National Library of Canada
Canadian Theses Service
Ottawa, Canada
K1A 0N4

Bibliothèque nationale du Canada
Services des thèses canadiennes

CANADIAN THESES

THÈSES CANADIENNES

NOTICE

The quality of this microfiche is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

Previously copyrighted materials (journal articles, published tests, etc.) are not filmed.

Reproduction in full or in part of this film is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30.

AVIS

La qualité de cette microfiche dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

Les documents qui font déjà l'objet d'un droit d'auteur (articles de revue, examens publiés, etc.) ne sont pas microfilmés.

La reproduction, même partielle, de ce microfilm est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30.

THIS DISSERTATION
HAS BEEN MICROFILMED
EXACTLY AS RECEIVED

LA THÈSE A ÉTÉ
MICROFILMÉE TELLE QUE
NOUS L'AVONS REÇUE

NL-339 (r.86/06)

Canada

The Effect of Speech Sample Length on Interrater and Intrarater Reliability

Claude Hamel

A Thesis

in

The Department

of

Applied Linguistics

Presented in Partial Fulfillment of the Requirements
for the Degree of Master of Arts at
Concordia University
Montréal, Québec, Canada

August 1986

© Claude Hamel, 1986

Permission has been granted to the National Library of Canada to microfilm this thesis and to lend or sell copies of the film.

The author (copyright owner) has reserved other publication rights, and neither the thesis nor extensive extracts from it may be printed or otherwise reproduced without his/her written permission.

L'autorisation a été accordée à la Bibliothèque nationale du Canada de microfilmer cette thèse et de prêter ou de vendre des exemplaires du film.

L'auteur (titulaire du droit d'auteur) se réserve les autres droits de publication; ni la thèse ni de longs extraits de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation écrite.

ISBN 0-315-32220-9

ABSTRACT

The Effect of Speech Sample Length on Interrater and Intrarater Reliability

Claude Hamel

This study investigates the reliability of holistic rater judgements of 30-, 60- and 120-second speech samples elicited from two relatively homogeneous groups, one young and at a "beginner" level, the other older and at an "intermediate" level. Samples are judged on the basis of overall oral proficiency by pools (n=6) of raters per speech sample length condition. Analysis of variance is used to estimate pooled rater reliability, interrater reliability and intrarater reliability. The effect of speech sample length on reliability is discussed. Also considered are the effect on reliability of pooled (n=6) versus individual rater judgements, and the effect of the speaker group. It is suggested that reliability estimates for the "intermediate" group are generally higher than those for the "beginner" group because of the comparative heterogeneity of the "intermediate" group. Most importantly, findings suggest that, given the length conditions involved, single-rater judgements give rise to unstable reliability estimates while pooled (n=6) rater judgements yield much more stable estimates. The investigator proposes tentatively that single judgements of speech samples which are less than two minutes in length be accepted with caution.

ACKNOWLEDGEMENTS

This research would not have been possible without the generosity of several individuals. I would like to express my gratitude especially to J.A. Upshur, without whose advice and patience interest in this project would never have been instilled nor been so sustained; to Patsy Lightbown and Molly Petrie for their comments and suggestions; to Patsy again, to Leslie Paris and to Iolanda Bolduc for making the speech sample collection for Group Y possible; to Kate Owens for allowing me the use of taped samples for Group O; to Anne Barkman for her nimble fingers at the computer; to Elizabeth Tadgell for her unwavering support; and to Joan, Anna and Murray for untold hours of personal encouragement.

TABLE OF CONTENTS

List of Figures

List of Tables

Chapter 1

Purpose of Research

A. Context for research question:

It seems reasonable to assume that the learning of a second language is most often prompted by a desire to be able to participate in active discourse with another speaker of the language, whether that speaker be a fellow learner or a native speaker. It seems equally reasonable to assume that for such second language learners, there exists a natural curiosity concerning one's general second language development, and specifically, one's oral proficiency relative to that of other learners of the same language. Teachers too require occasional indications of student progress for purposes of accountability towards student and colleague alike.

From time to time, the teacher of a second language must try to determine at what point along the oral proficiency continuum a learner's interlanguage might fall. But oral proficiency is a complicated amalgam of various skills. It incorporates fluency, a knowledge of vocabulary and syntax, a facility for pronunciation, and an awareness of the correctness of the language.

Tests, or portions of tests made up of discrete-point items enjoy a degree of reliability in scoring which is more difficult to establish for an oral interview. After

all, decontextualized discrete-point items usually elicit answers which are clearly either correct or incorrect. On the other hand, attempting to establish, even partly, an examinee's oral proficiency level on the basis of an "interview", is a much more delicate, much more subjective exercise.

This paper will not consider the question of how valid the interview is as a measure of oral proficiency. Nor will consideration be given to the question of whether an impressionistic, holistic judgement of a speech sample is more reliable than a judgement which assigns weightings to individual aspects of oral production. These are important questions but they are not within the scope of this paper. The study reported here deals exclusively with the reliability of holistic rater judgements of 30-second, 60-second and 120-second speech samples elicited using a face-to-face "interview" technique. A more detailed discussion of the concept of reliability and the inherent problems of error variance will preface the section on statistical analysis, but for the moment, it is important to stress that it is the <u>precision</u> of impressionistic ratings of short interviews which is of interest. Two major questions arise:

1. To what extent does an individual rater, judging a speech sample of a specific length by a speaker, agree with the judgement of the same speech sample made by another rater (interrater reliability)? How does this agreement

change given longer speech samples?

2. To what extent does an individual rater, judging a second speech sample of a specific length by a speaker, agree with his/her judgement of a different, previously heard speech sample of the same length by the same speaker (intrarater reliability)? How does this agreement change given longer speech samples?

The research question addressed in this study is therefore: What is the effect of speech sample length on interrater and intrarater reliability when raters are asked to make judgements based on overall oral proficiency? It follows that the two dependent variables are interrater and intrarater reliability. While speech sample length constitutes the main independent variable, other pertinent effects will be considered. The effect of the speaker group will be taken into account, one group being young and at a "beginner" level, the other group being older and at an "intermediate" level. The effect of rating a continuous speech segment of a particular length (for example, one segment of 60 seconds) versus two segments which total the same length (for example, two 30-second segments) will be investigated. As well, the effect on reliability of a pool of six raters versus individual raters will be considered.

The results of such a study would have implications for pedagogical or research situations in which a teacher or rater is asked to make holistic judgements of overall proficiency on the basis of short segments of speech. If

ratings performed under such conditions are seen to be highly reliable, results would suggest that speech segments of either 30, 60 or 120 seconds are perhaps adequate for student placement and evaluation purposes, or for purposes of research. If, on the other hand, reliability estimates of such holistic judgements are shown to be highly unstable, then the efficacy of ratings performed under such conditions must also be questioned.

B. Review of the literature:

Most manuals which accompany tests of language proficiency offer pertinent statistical information resulting from validity and reliability studies performed by or for authors of the tests. Because the question of validity will not be considered here, data pertaining to it will not be mentioned. Indications of reliability of various language tests, however, are pertinent to the study reported in this paper.

It is of some value to compare the reliability estimates of commonly-used objective tests of general language proficiency, commonly-used subjective tests of oral proficiency and finally, tests of an aspect of language proficiency whose results are sometimes based, like those of oral proficiency, on subjective rater judgements, namely tests of writing proficiency. Such a comparison might then allow speculative reasons for one type of test showing higher reliability estimates than

another type. Table 1 shows reliability figures for objective tests of general language proficiency and subjective tests of oral proficiency. As well, an objective and a subjective test of writing proficiency are cited.

TABLE 1

Description of and reliability estimates for tests of general language proficiency, oral proficiency and writing proficiency.

Title	Type	Time	Reliability	Level
		General language proficiency		
Test of English as a Foreign Language (TOEFL)	objective paper + pencil test; 4 scores: listening comprehension, structure and written expression, reading and vocab., total.	150 min.	reliability for 3 sections: .87 to .89, and for total score: .95. (Loyd 1985)	applicants from non-English countries to American colleges.
Michigan Test of English Language Proficiency (MTELP)	objective items on grammar, vocab. and reading comprehension; part of battery which tests writing plus aural or oral skills.	75 min.	K-R 21 reliability coefficients for 6 forms: .92 +.(Cervenka 1978)	college applicants from non-English countries.
Secondary Level English Proficiency Test (SLEP)	150 objective m/c items; 3 scores: listening comprehension, reading comprehension, total.	90 min.	reliability for two sections: .94 and .93 and for total score: 96. (Loyd 1985)	ESL students entering grades 7-11.
Modern Language Aptitude Test-Elementary (EMLAT)	objective; 5 scores: hidden words, matching words, finding rhymes, number learning, total.	75 min.	reliability estimates for various sections and total score: .93 to .96 (Hakstian 1972)	ESL speakers: grades 3-6.

(continued)

(continued)

Oral proficiency

Test of Spoken English (TSE)	responses taped and scored by Educ. Testing Service; 4 subjective scores: overall comprehensibility (range 0-300), pronunciation, grammar and fluency each on a scale 0.0 to 3.0	20 min.	interrater reliability estimates: from .87 to .92 (Subkoviak 1985)	adult speakers of ESL
Ilyin Oral Interview (IOI)	question/answer; each of 50 answers scored on 0-3 scale for appropriateness, grammaticality and intelligibility.	5-30 min.	no inter-rater/intra-rater data cited. (Guyette 1985)	junior high, secondary and adult ESL speakers.
Foreign Service Interview (FSI)	face-to-face interview; subjective holistic ratings; scale of 1-5.	20-40 min.	interrater reliability from .82 to .93 (Adams 1978) (Reschke 1978)	prospective government employees.

Writing proficiency

College Board Achievement Test in English Composition	objective; 90 5-choice items.	60 min.	reliability coefficient of .91 (Harris 1978)	college entrance.
Test of Written Language (TOWL)	3 objective subtests; 3 subjective subtests on spontaneous writing of story.	40 min.	interrater reliability for 3 subjective subtests: .93, .98, .76 (Williams 1985)	grades 3-12.

Before comparisons of reliability estimates are made, one should perhaps be reminded that when one talks about a test being objective or subjective, one refers to the procedure involved for scoring such tests. As can be seen, the reliability figures listed in Table 1 are consistently high for all three types of language proficiency (general, oral and writing) and for both objective and subjective types of test.

Let us first consider the objectively scored tests of general proficiency. With regard to the TOEFL, "the reliabilities reported for the three sections (.87 to .89) and for the total score (.95) are reasonably high." (Loyd 1985:1569) The MTELP estimates are "based on the scores of six groups of 150 randomly selected applicants to U.S. higher institutions from a variety of language backgrounds. For the six forms treated here, K-R 21 reliability coefficients were .92 or greater." (Cervenka 1978:190) For the SLEP, internal consistency reliabilities "based on the data from a group of 326 students tested at international test centers .... are .94 and .93 for the two section scores respectively and .96 for the total score." (Loyd 1985:1336) Finally, the EMLAT estimates are "based upon independently administered and timed half-tests, with the resulting correlations corrected with the Spearman-Brown formula. The obtained reliabilities, data on which were gathered from four schools, and which were estimated at

each of the four grade levels for each sex, are extremely high, ranging from .93 to .96." (Hakstian 1972:95)

One might suggest that the reliability estimates for the tests of general language proficiency are uniformly high because those tests are objective in nature, and test results can therefore be scored with a certain consistency. As well, one can see that the range of proficiency among subjects for computing reliabilities is wide, much wider than the range of proficiency one usually finds in a particular classroom of second language learners, as is the case for speakers used in the study reported here. One suspects that a wider range of proficiency among subjects would allow distinctions between them to be made more effectively, thus enhancing reliability estimates.

For the purpose of this paper, it is the reliability estimates for the subjective test of writing proficiency and most especially, the subjective tests of oral proficiency which merit closer attention, precisely because of their subjective nature.

Let us consider briefly the reliability of the two tests of writing proficiency mentioned in Table 1. The objective College Board Achievement Test reports "for an unspecified form of the English test - a reliability coefficient of .91." (Harris 1978:136) It is perhaps because it does not address the ability to write above the sentence level and because of its objective nature that the test enjoys a high degree of reliability. The Test of

Written Language (TOWL), a test which includes three objective and three subjective subtests, achieves high degrees of interrater reliability for each of the subjective subtests (.93, .98 and .76). It should be noted, however, that the scope of the study upon which these figures were based was somewhat limited; "both the number of scorers [15] and the number of stories [15], five at each of the three grade levels [3, 5 and 7], seem small." (Williams 1985:1603) As for the tests of general proficiency, the two tests of writing proficiency cited in Table 1 achieve reliability estimates based on studies involving groups of subjects whose ranges of ability are wide compared to the range of ability of subjects comprising the groups used in the study reported in this paper. Again, this may account in part for high estimates of reliability.

Let us next consider the reliability of tests of oral proficiency, the most important information in Table 1 for the purpose of the study reported here. It seems surprising, initially, that reliability figures for subjectively rated oral tests should be so similar to the reliability figures quoted for the objectively scored tests. For example, the Test of Spoken English (TSE) reliability estimates, based on a study of 134 foreign teaching assistants, fall within the range .87 - .92. "These interrater reliabilities are high for a subjective scoring procedure." (Subkoviak 1985:1593) It should be

noted, however, that TSE scores, scores resulting from subjective judgements, are all submitted to the Educational Testing Service for scoring; as a result, one would anticipate a certain consistency in scoring procedure and therefore high degrees of interrater reliability. Also, the judgements, while subjective, are based on speaker responses which have been elicited via taped and therefore consistent cues. Responses are similar in orientation, so comparisons of speech samples can presumably be established more effectively. As well, the range of ability of speakers is potentially wide. It should also be noted that final scores are not the result of an exclusively impressionistic judgement but rather the result of four combined scores reflecting various aspects of production.

The Ilyin Oral Interview (IOI), is another commonly-used subjective oral test. Unfortunately, "the author of the IOI has not demonstrated that scorer agreement can be obtained. Neither interjudge nor intrajudge reliability data are presented." (Guyette 1985:677) Any comment regarding rater reliability is therefore unwarranted. This test, though, is similar to the TSE in format, in elicitation cues and in breakdown of rater scores according to various aspects of a speaker's oral production.

Raters of the Foreign Service Interview (FSI) test, unlike raters for the TSE and IOI, make holistic judgements according to overall oral proficiency of speakers on the basis of a face-to-face interview. Also - and this is of

particular importance - like the subjectively scored samples of speech for the TSE and of writing for the three subjective subtests of the TOWL, FSI speech samples reflect a wide range of speaker ability and are the result of considerable test time, between twenty and forty minutes.

It would seem reasonable to suppose that such test lengths would elicit oral (in the case of the TSE or FSI) or written (in the case of the TOWL) samples of language that might offer raters more opportunity for reliable judgements than samples of language which are much shorter. Because none of the FSI reliability estimates is the result, specifically, of investigation into the effect of speech sample length on interrater and intrarater reliability, one can only speculate on that relationship. The FSI reliability figures, though, are of particular interest for the study reported in this paper. Estimates are based on judgements of relatively long segments of speech; they are also the result of holistic judgements of oral proficiency using a face-to-face interview as elicitation procedure. As will be described, speech samples used in the study reported here are comparatively short at 30- , 60- and 120 seconds, but they too are subjected to holistic rater judgements and are the result of a face-to-face "interview" procedure.

The reliability studies conducted by the U. S. Foreign Service Institute therefore warrant close attention. The Institute has investigated the interrater reliability of

judgements based on FSI interviews which typically last from twenty to forty minutes. These reliability studies are all based on results of FSI oral interview tests in which participants include a candidate, an interviewer and an examiner. The test procedure involves an examiner who is in charge of the test and an interviewer who directs the conversation. The proficiency scale used consists of eleven points, 0 (no proficiency), 0+, 1, 1+ ..... 4, 4+, 5 (native or bilingual proficiency). The examiner and the interviewer assign ratings independently. Two ratings which differ by half a point result in the lower rating being assigned. Two ratings which differ by a full point result in an arbitration procedure with the head of the testing unit.

