CHAPTER TEN

Reporting and interpreting test results

II. Conceptual exercises

A. Matching

1. f
2. g
3. h
4. j
5. i
6. k
7. a
8. e
9. b
10. c

B. True or false

1. T
2. F
3. F
4. T
5. T
6. T
7. T
8. F
9. F
10. T
11. F
12. T
13. T
14. T
15. F

C. Short answers

1. End-of-course language test

   i. Students

   Raw scores, because students may remember how many items there were, and could interpret the raw score.

   Percentage correct scores for mastery of the content domain.

   Percentile ranks, within each class, because students will be familiar with these from their standardized tests, and these will give them a way to compare their relative standing in their group.

   ii. Director of the program

   Percentile ranks, for the combined classes, because the administrator is not familiar with the tests, and these scores will enable him or her to interpret the relative standings of all the students in the course.

   Percentage correct, to provide information about students' mastery of the course content.

   iii. Parents

   Percentile ranks, within each class, because the parents won't be familiar with the test, and will most likely be familiar with percentile ranks from the standardized test results. Percentile ranks will enable them to understand the relative standing of their children in their class. Also, percentile ranks for the whole course would be helpful, as this would enable parents to understand the relative standing of their child in the whole course.

2. Placing students into a language program—their levels of knowledge appropriate to the course levels

   Need to use a domain-referenced test that is based on the content of the course. The results of this test would provide information about how much of the course material at different levels a given student has mastered.

3. Placing students into a language program—the same numbers of students in the different course levels in the program

   Use a norm-referenced test. This way the administrator could simply use the score distribution to identify cut points that would place the same numbers of students in the different course levels.

4. End-of-semester mastery test

   Use a domain-referenced test based on the course content. Since this is an assessment of students' achievement, we need a test that will provide information about how much of the course materials they have mastered. We would use percentage correct scores.

5. Teacher certification test consisting of three components

   i. Will you use a compensatory or non-compensatory composite score? Why?

   Non-compensatory, because each area is essential, and therefore in order to be certified, a prospective teacher must 'pass' each component at the specified cut score.

   ii. How will you weight the different component scores?

   First, we will calculate the variances and intercorrelations among the three tests. Then, we will use the z-scores and correlations to determine the effective weights of the components. If correlations among the tests are reasonably similar, we will then use the z-scores. If not, then we will divide each score by associated effective weights and multiply them by the corresponding nominal weights.

III. Hand calculations with small data sets

A. Calculate the following NR scores.

1. Calculate the means and standard deviations for the two sets of scores.

	Listening test	Reading test
$\overline{X}$	38.67	40.93
S	6.83	5.60

2. Calculate the percentile ranks, z-scores and linear T-scores for the two sets of test scores.

Listening

X_{list}	**f**	**cumf**	**%ile rank**	**Deviation score**	**z-score**	**Linear T-score**
47	1	15	100	8.33	1.220	62
44	4	14	93	5.33	0.780	58
43	2	10	67	4.33	0.634	56
40	1	8	53	1.33	0.195	52
39	1	7	47	0.33	0.048	50
36	3	6	40	-2.67	-0.391	46
35	1	3	20	-3.67	-0.537	45
29	1	2	13	-9.67	-1.416	36
20	1	1	7	-18.67	-2.734	23
Total	15					

Reading

X_{read}	f	cumf	%ile rank	Deviation score	z-score	Linear T-score
50	1	15	100	9.07	1.620	66
48	1	14	93	7.07	1.263	63
47	1	13	87	6.07	1.084	61
45	1	12	80	4.07	0.727	57
44	1	11	73	3.07	0.548	55
43	2	10	67	2.07	0.370	54
40	3	8	53	-0.93	-0.166	48
39	2	5	33	-1.93	-0.345	47
35	1	3	20	-5.93	-1.059	39
32	1	2	13	-8.93	-1.595	34
29	1	1	7	-11.93	-2.130	29
Total	15					

