

## CHAPTER FOUR

### Analyzing test tasks

#### II. Conceptual exercises

##### A. Matching

1. e
2. d
3. g
4. f
5. i
6. h
7. c
8. b

##### B. True or false

1. T
2. T
3. T
3. F
5. F
6. T
7. T
8. F
9. T
10. F

##### C. Brief responses

1. The three main purposes of classical item analysis:

- (1) Diagnostic feedback to test takers on how they performed on individual test tasks;
- (2) feedback to teachers and course developers relevant to the improvement of instruction;
- (3) feedback to test developers and test writers, to improve the usefulness of the test, enabling the test developer.

2. '*Rules of thumb*' for selecting test items in an NR test:

Select items that fall between a range of p-values between .25 and .75, or .20 and .80 and item discrimination indices equal to or greater than .30. In addition, select items that fit the content specifications of the test. Finally, use multiple sources of information in making a decision about items—quantitative and qualitative.

3. 'Rules of thumb' for selecting test items in a CR test:

Select items within a relatively narrow range of p-values around the percentage cut score and CR discrimination indices that are greater than zero. In addition, select items that fit the content specifications of the test. Finally, use multiple sources of information in making a decision about items—quantitative and qualitative.

4. Main limitations of classical item analysis:

First, an important limitation is that the item and score statistics obtained from tests developed with classical item analysis are essentially *sample-based descriptive* statistics. On the one hand, the item statistics that we obtain are dependent on the particular group, or sample, of test takers who take the test, while on the other hand, the test scores the test takers receive are dependent upon the particular set of *sample items* that make up the test. The dependency of item statistics on a specific sample of test takers means that if the test items are administered to different groups of test takers, the items may have different statistical characteristics. Thus, item statistics obtained through classical item analysis cannot be easily compared to item statistics based on the performance of different groups of test takers, and used to compare test takers whose scores are obtained from different tests.

A second important limitation is that the item statistics obtained with such analysis do not take into consideration the level of ability of any particular test taker. That is, classical item analysis makes no connection between the difficulty of a given item and the ability of a particular test taker. The classical difficulty index,  $p_i$ , tells us how a given *group* of individuals performed *on average* on a particular item. Thus, the only information we have for predicting how a given test taker will perform on a particular item is the average performance of a group on that item. Similarly, a test score tells us how well a given test taker did, *on average*, on a particular *set* of items. That is, a test taker's score is the number received out of some total possible score. The score that a test taker receives on a given test will depend on the difficulty level of that test. If different tests are of different levels of difficulty, then a test taker may appear to be of average ability on one test and high ability

on the other. What this means is that the inferences about ability that we make on the basis of test scores will depend on the particular set of items that make up the test.

Finally, classical item analysis allows us to look at only one aspect of the measurement procedure—the item. However, in many testing situations there are other aspects that we also need to investigate for purposes of test design. In an oral interview, for example, test takers may interact with different examiners, may be presented with multiple prompts, and their responses may be rated by different raters. In situations such as this, it would be useful to be able to obtain information about not only the relative difficulty of the different prompts, but also about the relative severity of the different examiners and raters.

5. Three important advantages of Item Response Theory over classical item analysis:

According to Hambleton and Swaminathan, the three advantages are (1) item parameter estimates are independent of the group of examinees. The  $a$  and  $b$  parameters (like item difficulty and item discrimination estimates in classical item analysis) are useful in practical test development; (2) test taker ability estimates are independent of the particular set of test items used; and (3) precision of ability estimates for specific levels of abilities is also available.

In addition, a practical application of IRT is in computer-adaptive testing. In this form of testing, a test taker is given a middle difficulty item to begin with, and if she gets this correct, the next item is slightly more difficult. If she continues to get the items correct, she is presented with increasingly difficult items. If she gets an item wrong, she is presented with a slightly easier item, and so on. This form of test administration, in which each test taker takes a slightly different set of items depending on how they respond, has the advantage that the test can be tailored to each individual test taker, for maximum efficiency.