One FSI reliability study involved twenty-one examiners and interviewers (6 in French, 4 in German and 11 in Spanish) who listened to and rated independently approximately fifty test tapes each. Tapes varied in proficiency level from 0 to 5. Correlations between pairs of raters for each of the languages mentioned ".... in all cases exceeded .82, with the average correlation .91." (Adams 1978:135)

A previous reliability study conducted by the FSI in 1973 also involved rater judgements of French, German and Spanish oral interviews. Taking all judgements into account, the study "yielded a reliability coefficient of .85. Other in-house reliability studies conducted by the

FSI, which were limited to only one language, have produced similar results, with one study, based on French tests given, showing a reliability coefficient of .93." (Reschke 1978:78)

It must be remembered that in the above-mentioned FSI studies,

a) the speech samples used were of between twenty and forty minutes in length, and

b) the eleven-point proficiency scale used (0, 0+, 1, 1+, ... 4, 4+, 5) reflected performances ranging from "no proficiency" (0) to "native or bilingual proficiency" (5).

These two factors render the FSI findings of little practical value to teachers who must operate within a more confined context. First of all, given the time, personnel, and resource constraints of the average classroom test situation, whether such testing be for evaluation or placement purposes, an examiner must often base judgements of oral proficiency on much shorter samples of speech. Secondly, it would seem comparatively easy to differentiate between oral proficiency levels which range from "no proficiency" to "native proficiency"; if one is asked to rate a group of learners whose abilities fall within such a wide range, one would anticipate reliable ratings based on widely discernible differences in proficiency levels. Most teachers, however, must make much finer distinctions between students who are part of a fairly homogeneous group of second language learners. In short, it would seem to be

more difficult to provide reliable proficiency ratings for a class of students whose linguistic abilities fall within a much narrower range along the second language oral proficiency continuum.

The following study considers these two crucial factors. To begin with, relatively short samples of speech (30, 60 and 120 seconds) are used in the rating procedure. As well, the speech samples are elicited from two relatively homogeneous groups of second language learners, each group being made up of speakers who would undoubtedly not fall within so wide a proficiency range as that reflected in the FSI rating scale of 1 (no proficiency) to 5 (native or bilingual proficiency). In fact, all students of one group were considered to be at a "beginner" level while all students of the other group were considered to be at an "intermediate" level.

As well as the comparative general effect of speech sample length on reliability, two major comparisons were anticipated:

1. a comparison of pooled (n=6) rater reliability, interrater reliability and intrarater reliability, and
2. a comparison of reliability of judgements of speakers at a "beginner" level and speakers at an "intermediate" level.

Chapter 2

Method

A. Speakers:

In order to investigate reliability of rater judgements for more than one age group and for more than one general level of second language oral proficiency, speech samples were collected from two groups.

Group Y

A Young Group (Group Y) was composed of twelve young native French speakers who had participated in an experimental, innovative and intensive five month (September 1984 - February 1985) Grade 5 ESL program, and whose previous formal exposure to English had been minimal. Range of age was ten to eleven years; seven speakers were female and five were male. These twelve students were part of a larger class of twenty-six students. Class members had been randomly selected from approximately 150 Grade 5 applicants to the intensive program. Children already fluent in English and children with severe learning or emotional disabilities had been excluded from the program. A questionnaire administered to the twenty-six students revealed that although most of them listened to English music or watched English television every day, the large majority had no occasion to speak English either at home with other members of their families or with friends. Exposure to prior ESL instruction had been one hour per

week from Grade 1. Two of the Group Y speakers, however, had been exposed to English schooling prior to the intensive program, one in kindergarten, the other in Grades 1, 2 and 3. "Intensity" of instruction involved twenty and a half hours per week for five months. Speech samples of all twenty-six students were collected during three taping sessions conducted within three weeks after completion of the intensive program. The elicitation procedure for all twenty-six students involved the Picture Card Game, which is described in Data Collection Procedure. Twelve students provided relatively uninterrupted speech samples of at least 180 seconds. These twelve students were drawn from the larger group and constitute Group Y.

Group O

Twelve Older speakers (age range 20-30 years; mean age 23.5; 8 male, 4 female) constitute Group O. All twelve speakers were native French speakers, ten from the province of Quebec, one from Manitoba and one from Zaire. All were members of a larger class of students enrolled in ESL C207 English Language - Intermediate 1, a Concordia University English course described in the University catalogue as being "for students who are not native speakers of English and who need further training in the effective use of English in the university setting." Placement in ESL C207 was on the basis of Test of English as a Foreign Language

(TOEFL) and/or Concordia English Diagnostic Language Test (CEDLT) results.

B. Data Collection Procedure:

Group Y

Because of the age of Group Y speakers, and because their exposure to English was more limited than that of Group O speakers, the elicitation procedure for Group Y involved a task which, it was hoped, would prompt speech samples of sufficient length and with few interlocutor interruptions. An extension of the Picture Card Game (PCG) used by Upshur (1971) and by Lightbown and Spada (1978) was therefore used to elicit speech from Group Y speakers.

The PCG involves an interlocutor and a subject. The subject is offered a set of four cards from which one card must be chosen. The interlocutor has facing him/her a set of four cards identical to the set of four offered to the subject. The four cards feature a graphic depiction of an object or event, each card differing only in some small detail from the other three. The subject is asked to talk about the picture s/he has chosen and told that on the basis of that description or narrative, the interlocutor will try to guess which one of the four cards the subject has chosen. The same procedure is followed for each set of four cards.

The eight original sets of cards used by Lightbown and Spada (1978) were also used for this data collection, but

they were used together with two additional sets (see Appendix A for examples of original and additional sets). Each new set of four cards was used following exactly the same game "rules" as those for the eight original sets. The two new picture sets, however, were much more complicated graphically in order to provide more detailed visual contexts, contexts which would prompt sufficiently lengthy, relatively uninterrupted samples of speech.

The order of presentation of each set of cards was the same for all speakers. The two more complicated visual cues were always the ninth and tenth picture sets to be played. As well as providing consistency in procedure, maintaining order of presentation meant that, by the ninth and tenth picture sets, the speaker was familiar with the game and was at ease playing it. As anticipated, it was these last two picture sets that provided speech samples which were of sufficient, relatively uninterrupted length. Twelve of the twenty-six students provided speech samples of at least 180 seconds based on picture sets nine and ten. These twelve students constitute Group Y. Speech samples were edited to eliminate the few interlocutor interruptions present, but hesitations, pauses and false starts of speakers were retained.

Each of the twelve Group Y speakers provided at least 120 seconds of speech based on picture set nine and at least sixty seconds of speech based on picture set ten (or vice versa). The 120-second segment (for example, from

picture set nine) was then broken down to provide a 30-second, a 60-second and a 120-second segment. The 60-second segment (for example, from picture set ten) was broken down to provide a 30-second and a 60-second segment. Figure 1 illustrates the procedure. Appendix B offers transcriptions corresponding to each segment for each speaker.

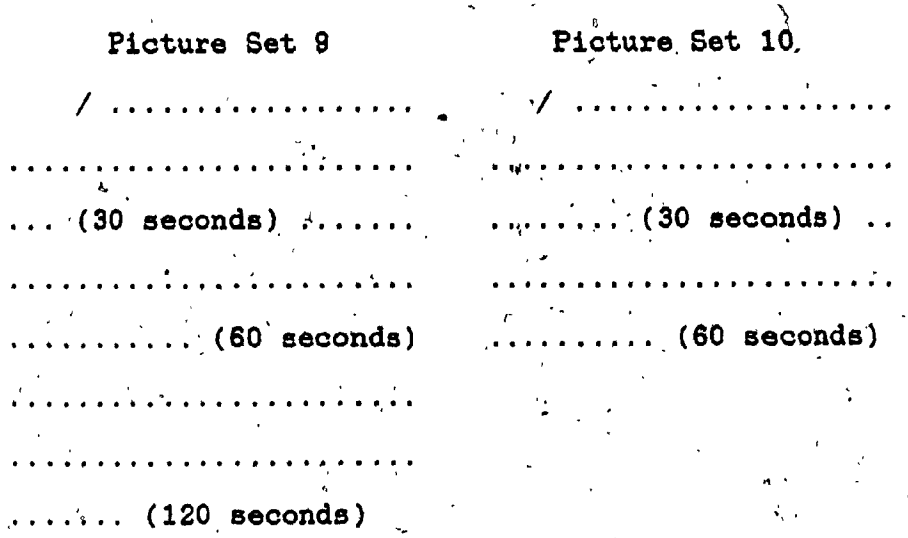


Figure 1. Graphic representation of transcriptions corresponding to picture sets 9 and 10

Two 30-second samples, two 60-second samples and one 120-second sample were thus obtained for each speaker in Group Y.

Group O

Taped speech samples of Group O speakers were kindly provided by Kate Owens, an instructor of ESL C207, whose elicitation technique had stipulated only that subjects decide on and talk about a topic of interest. These samples

provided the data, again with minimal editing, for two 30-second samples, two 60-second samples and one 120-second sample for each of the twelve Group O speakers. The breakdown of taped speech segments followed the same procedure as for Group Y in Figure 1. However, it must be remembered that while speech samples for Group Y speakers were all based on the same visual cues and therefore similar in content, the speech samples for Group O speakers reflected varied topics of conversation. Appendix B offers transcriptions corresponding to each segment for each Group O speaker.

C. Raters:

Students from three Concordia University TESL summer session classes (1985) were asked to participate in the research project. Two of these classes were introductory, one in Comparative Phonetics (TESL C221/1), the other in Modern English Grammar (TESL C231/1); the third was a Methodology class (TESL A298/1) offered to teachers who already had either a minimum of two years ESL teaching experience or credit in a TESL methodology class. Of the twenty-three volunteers, five were male, eighteen were female. Eighteen of the twenty-three raters designated English, while four listed French, to be their present dominant language. One rater designated Spanish to be her first and present dominant language.

Six raters were current full-time teachers of ESL;

eight raters were current part-time teachers of ESL. ESL teaching experience among raters ranged from a few months to twenty-three years, with a mean of 5.4 years.

Six raters designated their status to be full-time TESL undergraduate students while nine designated their status to be part-time. One rater designated her status as graduate student. Exposure to TESL-related classes ranged from a few weeks to four years.

For rating purposes, these wide differences in ESL teaching experience and TESL-related academic history were considered of minor importance. In fact, similar disparities exist in many second language evaluative situations where judgements must still be made regarding the oral proficiency of second language learners.

D. Rating Procedure:

As mentioned, twenty-three volunteers participated in the rating procedure. Of these, seventeen were randomly assigned to perform one rating task while six were randomly assigned to perform another rating task. A description of each task and the rationale for the use of a second task follow.

Rating Task 1

As can be seen from Figure 1, taped samples for each speaker from Groups Y and O included two 30-second samples, two 60-second samples and one 120-second sample. Five

raters were randomly assigned to rate the 30-second speech samples of Groups Y and O, while six raters performed the rating task for the 60-second samples and six raters did the same for the 120-second samples. For each length condition, in order to minimize order effect, half of the raters first rated Group Y samples while the other half first rated Group O samples. (Because five raters, rather than six, were assigned to the 30-second condition, three raters listened to Group Y samples first while only <u>two</u> listened to those of Group O first. However, for the 60- and 120-second conditions, three raters rated Group Y samples first while three raters rated Group O samples first.) Appendix C shows the random order of presentation of samples for each condition of length.

Speech samples of the least orally proficient speaker and the most orally proficient speaker from each Group were selected independently by the investigator and two colleagues, and these samples were used as bipolar models for each length condition. Each series of taped samples was then prefaced by the model of low proficiency and the model of high proficiency of pertinent length for each Group.

The rating procedure was as follows:

1. Raters were given a rating sheet (see Appendix D for examples). At the top of the sheet were two 11 cm lines. The first line had an intersecting mark at its extreme left hand side corresponding to the model speech sample of low proficiency for the Group in question. The second line had

an intersecting mark at its extreme right hand side indicating the highest rating possible. This highest rating corresponded to the model of high proficiency also heard at the beginning of the tape.

2. Raters were told that after hearing the two model speech samples, one low and one high, they would hear a series of speech samples of the same length as the two model samples. The rating sheet provided had an 11 cm line corresponding to each of the speech samples to be heard. Raters were asked to listen to each sample and assign a rating, on the basis of overall oral proficiency, to each sample by marking where, along an 11 cm line, that speaker might fall. For the 30-second and 60-second conditions, there were two speech samples per Young and Old speaker, which meant that for these length conditions, raters made twenty judgements for Young speakers and twenty judgements for Old speakers, two judgements per speaker. For the 120-second condition, where there was only one speech sample per speaker, raters made ten judgements for Young speakers and ten judgements for Old speakers, one per speaker.

3. As mentioned, raters first rated the ten or twenty speech samples of one Group. The tape was then flipped over to the other side where the ten or twenty speech samples of the other Group could be heard. The rating procedure, using a model of low proficiency, a model of high proficiency and another series of 11 cm lines, remained the same as the procedure adopted for the first Group rated. Intersecting

marks were subsequently assigned numerical values to the closest .5 cm.

During a pilot rating procedure using Rating Task 1, there emerged a potential threat to the validity of the procedure for ratings of 30-second and 60-second samples for Group O. It will be recalled that Group O speakers were asked to talk about a topic of personal interest. This meant that the rater might recognize Speaker X on hearing Speaker X's second speech sample, either because of voice quality of Speaker X or because of Speaker X's topic of conversation. The result would be assignment of the same rating for the second sample as for the first for Speaker X. Raters in the pilot procedure, however, agreed that it was not until they had heard most of the samples for Group O that they realized there were two samples per Group O speaker. Furthermore, it was agreed that the rating task itself, involving only an 11 cm line, frustrated easy recall of previously assigned ratings.

Rating Task 2

Because the first rating task meant that raters were asked to use the same scale (an 11 cm line) for both Groups, it was natural for raters to consider the same designated range from lowest to highest level of proficiency for both Groups. The "lowest" level for each Group had, after all, been designated by an intersecting mark at the extreme left hand side of one line, and the

"highest" level had been designated by an intersecting mark at the extreme right hand side of a second line. Although speakers in Group Y, for example, might have been perceived as falling within a narrower range of proficiency, raters would still tend to use the entire rating scale (the entire 11 cm line) while assigning judgements. Ratings of proficiency of Group Y speakers would therefore end up being marked along a scale that was identical to the scale used for Group O speakers, a group whose actual range of proficiency might well have been perceived by the raters to be much wider than the range of proficiency for Group Y. In short, observed scale differences between speakers might have been a result of the rating task.

This complication was anticipated, so a second rating task was devised for the 60-second sample length condition. This time, six raters were requested to listen to two 60-second speech samples. The first sample was of a randomly chosen speaker from Group Y. The sample had been assigned the arbitrary number 26. The second sample was of a randomly chosen speaker from Group O. That sample had been assigned the arbitrary number 114. Raters were subsequently asked to listen to randomly ordered 60-second samples of the remaining eighteen speakers (nine from Group Y, nine from Group O), and to assign a number to each speech sample on the basis of the arbitrarily assigned numbers (26 and 114). See Appendix D for an example of the rating sheet used for Rating Task 2.

It was speculated that a low perceived range of proficiency for Group Y speakers would result in a relatively tight cluster around the fixed number 26, and that a comparatively high perceived range of proficiency for Group O speakers would result in a correspondingly wider range of numbers below and above the fixed number 114.

Chapter 3

RESULTS

A. Pertaining to Rating Task 1

How clearly would the differences in speakers' oral proficiency be seen by the raters? How precisely would the raters agree in their judgements of oral proficiency of speakers?

Estimating how much agreement exists in rater judgements will offer a measure of how reliable those ratings are. A reliability coefficient will serve as a measure of the precision with which raters find true differences in proficiency among speakers. Before estimates of reliability can be made, though, one must be aware of the key concepts of true scores, true variance and error variance.

Consider the speaker who is rated for overall oral proficiency on a particular occasion by a particular rater. Applying the helpful explanation of Krzanowski and Woods (1984), the rating the speaker gets is a random rating from a whole population of possible ratings, a "superpopulation" of ratings. The true score for this speaker can be defined as the mean of this superpopulation of ratings and as such, is a purely hypothetical score. Equally hypothetical is true variance, an index of the true variability which exists among a group of speakers. It is impossible to establish true scores, and therefore true variance, because

the ratings each speaker receives are limited. The best one can do is estimate from samples what true score and true variance might be. The estimated true score of a speaker is then calculated to be the mean of all of the assigned ratings for that speaker, and estimated true variance is an index of the variability calculated to exist among the true scores of a group of speakers.

