of Measurement (CSEM) plots were used to evaluate the extent to which the test forms were parallel. The ten forms appeared to be close to parallel forms. For example, the raw scores associated with a scale score of 410 for target Form K is 41.2 and for the most difficult form, Form N, it is 37.7. This translated to about 6% difference between the easiest and hardest forms. The CSEMs were minimized around scale scores of 350 to 440.

[4] Because of the boundary constraints of the MLE theta estimates (4 and -4), the actual scale score mean and standard deviation are not exactly 400 and 40.

Figure 6.1. Test Characteristic Curves for English Forms

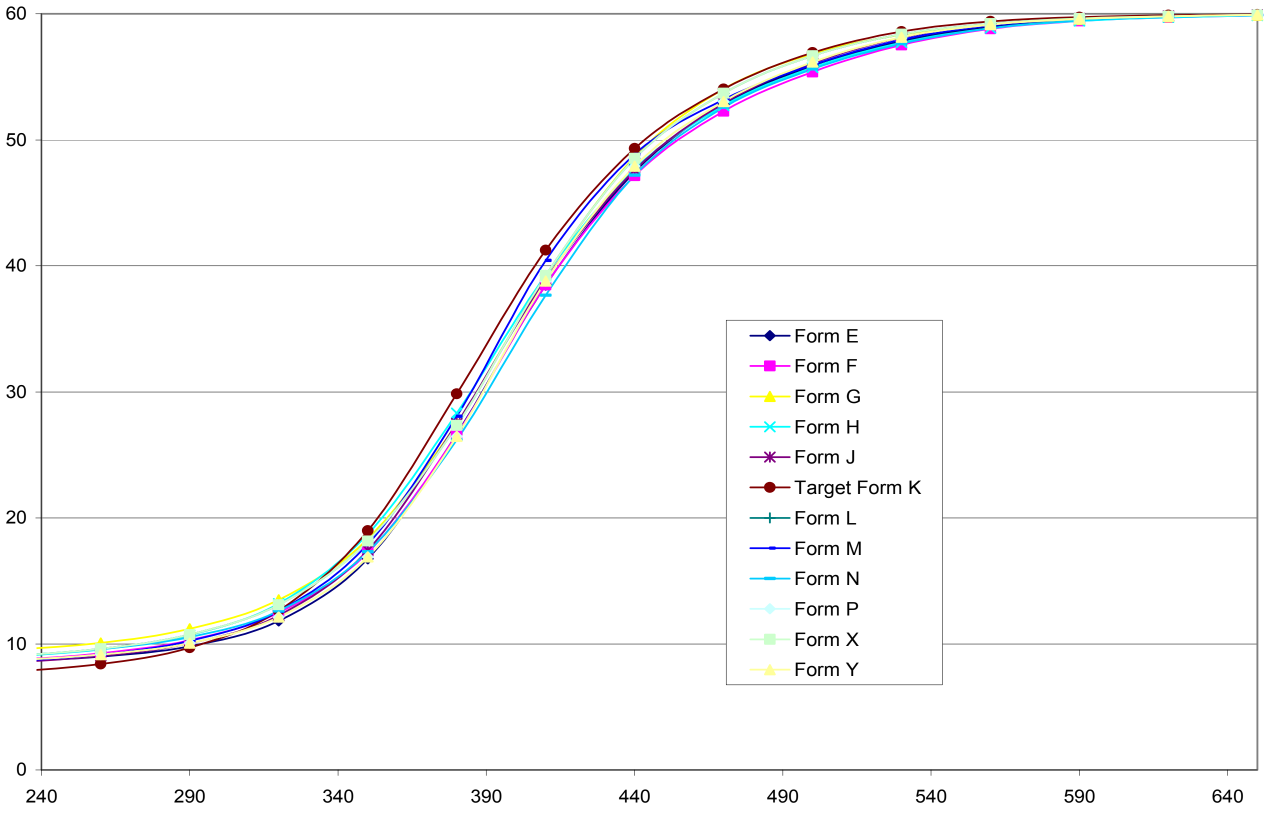
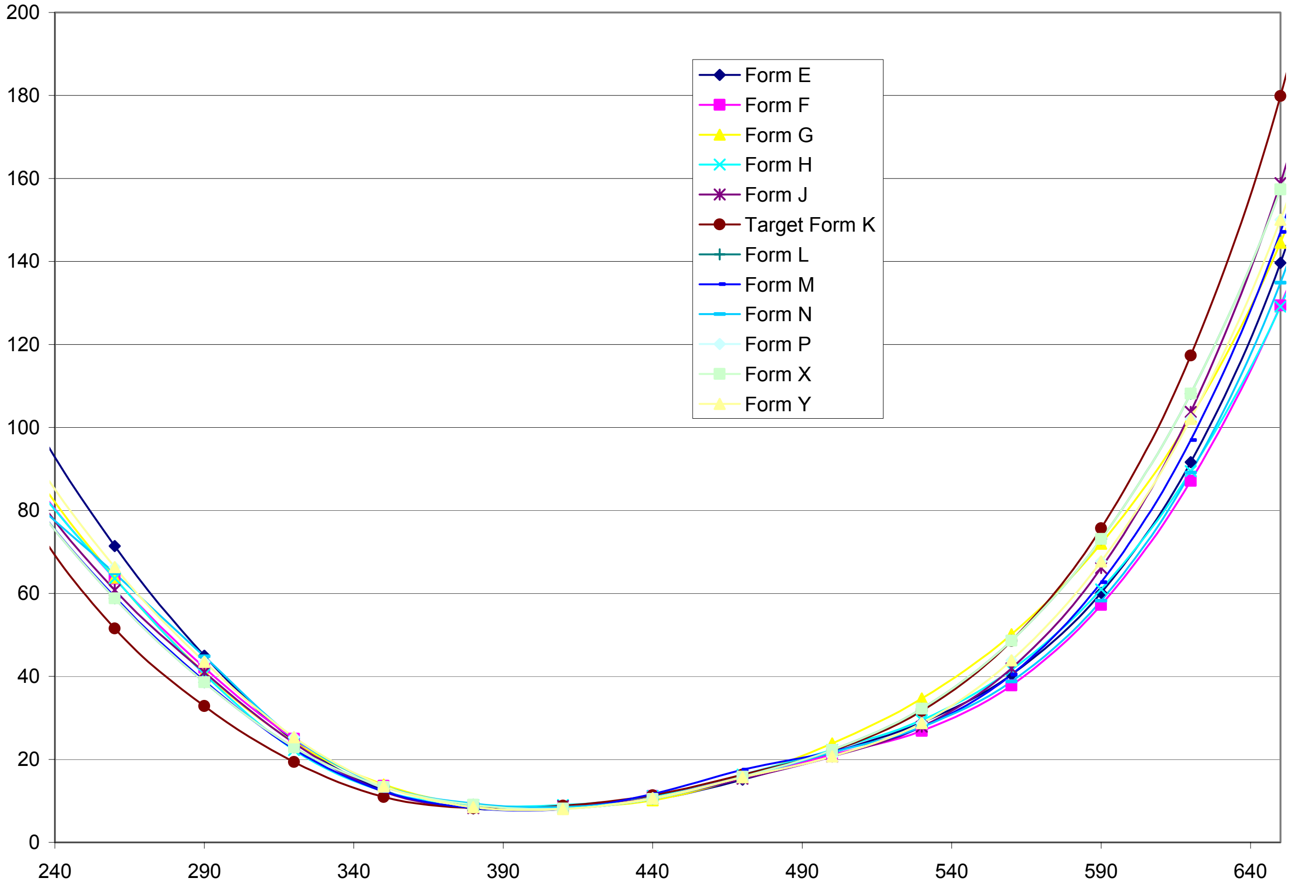


