DOCUMENT RESUME

ED 267 645 FL 015 627

AUTHOR Clark, John L. D.
TITLE A Study of the Comparability of Speaking Proficiency Interview Ratings across Three Government Language Training Agencies.
INSTITUTION Center for Applied Linguistics, Washington, D.C.
PUB DATE Jan 86
NOTE 58p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Comparative Analysis; Federal Government; French; German; *Interrater Reliability; *Interviews; *Language Proficiency; *Language Tests; Public Agencies; *Rating Scales; Speech Communication; *Test Reliability; Test Results
IDENTIFIERS Central Intelligence Agency; Defense Language Institute CA; Foreign Service Institute DC

ABSTRACT

A study of the reliability of the proficiency ratings scale and techniques used by three federal government agencies--the Central Intelligence Agency, the Defense Language Institute, and the Foreign Service Institute (FSI)--to test employees' oral language proficiency in French and German had two randomly selected two-person teams of testers from each agency test 20 subjects for each language. The ratings assigned to the subjects were compared with an expected rating distribution. Results indicated no statistical difference in the ratings across agencies, either for the combined languages or for each language separately. However, ratings in various sub-portions of the proficiency scale showed clear across-agency differences, generally reflecting relatively higher ratings on the part of FSI raters and, occasionally, wide discrepancies in scoring for individual examinees. Findings also indicated that, despite the feeling that their language proficiency had been adequately probed, a significant number of examinees felt that the FSI procedure was more anxiety-producing than the others and that it used some "unfair" eliciting techniques. It is concluded that, although the study examines only the test-retest comparability of the interviewing process across agencies, there is a need for further research if the procedure is to be used as an inter-agency measure. (MSE)

A Study of the Comparability of Speaking Proficiency Interview Ratings Across Three Government Language Training Agencies

John L. D. Clark

Center for Applied Linguistics

January 1986

1118 22nd Street, N.W.

Washington, DC 20037

A Study of the Comparability of Speaking Proficiency Interview Ratings Across Three Government Language Training Agencies

John L. D. Clark
Center for Applied Linguistics[1]

BACKGROUND

A pervasive question in the operational use and interpretation of the results of speaking proficiency interviews based on the "ILR" (Interagency Language Roundtable) proficiency level descriptions is the extent to which given examinees' performances would be evaluated in a similar manner across the variety of government agencies and other institutions that make use of this testing procedure. Although there has been a fair amount of conjecture and internal discussion of this topic on the part of examiners and administrators involved in the-day-to-day implementation of agency testing programs, there has not until recently been an opportunity to address the "comparability-of-rating" question in a straightforward empirical manner.

The following is a description of the procedures and major results of a direct experimental comparison of the proficiency ratings assigned to a common group of examinees by testers in each of three government language training agencies: CIA, DLI, and FSI; for each of two languages: French and German. Also discussed are the extent to which the results of this particular study might legitimately be extrapolated, cautions on areas in which extrapolation would not be appropriate, and recommendations for follow-up investigation of other aspects of reliability and validity of the interview testing process not formally addressed in the present study.

---

[1]The assistance of a number of other persons in the conduct of this study is gratefully acknowledged. Among the CAL staff, Lynn E. Thompson provided very effective administrative assistance during all phases of the project, Christina Garbacz had major responsibility for data entry as well as for various aspects of statistical processing, and Rebecca Oxford contributed substantially to project planning and procedures specification. Nina Levinson (CIA), Thea Bruhn (FSI), and Ellen Mitchell and Phillip White (DLI) coordinated the interviewing activities at their respective agencies with a high level of diligence and effectiveness, and a debt of appreciation is owed the many interviewees in the study who provided, on a voluntary basis, the time and personal interest needed to participate willingly in the interviewing process on three separate occasions. Finally, the major expression of appreciation must be reserved for the certified testers at each of the three agencies, who maintained throughout the six days of testing a seriousness of purpose and diligence of approach to their interviewing and rating tasks that fully demonstrate their high level of professionalism and competence in the proficiency testing role.

PROCEDURE

Overall study design. The basic experimental design for the study involved a "test-retest" procedure, in which each examinee was sequentially interviewed by a separate testing team from each of the three participating agencies. In conducting its own interviews, each team made use of the particular interviewing techniques and procedures for arriving at a final rating that were currently in use at that agency. On completion of the process, the team reported a single overall rating on the numerical proficiency scale and associated verbal descriptions of performance endorsed by the ILR member agencies in November 1981 as a "common metric" for speaking proficiency assessment and reporting. This scale comprises six major ratings--0, 1, 2, 3, 4, and 5--supplemented by five intermediate ("plus") ratings --0+, 1+, 2+, 3+, and 4+. The scale is intended to characterize the full range of possible learner proficiency levels, from no functional proficiency in the language (level 0) to proficiency indistinguishable in all respects from that of an educated native speaker (level 5).

Within the administrative and financial constraints involved, it was obviously not possible to carry out such a study for each of the numerous languages in which the agencies routinely test, nor, within a given language, to involve each and every one of the examiners/testers currently conducting interviews within that language. With regard to the selection of languages for the study, discussion with the testing coordinators at each of the three agencies, as well as with the ILR testing subcommittee, resulted in the identification of French and German as two languages for which an adequate number of examinees and testers for the study could be made available within each of the participating agencies, and for which the annual testing volume was sufficiently high to warrant priority attention from an administrative standpoint. With respect to the number of tester teams involved, staff time and travel cost considerations dictated a maximum of two teams per language, for each of the three agencies, i.e., the following configuration:

CIA French Team 1 (all teams are two-person)
CIA French Team 2

DLI French Team 1
DLI French Team 2

FSI French Team 1
FSI French Team 2

CIA German Team 1
CIA German Team 2

DLI German Team 1
DLI German Team 2

FSI German Team 1
FSI German Team 2

Selection of testers and examinees. In order to enhance the likelihood

that, for each agency and language, the testers actually selected for the study would be representative of the total group of individuals operationally testing in that agency/language, the testing coordinator at the agency was asked to provide a complete list of qualified, currently active testers in each language. From this list, the CAL project staff selected all study participants on a statistically random basis. It is thus considered that the composition of the tester groups from each of the three agencies constituted a rigorous random sampling of the population of testers in that language who had been identified by the agency as properly qualified and actively testing within that agency.

A second important design consideration was the selection of examinees. It was considered highly desirable by the project staff, as well as by the ILR testing committee, to investigate rating performance across the full range of proficiency levels by including in the examinee pool individuals covering the gamut from the lowest measurable level (0+) to the functional equivalent of an educated native speaker (5). At the same time, in view of the fact that the bulk of operational testing at each of the agencies is concentrated within a somewhat smaller band (roughly 1 to 3/3+ for CIA and DLI, 2 to 4 for FSI), it was considered important to insure that a reasonably large number of examinees within this "higher-volume" range would be included in the study sample. To help provide a distribution of examinees for the study that would satisfy both of these criteria, the testing coordinators at each agency were asked to locate and arrange for the participation, per agency, of 20 examinees in each language, and to select these individuals--on the basis of coordinator or language instructor judgments about their proficiency and/or recent interview scores in the agency files--so as to reflect as closely as possible the distribution of proficiency levels shown in Figure 1.

The coordinators were asked to employ, to the extent possible, a stratified random sampling procedure (for which detailed instructions were given) in identifying the particular examinees who would be asked to participate. The total pool from which the examinees at a given agency were to be drawn was defined to include, in addition to currently-enrolled students, other categories of individuals that the agency would typically have the occasion to test in the course of its ongoing testing activities (for example, instructor applicants at DLI, career officers at FSI). Due to a variety of factors, including scheduling conflicts on the part of potential examinees, the necessarily voluntary nature of participation, and the need to locate substitute interviewees on several occasions during the course of the testing, it was not possible to rigorously implement a statistically random process of examinee selection. However, since the major intent in selecting examinees was simply to provide an appropriate overall distribution of proficiency levels across examinees at each agency, departure from strict random selection of the examinee group was not considered a significant procedural drawback nor an impediment to the proper interpretation of the tester-specific information on which the study was primarily focused.

Scheduling of interviews. Interviews were conducted on a sequential basis, with two days of testing taking place at each agency. Testing dates were: FSI - September 9-10, 1985; CIA - September 11-12; DLI - September 17-18. On each of these dates, the "home" agency made available all necessary interviewing rooms and other facilities and was responsible for scheduling and contacting the examinees to be tested at that agency by all three tester groups.

