Reliability Tables from Technical Report 11

Reliabilities of Individual Items. Sixbury and Cashin (1995) reported reliabilities and standard errors of measurement for all items on the original IDEA form. Separate reports were made for classes with 10, 15, 20, 30, and 40 raters. No re-computations were done for these items. Reliabilities and standard errors of measurement were required, however, for all new items as well as for "scales" based on a combination of items (old and new). All computations were based on classes with 15-34 respondents. The average size for such classes was 22.5, and this figure was used to estimate reliabilities for classes of other sizes (10-14; 15-34; 35-49; 50-99; and 100+).

Reliabilities for individual items were estimated by randomly dividing the students in each class. The means for each half were correlated. This produced an estimate of the reliability for a class of 11 (half the average class size). The Spearman-Brown formula was used to estimate reliabilities for the other class sizes. Results are shown in Table 26.

TABLE 26

Reliabilities and Standard Errors of Measurement for "New" IDEA Items

Item Number		Size of Class									
		<15		15-34		35-49		50-99		100+	
New	Old	r_{11}	SE_M	r_{11}	SE_M	r_{11}	SE_M	r_{11}	SE_M	r_{11}	SE_M
1	60	.58	.39	.71	.32	.81	.26	.88	.21	.92	.17
3	64	.49	.39	.62	.34	.73	.29	.82	.23	.88	.19
5	61	.85	.42	.92	.31	.95	.24	.97	.18	.98	.14
9	67	.81	.30	.89	.23	.93	.17	.96	.13	.98	.11
14	62	.75	.42	.85	.33	.91	.26	.95	.20	.97	.16
15	68	.74	.33	.84	.25	.90	.20	.94	.15	.96	.12
16	72	.79	.33	.87	.25	.92	.19	.96	.15	.97	.12
17	69	.78	.30	.87	.23	.92	.18	.95	.14	.97	.11
18	71	.78	.34	.87	.26	.92	.20	.95	.15	.97	.12
20	70	.79	.39	.88	.30	.93	.23	.96	.17	.97	.14
25	50	.84	.36	.91	.27	.95	.20	.97	.16	.98	.12
29	51	.82	.36	.90	.27	.94	.21	.96	.16	.98	.12
30	53	.62	.35	.75	.29	.83	.23	.90	.18	.93	.15
31	55	.62	.36	.74	.30	.83	.24	.90	.19	.93	.15
32	54	.65	.37	.77	.30	.85	.24	.91	.19	.94	.15
40	38	.71	.35	.82	.28	.88	.22	.93	.17	.96	.13
41	44	.79	.31	.88	.23	.93	.18	.96	.14	.97	.11
42	45	.73	.32	.84	.25	.90	.20	.94	.15	.96	.12

Reliability is adversely affected if the range of ratings is constricted; such ranges also tend to reduce the standard error of measurement. Therefore, an item with a relatively high reliability will not necessary have a low standard error of measurement, as seen in Table 26.

In general, the standard error of measurement tended to be the lowest for the "Global outcomes" measures (new items 40-42), averaging .25 for classes in the 15-34 enrollment range. They were more variable for the methods items (items 1-20) than for the items assessing progress on objectives (items 25-32), but the averages for both sets were about the same (.28 and .29, respectively, for class size 15-34).

These estimates are a little larger than those reported by Sixbury and Cashin for the items on the original IDEA form. For a sample of 500 classes enrolling exactly 20 students, the standard errors of measurement for the 20 methods items averaged .247; for the 10 ratings of progress on specific

objectives, they averaged .249. From these results, we anticipate that the standard errors for the new items may be somewhat over-estimated from the pilot results. Of course, as a national database is developed, new and more stable estimates will be made.

