

Missouri End-of-Course Paper/pencil versus Online Comparability Study

Introduction and Data Collection

When implementing a computer-based or online assessment program, the comparability of the computer based assessment to its paper-and-pencil counterpart cannot be assumed. Conceivably, the nature of the construct being measured may change through computer administration, possibly showing up as a mean shift in difficulty or as an item by mode interaction. The purpose of this study is to describe a strategy for evaluating paper/pencil versus online comparability, and to provide a summary of several analyses performed to determine the comparability of the online and paper/pencil modes for the 2009 Spring administration of the Missouri End-of-Course (EOC) assessments. A specific challenge for the overall evaluation was the non-comparability of the student samples that were given the tests in each mode. The sample of students that took the online form was entirely determined by the willingness of the school or district to volunteer for the online administration.

Three separate comparisons were used to study the comparability of the online and paper/pencil assessments. The comparisons included:

1. all online test-takers versus all paper/pencil test takers,
2. a representative sample of online test-takes versus all paper/pencil test takers,
3. a matched sample of online test-takes versus paper/pencil test takers.

The first comparison (all online test-takers) was performed to set a baseline for the differences between the online and paper/pencil assessments. Students who participated in online testing were not representative of the population of students in Missouri. As such, differences between the two modes were expected. The second comparison (representative sample of online) was needed to more accurately determine the differences between the modes. This second comparison was made in an attempt to eliminate possible differences based on student demographics. Finally, the third comparison (matched sample of online) involved a tighter matching of a representative sample of online test-takers to a sample of paper/pencil test-takes. Matching for the third comparison was achieved using gender, ethnicity and age in years.

The following steps were used to match students:

1. Eliminate a targeted set of student records from the online sample until the sample is representative of the entire population of students. This “trimmed” set of students was also used for the second comparison (representative sample of online) mentioned above.
2. Sort the representative sample of online students and the entire set of paper/ pencil students by gender, ethnicity and age in years.
3. Combine the two datasets by merging students by gender, ethnicity and age in years.
4. Create a uniform random variable and sort by student ID and the random variable. This step was necessary because each online student matched with multiple paper/ pencil students (because of their larger numbers).
5. Select the *first* paper/pencil student matched with each online student. Because of the limited number of variables used for matching, virtually all online students were matched to a paper/pencil student.

Ethnicity was the most important of the available demographic variables for trimming the online sample to achieve representativeness as compared to the overall state population of students. Tables 1 through 4 provide distributions for ethnicity for each sample of students used in the three comparisons outlined above (i.e., all online test-takers, representative sample of online, and matched sample of online). Note that distributions by EOC test (i.e., Algebra I, Biology and English II) had similar percentages and thus are not reported individually.

Table 1 - Ethnicity Distribution for Total Sample of Paper/Pencil Test Takers

Ethn	Frequency	Percent	Cumulative Frequency	Cumulative Percent
AFAM	20523	13.92	20523	13.92
AIAN	714	0.48	21237	14.40
ASPI	2737	1.86	23974	16.26
HISP	4261	2.89	28235	19.15
WHIT	119222	80.85	147457	100.00

Table 2 - Ethnicity Distribution for Total Sample of Online Test Takers

Ethn	Frequency	Percent	Cumulative Frequency	Cumulative Percent
AFAM	4276	23.25	4276	23.25
AIAN	71	0.39	4347	23.63
ASPI	265	1.44	4612	25.07
HISP	555	3.02	5167	28.09
WHIT	13227	71.91	18394	100.00

Table 3 - Ethnicity Distribution for the “Trimmed” Sample of Online Test Takers

Ethn	Frequency	Percent	Cumulative Frequency	Cumulative Percent
AFAM	2100	13.69	2100	13.69
AIAN	62	0.40	2162	14.10
ASPI	230	1.50	2392	15.60
HISP	532	3.47	2924	19.07
WHIT	12412	80.93	15336	100.00

Table 4 - Ethnicity Distribution for the Matched Sample of Online Test Takers

Ethn	Frequency	Percent	Cumulative Frequency	Cumulative Percent
AFAM	2100	13.70	2100	13.70
AIAN	61	0.40	2161	14.09
ASPI	230	1.50	2391	15.59
HISP	531	3.46	2922	19.06
WHIT	12411	80.94	15333	100.00

Data Analyses

This section provides a summary of several analyses completed to determine how comparable the online and paper/pencil modes were for the 2009 Spring administration of the Missouri EOC assessments. Analyses generally fell into the following four categories:

1. Comparison of summary statistics for the three comparisons (i.e., all online test-takers, representative sample of online, and matched sample of online).
2. Comparison of item p-values and item means (PE items) for the three comparisons (i.e., all online test-takers, representative sample of online, and matched sample of online).
3. Comparison of the difference between Rasch values for matched samples and a comparison between the raw score to scale score results for matched samples.
4. A factor analytic comparison for matched samples.