Figure 6.2. Conditional Standard Error of Measurement for English Forms



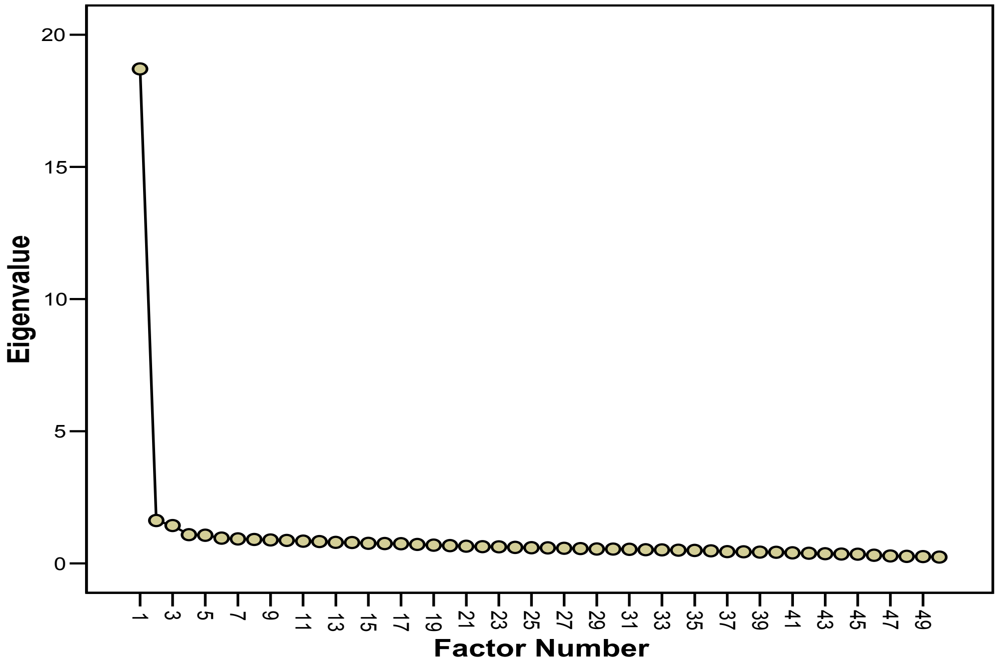
Factor Analysis Results

Factor analysis techniques were employed to investigate the dimensionality of the English MHSA primary forms. All students writing a particular form were used for the analyses. Given the ordinal nature of the item scores, matrices consisting of tetrachoric and polychoric correlations were produced for each form using PRELIS (Joreskog & Sorbom, 1993) and then analyzed using SPSS. The scree plots presented and discussed with respect to the eigenvalues and percentage of variation accounted for.

