CHAPTER 4
TEST DEVELOPMENT PROCESS

CONTENT ADVISORY COMMITTEES

The linchpins of the test development process are the Content Advisory Committee (CAC). These committees assure that the tests and scoring guides are appropriate for Kentucky students. Advisory committees in each area—Reading, Mathematics, Science, Social Studies, Writing, Arts & Humanities, and Practical Living/Vocational Studies—guide test development for open-response items, multiple-choice items, and the Writing and Alternate Portfolios. Members of the committees include Kentucky classroom teachers at each pertinent grade level, resource teachers, administrators, curriculum coordinators, university professors, and Kentucky Education Association representatives.

THE TEST DEVELOPMENT SEQUENCE

Test development is an iterative process involving repeated reviews, revisions, and analyses. Throughout this process there is considerable interaction among Content Advisory Committees, consultants and staff from the Kentucky Department of Education, and staff from the contractors. Table 4–1 presents the major steps in the test development process, and is followed by more in-depth descriptions of each step.

TABLE 4–1 MAJOR STEPS IN THE TEST DEVELOPMENT PROCESS	
1.	Analyze score data from previous year's test
2.	Select next year's common items from previous year's matrix items
3.	Select equating items from previous year's matrix items
4.	Develop concepts for new items
5.	Write initial draft of new items
6.	Review and revise initial item drafts
7.	Review and select items for field testing/pretesting by Content Advisory Committees
8.	Assemble field test booklets
9.	Review field test booklets
10.	Administer field tests/pretest
11.	Analyze field test/pretest results
12.	Select new items for upcoming year's test forms
13.	Review of selected items by Bias Review Committee
14.	Review of selected items by the contractor and KDE staff
15.	Final item selection by the contractor and KDE staff
16.	Revise selected items and assemble test forms
17.	Review of each assembled form for content and sensitivity balance by contractor staff
18.	Review of each assembled form for content and sensitivity balance by KDE staff
19.	Review of final typeset copy for potential typographical errors

The numbering of the following sections follows that of the table above.

1. **ANALYZE SCORE DATA FROM PREVIOUS YEAR'S TEST**. The last step of one year's test development process is the same as the first step of the next year's process. Item analysis provides information about how well each item performed. This information is used to select the previous years' items that will appear in the following year's test and for making modifications to previous years' items. Table 4–2 contains the primary information that could be produced for open-response items, and tables 4–3a through 4-3f show a sample of the item analysis output for open-response items. Summaries of item analyses are presented in Chapter 6. Members of Content Advisory Committees receive training on the interpretation of score data. The training ensures that each content area committee and individual member understands how information from the analyses is used to select or revise items.

TABLE 4–2 ANALYSES PERFORMED ON OPEN-RESPONSE ITEMS	
1. Analysis	Comments
2. Common item difficulty	Distribution of scores and average item score for the total group of examinees (on 0–4 scale).
3. Form item difficulty	Distribution of scores and average item score for common and matrix items based on group of examinees taking a particular form of the test (on 0–4 scale).
4. Item-total correlation	Correlation between item scores and scores on set of common items. For common items, this correlation is corrected for part-total contamination.
5. D_{N-A} biserial correlation	Discrimination between Novice and Apprentice. Biserial correlation between item score and student performance level, with performance level dichotomized as either Novice or above Novice.
6. D_{A-P} biserial correlation	Discrimination between Apprentice and Proficient. Biserial correlation between item score and student performance level, with performance level dichotomized as either Apprentice or below or Proficient or above.
7. D_{P-D} biserial correlation	Discrimination between Proficient and Distinguished. Biserial correlation between item score and student performance level, with performance level dichotomized as either Proficient or below or Distinguished.
8. DIF-SMD male-female	Dorans and Schmitt (1991)[1] standardized mean difference (DIF/SMD) for differential item functioning, comparing groups of male and female students matched on total common-item scores. A positive DIF-SMD score indicates, relative to other items, that an item favors males. A negative DIF-SMD indicates, relative to other items, that an item favors females.
9. DIF-SMD White-African American	Standardized mean difference for differential item functioning, comparing groups of White and African American students matched on total common item scores. A positive DIF-SMD score indicates, relative to other items, that an item favors White students. A negative DIF-SMD indicates, relative to other items, that an item favors African American students.
10. Coefficient Alpha	Internal consistency statistic calculated for common items and for each sub-form calculated for total set of items within the appropriate content category.

[1] Dorans, N., & Schmitt, A.P. (1991). *Constructed-response and differential item functioning: A pragmatic approach (ETS research report No. 91–47).* Princeton, NJ: Educational Testing Service.

TABLE 4–3a
GRADE 11 OPEN-RESPONSE ITEM ANALYSIS FOR 1998 READING[1]

				Percentage Distribution					
Form	N	Mean	Std. Dev	B[2]	0	1	2	3	4
Overall	40,269	2.27	0.75	1.27	2.21	18.19	36.70	30.67	10.96
1	3422	2.28	0.75	1.62	2.27	16.00	38.00	32.32	9.80
2	3377	2.26	0.75	1.39	1.79	19.47	36.31	30.19	10.84
3	3348	2.26	0.75	1.25	2.17	18.39	38.05	28.58	11.56
4	3358	2.33	0.73	1.22	1.73	16.63	35.91	33.55	19.96
5	3345	2.22	0.77	1.36	2.56	21.06	34.99	28.99	11.05
6	3348	2.20	0.72	1.19	3.34	18.16	39.19	29.48	8.64
7	3353	2.25	0.75	0.99	2.02	21.13	35.29	29.29	11.28
8	3334	2.20	0.78	1.37	3.25	21.44	35.02	27.57	11.34
9	3357	2.31	0.74	0.97	1.70	17.08	37.58	32.31	10.36
10	3338	2.28	0.75	1.42	2.18	18.66	36.31	29.29	12.13
11	3341	2.41	0.76	1.33	1.67	14.09	35.36	34.32	13.23
12	3378	2.31	0.73	1.15	1.89	16.16	38.34	32.13	10.33