Comparison of Mean Raw Scores

Tables 5 through 7 present summary statistics for the three comparisons. Note that mean raw score differences between paper/pencil and online modes are larger in Table 5 (all online test-takers) where the online sample of students is not representative of the total student population. Differences generally become smaller (less than one raw score point) for the last two comparisons (representative sample of online, and matched sample of online) where the samples of online students for comparison is more demographically representative.

Table 5 – Summary Statistics for the Original Sample of Paper/Pencil and Online Test Takers

Algebra I					
Mode*	N	Mean	Std Dev	Minimum	Maximum
No	48622	21.7694459	6.9978379	2.0000000	38.0000000
Yes	3956	20.3549039	6.9288914	4.0000000	38.0000000
Biology					
Mode*	N	Mean	Std Dev	Minimum	Maximum
No	48992	33.0761349	9.6685893	3.0000000	55.0000000
Yes	6343	32.1831941	9.6043550	5.0000000	55.0000000
English II					
Mode*	N	Mean	Std Dev	Minimum	Maximum
No	49843	27.4629336	6.2196791	3.0000000	39.0000000
Yes	6837	26.3125640	6.1330700	5.0000000	39.0000000

*Mode: No = Paper/pencil; Yes = Online.

Table 6 – Summary Statistics for the Trimmed Sample of Online Test Takers

Algebra I					
Mode*	N	Mean	Std Dev	Minimum	Maximum
No	48622	21.7694459	6.9978379	2.0000000	38.0000000
Yes	3525	20.8306383	6.8409839	4.0000000	38.0000000
Biology					
Mode*	N	Mean	Std Dev	Minimum	Maximum
No	48992	33.0761349	9.6685893	3.0000000	55.0000000
Yes	5672	32.7471791	9.5579002	5.0000000	55.0000000
English II					
Mode*	N	Mean	Std Dev	Minimum	Maximum
No	49843	27.4629336	6.2196791	3.0000000	39.0000000
Yes	6139	26.6388663	6.0743515	5.0000000	39.0000000

*Mode: No = Paper/pencil; Yes = Online.

Table 7 – Summary Statistics for the Matched Sample of Paper/Pencil and Online Test Takers

Algebra I					
Mode*	N	Mean	Std Dev	Minimum	Maximum
No	3524	21.8226447	6.8853420	2.0000000	38.0000000
Yes	3524	20.8331442	6.8403365	4.0000000	38.0000000
Biology					
Mode*	N	Mean	Std Dev	Minimum	Maximum
No	5672	32.6740127	9.8738316	4.0000000	55.0000000
Yes	5672	32.7471791	9.5579002	5.0000000	55.0000000
English II					
Mode*	N	Mean	Std Dev	Minimum	Maximum
No	6137	27.4551084	6.2633415	5.0000000	39.0000000
Yes	6137	26.6408669	6.0723913	5.0000000	39.0000000

*Mode: No = Paper/pencil; Yes = Online.

Comparison of Item P-values

Tables 7 through 12 present comparisons between item p-values and item means (PE items) for the third comparison only (matched sample of online). For each content area, the difference between the paper/pencil and online item p-value is provided (Tables 7, 9 and 11). In addition, the frequency of differences between p-values for each matched sample is given (Tables 8, 10 and 12). Differences between p-values were generally small, falling within the range of -.05 to .05. Some items did show larger differences perhaps suggesting somewhat of an interaction of item difficulty and mode of administration.

Table 7– Algebra I – Difference between P-values for Matched Samples

Oper Seq	Book Seq	Paper/ Pencil	Online	Difference
1	pv1	0.92	0.90	0.02
2	pv2	0.77	0.77	0.00
3	pv3	0.72	0.70	0.02
4	pv4	0.74	0.72	0.02
5	pv5	0.73	0.69	0.04
6	pv10	0.83	0.79	0.04
7	pv11	0.73	0.68	0.05
8	pv12	0.74	0.72	0.02
9	pv13	0.60	0.58	0.02
10	pv14	0.43	0.43	0.00
11	pv15	0.71	0.67	0.04
12	pv16	0.68	0.67	0.01
13	pv17	0.83	0.84	-0.01
14	pv18	0.81	0.78	0.03
15	pv19	0.66	0.64	0.02
16	pv20	0.58	0.58	0.00
17	pv21	0.55	0.55	0.00
18	pv26*	0.64	0.10	0.54*
19	pv27	0.50	0.41	0.09
20	pv28	0.55	0.53	0.02
21	pv29	0.53	0.52	0.01
22	pv30	0.52	0.52	0.00
23	pv31	0.58	0.58	0.00
24	pv32	0.48	0.46	0.02
25	pv33	0.41	0.32	0.09
26	pv34	0.49	0.49	0.00
27	pv35	0.61	0.60	0.01
28	pv36	0.38	0.32	0.06
29	pv37	0.57	0.56	0.01
30	pv38	0.38	0.33	0.05
31	pv43	0.32	0.29	0.03
32	pv44	0.58	0.59	-0.01
33	pv45	0.42	0.38	0.04
34	pv46	0.12	0.08	0.04
35	pv47	0.34	0.27	0.07
36	pv48	2.00	1.88	0.12

* Item #26 was scored differently for online test takers. It has been excluded from most analyses in the current study.