Figure 1

Intended Distribution of Examinee Proficiency Levels at Each Agency

Level	No. of Examinees
0+	1
1	1
1+	2
2	3
2+	3
3	3
3+	3
4	2
4+	1
5	1
	20

Project staff forwarded the testing coordinator at each agency a detailed schedule (See Figure 2) for allocating examiners to testing teams in such a way as to counterbalance the agency-order in which the examinees would be tested as well as to statistically randomize other (uncontrolled) effects attributable to examinees. The test administration schedule, which was followed with only very minimal exceptions at FSI and CIA, involved the administration of all three interviews to a given examinee within a single day. For example, as shown in Figure 2, "examinee 1" was interviewed by a CIA tester team during the first one-hour time period of Day 1, by a DLI team during the third time period, and by an FSI team during the fifth period. Examinee rest breaks of at least one hour were provided between interviews, as well as a one-hour lunch break between either the first and second or second and third interviews. In addition to the lunch period, each tester team had a further one-hour break at some point in the testing day.

In setting up the above testing schedule, it was understood and acknowledged that the per-day "interviewing load" on the part of the testers (six interviews on one day, four on the other) was in some cases more substantial than was typically the case in ongoing testing work at the agency. However, counterbalancing considerations of increased staff costs, additional travel/subsistence expenses, and potential inconvenience on the part of examinees who would be required to appear again on a second or even third testing day, dictated adoption of the indicated strategy. In a debriefing questionnaire completed at the end of the testing sessions, several examiners reported that they felt somewhat burdened by the overall quantity of interviews required over the available time span, but also for the most part noted that they considered their interviews and associated ratings given in the course of the study to be as thorough and as accurate as those carried out in regular agency testing.

At DLI, due to restrictions imposed on the scheduling arrangements by both the overall daily schedule at the agency and by individual examinees' classroom session assignments, it was necessary to adopt a somewhat modified procedure in which, for a given examinee, the three interviews were held over a two-day period, on either a 2-1 or 1-2 basis. This modification also resulted in a slightly easier and more uniform interviewing pace on the part of the testers, who, with very few exceptions arising from the occasional need to "catch up" for a student who had failed to appear at an assigned testing time, were required to test only 5 students on each of the two days.

Interviewing procedures. All tester teams were extensively advised, both in memoranda circulated prior to the testing and verbally at the beginning of the first testing day, to carry out each interview in strict conformance with the procedures currently in effect at the testers' agency, including, as appropriate, the use of any routine auxiliary materials (e.g., cue cards describing situations that the student is asked to deal with, background reading materials associated with the FSI "briefing" task, and so forth). In addition, the testers were to follow whatever procedures they normally used in arriving at a final interview rating, including, for example, jointly discussing the interviewee's performance; reviewing the verbal proficiency descriptions; and considering (and, if it was the operational procedure at the agency, rating) the speech sample with respect to specified sub-factors of performance. Each testing team was also asked to report the final global rating, as well as any factor scores or other routine annotations/feedback information, on the printed forms in use at their agency for this purpose. If separate forms were normally completed by each tester, both were to be

Figure 2

Interviewing Schedule

(Cell entries are examinee IDs; same sequence used for French and German)

	Day One						Day Two					
Time Slot:	A	B	C	D	E	F	G	H	I	J	K	L
CIA Team 1:	1	2	3	4	5	6	11	12	13	14		
CIA Team 2:	7	8	9	10			15	16	17	18	19	20
DLI Team 1:	5	6	1	2	3	4			11	12	13	14
DLI Team 2:			7	8	9	10	19	20	15	16	17	18
FSI Team 1:	3	4	5	6	1	2	13	14			11	12
FSI Team 2:	9	10			7	8	17	18	19	20	15	16

submitted; if there was any disagreement concerning the final rating, the testers were to resolve the issue among themselves and circle or otherwise indicate the official "final" rating on one or the other of these forms.

In the course of the interviewing process at all three agencies, the author and another professional project staff member separately sat in on a total of approximately twelve interview sessions, distributed fairly randomly across languages, agencies, and interviewer teams. All interviews conducted by the tester teams, whether or not they were also observed by project staff, were audio recorded on C-90 cassettes, using tape recorders with built-in microphones, with the recorders placed on a table between the examinee and testers. In most instances, the raters' post-interview discussion of the examinee's performance was also recorded. Spot-checking of a number of completed tapes indicated that the spoken material was in general clearly audible with respect to both the examinee and interviewers. The obtained total of over 300 interview recordings is considered to provide a valuable corpus for further linguistic analysis or other follow-up study.

Across all three agencies, 115 examinees were interviewed by testers from each of the three agencies, out of a design total of 120. This very high level of participation is due to both the diligence of the testing coordinators in making the initial administrative arrangments for the interviewing and their willingness and ability to readily locate appropriate substitute interviewees as the occasion required over the course of the six testing days.

RESULTS

Overall results. Two types of analysis, chi-square and analysis of variance, were conducted for the testing results as a whole, that is, for the scoring performance of testers across both language groups combined. Table 1 shows the observed and expected frequencies of ratings from 0/0+ (these two levels combined to provide adequate cell size) to 5 on the part of the CIA, DLI, and FSI rating teams. The overall chi square of 20.3, with a chance probability of .32, fails to demonstrate a statistically significant difference across agencies with respect to the rating of examinee performance on a global (combined languages) basis. Alternatively stated, this statistical test indicated an approximately 1 in 3 chance that the observed differences across agencies in interview scores assigned to given examinees were due simply to random statistical effects rather than to agency-specific differences in rating tendencies. It is customary not to consider differences between or among groups to be "significant" unless there is a less than 1 in 20 chance probability (usually abbreviated as $p < .05$) that the observed results are due to factors other than random variation. As shown in Table 5, nonsignificant results ($F = 2.27$; $p = 0.10$) for combined French and German interviews were also obtained for a between- and within-groups analysis of variance, a statistical procedure that also serves to determine the likelihood that the observed results are a consequence of random variation rather than true inter-group differences.

Chi-square analyses were also conducted separately for the French (Table 2) and German (Table 3) data. Nonsignificant differences were again found for both languages, with a quite high chance probability for French (.71) and a lower, but still nonsignificant probability (.10) for German. These results may be interpreted as indicating a 7 in 10 likelihood that the observed rating differences among the three agencies with respect to the French testing were

Table 1

Chi-Square by Agency and Interview Score Assigned

(French and German)

Observed Expected (O - E) Contribution	CIA	DLI	FSI	Row Totals
0, 0+	8 6.7 1.3 0.3	9 6.7 2.3 0.8	3 6.7 -3.7 2.0	20 3.1
1	19 14.7 4.3 1.3	15 14.7 0.3 0.0	10 14.7 -4.7 1.5	44 2.8
1+	16 14.0 2.0 0.3	17 14.0 3.0 0.6	9 14.0 -5.0 1.8	42 2.7
2	12 15.3 -3.3 0.7	13 15.3 -2.3 0.4	21 15.3 5.7 2.1	46 3.2
2+	10 14.7 -4.7 1.5	19 14.7 4.3 1.3	15 14.7 0.3 0.0	44 2.8
3	10 13.0 -3.0 0.7	11 13.0 -2.0 0.3	18 13.0 5.0 1.9	39 2.9
3+	16 12.7 3.3 0.9	11 12.7 -1.7 0.2	11 12.7 -1.7 0.2	38 1.3
4	10 10.7 -0.7 0.0	9 10.7 -1.7 0.3	13 10.7 2.3 0.5	32 0.8

Table 1 (cont.)

	8	6	9	
4+	7.7	7.7	7.7	23
	0.3	-1.7	1.3	0.6
	0.0	0.4	0.2	
	6	5	6	
5	5.7	5.7	5.7	17
	0.3	-0.7	0.3	0.1
	0.0	0.1	0.0	
Column Totals	115	115	115	345
	5.7	4.3	10.3	20.3

No. of Observations = 345 Degrees of freedom = 18

Chi square = 20.3 Chance probability = 0.32

Table 2

Chi-Square for French Interviews

Observed Expected (O - E) Contribution	CIA	DLI	FSI	Row Totals
0, 0+, 1	13 10.3 2.7 0.7	13 10.3 2.7 0.7	5 10.3 -5.3 2.8	31 4.1
1+	10 8.0 2.0 0.5	9 8.0 1.0 0.1	5 8.0 -3.0 1.1	24 1.8
2	8 9.0 -1.0 0.1	7 9.0 -2.0 0.4	12 9.0 3.0 1.0	27 1.6
2+	7 7.0 0.0 0.0	7 7.0 0.0 0.0	7 7.0 0.0 0.0	21 0.0
3	4 6.3 -2.3 0.9	7 6.3 0.7 0.1	8 6.3 1.7 0.4	19 1.4
3+	6 6.3 -0.3 0.0	5 6.3 -1.3 0.3	8 6.3 1.7 0.4	19 0.7
4	5 6.7 -1.7 0.4	7 6.7 0.3 0.0	8 6.7 1.3 0.3	20 0.7

Table 2 (cont.)