### III. Hand calculation exercises with small data sets

A. Data set 1: NR Multiple-choice test; N=75

1. Tally the responses
2. Calculate the  $D_i$  value

Item No.	Multiple choices	Upper group N=20 27 %	Tally number ( $R_U$ )	$p_u$	Lower group N=20 27 %	Tally number ( $R_L$ )	$p_l$	$p_i$	$D_i$
1	A	<del>////</del>			<del>////</del>	5	.25	.13	-.25
	B	<del>////</del>	5	.25	<del>////</del>	5	.25	.25	.00
	C*	<del>////</del> <del>////</del>	15	.75	<del>////</del> <del>////</del>	10	.50	.63	.25
	D							.00	
2	A*	<del>////</del> <del>////</del> //	12	.60	<del>////</del> <del>////</del> <del>////</del>	15	.75	.68	-.15
	B	<del>////</del> //	7	.35	<del>////</del>	5	.25	.30	.10
	C							.00	
	D	/	1	.05				.03	.05
3	A	<del>////</del>	5	.25	<del>////</del> //	7	.35	.30	-.10
	B	//	2	.10	<del>////</del>	3	.15	.13	-.05
	C	///	3	.15	//// //	7	.35	.25	-.20
	D*	<del>////</del> <del>////</del>	10	.50	///	3	.15	.33	.35
4	A	<del>////</del>	5	.25	//	2	.10	.18	.15
	B*	<del>////</del> <del>////</del>	10	.50	<del>////</del> <del>////</del> <del>////</del> //	17	.85	.68	-.35
	C	<del>////</del>	5	.25				.13	.25
	D				/	1	.05	.03	-.05

\* Key or correct answer

Bachman and Kunnan, Workbook for *Statistical Analyses for Language Assessment*  
Answers to Exercises

Fill in the following values in the Summary Table:

Item No.	Upper group ( $p_u$ )	Lower group ( $p_l$ )	Total group ( $p_i$ ) $D_i$		Comments
1	A= .00	A= .25	A= .13	A= -.25	Good distractor; p value ok; negative discrimination; no one in U group chose this option
	B= .25	B= .25	B= .25	B= .00	Poor distractor; same % of U & L groups chose this option
	C*= .75	C*= .50	C= .63	C= .25	Relatively easy item; Poor distractor; marginal discrimination
	D= .00	D= .00	D= .00	D= 00	No one chose this option
2	A*= .60	A*= .75	A= .68	A= -.15	Relatively easy item
	B= .35	B= .25	B= .30	B= .10	Poor distractor; more % in U group than L group chose this option
	C= .00	C= .00	C= .00	C= 00	Poor distractor; no one chose this option
	D= .05	D= .00	D= .03	D= .05	Poor distractor; very few chose this option
3	A= .25	A= .35	A= .30	A= -.10	Good distractor; negative discrimination
	B= .10	B= .15	B= .13	B= -.05	Ok distractor
	C= .15	C= .35	C= .25	C= -.20	Good distractor; negative discrimination
	D*= .50	D*= .15	D= .33	D= .35	Difficult item; good discrimination
4	A= .25	A= .10	A= .18	A= .15	Poor distractor; more in U than L group chose this option
	B*= .50	B*= .85	B= .68	B= -.35	Relatively easy item but negative discrimination
	C= .25	C= .00	C= .13	C= .25	Poor distractor; more in U group chose this option
	D= .00	D= .05	D= .03	D= -.05	Poor distractor; very few chose this option

--	--	--	--	--	--

3. Decisions regarding selecting, rejecting or revising of the four test items:

Item 1:

This item is relatively easy for this group. The group-wise p-values and  $D_i$  statistics show that the item is generally acceptable. However, B and C are poor distractors and no one chose distractor D; these distractors may need to be replaced. A final decision should only be made based on a review of the actual item and the content specifications that were used to write the item.

Item 2:

This item is relatively easy for this group but there are some problems. The p-values show that the item has acceptable values but the  $D_i$  statistic is negative as a higher percentage of test takers from the lower group got this item correct when compared to the upper group. In addition, B, C and D are all poor distractors. These distractors may need to be replaced and a thorough review of the actual item and the content specifications that were used to write the item are recommended.

