DOCUMENT RESUME

ED 358 718 FL 021 277

AUTHOR Ginsberg, Ralph B.; And Others
TITLE Listening Comprehension Before and After Study Abroad.
INSTITUTION Johns Hopkins Univ., Washington, DC. National Foreign Language Center.
PUB DATE Jul 92
NOTE 65p.; For a related document, see FL 021 276.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Achievement Gains; College Students; Higher Education; *Language Proficiency; Language Tests; *Listening Comprehension; Listening Comprehension Tests; Questionnaires; *Russian; Second Language Instruction; *Second Language Programs; *Study Abroad
IDENTIFIERS American Council of Teachers of Russian

ABSTRACT

This study examined listening comprehension in 82 university students who participated in an American Council of Teachers of Russian language program. Questionnaire data reveal that specific listening activities are not common in college Russian courses, and that students have little confidence in their ability to comprehend what they hear in a variety of situations. In addition, there are few relationships between the activities that do exist and either students' perceived listening competence or results on an objective listening test. The experience of study abroad led to substantial gains in listening comprehension. Six figures and 19 tables are included. (JP)

National Foreign Language Center Working Papers

Listening Comprehension Before and After Study Abroad

by

Ralph B. Ginsberg
National Foreign Language Center and University of Pennsylvania

Richard M. Robin
George Washington University

Paul R. Wheeling
Spatial Information Systems

NFLC

National Foreign Language Center

Listening Comprehension Before and After Study Abroad

by

Ralph B. Ginsberg
National Foreign Language Center and University of Pennsylvania

Richard M. Robin
George Washington University

and

Paul R. Wheeling
Spatial Information Systems

Acknowledgment

Work by Ralph Ginsberg and Paul Wheeling was supported in part by the grant "Language Learning during Study Abroad: The Case of Russian" from the Ford Foundation, which is gratefully acknowledged.

Listening Comprehension Before and After Study Abroad

1. INTRODUCTION

This paper is concerned with how students acquire skills in listening comprehension. It is based on a detailed questionnaire asking students to assess their own competence in a wide range of listening activities before and after a semester-long program of in-country Russian language study sponsored by the American Council of Teachers of Russian (ACTR), and on standardized tests and background data collected by ACTR as part of its ongoing research into the effects of study abroad. The data enable us to address such key issues as the place of listening comprehension in the college Russian curriculum, the level of listening competency achieved before and after study abroad, the relationship of previous classroom activities to listening competency, factors related to gains in competence during study abroad, and, at a more technical level, the reliability and validity of self-evaluations of listening comprehension, the relationship of self-evaluations to other measures of language proficiency, and the factors shaping ability to comprehend the language to which students are exposed while abroad.

Learning to Listen

Listening comprehension is a latecomer to the second-language pedagogical arena as a skill in its own right. Instructors assumed that listening comprehension was a function of learning to speak (Nord, 1981; Joiner, 1984; Heron and Seay, 1991). In college Russian courses, training in listening comprehension based on authentic samples (excerpts of speech produced by native speakers for other native speakers) was virtually non-existent until the 1980s. Practice in the comprehension of connected speech was limited largely to understanding the lectures of teachers who tailored their speech to the student audience. Practice with authentic scripts (overheard conversations, films, radio, television, and theater), when available at all, was limited to upper division courses. At beginning levels, particularly in programs influenced by audio-lingual approaches of the early 1960s, systematic practice in listening was limited to oral pattern drill in which students listened to stimulus statements (often available to them only on tape) and

produced the correct rejoinder.[1] While pattern drilling requires "listening," it places no demands on global communicative comprehension, inasmuch as the patterns are discrete sentences, often taken from previously memorized dialogs.

Today, despite the increased emphasis on real-world skills that gave rise to proficiency testing, first in the government, and later in educational institutions, training in listening comprehension as a separate skill is not normally an integral part of college Russian-language programs or commonly used materials. In fact, in programs in which "speaking" (i.e. grammatically accurate oral production) receives priority, students often find that they are able to ask a structurally complicated question but not understand a rather simple answer. Even in higher-level coursework, students who can follow college lectures prepared specifically for language learners, say a biographical sketch of a famous writer, are hard put to get the gist of straightforward broadcast news reports, movie schedules dictated over the phone, and announcements made over the public address system at airports or train stations. More complicated speech, such as movies and plays, is far beyond their reach.[2] To a large extent the only way for students to acquire any facility with authentic speech is during study abroad — the principal focus of this report — an opportunity possible for only a small percentage of students, where, the problem is rather "shoved under the rug" than addressed head on.

Incentives to re-examine our foreign-language curricula have come from grant-giving agencies, which have increasingly limited funding of programs involving foreign-language instruction to those institutions which can demonstrate adherence to established proficiency guidelines. As a result a proficiency teaching and testing infrastructure has now come into place. For students of Russian, the ACTFL oral proficiency interview is widely administered to test speaking. The Educational Testing Service offers tests in the remaining three skills and is developing a testerless speaking

---

[1]The original *ALM Russian, Level One* (Modern Language Materials Development Center, 1961) and Baker's *Russian for Everybody: Version for Americans* (1986) are the two best examples of audio-lingually-based textbooks containing dozens of hours of taped drill without student scripts.

[2]Programs in language instruction which stress "speaking" at the expense of practice in listening as an independent skill would appear to be at variance with the needs of the real world. Even after the break-up of the Soviet Union, one of the largest employers of recently graduated Russian majors is the federal government. Agencies, such as the State Department, the CIA, the National Security Agency (NSA), and the FBI Information Service (FBIS) have specific needs for proficient listeners and have developed specialized programs.

test. Since its release in 1986, the ETS Advanced Russian Listening Proficiency Test (referred to simply as the ETS Test where there is no danger of confusion) has been administered to approximately 4,500 examinees, many of them participants in ACTR programs where it is administered as part of ACTR's research. Despite their prevalence, how the tests should be constructed, and indeed what they really measure, remain highly controversial issues that animate the field.

The growth of proficiency testing has, in turn, fueled the development of proficiency-based curricula and materials. In the ETS Listening Test language situations are made as real as possible, with passages taken largely without adaptation from the Russian-language media. Partly as a result of the test, some teachers have begun to emphasize authenticity in class, e.g. by using Russian-language audio and video materials. Secondly, interest in proficiency has brought to the fore the role of communicative strategies — conversation management techniques, gisting, use of cultural and structural knowledge to set up expectations and schemata, etc. — in listening comprehension and how they might be acquired (Richards, 1983; Meyer, 1984; Dunkel, 1986; Bacon, 1989; Phillips, 1990; Long, 1990; Laviosa, 1991). Finally, interest in proficiency has changed the way we look at the "four skills." While previously, speaking (and sometimes reading in non-audio-lingual settings) served as the foundation for the others, with grammar the "engine" for moving all the skills forward — a picture which differs significantly from that of foreign language skills acquired abroad, where comprehension often precedes production — instructors have recently begun to examine each skill as a goal in its own right, and to consider how our "engine(s)" might be adjusted to move each skill forward effectively.

Research questions

These considerations raise a number of interconnected empirical and pedagogical issues that set an important part of the research agenda for the field. Study abroad plays a strategic role, both in terms of the learning opportunities it affords and as a research site where, as a matter of course, competence and its determinants are, as it were, put to the test. Based on data from our questionnaire and the ACTR database, this paper addresses the following specific questions:

- What training for listening comprehension is actually provided in college and university Russian programs? How important is it in the overall curriculum? How widely used are authentic materials? Are there attempts to teach communicative strategies?

- Does work with authentic materials in the classroom lead to better comprehension of authentic speech in-country? What particular kinds of speech are affected?

- Is there a place for inauthentic ("modified") materials in the pursuit of listening proficiency in the classroom (Pica and Young, 1987)? Can a student profit from "teacher talk," adapted materials, or the flawed Russian of other students?

- What is the relationship between listening comprehension and other skills? Are good listeners likely to be good readers or good speakers, as suggested by Coakley and Wolvin (1986)? Can training in speaking "leverage" training in listening?

- Is study abroad effective in improving listening comprehension? What listening competencies are particularly affected?

- Does previous intensive study at a home institution facilitate gain in listening comprehension over the course of an in-country experience, e.g. by sensitizing them to particular learning opportunities or imparting specific learning strategies?

- Are there any other factors besides pedagogical practices which are related to listening ability and changes during study abroad which might provide useful clues for improved selection, guidance, and organization for study abroad programs?

In order to answer these questions with the self-evaluations on our questionnaire, some technical questions of measurement, which nevertheless have very interesting substantive implications, have to be addressed. These include:

- Are students consistent in their evaluations of their own listening proficiency in authentic settings? Can their evaluations be used to construct useful scales of listening comprehension and its changes in authentic situations?

- How do self-evaluations compare with other, so-called "objective" measures of listening proficiency, and in particular with the ETS Listening Test? How do their self-evaluations relate to their performance on the Oral Proficiency Interview?

- What makes listening comprehension more or less difficult in different speech situations? Does the relative difficulty of different situations change after study abroad? If the objective is comprehension of "authentic" speech, are currently hierarchies of difficulty based on structural/thematic complexity (as originally suggested by Child, 1986) useful, or must other factors be taken into account?

Overview

The next section discusses the mechanics of our empirical methods: the respondents, the questionnaire and its administration, data management and data analysis. In Section 3 we present the picture of training for listening comprehension in American colleges and universities that emerges from the responses to Question 3 of the questionnaire, devoted to this topic. Section 4 is devoted to an analysis of the respondents' self-assessed competence, before and after their program abroad, in comprehending seventeen frequently encountered speech types. After examining how well students think they can handle the seventeen types (4.1), we take up the key question of combining the responses to form a single scale of listening comprehension (4.2). The construct validity of the scale is assessed by analying the calibrated item difficulties (4.3) and the scale's relationship to the ETS Listening Test and the Oral Proficiency Interview are examined (4.4). Section 5 is concerned with the relationship between the educational practices and listening comprehension. Both the scaled self-assessments and the ETS Test are used as criteria. Relationships to other student characteristics are also reported. In Section 6 factors affecting changes (or rather gains, since everyone improves) in listening comprehension are explored, again with emphasis on previous formal educational experiences. For self-assessments the results are largely negative, and for the ETS Test the results agree with the larger study of the ACTR data reported in another paper in this series (Ginsberg, 1992). Section 7 is a brief conclusion, relating the results to the research and pedagogical issues raised above.

2. DATA COLLECTION AND ANALYSIS

Our primary data derive from a questionnaire which was administered to participants in the Spring 1990 academic semester ACTR language programs at six Soviet host institutions, five in Moscow and one in Leningrad, in conjunction with ACTR's regular post-testing procedures. Students received the questionnaire shortly before the oral proficiency interview and were informed that the questionnaire was not required. Most filled it out while waiting for their interview or immediately thereafter and returned it to the OPI tester.

The questionnaire is reproduced as Appendix 1. Question 3 is the principal source of data on exposure to spoken Russian during formal training occurring before the study abroad experience. It asks the students to report, for each year of college Russian, summer program, and study abroad program, how often they were exposed to each of

eight specific kinds of speech. (The undergraduate school itself is identified in Question 1; the summer and in-country programs were to be noted on the questionnaire.) Question 4 asks the students to rate themselves, both at the beginning and at the end of their study abroad program, on their ability to comprehend seventeen speech types, using a five point scale ranging from 0 ("virtually impossible") to 4 ("easy: understand virtually all").[3] It is apparent in the format of the questionnaire, however, that the respondents are implicitly being asked to report changes resulting from their study abroad experience. The items were chosen to represent common situations encountered by students studying in Russia and to cover the range of difficulty commensurate with the listening abilities evidenced by American college students. Rationale for the individual items and the psychology of answering the questionnaire are discussed in Section 4 below in connection with the construction of a scale of listening competence.

Eighty-two of the 102 eligible students responded.[4] The respondents studied at 51 different colleges and universities, located in every region of the United States, before their ACTR program. There is considerable variety in the college programs reported. Even students from the same undergraduate institution generally report different classroom experiences on Question 3, evidently because pedagogical practices vary in different years and in different sections. The students' reports are, however, by and large consistent with what we know about the undergraduate programs with strong listening comprehension components. Nineteen respondents reported attending an intensive summer program (and a few reported more than one), and here too there is some variation in the experiences of students from different years. Several students reported an in-country immersion program, but these data are hard to use. Some reported their current ACTR program and others did not; among those reporting a

---

[3]Students could also respond "not applicable" (NA). These responses are interesting in themselves as an indication of what students did — or rather did not do — and what kinds of language they were exposed to while abroad. Table 1 gives the number of students who never experienced the types of speech enumerated in Question 4. By and large if an item was not experienced at the beginning it was not experienced at the end; but on items q. and n. there were quite a few more NA's at the beginning than at the end. There is clearly considerable variation by item. We leave the implications for program design to subsequent discussion.

[4]Comparison of the respondents with the twenty students who did not respond (either because they chose not to or because they did not receive the OPI and accordingly were not asked) on factors shown to be related to listening comprehension in previous studies (gender, knowledge of other languages, OPI scores, reading comprehension, etc.) indicated no non-response bias. It should be noted, however, that the respondents are only representative of ACTR's programs.

program, several failed to give the dates, making it impossible to tell which are prior to the present program. Nevertheless, as explained presently, in most cases, using notations by the students on the questionnaire and linked ACTR data, we were able to make plausible guesses as to whether or not the student had a true previous immersion.

All of the data were entered in a database for subsequent processing, with fields and values corresponding to the questionnaire. Students were not asked to give their names (although several did anyway). On the basis of their undergraduate institution, study Institute, years of Russian, previous immersions, and other collateral information, it was, however, possible in 79 of the 82 cases to link the records in our study to the ACTR database. The link enabled us to expand the scope of the study substantially by adjoining such variables as gender, age, knowledge of other languages, ETS scores, and OPI scores to the information derived from the questionnaire.[5] Moreover, by comparing responses on the questionnaire with the ACTR data, we were able derive more valid measures of two key variables in the analysis — years of college Russian and whether the student had a previous immersion. These variables are designated as *Years and *prevImm below.[6] The anonymity of the respondents was, of course, preserved.

We have tried to be inclusive in presenting the descriptive findings since there is so little empirical data available, especially for samples of the size of ours, on the factors addressed on the questionnaire. For the most part the analysis follows standard statistical and data analytic practice (graphical displays of distributions, ANOVA, Regression analysis, etc.). The only "advanced" technique relates to the construction of a scale from the items on in-country listening activities in Question 4 of the

---

[5]Because of missing data on some variables in the ACTR database, the number of cases on which analyses involving these variables are based, is generally less than the theoretical maximum of 79.

[6]Since there is only one student with one year of college Russian, and only five with more than four years, *Years recoded to 2 = (1 or 2), 3, and 4 = (4 and above). *prevImm is a 0/1 dummy variable. Cross-classification of *Years by *prevImm shows an expectable moderate relationship. The following table gives the base numbers for the plots in the Figures below:

		*Years			
		2	3	4	Total
*prevImm	No	24	30	18	72
	Yes	3	2	5	10
	Total	27	32	23	82

questionnaire, using Rasch-type models of item difficulty and individual ability. The implications and uses of the scaling models are quite intuitive and will be discussed in Section 4.2.

3. EXPOSURE TO SPOKEN RUSSIAN IN FORMAL EDUCATION

Question 3 of the questionnaire lists eight common ways in which students are exposed to spoken Russian in undergraduate and intensive summer programs and asks the respondents to report how frequently each figured in their previous training. That these activities are common does not necessarily mean that they are productive of the listening comprehension skills required of students living and studying in Russia and coping with native speech on a day to day basis. Indeed each differs from "authentic" speech in significant respects. Appendix 2 summarizes the differences in terms of the factors affecting listening comprehension noted in the Introduction. Thus it is an open empirical question (addressed in Sections 5 and 6) as to whether any of these activities affect the initial levels of what students can do, or, more subtly, whether they affect gains during study abroad by enabling students to take better advantage of in-country experiences.