English Form E

The results of the factor analysis for Form E show an initial eigenvalue of 18.71 for the first factor, accounting for 37.42% of the variance. There were four other eigenvalues greater than one, ranging from 1.62, accounting for 3.25 % of the variance, to 1.07, accounting for 2.14 % of the variance. The scree plot for Form E illustrates one dominant factor.

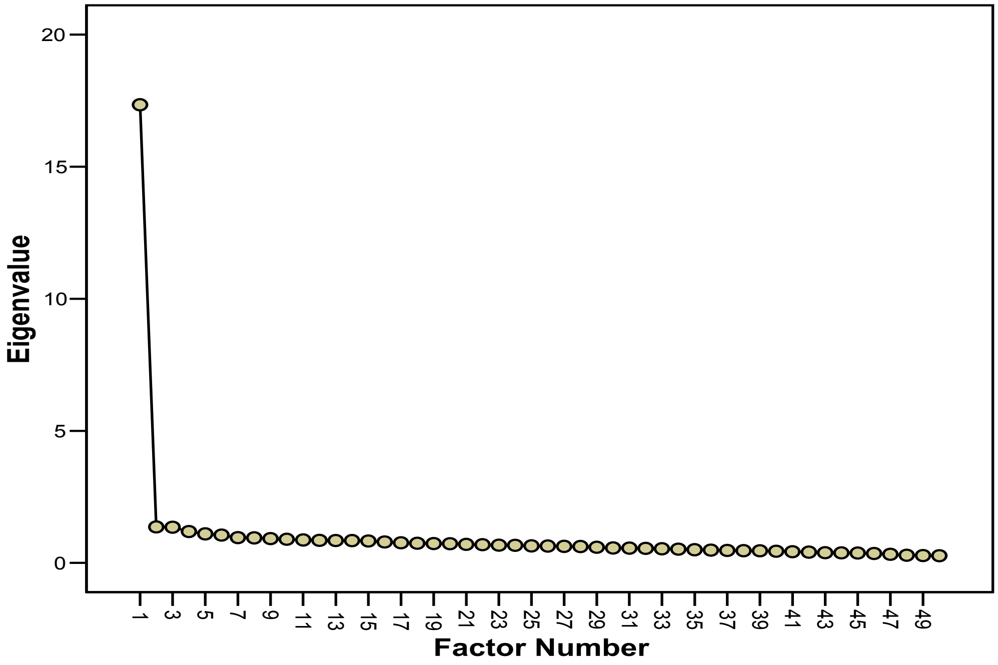
Figure 6.3 Form E Scree Plot



English Form F

The results of the factor analysis for Form F show an initial eigenvalue of 17.35, which accounts for 34.69% of the variance. There were six eigenvalues greater than one, although the remaining five eigenvalues were only slightly more than one, and accounted for less than 3% of the variance each. The scree plot for this factor analysis is provided below, indicating the presence of one dominant factor.

