DOCUMENT RESUME

ED 248 708 FL 014 564

AUTHOR Hoover, Wesley A.; And Others
TITLE A Longitudinal Look at Classroom Instruction and Reading Acquisition by Spanish-English Bilingual Students.
PUB DATE Apr 84
NOTE 56p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 23-27, 1984).
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Bilingual Students; English; *English (Second Language); Language Research; Longitudinal Studies; Primary Education; *Reading Achievement; *Reading Instruction; Second Language Learning; Spanish; *Spanish Speaking; Teacher Characteristics

ABSTRACT

A comprehensive six-year longitudinal study of the development of reading skills during the primary grades for a large sample of bilingual (Spanish-English) children and smaller samples of monolingual (English or Spanish) children is outlined at its midpoint. In this natural variation study, approximately 350 children taught by 200 teachers in 20 schools in six districts are tracked through the primary years. Their reading development and mastery of formal language is examined in detail each year through multiple measures, as is their instruction, through an array of indices, including classroom observations made throughout each academic year. In addition, information about the teachers' background, training, and language skills is gathered. Data available at this stage of the study, from a subsample of 63 children in grades 1-3, on several of the components of an interactive reading assessment in English and Spanish are analyzed and presented in detail, including charts of average growth and performance profiles for a variety of the measures used. (MSE)

A LONGITUDINAL LOOK AT CLASSROOM INSTRUCTION AND READING ACQUISITION BY SPANISH-ENGLISH BILINGUAL STUDENTS

Wesley A. Hoover
Southwest Educational Development Laboratory

Robert C. Calfee
Stanford University

Betty J. Mace-Matluck
Southwest Educational Development Laboratory

Paper presented at the meeting of the American Educational Research Association - New Orleans, Louisiana, April 1984.

ABSTRACT

This paper presents an overview of a comprehensive six-year longitudinal investigation of the development of reading skills during the primary grades for a large sample of children from bilingual (Spanish-English) backgrounds, and for smaller samples of children who are monolingual in English or Spanish. In this "natural variation" study, approximately 350 children taught by 200 teachers in 20 schools in six districts have been tracked through the early years of schooling. The study has carefully examined the children's development in language and reading on a yearly basis through multiple measures, coupled with an array of indices of the instruction received, including classroom observations made throughout each academic year. In addition, information has been gathered about each teacher's background, training, and language skills.

The paper also includes an exploratory analysis of growth in Spanish and English reading skill (along a number of dimensions as measured by a single instrument) in relation to the instruction received (as documented by the observation instrument) over grades 1 through 3 for a small, unrepresentation, subsample of the target students.

Introduction

Many children from second-language backgrounds have trouble learning to read in schools today, and many of these youngsters are from Spanish-language backgrounds, and are impoverished. The Bilingual Reading Study, now nearing completion at SEDL, is a comprehensive longitudinal investigation of the development of reading skills during the primary grades for a representative sample of more than 250 Texas children from bilingual backgrounds, and for smaller samples of children who are monolingual in English and in Spanish. In this "natural variation" study, teaching and learning have been carefully documented in field settings at several sites in order to (1) describe variations in both English and Spanish language competence for students living in bilingual communities; (2) document prevailing practices in classroom instruction for bilingual students; and (3) to allow a valid examination of the relations between instructional program and student achievement for students with differing entry profiles.

We are currently in the middle stages of the longitudinal data analyses, and in this paper, we will only present data from a subsample of the target students concerning their development in English and Spanish reading as measured by a single instrument in relation to the instructional program they have received as documented by the observation instrument over grades 1 through 3. The presentation will in general be non-technical, and is intended only as a sketch of the kinds of data collected with respect to reading development and instruction, with a glimpse at how the two are related. This relationship is not quantified here, but we are currently investigating a model capable of doing so via an index of the distance variance between the achievement and instructional profiles using standard-score transformations for each of the measures--a full complement of

technical reports will appear in December of 1984. We now turn to a general overview of the Bilingual Reading Study.

An Overview of the Design of the Study

To achieve the objectives of the study, considerable attention was given to the selection of schools, teachers and students, to the instruments for assessing language and reading achievement, and to the methods for documenting the classroom instruction.