Tables 2-5 show the prevalence of the eight listening related activities in the first, second, and third years of college Russian, and in intensive summer programs. In Table 2, for example, n = 79 students report a regular first year of college Russian course. Of these 14 (18%) say that they never had conversations with the teacher, while 18 (23 %) had such conversations practically every lesson. The tables document the characterization of the college Russian curriculum in the Introduction. In every single year, the majority of students (the overwhelming majority beyond first year) engage in classroom conversations with their teachers "sometimes," "regularly," or "often" (59%, 73%, and 88% in years one, two, and three or beyond, respectively). Most also report hearing Russian from other students more than "rarely" (72%, 68%, and 92%, as above). Before third year Russian, large majorities report some (more than "rare") listening input from grammar drills (73% and 67% in first and second years). Of the common activities involving extended flow-of-speech listening we find a large amount of input (again, more than "rare") through lectures by the teacher, although this form of instruction is slightly less predominant. Not quite half (46%) of second year respondents reported exposure to lectures by the teacher; in other years beyond first, the figure was a bit under two-thirds (63%).

Other activities in which listening comprehension plays a greater role appear more rarely. Beyond first-year Russian, where the figures are understandably (but perhaps not justifiably) low, only around 30% of the respondents were exposed to authentic video on more than a "rare" occasion, and only around 15% had listened to authentic audio recordings (without video). Even the traditional dictation, which requires major elements of the skills needed for listening comprehension, is not as widespread an exercise as might be expected. Of all respondents in all the levels only around 40% reported more than a "rare" dictation.

Intensive summer programs, some of which try to emphasize productive skills, seem to be similar in practice to year-long courses, with a stronger emphasis on teacher-student conversations and listening to other students but quite comparable figures for video, audio, and dictations.

In addition to the activities specifically designated on the questionnaire, many students reported at least one other listening related activity. These activities are listed in Table 6, and their frequencies are given as Items i and j of the tables. Unfortunately, these activities are too rare to analyze individually and too heterogeneous to analyze as a group. What is perhaps surprising is how few students mention anything at all, especially in the first two years of classroom study — a fact which emphasizes the relative neglect of instruction aimed specifically at listening comprehension. In the intensive summer programs too, activities not included in the eight specified on the questionnaire are relatively rarely mentioned.

4. SELF-ASSESSED RATINGS OF LISTENING COMPREHENSION

4.1. Response Frequencies

In this section we examine the respondents' self-assessments of their ability to comprehend spoken Russian in the seventeen different situations that make up Question 4 of the questionnaire. As noted in Section 2 these items were chosen both to be representative of the situations encountered by students during their stay abroad and to cover the range of difficulty commensurate with their initial and final abilities in the respondent group. Before we turn to the central question of whether the responses can be used to construct a scale measuring item difficulty as we have characterized it, and at the same time measuring individual ability before and after the program, it is of

some independent interest to look at the frequencies of responses to the pre- and post-program items (Tables 7 and 8, respectively).

As one would expect on all items there is significant variation among students in abilities reported, and there is significant variation among items in the number of students assigning themselves to each category. Consequently all items are useful in the measurement process. It is noteworthy that not even on the apparently easiest items (a. "teachers talking to you" and c. "friends talking to you in mixed company") do the majority of students feel that they can get most of what is said (i.e. categories 3, "not too hard: get most of it but miss some details," or 4, "easy: understand virtually all"). On the other hand, only on the hardest items do an appreciable number find comprehension "virtually impossible." For the most part students rate themselves in categories 1 or 2 ("hard to get" or "stressful, but can get the essentials"). Their self-assessments are dramatically different after their study abroad program, with 3's and 4's predominating except on the more difficult items (e.g. q. "street meeting and demonstrations" and n. "live plays"). Clearly, from the students' own point of view, substantial gains are made during study abroad. We return to the implications of these results in the concluding section of the paper.

4.2. Scaling: Measurement of Individual Ability and Item Difficulty[7]

As interesting as the individual items may be, for research (and administrative) purposes it is necessary to combine the responses to get an overall measure (or at most a few measures) of listening ability. Whether a set of items can be used to measure an underlying construct — e.g. listening comprehension — representing individual ability is an *empirical question* which has implications both for the items and for the people. Intuitively, people whose ability is greater than an item's difficulty should have a better than even chance of being able to "do" it, the probability increasing with the difference, while people whose ability is below the item difficulty should have a less than even

---

[7]The exposition of the logic of scaling in this section and the specific psychometric models employed are based on the work of Benjamin D. Wright and his colleagues. For an excellent tutorial presentation see Wright and Masters (1982) and Wright and Stone (1979). On the psychometric models themselves see also Andrich (1978a and 1978b). Estimation of item difficulties and individual abilities was carried out using the computer programs MSCALE by B. D. Wright, M. Rosner, and R. T. Congden, and MSTEPS by Rosner, Congden, and Wright, as modified by Norman Katz and RBG. The literature on scaling is vast and many alternative models have been proposed. While we cannot go into the issues here, suffice to say that the models we use have a cogent rationale and that our data happen to fit their assumptions extremely well.

chance, etc. Thus, in a probabilistic sense, most people should be able to "do" the easy items, but only the most able people should be able to "do" the difficult ones; conversely the easiest items should be "doable" by most people, but the most difficult one "doable" only by the most able. To put the matter from a negative point of view: against the background of the responses of all people to all items taken as a whole, an item that can be "done" by the least able people and not by the most able (again in a probabilistic sense), is measuring something else than intended, and hence does not contribute to the scale; and conversely, the responses of a person who can "do" the hard items but not the easy ones are determined by something other than his ability. It is the overall coherence of people and items — the overall pattern of the data — that really matters. Too many deviations from the expected pattern, or distinct subpatterns in the overall picture, imply that other factors besides the presumed underlying construct must be postulated to account for the data.[8]

A consequence of this intuitive reasoning can be seen clearly in Table 9 where the responses of each of the 82 students to each of the 17 pre-program items (i.e., the "starting" column of Question 4) are displayed. In the Exhibit the students are listed in decreasing order of their average score on all items, a crude measure of ability. The items are listed from left to right in increasing order of average reported difficulty, a crude measure of difficulty. If all people and all items "scale," there should be many 4's in the upper left-hand corner (high ability, easy items); the numbers should decrease (i.e. reported difficulty with an item should increase) as we move to the right (harder items) and down (less able people); and in the lower right-hand corner there should be mostly 0's and 1's. Deviations from this pattern indicate that the intuitive model does not hold. Remembering that expectations hold only in a probabilistic sense, the Exhibit is a textbook example of what the data ought to look like! A similar exhibit for the post-program data (not given here) shows the same pattern, again a textbook example.[9]

---

[8]For example, there would be evidence for specific knowledge and strategies about the Russian language media if some students were consistently better than expected on all media items, while other students were worse. Similarly, after overall levels of ability had been taken into account, there would be evidence for specific competencies in dealing with acoustically difficult environments if some students nevertheless were better than expected and others worse on all items were acoustics is salient.

[9]Another consequence of the basic intuition concerning item difficulty relates to pairs of items. If responses to any two items are compared by means of a two-way cross-tabulation, there will be more people rating themselves high on the easier item than the harder. This implies that in the table one item is harder, equal, or easier than the other depending on whether the preponderance of people is above, on, or

While for some purposes the average scores used in Table 9 would be sufficient to measure respondents' listening comprehension ability, for reasons discussed by Wright and Masters (1982) — which include stability across samples of items and people, robustness to extremes in the data, and measurements of goodness of fit, reliability, and validity — in the analysis that follows we employed the Rating Scale model developed by Andrich (1978a and 1978b), Wright and Masters (1982), and Masters (1980), as implemented in the computer program MSCALE. The parameters of the model, which are estimated by maximum likelihood by MSCALE, are the required measures, namely: the *item difficulties*, which can be compared with *a priori* characterizations to determine the validity of the scale and to suggest further hypotheses about comprehension difficulty; *"step" values* (analogous to thresholds in other ordinal variable models) which complement the item difficulties in determining choices among of the five response categories; and the *ability levels* of each of the respondents, which will be related to previous educational experience in Section 5, and are the basis of our analysis of the changes consequent to study abroad program, in Section 6. The ability measures are referred subsequently to as **preMSC** and **postMSC**.

It would take us too far afield to go into the details of the psychometric analysis here. For the interested reader the key results are explained in Appendix 3 and its associated Tables 10 and 11, and Figures 1 and 2. The upshot of the analysis is that

> *The model fits every single individual and item well: there clearly is an underlying scale that characterizes the students' self-assessments, both before and after the study abroad program.*

One of the most important properties of the Rating Scale model is its putative invariance to the sample of respondents on which it is calibrated; i.e., the same item difficulties should obtain regardless of the sample from which they are estimated. In particular, estimating item difficulties from two large *independent* samples, one studied before a study abroad and one after a study abroad program — different samples with the latter having arguably greater average listening comprehension ability — should give the same item difficulties. In our case the pre and post samples are obviously neither large nor independent, but still the item difficulties ought to be *more or less*

---

below the main diagonal. All two item comparisons should be consistent in their implicit item ranking. Examination of the data from this point of view again shows the pattern one would expect if the measurement model holds and produces the same ranking of item difficulty as the ranks of average scores.

consistent. Table 12, which gives rankings of item difficulty both before and after the study abroad program, and for students who have not had a previous in-country immersion before the current ACTR program, addresses this issue. The Table also gives the ranking based on average scores, the classical measure of item difficulty. Allowing for sampling variation, it is apparent that, no matter what the sample or model, the measures are indeed similar, a further confirmation of our scaling procedure. What differences there are, e.g. the exchanged positions of live plays and Russians talking among themselves, can be accounted for by differential exposure in the study abroad experience.

4.3 Construct Validity and Item Difficulty

We turn now to the *construct validity* of the self-assessed scale, i.e. whether it measures what it is intended to measure. With an item response model construct validity turns on the question of whether the measured difficulty of the items accords with our intuitive judgments of difficulty. Now, the ease of difficulty of comprehending speech depends, of course, among other things, on what kind of speech it is. To note a few of the salient factors differentiating so-called "authentic" speech situations encountered by students abroad:

- Comprehension is easier in interactive settings where the listener can interrupt the speaker for repetition and clarification than in settings such as lectures, the media, announcements and overheard conversations, both real and in the movies where such clarification is not possible.

- Comprehension is easier when supported by visual cues than when sound alone is involved (Coakley and Wolvin, 1986).

- Comprehension is easier in socially supportive or socially neutral environments (encounters with friends) than in socially tense environments (encounters with service personnel, the bureaucracy, etc.) or apprehensive environments (Meyer, 1984; Coakley and Wolvin, 1986; Bacon, 1989).

- Comprehension is easier in acoustically normal environments (small closed rooms) than in acoustically hostile environments (movie houses, theaters, large halls, loudspeakers, street conditions, bad phone connections).

- Comprehension is easier for short, scripted items (such as straightforward commercials and weather forecasts) than for longer items with complex, less predictable rhetorical structure (such as talk shows, detailed factual reports).

- Comprehension is easier when accents are standard and registers are neither too informal nor too formal.

In any given type of speech situation factors such as these combine with the listener's general level of ability, specific strategies he or she may have to cope with the source of the particular difficulties, factual knowledge of various sorts, and the structural complexity of the language itself, which underlies such categorizations as "novice," "intermediate," "advanced," and "superior" to determine how much he or she "understands." Selective attenuation of particular sources of difficulty accounts for the difference between "authentic" and "inauthentic" speech.

Appendix 4 summarizes how these general considerations apply to the speech types enumerated in Question 4. For any of the types, different respondents will bring to mind different prototypical experiences, as do we. Accordingly, hard and fast correspondences can not be expected. Still, for the most part the ordering of difficulty revealed by MSCALE conforms to our intuitive notions, and the discrepancies raise some intriguing questions. Interactive activities are generally easier than those involving "flow-of-speech" (broadcasts, movies, overheard conversations). The one exception is classroom lectures, rated third easiest pre and post by MSCALE, probably because the lecturers knew that their audience was foreign and modified their presentations accordingly. The item difficulty estimates also confirm that activities involving visual cues are consistently easier than similar audio-only counterparts; e.g., talking to friends and strangers face-to-face is easier than talking to them on the phone, and watching the news is easier than listening to it. It is important to remember, however, that the dimensions of difficulty are predictive only *ceteris paribus*, since other factors are also at work. Thus street meetings (Item q) are the hardest to understand, in spite their large visual component, because the poor acoustics and a lack of the necessary cultural and factual background. As to public affairs television, talking heads are not really usable video input, but the public affairs broadcasts mentioned as examples of the kinds of programs we had in mind, "Vzglyad" and "Pyatoe koleso" are usually visually rich (an apt comparison is "60 Minutes"). Finally, for the most part, acoustically normal environments are indeed easier to deal with than acoustically

hostile ones, again with the caveat that other factors also apply (as in the case of talking to friends on the telephone, for example, where friends are no doubt helping out).[10]

4.4 Relationship of Self-Assessments to the ETS Listening Test and the OPI

To validate the scale further it would be desirable to compare it with an external criterion. Unfortunately, the only measure available is the ETS Listening Test, a test whose validity is at least as problematic as the student self-assessments.[11] Normal plots for the distributions of scaled pre-program measure (preMSC) and the ETS Test are shown at the top of Figure 3. The straight line pattern indicates that both variables are normally distributed (their smoothed histograms look bell-shaped). The boxplots in the middle of the Figure show the distributions for each of the three levels of *Years. There is no apparent relationship between either variable and the number of years of Russian study. The scatter plot at the bottom relates these variables to one another; the correlation is .423 (based on n = 74 cases). The prominent points in the plot represent students with a previous in-country immersion program; the correlation would seem to be the same were they excluded. The correlation between the MSCALE post-program self-assessments and the post-program ETS is .292 (based on n = 75 cases). For cross-sectional data the correlations are respectable, although the two procedures are

---

[10]Quite idiosyncratic factors may be operating as well. For example, on acoustic grounds one would expect TV movies to be easier than movies in movie houses (unless Russian movie houses have drastically improved their sound systems), yet the two are nearly equivalent. To interpret the data one must take account of the specific movies involved. A sampling of the movie and television schedules in Moscow and Leningrad in the semester in question shows that the bulk of films shown on television were Soviet productions about Soviet realia, while the movie houses were billing dubbed American blockbusters such as *Star Wars*. Clearly, a studio-dubbed American film, with no foreign cultural baggage to get in the way, presents much less of a comprehension challenge.

[11]Educational institutions order the testing kit from ETS. The kits includes test booklets, forms, and a test tape. The acoustic quality of the recorded passages on tape varies significantly, and there is no control over the acoustics of the room in which the test is administered. Institutions such as Middlebury College have administered the test in the language lab with headphones. ACTR has administered the test in rooms of various sizes with a "boom box." In the current ETS Listening Test (1986) students listen to 17 passages recorded on tape, accompanied by multiple choice questions printed in the test booklet. The passages, arranged in order of assumed increasing difficulty, are read once. Most of the passages at the Novice and Intermediate Mid level are semi-authentic, that is, indistinguishable from authentic, but written specifically for this test. Many of these take the form of overheard conversations. Nearly all the passages beyond the Intermediate level are authentic and take the form of weather broadcasts; news reports; passages from college lectures, etc. Before each passage, students are given the opportunity to read the corresponding multiple choice questions in the test booklet. Then, after the passage has been read, time is allotted so that students can mark the correct answer.

obviously measuring different things. The reduced correlations pre and post indicate, in our view, that self-assessed ratings, with the descriptions of the response categories specified on the questionnaire, are more sensitive to changes during study abroad than the ETS Test, given its own special format, response categories, and administration. For a point of comparison with the scaled self-assessments, we examine initial levels and changes on the ETS Test along with the self-assessments in Section 5 and 6 below.

While the OPI as currently administered has no face validity as a listening test (especially for non-interactive activities), it is also interesting to relate the MSCALE self-assessments to pre and post-program OPI scores,[12] After all, the two skills should be related (e.g. Feyton, 1991). Figure 4 shows boxplots of scale score by OPI score, pre and post, and ANOVA's testing the significance of the difference. It is clear that there is a very strong relationship between the students' self-assessments and the OPI, with self-assessed listening increasing dramatically as OPI goes from 0+ to 2 and above. The F-test in the ANOVA is highly significant both pre and post ($p < .0001$ and $p < .0006$, respectively), with perhaps a slightly weaker relationship post due to the differential sensitivity of the response categories on the two instruments ("easy: get most" vs. "intermediate") to differences in average ability after a study abroad program. Cross-tabulation of OPI scores with self-assessed competence in the seventeen speech activities enumerated in Question 4 shows many significant associations. Although we cannot pursue the matter here, it is interesting to note (see Table 13) that by and large the associations between the OPI and specific competencies of Question 4 are stronger for interactive activities (teachers, friends talking to you, etc.), where speaking and listening are closely bound together, than for the flow-of-speech activities (Russians talking in your presence, radio, movies, public address announcements, demonstrations), where listening alone is involved. The TV Items (j. and k.) are notable exceptions. Whether this is an artifact of the varying difficulty of the items or reduced sample sizes due to lack of exposure, or whether it is a genuine effect, would require more data to determine.