Table 8 – Algebra I – Frequency of Differences between P-values for Matched Samples

Difference	Frequency	Percent	Cumulative Frequency	Cumulative Percent
-0.01	2	5.56	2	5.56
0.00	7	19.44	9	25.00
0.01	4	11.11	13	36.11
0.02	8	22.22	21	58.33
0.03	2	5.56	23	63.89
0.04	5	13.89	28	77.78
0.05	2	5.56	30	83.33
0.06	1	2.78	31	86.11
0.07	1	2.78	32	88.89
0.09	2	5.56	34	94.44
0.12	1	2.78	35	97.22
0.54	1	2.78	36	100.00

Table 9 – Biology – Difference between P-values for Matched Samples

Oper Seq	Book Seq	Paper/Pencil	Online	Difference
1	pv1	0.82	0.79	0.03
2	pv2	0.79	0.79	0.00
3	pv3	0.91	0.91	0.00
4	pv4	0.63	0.62	0.01
5	pv5	0.85	0.81	0.04
6	pv10	0.68	0.67	0.01
7	pv11	0.74	0.74	0.00
8	pv12	0.54	0.53	0.01
9	pv13	0.65	0.61	0.04
10	pv14	0.47	0.46	0.01
11	pv15	0.58	0.64	-0.06
12	pv16	0.77	0.75	0.02
13	pv17	0.93	0.93	0.00
14	pv18	0.69	0.69	0.00
15	pv19	0.48	0.48	0.00
16	pv20	0.31	0.34	-0.03
17	pv21	0.59	0.57	0.02
18	pv26	0.75	0.75	0.00
19	pv27	0.65	0.63	0.02
20	pv28	0.50	0.45	0.05
21	pv29	0.42	0.41	0.01
22	pv30	0.43	0.41	0.02
23	pv31	0.42	0.41	0.01
24	pv32	0.59	0.57	0.02
25	pv33	0.61	0.56	0.05
26	pv34	0.60	0.58	0.02
27	pv35	0.72	0.69	0.03
28	pv36	0.63	0.63	0.00
29	pv37	0.48	0.52	-0.04
30	pv38	0.72	0.69	0.03
31	pv43	0.61	0.63	-0.02
32	pv44	0.56	0.60	-0.04
33	pv45	0.43	0.45	-0.02
34	pv46	0.57	0.56	0.01
35	pv47	0.83	0.84	-0.01
36	pv48	0.67	0.69	-0.02
37	pv49	0.79	0.80	-0.01
38	pv50	0.73	0.76	-0.03
39	pv51	2.59	2.77	-0.18
40	pv52	0.74	0.70	0.04
41	pv53	0.93	1.06	-0.13
42	pv54	1.16	1.10	0.06
43	pv55	0.74	0.76	-0.02

44	pv56	0.66	0.67	-0.01
45	pv57	1.00	1.07	-0.07
46	pv58	0.86	0.75	0.11

Table 10 – Biology – Frequency of Differences between P-values for Matched Samples

Difference	Frequency	Percent	Cumulative Frequency	Cumulative Percent
-0.18	1	2.17	1	2.17
-0.13	1	2.17	2	4.35
-0.07	1	2.17	3	6.52
-0.06	1	2.17	4	8.70
-0.04	2	4.35	6	13.04
-0.03	2	4.35	8	17.39
-0.02	4	8.70	12	26.09
-0.01	3	6.52	15	32.61
0.00	8	17.39	23	50.00
0.01	7	15.22	30	65.22
0.02	6	13.04	36	78.26
0.03	3	6.52	39	84.78
0.04	3	6.52	42	91.30
0.05	2	4.35	44	95.65
0.06	1	2.17	45	97.83
0.11	1	2.17	46	100.00

Table 11 – English II – Difference between P-values for Matched Samples

Oper Seq	Book Seq	Paper/ Pencil	Online	Difference
1	pv1	0.74	0.70	0.04
2	pv2	0.74	0.74	0.00
3	pv3	0.89	0.86	0.03
4	pv4	0.88	0.81	0.07
5	pv5	0.90	0.88	0.02
6	pv6	0.38	0.33	0.05
7	pv7	0.85	0.86	-0.01
8	pv8	0.75	0.70	0.05
9	pv9	0.48	0.46	0.02
10	pv10	0.75	0.70	0.05
11	pv11	0.84	0.79	0.05
12	pv12	0.53	0.57	-0.04
13	pv24	0.58	0.59	-0.01
14	pv25	0.54	0.53	0.01
15	pv26	0.86	0.86	0.00
16	pv27	0.56	0.48	0.08
17	pv28	0.66	0.67	-0.01
18	pv29	0.45	0.48	-0.03
19	pv30	0.45	0.39	0.06
20	pv31	0.81	0.78	0.03
21	pv32	0.75	0.70	0.05
22	pv33	0.46	0.37	0.09
23	pv34	0.65	0.63	0.02
24	pv35	0.89	0.87	0.02
25	pv36	0.91	0.90	0.01
26	pv37	0.60	0.58	0.02
27	pv38	0.58	0.55	0.03
28	pv39	0.85	0.82	0.03
29	pv40	0.65	0.63	0.02
30	pv41	0.70	0.73	-0.03
31	pv43	0.87	0.86	0.01
32	pv44	0.54	0.54	0.00
33	pv45	0.80	0.81	-0.01
34	pv46	0.80	0.80	0.00
35	pv47	0.73	0.72	0.01
36	pv48	3.03	2.92	0.11