[1]Data in this table are for common and matrix items combined.
[2]B = No answer, 0 = Answer is totally incorrect or irrelevant, 1 = Answer shows minimal understanding or addresses only a small portion of the question, 2 = Completes some important components of the question, communicates clearly, but demonstrates gaps in conceptual understanding, 3 = Completes most components and communicates clearly, understands major concepts, but may misunderstand some less important ideas or details, and 4 = Completes all important components, communicates clearly. Demonstrates in-depth understanding of concepts and processes, where appropriate, chooses most efficient processes, offers insightful interpretations, generalizations, applications or strategies.

TABLE 4–3b
GRADE 11 OPEN-RESPONSE ITEM ANALYSIS FOR 1998 MATHEMATICS[1]

				Percentage Distribution					
Form	N	Mean	Std. Dev	B[2]	0	1	2	3	4
Overall	40,274	1.22	0.85	3.30	32.48	26.68	21.38	12.23	3.93
1	3417	1.33	0.89	3.39	30.04	25.70	20.93	14.03	5.91
2	3378	1.17	0.86	4.15	31.90	29.61	19.97	10.01	4.36
3	3349	1.30	0.89	2.88	30.54	26.51	22.42	11.68	5.97
4	3353	1.16	0.85	3.71	34.99	24.69	21.97	10.90	3.73
5	3348	1.19	0.81	3.18	31.77	29.32	20.77	11.81	3.16
6	3355	1.15	0.86	3.29	35.40	26.45	19.47	11.62	3.76
7	3354	1.19	0.79	2.85	31.48	28.46	23.52	11.32	2.37
8	3336	1.36	0.90	3.15	30.42	24.61	20.64	14.82	6.36
9	3355	1.16	0.87	3.23	36.24	23.85	21.04	12.79	2.85
10	3341	1.26	0.81	3.33	31.47	25.34	21.78	15.04	3.04
11	3342	1.16	0.79	2.93	33.31	28.36	22.09	10.68	2.82
12	3346	1.19	0.84	3.44	32.19	27.49	22.00	12.08	2.80

[1]Data in this table are for common and matrix items combined.
[2]B = No answer, 0 = Answer is totally incorrect or irrelevant, 1 = Answer shows minimal understanding or addresses only a small portion of the question, 2 = Completes some important components of the question, communicates clearly, but demonstrates gaps in conceptual understanding, 3 = Completes most components and communicates clearly, understands major concepts, but may misunderstand some less important ideas or details, and 4 = Completes all important components, communicates clearly. Demonstrates in-depth understanding of concepts and processes, where appropriate, chooses most efficient processes, offers insightful interpretations, generalizations, applications or strategies.

TABLE 4–3c
GRADE 11 OPEN-RESPONSE ITEM ANALYSIS FOR 1998 SCIENCE[1]

				Percentage Distribution					
Form	N	Mean	Std. Dev	B[2]	0	1	2	3	4
Overall	40,266	1.62	0.75	2.72	14.96	28.27	32.89	16.56	4.60
1	3417	1.61	0.77	2.53	15.21	29.02	32.40	16.16	4.68
2	3374	1.51	0.68	2.63	15.85	32.83	31.88	13.29	3.51
3	3345	1.79	0.76	2.25	12.14	25.07	33.96	20.21	6.37
4	3354	1.71	0.75	2.81	11.93	26.34	36.37	18.52	4.03
5	3348	1.51	0.71	3.56	15.26	31.67	32.42	13.51	3.59
6	3353	1.49	0.71	2.72	18.66	29.13	32.17	13.73	3.59
7	3357	1.51	0.77	3.14	20.25	26.87	30.24	14.50	5.00
8	3333	1.50	0.73	2.80	19.25	28.01	31.59	14.39	3.96
9	3356	1.71	0.74	2.21	13.68	28.49	33.49	19.50	4.64
10	3337	1.68	0.72	2.58	12.30	27.89	35.63	17.12	4.49
11	3341	1.77	0.74	2.28	12.40	25.74	33.72	19.76	6.11
12	3351	1.67	0.76	3.10	12.58	30.09	30.84	18.10	5.29

[1]Data in this table are for common and matrix items combined.
[2]B = No answer, 0 = Answer is totally incorrect or irrelevant, 1 = Answer shows minimal understanding or addresses only a small portion of the question, 2 = Completes some important components of the question, communicates clearly, but demonstrates gaps in conceptual understanding, 3 = Completes most components and communicates clearly, understands major concepts, but may misunderstand some less important ideas or details, and 4 = Completes all important components, communicates clearly. Demonstrates in-depth understanding of concepts and processes, where appropriate, chooses most efficient processes, offers insightful interpretations, generalizations, applications or strategies.