Item 3:

This item is a difficult item although the group-wise p-values and the  $D_i$  statistic are acceptable. A and C are good distractors but very few chose option B. A review of the actual item and the content specifications that were used to write the item are recommended before a final decision is made.

Item 4:

This item is quite problematic: first, the group-wise p-values indicate that the upper group has performed quite poorly when compared to the lower group; as a result, the  $D_i$  statistic is negative. Second, A, C and D are poor distractors; these distractors may need to be replaced.

Therefore, a thorough review of the item and the content specifications that were used to write the item are recommended.

Bachman and Kunnan, Workbook for *Statistical Analyses for Language Assessment*  
Answers to Exercises

**B. Data set 2: CR test with scores Before and After instruction (Pre- and Post-Instruction)**

Item	1		2		3		4		Total	
Student	Before	After	Before	After	Before	After	Before	After	Before	After
1	0	1	0	0	1	1	1	1	2	3
2	0	1	1	1	1	1	1	1	3	4
3	0	1	1	0	0	1	1	1	2	3
4	0	1	0	1	0	0	1	1	1	3
5	0	1	0	0	1	0	1	1	2	2
6	0	1	0	0	0	0	1	1	1	2
7	0	1	0	0	1	0	1	1	2	2
8	0	1	0	0	1	1	1	1	2	3
9	0	0	0	1	1	0	1	1	2	2
10	0	0	0	1	1	1	0	0	1	2
$\Sigma x_i$	0	8	2	4	7	5	9	9		
$p_i$	.00	.80	.20	.40	.70	.50	.90	.90		
DIS <sub>PPD</sub>	.80		.20		-.20		0.00			



3. Decisions regarding selecting, rejecting or revising of the four test items:

Item 1:

This item has appropriate item statistics ( $p_i$  and  $DIS_{PPD}$ ) and could be considered for selection in a CR test as long as it meets the content specifications that were used to write the item.

Item 2:

This item seems to be a difficult item; instruction did not improve the performance much. But the item statistics are acceptable. A thorough review of the actual item and the content specifications used to write the item are recommended.

Item 3:

This item is not suitable based on the item statistics. The performance 'after instruction' was not even as good as the 'before' performance. This is also reflected in the  $DIS_{PPD}$  statistic.

Item 4:

This item may not be an appropriate item because it was too easy before instruction and therefore cannot provide any new information about the test takers who take the item after instruction. If the content specifications require this item, a review of this item and the content specifications should be conducted.

## V. SPSS exercises

### 3. Computing the reliability of the Listening and Reading section of the ESLPE96

#### A. Listening section item and scale statistics (Items 1–30)

<b>Listening section of the ESLPE96</b>		
Number of cases	656	
Number of items in the scale	30	
Alpha	.723	
Mean for scale	22.35	
Standard deviation	4.10	
Two easiest items	Item: 28	Item: 2
p-values	.951	.930
SD	.216	.255
Item-total correlation	.356	.155
Two most difficult items	Item: 17	Item: 25
p-values	.398	.445
SD	.490	.497
Item-total correlation	.040	.212
Other items with low item-total correlation (below .200)	Items: 2 (.154) 4 (.174) 7 (.139) 12 (.125) 13 (.162) 19 (.117) 20 (.141) 23 (.161) 30 (.105)	

#### B. Comments on the Listening section

Based on the statistics above, the Listening section has high internal consistency (.723) for the 30 items and for this sample. The mean indicates that the score distribution is negatively

skewed and the standard deviation is low which means that this sample group is generally homogenous in this test section. The two easiest items have high p-values (above .900) and the two most difficult items have low p-values (below .500). The item-total correlations indicate that 10 items do not contribute towards test discrimination. In general, this test section (with the items) is acceptable if a review of the actual items and test specifications confirm the acceptability of the items.