In order to determine the reliability of rater judgements, one must try to establish how much of the total variance estimated to exist among speakers is due to real differences in proficiency (true variance) and how much is due to chance error (error variance). There is bound to be error variance in any situation where rater judgements are made. All raters will not agree completely regarding the variation of oral proficiency of speakers within a group. There will be errors of measurement. If such error variance is found to be as large, for example, as the true variance estimated to exist among speakers, then reliability of rater judgements must be considered to be low. If, however, true variance is shown to account for most of the total variance, then one can make stronger claims about the reliability of rater judgements.

In order to investigate reliability, one must be able to quantify true variance and error variance. Analysis of variance (ANOVA) will do exactly what its title states; it will analyze sources of variance. It will assign a proportion of total variance to the true variance estimated

to exist among speakers and will also assign a proportion of total variance to various sources of error variance. An ANOVA table will provide "source of variation" information with numerical values for the percentage of true variance among speakers and the percentage of error variance due to one or several sources. Sources of variance are listed in the ANOVA tables (see Appendix E for all ANOVA tables) as either MAIN EFFECTS or INTERACTION EFFECTS, references to the effect of variables involved in the analysis. In the analyses of variance used in this study, the possible MAIN EFFECTS are speaker, rater and rating occasion. An INTERACTION EFFECT is simply a combination of any two of these main effects or all three.

The effect of speaker on total variance is reflected in a numerical value which is an estimate of the real variation existing among speakers; as such, that numerical value must be considered the estimate of true variance. Variance, however, due to inconsistencies among raters or between two rating occasions must be considered error variance and numerical values will be assigned proportionately to these effects. It follows that any interaction effect will also be assigned a numerical value which reflects a proportion of error variance due to that interaction. It will become clear in the following discussion when and why a 2-way ANOVA (speaker, rater) or a 3-way ANOVA (speaker, rater, rating occasion) was used.

Having obtained from ANOVA tables numerical values for

true variance and error variance, one can then substitute those values in a ratio which defines reliability. The resulting reliability coefficient will in turn make possible claims concerning the reliability of rater judgements in question.

The ratio defining reliability is $\frac{V_t}{V_t + V_e}$

where V_t refers to true variance, and

V_e refers to error variance.

Looking at the formula, it becomes clear that if most of the total variance is due to true variance, then reliability will have a value close to 1. If, however, most of the variance is due to error variance, reliability will have a value close to 0. A reliability coefficient of .87, for example, would indicate that 87% of the estimated total variance is due to true variance among speakers and the remaining 13% is attributable to error variance.

The reader will recognize in the following discussion that the error variance (V_e) component of the ratio incorporates different sources of error variance, depending on the aspect of reliability being investigated. This study considers three such aspects of reliability:

i) the reliability of judgements of a speaker made by a pool of raters,

ii) the reliability of a judgement of a speaker made by an individual rater and a judgement of the same speaker made by any other rater (interrater

reliability), and

iii) the reliability of two judgements of a speaker made by any one rater (intrarater reliability).

Each aspect of reliability will be considered in turn.

i) the reliability of pooled ratings:

First consideration will be given to the comparative reliability of pooled ratings for each designated subgroup. Subgroups will be referred to as 30-second Young, 60-second Young, 120-second Young, 30-second Old, 60-second Old and 120-second Old.

It must be remembered that raters judged two samples per speaker for the 30-second and 60-second conditions, while raters judged only one sample per speaker for the 120-second condition. Therefore, for each of the subgroups 30-second Young, 60-second Young, 30-second Old and 60-second Old , three separate 2-way (speaker,rater) analyses of variance were performed. Three separate ANOVAs were performed for each of these subgroups in order to allow a comparison of reliability for pooled ratings where

a) both speech samples per speaker were included in the ANOVA,

b) only the first speech sample per speaker was included in the ANOVA, and

c) only the second speech sample per speaker was included in the ANOVA.

For each of the remaining two subgroups (120-second Young,

120-second, Old), only one ANOVA was necessary because raters judged only one speech sample per speaker.

Resulting ANOVA tables listed source of variation information with numerical values for:

- estimated true variance between speakers,
- estimated error variance due to the overall interaction effect of speaker and rater, and
- estimated error variance due to the effect of individual raters.

Because interest at this point was only in pooled ratings per subgroup, it was "estimated error variance due to the overall interaction effect of speaker and rater" which would reveal how reliable pooled ratings might be. The "estimated error variance due to the effect of individual raters" was not of immediate concern but would be used later in addressing the reliability of interrater agreement.

In calculating reliability of pooled rater judgements per subgroup then, the reliability formula

$$\frac{V_t}{V_t + V_e}$$

could be altered to read

$$\frac{V_s}{V_s + V_{s \times r}}$$

where V_s was the estimated true variance between speakers, and

$V_{s \times r}$ was the estimated error variance due to the

overall interaction effect of speaker and rater.

The reliability estimates for pooled ratings are presented in Table 2. As mentioned, results of all ANOVAs are included in Appendix E.

TABLE 2

Reliability of pooled ratings for two groups of speakers with six raters[a]

	Young	Old
30-second condition		
both samples:	.67	.69
sample 1 only:	.70	.66
sample 2 only:	.60	.70
60-second condition		
both samples:	.83	.91
sample 1 only:	.82	.88
sample 2 only:	.74	.85
120-second condition		
	.85	.82

[a] with the Spearman Brown prophecy formula applied to the 30-second condition where rater n=5.

ii) Interrater reliability:

The previous section reported findings which addressed the reliability of pooled ratings for each subgroup. Investigating the reliability of judgements between individual raters per subgroup meant a return to the ANOVA tables which listed source of variation information. By retaining the numerical values for "the estimated true variance between speakers" and "the estimated error variance due to the interaction effect of speaker and rater" and including "the estimated error variance due to the effect of individual raters", one could investigate how inclusion of the latter effect would alter reliability. In short, one could investigate interrater reliability. Expansion of the V_e component of the reliability ratio to include the portion of error variance due to inconsistencies of individual raters would mean changing the original ratio to read:

$$\frac{V_s}{V_s + V_{s \times r} + V_r}$$

where V_s was the estimated true variance between speakers,

$V_{s \times r}$ was the estimated error variance due to the overall interaction effect of speaker and rater, and

V_r was the estimated error variance due to the effect of individual raters.

Substitution in the ratio of appropriate numerical values

from the ANOVA tables per subgroup yielded estimates of interrater reliability, which are presented in Table 3.

TABLE 3

Interrater reliability coefficients for two groups of speakers

	Young	Old
30-second condition		
both samples:	.23	.27
sample 1 only:	.37	.22
sample 2 only:	.18	.39
60-second condition		
both samples:	.12	.20
sample 1 only:	.17	.23
sample 2 only:	.12	.17
120-second condition	.43	.62

iii) Intrarater reliability:

As mentioned, for the 30-second and 60-second conditions, two speech samples per speaker were rated. When one rater judged two speech samples of the same speaker, how much agreement was there between the two rating occasions? Investigating that agreement would reveal degrees of intrarater reliability for the four subgroups (30-second

Young, 60-second Young, 30-second Old, 60-second Old) in question.

Just as the 2-way (speaker, rater) ANOVA revealed source of variation information with corresponding numerical values allowing one to estimate pooled-rater reliability and interrater reliability, a 3-way (speaker, rater, rating occasion) ANOVA would take into account all listed sources of error variance in which occasion had an effect, allowing one to estimate intrarater reliability. One would again simply break down the V_e component of the original reliability ratio which would now read

$$\frac{V_s}{V_s + V_o + V_{s \times o} + V_{r \times o} + V_{r \times s \times o}}$$

where V_s was the estimated true variance between speakers,

V_o was the estimated error variance due to the effect of the rating occasion,

$V_{s \times o}$ was the estimated error variance due to the interaction effect of speaker and occasion,

$V_{r \times o}$ was the estimated error variance due to the interaction effect of rater and occasion, and

$V_{r \times s \times o}$ was the estimated error variance due to the interaction effect of rater, speaker and occasion.

Again, source of variation information per subgroup was

retreived from the 3-way ANOVA tables and appropriate numerical values were substituted in the formula. Estimates of intrarater reliability are given in Table 4.

TABLE 4

Intrarater reliability coefficients per subgroup

	Young	Old
30-second condition:	.44	.38
60-second condition:	.53	.74

B. Pertaining to Rating Task 2

The description of, and reason for a second rating task have been presented (see Rating Task 2 of Rating Procedure). Briefly, because raters were asked to use an 11 cm unmarked line for ratings of both Groups, it was natural for them to make use of the entire scale (the entire 11 cm line), having been given a model low sample designated by an intersecting mark at the extreme left hand side of one 11 cm line and a model high sample designated by an intersecting mark at the extreme right hand side of another 11 cm line. The raters may have perceived a wider range of proficiency for one of the Groups, but that perception would have been masked by the rating task. The second

rating task, involving the two arbitrary numbers 28 and 114, allowed raters the opportunity to rate speakers from either Group according to a fairly narrow perceived range of proficiency or a comparatively wide perceived range of proficiency, simply by assignment of appropriate numbers.

It is interesting to look only at the estimated values reflecting "true" variance in the ANOVA tables for Rating Task 1. The mean square value in an ANOVA table is an index of the variation due to a particular effect. The mean square value for SPEAKER is then an index of the variation which has been seen to exist among speakers in a group. That numerical value says nothing, by itself, about the reliability of ratings because it does not include information about error variance, but it does indicate how much variation has been seen to exist among speakers. The higher the numerical value corresponding to that "true" variance, the wider the perceived range of proficiency for that group of speakers. Looking only at the mean square numerical values reflecting estimated true variance among speakers for 30-, 60- and 120-second sample lengths, one can extract from ANOVA tables the data included in Table 5. If one then calculates true variance among speakers of Group Y and Group O using ratings of Rating Task 2, one can see whether greater true variance values result from that rating task and whether a greater discrepancy exists between the numerical values for Groups Y and O using that rating task. In other words, one can see how Rating Task 2

has affected the perceived range of proficiency for either Group. The true variance estimates based on Rating Task 2 are also listed in Table 5.

TABLE 5

Variance of pooled ratings for speakers using Rating Tasks 1 and 2

	Young	Old
Rating Task 1		
30-second condition		
both samples:	13.183	15.272
sample 1 only:	14.155	9.929
sample 2 only:	7.036	13.544
60-second condition		
both samples:	12.187	25.598
sample 1 only:	8.683	17.248
sample 2 only:	7.644	10.500
120-second condition		
	29.535	28.494
Rating Task 2		
60-second condition		
	20.68	162.14

One might speculate that the reason the true variance estimates for both Groups Y and O using Rating Task 1 are

comparatively close in value is that raters used the entire scale in assigning ratings. As discussed, the perceived range of proficiency among speakers could have been masked by the rating task. Rating Task 2, however, allowed raters to assign ratings according to a fairly narrow or a comparatively wider range of proficiency.

While the true variance value for Group Y using Rating Task 2 is slightly higher than the true variance values using Rating Task 1 for that Group at 60 seconds, the same comparison for Group O shows a much more marked increase in perceived range of proficiency using Rating Task 2. In short, raters seemed able to establish a wider range in oral proficiency among speakers using Rating Task 2, but most especially for Group O.

One might hypothesize how the reliability coefficients calculated via the task involving the 11 cm line (Rating Task 1) would be affected if the perceived range of proficiency for Group Y had been as wide as the perceived range of proficiency for Group O, assuming constant error variance. Such hypothetical extension is called "disattenuation". The formula which allows for it (see Gulliksen: 1958: 111) is:

$$R_{xx} = 1 - \frac{s^2_x}{S^2_x}(1 - r_{xx})$$

where s^2_x and S^2_x represent variance of two sets of numbers, in this case, variance among speakers for Group Y [20.68] and Group O [162.14] respectively as a result of Rating,

Task 2, and where r_{xx} is the original reliability coefficient, in this case, the coefficient originally calculated for the Group being corrected for attenuation, Group Y.

Because there is little difference in the original reliability estimates (see Table 2) for pooled ratings using the 120-second condition, and there are inconsistencies for the pooled rater estimates based on the 30-second condition, it is for the 60-second condition, where Group O coefficients are consistently higher, that the correction for attenuation seems appropriate.

As will be recalled, for results of the 60-second condition rating task involving the 11 cm line, three separate 2-way ANOVAs were performed to allow a comparison of reliability for pooled ratings where

a) both samples were included in the analysis,

b) only the first sample was included in the analysis, and

c) only the second sample was included in the analysis.

Because only one sample per subject was rated in the task involving assignment of numbers, two separate values for r_{xx} (reflecting conditions b and c above) were substituted in the formula.

The following results were obtained:

1. The original reliability coefficient calculated for pooled ratings for the 60-second Young subgroup where only sample 1 was used in the analysis (.82) increased to .98.

2. The original reliability coefficient calculated for pooled ratings for the 60-second Young subgroup where only sample 2 was used in the analysis (.74) increased to .97.

Table 6 shows original reliability coefficients for pooled ratings of the 60-second Young subgroup, reliability coefficients for pooled ratings of the 60-second Young subgroup corrected for attenuation, and original reliability coefficients calculated for pooled ratings of the 60-second Old subgroup.

TABLE 6

Original reliability estimates for pooled ratings of 60-second Y and O subgroups and reliability estimates for 60-second Y subgroup corrected for attenuation.

	Group Y		Group O
	original	corrected	original
sample 1:	.82	.98	.88
sample 2:	.74	.97	.85

Chapter 4

Discussion

It is perhaps prudent to preface this discussion with a reminder that the findings of the study in question can be interpreted only on the basis of a limited number and type of speech sample, and that the reliability figures listed here can be considered only within those limitations. It is clear that the numbers of speakers and raters are limited. The elicitation procedures used to obtain speech samples and the scoring procedure are also peculiar to this study. Given those restrictions, one can not assume that the results of the study are generally, much less universally, applicable. On the other hand, one can use the results to hypothesize about conditions which might affect the reliability of rater judgements. One might also offer tentative, practical pedagogical applications of the findings and suggest implications which those findings may have for further research. Having qualified generalizability and applicability, let us address the comparative general effect of speech sample length on reliability and consider the two anticipated comparisons mentioned in Chapter 1:

1. a comparison of pooled (n=6) rater reliability, interrater reliability and intrarater reliability, and
2. a comparison of reliability of judgements of speakers at a "beginner" level and speakers at an "intermediate" level.

1. Pooled rater reliability, interrater reliability and intrarater reliability:

To begin with, the reliability coefficients for pooled ratings per subgroup (Table 2) prompt two major observations:

i) There is a trend toward higher degrees of reliability for Group O judgements, but there is nonetheless a striking similarity between reliability coefficients per speech sample length for Young and Old speakers despite expectations based on the relative heterogeneity of the Old Group.

ii) The reliability coefficients in Table 2 reflect a certain consistency in rank ordering of speakers by pools of raters. While some raters scored all speakers within a limited range at one end of the rating scale, and other raters scored all speakers within an equally limited range at the opposite end of the scale, the raters, despite inconsistencies in use of the scale, have still rated speakers in comparable rank orders.

In general, there is increased reliability with increased length of speech sample. For example, for the Young Group, there is less reliability using either 30-second sample (.70 or .60) than using either 60-second sample (.82 or .74). Both 30-second samples give rise to approximately the same reliability (.67) as either 30-second sample (.70 or .60), so that using one or two 30-second samples seems to make little difference in

reliability. Further, using both 30-second samples gives rise to less reliability (.67) than using either 60-second sample (.82 or .74). Finally, reliability based on the 120-second sample (.85) exceeds the reliability based on either 60-second sample (.82 or .74) or both 60-second samples (.83).

There is one exception to the trend of increased reliability with increased length of sample. For the Old Group, the best estimate of reliability seems to be the result of judging two 60-second samples (.91). Surprisingly, there is slightly higher reliability for either 60-second sample (.88 or .85) than for one 120-second sample (.82). It would seem that, for the Old Group, one 60-second sample might provide as much opportunity as one 120-second sample for reliable pooled judgements, and that two 60-second samples might provide better opportunity for reliable judgement than one 120-second sample.