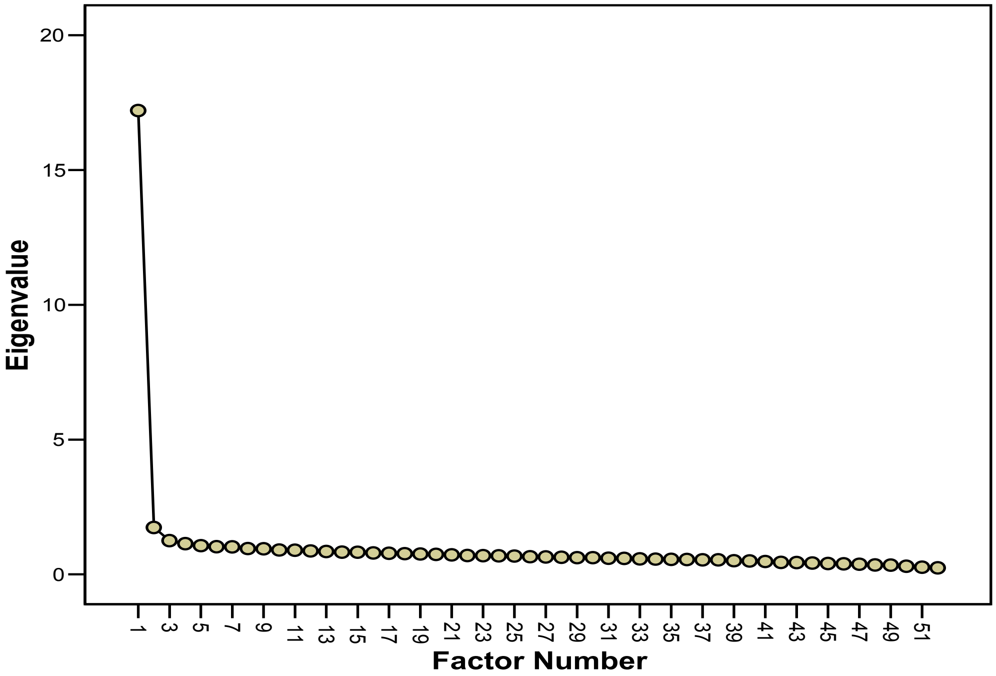
Figure 6.4. Form F Scree Plot



English Form G

The factor analysis results for Form G indicate an initial eigenvalue of 17.21 for the first factor, accounting for 33.09% of the variance. There were six other eigenvalues greater than one, ranging from 1.74 (3.34% of variance) to 1.02 (1.96% of variance). The scree plot for Form G indicates the presence of one dominant factor.

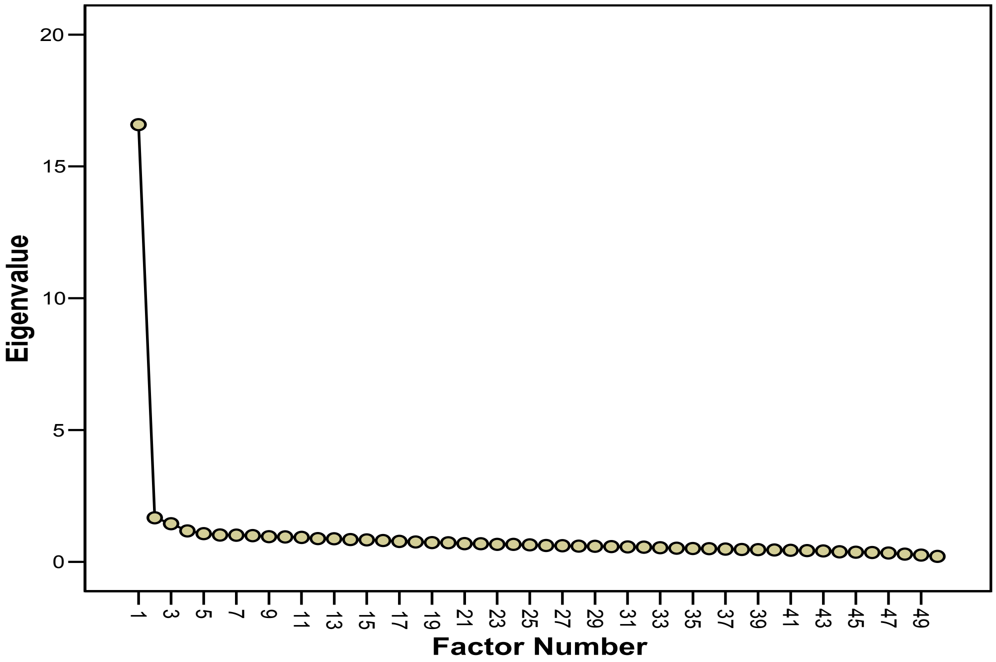
Figure 6.5. Form G Scree Plot



English Form H

The factor analysis results for Form H indicate an eigenvalue of 16.58 for the first factor, accounting for 33.17% of the variance. The remaining eigenvalues were less than two and accounted for less than 3.5% of the variance. The scree plot for Form H indicates the presence of one dominant factor.