3. Enter the values in the summary table below.

Summary table for comparing scores

Student	$\mathbf{X_{list}}$	$\mathbf{\%ile_{list}}$	$\mathbf{z_{list}}$	$\mathbf{T_{list}}$	$\mathbf{X_{read}}$	$\mathbf{\%ile_{read}}$	$\mathbf{z_{read}}$	$\mathbf{T_{read}}$
1.	40	53	0.195	52	39	33	-0.345	47
2.	43	67	0.634	56	43	67	0.370	54
3.	44	93	0.780	58	47	87	1.084	61
4.	29	13	-1.416	36	35	20	-1.059	39
5.	47	100	1.220	62	50	100	1.620	66
6.	44	93	0.780	58	45	80	0.727	57
7.	43	67	0.634	56	40	53	-0.166	48
8.	36	40	-0.391	46	39	33	-0.345	47
9.	44	93	0.780	58	48	93	1.263	63
10.	20	7	-2.734	23	43	67	0.370	54
11.	44	93	0.780	58	44	73	0.548	55
12.	36	40	-0.391	46	32	13	-1.595	34
13.	39	47	0.048	50	40	53	-0.168	48
14.	35	20	-0.537	45	40	53	-0.168	48
15.	36	40	-0.391	46	29	7	-2.130	29

B. Looking at the NR scores you have calculated for the two sets of test scores, answer the following questions:

1. Of the students who have the same raw scores on the two tests, how do their NR scores compare?

   Student 2 has the same raw score, 43, on both tests and Student 11 got 44 on both tests.

   Student 2's %ile ranks are the same across the two tests, while the z- and T-scores are higher for the Listening test, indicating that this student performed relatively better on Listening than Reading, with respect to the rest of the class.

Student 1's NR scores were all higher for the Listening test, indicating that this student performed relatively better on Listening than Reading, with respect to the rest of the class.

2. Of the students who have the same NR scores on the two tests, how do their raw scores compare?

   No students got exactly the same z- or T-scores on the two tests. Three students, Nos. 2, 5 and 9, got the same percentile rank, 67, 100 and 93, respectively.

   Student 2's performance was slightly above the means on both tests, and the raw scores were the same—43 on both tests.

   Student 5 got the highest score on both tests, but the raw scores were different—47 and 50 for Listening and Reading, respectively.

   Student 9 got the second highest score on both tests, but the raw scores were different—44 and 48 for listening and reading, respectively.

3. If you wanted to use these tests to provide feedback to your students, which score would be most appropriate? Why?

   We would use both the raw scores and the percentile ranks. We would use the raw scores because the students would probably know what the highest possible score is. We would use the percentile ranks so that they could compare their relative performance on the two tests as well as their relative standing in the group.

4. If you wanted to use these tests to provide feedback to the parents of your students, which score would be most appropriate? Why?

   We would use the percentile ranks because parents are probably familiar with these, and these would let the parents compare their child's relative performance on the two tests, and their relative standing in the group.

5. If you wanted to use these tests to provide information to school administrators about your students' Listening and Reading, which score would be most appropriate? Why?

We would use the percentile ranks because school administrators are probably familiar with these, and these would let them compare students' relative performance on the two tests, and their relative standing in the group.

C. Composite score based on listening comprehension score, speaking grammar rating, and speaking cohesion rating.

1. Calculate the self-weights (SW_i) and multiplier (w_i) for each component.

Listening: $SW_L = (3.754)^2 + .534 + .526 = 15.1525$

$$w_L = \frac{15.1525}{15.1525} = 1.000$$

Speaking grammar: $SW_{SG} = (1.389)^2 + .534 + .917 = 3.3803$

$$w_{SG} = \frac{15.1525}{3.3803} = 4.483$$

Speaking cohesion: $SW_{SC} = (.749)^2 + .526 + .917 = 2.0040$

$$w_{SC} = \frac{15.1525}{2.0040} = 7.561$$

Use the following table to summarize your results:

Component	**SW_i**	**w_i**
Listening	15.1525	1.000
Speaking grammar	3.3803	4.483
Speaking cohesion	2.0040	7.561

2. Using the multipliers obtained above, and nominal weights of 3 for listening, 2 for speaking grammar and 1 for speaking cohesion, calculate the average composite scores based on the effective weights for the three students. Use the table below to facilitate your calculations.