---

[12]Presumably the OPI only measures speaking. Thus, at the Novice and even Intermediate levels, where volume of language output, whether communicative or not, is critical, answering the "wrong" question comprehensibly does not necessarily detract from the final score. For example, in answering the question "What does your father do for a living?", an Intermediate Low *speaker* might say "He lives in Ohio. All my family lives in Ohio. I was born in Ohio too." Lowe (1985) finds data in government ILR-OPI testing to suggest a constant positive comprehension offset for French and Spanish, specific to language and level. Similar research for Russian would be desirable. Perhaps a modified form of the OPI can be developed to assess comprehension, at least through the Advanced level.

The whole issue of what the various instruments are measuring and how they are interrelated merits careful study, without prejudice to which of the available measures is the more valid.

5. RELATIONSHIP OF PREVIOUS TRAINING TO INITIAL LEVELS OF SELF-RATINGS AND ETS LISTENING PROFICIENCY

5.1. Summary Measures for Learning Activities

Now, one of our primary questions is whether the listening related learning activities in college bear any measurable relationship to students' (perceived) competence in listening tasks. Do students in programs which emphasize the listening related activities of Question 3 perform better on the listening competency criteria of Question 4 than students in programs in which these activities are not found? To address this issue, especially in a sample of moderate size, it is necessary to summarize the previous learning experiences reported on the questionnaire. We experimented with a number of different measures but ultimately chose, for each of the eight named and student specified activities in Question 3, to simply sum the numbers (0, 1, 2, 3, or 4) reported over the student's whole formal (college and intensive summer) learning career, excluding in-country programs.[13] These variables are labeled **rowA, ... , rowH** below.

Figure 5 shows distribution of these sums for each specified learning activity. The sums are correlated with *Years (of college Russian) and with one another, since students with more years of study have more opportunities to engage in the activities. The correlations are far from perfect (i.e. there is no colinearity problem), however, and the total of the activities is not equivalent to the total number of years, since the specified activities are by no means an exhaustive breakdown of all learning activities in college Russian. Our analytical question then becomes:

---

[13]Of course, in-country programs, summer or semester, provide learning experiences going far beyond what is possible in a college classroom, and their characterization would require a separate study. They do not enter the row sum variables but are explicitly taken into account in the analysis with the dummy variable *prevImm. Preliminary analyses showed that intensive summer programs did not have any special effect beyond their contribution to years of study and the row sums, and hence were not included in the baseline. In further preliminary analyses the sums of the frequencies of the designated activities in each year were computed, and interacted with the row sums, to explore the possibility that *when* learning activities occur, in addition to *what* they are, affects subsequent performance. The results were consistently negative.

Over and above the total years of study (whatever learning experiences that may have consisted of) and any previous study abroad experiences, do the specific pedagogical activities reported in Question 3 enhance the listening competencies reported in Question 4?

For example, when students with equivalent numbers of years of Russian and study abroad experience are compared, do the students whose programs included frequent teacher-student conversations do better than the students for whom teacher-student conversations were rare?

5.2. Regression Strategy

To assess the effects of the college-based listening comprehension activities and other background factors on our two measures of pre-program listening ability — called **preMSC** and **preETSL** in the outputs — and three measures of change in comprehension to be discussed in Section 6, a guided stepwise regression strategy was used.

First, a baseline of the grossest factors affecting the criteria was established. These are variables that must be controlled to arrive at meaningful assessments of activities effects. In the case of the pre-measures this was simply years of college Russian (***Years**) and whether or not the student had a previous in-country program (***prevImm**). In the case of the change measures the baseline consisted of ***Years**, ***prevImm**, and the pre-program ability level (**preMSC** or **preETSL**), which in this study, as in the ACTR data as a whole, dominates the prediction of change.

Second, all "row" variables (row sums in Question 3 of the questionnaire, referred to as **rowA**, ... , **rowH** below) were added to the baseline and an F test computed to determine their joint significance. Highly non-significant ($|t| << 1.$) row variables were eliminated, producing an intermediate model for closer examination.

Third, variables were dropped and added, singly and in pairs,[14] with F tests computed, to arrive at a final assessment of which row variables were significantly related to the criterion.

Fourth, any nonsignificant variables in the baseline were eliminated to determine a "good" model.

---

[14]Dealing with variables in combination (e.g. pairs) distinguished this guided procedure from the automated one-variable-at-a-time methods found in regression packages.

Finally, using the "good" model as a baseline, effects of the background factors **Gender, Age,** number of nonslavic languages studied (**nonSlav**), and initial levels of oral proficiency, measured by the OPI (**preOPI**), and reading proficiency (**preETSR**), measured by the ETS Reading test — variables which had been shown in the ACTR analysis to be related to changes in listening comprehension — were examined as a matter of general interest.

It should be emphasized that our whole strategy is distinctly exploratory, designed to summarize the data and suggest relationships rather than to test prespecified hypotheses.

5.3. Regression Results

The key regression results concerning educational and background factors affecting pre-program competency are presented in Tables 14 and 15. The full analyses may be summarized as follows:

preMSC

In the baseline ***Years** is not significant (t = .99 in Table 14.a). The eight row variables are jointly significant, but some are highly significant and some nonsignificant. Eliminating the nonsignificant ones, a plausible intermediate model to start step 3 above (Table 14.b) consists of **rowA** ("teacher-student conversations"), **rowD** ("language lab 'grammar' drills") and **rowH** ("video tapes from Soviet media"), the latter two having negative coefficients (i.e., they seem to be *counterproductive*). **rowA** is highly significant by any test. **rowD** and **rowH** are jointly significant (below the .05 level, comparing **A, D,** and **H** with **A** alone), but most of their significance resides in **rowD** (comparing **ADH** with **AD** and **AH**). Thus the strongly counterintuitive TV effect is an artifact of its chance correlations with the significant row variables (**A** and **D**). Dropping ***Years** from the analysis does not change these conclusions: only the "row" effects and a previous immersion program matter. The resulting "good" model (step 4) is given in Table 14.c. It implies that

- The quantity of teacher-student conversations *per se*, quite apart from anything else that went on in the classroom, has a positive effect on self-assessed listening ability.
- Language lab "grammar" drills, by contrast, seem to be counterproductive.
- A previous immersion program is (not surprisingly) beneficial over and above classroom study.

As for the background factors (Gender, Age, and nonSlav), none are significant, nor would they be expected to be. (In the ACTR database they affect change, not initial levels.) Reading Proficiency (preETSR) and oral proficiency (preOPI) are correlated with self-assessed listening proficiency (with the effects of previous training removed), i.e., taking into account significant factors in their educational history, students who are proficient in one skill are proficient in others.

preETSL

The results for preETSL are essentially negative. In the baseline, as was the case with preMSC, years of study is not significant but a previous in-country program is. None of the "row" variables is significant, singly or in combination, with the possible exception of rowF (taped texts just for language learners), the variable most closely related to the format of the test. The t statistics for rowA, rowB (negative), and rowF approach significance, although they are jointly not significant; as these variables are dropped the others lose significance, and none has a strong enough effect to enter by itself starting with the baseline. The t statistic for rowF by itself is 1.53, which is not quite significant at the .05 level (one tailed). Erring on the liberal side, our "good" model is given in Table 15.c. As was the case with preMSC, none of the background factors (Gender, Age, and nonSlav) is related to preETSL, while preOPI and preETSR are highly correlated with and without the effects of *prevImm and rowF removed.

5.4 Pedagogical Activities in Question 3 Related to Speech Types in Question 4

Before leaving the topic of effects of activities on skills, it is interesting to look briefly at the relationships between specific activities and specific competencies, relationships whose plausibility motivated the construction of the questionnaire. Simply from the definitions, teacher-student conversations (rowA of Question 3), ought to be related specifically to how well students understand teachers (Item a. of Question 4); similarly lectures read by the teacher (rowC of Question 3) should be related to understanding classroom lectures (Item i in Question 4); and exposure to audio and video tapes from the Soviet media (rowG and rowH of Question 3) ought to be related to handling the media abroad (Items j, k, l, o, and p in Question 4). Using the factors affecting listening comprehension noted in the Introduction, one might make a case that previous exposure to Soviet media (rowG and rowH) might help with movies and plays (Items m and n in Question 4), on the grounds that, like the media speech types, movies and plays cannot be interrupted, are often in acoustically hostile environments, etc.; and that teachers talking to you (rowA) should help with Russian friends talking to you

(Item b of Question 4), on the grounds that both are trying to facilitate understanding in a face-to-face, generally non-stressful environment. To the extent that no such specific case can be made, relationships between Question 3 activities and Question 4 speech types should be weak; e.g., there should be little relationship between use of the Soviet media (**rowG** and **rowH**) and understanding friends (Item c).

To explore these hypotheses partial correlations and regressions (controlling for *Years and ***prevImm**) for many combinations of row sums from Question 3 and self-assessments on items in Questions 4 were computed. Teacher-student conversations *are* clearly related to understanding teachers talking to you (t = 3.5), Russian friends talking to you (t = 3.5) and friends talking in mixed company (t = 3.0). But teacher-student conversations are *also* related to TV news programs (t = 3.2) and radio news (t = 2.6) and to a lesser extent (with t's around 1.8) most of the other items. Lack of sharp differences in the effects of **rowA** is a result of the multivariate nature of the data: responses to the items in Question 4 are highly interdependent and variables correlated with one tend to be correlated with them all (as is apparent in the analysis of **preMSC** in the previous section). Without much more data it is not possible to sort out the specific effects. Another problem arises in the analysis of the effects of use of Soviet audio and video tapes. These activities should prepare students to understand many speech types, but they are not related to *any* of them. A glance at Figure 5 reveals why no relationships would be detected, even if they existed: so few students had used the media that there is not enough variance in **rowG** and **rowH** to explain anything statistically. Lectures read by the teacher (**rowE**) turns out *not* to be related to understanding classroom lectures (t = .1), or to anything else: this activity may be simply ineffective. In sum, while some of the hypothesized relationships may hold, there is not sufficient data to establish very specific hypotheses. On the other hand, the data *do* support analyses involving combinations of items, i.e. **preMSC**, and one can according rely on results such as those presented in the previous subsection.

6. GAINS IN LISTENING COMPREHENSION

6.1. Measures

We turn now to the question of changes in listening comprehension consequent to the study abroad program. Table 16 shows the distribution of the amount of change on each of the seventeen speech types. It is apparent that in the students' own view improvement is the norm (i.e., no change is very rare, and no one feels he or she got

worse) on all of these criteria, a fact which accords with common perceptions of the benefits of language study abroad. Thus changes are gains. The two panels at the top of Figure 6 reinforce this conclusion by comparing the distributions of the scaled self-assessments and the ETS Test before and after the program. The whole distribution is moved up, i.e. increased. (For the ETS Test the figure understates the true change since the post test is known to be somewhat harder than the first.)

To examine the correlates of change — the final major question of the paper — it is necessary to develop overall measures of change at the individual level. There are two ways to approach the calculation of a single measure of change for each student, depending on how one construes the psychology of responding to the questionnaire:

- scale pre and post and take the difference, referred to as ΔMSC.
- calculate the difference between pre and post on each item and scale the differences, referred to as chngMSC

The first approach assumes that the student is making a straightforward (veridical) assessment of each pre and post item, with no elaboration to take account of change, and that the pre and post scores derive from the same scale. The estimated item parameters on the pre and post scales are not exactly the same since both are based on moderate sized samples. Accordingly, the pre and post student ability scores are not exactly equivalent. Nevertheless the differences are so small as to be negligible, so ΔMSC is a suitable measure for our purposes. The second approach assumes that the student is essentially reporting the change, which is the obvious intent of the questionnaire, and may have adjusted ("fudged") the pre or post ratings to make sure that his or her perceived changes are reflected in the answers. Change on a given criterion (e.g. listening to friends) could be small either because it is hard to change or because the item is so easy that there is not much room for change, and accordingly the items do not necessarily affect chngMSC in the same way. Consequently, this measure is somewhat less attractive than ΔMSC. Be that as it may, the two measures are so highly correlated ($r = .958$) that it hardly matters which one is used. Both measures were analyzed, with essentially equivalent results. As with pre-program levels, change

on the ETS Listening test (referred to as ΔETSL) is also analyzed, as a point of comparison and a matter of general interest.[15]

Table 17 gives the MSCALE analysis of "difficulty of change" for the seventeen comprehension items. (It is based on a double-sorted table of reported *changes* similar to Table 9.) All items are well fit by the procedure, but 4 of the 82 respondents seem to have rather deviant response patterns, probably resulting from the different ways in which items enter the scale. Change scores on all criteria are reasonably normally distributed, as shown by the normal plots in Figure 6, although ΔETSL has an outlier (observation number 6) which did not affect the analysis. Both MSCALE change measures are virtually uncorrelated with change in the ETS Test (r = .017 for both, based on n = 71 cases; see also the scatterplot in the bottom right panel of Figure 6), which is further evidence that the perceived abilities and the ETS Test are measuring different things. With regard to basic relationships it should also be noted that pre and post scores, on the MSCALEd self-assessments and on the ETS Test — or equivalently pre-scores and changes — are very highly correlated (r = .709 for pre and postMSC and r = .585 for pre and postETSL). Thus, not surprisingly, students who start the program above average are still above at the end of the program. The correlations between initial levels and changes are highly negative (r = -.797 for MSC and r = -.450 for ETS), i.e. the lower the initial competence the greater the gain. This too is not surprising since the pre measures are negative components of change by definition. The high correlations require that the pre levels be included in all of the regression models in our guided stepwise analysis.

6.2. Regression analysis of changes

ΔMSC and chngMSC.

Analysis of factors affecting change in self-assessed competencies, measured in either of the two ways described above, produced essentially the same negative results, a finding which is, nevertheless, interesting in itself. Except for the initial self-assessed level (preMSC), which is highly significant (with t statistics around 11 and 10 for the two criteria, respectively), all other variables, including all row variables and the baseline

---

[15]By contrast, with the ETS Test no change and losses do occur, even when the fact that postETSL is a harder test than preETSL is taken into account. Lack of parallelism in the test produces a downward bias in the change measure, which does not, however, affect the regression analysis because it is absorbed in the constant term, leaving the coefficients of interest unaffected.

variables *Years and *prevImm, are nonsignificant. (For chngMSC, row E (t = 1.71) and *Years (t = -1.67, *n.b.*) seem to be significant, but an F test shows that they are jointly nonsignificant and neither has a significant t statistic without the other.) Moreover, none of the background or preprogram language measures — Gender, Age, nonSlav, preOPI, and preETSR (reading) — is in the least bit significant.[16] Since there is substantial variation in the amount people change (starting at any given initial level), the factors affecting change must have to do with the specific experiences students have abroad. Any educational factors (years, previous programs, specific preprogram learning experiences, other language skills in Russian, experience in learning other languages) operate through the initial level (if at all). They do not have an independent effect on change, as they would, for example, if they prepared students in some specific way to take advantage of the experiences they have abroad. We return to the implications of this result in the concluding section.

ΔETSL.

In contrast to the self-assessments several factors affect change on the ETS Listening Test. (Recall that ΔETSL and ΔMSC are essentially uncorrelated and hence measure different things.) In the baseline (Table 18.a) *Years and *prevImm are nonsignificant, but rowA (teacher-student conversations), rowC (listening to other students), which has a negative effect, and rowF (taped texts for learners) survive step two of the guided procedure to warrant consideration in step three. The F statistic for these three variables jointly against the baseline is significant below the .05 level. Dropping the nonsignificant variables *Years and *prevImm to improve statistical power did not change any other effects, so the model in Table 18.b was explored. Testing rowA, and then rowC, and rowA and rowC jointly, shows that the learning practices effect is due primarily to rowF (the F statistic for rowA and rowC jointly is 2.11, n.s.), leaving preETSL and rowF (taped texts for learners) as a "good" model (Table 18.c). Exposure to tapes for language learners, which is related to the format of the ETS Test, possibly affected initial levels on the ETS Test, and its effect here may either reflect this fact or a similar artifact for the ETS postprogram test, or it may represent a genuine sensitization

---

[16] As noted in Section 2, students responding to the questionnaire are not all in the same ACTR programs; 14 of the 82 are either in a ten month program or have stayed on from the fall for a second four month program, which is in effect a ten month program. When the analysis is restricted to students who are clearly in a four month program only, the same results obtain, with the exception that for ΔMSC a previous immersion has a positive effect (t = 1.94) and women do somewhat better than men (t = -1.7 with initial level and previous immersion controlled). But the basic conclusions remain.

of students to in-country factors producing changes. Should the ETS Test continue to be used, the matter would merit further study.