TABLE 4–3d
GRADE 11 OPEN-RESPONSE ITEM ANALYSIS FOR 1998 SOCIAL STUDIES[1]

				Percentage Distribution					
Form	N	Mean	Std. Dev	B[2]	0	1	2	3	4
Overall	40,251	1.94	0.82	3.01	7.38	21.73	36.99	25.29	5.61
1	3418	1.86	0.81	3.16	7.35	23.97	38.38	23.10	4.03
2	3367	2.04	0.82	2.89	5.21	20.01	28.17	27.36	6.36
3	3346	1.87	0.81	3.29	10.21	21.61	34.20	25.49	5.21
4	3349	1.89	0.76	2.89	7.64	23.24	37.52	23.92	4.80
5	3344	1.90	0.81	3.27	8.53	21.91	36.26	24.70	5.33
6	3349	1.95	0.82	2.72	6.61	21.65	38.49	25.35	5.17
7	3357	2.13	0.83	2.70	5.13	17.68	35.73	30.78	7.98
8	3332	1.92	0.81	3.34	8.12	21.60	36.23	25.06	5.64
9	3359	1.91	0.82	2.81	7.59	22.39	37.90	23.99	5.33
10	3341	1.94	0.84	2.96	7.62	21.50	36.71	25.44	5.78
11	3339	1.90	0.81	2.92	7.64	23.53	36.69	23.95	5.28
12	3350	1.96	0.82	3.14	6.92	21.61	37.50	24.35	6.47

[1]Data in this table are for common and matrix items combined.
[2]B = No answer, 0 = Answer is totally incorrect or irrelevant, 1 = Answer shows minimal understanding or addresses only a small portion of the question, 2 = Completes some important components of the question, communicates clearly, but demonstrates gaps in conceptual understanding, 3 = Completes most components and communicates clearly, understands major concepts, but may misunderstand some less important ideas or details, and 4 = Completes all important components, communicates clearly. Demonstrates in-depth understanding of concepts and processes, where appropriate, chooses most efficient processes, offers insightful interpretations, generalizations, applications or strategies.

TABLE 4–3e
GRADE 11 OPEN-RESPONSE ITEM ANALYSIS FOR 1998
ARTS & HUMANITIES[1]

				Percentage Distribution					
Form	N	Mean	Std. Dev	B[2]	0	1	2	3	4
Overall	40,258	1.78	0.92	2.94	8.06	29.04	36.45	17.81	5.70
1	3423	1.44	0.83	2.76	12.23	39.51	32.83	10.87	1.80
2	3368	1.67	0.92	3.27	11.14	30.94	32.55	17.02	5.09
3	3347	1.74	0.85	1.86	7.70	30.86	38.17	18.07	3.34
4	3356	1.86	0.99	3.54	9.47	21.79	38.77	18.82	7.62
5	3347	1.89	0.86	2.89	5.49	25.94	39.77	19.07	6.84
6	3352	1.54	0.88	3.15	8.90	42.72	28.12	12.54	4.57
7	3354	2.13	0.97	3.15	5.34	18.00	36.59	24.42	12.51
8	3332	1.93	0.90	3.27	4.70	23.72	41.61	19.70	6.99
9	3351	1.85	0.90	2.38	6.50	26.05	41.03	19.38	4.66
10	3341	1.95	0.93	2.96	5.55	24.60	36.32	23.55	7.01
11	3341	1.61	0.89	3.37	11.17	30.22	37.38	15.56	2.31
12	3348	1.70	0.89	2.70	8.35	33.91	34.33	14.90	5.81

[1]Data in this table are for common and matrix items combined.
[2]B = No answer, 0 = Answer is totally incorrect or irrelevant, 1 = Answer shows minimal understanding or addresses only a small portion of the question, 2 = Completes some important components of the question, communicates clearly, but demonstrates gaps in conceptual understanding, 3 = Completes most components and communicates clearly, understands major concepts, but may misunderstand some less important ideas or details, and 4 = Completes all important components, communicates clearly. Demonstrates in-depth understanding of concepts and processes, where appropriate, chooses most efficient processes, offers insightful interpretations, generalizations, applications or strategies.

TABLE 4–3f
GRADE 11 OPEN-RESPONSE ITEM ANALYSIS FOR 1998
PRACTICAL LIVING/VOCATIONAL STUDIES[1]

				Percentage Distribution					
Form	N	Mean	Std. Dev	B[2]	0	1	2	3	4
Overall	40,238	1.98	0.88	2.67	4.70	23.32	39.54	23.65	6.14
1	3412	1.96	0.88	2.62	5.17	21.75	42.06	23.15	5.25
2	3374	2.08	0.85	3.02	2.64	17.56	45.08	25.40	6.20
3	3339	2.02	0.93	1.92	5.45	17.24	39.31	27.18	8.90
4	3350	2.07	0.90	2.75	3.75	22.88	35.12	27.91	7.60
5	3347	1.87	.080	2.55	2.88	35.02	32.15	22.02	5.38
6	3351	2.05	0.85	2.34	3.54	20.38	42.51	25.66	5.57
7	3353	1.93	0.89	2.31	3.52	29.16	37.78	21.03	6.20
8	3333	1.63	0.95	3.41	12.14	33.12	29.22	17.25	4.86
9	3357	1.97	0.86	3.27	6.39	18.61	41.67	24.38	5368
10	3342	2.06	0.97	3.16	6.71	18.07	37.74	25.06	9.27
11	3338	2.00	0.79	2.10	2.32	21.67	46.52	23.88	3.51
12	3342	1.98	0.79	2.56	1.87	24.29	45.19	20.83	5.27

[1]Data in this table are for common and matrix items combined.
[2]B = No answer, 0 = Answer is totally incorrect or irrelevant, 1 = Answer shows minimal understanding or addresses only a small portion of the question, 2 = Completes some important components of the question, communicates clearly, but demonstrates gaps in conceptual understanding, 3 = Completes most components and communicates clearly, understands major concepts, but may misunderstand some less important ideas or details, and 4 = Completes all important components, communicates clearly. Demonstrates in-depth understanding of concepts and processes, where appropriate, chooses most efficient processes, offers insightful interpretations, generalizations, applications or strategies.