4+, 5	8	6	8	
	7.3	7.3	7.3	22
	0.7	-1.3	0.7	0.4
	0.1	0.2	0.1	
Column Totals	61	61	61	183
	2.7	1.9	6.1	10.6

No. of observations = 183 Degrees of freedom = 18

Chi square = 10.6 Chance probability = 0.71

Table 3

Chi-Square for German Interviews

Observed / Expected / (O - E) / Contribution	CIA	DLI	FSI	Row Totals
0, 0+, 1	14	11	8	
	11.0	11.0	11.0	33
	3.0	0.0	-3.0	1.6
	0.8	0.0	0.8	
1+	6	8	4	
	6.0	6.0	6.0	18
	0.0	2.0	-2.0	1.3
	0.0	0.7	0.7	
2	4	6	9	
	6.3	6.3	6.3	19
	-2.3	-0.3	2.7	2.0
	0.9	0.0	1.1	
2+	3	12	8	
	7.7	7.7	7.7	23
	-4.7	4.3	0.3	5.3
	2.8	2.4	0.0	
3	6	4	10	
	6.7	6.7	6.7	20
	-0.7	-2	3.3	2.8
	0.1	1.	1.7	
3+	10	6	3	
	6.3	6.3	6.3	19
	3.7	-0.3	-3.3	3.9
	2.1	0.0	1.8	

Table 3 (cont.)

4, 4+, 5	11	7	12	
	10.0	10.0	10.0	30
	1.0	-3.0	2.0	1.4
	0.1	0.9	0.4	
Column Totals	54	54	54	162
	6.8	5.1	6.4	18.4

No. of observations	= 162	Degrees of freedom	= 12
Chi square	= 18.4	Probability of chance	= 0.10

Table 4

Chi-Square for Agency Pairs

	X2	N	df	p
French and German				
CIA - DLI	4.8	230	9	0.85
CIA - FSI	14.1	230	9	0.12
DLI - FSI	11.9	230	9	0.22
French				
CIA - DLI	1.6	122	7	0.98
CIA - FSI	8.3	122	7	0.30
DLI - FSI	7.1	122	7	0.42
German				
CIA - DLI	8.7	108	6	0.19
CIA - FSI	11.0	108	6	0.09
DLI - FSI	8.1	108	6	0.23

Table 5

Analysis of Variance and t-Test Comparisons

for Interview Scores across Three Agencies

Source of Variation	df	Sum of Squares	Mean Square	F	p
French and German					
Between groups	2	722.649	361.325	2.27	0.10
Within groups	342	54477.078	159.290		
Total	344	55199.728			
French					
Between groups	2	648.995	324.497	2.00	0.14
Within groups	180	29232.361	162.402		
Total	182	29881.355			
German					
Between groups	2	192.704	96.352	0.61	0.55
Within groups	159	25098.241	157.851		
Total	161	25290.944			

Table 5 (cont.)

t-Statistics

Comparison	t	p
French and German		
CIA - DLI	.517	0.64
CIA - FSI	1.531	0.22
DLI - FSI	2.048	0.13
French		
CIA - DLI	.078	0.94
CIA - FSI	1.691	0.19
DLI - FSI	1.769	0.18
German		
CIA - DLI	.674	0.55
CIA - FSI	.421	0.70
DLI - FSI	1.095	0.36

purely attributable to chance factors. Although there is a smaller probability (1 in 10) that the German differences were also simply a result of "chance," this figure still does not reach the commonly-accepted 1 in 20 criterion for a statistically significant difference. Analysis of variance for French and German groups considered separately (Table 5) also shows nonsignificant rating differences for both languages across the three participating agencies.

Additional chi-square analyses comparing the rating performance of individual pairs of agencies (CIA-DLI, CIA-FSI, and DLI-FSI) are shown in Table 4 for both whole-group and separate-language comparisons. All of these are statistically nonsignificant ($p > .05$). As shown in Table 5, similar results are found for t-tests of agency pairs (an analysis of variance-type procedure applicable to comparisons of pairs of groups), none of which comparisons reach statistical significance at the .05 level.

In summary of the overall analyses, it may be concluded that the ratings assigned during this study by CIA, DLI, and FSI tester teams, when considered across all the examinee proficiency levels taken as a whole, do not differ among the three agencies or between any pair of agencies in a statistically significant manner, either in combined (French and German) comparisons or in comparisons separately by language.

Inter-agency patterns of score distribution. Although the whole-group comparisons of scoring performance across the three agencies do not reach statistical significance, examination of the particular scores assigned to examinees within various portions of the overall proficiency range reveals some very interesting patterning. Table 6 shows the interview scores assigned to each examinee by the CIA, DLI, and FSI testers, listed in order of increasing mean score across the three agencies and including both French and German groups. For any given examinee, an asterisk in one of the columns indicates that that particular score is higher than the scores given by both of the other agencies. Of the 115 examinees interviewed, the CIA testers assigned, in 16 instances, a higher score than the other two tester teams. The DLI testers assigned higher ratings than their inter-agency colleagues on 8 occasions, and the FSI testers assigned higher scores in 43 cases. A fairly clear pattern is evident in the level 1, 1+, 2 range, with the FSI testers tending in many instances to assign a 1+ (or in a few instances, a 2) to examinees rated as level 1 by CIA and DLI testers. A similar tendency is noted a half-step higher on the scale, with FSI testers assigning level 2 to a number of examinees rated as 1 or 1+ by the other two agencies. A less marked tendency to assign 2+ vis-à-vis 1+ or 2 is also noted.

A tendency on the part of the FSI raters to assign level 3 scores to examinees rated lower than level 3 by the other two agencies is not evident in the data. While there are 6 such instances in the combined French and German data, there are 5 cases in which the CIA testers assigned 3 or 3+ to examinees rated as 2+ or lower by both the DLI and FSI teams. Beyond level 3, the distribution of assigned scores across the three agencies shows generally random differences, with no discernible agency-specific patterning.

Table 7 shows the distribution of ratings across agencies for the French testers separately. The tendency toward relatively higher ratings on the part of the FSI French raters is even more marked than for the combined language group, with higher-than-the-other-two-agency scores assigned by FSI to 30 of the 61 French examinees. Ratings of the CIA French testers were higher than

Table 6

Examinee Score Levels Assigned, by Agency

(French and German)

Legend: 00 = 0, 07 = 0+, 10 = 1, 17 = 1+, etc.

Asterisks indicate a score higher than that of the other two agencies.

CIA	DLI	FSI
07	00	07
07	07	07
07	07	10 *
07	10 *	07
07	07	10 *
07	07	10 *
07	10	10
10	10	10
10	10	10
10	10	10
10	10	10
17 *	07	10
10	07	17 *
10	07	17 *
07	17 *	10
10	07	17 *
10	10	17 *
10	10	17 *
10	10	17 *
10	10	17 *
10	10	17 *
10	10	17 *
10	10	20 *
10	10	20 *
10	17	17
10	17	20 *
17	10	20 *
10	17	20 *
17	10	20 *
10	20	20
10	17	27 *
17	17	20 *
17	17	20 *
17	17	20 *
17	17	20 *
17	17	20 *
20	17	20
17	20	20

Table 6 (cont.)

17	20	20
20	17	20
20	20	20
17	17	27 *
17	20	27
20	17	27 *
17	20	27 *
17	27 *	20
20	17	27 *
17	20	30 *
20	27 *	20
17	20	30 *
20	2	30 *
20	[illegible]	27
27	[illegible]	30 *
20	27	27
20	27	27
20	27	27
30 *	20	27
30 *	20	27
20	30	30
27	27	27
27	27	30 *
30 *	27	27
37 *	17	30
27	27	30 *
37 *	27	20
37 *	27	20
27	30	30
27	30	30
30	27	30
30	30	30
37 *	27	30
30	27	37 *
37 *	27	30
27	37 *	30
37 *	27	30
37	20	37
27	30	37 *
37	20	37
30	30	37 *
40 *	30	27
27	30	40 *
30	30	40 *
27	37	37
37	37	30
37	37	30
37	30	40 *
40 *	30	37
30	37	40 *

Table 6 (cont.)