Table 12 – English II – Frequency of Differences between P-values for Matched Samples

Difference	Frequency	Percent	Cumulative Frequency	Cumulative Percent
-0.04	1	2.78	1	2.78
-0.03	2	5.56	3	8.33
-0.01	4	11.11	7	19.44
0.00	4	11.11	11	30.56
0.01	4	11.11	15	41.67
0.02	6	16.67	21	58.33
0.03	4	11.11	25	69.44
0.04	1	2.78	26	72.22
0.05	5	13.89	31	86.11
0.06	1	2.78	32	88.89
0.07	1	2.78	33	91.67
0.08	1	2.78	34	94.44
0.09	1	2.78	35	97.22
0.11	1	2.78	36	100.00

Comparison of Differences for IRT Based Values

Tables 13 through 18 present comparisons between Rasch values and scale scores for the third comparison only (matched sample of online). Note that the scale scores for online are hypothetical because a decision to conduct a unique scaling for the online administration has not yet been made. However, these analyses are included to represent the impact on achievement level classifications should a decision be made to scale the online administrations based on this dataset.

For each content area, the difference between the paper/pencil (operational calibration) and online Rasch value is provided (Tables 13, 15 and 17). In addition, the difference between the paper/pencil (operational calibration) and online scale scores for the matched sample is given (Tables 14, 16 and 18). Differences between Rasch values were generally small, most less than .3 logits. More importantly, the cut scores for each mode for the Proficient and Advanced Achievement Levels were the same for each EOC assessment. Note that for Algebra I, the raw score cuts essentially dropped one raw score point because item #26 was dropped from the analysis.

Table 13 – Algebra I – Difference between Rasch values for Online Matched Sample

Oper Seq	Book Seq	P/P	Online	Difference
1	1	-2.23	-2.31	0.08
2	2	-1.04	-1.12	0.08
3	3	-0.71	-0.71	0.00
4	4	-0.81	-0.83	0.02
5	5	-0.78	-0.66	-0.12
6	10	-1.41	-1.28	-0.13
7	11	-0.68	-0.60	-0.08
8	12	-0.91	-0.84	-0.07
9	13	-0.11	-0.11	0.00
10	14	0.75	0.62	0.13
11	15	-0.55	-0.55	0.00
12	16	-0.44	-0.54	0.09
13	17	-1.43	-1.69	0.26
14	18	-1.29	-1.23	-0.06
15	19	-0.38	-0.38	0.00
16	20	0.08	-0.12	0.20
17	21	0.27	0.05	0.22
18	27	0.47	0.73	-0.26
19	28	0.19	0.15	0.04
20	29	0.33	0.20	0.13
21	30	0.31	0.21	0.09
22	31	0.05	-0.12	0.17
23	32	0.54	0.50	0.05
24	33	0.88	1.19	-0.31
25	34	0.45	0.32	0.13
26	35	-0.13	-0.18	0.05
27	36	1.03	1.20	-0.17
28	37	0.01	0.00	0.01
29	38	1.04	1.14	-0.10
30	43	1.37	1.38	-0.01
31	44	0.01	-0.17	0.18
32	45	0.84	0.86	-0.03
33	46	2.80	3.19	-0.39
34	47	1.26	1.49	-0.23
35	48	0.19	0.18	0.02

Table 14 – Algebra I – Difference between Scale Scores for Online Matched Sample

Score	P/P*	Online	Difference	Cut** Adjustment
0	100	100	0	
1	112	111	1	
2	127	126	1	
3	136	135	1	
4	143	142	1	
5	149	148	1	
6	154	153	1	
7	158	157	1	
8	162	161	1	
9	165	165	0	
10	169	168	1	
11	172	171	1	
12	175	174	1	
13	178	177	1	
14	181	180	1	
15	183	183	0	
16	186	186	0	
17	189	188	1	
18	191	191	0	
19	194	193	1	
20	196	196	0	
21	200 (199)	200 (199)	0	*
22	202	201	1	
23	204	204	0	
24	207	207	0	
25	210	209	1	
26	212	212	0	
27	215	215	0	
28	218	218	0	
29	221	221	0	
30	225 (224)	225	-1	*
31	228	228	0	
32	232	233	-1	
33	236	237	-1	
34	242	243	-1	
35	248	250	-2	
36	250	250	0	
37	250	250	0	
38	250	250	0	

* Paper/pencil calibration run on 38 items so raw score to scale score conversion will not be the same as for the operational raw score to scale score tables used for reporting.

** Raw score cuts were reduced by one score point from the operational test because item #26 was dropped from the analysis.