Schools, classes, and teachers. Some 20 schools and 200 teachers have participated in the SEDL study, providing variations in the nature of the reading program (a range from phonics-oriented to meaning-based), classroom organization (some self-contained, others team-taught), and grade structure (the range of grades in the individual school and the extent of cross-grading both vary). The schools differ in size, SES, urbanicity, and makeup of the student body (from medium to high concentrations of bilingual students).

Student cohorts. The study has been undertaken in three cohorts or "waves" of students. The first sample drawn was small (N=40) and of limited generality; the second was somewhat larger (N=80) and covered a slightly broader array of contexts. The third sample was both larger (N=250) and broader in its generality, and incorporated a number of procedural improvements based on experience gained from work with the first two cohorts.

All of the bilingual sites are from the state of Texas; included in the sample are smaller cohorts that are either monolingual in English (from the northern and central Texas area) or in Spanish (from Chihuahua, Mexico). Most students entered the study as kindergartners (the remaining students as first graders), and all will remain in the study through second

grade, 40% of the sample through third grade, and 25% through fourth grade--a critical period for the development of literacy.

Language assessment. Several types of data have been collected for each student on English and Spanish language proficiency. Each year, early in the Fall and late in both the Winter and Spring, we asked teachers to rate their students' language skills on a number of dimensions. We have also collected standardized oral language test data from Fall district-wide administrations. Finally, we have obtained recorded speech samples for most students in three settings--the classroom, the playground, and the home.

Reading assessments. Several instruments have been used to measure reading achievement. We have collected standardized reading achievement scores when available (mostly in English). More detailed information comes from a battery of individually-administered "performance-based" tests in both English and Spanish. In kindergarten or on entry to first grade, the Stanford Foundation Skills Test was employed to measure the child's pre-reading skills. From the end of first grade on, the Interactive Reading Assessment System was given during the Spring of each year; this instrument provides independent measures of a student's skill in decoding, word meaning, fluency in oral reading, and comprehension under listening and reading conditions. Finally, informal reading inventories were administered throughout the school year.

Classroom observations and teacher interviews. Monthly observations of the reading instruction in each classroom have been made, and teachers have been interviewed quarterly about their rationale for the program of instruction. The observation instrument covers staffing, grouping and organization, time allocation, the language of instruction, the character

of instruction and the materials and procedures employed, and the response of the students. The interviews focused on the teacher's general instructional objectives, as well as the objectives for individual target students. Together, these two instruments yield a rich characterization of the classroom environment for the target students.

In summary, the database established for the target sample provides a relatively comprehensive picture over the primary grades of (1) the development of language and reading skills in both English and Spanish, and (2) the instruction received during this developmental sequence. The next step to be taken is to link these two data sets, and we now turn to the analysis which explores this linkage.

The sample for the exploratory analysis discussed in this paper includes all bilingual students from Cohorts I and II who had completed third grade by the end of the fourth year of the study (Cohort I, N=36; Cohort II, N=27). The first cohort came from the South Texas-Mexico border area, while Cohort II was drawn from the West Texas-Mexico border area. Both areas are rural, of low socioeconomic status, and have large numbers of Spanish-dominant students.

Summarizing Progress in Reading

A primary purpose of the Bilingual Reading Study was the investigation of patterns of growth in reading achievement, and in the mastery of formal styles of language, of which reading is a special instance. This discussion will begin with a presentation of the concept of the growth track as a summary of the acquisition of reading skills over time. This concept will be linked to the separable-process theory, and to the design of the Interactive Reading Assessment System (Calfee & Calfee, 1979, 1981; Calfee, Calfee, & Peña, 1979), the instrument that will receive most attention

here. Next we will present the aggregate summaries of the reading achievement of the two cohorts, followed by a discussion of a series of protocols for individual students.

The Growth Track

The Bilingual Reading Study adopted as a foundational assumption the notion that reading is a dynamic, developmental process, and thus it was necessary to tailor both the design and the data analysis to be sensitive to the character of changes in student performance over time--more specifically, to trends that occur over the four or five years that comprise the primary reading program.