	40	27	40
	37	37	37
	37	37	40 *
	40 *	37	37
	40 *	37	37
	40	40	37
	37	40	40
	37	40	40
	30	47 *	40
	40	40	40
	47	27	47
	37	40	47 *
	40	47 *	40
	47 *	40	40
	40	40	47 *
	47 *	40	40
	47	37	47
	47	37	47
	40	47	47
	47	40	47
	47	50 *	47
	47	50	50
	50	47	50
	50	50	47
	50	47	50
	50	47	50
	50	50	50
	50	50	50
Mean:	26.0	25.2	28.8
S.D.:	13.4	12.5	11.8
N:	115	115	115

Table 7

Examinee Score Levels Assigned, by Agency

(French)

Legend: 00 = 0, 07 = 0+, 10 = 1, 17 = 1+, etc.

Asterisks indicate a score higher than that of the other two agencies.

CIA	DLI	FSI
07	00	07
07	07	10 *
07	07	10 *
07	07	10 *
10	10	10
10	07	17 *
10	07	17 *
10	07	17 *
10	10	17 *
10	10	17 *
10	10	20 *
10	10	20 *
17	10	20 *
10	20	20 *
17	17	20 *
17	17	20 *
17	17	20 *
20	17	20
17	20	20
20	17	20
20	20	20
17	17	27 *
20	17	27 *
17	20	27 *
17	27	20
17	20	30 *
17	20	30 *
20	20	30 *
20	27	27
27	17	30 *
20	27	27
20	30	30
27	27	27
27	27	30 *
30 *	27	27

Table 7 (cont.)

	37 *	17	30
	27	30	30
	30	27	37 *
	27	30	37 *
	30	30	37 *
	27	30	40 *
	2,	37	37
	37	30	40 *
	40 *	30	37
	37	37	37
	40 *	37	37
	40 *	37	37
	37	40	40
	37	40	40
	30	47 *	40
	37	40	47 *
	40	47 *	40
	47 *	40	40
	40	40	47 *
	47 *	40	40
	47	37	47
	47	40	47
	47	50	50
	50	50	47
	50	50	50
	50	50	50
Mean:	25.6	25.5	29.5
S.D.:	13.2	13.3	11.5
N:	61	61	61

those of their inter-agency colleagues in only 7 instances, and the ratings assigned by the DLI testers were virtually never higher (2 of 61 occasions) than both of the other agency teams. Rating patterns again show the higher FSI ratings to be most frequent at the 1, 1+, 2, and 2+ levels. However, in the case of French, there also appears to be some tendency toward the awarding of level 3 scores by FSI to examinees rated as 2+ or lower by the other two agencies (5 instances on the part of the FSI testers, with only one comparable rating by CIA and none by DLI). The inter-agency differences in mean interview scores for French, while not statistically significant, do show a clearly higher numerical value for FSI (29.5) than for CIA and DLI (25.6 and 25.5, respectively).

Table 8 shows the distribution of German ratings on an across-agencies basis. By contrast to the French data, an apparent tendency to half-point higher ratings on the part of the FSI testers is principally restricted to 1+ vs. 1 and 2 vs. 1+ comparisons, and is by no means as salient or as widespread across proficiency levels as is the case for the French group. Also noteworthy in the German ratings is a tendency to higher ratings on the part of the CIA testers in the middle level of the score range, with level 3 or 3+ assigned by CIA to four examinees rated as 2+ or lower by both DLI and FSI, and 3+ awarded to three other examinees who were considered to be no higher than level 3 by the other two agencies. Mean German interview scores (26.5, 24.9, and 27.5 for CIA, DLI, and FSI, respectively) did not differ significantly across agencies.

Across-agency differences in scoring patterns may also be examined by means of expectancy tables based on the frequencies which which raters from pairs of agencies assigned particular level scores to given examinees. Table 9 shows, for each of the levels assigned by the CIA French testers, the corresponding level assignments of the DLI testers. For example, for the total of 9 interviewees who were rated as level 1 by CIA, 56 percent of these examinees were also rated as level 1 by DLI, 33 percent were rated as level 0+, and 11 percent, as level 2. For the 10 examinees rated as 1+ by CIA, the DLI ratings were split at 40 percent each for level 1+ and 2, and 10 percent for levels 1 and 2+. The discrepancies are more marked for the comparison of CIA and FSI ratings in French (Table 10), which shows, for example, that examinees considered to be at level 1 by the CIA testers were in a majority of cases rated as 1+ (56 percent) by the FSI testers and in third of the cases, as level 2. At this level, 89 percent of the "level 1" examinees by CIA standards were rated as level 1+ or higher by the FSI testers. The tendency continues through "CIA levels" 1+, 2, 2+, 3, and 3+, with the majority of FSI ratings being at least a half-level higher in all five comparisons. With the exception of "DLI 2+," comparisons of DLI and FSI French scores (Table 12) reveal an essentially similar pattern across DLI levels 0 through 3, with the bulk of the FSI scores consistently a half-level or more higher than the scores assigned to the same examinees by the DLI testers.

For German, there is no consistent pattern of higher or lower ratings between the CIA and DLI raters from levels 0+ through 2+ (Table 15), but at "CIA levels" 3 and 3+, the DLI raters were seen to assign somewhat lower ratings on the whole, with an appreciable spread at CIA 3+, where 50 percent of the corresponding DLI ratings were a full level lower and 20 percent, a level and a half lower. For CIA-FSI comparisions in German (Table 16), there is a clear pattern of at least half-point higher FSI ratings at CIA levels 0+ through 2+, and a similar pattern for DLI-FSI comparisons (Table 18). A particularly large discrepancy is noted for DLI level 2+, which shows corresponding FSI scores ranging from 2 to 4+.

Table 8

Examinee Score Levels Assigned, by Agency

(German)

Legend: 00 = 0, 07 = 0+, 10 = 1, 17 = 1+, etc.

Asterisks indicate a score higher than that of the other two agencies.

CIA	DLI	FSI
07	07	07
07	10	07
07	10	10
10	10	10
10	10	10
10	10	10
17 *	07	10
07	17	10
10	10	17 *
10	10	17 *
10	10	17 *
10	17	17
10	17	20 *
17	10	20 *
10	17	20 *
10	17	27 *
17	17	20 *
17	17	20 *
17	20	20
17	20	27 *
20	17	27 *
20	27 *	20
20	27	27
20	27	27
30 *	20	27
30 *	20	27
27	27	30 *
37 *	2[illegible]	20
37 *	27	20
27	30	30
30	27	30
30	30	30
37 *	27	30
37 *	27	30
27	37 *	30
37 *	27	30
37	20	37

Table 8 (cont.)

	37	20	37
	40 *	30	27
	30	30	40 *
	37	37	30
	37	37	30
	30	37	40 *
	40	27	40
	37	37	40 *
	40	40	37
	40	40	40
	47	27	47
	47	37	47
	40	47	47
	47	50	47
	50	47	50
	50	47	50
	50	47	50
Mean:	26.5	24.9	27.5
S.D.:	13.6	11.7	12.0
N:	54	54	54

Table 9

Expectancy Table for DLI from CIA Scores

(French)

Cell entries show the percentage of examinees assigned given scores by DLI for each level assigned by CIA.

CIA	DLI 0	0+	1	1+	2	2+	3	3+	4	4+	5	N
0+	25	75										4
1		33	56		11							9
1+			10	40	40	10						10
2				38	25	25	13					8
2+				14		29	43	14				7
3						50	25			25		4
3+				17			17	17	50			6
4							20	40	20	20		5
4+								20	60		20	5
5											100	3

Table 10

Expectancy Table for FSI from CIA Scores

(French)

Cell entries show the percentage of examinees assigned given scores by FSI for each level assigned by CIA.

		FSI										
		0+	1	1+	2	2+	3	3+	4	4+	5	N
	0+	25	75									4
	1		11	56	33							9
	1+				60	20	20					10
C	2				38	38	25					8
I	2+					14	43	29	14			7
A	3					25		50	25			4
	3+						17	17	50	17		6
	4							60	20	20		5
	4+								40	40	20	5
	5									33	67	3

Table 11

Expectancy Table for CIA from DLI Scores

(French)

Cell entries show the percentage of examinees assigned given scores by CIA for each level assigned by DLI.