	Listening	**Speaking grammar**	**Speaking cohesion**	**Weighted average**
X_1	9	4.0	3.0	
w_1	1.000	4.483	7.561	
NW_1	3	2	1	
EW_1	3.000	8.966	7.561	$(\sum EW_i = 19.527$
$X_1(EW_1)$	27.000	35.864	22.683	4.38
X_2	20	7.0	4.0	
w_2	1.000	4.483	7.561	
NW_2	3	2	1	
EW_2	3.000	8.966	7.561	
$X_2(EW_2)$	60.000	62.762	30.244	7.84
X_3	4	5.5	3.3	
w_3	1.000	4.483	7.561	
NW_3	3	2	1	
EW_3	3.000	8.966	7.561	
$X_3(EW_3)$	12.000	49.313	24.951	4.42

3. Now calculate the average composite scores based on the effective weights for the three students, using nominal weights of 1 for all three component scores. Use the table below to facilitate your calculations:

Component	**Listening**	**Speaking grammar**	**Speaking cohesion**	**Weighted average**
X_1	9	4.0	3.0	
w_1	1.000	4.483	7.561	
NW_1	1	1	1	
EW_1	1.000	4.483	7.561	$\sum EWi = 13.044$
$X_1(EW_1)$	9.000	17.932	22.683	3.80
X_2	20	7.0	4.0	
w_2	1.000	4.483	7.561	
NW_2	1	1	1	
EW_2	1.000	4.483	7.561	
$X_2(EW_2)$	20.000	31.381	30.244	6.26
X_3	4	5.5	3.3	
w_3	1.000	4.483	7.561	
NW_3	1	1	1	
EW_3	1.000	4.483	7.561	
$X_3(EW_3)$	4.000	24.657	29.951	4.49

4. Use the following table to summarize the results of the three different sets of calculations (unweighted average, average of equally weighted components, average of unequally weighted components):

	Weighted average, based on X_i	**Weighted average, based on EW_i**	**Unweighted average, based on X_i**	**Unweighted average, based on EW_i**
Student 1	6.33	4.38	5.33	3.80
Student 2	13.00	7.84	10.33	6.26
Student 3	4.38	4.42	4.27	4.49

5. Discuss the differences in the composite scores obtained by the four different sets of calculations.

   The first thing we might notice is that the rank order of these three students changes, depending on how we calculate the composite average. Although Student 1 always ranks first, when we use the raw scores as a basis, with both the weighted and unweighted average, Student 1 is second and Student 3 is third. On the other hand, when we use the effective weights to calculate the average, Student 3 is second and Student 1 scored much higher than Student 3 on the component test—Listening—that had the highest self-weight. Since the SW of the Listening scores is five times as large as the next largest SW in any composite based on the raw scores, Student 1 will be higher than Student 3. We can also see that multiplying the raw Listening scores by 3 makes this effect even greater. When we base the composite on the effective weights, the multipliers equalize relative importance of the three component scores, so that the extreme effect of the Listening scores on the composite scores is reduced.

   A second observation we can make is that the magnitudes in the differences among the three students' composite scores is greatly reduced when we base the composite on the effective weights. Thus, with the composites based on the raw

scores, Student 1's score is much higher than that of the other two students, while with the composites based on the effective weights, the difference between Student 1 and the other two students is much smaller.

6. Which composite do you think best represents the test designer's intended weights? Why?

   In this example we can see the effects of averaging raw scores from tests that are obviously on quite different scales. The total scores possible for these three measures were: Listening—20, Cohesion rating—3, Grammar rating—7. These scale differences are reflected in the different standard deviations of these scores, and are in turn reflected in the self-weights. Given these differences, we can see that using the raw scores distorts the designer's intended importance, with the Listening being self-weighted as seven times as important as the Cohesion rating, and five times as important as the Grammar rating. We thus believe that the weighted average based on the effective weights best implements the test designer's intended weights, since this effectively weighs each component according to the test developer's nominal weights.

V. SPSS exercises

A. Use SPSS FREQUENCIES AND DESCRIPTIVES commands to calculate percentile ranks and z-scores, and COMPUTE to calculate linear T-scores for the variables 'sproav' and 'svocav'. Use the data set 'SPAAL93-total scores only.sav' on the CD.