Table 15 – Biology – Difference between Rasch values for Online Matched Sample

Oper Seq	Book Seq	P/P	Online	Difference
1	1	-1.12	-0.94	-0.18
2	2	-0.90	-0.92	0.02
3	3	-1.98	-2.05	0.07
4	4	0.00	0.01	-0.01
5	5	-1.32	-1.08	-0.24
6	10	-0.27	-0.22	-0.05
7	11	-0.63	-0.59	-0.04
8	12	0.43	0.47	-0.04
9	13	-0.12	0.08	-0.20
10	14	0.85	0.78	0.07
11	15	0.31	-0.05	0.35
12	16	-0.76	-0.67	-0.08
13	17	-2.38	-2.39	0.01
14	18	-0.36	-0.35	-0.01
15	19	0.69	0.71	-0.02
16	20	1.57	1.41	0.16
17	21	0.19	0.27	-0.09
18	26	-0.66	-0.68	0.03
19	27	-0.15	0.00	-0.15
20	28	0.61	0.84	-0.22
21	29	1.04	1.04	0.00
22	30	1.00	1.05	-0.06
23	31	0.98	1.04	-0.06
24	32	0.18	0.28	-0.10
25	33	0.13	0.31	-0.17
26	34	0.16	0.22	-0.06
27	35	-0.49	-0.32	-0.17
28	36	0.00	0.01	0.00
29	37	0.72	0.52	0.20
30	38	-0.51	-0.35	-0.16
31	43	0.08	-0.01	0.09
32	44	0.34	0.14	0.20
33	45	0.90	0.86	0.04
34	46	0.30	0.31	-0.01
35	47	-1.30	-1.34	0.04
36	48	-0.23	-0.31	0.08
37	49	-0.89	-1.00	0.11
38	50	-0.50	-0.70	0.20
39	51	-0.03	-0.27	0.24
40	52	1.31	1.33	-0.01
41	53	0.79	0.49	0.31
42	54	0.30	0.39	-0.10
43	55	-0.64	-0.71	0.08
44	56	-0.17	-0.23	0.06
45	57	1.41	1.30	0.11
46	58	1.08	1.35	-0.27

Table 16 – Biology – Difference between Scale Scores for Online Matched Sample

Score	P/P	Online	Difference	Cut Adjustment
0	100	100	0	
1	107	107	0	
2	121	121	0	
3	130	130	0	
4	137	137	0	
5	142	142	0	
6	146	146	0	
7	150	150	0	
8	153	153	0	
9	156	156	0	
10	159	159	0	
11	162	161	1	
12	164	164	0	
13	166	166	0	
14	169	168	1	
15	171	170	1	
16	173	172	1	
17	175	174	1	
18	177	177 (176)	1	*
19	178	178	0	
20	180	180	0	
21	182	181	1	
22	184	183	1	
23	185	185	0	
24	187	187	0	
25	189	188	1	
26	190	190	0	
27	192	191	1	
28	193	193	0	
29	195	195	0	
30	197	196	1	
31	198	198	0	
32	200	200	0	
33	202	201	1	
34	203	203	0	
35	205	204	1	
36	207	206	1	
37	208	208	0	
38	210	210	0	
39	212	212	0	
40	214	213	1	
41	216	215	1	
42	218	218	0	
43	220	220	0	
44	223	222	1	
45	225	225	0	
46	228	227	1	
47	231	230	1	
48	234	234	0	

DRAFT - Missouri End-of-Course Paper/pencil versus Online Comparability Study

49	238	237	1
50	242	242	0
51	247	247	0
52	250	250	0
53	250	250	0
54	250	250	0
55	250	250	0

Table 17 – English II – Difference between Rasch values for Online Matched Sample

Oper Seq	Book Seq	P/P	Online	Difference
1	1	-0.12	-0.04	-0.08
2	2	-0.15	-0.29	0.13
3	3	-1.28	-1.13	-0.14
4	4	-1.22	-0.69	-0.53
5	5	-1.44	-1.32	-0.12
6	6	1.79	1.84	-0.05
7	7	-0.92	-1.11	0.19
8	8	-0.15	-0.02	-0.13
9	9	1.19	1.17	0.02
10	10	-0.19	-0.03	-0.16
11	11	-0.79	-0.61	-0.17
12	12	0.97	0.64	0.34
13	24	0.74	0.54	0.20
14	25	0.93	0.84	0.09
15	26	-1.05	-1.15	0.10
16	27	0.89	1.07	-0.18
17	28	0.35	0.15	0.20
18	29	1.40	1.07	0.32
19	30	1.43	1.52	-0.09
20	31	-0.64	-0.52	-0.12
21	32	-0.18	-0.05	-0.13
22	33	1.37	1.61	-0.24
23	34	0.42	0.35	0.07
24	35	-1.36	-1.22	-0.14
25	36	-1.64	-1.61	-0.03
26	37	0.65	0.61	0.04
27	38	0.75	0.77	-0.02
28	39	-0.90	-0.83	-0.07
29	40	0.41	0.34	0.07
30	41	0.07	-0.19	0.27
31	43	-1.10	-1.14	0.04
32	44	0.95	0.81	0.14
33	45	-0.54	-0.77	0.22
34	46	-0.50	-0.66	0.17
35	47	-0.08	-0.13	0.05
36	48	-0.07	0.19	-0.27