DLI	CIA 0+	1	1+	2	2+	3	3+	4	4+	5	N
0	100										1
0+	50	50									6
1		83	17								6
1+			44	33	11		11				9
2		14	57	29							7
2+			14	29	29	29					7
3				14	43	14	14	14			7
3+					20		20	40	20		5
4							43	14	43		7
4+						50		50			2
5									25	75	4

Table 12

Expectancy Table for FSI from DLI Scores

(French)

Cell entries show the percentage of examinees assigned given scores by FSI for each level assigned by DLI.

		FSI										
		0+	1	1+	2	2+	3	3+	4	4+	5	N
DLI	0	100										1
	0+		50	50								6
	1		17	33	50							6
	1+				56	22	22					9
	2				43	14	43					7
	2+				14	57	14	14				7
	3						29	43	29			7
	3+							80		20		5
	4								57	43		7
	4+								100			2
	5									25	75	4

Table 13

Expectancy Table for CIA from FSI Scores

(French)

Cell entries show the percentage of examinees assigned given scores by CIA for each level assigned by FSI.

		C I A																	
		0+		1		1+		2		2+		3		3+		4		4+		5	N
F S I	0+	100																			1
	1	75		25																	4
	1+			100																	5
	2			25		50		25													12
	2+					29		43		14		14									7
	3					25		25		38				13							8
	3+									25		25		13		38					8
	4									13		13		38		13		25			8
	4+													20		20		40		20	5
	5																	33		67	3

Table 14

Expectancy Table for DLI from FSI Scores

(French)

Cell entries show the percentage of examinees assigned given scores by DLI for each level assigned by FSI.

		DLI 0	0+	1	1+	2	2+	3	3+	4	4+	5	N
	0+	100											1
	1		75	25									4
	1+		60	40									5
F	2			25	42	25	8						12
S	2+				29	14	57						7
I	3				25	38	13	25					8
	3+						13	38	50				8
	4							25		50	25		8
	4+								20	60		20	5
	5											100	3

Table 15

Expectancy Table for DLI from CIA Scores

(German)

Cell entries show the percentage of examinees assigned given scores by DLI for each level assigned by CIA.

CIA	DLI 0+	1	1+	2	2+	3	3+	4	4+	5	N
0+	25	50	25								4
1		60	40								10
1+	17	17	33	33							6
2			25		75						4
2+					33	33	33				3
3				33	17	33	17				6
3+				20	50		30				10
4					20	20		40	20		5
4+					33		33			33	3
5									100		3

Table 16

Expectancy Table for FSI from CIA Scores

(German)

Cell entries show the percentage of examinees assigned given scores by FSI for each level assigned by CIA.

		FSI 0+	1	1+	2	2+	3	3+	4	4+	5	N
	0+	50	50									4
	1		30	40	20	10						10
	1+		17		67	17						6
C	2				25	75						4
I	2+						100					3
A	3					33	33		33			6
	3+				20		50	20	10			10
	4					20		20	40	20		5
	4+									100		3
	5										100	3

Table 17

Expectancy Table for CIA from DLI Scores

(German)

Cell entries show the percentage of examinees assigned given scores by CIA for each level assigned by DLI.

		CIA 0+	1	1+	2	2+	3	3+	4	4+	5	N
	0+	50		50								2
	1	22	67	11								9
	1+	13	50	25	13							8
D	2			33			33	33				6
L	2+				25	8	8	42	8	8		12
I	3					25	50		25			4
	3+					17	17	50		17		6
	4								100			2
	4+								25		75	4
	5									100		1

Table 18

Expectancy Table for FSI from DLI Scores

(German)

Cell entries show the percentage of examinees assigned given scores by FSI for each level assigned by DLI.

		FSI																				
		0+		1		1+		2		2+		3		3+		4		4+		5		N
	0+	50		50																		2
	1	11		44		33		11														9
	1+			13		13		50		25												8
D	2							17		50				33								6
L	2+							25		17		42				8		8				12
I	3									25		50				25						4
	3+											50				33		17				6
	4													50		50						2
	4+																	25		75		4
	5																	100				1

Table 19

Expectancy Table for CIA from FSI Scores

(German)

Cell entries show the percentage of examinees assigned given scores by CIA for each level assigned by FSI.

FSI	CIA 0+	1	1+	2	2+	3	3+	4	4+	5	N
0+	100										2
1	33	50	17								6
1+		100									4
2		22	44	11			22				9
2+		13	13	38		25		13			8
3					30	20	50				10
3+							67	33			3
4						40	20	40			5
4+								25	75		4
5										100	3

Table 20

Expectancy Table for DLI from FSI Scores

(German)

Cell entries show the percentage of examinees assigned given scores by DLI for each level assigned by FSI.

		DLI										
		0+	1	1+	2	2+	3	3+	4	4+	5	N
	0+	50	50									2
	1	17	67	17								6
	1+		75	25								4
F	2		11	44	11	33						9
S	2+			25	38	25	13					8
I	3					50	20	30				10
	3+				67				33			3
	4					20	20	40	20			5
	4+					25		25		25	25	4
	5									100		3

Tables 13-14 (for French) and Tables 19-20 (German) show the variation in scores observed for the CIA and DLI interviews for given score levels on the FSI-conducted interviews. These data may be "read" in the same manner as those shown in the other expectancy tables. For example, as shown in Table 13, of the 12 French interviewees assigned a rating of 2 by the FSI testers, 25 percent received a score of 1 in the interviews conducted by CIA; 50 percent received a rating of 1+; and 25 percent, a rating of 2.

Three major considerations should be kept in mind in evaluating the observed results. First, at issue in this study is the test-retest reliability of the interviewing process, in which the intent is to determine the extent to which given examinees, undergoing separate, independent interviews by each of the three agencies, will be assigned similar level scores in each instance. Observed variation in examinee score levels may be attributable--in proportions that it is not statistically possible to determine on the basis of the present study--to actual performance differences on the part of the examinee across the three interviewing occasions, as well as to agency-specific differences in the manner in which a given examinee performance would tend to be evaluated across the three agencies. It is, therefore, possible to suggest that at least some of the scoring differences observed in this study may be attributable to interview-to-interview variation in performance on the part of the examinees, rather than to rater unreliability per se. However, if the intended operational assumption is that the face-to-face interviewing technique (assuming good will and serious communicative effort on the part of the examinee and diligence and proper attention to elicitation procedures on the part of the examiners) should result in the awarding of similar ratings on closely contemporaneous interviewing occasions, the procedure used in this study may be considered an appropriate empirical approach to determining the validity of this assumption, within the general linguistic and personnel parameters involved (a sampling of two interviewer pairs for two languages across the three participating agencies).

Second, although the total number of interviews obtained in the study was as large as practicable within the financial and administrative constraints involved, and may be considered to provide a reasonably stable and accurate indication of the results that would be secured in a similar but larger study, some caution in interpretation and extrapolation should be exercised, especially in analyzing those expectancy table columns and associated data that are based on a relatively smaller number of interviews.

Third, the expectancy table data should not be viewed as representing in any sense "true" level ratings on the vertical axis. These tables simply show the extent to which the agencies in question tended to vary in the frequencies with which they assigned a given level score to a particular examinee. Any determination of which, if any, of the ratings assigned should be considered to reflect the "true" proficiency level of the examinee is beyond the scope of this study and, indeed, represents a question for which statistical data per se are, at best, of very limited value. Although there is some indication that, for the two languages involved, the interview ratings assigned by CIA and DLI were for certain portions of the overall proficiency scale more similar to each other than they were to the corresponding ratings assigned by FSI, it cannot and should not be concluded from these results that the former ratings were found to be "correct" and the latter "incorrect," in any useful external or criterial sense of the term.

Analysis of rating "factor" data. Some additional information, especially for the interviews conducted by the FSI testing teams, is available concerning the statistical interrelationships of the raters' scoring of various linguistic categories or "factors" that are generally considered to contribute collectively to overall proficiency as expressed in the global rating, but at the same time to provide a certain amount of diagnostic feedback concerning particular sub-aspects of performance (within a given global level) exemplified by a given examinee. Table 21 shows, for combined French and German data, the observed intercorrelations of the FSI global rating and each of the five "factor" scores--"listening comprehension," "discourse," "structure," "lexicalization," and "fluency"-- regularly assigned by FSI testers for interviews conducted by that agency, as an aid in focusing on component aspects of the global rating and in providing for greater objectivity in the rating process overall.