As we have seen in tables 4-3a though 4-3f Item analysis provides important information about how well each of the grade 11 forms performed across all students. However, additional analysis is necessary to see if items within and across these forms function differently between the different sub-populations that took the KIRIS test. This additional analysis is called Differential Item Functioning (DIF). DIF looks at the psychometric difference in how an item functions for two groups. The requirements for DIF analysis are: 1) There are two comparable groups, 2) The groups must have taken the same items, and 3) The test items are located within the same item context for both groups. It is extremely important to distinguish the difference between how an item functions between groups (DIF) and the difference between content ability levels between groups. For DIF to be an issue an item must function differently between groups. The overall ability of either group does not contribute in any DIF analysis.

DIF is an important concern because it directly goes to the validity of the test. One cannot draw truly valid conclusions about different population performance on the test as a whole if the items within the test show DIF. If DIF occurs total test scores may be measuring different attributes for these different groups.

Dorans and Schmitt use a standardized approach to see if an item exhibits DIF. In their approach it is not necessary to specify an Item Response Model. An item shows DIF "when the expected performance on an item differs for matched examinees from different groups" (Dorans and Schmitt, 1991). Using all available data, a calculation of the non-parametric item-test regressions on both the reference and focal groups are performed. Since the focal group is the focus of the analysis we have defined Women and African American as our focal groups. Since we are referring to or referencing another group when doing DIF analysis we will place Men and Whites, respectively as reference groups for each of Women and African American focal groups.

Tables 4-4a though 4-4d indicate any possible DIF based on the gender of the examinee, while Tables 4-4e though 4-4h report DIF based on the examinee's race. Standardized DIF scores can range from –1 to +1. A positive value of DIF indicates that the item favors the focal group and conversely a negative value of DIF favors the reference group. DIF values between -.05 and +.05 are considered negligible, while DIF values between -.10 and -.05 & +.10 and +.05 are to be inspected. DIF values outside -.10 and +.10 range are more unusual and should be examined very carefully. As can be seen from Tables 4-4a though 4-4h the common items for grade 11 content areas given in 1998 are DIF free. Based on these analyses there was no differential item functioning between groups for either group; gender or ethnicity.

TABLE 4–4a DIFFERENTIAL ITEM FUNCTIONING FOR 1998 STANDARDIZED MEAN DIFFERENCES – GENDER[1] GRADE 11				
Reading – Common Items Only				
Form	Item 1	Item 2	Item 3	Item 4
Overall	-0.004	0.024	-0.003	-0.017
1	-0.003	0.025	-0.010	-0.012
2	-0.010	0.021	0.000	-0.011
3	-0.009	0.028	-0.006	-0.013
4	0.001	0.022	.0.006	-0.017
5	0.001	0.030	-0.016	-0.016
6	-0.004	0.015	0.001	-0.012
7	-0.004	0.025	-0.002	-0.019
8	-0.003	0.030	-0.005	-0.022
9	0.004	0.026	0.001	-0.030
10	-0.006	0.022	0.002	-0.018
11	-0.005	0.019	0.005	-0.020
12	-0.003	0.024	-0.000	-0.021

[1] Reference group is male and the focal group is female.

TABLE 4–4b DIFFERENTIAL ITEM FUNCTIONING FOR 1998 STANDARDIZED MEAN DIFFERENCES – GENDER [1] GRADE 11				
Mathematics – Common Items Only				
Form	Item 1	Item 2	Item 3	Item 4
Overall	0.009	-0.033	0.011	0.013
1	0.008	-0.032	0.007	0.016
2	0.004	-0.028	0.008	0.016
3	0.009	-0.035	0.008	0.019
4	0.008	-0.034	0.013	0.013
5	0.003	-0.034	0.014	0.016
6	-0.006	-0.026	0.016	0.016
7	0.014	-0.035	0.012	0.009
8	0.006	-0.030	0.007	0.017
9	0.017	-0.035	0.014	0.004
10	0.006	-0.041	0.016	0.019
11	0.015	-0.023	0.007	0.002
12	0.018	-0.038	0.009	0.011

[1] Reference group is male and the focal group is female.

TABLE 4–4c DIFFERENTIAL ITEM FUNCTIONING FOR 1998 STANDARDIZED MEAN DIFFERENCES - GENDER [1] GRADE 11				
Science – Common Items Only				
Form	Item 1	Item 2	Item 3	Item 4
Overall	0.039	-0.023	-0.018	0.008
1	0.030	-0.027	-0.009	0.005
2	0.047	-0.016	-0.026	-0.002
3	0.036	-0.023	-0.030	0.018
4	0.032	-0.018	-.0.027	0.013
5	0.033	-0.019	-0.030	0.013
6	0.041	-0.024	-0.016	-0.001
7	0.035	-0.027	-0.017	0.009
8	0.027	-0.031	-0.004	0.007
9	0.025	-0.017	-0.015	0.011
10	0.034	-0.021	-0.021	0.007
11	0.036	-0.023	-0.009	-0.002
12	0.035	-0.032	-0.010	0.010

[1] Reference group is male and the focal group is female.

TABLE 4–4d DIFFERENTIAL ITEM FUNCTIONING FOR 1998 STANDARDIZED MEAN DIFFERENCES - GENDER [1] GRADE 11				
Social Studies – Common Items Only				
Form	Item 1	Item 2	Item 3	Item 4
Overall	-0.008	-0.017	0.009	0.016
1	-0.007	-0.015	0.012	0.010
2	-0.016	-0.018	0.011	0.024
3	-0.004	-0.018	0.008	0.013
4	-0.003	-0.022	.0.007	0.018
5	-0.008	-0.019	0.008	0.019
6	-0.016	-0.014	0.008	0.023
7	-0.001	-0.017	0.008	0.010
8	-0.009	-0.023	0.013	0.019
9	-0.009	-0.025	0.019	0.015
10	-0.009	-0.008	-0.000	0.017
11	-0.012	-0.014	0.010	0.016
12	-0.007	-0.014	0.008	0.014

[1] Reference group is male and the focal group is female.