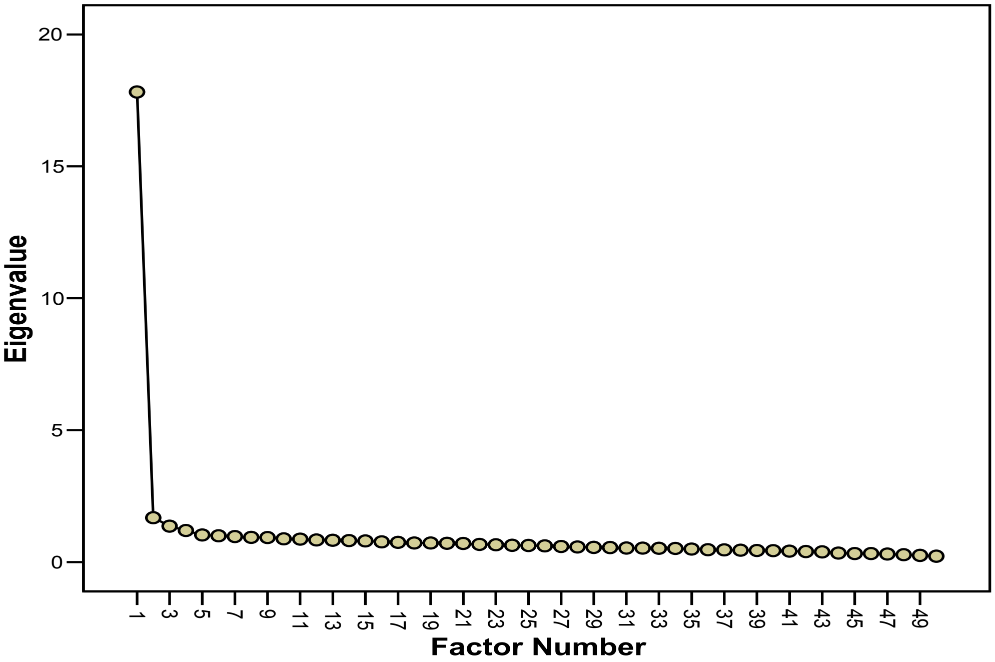
Figure 6.6. Form H Scree Plot



English Form J

The factor analysis results for Form J reveal an initial eigenvalue of 17.82 for the first factor, accounting for 35.64% of the variance. The remaining 4 eigenvalues with values greater than 1 ranged from 1.69, accounting for 3.37% of the variance, to 1.03, accounting for 2.06% of the variance. The scree plot for Form J illustrates one dominant factor.

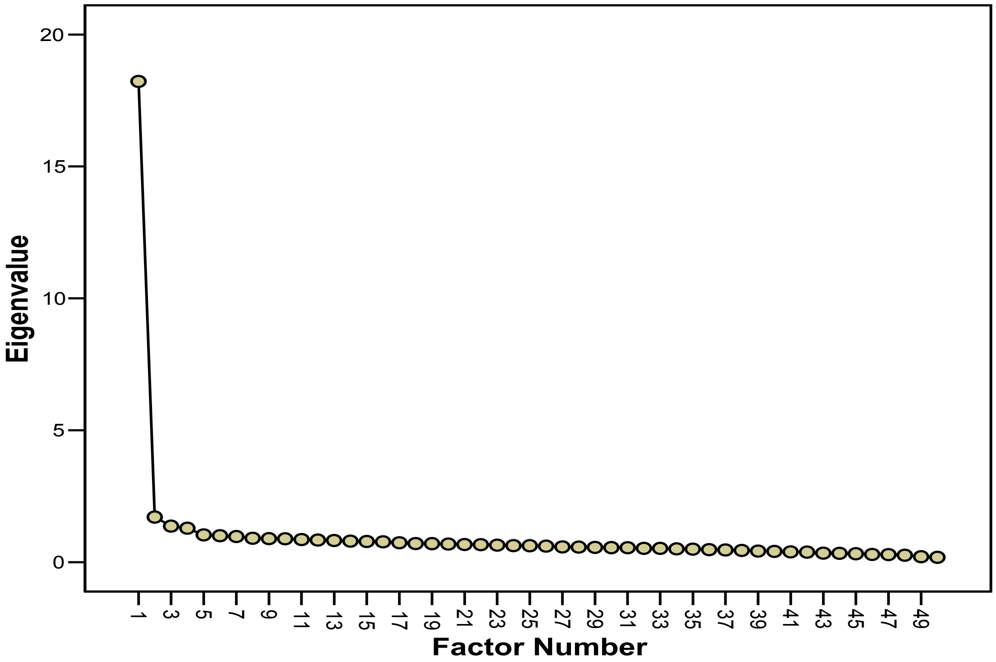
Figure 6.7. Form J Scree Plot



English Form K

The results of the factor analysis for Form K indicated an initial eigenvalue of 18.22 for the first factor, accounting for 36.45% of the variance. There were six eigenvalues greater than or equal to 1. The remaining 5 eigenvalues had values ranging from 1.71, accounting for 3.43% of the variance, to 1.00, accounting for 2.01% of the variance. The scree plot for Form K illustrates the dominance of one factor.

