BEST Plus™ 2.0

Special Report

Item Overlap

The selection of items for administration is particularly important in order to control for item overlap between a pre- and post- test administration. On the computer-adaptive version of BEST Plus, neither the test administrator nor the examinee knows in advance which items will be administered. While it is possible for an examinee to receive the exact same questions in the exact same order on two administrations of the computer-adaptive version, this is highly unlikely. When the print version of BEST Plus is used, the administrator is responsible for ensuring that each examinee receives different forms for the pre- and post-test (per the Test Administrator Guide), and therefore there should be no overlap between test administrations for the print-based version. However, the amount of item overlap between two administrations of the computer-adaptive version of BEST Plus can be empirically investigated.

In the fall of 2014, CAL conducted a study in which 33 examinees took the computer-adaptive version of BEST Plus 2.0 in one room and then were immediately retested in another room. Overlap was examined from two perspectives: folder overlap and item overlap. Twenty-two of the 33 examinees (67%) experienced some degree of item overlap. Table 1 shows the rates of overlap of folders between the two test administrations in this study. (Note: The data in the tables do not include the warm-up items because these are intentionally the same for all examinees.)

Table 1. Rates of folder overlap between two test administrations

Number of Folders That Overlapped	**Number of Examinees**	**% of Total Examinees (N=33)**
0	11	33%
1	10	30%
2	10	30%
3	2	6%
Total with Overlap	**22**	**67%**

At the folder level, 11 examinees (33%) received no common folders across the two administrations, ten examinees (30%) received one folder in common, ten examinees (30%) received two folders in common, and two examinees (6%) received three folders in common across the two administrations. These results demonstrate that it is most likely that an examinee would have no or at most one folder in common across two test administrations (63% of examinees in this study) given back-to-back.

Table 2 shows the rates of overlap of items and folders across which the overlapping items are dispersed between test administrations in this study.

Table 2. Rates of item overlap between two test administrations

Number of Items That Overlapped	**Number of Examinees**	**% of Total Examinees (N=33)**
0	11	33%
1	1	3%
2	6	18%
3	4	12%
4	8	24%
5	2	6%
6	0	0%
7	1	3%
Total with Overlap	**22**	**67%**

At the item level, 11 examinees (33%) received no items in common, one examinee (3%) received one item in common, six examinees (18%) received two items in common, four examinees (12%) received three items in common, eight examinees (24%) received four items in common, two examinees (6%) received five items in common, and one examinee (3%) received seven items in common across the two administrations. These results suggest that it is most likely than an examinee would have no or at most three items in common across two test administrations (66% of examinees in this study) given back-to-back.

An additional analysis was conducted using only data from examinees who experienced some degree of item overlap (22 out of 33 examinees) from the current study. Table 3 presents the descriptive statistics for the 22 examinees across the two computer-adaptive administrations, in which there was some degree of item overlap. It demonstrates the similarity of mean performance in terms of scale scores across the examinees between the first administration of BEST Plus (Test 1) and the second administration of the test (Test 2).

Table 3. Descriptive statistics for computer-adaptive test administrations with item overlap

	N	Mean	Std. Deviation	Min	Max
Computer-adaptive Test 1	22	497.64	69.55	380	659
Computer-adaptive Test 2	22	492.18	61.70	376	619

As could be expected, the correlation between the two performances was high (.85). However, as seen in Table 3, average performances between the first and second administration did not increase. Indeed, they showed a slight decrease. Table 4 presents the scale score difference between the two computer-adaptive test administrations with item overlap. To examine whether there was a statistically significant difference between scale scores on the two computer-adaptive test administrations with item overlap, CAL staff conducted a paired-sample t-test. The results, presented in Table 4, show that the difference between performances on the two test administrations was not statistically significant, suggesting that the administration of a few overlapping items ultimately does not have an impact on observed scores.

Table 4. Paired difference of scale scores for computer-adaptive test administrations with item overlap

	Paired-difference mean	*t*	*df*	*P*
Computer-adaptive Test 1 vs. Computer-adaptive Test 2	5.46	.69	21	.50

It is important to note that the situation in the above test-retest study does not reflect operational test administration at the program level. Programs do not, and should not, administer BEST Plus as a post-test immediately after the pre-test. Nevertheless, the study was for research purposes and the results serve to inform testing programs in terms of the situation in which the rate of item overlap would mostly likely have been the largest across computer-adaptive test administrations (i.e., no growth in student proficiency between administrations).

To examine the degree of item overlap in operational conditions, CAL staff also looked at the degree of item overlap from program data. One operational sample (n=1,353) of tests administered between program years 2011–2013 was examined (henceforth referred to as the 2011-2013 operational data set). The sample drew from testing sites in different locations in the United States that have shared their data with CAL for research purposes. Among this operational sample, 500 examinees had complete pre- and post-test records, and therefore were included in the analysis for item overlap. Table 5 and Table 6 show the rates of overlap by the number of folders and by the number of items, respectively.

Table 5. Rates of folder overlap for pre- and post-administrations in operational programs

Number of Folders That Overlapped	Number of Examinees	% of Total Examinees (N=500)
0	248	50%
1	170	34%
2	63	13%
3	13	3%
4	6	1%
Total with Overlap	**252***	**50%**

Table 6. Rates of item overlap for pre- and post-administrations in operational programs

Number of Items that Overlapped	Number of Examinees	% of Total Examinees (N=500)
1	41	8%
2	69	14%
3	75	15%
4	28	6%
5	20	4%
6	6	1%
7	5	1%
8	2	<1%
9	0	<1%
10	2	<1%
11	1	<1%
Total with Overlap	**249***	**50%**

*Note: the difference between rates of overlap at the folder and item levels suggests that three examinees received folders that overlapped between test administrations, but the items drawn from the folders were entirely different.

As expected, rates of overlap were less in the operational program where there was a time lapse (and potentially growth in student proficiency) than when the two administrations occurred back-to-back. In the operational data, at the folder level, 248 examinees (50%) received no folders in common, compared to only 33% in the back-to-back study. In the operational programs, 170 examinees (34%) received one folder in common, 63 examinees (13%) received two folders in common, 13 examinees (3%) received three folders in common, and 6 examinees (1%) received four folders in common.

At the item level, 251 examinees (50%) received no items in common, whereas, again, only 33% received no items in common when two administrations of BEST Plus occurred back-to-back. In the operational data, 41 examinees (8%) received one item in common, 69 examinees (14%) received two items in common, 75 examinees (15%) received three items in common, 28 examinees (6%) received four items in common, 20 examinees (4%) received five items in common, and about 3% of the examinees received six or more items in common. Overall, these results demonstrate that in operational testing, it is most likely that an examinee would have no or at most one folder in common (84% of examinees in the operational data) and would have no or at most three items in common (87% of examinees) across computer-adaptive test administrations of the BEST Plus.

Based on the research presented here, item exposure does not appear to be a significant influence on BEST Plus performances. Although there was a higher rate of item exposure, as expected, in the two BEST Plus administrations given back-to-back in the Fall 2014 study than in the operational program, that level of item exposure did not appear to influence average test performance. Thus, it is unlikely to be a factor given the lower rates of item exposure that occur in operational testing when the computer-adaptive BEST Plus is used as a pre-test and a post-test.

CAL
CENTER FOR APPLIED LINGUISTICS

www.cal.org