Again in contrast to self-assessed changes, and in agreement with the larger ACTR database, **Gender** and **nonSlav** have relationships to ΔETSL (men change more than women, and the more languages known the better, all other things being equal). Furthermore **preETSR** (reading) has a highly significant effect, but **preOPI** does not, with **preETSL** and **rowF** controlled. These results are shown in panels a and b of Table 19, which constitutes a "good" model for ΔETSL in this sample. Confining the analysis to four month students only leads to the same conclusions, presented in panel c of Table 19. Interpretation of the **Gender, nonSlav** and **preETSR** effects follows along the lines of the working paper on the ACTR data (Ginsberg, 1992).

7. DISCUSSION

The results reported in Section 3 indicate that, current trends in the literature on pedagogy to the contrary, specific listening activities are not common in college Russian courses. Not surprising, then, at the beginning of their study abroad program, students have little confidence in their ability to comprehend what they hear in a wide range of situations (Section 4.1). Moreover, as shown in Section 5, there are few relationships between the specific activities that do exist and either the students own perceptions of their listening competence or the more "objective" ETS Listening Test, nor do college courses seem to prepare students to take advantage of the study abroad experience. The strong relationship of teacher-student conversations (whatever that may involve) and self-evaluations is an important result, simply because it implies that *something* matters. As for study abroad itself, both according to the students themselves and on the basis of the ETS Test (Ginsberg, 1992), substantial gains are made. This too is not surprising, since the ACTR program requires students to attend Russian classes in their Institutes for five hours a day for a whole semester, along with the large amount of listening they do in a variety of situations outside of class. Clearly there are many positive factors at work which could be built into both domestic and study abroad programs.

On the one hand our negative results concerning the effects of domestic training may simply indicate a statistical constraint, namely that there is as yet not enough variation in college listening activities to detect significant relationships. Obviously more extensive training in listening, and more research based on it, are required. On the

other hand, the lack of relationships may indicate that what instruction there is is not as effective as one would desire. For example, authentic audios and videos presented without carefully developed accompanying materials to make them comprehensible are likely to be wasted. Students who attempt to view and understand movies or public affairs programs without the necessary background information, knowledge of the schema, and a familiarity with the specifics of the lexicon to be encountered are sure to fail. The counterproductive effects of taped pattern drills, which seem to rob students of the need to develop global comprehension skills, argues for much closer attention to listening *per se*. By contrast, the positive effect of teacher-student conversations indicates a place for "inauthentic" materials, although when one considers supporting the use of "authentic" materials, the whole distinction becomes moot. Thus research, in both domestic and study abroad environments, directed at what students actually do with authentic materials, what practices seem to be effective, and how those practices can be supported, would seem to be fruitful. With regard specifically to study abroad, steps in that direction will be reported in subsequent papers in this series.

Finally, our results bear on a number of issues concerned with testing and measurement. The nearly canonical results of our scaling procedures clearly indicate that students are consistent in their evaluations of their own listening skills over a wide range of situations and that a reliable scale can be constructed based on their responses. The face validity of the scale, in the sense that measured item difficulty corresponds to our intuitive notions, is persuasive, but it would be valuable to be able to relate the scale to more objective measures. The strong correlation of the scale to the OPI, and the pattern of relationships between the OPI and the individual items, discussed in Section 4.4, is evidence in the right direction, as is the positive correlation with the ETS Listening Test. That the ETS Test is not more highly correlated with the scale than it is says as much about the ETS Test as about self-evaluations. Close comparison of the items included in the two measures, in conjunction with analysis of the way the ETS Test is administered, suggests that the assumptions about comprehension underlying the ACTFL scale may require reexamination, to take account explicitly of redundancy, acoustics, interactivity, context, stress, etc., and/or that more flexible and varied testing procedures may be required, to obtain useful assessments of listening comprehension in authentic situations. This too will be the subject of future research.

REFERENCES

Andrich, D. (1978a). A rating formulation for ordered response categories. *Psychometrika*, 43, 561-573.

Andrich, D. (1978b). Scaling attitude items constructed and scored in the Likert tradition. *Educational and Psychological Measurement*, 38, 665-680.

American Council for the Teaching of Foreign Languages (1988). *ACTFL Proficiency Guidelines: Russian*. Yonkers, NY.

Bacon, S. (1989) Listening for real in the foreign language classroom. *Foreign Language Annals*, 22, 543-551.

Baker, R. L. (1986) *Russian for Everybody: Version for Americans*. Moscow: Russkiy Yazyk.

Coakley, C. G. and Wolvin, A.D. (1986). Listening in the native language. In Barbara H. Wing (Ed.), *Listening, Reading, Writing: Analysis and Application*. Middlebury, VT: Northeast Conference.

Child, J. (1987). Language proficiency levels and the typology of texts. In H. Byrnes and M. Canale (Eds.), *Defining and Developing Proficiency: Guidelines, Implementations, and Concepts*. Lincolnwood, IL: National Textbook Company.

Dunkel, P. (1986). Developing listening fluency in L2: Theoretical principles and pedagogical considerations. *Modern Language Journal*, 70, 99-106.

Feyton, C. (1991). The power of listening ability: An overlooked dimension in language acquisition. *Modern Language Journal*, 75, 173-180.

Ginsberg, R. B. (1992) *Language Gains during Study Abroad: An Analysis of the ACTR Data*. National Foreign Language Center Working Paper, Washington, DC.

Heron, C. A. and Seay, I. (1991). The effect of authentic oral texts on student listening comprehension in the foreign language classroom. *Foreign Language Annals*, 24, 487-495.

Joiner, E. (1984). Listening from the inside out. *Foreign Language Annals*, 17, 335-338.

Laviosa, F. (1991). An investigation of the listening strategies of advanced learners of Italian as a second language. Paper presented at the Conference of Bridging Theory and Practice in the Foreign Language Classroom, Loyola College of Baltimore, Maryland, October 18-20.

Long, D. R. (1990). What you don't know can't help you: An exploratory study of background knowledge and second language listening comprehension. *Studies in Second Language Acquisition*, 12, 65-80.

Lowe, P., Jr. (1985). The ILR proficiency scale as a synthesizing research principle: The view from the mountain. In C. J. James (Ed.), *Foreign Language Proficiency in the Classroom and Beyond*. Lincolnwood, IL: National Textbook Company.

Martin, C. L., Robin, J., and Jarvis, D. K. (1991) *The Russian Desk: A Listening and Conversation Course*. Columbus, OH: Slavica Publishers.

Meyer, R. (1984). Listen my children and you shall hear. *Modern Language Journal*, 73, 333-344.

Modern Language Materials Development Center (1961). *ALM Russian , Level One*. New York: Harcourt, Brace, and World.

Nord, J. R. (1981). Three steps to listening fluency: A beginning. In H. Winitz (Ed.), *The Comprehension Approach to Foreign Language Instruction*. Rowley, MA: Newbury House.

Phillips, J. K. (1990). An analysis of text in video newscasts: A tool for schemata building in listening. Georgetown University Roundtable on Languages and Linguistics, Washington, DC.

Pica, T., Young, R., and Doughty, C. (1987). The impact of interaction on comprehension. *TESOL Quarterly*, 21, 737-758.

Richards, J. (1983). Listening comprehension: Approach, design, procedures. *TESOL Quarterly*, 17, 219-140.

Robin, R. M. (1987) *Exemplary Russian Listening Comprehension Materials*. Distributed by American Association for the Teaching of Foreign Languages, Yonkers, NY.

Robin, R. M. and Lekic, M. (1990). *Russian Listening Comprehension, Part I*. Columbus, OH: Ohio State University Slavic Language Materials.

Wright, B.D. and Stone, M.H. (1979). *Best Test Design*. Chicago: Mesa Press.

Wright, B.D. and Masters, G.N. (1982). *Rating Scale Analysis: Rasch Measurement*. Chicago: Mesa Press.

Appendix 1

Listening Comprehension Questionnaire

1. Your current school or alma mater? _____

2. Years of classroom study of Russian? _____

3. Here is a list of ways of exposing one to spoken Russian. For each year of Russian in your home school, as well as other programs (e.g. Middlebury summer or previous ACTR programs), indicate which activities were used and how intensively according to the following scale: 0-Never; 1-Rarely: once or twice a semester; 2-Sometimes: once or twice a month; 3-Regularly: once a week; 4-Quite Often: part of nearly every class/home session; N/A or blank-Not applicable.

	YEAR OF RUSSIAN STUDY					OTHER PROGRAMS Give level(s) & program name(s)		
	1st	2nd	3rd	4th	5th	Intens. summer progs	USSR(1)	USSR(2)
a. Teacher-student conversations								
b. Dictations								
c. Listening to other students								
d. Language lab "grammar" drills								
e. Lectures read by the teacher								
f. Taped texts just for language learners								
g. Audio tapes from Soviet media								
h. Video tapes from Soviet media								
i. Other listening practice (what?)								
j. Other listening practice (what?)								

4. Rate these types of speech based on ease of comprehension at the beginning and end of your ACTR program. Use the following scale: 4-Easy: understand virtually all; 3-Not too hard: get most of it, but miss some details; 2-Stressful, but can get the essentials; 1-Hard to get the gist; 0-Virtually impossible; N/A-No exposure or not applicable.

	Starting	Ending	Speech type
a.	____	____	Teachers talking to you
b.	____	____	Friends talking to you (Russian friends only)
c.	____	____	Friends talking to you (mixed company: Russians, Americans)
d.	____	____	Strangers talking to you (e.g. clerk to customer)
e.	____	____	Russians talking among themselves in your presence
f.	____	____	Telephone conversations with friends
g.	____	____	Telephone conversations with strangers
h.	____	____	Public address announcements (as in airports, train stations)
i.	____	____	Classroom lectures
j.	____	____	TV news programs such as "Vremya"
k.	____	____	Public affairs TV such as "Vzglyad," "5-oe koleso"
l.	____	____	TV movies without subtitles
m.	____	____	Movies in movie theaters
n.	____	____	Live plays
o.	____	____	Radio news reports
p.	____	____	Radio discussion programs
q.	____	____	Street meetings and demonstrations

Appendix 2
Authenticity of Classroom Based Practices

a. Teacher-student conversations. Students probably get most of their classroom language input from teacher talk, directed either to the class as a whole or to individual students. Teacher talk is almost always inauthentic. It is characterized by frequent repetition, even when the student indicates no need for it, deliberateness, and questions with no information gap. Further, teacher-student conversations rarely involve utterances beyond a sentence in length on the part of either participant, thus offering little opportunity for practice in comprehension of paragraphed, connected speech.

b. Dictations are viewed as helping to improve listening comprehension and spelling. If the language of a dictation is authentic (or passably semi-authentic), then the slowed rate of speech and repetition renders it similar to a commonly used kind of cloze exercise in which students try to reproduce missing portions of a text by listening to them on tape a number of times.

c. Listening to other students. What listening input does not come from teacher-student conversations is most likely to come from other students, whose speech is even slower and more inauthentic than that of the teacher.

d. Language lab "grammar" drills at first glance have little to recommend from the standpoint of listening proficiency *per se*. Nevertheless, when students do the drills without a tape script, they are forced to rely on short-term memory to remember the pattern and its components, a listening skill vital to catching details such as times, dates, names, and places in a news report or weather forecast.

e. Lectures read by the teacher are often the students' first experience with connected paragraphed speech. Inauthenticity is due to the fact that such lectures are tailored to a foreign audience.

f. Taped texts just for language learners. Many programs include stories and other short texts, such as those recorded for the *Temp* series (Pushkin Institute, 1978). Some teachers use these specifically for listening comprehension, some for dictation, and still others as an occasional diversion.

g and h. Audio and video tapes from Soviet media. Spurred by the availability of satellite reception, anthologies on videotape (Vzglyad, Ogonek, as well as material collected specifically for distribution to classrooms accompanied by special exercises*), radio broadcasts and television soundtracks have made some inroads into Russian-language classrooms and language labs over the past few years. The language is of course authentic, although the conditions under which students listen may not be.

---

* The producers of "Vzglyad" produced a series called "Vzglyad-rok," marketed in Russia. Since 1989, the weekly magazine "Ogonek" has marketed an extensive series of expose-style reports (some of which had appeared earlier on Soviet television). That series is available in this country. Compilations of Soviet video and audio along with language exercises or texts have come from R. Robin (1987), R. Robin and Lekic (1990), Martin and J. Robin (1991), and the Satellite Communications for Learning Project, which has used SCOLA since 1991).

Appendix 3
The Rating Scale Model

In the briefest terms, the Rating Scale model is a generalization of Rasch's logit model for dichotomous choices, as described in Wright and Stone (1978). For dichotomous items the probability that individual k, with ability a_k, gets item i, with difficulty d_i, correct is $\exp[a_k - d_i]/(1+\exp[a_k - d_i])$. In the Rating Scale model the probability that individual k scores better than category j *given that he is at least in category j* is $\exp[a_k - d_i - b_j]/(1+\exp[a_k - d_i - b_j])$, where the "step" parameters b_j ($j = 0,1, \ldots$), are the same for all items. The Rating Scale model is a special case of the "partial credit" model, in which the step parameters differ for each item. We fit the Partial Credit model (Wright and Master, 1982) using the program MSTEPS, but the Rating Scale model, which is appropriate for our questionnaire, fits the data just as well with many fewer parameters. Given estimates of the parameters a_k, d_i, and b_j, other key probabilities may be derived. For example Figures 1 and 2 give the probabilities of each response and the most probable response to each item. To use Figure 1, for individual k and item i, calculate $a_k - d_i$ and read up to get the probabilities of each category. At the bottom of the Figure the distribution of the individual abilities is shown (with a "2" indicating 2 respondents at that point). In Figure 2, to get the most probable response (which is also the highest curve in Figure 1) read across from the item and up from the ability: the closest curve to the left gives the most probable response. Figure 2 could, for example, be used to calculate predicted responses or to impute missing data.

Tables 10 and 11 show the estimated item difficulties and associated statistics for pre- and post-program ratings, respectively. For convenience, in each table the estimates are shown twice, first with the items in the order in which they appear on the questionnaire, and then with the items ordered hardest to easiest. The difficulty estimates plus "step 4" are in the column labeled "MEASURE." The absolute numbers are on a logit scale which can only be interpreted in the context of the model, but the differences between the numbers make clear the relative location and distances between items on the underlying scale. Thus, in Table 10.b the items are well-spread on the difficulty scale; Items q, "demonstrations ...", and c, "friends ...", are at the extremes and are very different in difficulty; Item f, "telephone with friends," is of moderate difficulty; and Items n, "live plays", and j, "TV news", are essentially equivalent in difficulty. "ERROR" is the standard error of the estimated difficulties; items less than, say, 1.5 standard errors apart are not significantly different from a statistical point of view. (Remember we are dealing with estimates based on a sample of 82 students.)

The last three columns of the MSCALE calibration outputs relate to the goodness of fit of the items to the model, an important reason for using a precise psychometric model as against the more familiar scores of Table 9. LASTIT is the discrepancy between observed and expected (predicted by the model) scores; it is small in all cases, with discrepancies of less than 3 percent. OUTFIT and INFIT are two standardized residuals (mean square and information, respectively); values greater than 2 indicate lack of fit for the item, and negative values indicate greater than expected orderliness, which does not present a problem here. Clearly the fit is very good for all items, pre and post.