TABLE 4–4e DIFFERENTIAL ITEM FUNCTIONING FOR 1998 – ETHNICITY[1] STANDARDIZED MEAN DIFFERENCES GRADE 11				
Reading – Common Items Only				
Form	Item 1	Item 2	Item 3	Item 4
Overall	0.002	0.005	-0.005	-0.002
1	0.000	0.002	0.002	-0.004
2	0.012	0.013	0.004	-0.018
3	-0.011	0.002	-0.005	0.003
4	0.004	0.002	-0.014	0.008
5	-0.006	0.009	-0.011	0.008
6	0.023	0.014	-0.012	-0.025
7	-0.002	0.006	-0.015	0.011
8	0.008	0.002	-0.010	-0.000
9	0.010	0.010	-0.012	-0.008
10	0.0092	-0.009	0.020	-0.012
11	-0.008	0.008	-0.013	0.012
12	-0.005	0.010	-0.001	-0.004

[1]Reference group is White and the focal group is African American

TABLE 4–4f DIFFERENTIAL ITEM FUNCTIONING FOR 1998 - ETHNICITY[1] STANDARDIZED MEAN DIFFERENCES GRADE 11				
Mathematics – Common Items Only				
Form	Item 1	Item 2	Item 3	Item 4
Overall	-0.003	-0.014	0.013	0.031
1	0.001	-0.007	0.001	0.004
2	-0.007	0.003	0.005	-0.000
3	-0.007	-0.005	0.009	0.003
4	-0.003	-0.013	.0.026	-0.010
5	0.005	-0.020	0.020	-0.005
6	-0.014	-0.021	0.028	0.008
7	-0.002	-0.014	-0.000	0.017
8	0.012	-0.022	0.011	-0.001
9	-0.003	-0.029	0.018	0.015
10	-0.001	0.002	-0.005	-0.006
11	0.005	-0.030	0.020	0.005
12	-0.020	-0.012	0.023	0.009

[1]Reference group is White and the focal group is African American

TABLE 4–4g DIFFERENTIAL ITEM FUNCTIONING FOR 1998 - ETHNICITY[1] STANDARDIZED MEAN DIFFERENCES GRADE 11				
Science – Common Items Only				
Form	Item 1	Item 2	Item 3	Item 4
Overall	0.010	-0.010	-0.006	0.006
1	0.023	0.000	-0.006	-0.008
2	0.036	0.001	-0.013	-0.009
3	0.011	-0.024	0.008	0.005
4	0.006	-0.024	-0.006	0.024
5	0.003	-0.005	-0.015	0.017
6	0.016	-0.011	-0.011	0.005
7	0.008	-0.017	-0.013	0.022
8	0.017	-0.023	-0.001	0.007
9	0.016	-0.017	-0.002	0.003
10	-0.001	0.007	0.005	-0.011
11	-0.006	-0.003	-0.015	0.023
12	-0.000	0.006	-0.003	-0.003

[1]Reference group is White and the focal group is African American

TABLE 4–4h DIFFERENTIAL ITEM FUNCTIONING FOR 1998 – ETHNICITY[1] STANDARDIZED MEAN DIFFERENCES GRADE 11				
Social Studies – Common Items Only				
Form	Item 1	Item 2	Item 3	Item 4
Overall	-0.007	-0.020	0.013	0.015
1	-0.008	-0.016	0.004	0.020
2	-0.017	-0.023	0.020	0.020
3	-0.015	-0.016	0.014	0.017
4	-0.010	-0.025	.0.020	0.015
5	0.000	-0.026	0.017	0.009
6	-0.006	-0.025	0.011	0.020
7	-0.005	-0.018	0.010	0.013
8	-0.006	-0.033	0.016	0.023
9	-0.010	-0.012	0.006	0.015
10	0.004	-0.007	-0.000	0.003
11	-0.008	-0.028	0.021	0.014
12	-0.002	-0.018	0.010	0.010

[1]Reference group is White and the focal group is African American

TABLE 4–5a OPEN RESPONSE ITEM ANALYSIS FOR 1998 GRADE 11					
Reading					
Form	Alpha	Median Corrected Item-Total Correlation	N vs. APD Median Biserial Correlation	NA vs. PD Median Biserial Correlation	NAP vs. D Median Biserial Correlation
Overall	.872	.676	.828	.769	.651
1	.883	.695	.830	.763	.636
2	.878	.677	.828	.791	.657
3	.874	.676	.828	.782	.665
4	.865	.670	.830	.748	.609
5	.877	.684	.828	.776	.675
6	.865	.661	.829	.748	.662
7	.868	.663	.828	.782	.661
8	.870	.676	.828	.764	.642
9	.879	.684	.828	.777	.666
10	.868	.671	.828	.760	.646
11	.882	.686	.833	.756	.622
12	.868	.660	.828	.775	.646

TABLE 4–5b OPEN RESPONSE ITEM ANALYSIS FOR 1998 GRADE 11					
Mathematics					
Form	Alpha	Median Corrected Item-Total Correlation	N vs. APD Median Biserial Correlation	NA vs. PD Median Biserial Correlation	NAP vs. D Median Biserial Correlation
Overall	.865	.672	.728	.882	.924
1	.845	.631	.744	.796	.770
2	.873	.704	.701	.914	.980
3	.865	.686	.751	.894	.858
4	.866	.676	.717	.902	.924
5	.849	.666	.701	.888	.924
6	.864	.668	.701	.885	.964
7	.847	.646	.751	.828	.803
8	.865	.678	.751	.849	.867
9	.874	.704	.728	.914	.924
10	.847	.650	.728	.843	.924
11	.849	.658	.701	.872	.924
12	.865	.678	.731	.879	.890