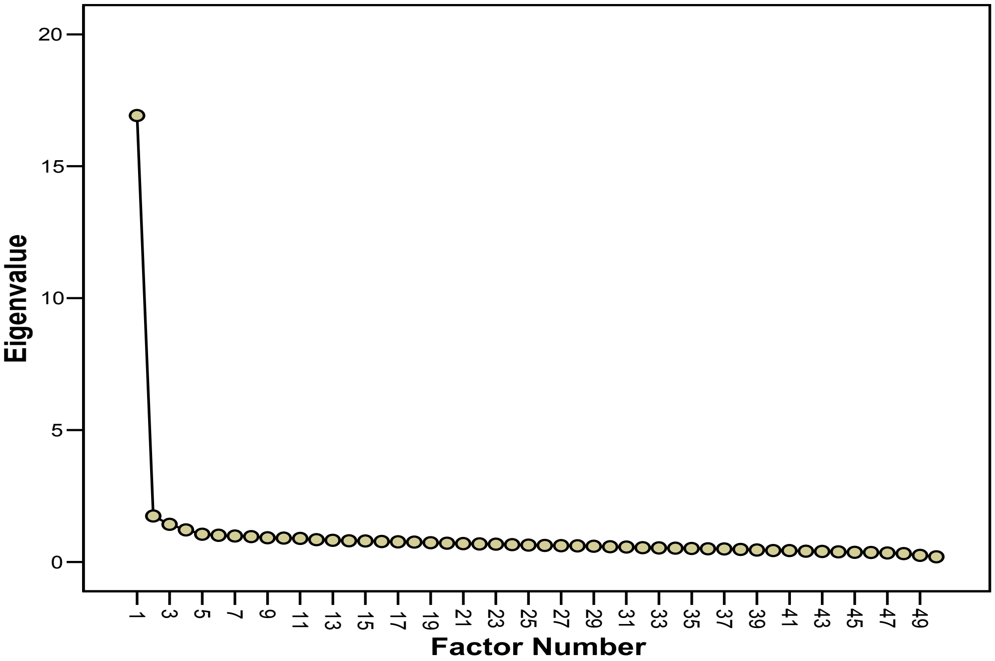
Figure 6.8. Form K Scree Plot



English Form L

The factor analysis results for Form L reveal an initial eigenvalue of 16.92, which accounts for 33.84% of the variance. There were six eigenvalues greater than one, although the remaining eigenvalues were less than 2, with variances ranging from 3.5 to 2%. The scree plot for this factor analysis is provided below, indicating the presence of one dominant factor.

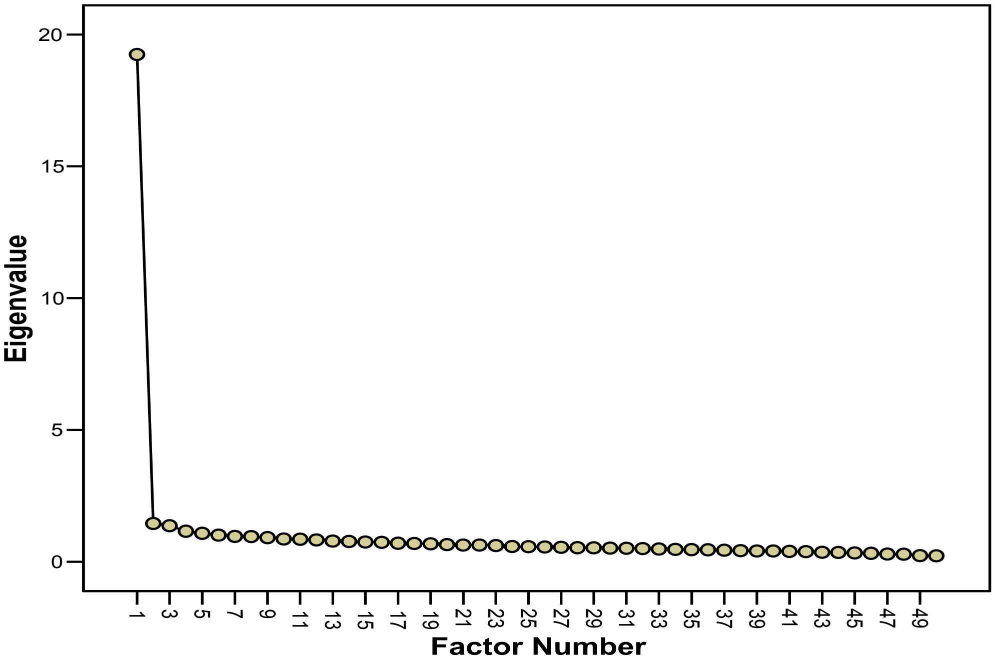
Figure 6.9. Form L Scree Plot



English Form M

The factor analysis results for Form M indicate an initial eigenvalue of 19.24, which accounts for 38.49% of the variance. Of the remaining 5 eigenvalues greater than 1, all were less than 1.5 and accounted for less than 3% of the variance. The scree plot for this factor analysis is provided below, indicating that one dominant factor is present.