The observed high correlations may be considered attributable to the combined effects of at least two possible sources of correspondence: "true" close relationships among the factors as exemplified in the examinee's performance; and a potential "halo effect" arising from the fact that all factor scores are assigned by the same testers, who may be influenced to some extent by examinee performance on one or more of the other factors while attempting to objectively rate a given factor. Although the correlational data suggest that, on a total-group basis, relatively little additional information is provided by the individual factor scores that is not already statistically captured in the global rating, the scoring profiles of particular examinees whose pattern of factor ratings shows an appreciable departure from linearity may be of interest from a diagnostic or pedagogical standpoint. For example, the scatterplot of "structure" vs. "lexicalization" scores shown in Table 22 shows three examinees whose factor ratings for "structure" were proportionately appreciably higher than their ratings for "lexicalization"; and three other examinees for whom the "lexicalization" scores were noticeably higher than the "structure" scores. Although detailed linguistic review of the interviewing performance of particular examinees is beyond the scope of the present report, the scoring data obtained in the study can serve to identify these and other "discrepant" cases for further clinical analyses addressing, for example, the so-called "street learner/school learner" performance differences frequently reported in operational testing activities.

Detailed factor score data are not available for the CIA or DLI interviews in that, for the most part, testers from these two agencies followed the current operational procedure of providing only the overall global rating, with the single exception of a separate "listening comprehension" score that was consistently awarded by by the CIA raters and in about two-thirds of the cases by the DLI raters. The obtained "listening" vs. "global" correlations were .97 for the CIA interviews (N = 114) and .98 for the DLI interviews (N = 85). These data again suggest that, on a whole-group basis, very little "new" information is provided by the separate listening score. Analysis of individual discrepant cases for clinical or pedagogical purposes would of course be possible for the CIA and DLI data as well as for the FSI interviews.

Examinee and tester feedback on interviewing process. The observations and opinions of both examinees and interviewers concerning various aspects of the interviewing procedures as exemplified during the study were solicited through two separate questionnaires (Appendices A and B). The examinee questionnaire requested information on the examinee's affiliation and test language, and both "yes-no" and open-ended comment responses to the following

Table 21

Intercorrelations of FSI Factor Ratings

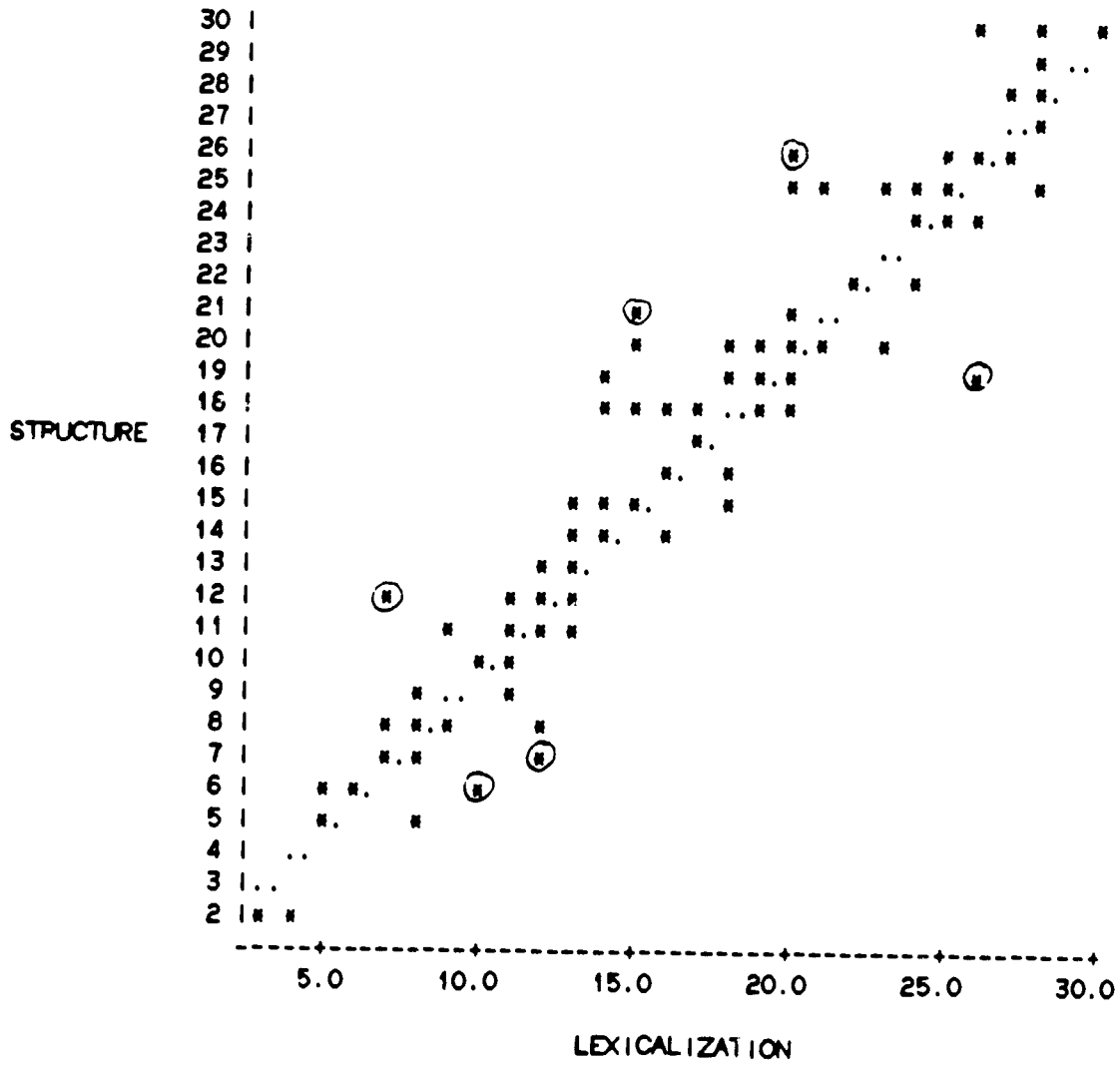
(N = 115)

	Global Rating	Comprehension	Discourse	Structure	Lexicalization
Comprehension	.94				
Discourse	.97	.92			
Structure	.97	.92	.96		
Lexicalization	.97	.94	.95	.96	
Fluency	.95	.94	.94	.93	.94

Table 22

FSI Structure vs. Lexicalization

(r = .96; N = 115)



questions for each of the three interviews taken:

"Did the opportunities the testers provided you to speak the language during the [first, second, third] interview (in terms of the type and number of topics covered, range of performance required) adequately probe your maximum proficiency level?"

"Did the testers during the [first, second, third] interview use any elicitation techniques or cover any kinds of topics that you thought were in any way "unfair" or in some other way not a valid test of your speaking proficiency?"

"During the [first, second third] interview, did the testers appear to make a conscious effort to put you at ease?"

Four additional summary questions involved forced-choice judgments as follows:

"In which of the three interviews do you feel you...

were most relaxed and at ease?

were the most anxious or nervous?

best demonstrated your optimum speaking proficiency?

least well demonstrated your optimum speaking proficiency?"

Questionnaires were distributed to the examinees by the testing coordinator at each agency within about one week following completion of the interviewing process, an approach intended to avoid the possibility that examinees filling out the questionnaire immediately on completion of the testing might be disproportionately influenced by their experience in the most recently-taken interview. The examinee was asked to provide his or her name on an attached slip in order to properly categorize the 'first,' 'second,' and 'third' interviews taken, but was assured that the slip would be removed when the results were summarized and that all data would be analyzed and reported on an anonymous basis. A self-addressed, postpaid envelope was provided for return of the questionnaire. Of the 115 examinees participating in the study, questionnaires were returned by 83, a response rate of 72 percent.

Table 23 provides a summary of the examinee questionnaire responses. To the question "Did the opportunities the testers provided you to speak the language during the interview...adequately probe your maximum proficiency level?," by far the greatest number of responses (87 percent overall) were in the affirmative. Chi-square analysis for CIA, DLI, and FSI interviews showed no significant differences across agencies in the frequency with which the examinees reported an adequate probing of maximum proficiency level. To the question of elicitation techniques or coverage of topics that the examinee considered "'unfair' or in some other way not a valid test of your speaking proficiency," 82 percent of the total judgments across agencies were that "unfair" techniques had not been used. However, on an agency-specific basis, the corresponding chi-square is highly significant ($p = .007$) with 29 percent of the FSI interviews being judged as "unfair" in procedure or topical coverage, as contrasted to 13 percent and 12 percent for the CIA and DLI interviews, respectively.