TABLE 4–5c OPEN RESPONSE ITEM ANALYSIS FOR 1998 GRADE 11					
Science					
Form	Alpha	Median Corrected Item-Total Correlation	N vs. APD Median Biserial Correlation	NA vs. PD Median Biserial Correlation	NAP vs. D Median Biserial Correlation
Overall	.807	.568	.701	.701	.629
1	.826	.600	.700	.705	.629
2	.776	.536	.702	.697	.648
3	.814	.575	.718	.697	.607
4	.830	.613	.769	.710	.655
5	.816	.585	.695	.739	.685
6	.795	.534	.631	.697	.669
7	.805	.566	.631	.714	.669
8	.803	.562	.631	.697	.629
9	.809	.571	.703	.705	.629
10	.797	.553	.703	.697	.614
11	.799	.551	.704	.697	.608
12	.809	.572	.676	.705	.623

TABLE 4–5d OPEN RESPONSE ITEM ANALYSIS FOR 1998 GRADE 11					
Social Studies					
Form	Alpha	Median Corrected Item-Total Correlation	N vs. APD Median Biserial Correlation	NA vs. PD Median Biserial Correlation	NAP vs. D Median Biserial Correlation
Overall	.886	.696	.828	.773	.744
1	.893	.711	.828	.791	.749
2	.890	.696	.828	.772	.720
3	.880	.690	.828	.789	.740
4	.863	.676	.828	.760	.722
5	.883	.702	.828	.779	.750
6	.898	.721	.847	.811	.756
7	.891	.705	.843	.764	.723
8	.883	.696	.828	.766	.776
9	.889	.693	.832	.774	.739
10	.894	.715	.828	.798	.752
11	.882	.688	.828	.771	.758
12	.881	.689	.828	.760	.734

2. **SELECT NEXT YEAR'S COMMON ITEMS FROM PREVIOUS YEAR'S MATRIX ITEMS.** At their first meeting, the Content Advisory Committees select and make revisions to next year's common items. Decisions are based upon both the breadth of coverage of content and Academic Expectations as well as the statistical quality of the items.

   When committee members review test questions, they sort the items by content categories defined by the content guidelines and Academic Expectations. The contractor's test developers and the Content Advisory Committees hold joint responsibility to assure that the subset of questions within a category provide breadth of coverage of the important concepts and skills in that sub domain. With materials organized this way and with approximate quotas for the numbers of items in different categories required in a final test, the developers and committees easily maintain a relatively stable distribution and balance of different items across the categories.

3. **SELECT EQUATING ITEMS FROM PREVIOUS YEAR'S MATRIX ITEMS.** Similarly, at their first meeting the Content Advisory Committees select matrix items that will be reused to provide the equating link that maintains standards across test forms and years. Decisions are based upon both breadth of coverage of content and Academic Expectations, statistical quality of the items, and equating needs.

4. **DEVELOP CONCEPTS FOR NEW ITEMS.** The Content Advisory Committees review the existing item pool for areas that need to be replenished. For identified areas, the appropriate Content Advisory Committee generates ideas for several possible items. The goal is always to develop more items than are needed for each content area. A larger number of items allow the advisory committees more choice and flexibility when selecting new items to match the *Core Content for Assessment*. The development of the *Core Content for Assessment* was described in Chapter 2 on page 2-17.for the upcoming year's test forms. Teachers draft items to address every bulleted item in the

   Beginning in 1996. multiple-choice items were pretested. Multiple-choice items may be written to elicit factual (recall) or higher order thinking. In KIRIS, since the open-response item seeks to elicit higher order thinking, the multiple-choice items seek to include both objectives, with an eye mainly toward expanding the coverage of the core content.

5. **WRITE INITIAL DRAFT OF NEW ITEMS.** Content Advisory Committee members receive resource materials and training to assist their work during the test development process. The materials include each of the following: development manuals for open-response and multiple-choice items, content information such as the Academic Expectations, *The Core Content for Assessment*, sections from *Transformations: Kentucky's Curriculum Framework;* and assessment information such as Kentucky's holistic scoring guide, released items from previous KIRIS assessments, and *Guidelines for Handling Sensitive Issues in Kentucky's State Assessment Development.* Training is designed around the use of the manuals and other materials during the development of assessment items.

   During the first meeting, advisory members are trained and begin generating ideas for new items and/or write initial drafts of new items. After the meeting, a contractor's staff member (typically the test developer assigned to that committee) takes the ideas generated by the Content Advisory Committees and develops initial drafts of new items or edits the initial drafts of new items produced by committee members.

6. **REVIEW AND REVISE INITIAL ITEM DRAFTS.** The contractor's test development supervisors review the initial item drafts. Potential improvements are discussed with the item writers.

7. **REVIEW AND SELECTION OF ITEMS FOR FIELD TESTING/PRE-TESTING BY CONTENT ADVISORY COMMITTEES.** At their second meeting, the Content

Advisory Committees review items that have been edited or revised by the contractor, suggest additional revisions, and select items for field-testing. As committees make their selection, they consistently examine the item to ensure it reflects Kentucky's Academic Expectations and *Core Content for Assessment.*

8. **ASSEMBLE FIELD TEST BOOKLETS.** A test assembler and a test developer select the items that will constitute each field test form. The last year for this step was 1995. In subsequent years items were pretested in the regular KIRIS assessment.

9. **REVIEW FIELD TEST BOOKLETS.** In 1995 the contractor's test development supervisor reviewed the assembled field test booklet. After 1995 this step became a review of the pretest items in the assembled KIRIS test booklets.

10. **ADMINISTER FIELD TESTS/PRETESTS.** New questions are field tested, gathering 50 to 100 student responses per question using students from schools in other states. The final year for this process was 1995. Beginning in 1996, items were pretested by inclusion in the operational KIRIS assessment. The items are not included in the accountability index for schools, nor are they reported publicly. This change allows for large samples of data collected from Kentucky students to be the basis for item analysis.