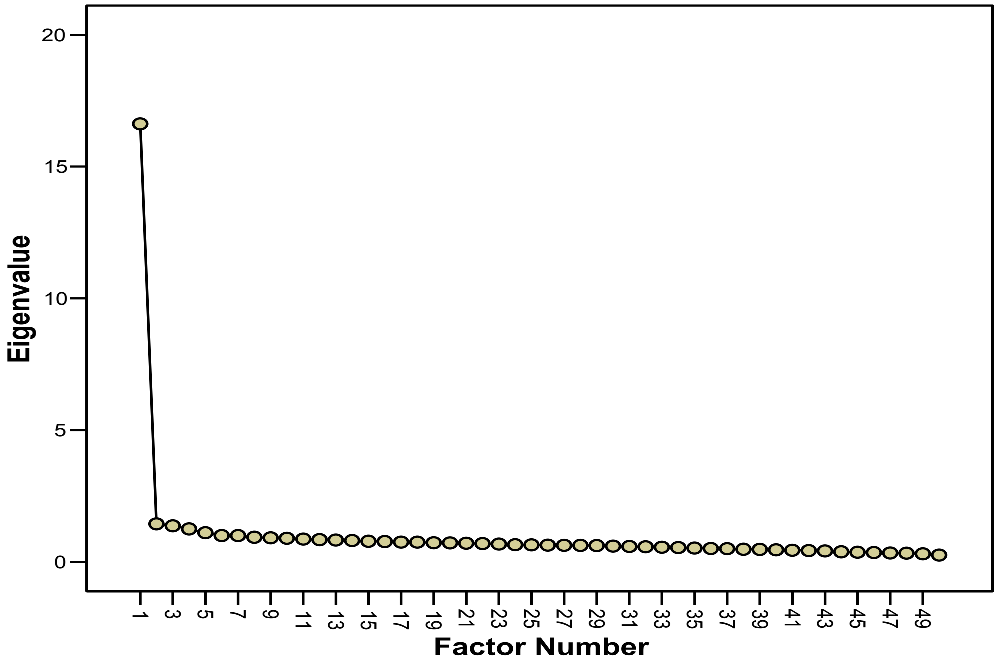
Figure 6.10. Form M Scree Plot



English Form N

The results of factor analysis for Form N shows an initial eigenvalue of 16.62, which accounts for 33.25% of the variance. There were 7 eigenvalues with values greater than 1. Of the remaining 6 eigenvalues, all were less than 1.5 and accounted for between 2 and 3% of the variance. The scree plot, provided below, demonstrates the presence of one dominant factor.

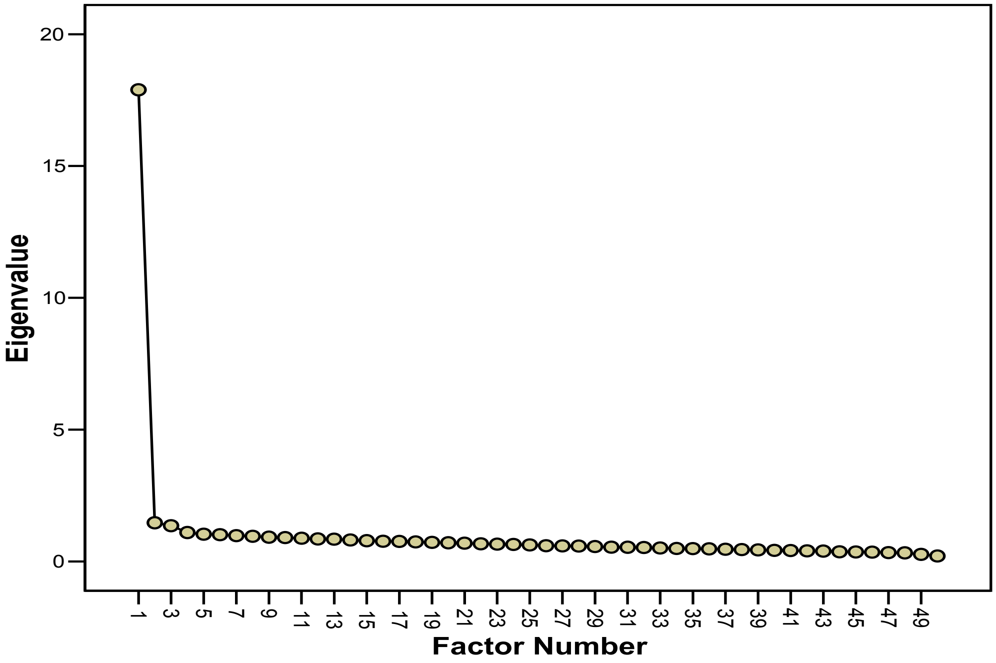
Figure 6.11. Form N Scree Plot



English Form P

The factor analysis results for Form P show an initial eigenvalue of 17.89, which accounts for 35.78% of the variance. Of the six eigenvalues that were greater than 1, the remaining five were less than 1.5 and accounted for between 2 and 3% of the variance. The scree plot for this factor analysis, provided below, illustrates one dominant factor.

Figure 6.12. Form P Scree Plot



Conclusion

The factor analyses results of the 10 primary forms indicate that one dominant factor underlies the MHSA English exams. In all cases, the first factor accounted for one-third or more of the total variance. The remaining factors accounted for considerably smaller percentage of the variance.

Summary Statistics of Student Achievement

This section summarizes the test-level statistics obtained for the English 2005 administration of the MHSA. The test-level analyses include demographic distributions, scale score information, and reliability analyses. The demographic characteristics of the students are presented in Table 6.14, whereas the scale score statistics and reliability analyses are presented in Table 6.15 for the primary forms and Table 6.16 for the make-up forms.