C. Reading section item and scale statistics (Items 31-70)

<b>Reading section of the ESLPE96</b>		
Number of cases	656	
Number of items in the scale	40	
Alpha	.845	
Mean for scale	29.428	
Standard deviation	6.438	
Two easiest items	Item: 56	Item: 36
p-values	.921	.901
SD	.270	.299
Item-total correlation	.457	.261
Two most difficult items	Item: 67	Item: 32
p-values	.354	.521
SD	.479	.500
Item-total correlation	.340	.156
Other items with low item-total correlation (below .200)	Items: 38 (.179) 43 (.173) 45 (.184) 61 (.142)	

D. Comments on the Reading section of the ESLPE96

Based on the statistics above, the Reading test section has high internal consistency (.845) for the 40 items and for this large sample. The mean indicates that the score distribution is negatively skewed and the standard deviation is low which means that this sample group is generally homogenous in these two abilities. The two easiest items have high p-values (above .900) and one of the two most difficult items has low p-values (below .500). The item-total correlations indicate that a few items do not contribute towards test discrimination. In general, this test section (with the items) is acceptable if a review of the actual items and test specifications confirm the acceptability of the items.

E. Evaluating the Listening section of the ESLPE96 (Items 1-20)

1. Are the items appropriate from the point of view of the purpose of the test?

As the purpose of the ESLPE is to place or exempt undergraduate or graduate students who have been admitted to UCLA into appropriate ESL classes, the Listening section test items have to be evaluated in this context. The first passage, 'Lecture 1 Music History', is a mock first-day introduction to the course. The lecture includes lots of detailed information about the course (such as the instructor's phone numbers, the dates for the midterm, quizzes, the types of papers, etc.) and a brief introduction to the subject of the course. The test items that follow require the ability to remember lots of information provided in the lecture and a general understanding of the subject matter of the course.

Generally speaking, it appears that the lecture material is appropriate to undergraduate and graduate students although whether the subject matter (music history) and cultural context (European culture) are appropriate for the purpose of the test can only be judged after conducting studies based on test performance (for example, this can be done through differential item functioning analyses that check for differences in performance in the Listening section by different academic major groups and salient cultural groups). In terms of the test items, an analysis of the types (and number) of items that deal with different listening abilities needs to be conducted so that a list of abilities tested can be used to check for balance and appropriacy in terms of the purpose of the test (for example, this

can be done by comparing the skills tested in the test with a checklist or taxonomy of Listening micro-skills; Richards (1983, cited in Buck, 2001, pp. 56–57) proposes such a taxonomy for academic listening although not all skills may be appropriate for the context of this test).

## 2. What are the descriptive statistics?

Here are the descriptive statistics for the 20 items.

<b>Listening section of the ESLPE96—(Items 1-20)</b>	
Number of cases	656
Number of items in the scale	20
Alpha	.623
Mean for scale	14.652
Standard deviation	2.933

Although the mean and standard deviation indicate that the 20-item set is acceptable, the alpha estimate indicates that the internal consistency reliability is not very high. A further examination of individual items should provide more information.

## 3. Comment on the easiest and most difficult items

### a. Two easiest items

Two easiest items	Item: 2	Item: 8
p-values	.930	.906
SD	.256	.293
Item-total correlation	.137	.276

The two easiest items for this test-taking group are Item 2 and Item 8. Item 2 is a general question on the subject matter of the course in a True–False format; in Richards’ taxonomy, ‘ability to identify scope of lecture’. Item 8 is about whether there is more Reading than Listening in the course in a True–False format; in Richards’ taxonomy, ‘ability to identify

specific information’. The first question is easy because of its general nature and the True–False format; the standard deviation is also low, indicating that the variation on this item is low. The second question is specific but as listening activities are mentioned many times in the lecture, it turns out to be easy. In terms of the item-total correlation (item discrimination), Item 2 does not seem to contribute much. Based on these statistics and the skills tested, Item 2 could be dropped unless an easy item with this skill is essential. Item 8, on the other hand, seems a useful item to the Listening section.

b. Two most difficult items

The two most difficult items	Item: 17	Item: 10
p-values	.398	.527
SD	.490	.500
Item-total correlation	.029	.274

The two most difficult items are Item 17 and Item 10. Item 17 requires a specific piece of information that is mentioned just once: that Leipzig is the most important musical center; ‘ability to identify specific information’. Item 10 too requires a specific piece of information about what is the most important information of the first paper; ‘ability to identify specific information’. These two items assess the same ability and appear to be difficult and have very high standard deviation. In terms of the item-total correlation, Item 17 does not contribute at all while Item 10’s contribution is satisfactory. Based on these statistics and the skills tested, Items 17 and 10 could be dropped without loss of information as there are many other items that test similar skills and have better item statistics.