11. **ANALYZE FIELD TESTING/PRETESTING RESULTS.** After field testing or pretesting, student answers to open-response questions are scored by the contractor using a scoring rubric. In addition, the contractor brings samples of student responses to the next Content Advisory Committee meeting for review. Multiple-choice data are analyzed to produce item statistics used to identify potential problems with items. Questions are revised based on pretest results prior to and during committee meetings as the actual student pretest committees review responses.

12. **SELECT NEW ITEMS FOR UPCOMING YEAR'S TEST FORMS.** At the third meeting, the Content Advisory Committees select the pretest items needed to replace deleted test items. Scoring data and sample student responses, for open-response items, are used by committee members as they select items for the test. Also, the breadth of coverage of core content and Academic Expectations provides decision points for selection.

13. **REVIEW OF SELECTED ITEMS BY BIAS REVIEW COMMITTEE.** After final item sets are selected, all items are submitted to the Bias Review Committee. The role of this committee is to ensure that the content of the test is fair and equitable for all students, and that the tests contain no material that could be construed to be racially, ethnically, or gender stereotyped or biased toward any group. Each testing year this committee meets to review all open-response and multiple-choice items for every grade level. Committee members use response forms to record their specific concerns about each item and to provide suggestions for revision.

14. **REVIEW OF SELECTED ITEMS BY THE CONTRACTOR AND KDE STAFF.** The Content Advisory and Bias Review Committee's concerns are reviewed, and any appropriate revisions are made.

15. **FINAL ITEM SELECTION BY THE CONTRACTOR AND KDE STAFF.** Face-to-face meetings with KDE and the contractor are conducted. During these meetings, KDE and the contractor's staff give a final review to the coverage of core content and Academic Expectations, as well as the statistical qualities of the items. Items and scoring guides are reviewed and the contractor makes any appropriate revisions.

16. **ASSEMBLE TEST FORMS.** The contractor builds the 12 test forms for each grade level.

17. **REVIEW OF EACH ASSEMBLED FORM FOR CONTENT AND SENSITIVITY BALANCE BY THE CONTRACTOR.** While each individual item has been reviewed, the mix of items on each form is now assessed to ensure that the *Core Content for Assessment*, Academic Expectations, and politically and socially sensitive issues are in balance within each test form.

18. **REVIEW OF EACH ASSEMBLED FORM FOR CONTENT AND SENSITIVITY BALANCE BY KDE STAFF.** KDE staff review the final draft of each form before approving the forms for typeset. KDE staff must ensure that final revisions agreed upon at the face-to-face meetings have been made to each item.

19. **REVIEW OF FINAL TYPESET COPY FOR POTENTIAL TYPOGRAPHICAL ERRORS.** The contractor and KDE staff performs final reviews of typeset copy before giving approval to print.

SPECIAL STUDIES

Over the course of the third accountability cycle, various entities conducted several special studies as part of continuing test development. A summary of the results of those studies follows.

THE NITKO STUDY[2]. A study by Anthony J. Nitko of the University of Pittsburgh, published in May, 1997, investigated the match between the assessment and the core content and compared KIRIS with the CTBS/5, CTBS/4 (to a lesser extent) and the CAT5. The study identified some elements of the core content that were not being adequately covered or treated in a balanced fashion. Adjustments were made in subsequent assessments.

Nitko's study was primarily based upon limited interviews and the technical manuals for the tests evaluated. Nitko found that Goals 3 and 4 were not assessed by KIRIS (by

[2]Nitko, A.J. (May, 1997). *A guide to tests in Kentucky*. Frankfort, KY: Kentucky Institute for Educational Research.

law), and that only portions of the other four Goals were assessed, primarily because they did not lend themselves to paper and pencil tests. He found that the CTBS/5 matched Kentucky's core content better than the CTBS/4 or the CAT5, with the latter being more traditional and less affected by national standards set by professional organizations such as the National Council of Teachers of Mathematics. None of the multiple-choice tests matched the core curriculum well. Nitko concluded that the KIRIS assessment was adequate for assessing schools, but had limitations for assessing students. The CTBS/5 was deemed the most appropriate for comparing Kentucky students to a national sample. The complementary nature of the tests was emphasized. The different purposes and content made a clear case for using more than just KIRIS for evaluation of educational progress in Kentucky. The multiple-choice tests supplemented the school level orientation of KIRIS with the student level orientation of the other tests. Nitko found KIRIS somewhat less reliable than the other tests because of hand scoring written questions, but assumed the reliability was probably adequate for school accountability, but not for student decisions. The CAT5 and the CTBS/4 lack comparability to the KIRIS test and are not appropriate substitutes. The study answered thirteen questions about the assessments in a question and answer format. There were no tables or analytical data provided.

RESETTING STANDARDS. A second major effort to improve Kentucky's assessment system concluded during Accountability Cycle 3. That was a resetting of KIRIS performance standards, originally set in 1992, for the four classifications (Novice, Apprentice, Proficient and Distinguished) used in evaluating student accomplishment. More than 200 Kentuckians participated in evaluating the standards in October and November 1995. Following advice from the Office of Education Accountability (a legislative office assigned by law to oversee the implementation of KERA) to reset the standards, the various advisory committees (including the National Technical Working Group (NTWG) which advises KDE on technical issues) recommended to KDE to validate the standards rather than start from the beginning again. KDE developed a standards validation plan, which was reviewed by OEA and NTWG. Following implementation, the resulting data were presented to the NTWG on December 1-2, 1995. A report was presented to the Kentucky Board of Education for approval at its December 5-6, 1995 meeting.