Appendix 4
A Priori Analysis of Difficulty of Items in Question 4 of the Questionnaire

a. Teachers talking to you. Because teacher talk was so common a form of input, most students would rate their abilities in this category among the greatest.

b. Friends talking to you (Russian friends only). Similar to Item a. Russian-speaking friends certainly tailor their speech for foreigners, albeit perhaps not as carefully as teachers. Perhaps more than teacher talk, talk among friends is interactional, involving real turn-taking, no pretense in listener interest, and true information gaps.

c. Friends talking to you (mixed company: Russians, Americans). Similar to b, above. The presence of other Americans is likely to mitigate some difficulties. Therefore, we hypothesized that students who have trouble in all-Russian company might understand more if a few Americans were present, even if Russian only was spoken.

d. Strangers talking to you (e.g. clerk to customer). Students who had done a great deal of work with role-play situations would be expected to have fewer difficulties in this category. Otherwise, we would expect students to rate this category slightly harder than conversations with Russian friends due to the need for greater semantic accuracy under more stressful conditions, given the state of the service sector.

e. Russians talking among themselves in your presence. These are essentially overheard conversations, and as such, can be expected to be rated very difficult, especially because of the small likelihood of students having had much practice before an in-country experience.

f. Telephone conversations with friends. The interactional nature of these conversations should make them relatively easy. On the other hand, we would expect hostile acoustics to raise the difficulty level. Given the fact that few students have the opportunity in their home institutions to practice this kind of listening comprehension, we could expect a fairly high difficulty rating to begin with. But we would also expect students who had had many telephone conversations to improve (or believe that they had improved) their listening comprehension dramatically.

g. Telephone conversations with strangers. This category represents an overlay of Items d and f above.

h. Public address announcements (as in airports, train stations). On the ACTFL scale these are usually pegged at Novice High to Intermediate Mid. However, we surmise that the lack of attention given to acoustics is critical. We therefore assumed that students would rate this a difficult category, at least initially.

i. Classroom lectures. Despite the non-interactive nature of this category, we expected students to rate this as a relatively easy kind of comprehension because many will have had practice, albeit with inauthentic lectures, at their home institutions. Furthermore, most (but by no means all) Russian lecturers in-country make some attempts to modify their lecture style for the foreign audience.

j. TV news programs such as "Vremya." Most of the content of "Vremya" could be pegged at ACTFL Advanced, Advanced Plus, and Superior (levels ranging from straightforward factual narration on familiar topics to speech on abstract topics involving supported opinions, hypothetical suppositions, and requiring inferencing on the part of the listener. Those students who had been exposed to such broadcasts systematically might be expect to rate these as something less than impossible. The presence of the accompanying video could be expected to mitigate the difficulty level, especially compared to radio (Item o, below). However, unlike certain categories (movies in movie theaters, public address announcements), the acoustic environment is usually friendly.

k. Public affairs TV such as "Vzglyad," "5-oe koleso." Nearly all these broadcasts involve heavy cultural referencing, supported opinion, and other components characteristic of the ACTFL Superior level.

l. TV movies without subtitles. These are essentially overheard conversations, but in acoustically friendly environments. Overall, we would expect students to find these difficult. The difficulty of movies is made even greater by the presence of heavy cultural referencing, non-standard speech, and schematic and stylistic complexity. (Predictability is a chief component of comprehension, but it is the filmmaker's nemesis.) On the other hand, students who limit their movie-watching to films, with whose story lines they are familiar might have an easier time. Similarly a diet of dubbed American films, with their lack of foreign cultural referencing, and their easier soundtrack, simpler both stylistically and acoustically, might lead students to rate this category as being not so difficult.

m. Movies in movie theaters. We can expect to find all the difficulties of Item l, above, but with the added problem of bad acoustics.

n. Live plays. This category is similar to Item m, above, although in small theaters the acoustics may be better.

o. Radio news reports. These are similar to Item j, above, but without the visual component.

p. Radio discussion programs. These are similar to Item k, above, but with no visuals.

q. Street meetings and demonstrations. Street demonstrations take place in acoustically unfriendly environments. However, students who show up at demonstrations not accidentally, are likely to have a background familiarity with what is occurring. Therefore, they may be able to construct a useful schema to aid in comprehension.

Table 1
Number of Students who Never Engaged in Particular Listening Activities during their Stay in Russia

Item	Both NA	Item	Both NA
a. Teachers talking to you	0	j. TV news programs	5
b. Friends talking to you (Russian friends only)	0	k. Public affairs TV	28
c. Friends talking to you (mixed company; Russian, American)	3	l. TV movies without subtitles	20
d. Strangers talking to you (e.g. clerk to customer)	0	m. Movies in movie theater	8
e. Russians talking among themselves in your presence	0	n. Live plays	14
f. Telephone conversations with friends	4	o. Radio news reports	13
g. Telephone conversations with strangers	8	p. Radio discussion program	32
h. Public address announcements	0	q. Street meetings and demonstrations	13
i. Classroom lectures	0		

Table 2
Exposure to Spoken Russian in First Year
(number/*percent*)

	Frequency				
Activity	Never	Rarely	S'times	Reg'ly	Often
Teacher-student conversations	14 *18*	19 *24*	17 *22*	11 *14*	18 *23*
Dictations	28 *35*	17 *22*	11 *14*	16 *20*	7 *9*
Listening to other students	15 *19*	7 *9*	9 *11*	14 *18*	34 *43*
Language lab "grammar" drills	17 *22*	5 *6*	11 *14*	25 *32*	21 *27*
Lectures read by teacher	37 *47*	12 *15*	13 *16*	12 *15*	5 *6*
Taped texts just for language learners	35 *44*	8 *10*	16 *20*	14 *18*	6 *8*
Audio tapes from Soviet media	70 *89*	2 *3*	7 *9*	0 *0*	0 *0*
Video tapes from Soviet media	53 *67*	11 *14*	13 *16*	2 *3*	0 *0*
Other (1)	63 *80*	5 *6*	2 *3*	7 *9*	2 *3*
Other (2)	76 *96*	2 *3*	0 *0*	1 *1*	0 *0*

n = 79 programs described/students reporting

Table 3
Exposure to Spoken Russian in Second Year
(number/*percent*)

	Frequency				
Activity	Never	Rarely	S'times	Reg'ly	Often
Teacher-student conversations	7 *9*	13 *17*	16 *21*	14 *18*	26 *34*
Dictations	23 *30*	22 *29*	15 *20*	10 *13*	6 *8*
Listening to other students	11 *14*	6 *8*	10 *13*	12 *16*	37 *49*
Language lab "grammar" drills	19 *25*	6 *8*	18 *24*	23 *30*	10 *13*
Lectures read by teacher	27 *36*	14 *18*	13 *17*	12 *16*	10 *13*
Taped texts just for language learners	40 *53*	6 *8*	14 *18*	10 *13*	6 *8*
Audio tapes from Soviet media	64 *84*	2 *3*	8 *11*	2 *3*	0 *0*
Video tapes from Soviet media	47 *62*	8 *11*	15 *20*	6 *8*	0 *0*
Other (1)	59 *78*	4 *5*	2 *3*	5 *7*	6 *8*
Other (2)	69 *91*	1 *1*	2 *3*	2 *3*	2 *3*

n = 76 programs described/students reporting

Table 4
Exposure to Spoken Russian in Third Year and Above
(number/*percent*)

	Frequency				
Activity	Never	Rarely	S'times	Reg'ly	Often
Teacher-student conversations	4 *6*	4 *6*	15 *23*	16 *25*	26 *40*
Dictations	28 *43*	15 *23*	14 *22*	6 *9*	2 *3*
Listening to other students	3 *5*	2 *3*	8 *12*	14 *22*	38 *58*
Language lab "grammar" drills	33 *51*	10 *15*	8 *12*	9 *14*	5 *8*
Lectures read by teacher	15 *23*	9 *14*	12 *18*	14 *22*	15 *23*
Taped texts just for language learners	45 *69*	5 *8*	4 *6*	8 *12*	3 *5*
Audio tapes from Soviet media	43 *66*	11 *17*	8 *12*	3 *5*	0 *0*
Video tapes from Soviet media	33 *51*	9 *14*	13 *20*	9 *14*	1 *2*
Other (1)	45 *69*	5 *8*	2 *3*	9 *14*	4 *6*
Other (2)	57 *88*	0 *0*	1 *2*	4 *6*	3 *5*

n = 65 program/years described

Table 5
Exposure to Spoken Russian in Summer Programs
(number/*percent*)

Activity	Frequency				
	Never	Rarely	S'times	Reg'ly	Often
Teacher-student conversations	1 *3*	0 *0*	5 *16*	2 *6*	24 *75*
Dictations	15 *47*	5 *16*	6 *19*	6 *19*	0 *0*
Listening to other students	1 *3*	0 *0*	0 *0*	2 *6*	29 *91*
Language lab "grammar" drills	10 *31*	3 *9*	3 *9*	9 *28*	7 *22*
Lectures read by teacher	10 *31*	3 *9*	2 *6*	7 *22*	10 *31*
Taped texts just for language learners	20 *63*	1 *3*	2 *6*	6 *19*	3 *9*
Audio tapes from Soviet media	27 *84*	0 *0*	0 *0*	2 *6*	3 *9*
Video tapes from Soviet media	15 *47*	2 *6*	4 *13*	7 *22*	4 *13*
Other (1)	21 *66*	2 *6*	2 *6*	4 *13*	3 *9*
Other (2)	30 *94*	0 *0*	1 *3*	0 *0*	1 *3*

n = 32 programs described

Table 6
Formal Learning Activities Mentioned in the "Other" Categories
(number of students mentioning it in parentheses)

Year 1	Year 2
Russian table (4) Russian house Russian friend Video tapes for language learners (2) Songs, etc Oral exams in language lab Movies (2) Stories	Russian table (2) Russian house (4) Work Slavic Dept. Poetry Conversations w/ Soviets (3) Video tapes for language learners (2) Russian friend Oral exams in language lab Listening to native speakers Movies (2)
Years 3 , 4, and 5	Summer Programs
Russian house (3) Work Slavic Dept. Russian table (2) Poetry Films, plays (8) Conversations w/ Soviets (4) Conversation hour (2) Conversation w/ TA Students giving reports Skits, singing (3) Listening to native speakers Guest speakers (2)	Conversations w/ Soviets Conversation w/ TA Teacher talking Living (Russian only rule) Movies Play Russian table

Table 7
Frequencies of PreProgram Self-Assessments
(number/*percent*)

Item	0	1	2	3	4	n/a
a. Teachers talking to you	4 *5*	11 *13*	30 *37*	29 *35*	7 *9*	1 *1*
b. Friends talking to you (Russian friends only)	6 *7*	16 *20*	40 *49*	17 *21*	2 *2*	1 *1*
c. Friends talking to you (mixed company; Russian, American)	1 *1*	10 *12*	33 *40*	29 *35*	5 *6*	4 *5*
d. Strangers talking to you (e.g. clerk to customer)	10 *12*	24 *29*	36 *44*	12 *15*		
e. Russians talking among themselves in your presence	33 *40*	26 *32*	19 *23*	4 *5*		
f. Telephone conversations with friends	11 *13*	29 *35*	28 *34*	10 *12*		4 *5*
g. Telephone conversations with strangers	16 *20*	33 *40*	21 *26*	4 *5*		8 *10*
h. Public address announcements	32 *39*	30 *37*	14 *17*	5 *6*	1 *1*	
i. Classroom lectures	4 *5*	14 *17*	37 *45*	24 *29*	3 *4*	
j. TV news programs	21 *26*	29 *35*	23 *28*	3 *4*		6 *7*
k. Public affairs TV	17 *21*	17 *21*	15 *18*	2 *2*		31 *38*
l. TV movies without subtitles	11 *13*	27 *33*	19 *23*	4 *5*		21 *26*
m. Movies in movie theater	13 *16*	28 *34*	27 *33*	3 *6*		9 *11*
n. Live plays	14 *17*	27 *33*	20 *24*	1 *1*		20 *24*
o. Radio news reports	16 *20*	26 *32*	23 *28*		1 *1*	16 *20*
p. Radio discussion program	15 *18*	18 *22*	13 *16*	2 *2*		34 *42*
q. Street meetings and demonstrations	24 *29*	21 *26*	18 *22*			19 *23*

Response Codes: **0** ("virtually impossible"), **1** ("hard to get gist"), **2** ("stressful"), **3** ("get all but details"), **4** ("easy: understand all").

Table 8
Frequencies of PostProgram Self-Assessments
(number/*percent*)

Item	0	1	2	3	4	n/a
a. Teachers talking to you			3	27	52	
			4	*33*	*63*	
b. Friends talking to you (Russian friends only)			9	43	30	
			11	*52*	*37*	
c. Friends talking to you (mixed company; Russian, American)			2	33	44	3
			2	*40*	*54*	*4*
d. Strangers talking to you (e.g. clerk to customer)			6	62	14	
			7	*76*	17	
e. Russians talking among themselves in your presence		8	33	37	4	
		10	*40*	*45*	5	
f. Telephone conversations with friends		1	7	48	22	4
		1	*9*	*59*	*27*	*5*
g. Telephone conversations with strangers		4	25	40	5	8
		5	*31*	*49*	*6*	*10*
h. Public address announcements	2	10	32	31	7	
	2	*12*	*39*	*38*	*9*	
i. Classroom lectures		1	6	36	39	
		1	*7*	*44*	*48*	
j. TV news programs		6	28	40	3	5
		7	*34*	*49*	*4*	*6*
k. Public affairs TV		6	24	21	3	28
		7	*29*	*26*	*4*	*34*
l. TV movies without subtitles		3	23	33	2	21
		4	*28*	*40*	*2*	*26*
m. Movies in movie theater		4	23	40	4	11
		5	*28*	*49*	*5*	*13*
n. Live plays	1	5	34	25	1	16
	1	*6*	*42*	*31*	*1*	*20*
o. Radio news reports		6	25	34	2	15
		7	*31*	*42*	*2*	*18*
p. Radio discussion program		6	22	18	4	32
		7	*27*	*22*	*5*	*39*
q. Street meetings and demonstrations	3	10	38	17	1	13
	4	*12*	*46*	*21*	*1*	*16*

Response Codes: **0** ("virtually impossible"), **1** ("hard to get gist"), **2** ("stressful"), **3** ("get all but details"), **4** ("easy: understand all").