These differences in pooled-rater reliability estimates seem minor, however, especially when the same pooled rater estimates are considered in comparison with interrater and intrarater estimates.

Comparing the reliability coefficients for pooled ratings (Table 2) with those affected by interrater differences (Table 3), one can notice a drastic decline in values. Estimates of interrater agreement based on speech segments of 120 seconds are more stable (.43 and .62) than estimates based on 30- or 60-second segments, but there

seems to be no pattern in the estimates of interrater reliability for the latter two conditions of length. Clearly, there is much less interrater agreement than agreement in the average of ratings assigned by pools of (n=6) raters.

Finally, one should consider the reliability coefficients resulting from the investigation into intrarater agreement (Table 4). Again, pooled ratings seem to give rise to much higher degrees of reliability than two ratings of a speaker performed by any one rater. Also, an individual rater will seem to agree more with his/her previously assigned rating of a speaker (intrarater agreement) than with the rating of that same speaker by another rater (interrater agreement). Higher intrarater estimates than interrater estimates would suggest that raters, when judging alone, used more consistent criteria for scoring. When an individual rater's scores were compared with those of any other rater, however, discrepancies in scoring criteria gave rise to greater differences in ratings.

One should also note that while there is no significant difference between the intrarater reliability estimates between the two Groups for the 30-second condition, there is a marked difference in the estimates for the 60-second condition between the two Groups. A possible explanation for this might be that when raters judged two samples of a Group Y speaker, there was a topic

change involved; one sample was the result of a description of a picture from picture set 9 and the other the result of a description of a picture from picture set 10, or vice versa. A speaker was perhaps more familiar with the vocabulary or the context of one of the two pictures. This could have resulted in a difference in facility or fluency of speech for one of the two pictures. This in turn might have affected the ratings of that speaker. When the raters judged two samples of a Group O speaker, no topic change was involved. It was perhaps easier for raters to maintain relative consistency in ratings when the topic remained constant, as it did for Group O speakers. The shorter length condition of 30 seconds may not have been of sufficient duration for this situation to affect intrarater reliability but the longer 60-second condition may have been. This would account for the substantially higher degree (.74) of intrarater agreement for the 60-second samples of Group O.

In general, results would suggest that, given relatively short speech segments of speakers who have been designated at the same general level of proficiency, a pool of six raters can make distinctions among those speakers which are much more reliable than the distinctions made by individual raters. For pooled ratings, rater differences seem to compensate for each other, so that although there are certain inconsistencies in rank ordering and estimated

differences in proficiency among speakers, there is still a high degree of general overall agreement. In keeping with the discussion of variance, one might be reminded that for pooled ratings, true differences (viz. true variance) among speakers can be established much more reliably than when any individual rater judgements are considered. For the latter rating condition, the error variance involved seems to preclude stable estimates of reliability.

Given the speech sample lengths investigated, one should not assume high reliability for holistic judgements by individual raters. This has both pedagogical and research implications.

The teacher who monitors the progress of a group of students over an extended period of time has the opportunity to make repeated evaluations and, presumbably therefore, more reliable judgements. It is certainly not unusual, however, for teachers of a second language to have to make judgements of overall proficiency on the basis of short segments of speech, particularly for placement purposes. Most teachers are not unfamiliar with a situation involving a group of second language speakers who have been "placed" in a particular class. A few of those speakers, suspected of being at the same general level as that of the rest of the class, have clearly been misguided. Such indiscretions in placement are perhaps minor because they are reversible; a student can be replaced in a setting more conducive to his/her level of proficiency.

What is more unsettling is the evaluative situation where individual raters must make judgements upon which a final grade, accreditation or further advancement depends. In any such situation, where the raters are unfamiliar with the students in question, it would seem that speech samples longer than those used in this study should be the basis for evaluation.

Findings of this study have implications for further research as well. Research data based on holistic judgements of speech samples which are less than 120 seconds in length and which have been rated by individual raters must be viewed with suspicion. Further research might suggest whether such single-rater judgements should perhaps be based on longer speech samples before they can be considered satisfactorily reliable.

2. The reliability of judgements for Group O speakers versus the reliability of judgements for Group Y speakers:

Results reveal a trend towards greater reliability estimates for Group O than for Group Y ratings. This applies to pooled ratings as well as individual ratings. Why should this be the case? On the basis of the reliability coefficients corrected for attenuation, it seems that given a wider range of perceived proficiency for Group Y, raters would probably have made judgements of proficiency for that Group which would be more comparable in reliability to the judgements of proficiency for Group O

speakers. It is reasonable to speculate that a more heterogeneous group of Y speakers would have prompted a more highly pronounced perceived range of proficiency, which would have generated a higher degree of reliability.

But why, then, would the Group Y reliability estimates corrected for attenuation be substantially higher than the original reliability estimates for Group O? The corrected reliability estimates allow one to speculate that if raters had perceived as wide a range of proficiency for Group Y as they did for Group O, original reliability estimates for Group Y would have been much higher. One wonders, though, why the Group Y corrected estimates are so much higher than the original reliability estimates for Group O (.98 and .97 versus .88 and .85 respectively). What could account for such a difference?

One must remember that the elicitation procedure for the two groups was different, resulting in varied topics of conversation for Group O speakers but consistent contexts for Group Y conversations. As well, Group Y speakers, at a "beginner" level, were part of a group whose language skills and strategies were less divergent and developed than those of Group O speakers, speakers at an "intermediate" level. Group Y samples reflected context and therefore vocabulary, and language skills, which were more uniform. Such a situation would perhaps have allowed raters the opportunity to use comparable criteria for scoring and would consequently have allowed differences between

speakers to be established more effectively. On the other hand, raters, when listening to samples of Group O speakers, would perhaps have tried to establish differences among speakers using more complex criteria. Freedom of selection of topic would necessarily have involved varied contexts and varied vocabulary; more developed language would have involved more complicated strategies. Raters, consequently, might have judged Group O samples according to more widely divergent criteria than Group Y samples. If one considers for Group Y a perceived range of proficiency comparable to that of Group O, combined with the effect of consistency in elicitation cues for Group Y and the general level of language proficiency of Group Y, one might account for the high reliability estimates corrected for attenuation for that group.

Such high corrected reliability estimates lead one to suspect that elicitation of speech should perhaps involve a procedure prompting speech samples which can be compared and judged using the same scoring criteria. The same visual or oral stimuli for verbal responses, for example, would elicit speech samples which contain similar descriptions, narratives and vocabulary. It seems logical to speculate that such a procedure would allow applications of consistent criteria in scoring and thereby enhance the reliability of rater judgements.

In conclusion, it is hoped that further research might be conducted using speech sample segments of the same length as those used in this study or of slightly longer lengths and that these samples be elicited from speakers who are deemed to be at the same general proficiency level. Such research might corroborate or put into question the validity of this study's findings. Until such research data can be reported, it is this investigator's tentative suggestion that holistic judgements of speech samples of less than two minutes in length, made by individual raters, be accepted with caution.

References

Adams, M. 1978. Measuring foreign language speaking proficiency: A study of agreement among raters. In John L. D. Clarke (Ed.), Direct testing of speaking proficiency, theory and application. Princeton, N.J.: Educational Testing Service, 129-149.

Cervenka, E. 1978. Review of Michigan Test of English Language Proficiency. In O.K. Buros (Ed.), The Eighth Mental Measurements Yearbook. Highland Park, N.J.: The Gryphon Press, 188-190.

Gulliksen, H. 1958. Theory of Mental Tests. New York, N.Y.: John Wiley & Sons, Inc.

Guyette, T. 1985. Review of Ilyin Oral Interview. In J.V. Mitchell, Jr. (Ed.), The Ninth Mental Measurements Yearbook. Lincoln, Nebraska: The Buros Institute of Mental Measurements, 677-678.

Hakstian, R. 1972. Review of Modern Language Aptitude Test - Elementary. In O.K. Buros (Ed.), The Seventh Mental Measurements Yearbook. Highland Park, N.J.: The Gryphon Press, 94-96.

Harris, D. 1978. Review of College Board Achievement Test in English Composition. In O.K. Buros (Ed.), The Eighth Mental Measurements Yearbook. Highland Park, N.J.: The Gryphon Press, 135-136.

Krzanowski, W. and Woods, A. 1984. Statistical aspects of reliability in language testing. Language Testing, vol.1, no. 1, 1-20.

Lightbown, P. and Spada, N. 1978. Performance on an oral communication task by Francophone ESL learners. SPEAQ Journal, vol.2, no. 4, 35-54.

Loyd, B. 1985. Review of Secondary Level English Proficiency Test. In J.V. Mitchell, Jr. (Ed.), The Ninth Mental Measurements Yearbook. Lincoln, Nebraska: The Buros Institute of Mental Measurements, 1335-1336.

Loyd, B. 1985. Review of Test of English as a Foreign Language. In J.V. Mitchell, Jr. (Ed.), The Ninth Mental Measurements Yearbook. Lincoln, Nebraska: The Buros Institute of Mental Measurements, 1568-1569.

Reschke, C. 1978. Adaptation of the FSI interview scale for secondary schools and colleges. In John L.D. Clarke (Ed.), Direct testing of speaking proficiency, theory and application. Princeton, N.J.: Educational Testing Service, 77-88.

Subkoviak, C. 1985. Review of Test of Spoken English. In J.V. Mitchell, Jr. (Ed.), The Ninth Mental Measurements Yearbook. Lincoln, Nebraska: The Buros Institute of Mental Measurements, 1592-1593.

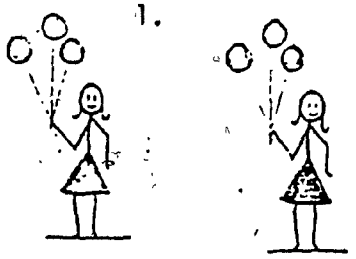
Upshur, J.A. 1971. Objective evaluation of oral proficiency in the ESOL classroom. TESOL Quarterly, 5, 47-59.

Williams, R. 1985. Review of Test of Written Language. In J.V. Mitchell, Jr. (Ed.), The Ninth Mental Measurements Yearbook. Lincoln, Nebraska: The Buros Institute of Mental Measurements, 1602-1604.

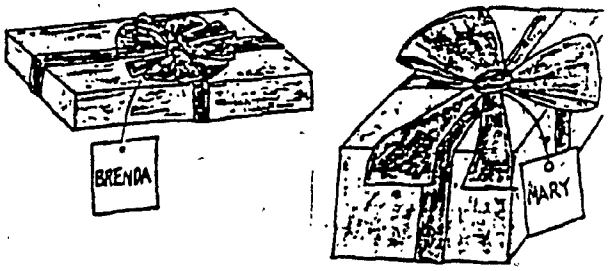
Appendix A

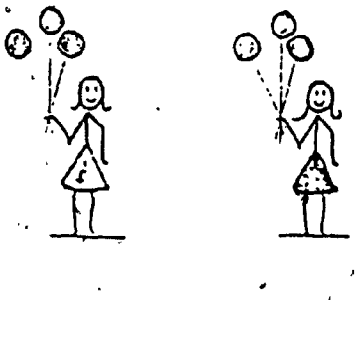
Eight original sets of PCG cards.

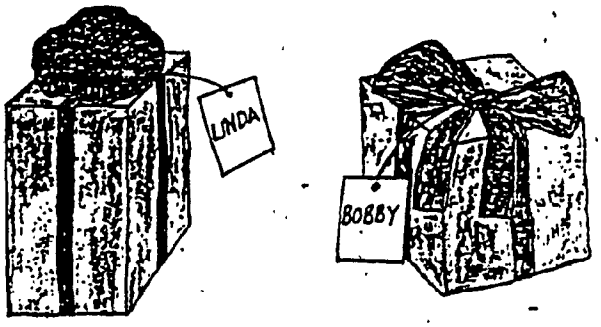
1.



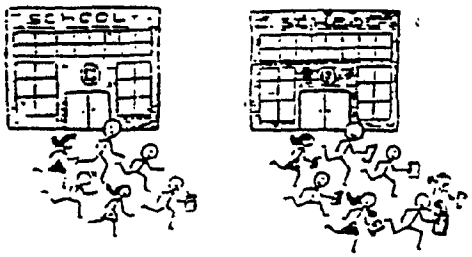
2.



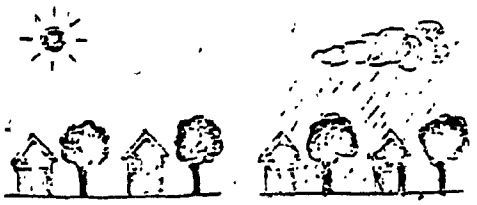


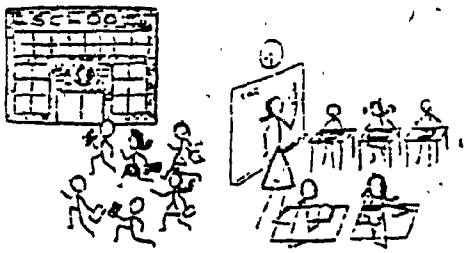


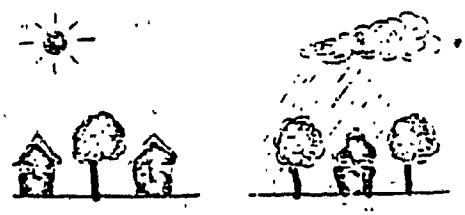
3.



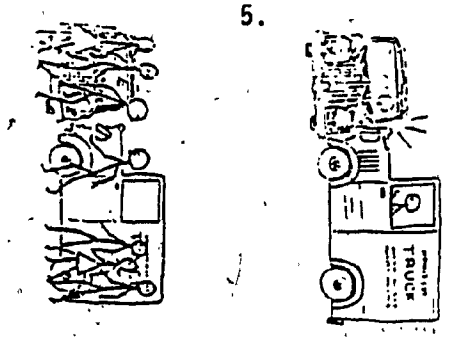
4.

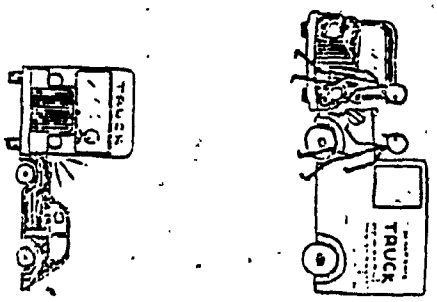




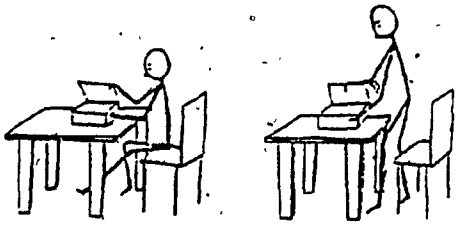


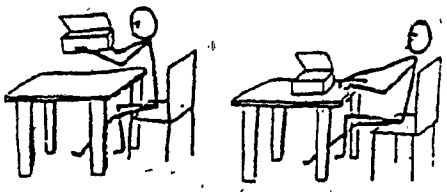
5.



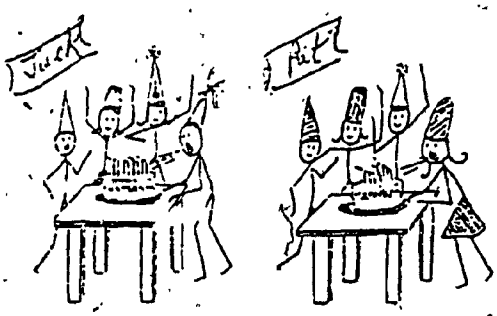


6.

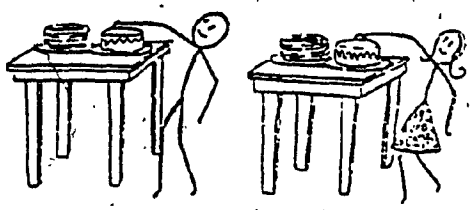


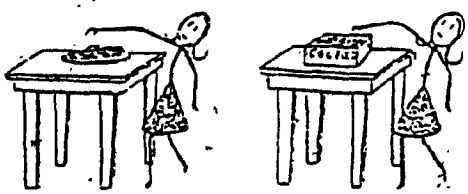


7.