Table 18 – English II – Difference between Scale Scores for Online Matched Sample

Score	P/P	Online	Difference	Cut Adjustment
1	105	106	-1	
2	125	126	-1	
3	137	138	-1	
4	145	146	-1	
5	150	151	-1	
6	155	155	0	
7	158	159	-1	
8	162	163	-1	
9	165	166	-1	
10	168	168	0	
11	171	171	0	
12	173	173	0	
13	176	176	0	
14	178	178	0	
15	180	180	0	
16	182	183	-1	
17	185	185	0	
18	187	187	0	
19	189	189	0	
20	191	191	0	
21	193	193	0	
22	196	195	1	
23	198	198	0	
24	200	200	0	
25	202	202	0	
26	205	204	1	
27	207	207	0	
28	210	209	1	
29	212	212	0	
30	215	215	0	
31	218	218	0	
32	221	221	0	
33	225	225	0	
34	229	229	0	
35	234	234	0	
36	240	240	0	
37	248	248	0	
38	250	250	0	
39	250	250	0	

Factor Analytic Comparison

English II

The extraction method utilized was Iterated Principal Factors with an Oblique Varimax rotation. The initial runs with paper/pencil data yielded 4 eigenvalues that were greater than 1 whereas the online data yielded 6 eigenvalues that were greater than 1. Upon review of the initial eigenvalue scree plot as well as the difference between reduced correlation matrix eigenvalues, it was decided to retain 4 factors in each of the two data sets. Additionally, because our primary objective was to compare the factor structures of the paper and online administrations, it seemed appropriate to focus on the same number of factors for both analyses to enhance the interpretability of the comparison.

Table 19 - Percent Agreement of Items with High P/P Factor Loadings that also Loaded Highly with Similarly Defined Online Factors

P/P Factor #	Similarly Defined Online Factor #	# High Load OnLine Items / # P/P Items	Percent Agreement
1	2	15/16	93.8%
2	1	14/19	73.7%
3	3	9/11	81.8%
4	4	11/12	91.7%

The greatest agreement across administrations occurred with paper/pencil factor 1 and online factor 2. Of the 16 items with high factor loadings on operational factor 1, 15 or 93.8% also loaded highly with online factor 2. The least agreement across administrations was with operational factor 2, where 73.7% of the items also loaded highly with online factor 1.

Table 20 - English II Paper/pencil Factor Analysis of Tetrachoric Correlation Coefficients

Factor Structure (Loadings/Correlations)

	Factor1	Factor2	Factor3	Factor4
v1	55 *	41 *	49 *	33
v2	41 *	33	26	26
v3	49 *	28	18	23
v4	81 *	50 *	31	38
v5	59 *	38	25	34
v6	10	10	33	14
v7	51 *	43 *	43 *	36
v8	56 *	45 *	40	35
v9	29	23	50 *	25
v10	46 *	41 *	33	32
v11	68 *	59 *	43 *	41 *
v12	35	30	37	26
v13	32	27	37	23
v14	35	31	40	26
v15	57 *	57 *	43 *	46 *
v16	57 *	47 *	63 *	37
v17	32	34	29	23
v18	15	21	1	11
v19	34	36	53 *	33
v20	43 *	57 *	40	43 *
v21	50 *	60 *	50 *	46 *
v22	25	26	45 *	23
v23	31	39	37	29
v24	45 *	66 *	34	52 *
v25	48 *	83 *	33	60 *
v26	31	42 *	30	28
v27	39	42 *	43 *	35
v28	44 *	64 *	26	51 *
v29	33	45 *	35	37
v30	38	53 *	28	43 *
v31	38	52 *	32	58 *
v32	35	34	43 *	37
v33	24	33	15	68 *
v34	19	23	17	58 *
v35	35	40 *	36	50 *
v36	*minimum score=1 with 0 variance as a result 'no' factor loadings			

Printed values are multiplied by 100 and rounded to the nearest integer. Values greater than 0.403429 are flagged by an '*'.

Algebra I

The extraction method utilized was Iterated Principal Factors with an Oblique Varimax rotation. The initial runs with paper/pencil data yielded 5 eigenvalues that were greater than 1 whereas the online data yielded 7 eigenvalues that were greater than 1. Though the scree plot indicated that there could be 5 factors present, initial runs proved difficult to interpret. Upon closer examination of the differences between reduced correlation matrix eigenvalues (successive eigenvalues showed little change after the fourth factor), and in consideration of factor loading interpretability issues, it was decided to retain 4 factors in each of the two data sets. As with the English II analysis, because our primary objective was to compare the factor structures of the paper and online administrations, it seemed appropriate to focus on the same number of factors for both datasets to enhance the interpretability of the comparison.