Table 9
Pre-Program Self-Assessments Sorted by Person and Item Scores

ID	c1	a1	i1	b1	d1	f1	m1	l1	g1	o1	n1	j1	p1	k1	h1	e1	q1	Average	Rank
15	4	4	4	3	3	3	3	3	3	4	3	3	9	3	4	3	1	3.19	1
34	4	4	4	4	3	3	2	9	3	2	9	2	2	2	2	2	2	2.73	2
12	3	3	3	3	3	2	3	3	2	2	2	3	3	3	1	2	9	2.56	3
42	3	4	3	3	3	3	3	2	2	2	2	2	2	1	2	3	2	2.47	4
13	9	3	3	3	3	3	2	2	3	2	2	2	2	2	3	2	2	2.44	5
82	4	3	3	3	3	2	2	2	2	2	2	2	2	2	3	2	2	2.41	6
23	3	3	3	3	2	2	3	3	2	2	2	2	3	2	2	2	2	2.41	7
80	3	3	3	3	2	2	3	2	2	2	2	2	9	2	3	2	2	2.38	8
74	3	3	3	2	2	3	9	2	3	9	1	2	9	2	3	2	2	2.36	9
58	4	4	3	4	3	2	2	2	2	2	2	2	2	2	0	3	1	2.35	10
64	4	3	3	3	2	3	9	2	2	1	1	9	9	9	1	3	9	2.33	11
61	3	3	3	3	2	2	2	2	9	2	2	2	9	9	1	2	9	2.23	12
18	3	3	2	3	2	9	2	9	9	2	2	2	2	9	2	2	2	2.23	13
70	3	3	2	3	3	2	2	9	1	2	9	9	9	9	2	1	2	2.17	14
66	2	3	3	3	2	2	2	2	2	1	2	2	9	2	2	2	2	2.13	15
28	3	4	4	2	3	2	2	9	2	1	1	1	1	1	3	2	2	2.13	16
24	2	3	2	3	2	3	2	2	2	2	2	2	2	2	1	2	2	2.12	17
69	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2.00	18
21	3	3	3	3	2	2	2	1	2	1	9	1	2	1	2	2	1	1.94	19
9	9	3	3	2	2	2	2	9	1	1	9	2	2	9	2	1	1	1.85	20
55	3	3	3	2	2	2	2	1	2	2	1	1	1	1	1	2	2	1.82	21
72	3	2	3	3	2	3	1	1	1	2	9	9	9	9	2	2	0	1.77	22
31	3	3	2	2	2	2	9	1	1	1	9	2	9	2	2	2	9	1.77	23
16	2	2	2	2	2	2	1	2	2	2	1	2	2	2	2	2	0	1.76	24
79	3	2	2	3	3	9	2	9	1	9	2	1	9	9	0	0	2	1.75	25
30	3	3	1	3	1	1	2	2	1	2	2	3	1	2	0	1	9	1.75	26
11	3	2	2	2	2	1	2	2	1	2	1	2	2	9	1	2	1	1.75	27
40	2	3	3	2	2	2	1	1	1	2	1	2	2	2	1	1	1	1.71	28
54	3	3	3	2	3	2	2	9	1	1	1	1	1	1	1	1	1	1.69	29
73	2	2	2	2	1	1	9	2	1	2	2	2	9	9	1	1	2	1.64	30
8	3	3	2	2	2	2	2	9	2	9	9	0	1	9	1	0	1	1.62	31
17	3	3	2	2	2	3	2	9	2	1	1	1	1	9	0	1	0	1.60	32
19	3	3	2	2	2	2	1	2	2	1	1	9	9	9	0	1	0	1.57	33
3	9	2	2	1	1	2	2	2	0	2	2	2	9	9	1	1	9	1.54	34
68	2	2	2	2	2	1	2	2	1	9	9	1	9	9	1	1	1	1.54	35
7	3	3	3	2	2	1	1	1	9	1	9	1	1	1	1	0	9	1.50	36
59	3	2	3	2	2	2	1	1	1	0	1	2	9	9	1	1	0	1.47	37
81	2	2	1	2	2	2	1	1	2	9	9	1	9	9	1	1	1	1.46	38
46	2	4	2	2	1	2	1	9	1	9	2	0	9	0	1	0	2	1.43	39
32	3	2	2	2	2	1	2	1	1	1	1	1	1	1	1	1	1	1.41	40
43	2	3	2	2	1	1	2	9	1	9	2	1	9	1	0	0	9	1.38	41
26	2	2	2	2	2	2	9	1	1	0	9	1	0	9	1	2	9	1.38	42
41	2	1	2	2	1	3	9	1	9	1	9	0	9	9	1	1	9	1.36	43
1	2	2	2	2	3	2	1	1	2	1	2	1	0	1	1	0	0	1.35	44
65	2	2	2	2	1	1	1	9	1	9	9	9	9	9	1	0	9	1.30	45
62	2	2	2	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1.29	46
10	2	2	2	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1.29	47
5	2	2	2	2	1	2	2	3	0	2	1	1	1	1	0	0	0	1.29	48
53	3	3	3	2	1	1	1	1	0	1	9	1	1	1	0	0	9	1.27	49
29	2	1	1	2	2	1	1	2	1	0	1	1	1	2	0	1	2	1.24	50
27	2	2	2	1	1	1	1	9	1	1	1	2	9	9	1	0	9	1.23	51
39	3	3	3	2	2	1	9	9	1	0	0	0	1	9	0	1	0	1.21	52
22	2	9	2	2	1	1	1	1	1	1	1	1	9	1	1	1	9	1.21	53
2	3	2	2	1	2	1	1	1	1	0	2	2	0	9	1	0	0	1.19	54
77	3	2	2	2	1	0	1	9	0	9	9	1	9	9	0	1	9	1.18	55
20	1	2	2	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1.18	56
6	2	4	2	2	1	1	1	0	1	1	1	1	0	0	1	0	1	1.12	57
48	3	3	3	2	2	1	1	0	1	1	0	0	0	0	0	1	1	1.12	58
47	2	3	3	1	2	1	1	1	1	1	0	0	1	0	2	0	0	1.12	59
57	2	2	2	2	1	1	9	1	2	1	9	0	9	0	0	0	1	1.07	60
50	2	1	2	2	1	2	2	0	1	1	1	1	9	9	0	0	0	1.07	61
36	2	2	1	1	1	0	1	1	0	1	1	1	1	1	2	0	1	1.00	62
83	2	2	1	1	2	9	0	0	9	0	1	1	9	9	1	1	9	1.00	63
56	2	2	2	2	1	2	0	0	1	2	1	1	0	0	0	0	0	0.94	64
78	2	2	1	2	2	1	1	1	1	0	0	1	0	1	0	1	0	0.94	65
44	2	3	3	2	1	0	0	9	0	0	0	0	0	0	2	0	0	0.81	66
71	1	2	2	1	0	1	0	0	9	9	9	1	9	9	0	0	9	0.73	67
76	1	1	2	0	0	1	1	1	1	9	0	1	9	9	1	0	0	0.71	68
25	3	2	2	1	0	1	0	0	1	0	1	0	0	0	0	0	1	0.71	69
60	2	1	1	1	1	1	1	0	1	1	0	0	1	0	0	0	1	0.71	70
75	9	1	1	9	1	1	1	1	0	9	9	0	9	9	0	1	9	0.70	71
63	2	2	2	1	0	1	0	1	0	0	0	0	0	0	0	0	9	0.56	72
4	1	1	1	1	1	9	0	9	9	9	9	0	9	9	0	0	0	0.50	73
35	2	1	1	1	0	0	1	1	0	0	1	0	0	0	0	0	0	0.47	74
45	1	1	0	0	1	0	2	9	0	9	0	0	9	0	0	0	0	0.36	75
67	1	1	1	0	0	1	0	0	0	9	9	9	9	9	0	0	0	0.33	76
49	1	0	1	1	0	0	0	9	0	0	1	0	0	0	0	0	1	0.31	77
37	1	2	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0.29	78
38	1	1	1	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0.29	79
51	2	0	0	1	0	0	0	9	0	0	0	0	9	0	0	0	0	0.20	80
33	1	0	0	1	0	0	9	1	0	0	0	0	0	0	0	0	0	0.19	81
52	0	0	0	0	1	0	0	9	9	9	0	0	9	9	0	0	0	0.08	82
aver.	2.3	2.3	2.1	1.9	1.6	1.5	1.3	1.3	1.2	1.2	1.1	1.1	1.0	1.0	0.9	0.9	0.9		
n	78	81	82	81	82	78	73	61	74	66	62	76	48	51	82	82	63		
rank	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17		

Table 10
MSCALE Calibration of Pre-Program Self-Assessments
Item+Step 4 Statistics

a. Order on Questionnaire

NUM	NAME	SCORE	SAMPLE	MEASURE	ERROR	LASTIT	OUTFIT	INFIT
1	A1	186	81	1.56	0.19	-2.2	0.48	0.42
2	B1	155	81	2.60	0.19	-1.6	-2.66	-2.70
3	C1	183	78	1.39	0.19	-2.3	-1.63	-1.70
4	D1	132	82	3.37	0.18	-0.7	0.02	-1.19
5	E1	76	82	5.12	0.19	1.7	-0.53	-1.14
6	F1	115	78	3.74	0.19	-0.3	-1.43	-1.40
7	G1	87	74	4.55	0.19	0.8	-2.46	-2.45
8	H1	77	82	5.09	0.19	1.7	1.07	1.80
9	I1	172	82	2.09	0.19	-2.0	-0.78	-0.82
10	J1	84	76	4.64	0.19	1.0	-1.81	-1.85
11	K1	53	51	4.99	0.24	0.9	-1.95	-2.15
12	L1	77	61	4.36	0.21	0.5	-0.77	-1.63
13	M1	97	73	4.08	0.19	0.2	-1.65	-2.77
14	N1	70	62	4.66	0.21	0.8	-1.07	-1.79
15	O1	76	66	4.77	0.20	1.0	-1.78	-1.81
16	P1	50	48	4.95	0.24	0.8	-2.16	-2.52
17	Q1	57	63	5.26	0.22	1.4	0.44	0.38

b. Hardest to Easiest

NUM	NAME	SCORE	SAMPLE	MEASURE	ERROR	LASTIT	OUTFIT	INFIT
17	Q1	57	63	5.26	0.22	1.4	0.44	0.38
5	E1	76	82	5.12	0.19	1.7	-0.53	-1.14
8	H1	77	82	5.09	0.19	1.7	1.07	1.80
11	K1	53	51	4.99	0.24	0.9	-1.95	-2.15
16	P1	50	48	4.95	0.24	0.8	-2.16	-2.52
15	O1	76	66	4.77	0.20	1.0	-1.78	-1.81
14	N1	70	62	4.66	0.21	0.8	-1.07	-1.79
10	J1	84	76	4.64	0.19	1.0	-1.81	-1.85
7	G1	87	74	4.55	0.19	0.8	-2.46	-2.45
12	L1	77	61	4.36	0.21	0.5	-0.77	-1.63
13	M1	97	73	4.08	0.19	0.2	-1.65	-2.77
6	F1	115	78	3.74	0.19	-0.3	-1.43	-1.40
4	D1	132	82	3.37	0.18	-0.7	0.02	-1.19
2	B1	155	81	2.60	0.19	-1.6	-2.66	-2.70
9	I1	172	82	2.09	0.19	-2.0	-0.78	-0.82
1	A1	186	81	1.56	0.19	-2.2	0.48	0.42
3	C1	183	78	1.39	0.19	-2.3	-1.63	-1.70

Table 11
MSCALE Calibration of Post-Program Self-Assessments
Item+Step 4 Statistics

a. Order on Questionnaire

NUM	NAME	SCORE	SAMPLE	MEASURE	ERROR	LASTIT	OUTFIT	INFIT
1	A2	295	82	0.79	0.23	-1.5	-0.43	0.00
2	B2	267	82	2.03	0.21	-1.4	-1.92	-1.82
3	C2	279	79	1.04	0.23	-1.5	-0.69	-0.51
4	D2	254	82	2.55	0.20	-1.2	-2.94	-2.87
5	E2	201	82	4.25	0.17	0.0	-1.42	-1.64
6	F2	247	78	2.33	0.21	-1.2	-0.90	-0.75
7	G2	194	74	3.89	0.19	-0.3	-0.79	-0.89
8	H2	195	82	4.41	0.16	0.2	0.42	0.23
9	I2	277	82	1.62	0.21	-1.5	-0.17	0.34
10	J2	194	77	4.09	0.18	-0.1	-3.24	-2.94
11	K2	129	54	4.43	0.20	0.1	-3.70	-3.88
12	L2	156	61	4.03	0.20	-0.2	-2.74	-2.52
13	M2	186	71	3.99	0.19	-0.2	-2.74	-2.68
14	N2	152	66	4.61	0.18	0.3	0.70	-0.48
15	O2	166	67	4.24	0.19	0.0	-1.04	-1.20
16	P2	120	50	4.46	0.21	0.1	-2.09	-2.55
17	Q2	141	69	5.04	0.17	0.8	-1.85	-2.06

b. Hardest to Easiest

NUM	NAME	SCORE	SAMPLE	MEASURE	ERROR	LASTIT	OUTFIT	INFIT
17	Q2	141	69	5.04	0.17	0.8	-1.85	-2.06
14	N2	152	66	4.61	0.18	0.3	0.70	-0.48
16	P2	120	50	4.46	0.21	0.1	-2.09	-2.55
11	K2	129	54	4.43	0.20	0.1	-3.70	-3.88
8	H2	195	82	4.41	0.16	0.2	0.42	0.23
5	E2	201	82	4.25	0.17	0.0	-1.42	-1.64
15	O2	166	67	4.24	0.19	0.0	-1.04	-1.20
10	J2	194	77	4.09	0.18	-0.1	-3.24	-2.94
12	L2	156	61	4.03	0.20	-0.2	-2.74	-2.52
13	M2	186	71	3.99	0.19	-0.2	-2.74	-2.68
7	G2	194	74	3.89	0.19	-0.3	-0.79	-0.89
4	D2	254	82	2.55	0.20	-1.2	-2.94	-2.87
6	F2	247	78	2.33	0.21	-1.2	-0.90	-0.75
2	B2	267	82	2.03	0.21	-1.4	-1.92	-1.82
9	I2	277	82	1.62	0.21	-1.5	-0.17	0.34
3	C2	279	79	1.04	0.23	-1.5	-0.69	-0.51
1	A2	295	82	0.79	0.23	-1.5	-0.43	0.00

Table 12
Ranking of Item Difficulties by Average Score and MSCALE
Pre- and Post-Program Self-Assessments

	Pre Program			Post Program	
Item	Average Score	MSCALE	Av No previm	Average Score	MSCALE
a. Teachers talking to you	2	2	2	1	1
b. Friends talking to you (Russian friends only)	4	4	4	4	4
c. Friends talking to you (mixed company; Russian, American)	1	1	1	2	2
d. Strangers talking to you (e.g. clerk to customer)	5	5	5	6	6
e. Russians talking among themselves in your presence	16	16	16	12	12
f. Telephone conversations with friends	6	6	6	5	5
g. Telephone conversations with strangers	9	9	9	7	7
h. Public address announcements	15	15	14	15	13
i. Classroom lectures	3	3	3	3	3
j. TV news programs	12	10	13	10	10
k. Public affairs TV	14	14	15	14	14
l. TV movies without subtitles	8	8	8	9	9
m. Movies in movie theater	7	7	7	8	8
n. Live plays	11	11	11	16	16
o. Radio news reports	10	12	10	11	11
p. Radio discussion program	13	13	17	13	15
q. Street meetings and demonstrations	17	17	12	17	17

Table 13
Chi-square values for Pre-Preogram OPI vs. Self-Assessed Competencies*

Item	χ^2	Item	χ^2
a. Teachers talking to you	16.96	j. TV news programs	19.05
b. Friends talking to you (Russian friends only)	9.41	k. Public affairs TV	14.00
c. Friends talking to you (mixed company; Russian, American)	13.44	l. TV movies without subtitles	5.17
d. Strangers talking to you (e.g. clerk to customer)	9.57	m. Movies in movie theater	6.36
e. Russians talking among themselves in your presence	5.96	n. Live plays	2.18
f. Telephone conversations with friends	6.72	o. Radio news reports	5.18
g. Telephone conversations with strangers	9.93	p. Radio discussion program	7.36
h. Public address announcements	8.14	q. Street meetings and demonstrations	3.77
i. Classroom lectures	13.64		

* All tables have 2 degrees of freedom. OPI was grouped 1 and below vs. 1+ and above, and self-assessments were grouped 0,1,2 and above, to reduce the number of small cells.

$\chi^2 < 5.99$ is significant at the .05 level; $\chi^2 < 9.21$ is significant at the .01 level.