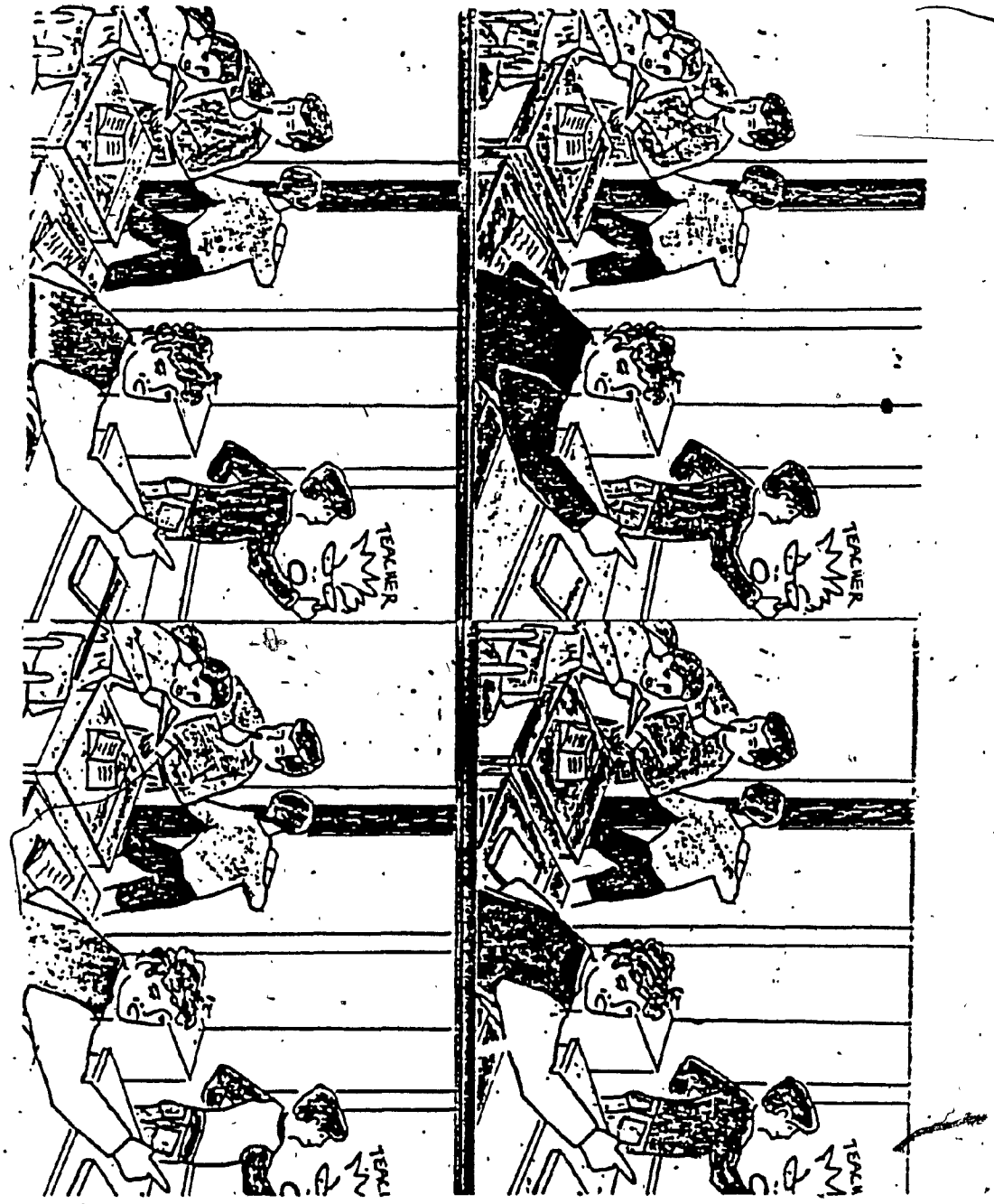


8.

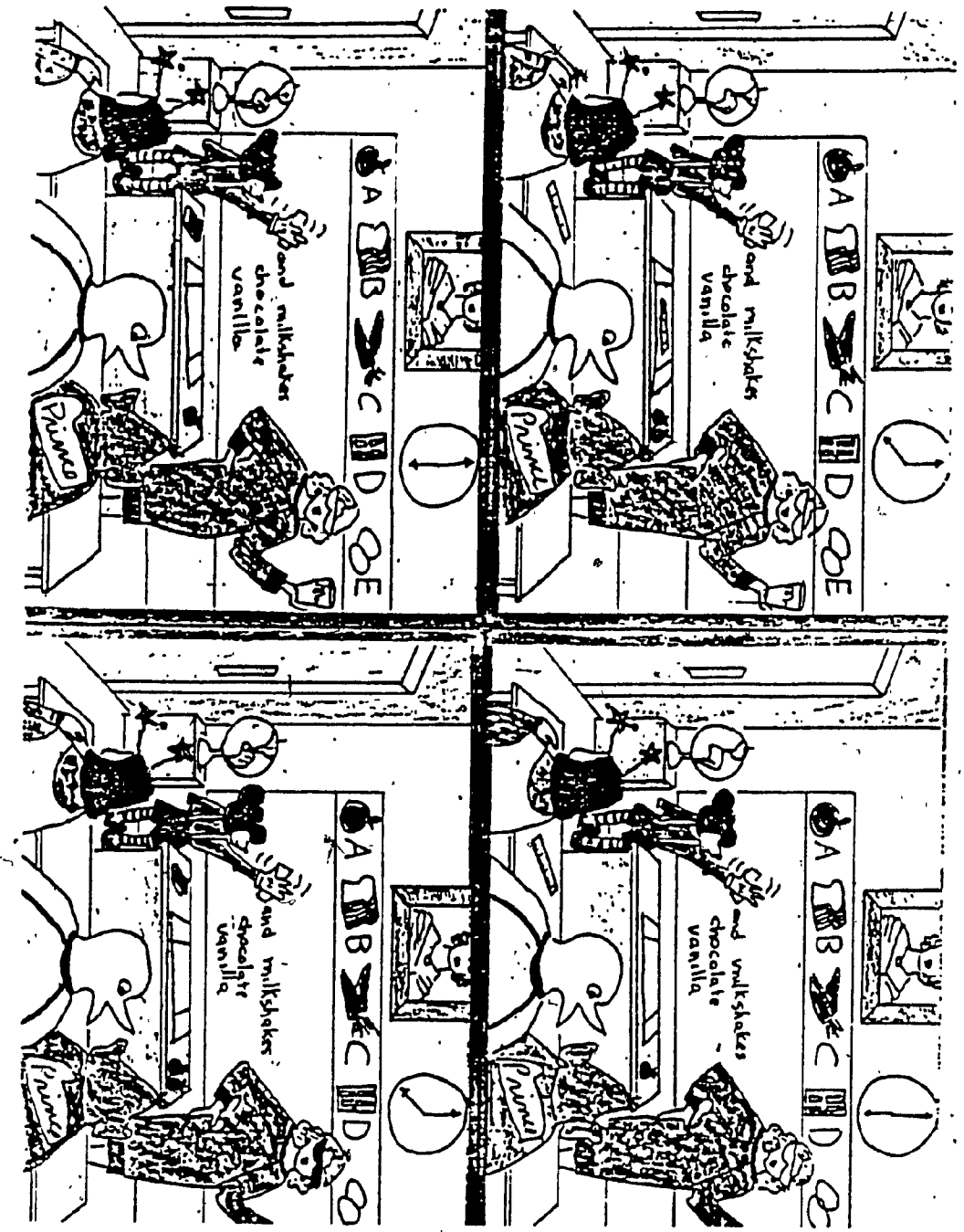




Two additional sets of PCG cards



Renderings for two additional sets by M. Hebert.


A
B
C
D
E
and milkshakes
chocolate
vanilla
Prince

Appendix B

Transcripts of speech segments for Group Y and Group O speakers.

Group Y speakers

Model Sample (low)

Is uh ... old student with uh ... There's one uh ... stand up and uh ... draw the ... face of the teacher ... (30 seconds) ... and uh ... another uh ... has uh ... a airplane uh ... made uh ... with uh ... paper ... (60 seconds) ... and uh ... there's ... five student ... mmm ... One look uh ... at the door if uh ... I think the teacher come ... uh ... The ... There's four student stand up and one uh ... he's ... he's sit down ... uh ... There's uh ... on this picture the ... the book is open and uh ... this one is uh closed ... (120 seconds)

Model Sample (High)

It's uh ... another teacher with her club, her class ... uh ... There's a globe ... a board, and on the board, there ... it's uh ... write "and milkshake, chocolate, vanilla" ... and uh ... there's a little boy ... he uh ... erasing the board (30 seconds) ... Over the board, there's a clock and it's six o'clock ... and three students are sit- ... sitting on their chair and one student has a ... a sweater of Prince ... and the teacher uh ... is wearing ... a purple dress, and she ... have earrings and ... glasses ...(60 seconds)... She hold a glass ... and over the board uh ... beside the clock ... There's a picture ... uh ...

One student is uh ... wearing in her hair ... She has long hair and in her hair, she has uh ... little springs with uh ... stars on it ... and there's a ruler on the table ... and one student is like a duck ... There's a apple ... an apple on the ... the, the desk, oh yeah ... and there's a ... and the left wall is yellow ... with a door (120 seconds).

Speaker a:

It's uh ... a teacher ... She uh ... teach ... uh ... to the student uh ... the good thing for eat ... She hold a glass on uh ... sh ... uh ... its hand ... In the class, there are four student ... (30 seconds) ... One is a penguin and another is uh ... a elephant, and the name of the elephant is Prince ... A girl ... she wearing a blue ... uh ... not sweater but uh ... jumper, and uh ... yellow and I don't know what ... brown uh blouse ... (60 seconds) and uh ... another ... the last student is uh ... uh ... I don't know what it is, but uh ... is a person ... uh ... She write uh ... she erase the board and uh ... she wearing, he wearing, or she... a ... green uh ... short with uh ... yellow uh sweater ... non ... uh ... red sweater ... uh ... Its uh ... its hair is uh ... black, and uh ... black ... uh ... It's six uh ... it uh ... six o'clock ... (120 seconds).

***

In this class, they are uh ... three, three boys and two

girl ... A boy's writing the board ... and uh ... she uh he ... draw a monster ... and uh ... the ... this monster ... she ... he write uh "Teacher" ... (30 seconds) uh ... This boy, sh-, he wearing uh ... blue, a red sweater and uh ... uh blue pants and he hair is uh ... short hair ... short black hair ... uh ... The girl ... one girl ... she wearing a skirt, blue skirt and a ... blue uh ... sweater (60 seconds).

Speaker b:

There a board ... uh ... Is say ... on the board, there are uh ... up of the board ... there are a clock ... and uh ... a little decoration ... It's a person on it, it's a ... a cable (un cable) ... uh ... uh ... on the top of the, on the bulletin board, I think ... there apple ... (30 seconds) ... apple... an apple, and uh ... behind uh ... in the right of the apple, there are ... A ... After there are bread, a B, carrot, C, door, D, eggs, E ... uh ... It's about uh ... seven, seven o'clock ... (60 seconds) ...uh ... There's a ... I don't know if it's a girl or a boy ... He is uh ... try to uh erase the board ... It have uh ... black shoes ... black hair ... red uh ... sweater and uh ... black uh short with uh ... uh ... uh ... suspenders ... and a globe uh ... just uh ... at the left ... of him ... on a little table, I think ... There are a little girl ... uh ... There are a table ... and there's three person ... One person uh ... is uh ... is saying "Prince" on it and

the other ... It's a duck, I think ... The other, it's a girl ... and uh ... in front of uh ... behind the desk of the teacher, there ... the teacher ... she take it ... a glass in her hand ... (120 seconds).

***

Uh ... There are two girls, I think ... uh three boys, three boys ... One boys is on the board, and he ... he uh ... he write on the boar-, board "Teacher" and he do a teacher with a glass and a big mouth ... and hair up ... and another boys is sit on his chair (30 seconds) ... and is uh point uh the board ... and uh ... and a girls, I think, she look of the door for uh ... look if the teacher coming ... and uh ... she have brown hair and a purple sweater or shirt ... and uh... uh ... short ... it's short, but it's go uh ... uh ... brown knickers (60 seconds).

Speaker c:

Uh ... I ... Is in the classroom again ... and uh ... the teacher is uh ... is not there ... The children is uh ... is do a, a ... (des folies) and uh ... he has a boy uh ... boys ... uh ... uh ... He write in the blackboard uh "Teacher" (30 seconds) ... and he uh ... he do a ... monster and uh ... he has a ... girl with a airplane ... and uh ... a airplane in paper ... and uh ... he has a boy ... uh check uh ... uh ... to uh... when uh the teacher uh ... arrive (60 seconds) ... and uh ... and there's ... uh ... other boy uh ... He do "Shhhht" ... or something ...

"Don't do uh ... " ... In the desk of the children, has a book ... and uh ... on the ... teacher's decks too ... The little girl uh ..., with a airplane in her hand ... he has a dress, blue ... blue, blue dress and uh ... the boy uh ... with uh .. an uh ... He's write uh ... in the blackboard uh ... He's wearing uh ... a red uh ... sweater, blue pants and uh ... (un veston) ... (120 seconds).

***

Uh ... Is in the classroom ... Is uh ... uh ... He has a children uh ... erase uh ... the ... erase the tableau ... (30 seconds) ... erase uh ... uh ... in the blackboard, he has ... is write uh ... "and milkshake, chocolate, vanilla, vanille" ... and uh ... he has uh ... is uh the Hallowe'en, I think ... He has uh ... costume ... He has one is uh ... Her costume is a duck ... (60 seconds).

Speaker d:

Uh ... Is in the classroom ... uh ... There is uh one clock, and it's uh ... it's six, it's six o'clock ... The ... There uh ... It has uh ... There is a board with uh ... some word on it ... and it's write uh ... "and milkshake, chocolate, vanilla" ... (30 seconds) ... uh ... and uh ... a little girl with uh ... green short and red sweater erase the words ... and next to the little girl, there's a ... globe ... on a ... a little table uh ... or a box (60 seconds), and uh ... next uh ... on the centre, there is, there is a bureau with an, an apple a book ... and next to

the bureau, there, there is the teacher with uh ... with uh ... a purple dress ... and uh ... uh ... she uh ... she have uh ... a glass on her head, h-, h-, hand ... and uh ... facing the teacher, there is one person with a green head ... and a ... purple sweater. His name is Prince ... Next to Prince, it has uh ... another student. It's like a duck ... and next to the uh ... duck, it, it has a girl with uh ... uh, black hair and uh ... some antennas on her head (120 seconds),

***

There's a ... a person write on the board ... draw on the board the, the face of the teacher ... and he write uh ... uh "Teacher" ... After, there's a girl look uh ... open the door and look if the teacher come ... and uh ... she have a ... a purple sweater ... brown pants (30 seconds) ... and next to the girl, there is, there is a boy with uh ... blue pants and a ... green sweater with uh ... and uh ... brown hair ... and uh ... next to the litt-, the, the boy, there's a girl with uh ... blond hair and a ... blue dress ...(60 seconds).

Speaker e:

The ... in the picture, there is a ... a mouse and ... a mouse uh ... in front of the board ... uh ... It's wear in uh ... he has ... uh ... a red sweater and a black shirt ... (30 seconds) ... uh ... In the picture, there is uh ... five people, one teacher uh ... dressed in uh ... a, a

black-purple uh ... dress ... There is a ... a ... a girl with uh ... stars ... star on her head ... (60 seconds) ... and she has a little watch ... She's wear a ... a blue jumper with uh ... an uh ... a yellow blouse with ... a long-sleeved yellow blouse ... On the board, on the board, it's write ... "and milkshake, chocolate, vanilla ... vanille, vanilla" ... uh ... It ... on the clock, it's uh ... six o'clock, and uh ... the teacher ... she has a ... glass ... and uh ... she has uh ... like a pot in her hand ... Beside the little mi-, mouse, it has a globe... (120 seconds).

***

Uh ... On the picture, there is uh ... five student ... One student, it's wear a ... blue, blue jeans and a red sweater ... and it, he write uh ... "Teacher" and with a ... a big face ... uh ... There is a little girl with uh ... yellow uh ... hair (30 seconds) ... a blue sweater and a blue skirt ... She has a little airplane in uh ... paper ... made in paper ... uh ... A boy look uh ... behind the door ... uh ... Another boy ... he has uh ... red hair (60 seconds).