Table 22 - Percent Agreement of Items with High P/P Factor Loadings that also Loaded Highly with Similarly Defined Online Factors

P/P Factor #	Similarly Defined Online Factor #	# High Load OnLine Items / # P/P Items	Percent Agreement
1	1	12/14	85.7%
2	2	14/18	77.8%
3	4	12/15	80.0%
4	3	4/6	66.7%

The greatest agreement across administrations occurred with paper/pencil factor 1 and online factor 1. Of the 14 items with high factor loadings on paper/pencil factor 1, 12 or 85.7% also loaded highly with online factor 1. The least agreement across administrations occurred with operational factor 4, where 66.7% of the items also loaded highly with online factor 3.

Table 23 – Algebra I - Paper/pencil Factor Analysis of Tetrachoric Correlation Coefficients

Factor Structure (Loadings/Correlations)				
	Factor1	Factor2	Factor3	Factor4
v1	41 *	43 *	24	32
v2	23	23	12	14
v3	50 *	43 *	44 *	36
v4	34	37	25	26
v5	57 *	37	48 *	39
v6	40 *	27	38	33
v7	42 *	38	27	29
v8	48 *	45 *	38	36
v9	39	51 *	44 *	39 *
v10	22	57 *	39	34
v11	47 *	66 *	38	43 *
v12	44 *	57 *	37	40 *
v13	50 *	46 *	32	39
v14	60 *	41 *	38	36
v15	63 *	59 *	47 *	46 *
v16	24	49 *	27	30
v17	34	59 *	40 *	39
v18	45 *	38	37	33
v19	33	41 *	46 *	37
v20	27	30	56 *	31
v21	37	32	44 *	32
v22	23	31	28	37
v23	38	39 *	44 *	37
v24	27	55 *	46 *	38
v25	24	32	64 *	33
v26	**dropped**			
v27	24	32	46 *	34
v28	15	41 *	57 *	35
v29	44 *	38	43 *	32
v30	20	26	37	27
v31	30	35	33	31
v32	23	42 *	27	36
v33	4	11	13	65 *
v34	-4	26	31	28
v35	21	32	45 *	29
v36	72 *	48 *	33	47 *

Printed values are multiplied by 100 and rounded to the nearest integer. Values greater than 0.38974 are flagged by an '*'.

Table 24 – Algebra I - Online Factor Analysis of Tetrachoric Correlation Coefficients

	Factor Structure (Loadings/Correlations)			
	Factor1	Factor2	Factor3	Factor4
v1	48 *	39 *	21	26
v2	26	22	17	5
v3	44 *	34	39 *	43 *
v4	37	31	39 *	26
v5	52 *	11	39 *	43 *
v6	43 *	28	31	34
v7	40 *	31	29	16
v8	44 *	41 *	40 *	36
v9	37	47 *	39 *	35
v10	27	52 *	39 *	39 *
v11	49 *	64 *	37	35
v12	36	44 *	38 *	36
v13	51 *	46 *	29	31
v14	67 *	35	34	45 *
v15	64 *	53 *	49 *	50 *
v16	22	47 *	36	24
v17	39 *	55 *	42 *	39 *
v18	-31	-16	-18	-22
v19	31	41 *	38 *	47 *
v20	16	25	23	50 *
v21	40 *	12	39 *	38 *
v22	22	36	14	24
v23	48 *	27	39 *	40 *
v24	30	54 *	50 *	42 *
v25	18	18	35	67 *
v26	dropped			
v27	31	31	36	33
v28	23	32	47 *	45 *
v29	49 *	27	45 *	38 *
v30	23	9	42 *	25
v31	37	44 *	40 *	29
v32	27	40 *	28	26
v33	9	20	9	12
v34	13	21	64 *	26
v35	29	27	46 *	36
v36	76 *	38 *	44 *	24

Printed values are multiplied by 100 and rounded to the nearest integer. Values greater than 0.374458 are flagged by an '*'.

Biology

The extraction method utilized was Iterated Principal Factors with an Oblique Varimax rotation. The initial runs with both the paper/pencil as well as online data yielded 6 eigenvalues that were greater than 1. Initial runs with 6 factors posed problems with interpretability. After review of the initial eigenvalue scree plot, differences between reduced correlation matrix eigenvalues (successive eigenvalues showed little change after the fourth factor), and factor loading interpretability issues, it was decided to retain 4 factors in each of the two data sets. As with the English II and Algebra analyses, because our primary objective was to compare the factor structures of the paper/pencil and online administrations, it seemed appropriate to focus on the same number of factors for both datasets to enhance the interpretability of the comparison.

Table 25 - Percent Agreement of Items with High P/P Factor Loadings that also Loaded Highly with Similarly Defined Online Factors

P/P Factor #	Similarly Defined Online Factor #	# High Load OnLine Items / # P/P Items	Percent Agreement
1	1	9/11	81.8%
2	2	16/17	94.1%
3	3	20/22	90.9%
4	4	18/21	85.7%

The greatest agreement across administrations occurred with paper/pencil factor 2 and online factor 2. Of the 17 items with high factor loadings on operational factor 2, 16 or 94.1% also loaded highly with online factor 2. The least agreement across administrations was with operational factor 1, where 81.8% of the items also loaded highly with online factor 1.