Table 14
Regresssion Models for Factors Affecting PreProgram Self-Assessments
(n = 82)

a. **Baseline Model**

Variable	Coefficient	s.e. of Coeff	t-ratio
Constant	-2.65620	0.7690	-3.45
*Years	0.252134	0.2537	0.994
*prevImm	1.45657	0.6042	2.41

R^2= 8.6% R^2(adjusted) = 6.3%
s = 1.778 with 82 - 3 = 79 degrees of freedom

b. **Intermediate Model**

Variable	Coefficient	s.e. of Coeff	t-ratio
Constant	-2.98572	0.7149	-4.18
*Years	-0.030090	0.2499	-0.120
*prevImm	1.30746	0.5925	2.21
rowA	0.217961	0.0506	4.31
rowD	-0.081830	0.0483	-1.69
rowH	-0.092175	0.0635	-1.45

R^2 = 27.8% R^2(adjusted) = 23.0%
s = 1.611 with 82 - 6 = 76 degrees of freedom

c. **"Good" Model**

Variable	Coefficient	s.e. of Coeff	t-ratio
Constant	-3.03565	0.4488	-6.76
*prevImm	1.43142	0.5868	2.44
rowA	0.191618	0.0449	4.27
rowD	-0.096339	0.0470	-2.05

R^2 = 25.8% R^2(adjusted) = 22.9%
s = 1.613 with 82 - 4 = 78 degrees of freedom

Table 15
Regresssion Models for Factors Affecting
PreProgram ETS Listening Proficiency
(n = 74)

a. Baseline Model

Variable	Coefficient	s.e. of Coeff	t-ratio
Constant	14.9696	3.496	4.28
*Years	0.296749	1.140	0.260
*prevImm	5.66273	2.688	2.11

R^2 = 6.1% R^2(adjusted) = 3.4%
s = 7.542 with 74 - 3 = 71 degrees of freedom

b. Intermediate Model

Variable	Coefficient	s.e. of Coeff	t-ratio
Constant	13.7284	3.552	3.87
*Years	-0.004561	1.194	-0.004
*prevImm	4.77927	2.749	1.74
rowA	0.292972	0.2281	1.28
rowB	-0.452081	0.2783	-1.62
rowF	0.421107	0.2422	1.74

R^2 = 13.4% R^2(adjusted) = 7.1%
s = 7.399 with 74 - 6 = 68 degrees of freedom

c. "Good" Model

Variable	Coefficient	s.e. of Coeff	t-ratio
Constant	14.6672	1.199	12.2
*prevImm	4.81062	2.705	1.78
rowF	0.366659	0.2390	1.53

R^2 = 9.0% R^2(adjusted) = 6.4%
s = 7.424 with 74 - 3 = 71 degrees of freedom

Table 16
Frequencies of Self-Assessments Changes in Seventeen Speech Types
(number/*percent*)

Item	0	1	2	3	4	Total
a. Teachers talking to you	9 *11*	42 *52*	27 *33*	1 *1*	2 *2*	81 *100*
b. Friends talking to you (Russian friends only)	6 *7*	45 *56*	26 *32*	3 *4*	1 *1*	81 *100*
c. Friends talking to you (mixed company; Russian, American)	14 *18*	39 *50*	21 *27*	4 *5*		78 *100*
d. Strangers talking to you (e.g. clerk to customer)	7 *9*	38 *46*	29 *35*	6 *7*	2 *2*	82 *100*
e. Russians talking among themselves in your presence	5 *6*	38 *46*	31 *38*	7 *9*	1 *1*	82 100
f. Telephone conversations with friends	4 *5*	31 *40*	29 *37*	13 *17*	1 *1*	78 *100*
g. Telephone conversations with strangers	9 *12*	29 *39*	30 *41*	6 *8*		74 *100*
h. Public address announcements	11 *13*	31 *38*	34 *42*	5 *6*	1 *1*	82 *100*
i. Classroom lectures	8 *10*	47 *57*	23 *28*	4 *5*		82 *100*
j. TV news programs	6 *8*	35 *46*	33 *43*	2 *3*		76 *100*
k. Public affairs TV	4 *8*	24 *47*	23 *45*			51 *100*
l. TV movies without subtitles	7 *8*	28 *47*	24 *40*	1 *2*		60 *100*
m. Movies in movie theater	9 *13*	35 *50*	24 *34*	2 *3*		70 *100*
n. Live plays	7 *8*	34 *57*	18 *30*	1 *2*		60 *100*
o. Radio news reports	6 *9*	33 *52*	20 *31*	5 *8*		64 *100*
p. Radio discussion program	4 *8*	23 *48*	20 *42*	1 *2*		48 *100*
q. Street meetings and demonstrations	13 *21*	28 *44*	21 *33*	1 *2*		63 *100*

Responses: **0** (no change), **1** (one notch, e.g. "hard" to "stressful"), ... , **4** (four notches)

Table 17
MSCALE Calibration of Self-Assessed Changes
Item+Step 4 Statistics

a. Order on Questionnaire

NUM	NAME	SCORE	SAMPLE	MEASURE	ERROR	LASTIT	OUTFIT	INFIT
1	A	107	81	2.37	0.18	4.4	0.69	0.82
2	B	110	81	2.28	0.18	4.2	-2.29	-2.18
3	C	93	78	2.73	0.19	4.7	0.21	0.30
4	D	122	82	1.95	0.18	3.4	0.21	0.26
5	E	125	82	1.86	0.18	3.2	-1.31	-1.26
6	F	132	78	1.43	0.18	1.7	0.51	0.47
7	G	107	74	2.01	0.19	3.3	-0.09	-0.05
8	H	118	82	2.07	0.18	3.7	1.92	1.87
9	I	105	82	2.48	0.18	4.6	-1.46	-1.38
10	J	107	76	2.15	0.19	3.6	-1.06	-1.04
11	K	70	51	2.19	0.23	2.5	-3.01	-3.14
12	L	79	60	2.29	0.21	3.3	-1.80	-1.88
13	M	89	70	2.62	0.20	3.8	-1.44	-1.37
14	N	73	60	2.61	0.22	3.6	-0.24	-0.26
15	O	88	64	2.07	0.21	3.4	1.12	1.07
16	P	66	48	2.22	0.24	2.4	-2.03	-2.05
17	Q	73	63	2.79	0.21	3.9	0.71	0.71

b. Hardest to Easiest

NUM	NAME	SCORE	SAMPLE	MEASURE	ERROR	LASTIT	OUTFIT	INFIT
17	Q	73	63	2.79	0.21	3.9	0.71	0.71
3	C	93	78	2.73	0.19	4.7	0.21	0.30
13	M	89	70	2.62	0.20	3.8	-1.44	-1.37
14	N	73	60	2.61	0.22	3.6	-0.24	-0.26
9	I	105	82	2.48	0.18	4.6	-1.46	-1.38
1	A	107	81	2.37	0.18	4.4	0.69	0.82
12	L	79	60	2.29	0.21	3.3	-1.80	-1.88
2	B	110	81	2.28	0.18	4.2	-2.29	-2.18
16	P	66	48	2.22	0.24	2.4	-2.03	-2.05
11	K	70	51	2.19	0.23	2.5	-3.01	-3.14
10	J	107	76	2.15	0.19	3.6	-1.06	-1.04
8	H	118	82	2.07	0.18	3.7	1.92	1.87
15	O	88	64	2.07	0.21	3.4	1.12	1.07
7	G	107	74	2.01	0.19	3.3	-0.09	-0.05
4	D	122	82	1.95	0.18	3.4	0.21	0.26
5	E	125	82	1.86	0.18	3.2	-1.31	-1.26
6	F	132	78	1.43	0.18	1.7	0.51	0.47

Table 18
Regresssion Models for Effects of Educational Factors
on Changes in ETS Listening Proficiency
(n = 71)

a. Baseline Model

Variable	Coefficient	s.e. of Coeff	t-ratio
Constant	17.6146	3.425	5.14
ETSL1	-0.384379	0.1033	-3.72
*Years	0.624572	1.002	0.623
*prevImm	-2.71491	2.697	-1.01

R^2 = 21.8% R^2(adjusted) = 18.3%
s = 6.468 with 71 - 4 = 67 degrees of freedom

b. Intermediate Model

Variable	Coefficient	s.e. of Coeff	t-ratio
Constant	19.8842	2.360	8.43
ETSL1	-0.471500	0.0974	-4.84
rowC	-0.399345	0.2003	-1.99
rowF	0.573962	0.2042	2.81
rowA	0.317915	0.2081	1.53

R^2 = 31.3% R^2(adjusted) = 27.2%
s = 6.107 with 71 - 5 = 66 degrees of freedom

c. "Good" Model

Variable	Coefficient	s.e. of Coeff	t-ratio
Constant	18.8324	1.767	10.7
ETSL1	-0.468232	0.0976	-4.80
rowF	0.508155	0.2047	2.48

R^2 = 26.9% R^2(adjusted) = 24.7%
s = 6.209 with 71 - 3 = 68 degrees of freedom

Table 19
Regresssion Models for Correlates of Changes in ETS Listening Proficiency

a. **Effects of Gender and nonSlav (n = 70)**

Variable	Coefficient	s.e. of Coeff	t-ratio
Constant	10.8442	3.559	3.05
ETSL1	-0.446981	0.0977	-4.58
rowF	0.540549	0.2036	2.66
Gender	3.49873	1.558	2.25
nonSlav	1.77690	0.9169	1.94

$R^2 = 33.9\%$ R^2(adjusted) = 29.8%
s = 6.036 with 70 - 5 = 65 degrees of freedom

b. **Effects of Gender, nonSlav and Reading Proficiency (n = 70)**

Variable	Coefficient	s.e. of Coeff	t-ratio
Constant	11.0205	3.247	3.39
ETSL1	-0.665769	0.1065	-6.25
rowF	0.475666	0.1865	2.55
Gender	2.32520	1.456	1.60
nonSlav	1.44705	0.8413	1.72
ETSR1	0.356082	0.0949	3.75

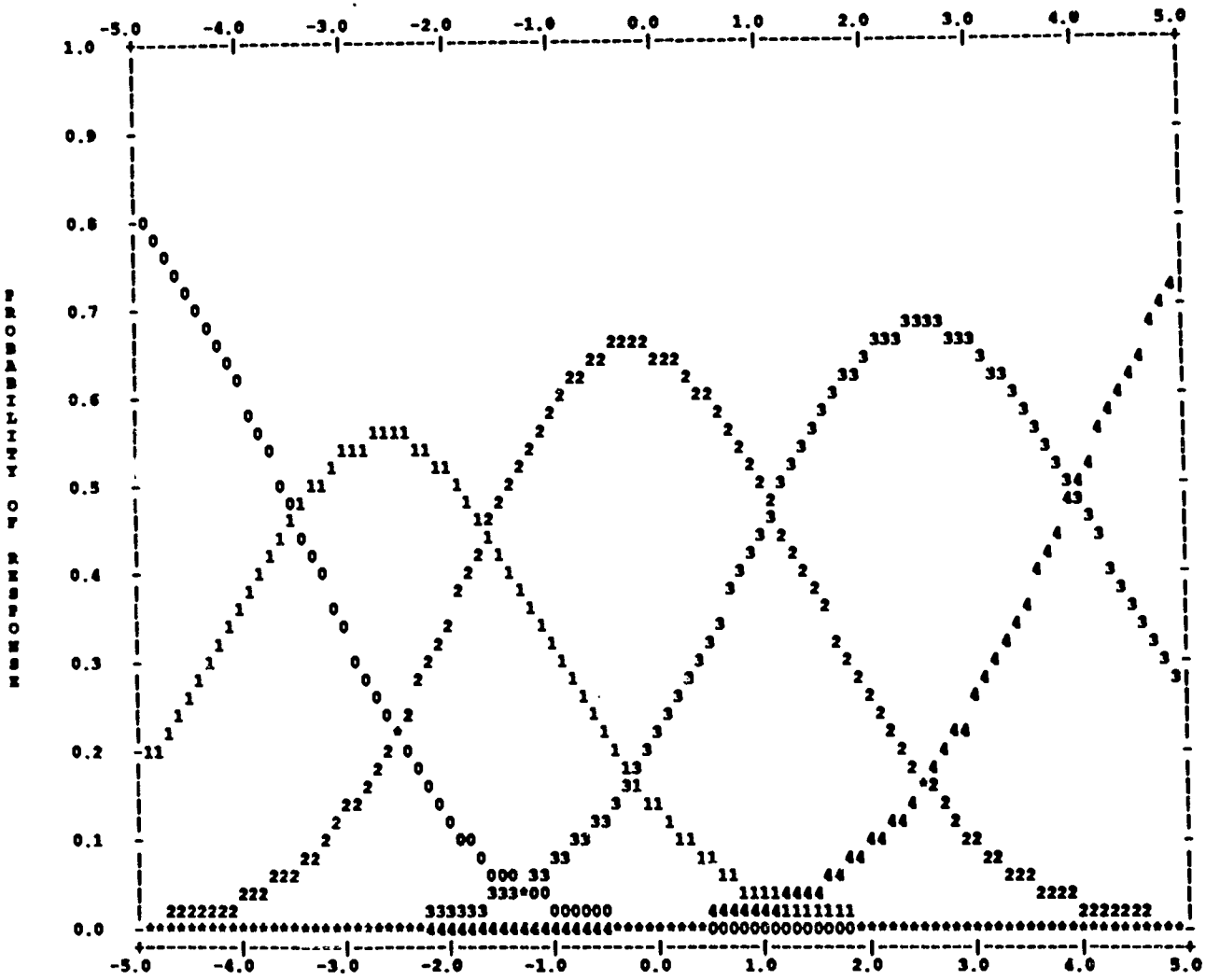
$R^2 = 45.8\%$ R^2(adjusted) = 41.6%
s = 5.507 with 70 - 6 = 64 degrees of freedom

c. **Effects of Gender, nonSlav and ETS Reading Proficiency, Four month students only (n = 59)**

Variable	Coefficient	s.e. of Coeff	t-ratio
Constant	10.2199	3.526	2.90
ETSL1	-0.715775	0.1189	-6.02
Gender	2.90623	1.624	1.79
nonSlav	1.67234	0.9321	1.79
ETSR1	0.340874	0.1097	3.11
rowF	0.565075	0.2186	2.59

$R^2 = 48.0\%$ R^2(adjusted) = 43.1%
s = 5.665 with 59 - 6 = 53 degrees of freedom

Figure 1
MSCALE Calibration of PreProgram Self-Assessments
Response Category Probability Curves



STEP VALUES: -3.46 -1.60 1.11 3.95

Figure 2
MSCALE Calibration of PrePregram Self-Assessments
Most Probable Response Curves