c. Other items

Item-total correlation (below .200)	Items: 4 (.168) 7 (.154) 12 (.102) 13 (.145) 19 (.110) 20 (.145)
--	---

The other 15 items related to this lecture of the Listening section seem appropriate in terms of item statistics although in terms of item-total correlation statistics, six items have values below .200. Items 12 and 19 have the lowest estimates and need to be reviewed: Item 12 (Mean=.825; SD=.381) and Item 19 (Mean=.742; SD=.438). Both items test the 'ability to identify specific information' and both are relatively easy. So, the items could be retained although they do not contribute much by way of item discrimination. The other items may be reviewed in a similar manner.

In terms of the skills tested, most of the items test the 'ability to identify specific information'. An analysis of the skills tested in the section needs to be conducted so that the appropriate balance of skills can be achieved for the purpose of the test. [Similar analyses of the other ten items in the Listening section (based on the 'Greenhouse' lecture) should be done so that the information obtained from the 40 test items is appropriate for the purpose of the test.]

#### F. Evaluating the Reading section of the ESLPE96 (Items 31-50)

##### 1. Are the items appropriate from the point of view of the purpose of the test?

As the purpose of the ESLPE is to place or exempt undergraduate or graduate students who have been admitted to UCLA into appropriate ESL classes, the Reading section test items have to be evaluated in this context. The first passage – Passage 1 'Scarcity Choice' is an introduction to the topic. The passage includes definitions of concepts and a discussion of relevant issues, typical of introductory reading material in undergraduate or graduate studies. The test items that follow require the ability to 'draw inferences from content', 'draw inferences about the meaning of words in context', 'draw inferences from author's purpose and attitude', 'understand conceptual meaning', and 'understand explicitly stated and unstated information'.

Generally speaking, it appears that the Reading passage is appropriate to undergraduate and graduate students although whether the subject matter (economics) is appropriate for the purpose of the test can only be judged after conducting studies based on test performance (for example, this can be done through differential item functioning

analyses that check for differences in performance in the Reading section by different academic major groups). In terms of the test items, an analysis of the types (and number) of items that deal with different reading abilities needs to be conducted so that a list of abilities tested can be used to check for balance and appropriacy in terms of the purpose of the test (for example, this can be done by comparing the skills tested in the test with a checklist or taxonomy of Reading micro-skills; Davies (1968) and Munby (1978), cited in Alderson (2000, pp. 9–11; 93–97), propose such a taxonomy for reading although not all skills may be appropriate for the context of this test).

## 2. What are the descriptive statistics?

Here are the descriptive statistics for the 20 items.

<b>Reading section of the ESLPE96—reliability (Items 31-50)</b>	
Number of cases	656
Number of items in the scale	20
Alpha	.688
Mean for scale	15.017
Standard deviation	3.196

Although the mean and standard deviation indicate that the 20-item set is acceptable, the alpha estimate indicates that the internal consistency reliability is not very high. A further examination of individual items should provide more information.

## 3. Comment on the easiest and most difficult items.

### a. Two easiest items

Two easiest items	Item: 36	Item: 43
p-values	.901	.887
SD	.299	.317
Item-total correlation	.312	.207



The two easiest items for this test-taking group are Item 36 and Item 43. Item 36 requires a specific piece of information in a multiple-choice format; the ability tested is to ‘understand explicitly stated information [see first sentence in Paragraph 6]’. Item 43 is a general question on the subject matter of the passage in a True–False format; the ability tested is to ‘draw inferences from content’. Both questions are easy because of the general nature of the items; the standard deviation is not too high, indicating that the variation on this item is quite low. In terms of item-total correlation, Item 43’s contribution is low. Based on these statistics and the skills tested, Item 36 can be retained and Item 43 could be dropped.

b. Two most difficult items

Two most difficult items	Item: 32	Item: 47
p-values	.521	.537
SD	.500	.499
Item-total correlation	.131	.242