Table 26 – Biology - Paper/pencil Factor Analysis of Tetrachoric Correlation Coefficients

	Factor Structure (Loadings/Correlations)			
	Factor1	Factor2	Factor3	Factor4
v1	35	37	58 *	39 *
v2	38	39	54 *	47 *
v3	17	20	24	17
v4	29	28	23	39
v5	30	36	43 *	30
v6	34	40 *	49 *	37
v7	33	37	55 *	36
v8	28	25	25	43 *
v9	40 *	39	46 *	56 *
v10	25	25	32	37
v11	17	14	23	27
v12	36	39	56 *	39
v13	44 *	56 *	69 *	36
v14	40 *	42 *	60 *	45 *
v15	34	32	42 *	49 *
v16	20	16	19	33
v17	18	17	33	24
v18	42 *	44 *	47 *	56 *
v19	32	34	36	39 *
v20	51 *	54 *	48 *	62 *
v21	34	31	37	51 *
v22	20	18	30	28
v23	32	30	41 *	47 *
v24	35	37	51 *	44 *
v25	38	37	42 *	53 *
v26	35	33	45 *	49 *
v27	26	31	36	27
v28	36	40 *	39	48 *
v29	28	24	29	47 *
v30	33	36	40 *	42 *
v31	19	23	31	22
v32	30	30	31	43 *
v33	33	31	41 *	52 *
v34	33	34	40 *	41 *
v35	40 *	45 *	52 *	45 *
v36	35	55 *	35	23
v37	93 *	37	18	23
v38	96 *	42 *	21	29
v39	49 *	64 *	45 *	35
v40	33	41 *	35	26
v41	29	40 *	35	22
v42	40 *	57 *	46 *	31
v43	38	62 *	29	31
v44	43 *	74 *	34	32
v45	34	52 *	36	30
v46	26	42 *	16	16

Printed values are multiplied by 100 and rounded to the nearest integer. Values greater than 0.393858 are flagged by an '*'.

Table 27 – Biology - Online Factor Analysis of Tetrachoric Correlation Coefficients

	Factor Structure (Loadings/Correlations)			
	Factor1	Factor2	Factor3	Factor4
v1	34	34	57 *	41 *
v2	34	44 *	52 *	48 *
v3	18	20	24	18
v4	31	33	21	45 *
v5	29	35	41 *	30
v6	32	39 *	45 *	36
v7	28	32	52 *	37
v8	23	21	19	41 *
v9	37	32	46 *	52 *
v10	25	30	33	42 *
v11	16	19	24	30
v12	32	40 *	55 *	37
v13	40 *	56 *	68 *	34
v14	40 *	48 *	61 *	47 *
v15	35	39 *	42 *	53 *
v16	20	19	18	30
v17	16	18	31	23
v18	42 *	51 *	46 *	55 *
v19	30	29	35	39 *
v20	46 *	53 *	45 *	59 *
v21	27	26	35	46 *
v22	18	20	38	28
v23	30	31	44 *	52 *
v24	35	38	48 *	39 *
v25	32	27	39	55 *
v26	34	38	44 *	47 *
v27	24	32	31	27
v28	36	44 *	38	45 *
v29	26	26	33	45 *
v30	33	37	36	40 *
v31	14	24	26	18
v32	27	28	34	37
v33	29	27	40 *	51 *
v34	29	38	42 *	37
v35	38	49 *	56 *	37
v36	36	50 *	34	25
v37	94 *	32	11	20
v38	94 *	36	15	23
v39	48 *	62 *	41 *	37
v40	31	40 *	32	25
v41	29	44 *	38	21
v42	45 *	54 *	46 *	33
v43	37	61 *	28	30
v44	46 *	74 *	33	30
v45	37	60 *	34	33
v46	31	52 *	26	24

Printed values are multiplied by 100 and rounded to the nearest integer. Values greater than 0.388936 are flagged by an '*'.

Preliminary Summary of Results

While we have not pushed the factor analyses beyond this point, or attempted to identify specific factors, we believe that the results support an interpretation of comparability of factor structures for the two modes of administration. This statement applies to all three EOC tests (English II, Algebra I and Biology). For all three tests, over 80% of the items seemed to load on similarly defined factors across the two modes of administration.

If the results tend to support an assumption that the factor structures of the two administration modes are comparable, what does that mean for an overall decision of comparability? This result seems suggestive, but given the apparent differences in the paper/pencil and online samples, it does not seem sufficient for a determination that there is no mode effect. Clearly the paper/pencil and the voluntary online samples were not comparable, and this result confounds our ability to interpret differences in student performance across the modes. Efforts were made to remove some of the demographic differences between the samples and the performance of the two groups moved closer together. However, they still were different. For these samples, the online group tended to perform slightly less well than the paper/pencil group for most of the test items. Sufficient information may still not be there to disentangle the effects of mode and the non-random effects of sample selection. An effort to expand on the matching to include an SES indicator or performance on similar cognitive measures (MAP?) might help in building subsamples of paper/pencil and online examinees that better approximate equivalent groups and could better support an evaluation of comparability.