Speaker f:

Is again at, in the school ... There is ... five children ... One uh ... is a ... I think is a girl because uh ... she have uh (tresses) ... and she look uh ... she open the door and she look like that ... (30 seconds) ... and uh ...

her uh ... her uh ... sweatshirt is purple with uh ... brown pants ... and the other uh ... the other uh ... boy uh ... writing on the,the, the board ... (60 seconds) ... and he ... and he writing ... he, he draw the, the teacher ... and she's not very beautiful ... and uh ... his sweatshirt is red ... and uh ... his hair is ... black ... black-green ... and uh ... his pants uh ... are blue, and he have two pocket in, in the back ... and another ... he have uh ... short hair with uh ... uh, his hair is brown, and he do like this with his finger ... and uh ... the color of his uh ... sweatshirt is green and his pants is uh ... blue (120 seconds).

***

I think is Mickey Mouse writing on the, the board ... uh ... He have a ... a, a green ... a red uh sweatshirt ... and uh ... a ... short ... a black short with white and black uh ... uh ... socks ... (30 seconds) and the shoes is black ... and next to it, is uh ... the ... the bureau of the teacher ... There is on it a book, black book and uh ... apples and uh ... like uh ... a book open or uh ... I don't know ... (60 seconds).

Speaker g:

It's a ... another class ... The ... the person, it's uh ... (.....) and uh ... and after uh ... students draw on the board a ... a teacher uh ... very uh ... not very good ... and the ... other students play ... (30 seconds)

... One look on ... at the door ... Another make a ... another girl make a, a, a airplane with paper ... and the students uh ... don't uh ... uh ... we ... he don't uh ... listening, it's not a, a good class, I think ... when the teacher is not there ... (60 seconds) ...I think ... uh ... the teacher uh ... will come in the class, and he ... he don't want to uh ... don't want to uh ... know the teacher what she, we do and uh ... they have ... they are three boy and two girls ... and the desk at the teacher ... he have a English book ... There are three decks ... three decks, desk ... and uh one chair ... and there are two book open ... mmm ... uh ... The, the boy uh ... draw the teacher is wearing uh ... uh ... jeans and a, a red sweater ... (120 seconds).

***

There uh ... there at a, a class ... and the teacher uh ... take a ... has a glass on her hand ... She has glass ...-es ... There are four student ...One, I think ... he has uh ... he has a headage ... (30 seconds) ... Another, I think it's a, a bird and another, it's a ... a, a frog ... and there are one student who erases the board, I think it's a mouse ... There ... there are apple, pear, carrots, door egg, and a clock ... (60 seconds).

Speaker h:

Uh ... There are uh ... five uh people ... uh ... They uh ... are in a class ... One boy uh ... uh, write uh writ-,

write, writing uh ... writing uh ... it's, I think uh "Teacher", and the face ... (30 seconds) ... uh funny with uh ... glasses... and uh ... There are uh ... two girls ... and uh ... three boy, boy ... We see a little bit ... (ben) we see a ... a ... three decks, decks ... and uh ... one chair ... A girl ... the ... open ... opening a, a door (60 seconds) ... On the left of the girl who uh ... opening the door ... uh has a finger on her, his mouth ... and next there are uh ... next there are a girl ... the s- ... uh, uh ... beside a desk ... and uh ... she holding a ... uh ... an avion in the hand, and uh, and on the right of the girl ... there h-, h- ...is uh ... a boy uh ... who uh ... showing uh ... the board ...and the other uh ... boy uh ... writing on the board (120 seconds).

***

Uh ... There is uh ... five uh ... people ... uh, I think is the, the day of uh ... Hallowe'en because uh ... they all uh ... with uh ... they wearing uh ... costumes (là) ... (30 seconds) uh ... and one boy ... uh ... Mickey (là) ... he, or she ...it's erase on the board and uh ... a woman uh ... on the front has a glasses ... and uh ... she uh ... hold a ... a glass, I think ... (60 seconds).

Speaker i:

A boys with uh uh ... green uh ... pant, pants and a ... a red uh ... sweater with uh ... a glove ... a white glove uh ... in his hand ... is erase on the board ... erase the

board and uh ... the ... the teacher have a ... a purple uh dress with uh ... glass ... (30 seconds) and uh ... she uh ... she hold the glass, a glass in her hand and uh ... On the decks ... the decks ... non ... desk ... there uh ... a apple, a book uh ... It's uh ... six o'clock (60 seconds) and uh ... There's a children with uh ... a green head (laughter), a ... purple uh ... purple sweater ... and uh ... it's write on the, the back uh ... "Prince" ... uh ... It's another children ... is uh ... uh ... a head, the head is a, a duck and uh ... (laughter) the other, she have a ... a yellow sweater with little flower ... uh ... She, she have uh ... a ... blue, another sweater with uh ... (ben) ... blue and uh ... long hair, black ... and uh ... she have a ... star in uh ... her hair ... (120 seconds).

***

There's uh ... five children ... uh ... One children is uh ... a girl and she look out uh ... the door ... uh ... She's wearing a brown uh pants and a purple uh ... sweater ... Another boy uh ... is uh ... there ... He has a, a green sweater with uh ... blue pants ...(30 seconds) ... uh ... Oh ... There's two decks ... uh ... (ben) two little decks and a, a big, a big desk ... uh ... Has a, a girl with blond hair and a skirt and sweater ... blue ... (60 seconds).

Speaker j:

Is ... uh ... in the, in the school ... They have ... the

... they have one boy (ben) M- M- Martian ... His name is Prince ... is uh ... They have a, a girl ... They have a star ... here ... Is in the ... is uh six o'clock ... (30 seconds) ... They have ... on the board is uh ... write "and milkshake, chocolate, vanilla" ... They have apple on the desk ... They have ruler on the student desk ... They have a globe ... uh ... They have a picture of a woman on the top of the picture ... uh ... The ... the woman, the teacher ... have a purple dress (60 seconds) ...with, I don't know what she have it in her hand, but she have, I think it's the glass ... The girl's look the door ... They have one girl look the door ... She has a ... a yellow sweater, blue jumper ... and brown hair ,,, At her right , they have a ... a bird, like bird ... I don't ... I ... and uh ... The, the wall of the class in ... on my left is yellow ... The ... and ... on the board, they have apple, A, bread, B, (ben) bread, B ... a carrot and C ...D ... a door and it's write D ... uh ... eggs and is write E ... and the teacher have uh ... glasses (120 seconds).

***

They have ... they have one boy draw the teacher, but uh ... not in the ... He have ... When he draw the, the teacher, he have the hair like that (laughter) ... and uh ... They have one boy's ... look at the door when the teacher not come (là) ... (30 seconds) ... They have one boy said, uh ... I don't know but, I think said uh ... "Look at the ... in ... in the board" ... They have one boy

said uh ... "Shut up", I think that Shhhhh ... okay ... They have one girls, I don't know what she make ... I think she make the jet, or she read something ... She have ... They have one ... The boy write on the board ... She, he have uh ... uh ... a red ... sweater ... (60 seconds).

Group O speakers

Model Sample (low)

I am from Zaire ... uh ... in middle of uh .., uh ... Africa ... It's a uh ... a warm country ... uh ... We have ... uh ... two season ... uh ... a rain- ... uh ... a rainy season and uh ... what you say ... wet ... yeah, we-, wet uh ... sea-, season ... uh ... Our population is uh ... uh ... thirty millions people (30 seconds), ... yeah, and uh ... uh ... the area is two, three, four, five uh ... miles ... non ... thousand ... uh ... no ... I think mil- ... kilomètres uh (laughter) square ... per square ... yeah ...(60 seconds) ... Ah, what kind of industry ... You know uh ... uh ... we didn't have ... uh ... more industry, but ... uh ... when we ... we was uh ... uh ... colonisated by uh ... by the Belge ... uh the Belgium, you know and, and they ... they developed uh ... the industry of uh ... what you say, mines? ... yeah, m- .. mines and uh ... in the, the east of uh ... m- ... my uh ... my country, uh ... uh ... there are many industries of uh ... uh ... mines ... uh ... We produce ... uh we ... we are the first uh ... productor of uh ... cobalt ... yeah, cobalt ... uh ... uh ...industriament ... (120 seconds).

Model Sample (High)

Well, in Canada ... there's a fed- ... okay ... federal government ... which is uh ... responsible to get uh ... which is responsible to collect all the taxes, and all the expenditures for the public. In Quebec, the people are

affected by Quebec government and ... as well as Canada, so Quebec government ... in other words, Quebec government collects its own taxes ... from Quebeckers ... (30 seconds) ... but uh ... approximately fifty percent of the rules and regulations are same, like ... something, s- ... some, some rules uh for example ... apply to Quebeckers which also apply to Ontarions. (60 seconds) ... Now, it's uh ... somewhat difficult, not difficult but somewhat uh ... more expense ... in the sense of taxation, for the people who make a lot of money, in Quebec. Their ... I think, combined marginal, marginal tax rate, for Quebeckers is close to fifty-five percent ... so ... more you work, more you pay taxes ... Conversely, it is much better for the people who make less money. For example, ten thousand dollars if you have two children ... uh ... you get child tax credit ... then uh ... from Quebec government as well as from ... federal. You get uh ... real estate taxes and ... which is not uh ... given other provinces, but they give indirectly by ... another form ... (120 seconds).

Speaker A:

Oh boy, this is very uh ... it's not necessary difficult , but it's uh ... a personal, a personal uh ... way of working ... I can do sss ... I can do some editing ... uh ... and they can be different than some other people can do ... this editing ... (30 seconds) ... and uh ... it's a personal choice ... and uh ... yeah ... it's sometime, it's

difficult to choose the exact scene uh ... of what you want really ... because there is some problem ... uh ... sometime the voice is less uh ... oh less (.....) less, yeah .. less loud than uh ... you want ... or uh ... something like that ... (60 seconds) ... The music is too loud, or uh ... you know, and the music doesn't fit with uh ... the next scene ... of what you want ... you know, thirty second is very short ... (laughter) ... and uh ... okay ... The film is uh ... it's film uh ... it's like picture ... okay ... you have grain ... grain of the picture, you have uh ... thousand of thousand of thousand of grain in the picture and video is line ... you can have uh ... five hundred lines per inch ... for video ... so the quality's is very different than uh ... film... That's the reason why uh ... your video is look like cheaper than film, because it's line ... but uh ... in probably ... ten years, oh no, less than that ... two, three, or four years, I don't know uh ... you will have uh ... a video with uh ... with twenty thousand lines per inch ... (120 seconds)

***

There was ... there will always have somebody to do film, you know ... like there is people who work ... would ... like uh ... two hundred years ago, so ... but uh ... by now, the video is very popular than film because uh ... you can do a lot of effects in video uh ... would take uh just uh ... one hour or two ... (30 seconds) ... and film, it will take uh ... probably four days ... because you have to

process the film, to develop the negative and to do a copy ... and after that, to edit the copy ... and to return to the lab to do another copy, a final copy ... It's very long ... In video, you have the result uh ... right now, if I can say uh ... right now ... It just take one hour, two hour, to depend uh ... of the, of the shot (60 seconds).

Speaker B:

It's because what I, what I study in course sometime ... I, I think what I did in such ... in like teacher, and it's help me, the, the behaviour of certain k- ... uh ... some kid in class and the answer or the s- ... one after one years, I have, I met a little girl, five years old I think, and the morning, she was really, really sad, always the tear in eyes and I didn't know what she had ... (30 seconds) ... Before she, she left the house, the morning, the ... their parents said to her that they was divorced ... and she was really sad for the day ... I really don't understand when, when the parent ... left her to, to go to school ... She didn't listen what we ... what we said and uh ... at the first recreation, she, she left the school and ... run to the house ... (60 seconds) ... and, I'd like to work with that ... If you don't help this kid right now, I think always, always they will carry, carry that with them ... so ... we have to ... to do something ... pretty soon ... The ... What's happened with that example ... the, the sociable or something, like in, in class, the teacher

sometime can know uh ... something wrong, going wrong with this kid, but he doesn't have the time to take care ... He know ... he, he know he have to do something, but what? ... Maybe five minutes a day, and this kid need maybe three hours a day ... It's for that, and I think it's like sociable in Quebec ... The kid can be, can be in the group sociable, but ... something, sometime that is help ... that is help him, but sometime he need more, more attention, more uh ... exception just for him, just uh ... just to, to improve ... (120 seconds)

***

Because I don't believe at the ... straight therapy with children, sometime, I, I think if I am a children ... maybe five or six years old ... and I have to meet psychologist, and when I meet the psychologist in ... is in its ... uh ...straight room with the desk and two chair ... If I am a children,I won't, I won't want to go there ... (30 seconds) ... so, for me ... I think it's better to ... the children have to be welcome ... and the (.....) ... the play with ... every, every toys ... because it's, it's better for him. It's my ... it's my idea, and also, if you want to ... to talk with the children, it's not ... is if you want to have better result, you have to, to play with him ... Is not to ... in asking "How are you? What do you think of your mother?" ... (60 seconds).

Speaker C:

She's good, but she treating us like kids and, you know, when we pay for the course and we are student uh ... university student ... and we have the right to do what we want with the course ... I, I think so ... and I'm not the only one ... The majority of, of the people in the class think that ... and uh ... when we uh ... we are late, you know, she uh ... look at us with uh big eyes ... (30 seconds) ... and you know, she's very repressive ... with us ... when we are late, and when we uh ... miss classes ... and all this ... and somebody uh ... this morning ... tell her uh ... okay ... she uh ... begin to talk about the, the ... you know, the ... people who were late and she ... she's just not, you know ... she don't like it ... (60 seconds) ... and she ... she ... she was telling us that she don't like it ... and somebody tell her ... but you know uh ... somebody tell her uh ... that we have the right to, to ... you know, arrive when we want to ... and she don't ... she didn't want to talk about ... she uh .. became uh ... uh ... very uh ... nervous and uh ... didn't want to talk about it ... and you know, tell in front of the class that she don't want to uh ... you know uh ... uh more than that ... uh ... that, I don't know how to say it but ... she tell the, the ... Alise that uh ... she didn't want to insult her, that's it ... insult her in front of the class, but she did it all the time ... (120 seconds).

***

I don't know, I wish I could do something, because I don't want her to, you know uh ... to be out of the, the university ... nothing like that but, ... I think it's important that she know and she don't take critics ... She don't want to take critics ... I don't ... I don't know ... maybe the department, the English department, but ... I don't want to make a, a big deal of it ... (30 seconds) ... I just want to tell her, but if I tell her, I, I'm afraid she will, you know uh ... I don't know ... not be fair after that ... and I think she did ... She's not fair when, when people are ... got late or ... get late or ... miss a class ... She uh ... like I said before, she, she find a way to remind us ... indirectly ... (60 seconds).

Speaker D

Okay, I like to work in the restaurant because I like to work with the public and uh ... I like to serve the people, because ... I like to give my service, my personal service ... okay ... and I think that everybody in the life have to go in the restaurant for to eat something ... and uh ... I meet a lot of people ... and I like to work for, just for the weekend ... okay ... because I go to the school ... and uh ... because I meet a lot of people ... and uh ... (30 seconds) ... I'm very enjoy when I'm work, when I work like waiter ... and I would like, after my university, I would like to be uh ... management hotel ... okay ... because I think that it's very important for uh ... I would like to

do uh ... another kind of restaurant ... something special with uh ... it's like a ... a little hostel? ... okay ... somewhere in the ... when the environment is beautiful ... okay ... (60 seconds) ... and I would like to have a little hotel with uh ... maybe ten, ten rooms ... and a little dining room, and something special, you know ... with a little farm ... with uh ... chicken and something like that, you know what I mean, okay? ... but I can't be alone for to, for to do this restaurant ... because it's ... it's uh ... too big for me ... okay ... I would work twenty-four hours in a day ... and uh ... I don't know ... I would like something like uh ... when you go in, in England or in France, you know ... you can meet something like that ... It's calling uh ... farming holiday or something like that ... (120 seconds).