The two most difficult items are Item 32 and Item 47. Item 32 requires the ability to ‘recognize a writer’s purpose’ but the item is presented in a somewhat confusing manner, with a long, complex statement. This may have accounted for the difficulty of the item. Item 47 requires the ability to ‘draw inferences about the meaning of words in context’; the specific word is ‘channeled’ with four multiple-choice answers. Both items also have high standard deviation indicating that there is much variation on these items. In addition, Item 32 has very low item-total correlation. Based on these statistics and the skills tested, Item 32 does not contribute much, and could be dropped without loss of information as there are many other items that test similar skills and have better item statistics. Item 47 too could be dropped unless there is interest in retaining it as one of the four ‘words in context’ items.

c. Other items

The other 16 items related to this passage of the Reading section seem appropriate in terms of item statistics. Only Item 38 has an item-total correlation below .200; the item requires the ability to ‘draw inferences about the meaning of words in context’. An analysis of the

skills tested in this section needs to be conducted so that the appropriate balance of skills can be achieved for the purpose of the test.

Item-total correlation (below .200)	38 (.168)
--	-----------

[Similar analyses of the other 20 items based on the second passage ('Class before birth') should be done so that the information obtained from the 40 test items is appropriate for the purpose of the test.]

G. Recomputing reliability of the test sections

1. Listening section statistics recomputed after dropping Items 2, 10, & 17

<b>Listening section of the ESLPE96</b>	<b>Original statistics</b>	<b>Recomputed statistics</b>	
Number of cases	656	656	
Number of items in the scale	20	17	
Alpha	.623	.615	
Mean for scale	14.652	12.797	
Standard deviation	2.933	2.650	
Two easiest items		Item: 8	Item: 14
p-values		.906	.895
SD		.293	.307
Item-total correlation		.270	.320
Two most difficult items		Item: 4	Item: 16
p-values		.558	.564 .49
SD		.497	6
Item-total correlation		.159	.288
Other items with low item-total correlation (5 items with values below .200)		Items: 7 (.153) 12 (.087) 13 (.155) 19 (.106) 20 (.137)	

2. Reading section statistics recomputed after dropping Items 32, 43, & 47.

Reading section of the ESLPE96	Original statistics	Recomputed statistics	
Number of cases	656	656	
Number of items in the scale	20	17	
Alpha	.688	.677	
Mean for scale	15.017	13.072	
Standard deviation	3.196	2.849	
Two easiest items		Item 36	Item 37
p-values		.901	.875
SD		.299	.331
Item-total correlation		.327	.338
Two most difficult items		Item 42	Item 33
p-values		.654	.680
SD		.476	.467
Item-total correlation		.239	.236
Other items with low item-total correlation (2 items with values below .200)		Items: 38 (.177) 46 (.183)	

3. Are your test sections better than Listening and Reading test sections now? What are your reasons?

The parts of the two test sections that were examined carefully now have better test items in general: the easiest items have p-values near or above .900 and the most difficult items are now above .500. The standard deviations continue to be quite high for the most difficult items but the item-total correlations are not as low now. The overall alpha estimate for both tests did not improve; in fact it dropped a bit (for Listening from .623 to .615 and for

Reading from .688 to .677), largely due to the 10-15 per cent reduction in test length for both tests.

[Looking ahead to Chapter 5 in the book, we could use the Spearman–Brown prophecy formula (Equation 5.8) and substitute coefficient alpha for the split-half estimate to estimate the reliability of the revised tests if they were 20 items long. This procedure yields an estimate of .654 for Listening and .712 for Reading. Thus, if we were to add another three items of comparable quality to these tests, the reliability would be improved.]

However, the Listening section still seems to have a few problems: a few items (7, 12, 13, 19, 20) still have very low item-total correlation values. These items need to be reviewed along with Item 4, which is one of the most difficult items. The Reading part seems in quite good shape; the only interest might be to review Items 38 and 46. Once these reviews are completed, the other items in the test sections (Lecture 2 and the 10 related items in the Listening section, and Passage 2 and the 20 related items in the Reading section) need to be reviewed for a fully complete evaluation of these test sections. Finally, we would like to emphasize that when these item reviews are conducted, both statistical specifications as well as content specifications should be used in evaluating the lectures/passages and related items.