Table 23

Summary of Responses to Examinee Questionnaire

"Did the opportunities the testers provided you to speak the language during the interview (in terms of the type and number of topics covered, range of performance required) adequately probe your maximum proficiency level?"

	CIA Interview	DLI Interview	FSI Interview	Total
YES	90%	86%	85%	87%
NO	10%	14%	15%	13%
Total Responses:	72	72	75	219

Chi square = .92; p = .63

"Did the testers use any elicitation techniques or cover any kinds of topics that you thought were in any way "unfair" or in some other way not a valid test of your speaking proficiency?

	CIA	DLI	FSI	Total
YES	13%	12%	29%	18%
NO	87%	88%	71%	82%
Total responses:	83	83	83	249

Chi square = 9.93; p = .007

Table 23 (cont.)

"Did the testers appear to make a conscious effort to put you at ease?"

	CIA	DLI	FSI	Total
YES	89%	95%	82%	89%
NO	11%	5%	18%	11%
Total responses:	83	83	82	248

Chi square = 7.50; p = .024

"In which of the three interviews do you feel you were most relaxed and at ease?"

CIA	DLI	FSI
27	40	12

"In which of the interviews do you feel you were the most anxious or nervous?"

CIA	DLI	FSI
15	9	50

"In which of the three interviews do you feel you best demonstrated your optimum speaking proficiency?"

CIA	DLI	FSI
23	29	25

"In which of the three interviews do you feel you least well demonstrated your optimum speaking proficiency?"

CIA	DLI	FSI
22	20	27

The question "Did the testers appear to make a conscious effort to put you at ease?" was answered affirmatively in almost 9 out of 10 cases overall (89 percent), but chi-square analysis again shows a significant across-agency difference ($p = .024$), with a somewhat smaller proportion of the FSI interviews (82 percent) being judged as consciously directed toward putting the examinee at ease, by comparison to the corresponding CIA (89 percent) and DLI (95 percent) figures.

Although the total number of data elements for the forced-choice questions (one rather than three per examinee) are insufficient for across-agency statistical comparison, the absolute frequencies of response to these questions appear to corroborate rather closely the results of the earlier questions. To the question, "In which of the three interviews do you feel you were most relaxed and at ease?," 40 interviewees indicated "DLI"; 27, "CIA"; and 12, "FSI". The conversely-phrased question, "In which of the three interviews do you feel you were the most anxious or nervous?," showed even greater differentiation across agencies, with only 9 interviewees identifying "DLI"; 15, "CIA"; and 50, "FSI." Notwithstanding an apparent clear discrimination on the examinees' part as to the relative ease/anxiety producing qualities of the interview as conducted by each of the three agencies, no appreciable across-agency differences are shown in their judgments of the agency providing the best or worst opportunity to demonstrate their optimum speaking proficiency.

To determine possible differences in questionnaire response tendencies attributable to an interaction between the agency affiliation of the examinees and that of the tester teams--that is, to investigate the possibility that, for example, DLI students might have reported different experiences or opinions concerning their participation in the DLI-conducted interviews than did interviewees from CIA or FSI being tested by the DLI teams (or analogously for other examinee/agency combinations)--additional chi square analyses of each of the questions summarized in Table 23 were carried out for the crosstabulations of interviewee agency and tester agency. All of these analyses showed nonsignificant ($p > .05$) interaction effects, suggesting that reported examinee reactions to their experiences in being tested by each of the three agencies did not vary to any meaningful extent as a consequence of their own agency affiliation. These results must be considered only suggestive in view of the fact that, at all three agencies, a few of the examinees (particularly at the higher proficiency levels) were necessarily drawn from agency alumni or other sources. As such, their own reactions to the interviewing process may not have been fully typical of those of the current students; however, to exclude these non-student cases from the interaction analysis would have reduced the already small cell sizes to statistically inappropriate levels.

The questionnaire completed by the examiners themselves (Appendix B) was somewhat less formal than the examinee questionnaire and requested open-ended comments by the testers concerning several aspects of their interviewing in the course of the project. Of the 24 testers participating in the study, 18 returned completed questionnaires (75 percent). Responses were on an intentionally anonymous basis, with only the tester's "language and agency affiliation" being requested on the questionnaire form. As shown in Table 24, based on the project staff's categorizations for analysis purposes of the free-response answers, the great majority of testers felt that the interviewing procedures they had used during the study were the same as those used during "routine, day-to-day testing" at their agency; and that the ratings which they assigned were, on the whole, as accurate as those typically made during

Table 24

Summary of Responses to Tester Questionnaire

"Do you feel that the interviewing procedures (elicitation techniques, use of props, role-plays, etc.) you used during the study were the same as those you use in routine, day-to-day testing?"

SOMEWHAT	2
DEFINITELY	16

"Do you feel that the ratings you assigned during the study were, on the whole, more accurate, about as accurate, or less accurate than ratings you typically make in routine testing at your agency?"

NOT AS ACCURATE	1
ABOUT AS ACCURATE	17

"Do you feel that the accuracy of your ratings varied at certain times or points during the six-day testing period?"

NOT AT ALL	9
A LITTLE	6
SOMEWHAT	3

"Did you notice any differences in the composition of the examinee groups at the different agencies with respect to overall levels of proficiency, examinee reactions to interview techniques, etc.?"

NOT AT ALL	4
A LITTLE	4
SOMEWHAT	10

Table 24 (cont.)

"Do you feel that participation in the project was in any way interesting or beneficial to you?"

NOT AT ALL	1
SOMEWHAT	4
DEFINITELY	13

operational testing. To the question, "Do you feel that the accuracy of your ratings varied at certain times or points during the six-day testing period?," most respondents were of the opinion that their judging accuracy had not varied appreciably over the course of the testing, but some cited the relatively intensive testing schedule (involving in some cases up to six interviews per day) as a potential source of end-of-day fatigue and consequent lack of full and "fresh" attention to the interviewing and rating tasks. With respect to the possible effects of "examiner fatigue" on the overall study results, it should be emphasized that the counterbalanced scheduling of the interviewing sessions was designed to adjust operationally for this and other possible sequence-of-interviews-related factors insofar as the inter-agency comparisons at issue in the study are concerned.

Some differences in the overall composition of the examinee groups at the three different agencies were also noted by the testers, with the FSI and CIA examinees, in general, considered to be more proficient on a total-group basis than the DLI interviewees. Again, the balanced nature of the study design, in which testers from each agency interviewed the same examinees at all three testing locations, would be expected to rule out any effects of inter-agency differences in examinee populations with respect to the project results per se.

Despite the fairly rigorous testing schedule, which involved both concentrated interviewing on a day-to-day basis and travel between Washington and Monterey within a relatively brief time span, the great majority of interviewers felt that their participation in the project had been of interest and benefit to them. Cited especially in this regard were the opportunities to meet and interact with testers from other agencies and to "share notes" on both a personal and professional basis. Several examiners expressed the hope that similar projects undertaken in the future could have built into them more extensive and more formally-structured opportunities for this type of interaction.

SUMMARY

The major results of the study may be summarized as follows. With respect to the testing of French and German by trained CIA, DLI, and FSI interviewer/raters, as represented by two randomly selected two-person teams for each agency and language, who interviewed and rated a total of 20 examinees each across essentially the full spectrum of proficiency levels, the ratings assigned did not differ across agencies in a statistically significant way, either on a combined (French plus German) or individual-language basis. Notwithstanding these overall results, examination of the rating performance for various sub-portions of the proficiency scale showed fairly clear across-agency differences for both languages, primarily at the lower and middle ranges of the scale, with these differences for the most part reflecting relatively higher rating assignments on the part of the FSI raters by comparision to the ratings given by the other two agencies. As shown both in the distributions of test scores for the same examinees across agencies and in a series of two-way expectancy tables derived from these distributions, there are occasional fairly wide discrepancies in scoring for individual examinees, which suggests the advisability, on a follow-up basis, of clinically studying the most discrepant cases from both linguistic and interviewing-procedure standpoints, to attempt to identify common factors that may have contributed to these scoring differences.

Analysis of the intercorrelations of the FSI "factor" scores among themselves and with the global ratings shows very high correspondence among all of these variables. Correlations of the CIA and FSI "listening" scores with the global ratings were also extremely high. These results suggest that, notwithstanding the possible utility of the factor scoring process in facilitating the interviewers' overall rating task, relatively little new or different statistical information is provided by the factor scores by comparison to the information already contained in the global ratings. However, factor score analysis does make it possible to identify individual examinees showing atypical (non-linear) factor score patterns, and detailed linguistic analysis of the interview performance of individuals showing such patterns may be of both research and pedagogical interest.