Reliabilities of More Complex Measures. As noted earlier, the 20 methods items included on the new form (10 from the original form; 10 from the pilot items) were grouped into five scales representing dimensions of teaching styles: *Student-Faculty Contact* (new form Items 1, 2, and 20); *Involving Students* (new form Items 5, 9, 14, 16, and 18); *Establishing Expectations* (new form Items 3, 4, 8, 13, and 15); *Clarity of Communication* (new form Items 6, 10, and 11); and *Assessment-Feedback* (new form Items 7, 12, 17, and 19). The reliability of scores derived by averaging ratings on these items were computed using data from classes enrolling 15-34 students in the pilot sample. Cronbach's Coefficient Alpha was computed for each scale; results were used to estimate scale reliability for classes averaging 22.5. The Spearman-Brown formula was then employed to estimate reliabilities for classes enrolling 10-15, 35-49, 50-99, and 100+. These figures, together with the standard deviations calculated from classes enrolling 15-34, were used to estimate standard errors of estimate for each size range. Results are shown in Table 27.

TABLE 27

Reliability and Standard Error of Measurement for Methods Scales

Scale	Size of Class									
	10-14		15-34		35-49		50-99		100+	
	r_{11}	SE_M	r_{11}	SE_M	r_{11}	SE_M	r_{11}	SE_M	r_{11}	SE_M
S-Fac. Contact	.75	.21	.85	.16	.90	.13	.94	.10	.96	.08
Involving Students	.78	.21	.88	.16	.92	.13	.95	.10	.97	.08
Estab. Expectations	.83	.20	.91	.15	.94	.12	.97	.09	.98	.07
Clarity Commun.	.78	.21	.88	.16	.92	.13	.95	.10	.97	.08
Assessment/Feedback	.63	.23	.76	.19	.84	.16	.90	.12	.93	.10

As expected, reliabilities were higher, and standard errors lower, for these combinations of items than for the individual items reported in Table 26. These results suggest that the items within the scales were internally consistent, offering reason to conduct further research to establish the meaning and usefulness of the scales.

One other important outcome included on the IDEA Report to the Instructor is that for *Progress on Relevant Objectives.* This measure considers the student progress ratings on objectives chosen by the instructor as "Important" or "Essential." To construct it, the average progress rating on each selected objective was first converted to a "T Score," a standardized score which not only makes it easy to compare a given rating with those for others choosing the objective as "Important" or "Essential" but also puts all progress ratings on a scale with the same mean (50) and standard deviation (10). In computing an average rating for *Progress on Relevant Objectives*, standard scores for "Essential" ratings were given a double weight while those for "Important" objectives received a single weight (and those for objectives not chosen were omitted; i. e., given no weight).

Using data from the pilot sample, reliabilities were computed for both raw and "adjusted" scores for classes enrolling 13-17, 22-29, and 35-49. To do this, students in each class were randomly divided, and the average rating for each of these two halves were correlated. The Spearman-Brown formula was applied to estimate the reliability for classes enrolling an average of 15, 25, and 42 students. It was also

used to project reliabilities for classes of 10-14, 15-34, 35-49, 50-99 and 100 or over. Table 28 summarizes these reliabilities and the associated standard errors of measurement.

As expected, the reliability of this summary measure was generally higher than those for individual progress ratings (Table 26). Standard errors are not comparable, since those in Table 28 are presented in T score, rather than raw score, units. Estimates for adjusted ratings were somewhat less accurate than for raw ratings, reflecting the lack of perfect reliability of measures used to make adjustments. The largest proportion of classes participating in the IDEA program is in the 15-34 range; for such classes, standard errors on this measure were 3.2 and 3.6 T score points for raw and adjusted ratings, respectively.

Table 28

Reliability and Standard Error of Measurement for "Progress on Relevant Objectives" (Raw and Adjusted)

Size of Class	Raw Progress Rating		Adjusted Progress Rating	
	r_{11}	SE_M	r_{11}	SE_M
10-14	.78	4.5	.74	5.1
15-34	.87	3.2	.85	3.6
35-49	.92	2.5	.90	2.7
50-99	.95	1.9	.94	2.2
100+	.97	1.6	.96	1.8