***

Okay ... because sometime it's very hard to do okay ... and you have to remember the order ... and sometime, you have a quarrel? with, non, a fight with someone ... and ten minutes after, you can be very happy because ... someone make, makes, makes you uh smile ... okay ... and uh ... it's good for your character ... okay (30 seconds) ... I work, I w- ... when I was uh ... fourteen years old, I began to work in the restaurant ... and I have twelve year ... twelve years experience ... and uh ... when I go to work, it's like a habits, that's all ... I don't go to work ... just uh ... I feel like uh ... when I go to do ski, or

something like that, but ... but sometime, it's not very funny, because when you meet someone uh ... how you say ... ugly? non, yes ... (60 seconds).

Speaker E:

And uh ... I'm practising track and field since 1978 ... what means six years ... I really like it ... I'm training about uh ... five to seven hours a day ... It's not too bad ... Could be better ... Well, it's improving every year ... We're having more administration problem than any other problem ... The faci-, facilities are good ... It's just .. they, like, try to ... they try to ... I don't know ... to don't let us do whatever we want ... (30 seconds) ... It's always problems to get whatever ... we would like to do or ... like we have to do gymnastic ... The city is there ... "Aw, you should not do gymnastic; it's not really good, you know. Gymnastic is for the gymnastic athletes" ... and ... but gymnastic is the most important part of the training, so we have to do gymnastic all the time ... and we're having problem with that ... We need new mats, and this is a thousand bu-, dollars ... and they don't want to buy it and ... yeah ... you need it, without training, you're not going very far ... It's fun. That's why I'm doing it ... (60 seconds) ... The trips, cause I'm always going all over the United States, Canada and Europe ... It's really fun ... and uh ... I want to make the Olympic ... That's, you know, after that, we'll see ... I don't know if I would

like to get a world record or whatever, but I'd like just to get in a certain point where I'll be happy and see what, what's there, what is different ... It's really fun, because you meet people all the time, and they ... You meet people all want to do the same thing than you ... want ... and they are human like you are, you know ... When you watch them on T.V., you think they are not human, but when you meet them, it's like ... I mean, like you, they train the same thing, they do the same thing, they drink the same thing, they dream about the same thing ... It's really fun ... It's ... It's really a lot like life too ... In life, the training is like going to school ... and after you achieve your school, you ... can go for your goal, what is having uh ... you know, doing your life, whatever it is ... It could be professional or anything else ... (120 seconds).

***

Pole-vault is a different event than all the other because it takes more guts, because you're jumping with ... taking off from sixteen feet, you know, where you hold the pole ... so you have to have guts ... It's a more ... you have to train a bit more because it's a, a job which is more near gymnastic ... and then ... at the same time, it's a jump, so you have to combine two sports almost, you know, gymnastic and at the same time, running and jumping ... (30 seconds) ... and uh ... also it takes, it's a different athlete ... It's somebody who's really, really realxed,

because in a competition, you spend ... A competition can last for six to eight hours, and you might have to do like ... uh five jumps in these six hours ... so you have to be really, mentally really ... well-prepared, because you have to be at your peak at every jump, because if you miss a bar, if you're in a big meet, if you miss a bar , that's mean you might miss the team for, I don't know, if it's an Olympic session or something ... (60 seconds).

Speaker F:

Amnesty International is uh ... an unpolitical groups ... and it's not supposed to work for the government ... It's just uh ... some members of uh .. of ... are working for uh ... the prisoner politics ... I mean the prisoner or who ... not use violence ... (30 seconds) ... who are in uh prison? what is prison? ... penitentiary? ... for uh ... for their believe ... like uh ... their sex or uh ... ethnic or uh ... color, and especially their believe uh politic ... and uh ... what we do, in Amnesty International, it's uh ... we write a lot of letter to uh ... the government of uh ... the country where from uh ... the prisoner ... (60 seconds) ... and uh ... we uh ... we try with our letters to uh ... to, to ... to libère ... to (laughter) ... okay ... to get out the prisoner ... of the prison ... penitentiary ... Then uh ... often we write a letter, and it's not useful, because the government didn't take care about it ... but we try ... and we are not

supposed to help a prisoner if he is in Canada, because I am in Canada ... We have to help the people, the prisoner out of our country ... After uh ... uh ... we make a lot of conference and we make a lot uh ... I means uh ... I can't explain it ... (120 seconds).

***

I work a lot last uh ... last month about a concert ... This is a pianist Argentin ... He went in prison uh ... three years, and he get out last year ... Then uh ... he is a pianist Argentin, and we make for him here a concert ... (ben) I means he make a concert for us ... for uh ... Amnesty International ... (30 seconds) ... yeah ... and it was a bit hard to organize it because uh ... it's something difficult and if you don't have the experience, like us, like me, I didn't do that before, and uh ... it was hard to uh ... to reserve the, the, the place and uh ... the tickets sold it ... sell it, and everything like that ... and we try also to contact some uh artist? ... for this concert ... (60 seconds).

Speaker G:

Uh ... even though the United States is the uh ... the ... the most uh ... is the world's richest nation, and the most uh ... sophisticated one uh ... about uh ... medical care ... uh ... about forty thousand babies uh ... perish uh ... every, every year ... (30 seconds) ... and uh ... nine out of ten uh ... die because they are either uh ... premature

babies or uh ... low weight babies ... which is about, which is under ... five pounds ... uh ... uh ... The uh ... In the United States, they don't have uh ... the medic-, the medical care as we have here ... The government doesn't pay for it ... (60 seconds) ... so the, the people have to pay for visiting the doctor ... which is uh ... not quite expensive, but uh ... for a lot of people, it is expensive ... and uh ... most of the people uh ... get an insurance as uh ... they work ... but uh ... if they lose their, their job, the insurance will be over after thirty days ... so, for uh ... a lots of people ... visiting a doctor is a lux- a luxury? ... so ... uh ... uh ...if most of the mothers uh ... had seen a doctor during ... well, at the beginning of uh ... her pregnancy, it could have uh ... detect uh ... signs of uh trouble and could have prevent ... prevented uh ... the premature labour (120 seconds).

***

Uh ... with the recession ... the uh ... the government uh ... cut down on uh ... many expenses ... like uh ... on social workers and everything like that ... so the, the people ... even though the people want to have access to uh ... those kind of social uh ... care ... uh ... there is a big selection ... (30 seconds) ... uh ... done uh ... being done on the people that uh ... can see or not see a doctor ... like there, there, there is uh ... a medical centre which is uh ... which they call it Medicaid ... which is uh ... for the poor people, but the very poor (60 seconds).

Speaker B:

Mmmm ... This year I'm a student, I'm independent students ... uh ... I was working as a secretary for uh ... eight years, and I quit my job in last June, so uh ... to come, to return to school was a big, big change in my life and I still feel a litlle bit uh ... I know what to say ... dizzy, you know, dizzy ...(laughter) ... (30 seconds) ... a new ... new way to live, you know ... and uh ... I hope I will ... feel better in some ... in few months ... I hope so .. uh ... uh ... Up to date, my life has been uh ... quite ... quiet life ... (60 seconds) ... and ... I was working ... good girl, nothing happened ... I didn't travel ... and uh ... I ... I was feeling bored, really bored by my life ... and uh ... I have many questions too about life ... Why do we live? Why we are here? (laughter) ... because when I see all the ... the ... the bad things in the world, all over the world, I just feel uh ... unhappy. When I heard about uh ... acid rain ... when I heard about uh ... nuclear armaments ... when I heard about uh ... fights and wars between countries ... when I heard about uh ... uh ... people murders like ... the one ... the big one we had last week in India ... (120 seconds).

***

I ... I'm just wondering what's happening now around us ... I ... I feel very sad about that ... I would like to ... to uh ... escape this world to go on ... on an ... a desert

... on a desert island and forget everything ... alone ... Every ... okay ... I'm sure I'm not alone to feel like that ... Many people feel that ... are very sorry for what is happening in this world ... (30 seconds) ... Is completely crazy, and there ... there is so vio- ... violence now ... You just look in ... on T.V. ... the videos ... rock videos ... They are so violent, you know, and all the ... the policies program on T.V. and ... always vi- violent ... (60 seconds).

Speaker I:

Okay ... uh ... right now, I'm looking for a, a summer job for next sum- ... next summer ... and uh ... I thought to try to find a job in the ... uh environmental uh ... study ... I went to the, to Parc Canada, and I ask uh ... ask to the guy who was there ... (30 seconds) ... uh ... what kind of job that I can get in a park ... so he told me then is uh ... there are three different jobs ... uh ... and the one who interest more me is uh ... the one is uh ... guide in a park uh ... but it's an exchange between the provinces (60 seconds) ... and you ... the ... the goal of this job is to ... to go in one other province and uh ... they exchange the guide, so like this, I will uh ... be able to uh ... learn English, and uh ... to see a ... a different country ... not country, but uh ... different are-, area and uh ... it's why this afternoon, I will have a ... I will have to meet a person from Parc Canada for a ... He

will give me all the uh ... the paper and the information for a fill ... the sheet for the job ... and uh ... I have also to give him my uh ... curriculum vitae ... and he explain me ... where I can work and what kind of job exactly what it will be ... (120 seconds).

***

I went there one time, and it, it was too fast for ... for then ... for appreciate all the nice uh ... nature and uh ... I would like to ... be able to st-, not study but live in this kind of uh ... environmental ... uh ... environment ... (30 seconds) ... for uh ... be able to ... re- rep- reproduce something quite the same here in Quebec in uh ... in future ... and uh ... for, you have to live in for ... really know what's, what's going on ... and ... like what are the real uh ... uh ... objectives and uh ... for feel, just read about something, you, you don't get the ... the right information ... (60 seconds).

Speaker J:

My job is not only to, to sell ... it's to uh ... to organize the, the store ... uh ... all stock we receive uh ... you know ... y- ... see , everything uh ... that we receive here in uh ... sports uh ... every kind of things ... (30 seconds) ... and uh ... it give me uh ... more knowledge about every, every kind of sport .... is what I like from that, from that job ... uh ... before, I, I know only about uh ... major sport, like baseball and uh ...

hockey, everything ... Now, uh ... is more uh ... more about camping and uh ... yeah, yeah ... uh ... We import many tents from Europe ... (60 seconds) ... and we are the uh ... the only store here in Quebec, in Quebec to do it, so uh ... everybody who wants uh ... new tents, new models ... they come uh ... to us ... so uh ... we know about our stock ... see, it's easy to, to sell about ... and uh ... I, I like it ... Running shoes ... uh ... only few models ... because it's too complex ... uh ... if you uh ... don't know how uh ... if you have a ... running shoe from uh ... each company ... it's a definite (ben) ... you know uh ... about uh ... like uh ... a guy uh sport who have uh ... I don't know, a hundred kind of running shoes ... (120 seconds).

***

We buy a lot of products from the U.S.A. ... Our dollars is cheaper eh? ... uh ... I think it's the major reasons ... yeah, and uh ... every products uh ... if they don't ... if we don't import them, uh ... from the U.S.A., it's from Ontario ... so we have always a transport to pay ... (30 seconds) ... Here in Quebec, it's uh ... it's a bit expensive ... uh ... This summer, when I go to uh ... I uh ... gone, go, go ... went to uh Plattsburg ... yeah ... I compared the, the products ... uh ... In the U.S.A. okay, it's in uh ... U.S. cash ... it's only for uh ... only a running shoe ... we sell about thirty bucks ... it's uh fifteen bucks U.S. ... (60 seconds).

Appendix C

Order of presentation of speech samples for each condition of length and for both rating tasks.

Rating Task 1

30-second condition

Group Y speakers	Group O speakers
1. f	1. B
2. d	2. G
3. e	3. E
4. c	4. F
5. j	5. H
6. h	6. C
7. g	7. J
8. a	8. I
9. c	9. G
10. j	10. A
11. a	11. D
12. e	12. E
13. d	13. A
14. g	14. B
15. h	15. I
16. b	16. H
17. i	17. F
18. f	18. J
19. b	19. C
20. i	20. D

60-second condition

Group Y speakers	Group O speakers
1. b	1. J
2. f	2. A
3. a	3. I
4. b	4. E
5. e	5. B
6. i	6. F
7. g	7. C
8. c	8. H
9. d	9. A
10. e	10. B
11. h	11. E
12. d	12. I
13. j	13. G
14. a	14. D
15. i	15. F
16. j	16. C
17. f	17. H
18. h	18. J
19. c	19. D
20. g	20. G

120-second condition

Group Y speakers	Group O speakers
1. e	1. D
2. g	2. F
3. f	3. A
4. h	4. h
5. c	5. J
6. i	6. B
7. b	7. E
8. a	8. I
9. j	9. G
10. d	10. C

Rating Task 2

60-second condition

Speakers

1. h
2. D
3. i
4. F
5. C
6. g
7. E
8. a
9. C
10. I
11. b
12. G
13. j
14. A
15. f
16. J
17. B
18. d

Appendix D

Rating sheets for Rating Tasks 1 and 2.

TAPE

NUMBER ...

Side A

Example 1.

Example 2.

1.

2.

3.

4.

5.

6.

7.

8.

9.

10.

11.

12.

13.

14.

15.

16.

17.

18.

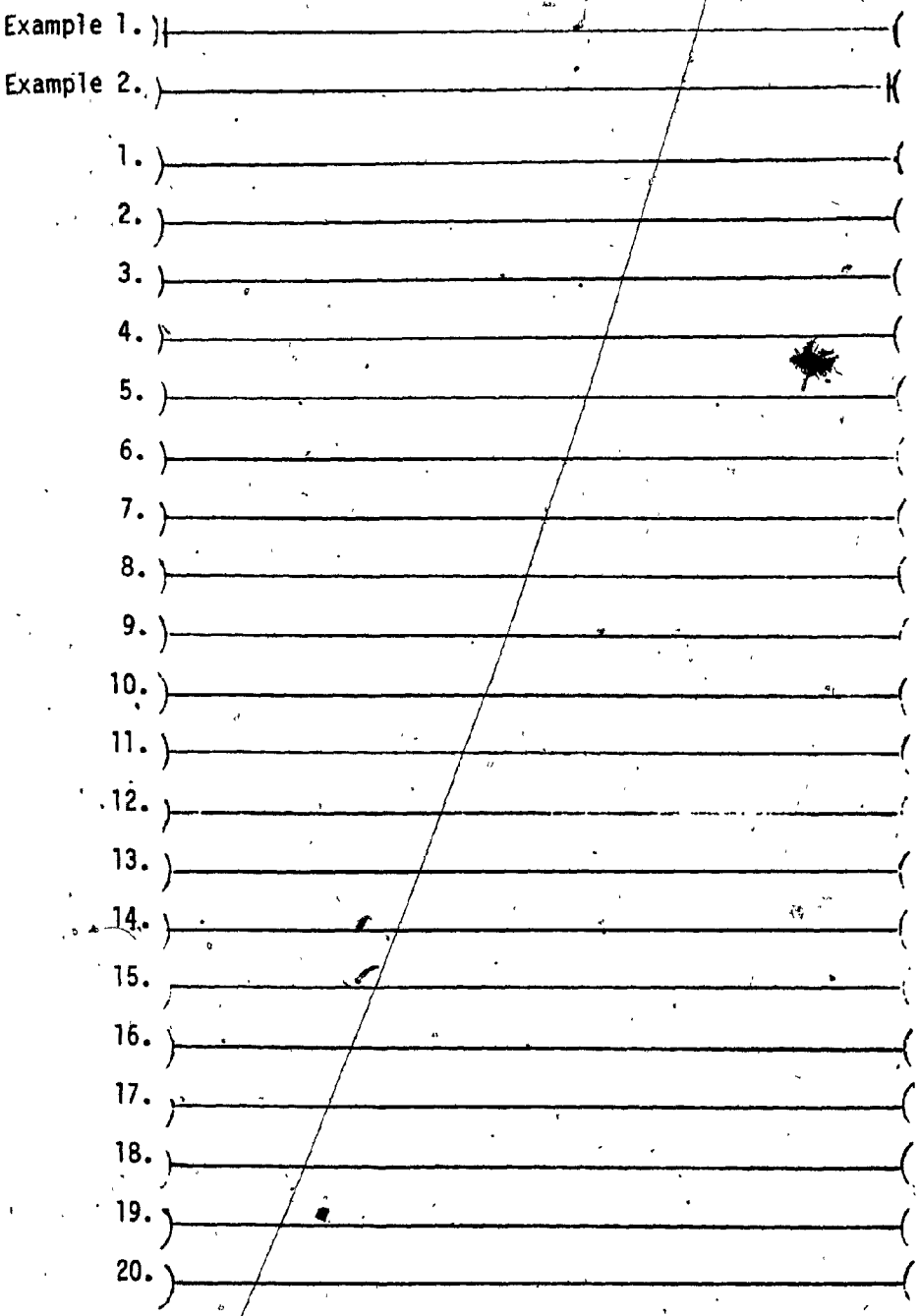
19.