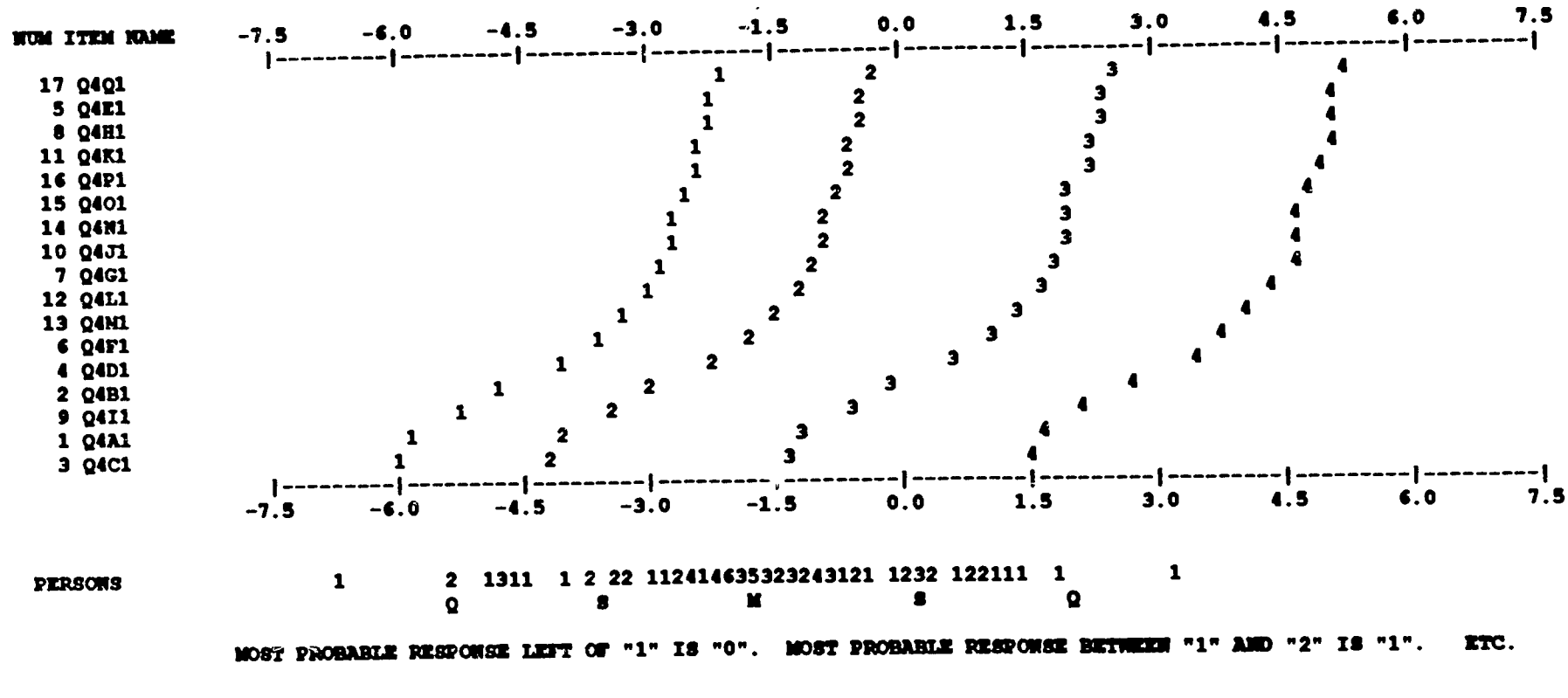


Figure 3
Distributions of Pre-Program Measures

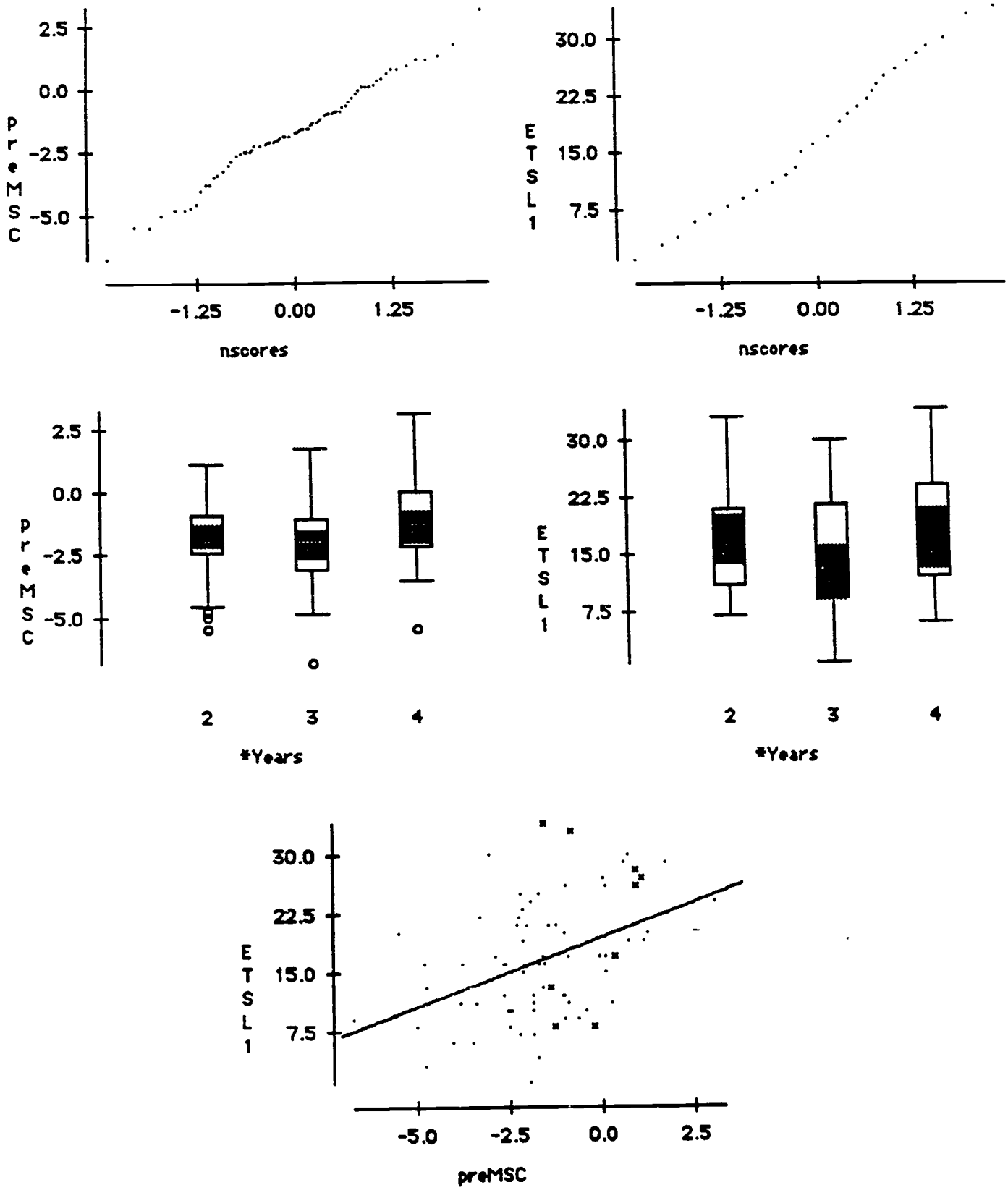
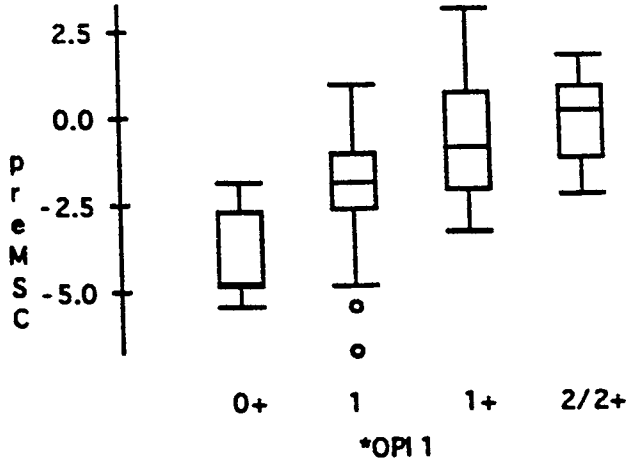


Figure 4
Relationship between OPI and Scaled Self-Assessments

a. Pre-Program

Frequency breakdown of *OPI1

Group	Count	%
0+	9	11.5
1	45	57.7
1+	16	20.5
2/2+	8	10.3
Total	78	



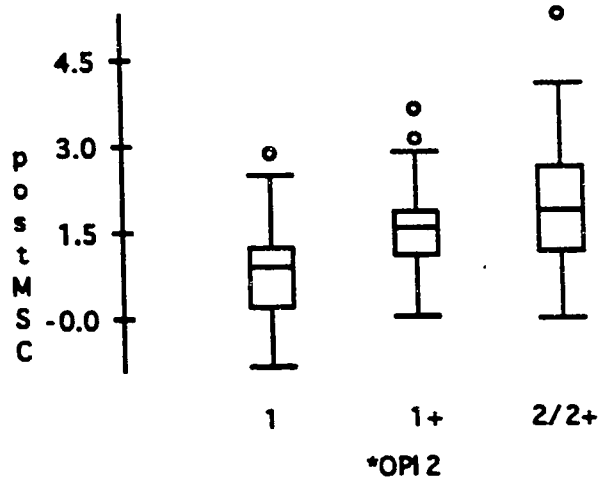
Analysis of Variance For preMSC
82 total cases of which 4 are missing

Source	df	Sum of Squares	Mean Square	F-ratio	Prob
*OPI1	3	88.1344	29.3781	12.283	0.0000
Error	74	176.986	2.39171		
Total	77	265.121			

b. Post-Program

Frequency breakdown of *OPI2

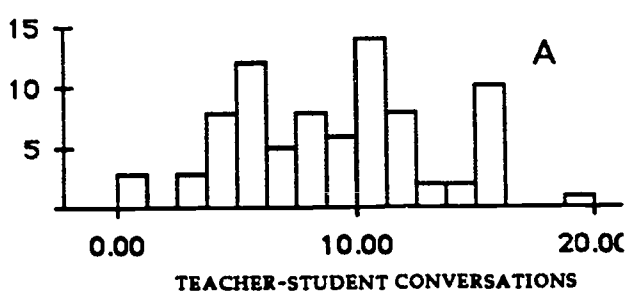
Group	Count	%
1	27	38.6
1+	29	41.4
2/2+	14	20
Total	70	



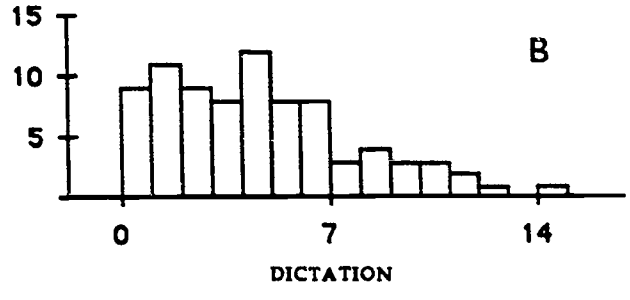
Analysis of Variance For postMSC
82 total cases of which 12 are missing

Source	df	Sum of Squares	Mean Square	F-ratio	Prob
*OPI2	2	16.2365	8.11824	8.2847	0.0006
Error	67	65.6540	0.979910		
Total	69	81.8904			

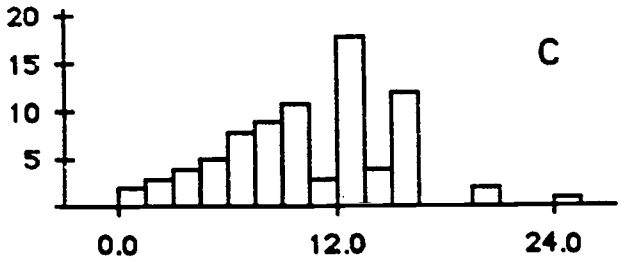
Figure 5
Distribution of Listening Activities



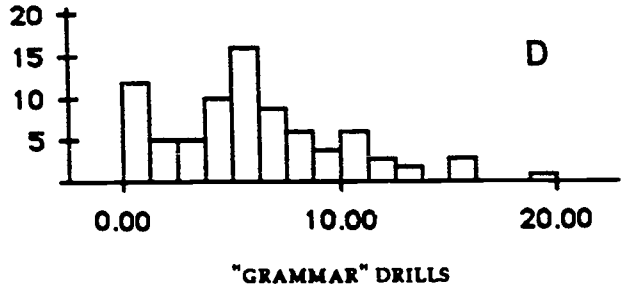
TEACHER-STUDENT CONVERSATIONS



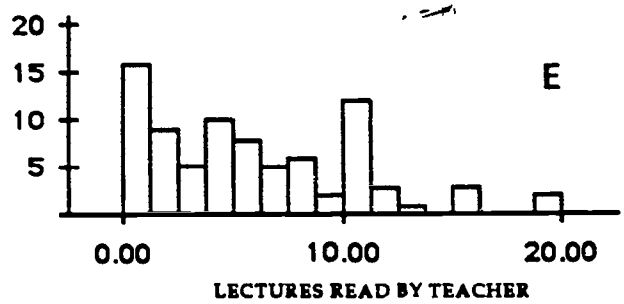
DICTATION



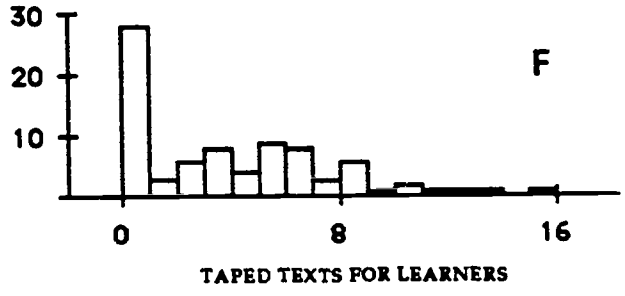
LISTENING TO OTHER STUDENTS



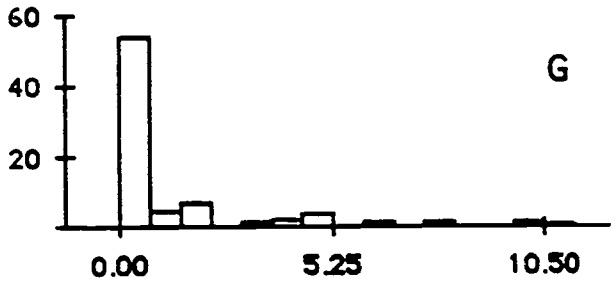
"GRAMMAR" DRILLS



LECTURES READ BY TEACHER



TAPED TEXTS FOR LEARNERS



AUDIO TAPES FROM SOVIET MEDIA

VIDEO TAPES FROM SOVIET MEDIA

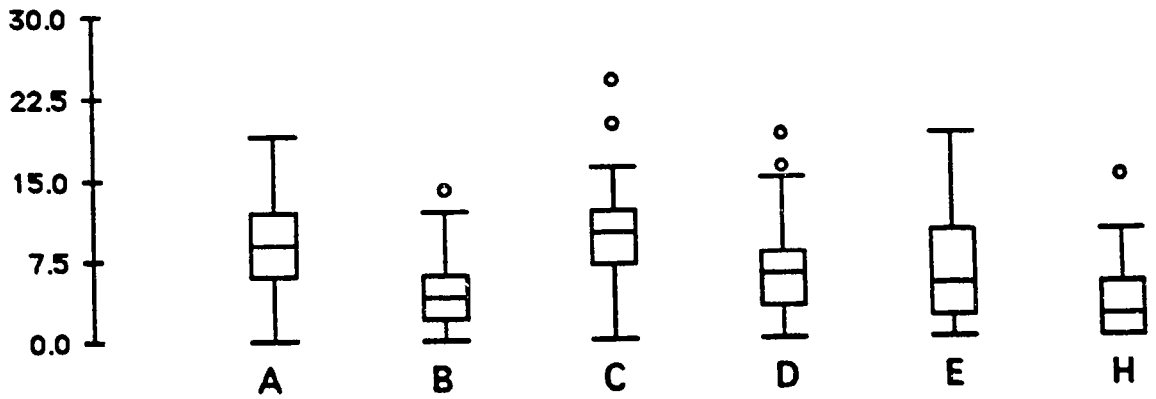
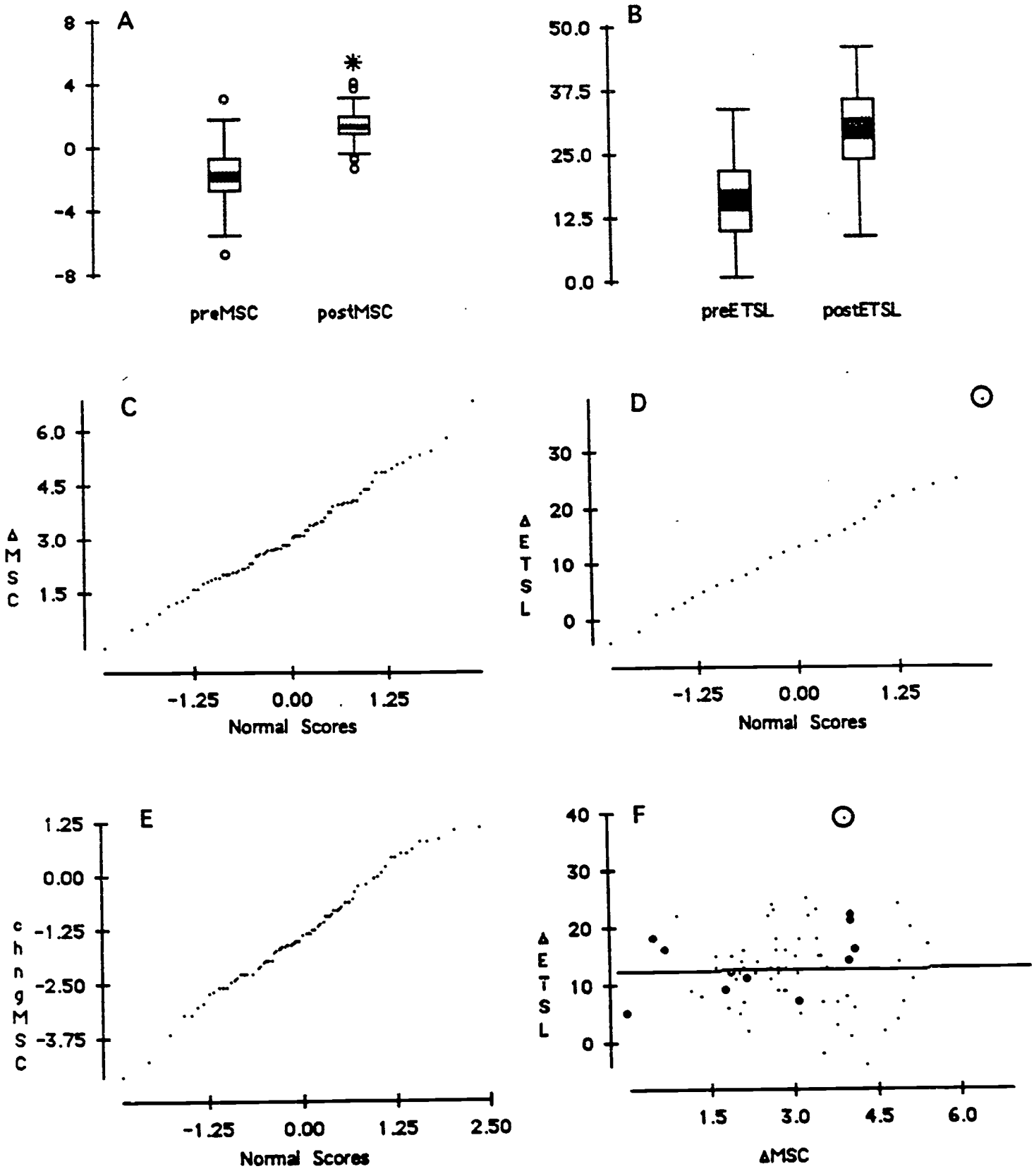


Figure 6
Plots of Change Measures



The National Foreign Language Center
at the Johns Hopkins University
1619 Massachusetts Avenue NW
Suite 400
Washington DC 20036
Telephone 202/667-8100