Table 6.14 Demographic Information for the English Exam

		May Primary Forms		May Make-Up Forms	
		N	%	N	%
Overall		54643		3250	
Gender					
	Male	27000	49.4	1771	54.5
	Female	27642	50.6	1478	45.5
	Missing	1	0.0	1	0.0
Special Education					
	Yes	5251	9.6	425	13.1
	No	48492	88.7	2765	85.1
	504	900	1.6	60	1.8
Ethnicity					
	American Indian	191	0.3	10	0.3
	Asian/Pacific Islander	3118	5.7	106	3.3
	African American	20546	37.6	1526	47.0
	White	27659	50.6	1396	43.0
	Hispanic	3128	5.7	211	6.5
	Missing	1	0.0	1	0.0
Limited English Proficient					
	Yes	920	1.7	61	1.9
	No	53256	97.5	3146	96.8
	Exited	467	0.9	43	1.3

Table 6.15. Summary Statistics for English Primary Forms

		May			
		Mean	SD	N	Alpha[a]
Overall		401.07	40.38	54643	0.93
Gender					
	Male	393.16	42.60	27000	
	Female	408.79	36.47	27642	
	Missing	*	*	1	
Special Education					
	Yes	359.42	40.47	5251	
	No	405.68	37.72	48492	
	504	395.49	38.61	900	
Ethnicity					
	American Indian	393.25	38.51	191	
	Asian/Pacific Islander	419.15	38.86	3118	
	African American	384.24	36.70	20546	
	White	412.80	38.58	27659	
	Hispanic	390.32	36.89	3128	
	Missing	*	*	1	
Limited English Proficient					
	Yes	369.25	31.02	920	
	No	401.73	40.37	53256	
	Exited	388.19	29.28	467	

* Statistics not reported for sample size less than 50 (N<50)

[a] alpha values ranged from 0.9239 to 0.9392 (M = 0.9300) across the 10 primary forms

Table 6.16. Summary Statistics for English Make-Up Forms

		May Make-Up Forms							
		X				Y			
		Mean	SD	N	Alpha	Mean	SD	N	Alpha
Overall		368.19	47.74	2782	0.93	366.57	42.10	468	0.92
Gender									
	Male	357.30	49.33	1501		355.85	45.11	270	
	Female	380.98	42.40	1280		381.18	32.41	198	
	Missing	*	*	1				0	
Special Education									
	Yes	335.09	46.32	352		338.47	42.39	73	
	No	372.88	46.03	2379		371.80	39.99	386	
	504	377.98	45.37	51		*	*	9	
Ethnicity									
	American Indian	*	*	9		*	*	1	
	Asian/Pacific Islander	379.94	46.49	90		*	*	16	
	African American	356.61	43.86	1314		360.92	40.80	212	
	White	380.29	49.71	1183		371.99	42.31	213	
	Hispanic	366.89	42.12	185		*	*	26	
	Missing	*	*	1				0	
Limited English									
Proficient	Yes	*	*	49		*	*	12	
	No	368.47	48.15	2694		367.00	42.29	452	
	Exited	*	*	39		*	*	4	

* Statistics not reported for sample size less than 50 (N<50)

Table 6.17 indicates the percent of students achieving the basic, proficient, and advanced levels. Results indicated that 56.3 percent of students achieved proficiency on the exam.

Table 6.17. Percent of Students by Classification

	2005
Basic	42.7
Proficient	34.7
Advanced	22.6

Decision Accuracy and Consistency

The accuracy of decisions based on specified cut-scores was assessed for Reliability of Classification using the computer program RelClass, ETS proprietary software. RelClass provides two statistics that describe the reliability of classifications based on test scores (Livingston & Lewis, 1995). More specifically, information from an administration of one form is used to estimate the following:

3) Decision Accuracy describes the extent to which examinees are classified in the same way as they would be on the basis of the average of all possible forms of a test. Decision accuracy answers the question: How does the actual classification of test takers, based on their single-form scores, agree with the classification that would be made on the basis of their true scores, if their true scores were somehow known.

4) Decision Consistency describes the extent to which examinees are classified in the same way as they would be on the basis of a single form of a test other than the one for which data are available. Decision consistency answers the question: What is the agreement between the classifications based on two non-overlapping, equally difficult forms of the test.

Table 6.18 provides the results for the decision classification of the proficient and advanced cut-scores. The decision accuracy indices were both greater than 0.90, indicating high agreement between classifications based on the observable variable (scores on one form of a test) and classifications based on an unobservable variable (the test takers' true scores). The decision consistency indices approached 0.90, which also indicate a high agreement between classifications based on two variables (scores on the form students have taken and score from a parallel form of the same test that is not administered to the students).

Table 6.18. Decision Statistics for the English Exam

	Decision Accuracy		Decision Consistency	
	Proficient	Advanced	Proficient	Advanced
English	0.914	0.920	0.884	0.886