At each grade level and for each subject area the study participants included teachers, generalists and parents. For high school standards, businesspersons and policy makers were added. The participants were divided into two groups, a confirmation group, and a replication group. For the confirmation group, participants were shown a description of each standard and told how these standards apply to the work sample. They were asked to decide whether the described standard applied to the work sample or if a different standard was more appropriate. The replication group performed the same activity, but was not told the performance level of the work according to the current standards.

NTWG confirmed the process and results of the standard setting. Following are four main results of the study. For the confirmation group, all standards were judged to be

reasonable and valid. For the confirmation group, which knew how the contractor scored the items, there were no significant differences among the five groups of judges (teachers, generalists, parents, businesspersons, policy makers). For the replication group, where the participants did not know the level that the example was scored by the contractor, results were consistent with the confirmation group, but a larger number felt the standards should be harder. Among those who disagreed with the standards, 20 percent felt they should be more difficult, while 5 percent suggested they should be easier. There was only partial consistency as to what standards were too easy across grade levels, although reading appeared less often than the other content areas. Tables 4-6 and 4-7 display the numbers of judges who agreed and disagreed with the standards.

TABLE 4-6
Percentage of Cases Judged in Confirmation Group to Agree with Standards

Standards Should Be	Grade 4				Grade 8				Grade 11			
	Reading	Math	Science	Social Studies	Reading	Math	Science	Social Studies	Reading	Math	Science	Social Studies
Lower	1	11	2	3	1	0	7	7	13	2	5	3
Same	86	66	70	79	89	79	78	59	61	82	77	82
Higher	14	23	28	18	10	21	15	34	26	17	17	15

Table 4-7
Percentage of Cases Judged in Replication Group to Agree with Standards

Standards Should Be	Grade 4				Grade 8				Grade 11			
	Reading	Math	Science	Social Studies	Reading	Math	Science	Social Studies	Reading	Math	Science	Social Studies
Lower	0	4	6	4	5	12	2	9	12	5	2	3
Same	70	73	69	68	80	74	67	60	68	80	50	67
Higher	27	26	25	28	15	14	31	31	20	15	48	30

Table 4-6 suggests that for grade 4 a sizable minority of participants believed that Mathematics and Science (23 percent and 28 percent respectively) should have higher standards. At the grade 8 a sizable minority of participants indicated Mathematics and Social Studies standards (21 percent and 34 percent respectively) were low. The result for grade 11 was different with Reading drawing the most attention for being too low (26 percent).

Table 4-7 indicates that the number of judges in the replication group suggested somewhat higher standards for Reading and Social Studies at the elementary level than in the confirmation group. Moreover, no subject appeared in greater need for higher standards than the others, with percentages ranging from 25 percent to 28 percent over all content areas. Among the replication group for grade 8, a sizable minority recommended higher standards in Science and Social Studies. Although both minority groups wanted higher standards for Social Studies, the replication group doubled the percentage wanting higher standards for Science (from 15 percent to 31 percent). For

grade 11 unlike the confirmation group, the replication group saw little need for higher standards in reading. Instead, a near majority (48 percent) recommended higher standards for science and a sizable minority (30 percent) for social studies. Both the confirmation and replication groups show a pattern in their responses. In all cases there are a much higher percentage of judges who believe that standards should be raised (ranging from 14 percent to 48 percent) than should be lowered (ranging from 1 percent to 13 percent). However, the vast majority (ranging from 70 percent to 89 percent) of both the confirmation and replication judges felt the standards should remain the same.

While resetting standards was the focus for Reading, Science, Mathematics, and Social Studies, the setting of initial standards was the focus in Arts & Humanities (A & H) and Practical Living/Vocational Studies (PL/VS). Prior to this initial setting of standards for these subjects, they had been tested in conjunction with one of the core subjects. The new standards for A & H and P L/V S allowed the independent testing of these subjects, and established the descriptors of Novice, Apprentice, Proficient and Distinguished for the first time.

For a complete description of the process, the selection of participants, tables of results and the analysis, see *KIRIS Standards Validation Study: Mathematics, Reading, Science, & Social Studies,* in the Occasional Paper Series, Bureau of Learning Results Services, Kentucky Department of Education, May 15, 1996 (LRS96-2), and KIRIS Standards Setting Study: Arts & Humanities, Practical Living/Vocational Studies, October 14, 1996 (LRS96-4).

For summaries of other studies completed before the beginning of Accountability Cycle 3 see the *KIRIS Accountability Cycle 2 Technical Manual.*

DEVELOPMENT OF SCORING GUIDES

Scoring guides and training materials for open-response questions that were carried over from Cycle 2 testing remained intact to ensure comparability of scores. New training materials were developed by test development staff, and the contractor's scoring staff, for those items new to the assessment during the year's from1995 through 1998.

During initial test development meetings, the advisory committees discussed and drafted scoring guides for open-response questions. These guides were finalized by the contractor's staff and scorers and reviewed by Department of Education content specialists. Scoring guides were then compared to actual papers returned from testing. After this review, scoring guides were revised and refined yet again. Finally, student papers were selected for training packs. Typically, two sets of training packs per question were assembled. One contained clear exemplars of student work at each score point and was used to orient scorers to the range of responses and associated score points for each question. The second pack contained less clear examples of score points and was used in training to gauge readers' internalization of the score points and readiness to begin scoring.

DEVELOPMENT OF MULTIPLE-CHOICE ITEMS

Multiple-choice items for the 1996 KIRIS assessment were included in a combined subjects pre-test section within each form. These items were used as part of a large open-response field test and were not included in the accountability index. In each case, multiple-choice items were developed following essentially the same procedures described above for open-response items. Beginning with the 1997 and 1998 test, the multiple-choice items were included in the test booklets with the intention of including the multiple-choice items in the school index for Accountability Cycle 4.

This page was intentionally left blank.