Questionnaire-based information obtained from the participating examinees indicates that, for the most part, the examinees felt that their optimum level of proficiency had been adequately probed in interviews conducted by all three agencies. There were, however, appreciable differences in the examinees' affective reactions to the interviewing process, with a statistically significant tendency for the the examinees to view the FSI interviewing procedure as both more anxiety-producing and making more frequent use of what they considered to be "unfair" elicitation techniques. Also on the basis of questionnaire responses, the great majority of participating testers reported that, in their opinion, the interviews which they had conducted during the study were quite similar to the operational interviews given at their home agency with respect to interviewing procedures and accuracy of ratings, although the atypically long testing day was cited in some instances as a potential source of differences in both areas. Virtually all testers found their own involvement in the study to have been quite rewarding to them from personal and/or professional standpoints.

With regard to extrapolation of study results, it is reasonable to assume, as a consequence of the study design, that the testers chosen for the study represented a random sample of the population of testers currently interviewing in that language at each agency. As such, their performance may be considered indicative of the probable total group characteristics of testers in that language/agency combination, without, however, ruling out the possibility that the "luck of the draw" may have in some instances placed in the sample individuals having atypical characteristics in terms of their elicitation procedures or accuracy of rating vis-a-vis those of their colleagues.

Considerable caution should be exercised in extrapolating the observed results for French and German testing to testing in other languages not formally investigated in the study, both in view of the fact that the non-studied languages have different populations of testers, and in consideration of possible linguistically-based differences across languages that would have an operational bearing on the interviewing process and/or on the reliability of the ratings assigned. It should also be emphasized that the present study provides information about the test-retest comparability of the interviewing process on an across-agencies basis, and does not directly examine the question of rating reliability within a given agency (i.e., the extent to which each of several raters within one agency would agree with one another in repetitive interviewing of a given examinee), and it is quite possible to suggest that the level of scoring agreement within any one agency would be greater than that observed on an inter-agency basis. However, to the extent that the ILR scale-based interview is intended to represent a "common metric" of examinee performance, with identical meaning and interpretation across using agencies,

the results of the present study warrant close examination for possible conceptual or procedural implications that would arise from holding such an objective.

APPENDIX A

QUESTIONNAIRE FOR PARTICIPANTS IN INTERVIEW TESTING STUDY

We would first like to take this opportunity to thank you for participating as an examinee in our study of proficiency testing and scoring procedures across three government language-teaching agencies. In order to derive the greatest possible amount of useful information from the study, we would very much appreciate it if you would take a few minutes to answer the questions below, based on your own experiences as an interviewee for this project.

In order to properly categorize the "first," "second," and "third" interviews you took, we would ask you to indicate your name on the slip attached to the front of the questionnaire. This slip will be removed when the results are summarized, and all data will be analyzed and reported on an anonymous basis.

A preaddressed, postpaid return envelope is enclosed for your convenience. In order for us to be able to prepare the final report on a timely basis, we would request that you return the completed questionnaire to us within one day of receipt if at all possible.

Information concerning the proficiency level ratings that you were assigned during the study will be forwarded to you within approximately 5 days.

Thank you again for your much-appreciated interest and participation in this important measurement study.

Please answer each of the questions below by marking the correct space and/or by filling in a response as appropriate:

(1) At which agency are you a student (or otherwise affiliated)? Check <u>one</u>:

[ ] CIA
[ ] DLI
[ ] FSI

(2) In which language were you tested?

[ ] French
[ ] German

PLEASE ANSWER THE FOLLOWING QUESTIONS IN TERMS OF THE <u>FIRST</u> OF THE THREE INTERVIEWS YOU TOOK DURING THE STUDY.

(3) Did the opportunities the testers provided you to speak the language during the FIRST interview (in terms of the type and number of topics covered, range of performance required) adequately probe your maximum proficiency level?

[ ] Yes
[ ] No
[ ] Not Sure

Comments?

(4) Did the testers during the FIRST interview use any elicitation techniques or cover any kinds of topics that you thought were in any way "unfair" or in some other way not a valid test of your speaking proficiency?

[ ] Yes
[ ] No

If "yes," please describe briefly:

(5) During the FIRST interview, did the testers appear to make a conscious effort to put you at ease?

[ ] Yes
[ ] No

Comments?

PLEASE ANSWER THE FOLLOWING IN TERMS OF THE SECOND INTERVIEW YOU TOOK.

(6) Did the opportunities the testers provided you to speak the language during the SECOND interview (in terms of the type and number of topics covered, range of performance required) adequately probe your maximum proficiency level?

[ ] Yes
[ ] No
[ ] Not Sure

Comments?

(7) Did the testers during the SECOND interview use any elicitation techniques or cover any kinds of topics that you thought were "unfair" or in some other way not a valid test of your speaking proficiency?

[ ] Yes
[ ] No

If "yes," please describe briefly:

(8) During the SECOND interview, did the testers appear to make a conscious effort to put you at ease (regardless of whether it "worked")?

[ ] Yes
[ ] No

Comments?

PLEASE ANSWER THE FOLLOWING IN TERMS OF THE THIRD INTERVIEW

(9) Did the opportunities the testers provided you to speak the language during the THIRD interview (in terms of the type and number of topics covered, range of performance required) adequately probe your maximum proficiency level?

[ ] Yes
[ ] No
[ ] Not Sure

Comments?

(10) Did the testers during the THIRD interview use any elicitation techniques or cover any kinds of topics that you thought were in any way "unfair" or in some other way not a valid test of your speaking proficiency?

[ ] Yes
[ ] No

If "yes," please describe briefly:

(11) During the THIRD interview, did the testers appear to make a conscious effort to put you at ease?

[ ] Yes
[ ] No

Comments?

---

(12) In which of the three interviews do you feel you were most relaxed and at ease? [ ] FIRST [ ] SECOND [ ] THIRD

Comments?

(13) In which of the three interviews do you feel you were the most anxious or nervous? [ ] FIRST [ ] SECOND [ ] THIRD

Comments?

(14) In which of the three interviews do you feel you best demonstrated your optimum speaking proficiency? [ ] FIRST [ ] SECOND [ ] THIRD

Comments?

PLEASE CONTINUE ON BACK PAGE.

(15) In which of the three interviews do you feel you <u>least well</u> demonstrated your optimum speaking proficiency?

Comments?

Please use the space below to give any additional information, comments, or suggestions concerning the interviewing procedures or other aspects of the study, or your performance on the interviews. Where necessary, please identify the interview(s) as FIRST, SECOND, etc. <u>Thank you very much for your help!</u>

APPENDIX B

INTERVIEW RATING COMPARABILITY STUDY

EXAMINER FEEDBACK FORM

We would like to take this opportunity to express our appreciation for your diligent and conscientious participation in the interview rating comparability study that will be completed with the third-agency testing at DLI today and tomorrow. Because of the quite busy schedule, which is necessitated for logistic reasons, it will not be possible for us to arrange for formal group discussions and information sharing concerning the interviewing process and other aspects of the study (even though some interaction has been possible on a more informal basis). In lieu of a formal feedback meeting as part of the "testing day" itself, we would greatly appreciate your taking the opportunity at some point over the next two days to respond to the questions below. In addition to answering the specific questions, we would appreciate any more general feedback or suggestions that you would care to provide concerning any aspect of the study. We would ask you not to give your name when filling out the questionnaire, but we would appreciate your marking your language and agency affiliation in the space provided at the end of the questionnaire.

---

1. Do you feel that the interviewing procedures (elicitation techniques, use of props, role-plays, etc.) you used during the study were the same as those you use in routine, day-to-day testing? Please explain briefly.

2. Do you feel that the ratings you assigned during the study were, on the whole, more accurate, about as accurate, or less accurate than the ratings you typically make in routine testing at your agency? Please explain.

3. Do you feel that the accuracy of your ratings varied at certain times or points during the six-day testing period?

4. Did you notice any differences in the composition of the examinee groups at the different agencies with respect to overall levels of proficiency, examinee reactions to interview techniques, etc.?

5. Do you feel that participation in the project was in any way interesting or beneficial to you?

6. If a similar or expanded study of rating comparability were to be conducted in the future, do you have any recommendations on additional factors that might be included in planning or carrying out the study?

Your Language _____ Agency _____