Enter these scores for the first 15 students in the data set, using the table below:

ID	sproav	sprov %ile*	zsproav	Tsproav	svocav	svocav %ile*	zsvocav	Tsvocav
5594163	2.8	20	- .680	43	2.5	25	- .706	43
1986249	3.5	60	.414	54	3.0	50	- .002	50
6222871	3.0	35	- .315	47	2.8	30	- .354	46
6122802	3.5	60	.414	54	2.8	30	- .354	46
6284034	3.5	60	.414	54	3.0	50	- .002	50
5529165	3.5	60	.414	54	3.0	50	- .002	50
5220155	3.3	50	.122	51	3.0	50	- .002	50
6244195	4.0	85	1.143	61	3.8	80	1.055	61
1016295	2.5	15	-1.044	40	2.0	10	-1.411	36
5750752	3.0	35	- .315	47	4.0	90	1.407	64
5631608	4.0	85	1.143	61	4.0	90	1.407	64
5767363	2.0	10	-1.774	32	2.0	10	-1.411	36
6141845	3.5	60	.414	54	3.3	65	.351	54
5583673	3.0	35	- .315	47	2.3	20	-1.058	39
5649776	3.5	60	.414	54	3.0	50	- .002	50

* These are approximate percentile ranks, based on the table above, generated by the DESCRIPTIVE procedure in SPSS. Using this table, we find the percentile rank that is

nearest the score. If a score falls between two scores in the table, then we take the middle percentile. Thus a pronunciation average score of 2.5 is between the table scores of 2.3 and 2.8, with percentile ranks of 10 and 20, respectively, so it has an approximate percentile rank of 15. For multiple scores on the table (in bold), the percentile rank is the average of the percentile ranks associated with the score. Thus, a score of 3.0 on the pronunciation average rating would have an approximate percentile rank of 35, while a score of 3.0 on the vocabulary average rating would have a percentile rank of 50.

1. Of the students who have the same raw scores for pronunciation and vocabulary, how do their NR scores compare?

   Two students, Nos. 5631608 and 5767363, have the same raw scores, 4.0 and 2.0, respectively.

   Student 5631608's NR scores are slightly higher for vocabulary than for pronunciation, so this student performed better on the vocabulary ratings than on pronunciation, relative to the group.

   Student 5767363's percentile ranks are also the same on both measures, while this student's z- and T- scores are slightly higher for vocabulary than for pronunciation. Given the similarities of the NR scores this student performed about the same on both measures, relative to the group.

2. Of the students who have the same NR scores for pronunciation and vocabulary, how do their raw scores compare?

   Two students, Nos. 5220155 and 5767363, have the same percentile ranks on both measures. Student 5220155's raw scores are different, while student 5767363's raw scores are the same.

   Two students, Nos. 5594163 and 6244195, have the same linear T-scores on both measures. For both students, their average ratings are very close, differing by only .3 and .2.

3. If you wanted to use these tests to provide feedback to the students, which score would be most appropriate? Why?

   If the students know the scales that the average ratings are based on, we would report these, since these might be linked to level descriptors that would be

informative to the students. We would also report the percentile ranks, as these provide a basis for students to compare their performance on the two ratings, relative to the group. In this case, we would report both a CR score (level descriptor) and an NR score.

4. If you wanted to use these tests to provide feedback to Spanish instructors, which score would be most appropriate? Why?

   We would report the average ratings, relating these to level descriptors, which would give instructors a sense of the levels of pronunciation and vocabulary. We would also report the percentile ranks, in case instructors wanted to know the relative standing of their students on the two measures, relative to the entire group. In this case, we would report both a CR score (level descriptor) and an NR score.

5. If you wanted to use these tests to provide information to EAP administrators about students' pronunciation and vocabulary, which score would be most appropriate? Why?

   EAP administrators might be interested in both a CR indicator of levels of speaking pronunciation and oral vocabulary use, and in being able to compare the relative standing of students from different campuses in the UC system. Thus, we would report both the average ratings, relating these to level descriptors, and the percentile ranks, which would enable administrators to compare the relative standings of students from different campuses.