20.

TAPE

NUMBER ...

Side B



Age: ..... Sex: M ... F ...

Languages spoken: English ... French ... Other .....

First language(s) acquired: .....

Present dominant language:.....

Current status: (please indicate where more than one applies)

ESL teacher: Full time ... Part time ...

TESL teacher: Full time ... Part time ...

TESL undergraduate student: Full time... Part time ...

TESL graduate student: Full time ... Part time ...

ESL-related experience:

Teaching ESL: ..... years Studying TESL: ..... years

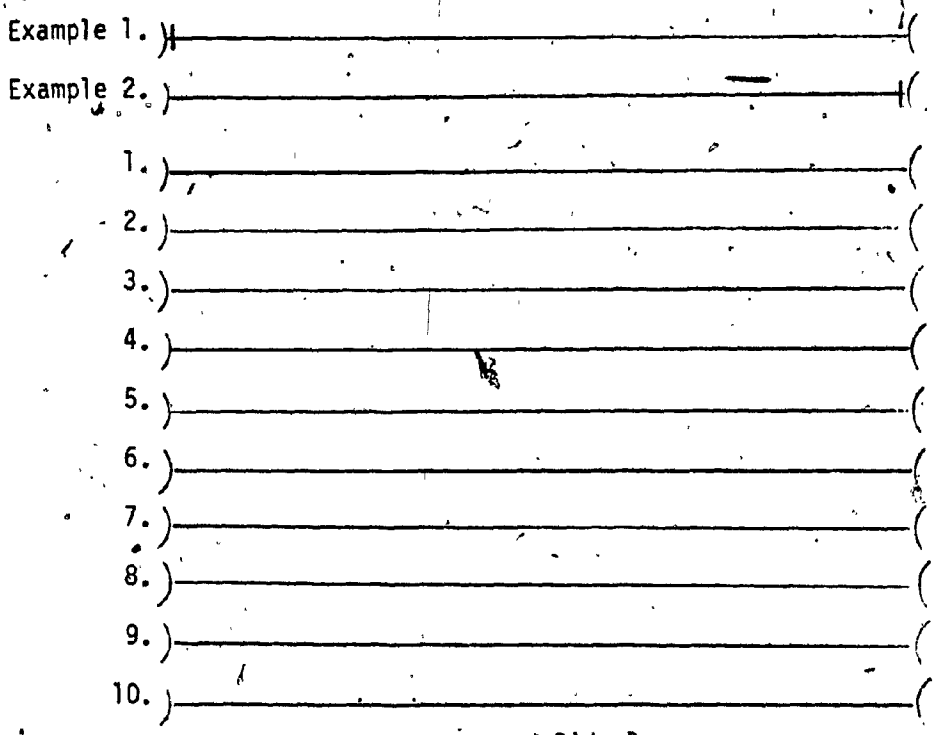
Level .....

Thank you very much for the time you have taken to participate in this project.

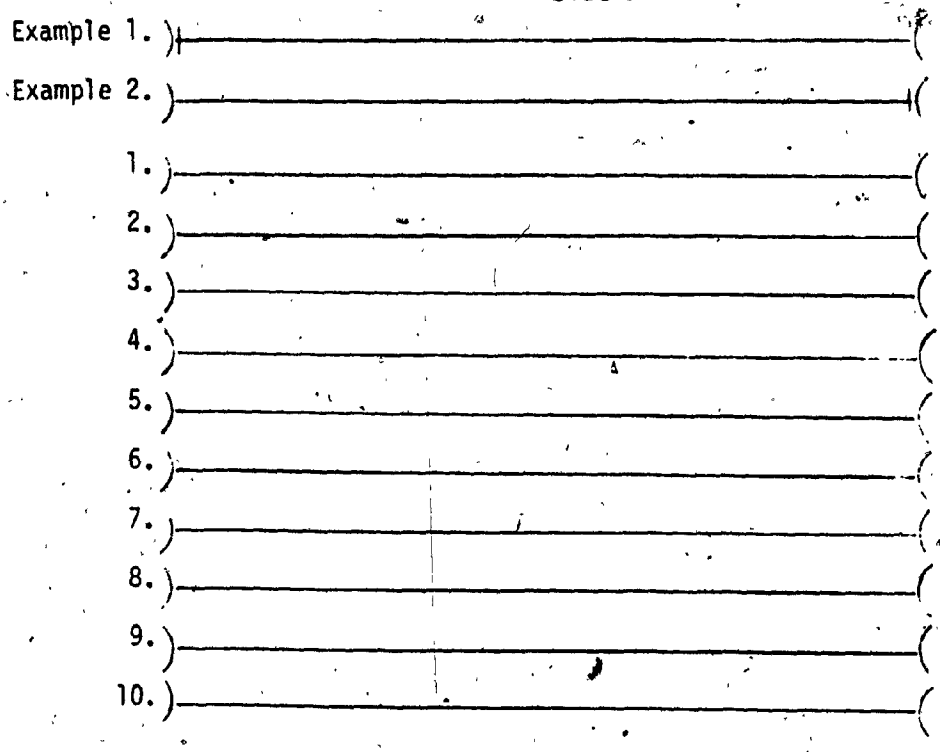
TAPE

NUMBER ...

Side A



Side B



Age: ..... Sex: M ... F ...

Languages spoken: English ... French ... Other .....

First language(s) acquired: .....

Present dominant language: .....

Current status: (please indicate where more than one applies)

ESL teacher: Full time ... Part time ...

TESL teacher: Full time ... Part time ...

TESL undergraduate student: Full time ... Part time ...

TESL graduate student: Full time ... Part time ...

ESL-related experience:

Teaching ESL: ..... years Studying TESL: ..... years

Level .....

Thank you very much for the time you have taken to participate in this project.

Example 1. 26 .....

Example 2. 114 .....

TAPE NUMBER ....

1. .....
2. .....
3. .....
4. .....
5. .....
6. .....
7. .....
8. .....
9. .....
10. .....
11. .....
12. .....
13. .....
14. .....
15. .....
16. .....
17. .....
18. .....

Age: ..... Sex: M ... F ...

Languages spoken: English ... French ... Other .....

First language(s) acquired: .....

Present dominant language: .....

Current status: (please indicate where more than one applies)

ESL teacher: Full time ... Part time ...

TESL teacher: Full time ... Part time ...

TESL undergraduate student: Full time ... Part time ...

TESL graduate student: Full time ... Part time ...

ESL-related experience:

Teaching ESL: ... years Studying TESL: ... years

Level .....

Thank you for the time you have taken to participate in this project.

Appendix E

ANOVA tables

2-way ANOVAs (rating by speaker, rater)

A. Where both speech samples per speaker are included in the ANOVA:

30-second Young

Source of variation	Sum of squares	DF	Mean Square
MAIN EFFECTS	300.175	13	23.090
SPEAKER	118.650	9	13.183
RATER	181.525	4	45.381
2-WAY INTERACTIONS	269.825	36	7.495
SPEAKER RATER	269.825	36	7.495
EXPLAINED	570.000	49	11.633
RESIDUAL	246.250	50	4.925
TOTAL	816.250	99	8.245

30-second Old

Source of variation	Sum of squares	DF	Mean Square
MAIN EFFECTS	293.592	13	22.584
SPEAKER	137.452	9	15.272
RATER	156.140	4	39.035
2-WAY INTERACTIONS	289.510	36	8.042
SPEAKER RATER	289.510	36	8.042
EXPLAINED	583.102	49	11.900
RESIDUAL	274.125	50	5.483
TOTAL	857.228	99	8.659

60-second Young

Source of variation	Sum of squares	DF	Mean Square
MAIN EFFECTS	538.654	14	38.475
SPEAKER	109.685	9	12.187
RATER	428.969	5	85.794
2-WAY INTERACTIONS	108.552	45	2.412
SPEAKER RATER	108.552	45	2.412
EXPLAINED	647.206	59	10.970
RESIDUAL	140.625	60	2.344
TOTAL	787.831	119	6.620

60-second Old

Source of variation	Sum of squares	DF	Mean Square
MAIN EFFECTS	721.625	14	51.545
SPEAKER	230.383	9	25.598
RATER	491.242	5	98.248
2-WAY INTERACTIONS	113.592	45	2.524
SPEAKER RATER	113.592	45	2.524
EXPLAINED	835.217	59	14.156
RESIDUAL	100.750	60	1.679
TOTAL	935.967	119	7.865

B. Where only sample 1 per speaker was included in the ANOVA:

30-second Young

Source of variation	Sum of squares	DF	Mean Square
MAIN EFFECTS	211.795	13	16.292
SPEAKER	127.397	9	14.155
RATER	82.519	4	20.630
2-WAY INTERACTIONS	249.714	36	6.937
SPEAKER RATER	249.714	36	6.937
EXPLAINED	461.510	49	9.419
RESIDUAL	.000	1	.000
TOTAL	461.510	50	9.230

30-second Old

Source of variation	Sum of squares	DF	Mean Square
MAIN EFFECTS	237.300	13	18.254
SPEAKER	89.360	9	9.929
RATER	142.451	4	35.613
2-WAY INTERACTIONS	211.524	36	5.876
SPEAKER RATER	211.524	36	5.876
EXPLAINED	448.824	49	9.160
RESIDUAL	.000	1	.000
TOTAL	448.824	50	8.976

60-second Young

Source of variation	Sum of squares	DF	Mean Square
MAIN EFFECTS	273.663	14	19.547
SPEAKER	78.146	9	8.683
RATER	195.592	5	39.118
2-WAY INTERACTIONS	85.140	45	1.892
SPEAKER RATER	85.140	45	1.892
EXPLAINED	358.803	59	6.081
RESIDUAL	.000	1	.000
TOTAL	358.803	60	5.980

60-second Old

Source of variation	Sum of squares	DF	Mean Square
MAIN EFFECTS	422.159	14	30.154
SPEAKER	155.230	9	17.248
RATER	266.788	5	53.358
2-WAY INTERACTIONS	101.390	45	2.253
SPEAKER RATER	101.390	45	2.253
EXPLAINED	523.549	59	8.874
RESIDUAL	.000	1	.000
TOTAL	523.549	60	8.726

C. Where only sample 2 per speaker was included in the ANOVA:

30-second Young

Source of variation	Sum of squares	DF	Mean Square
MAIN EFFECTS	186.220	13	14.325
SPEAKER	63.325	9	7.036
RATER	122.381	4	30.595
2-WAY INTERACTIONS	195.427	36	5.429
SPEAKER RATER	195.427	36	5.429
EXPLAINED	381.647	49	7.789
RESIDUAL	.000	1	.000
TOTAL	381.647	50	7.633

30-second Old

Source of variation	Sum of squares	DF	Mean Square
MAIN EFFECTS	202.765	13	15.597
SPEAKER	121.897	9	13.544
RATER	72.935	4	18.234
2-WAY INTERACTIONS	239.774	36	6.660
SPEAKER RATER	239.774	36	6.660
EXPLAINED	442.539	49	9.031
RESIDUAL	.000	1	.000
TOTAL	442.539	50	8.851

60-second Young

Source of variation	Sum of squares	DF	Mean Square
MAIN EFFECTS	335.933	14	23.995
SPEAKER	68.792	9	7.644
RATER	264.461	5	52.892
2-WAY INTERACTIONS	115.878	45	2.575
SPEAKER RATER	115.878	45	2.575
EXPLAINED	451.811	59	7.658
RESIDUAL	.000	1	.000
TOTAL	451.811	60	7.530

60-second Old

Source of variation	Sum of squares	DF	Mean Square
MAIN EFFECTS	334.232	14	23.874
SPEAKER	94.504	9	10.500
RATER	239.619	5	47.924
2-WAY INTERACTIONS	79.571	45	1.768
SPEAKER RATER	79.571	45	1.768
EXPLAINED	413.803	59	7.014
RESIDUAL	.000	1	.000
TOTAL	413.803	60	6.897

D. Where only one sample was available per speaker:

120-second Young

Source of variation	Sum of squares	DF	Mean Square
MAIN EFFECTS	421.942	14	30.139
SPEAKER	265.816	9	29.535
RATER	163.386	5	32.677
2-WAY INTERACTIONS	232.804	45	5.173
SPEAKER RATER	232.804	45	5.173
EXPLAINED	654.746	59	11.097
RESIDUAL	.000	1	.000
TOTAL	654.746	60	10.912

120-second Old

Source of variation	Sum of squares	DF	Mean Square
MAIN EFFECTS	296.308	13	22.793
SPEAKER	256.446	9	28.494
RATER	43.569	4	10.892
2-WAY INTERACTIONS	221.731	36	6.159
SPEAKER RATER	221.731	36	6.159
EXPLAINED	518.039	49	10.572
RESIDUAL	.000	1	.000
TOTAL	518.039	50	10.361

3-way ANOVAs (rating by speaker, rater, rating occasion)

30-second Young

Source of variation	Sum of squares	DF	Mean Square
MAIN EFFECTS	307.894	14	21.992
SPEAKER	121.055	9	13.451
RATING OCCASION	.945	1	.945
RATER	185.256	4	46.314
2-WAY INTERACTIONS	358.195	49	7.310
SPEAKER RATING OCCASION	66.049	9	7.339
SPEAKER RATER	273.096	36	7.586
RATING OCCASION RATER	16.213	4	4.053
3-WAY INTERACTIONS	162.639	36	4.518
SPEAKER RATING OCCASION RATER	162.639	36	4.518
EXPLAINED	828.728	99	8.371
RESIDUAL	.000	1	.000
TOTAL	828.728	100	8.287

30-second Old

Source of variation	Sum of squares	DF	Mean Square
MAIN EFFECTS	310.509	14	22.179
SPEAKER	142.742	9	15.860
RATING OCCASION	1.398	1	1.398
RATER	161.052	4	40.263
2-WAY INTERACTIONS	402.391	49	8.212
SPEAKER RATING OCCASION	62.775	9	6.975
SPEAKER RATER	290.616	36	8.073
RATING OCCASION RATER	49.468	4	12.367
3-WAY INTERACTIONS	160.288	36	4.452
SPEAKER RATING OCCASION RATER	160.288	36	4.452
EXPLAINED	873.188	99	8.820
RESIDUAL	.000	1	.000
TOTAL	873.188	100	8.732

60-second Young

Source of variation	Sum of squares	DF	Mean Square
MAIN EFFECTS	538.313	15	35.888
SPEAKER	109.291	9	12.143
RATING OCCASION	.188	1	.188
RATER	428.300	5	85.660
2-WAY INTERACTIONS	162.519	59	2.755
SPEAKER RATING OCCASION	34.770	9	3.863
SPEAKER RATER	107.208	45	2.382
RATING OCCASION RATER	22.001	5	4.400
3-WAY INTERACTIONS	87.082	45	1.935
SPEAKER RATING OCCASION RATER	87.082	45	1.935
EXPLAINED	787.913	119	6.621
RESIDUAL	.000	1	.000
TOTAL	787.913	120	6.566

60-second Old

Source of variation	Sum of squares	DF	Mean Square
MAIN EFFECTS	723.397	15	48.226
SPEAKER	229.837	9	25.537
RATING OCCASION	2.253	1	2.253
RATER	491.299	5	98.260
2-WAY INTERACTIONS	152.927	59	2.592
SPEAKER RATING OCCASION	20.267	9	2.252
SPEAKER RATER	118.993	45	2.644
RATING OCCASION RATER	15.973	5	3.195
3-WAY INTERACTIONS	61.681	45	1.371
SPEAKER RATING OCCASION RATER	61.681	45	1.371
EXPLAINED	938.004	119	7.882
RESIDUAL	.000	1	.000
TOTAL	938.004	120	7.817