Although reading research (and educational research in general) has given little attention in recent years to the measurement of the course of learning (e.g., "What is the shape of the learning curve?" is a question that seldom arises in educational research at present), instructional programs still reflect this dimension. For instance, basal reading materials are carefully graded to present the student with a set of learning materials and experiences that gradually increase in difficulty as the student moves through the program.

The Interactive Reading Assessment System (IRAS) incorporates the developmental dimension of the basal reading series for all components of the separable-process design. For instance, the series of words at the beginning of the test, which the student is asked first to decode and then to define, is graded by reference to several of the standard word counts used in basal-reader designs. The synthetic word lists for assessment of decoding in IRAS are ordered according to several factors known to affect difficulty of pronunciation as these are reflected in the typical scope-and-sequence charts. The sentences used for assessment of oral reading

increase in a regular fashion on the factors of length (number of words) and syntactic complexity. The texts used to assess comprehension increase in overall length (number of words), propositional "load" (Kintsch & van Dijk, 1978; for practical purposes, this factor is the number of distinctive ideas), and text structure (Calfee & Curley, 1979); expository texts of increasing formality are introduced at the second grade level in addition to the narrative texts appearing at all levels.

The construction of the materials in IRAS was graded with the aim of introducing a one-year increase for every two levels on the test. Thus, success on Level A for each of the IRAS components should correspond more or less to the curriculum halfway through the end of first grade, success on Level B should identify a student who could handle the materials at the end of first grade, and so on.

The design of IRAS into components and levels for each component was coupled with an informative, but efficient, technique for determining the student's proficiency level for each component--in essence, the technique was to locate as quickly as possible two critical levels for each component: the level at which the student did relatively well, and the level at which the student did relatively poorly. These two levels were usually adjacent to one another.

The details of this strategy are described in the IRAS manual, but a couple of examples will help the reader who is not familiar with the instrument. The first task for the student was to scan a series of graded-word lists, six words per list. The student was informed that the words increased in difficulty from one list to the next, and was asked to scan through the series until he or she encountered a list that was too difficult to "read" (i.e., to decode). Virtually every student understood the

task without apparent difficulty, and most students quickly went about searching through the lists to find the limits of their mastery. Once a selection was made, the student was asked to pronounce each word on the immediately preceding list. If the student did reasonally well, the next more difficult list was presented for pronunciation; if the student did too poorly on the first list, the next easier list was presented. This procedure was continued until failure was found for students presented with more difficult lists than the initial one or until success was found for students presented with lists less difficult than the initial list presented.

The second example is the comprehension task. The critical vocabulary level provides an estimate of the level of text at which the student can read aloud with a reasonable degree of fluency. This estimate needs to be off only slightly to substantially increase the amount of testing time required for comprehension assessment--the vocabulary definition and the text comprehension tasks required a disproportionate amount of time because a free-response mode was employed. A more precise estimate of the critical text level, along with an erficient sample of oral readi fluency, was gained by having each student read aloud a series of sentences of graded difficulty. Each sentence could be read by a proficient reader in less than 20 seconds. If a student made too many mistakes or took too long on one of the sentence sets, the tester stopped the task at that level, and presented the narrative text at the next lower level for assessment of comprehension.

The critical-levels technique generated two types of information on each of the component tasks. One measure was the student's highest level of success, where "level" refers to the IRAS levels described previously.

The second measure was an index of the quality of the response on the two critical levels--highest level of success, and the next level where performance dropped below the critical value required for success. Throughout IRAS, these criteria were generally set quite low, so that if a student made 'correct' responses to half or more of the items contained in a given level of a task, this was considered success.

In deriving a score for a given task, the quality of individual reponses within each level attempted were first scaled for their degree of 'correctness.' For example, in Definitions, a response providing a complete formal definition was assigned a score of '3'; a poor, but acceptable, definition was given a value of '2'; a correct multiple-choice response received a '1'; and an incorrect multiple-choice response was assigned a value of '0.' For purposes of determining success at a given level, any item receiving a value above 0 was counted as a successful response. An index of the quality of response at a given level was formed by calculating the proportion of points received relative to the total number of possible points at that level. To summarize a student's performance on a given task, both level and quality indices were included by taking the ordinal value of the level of highest success, and adding to it, the average of the quality indices at that level and the next level where performance was not successful (e.g., a student who passed level E with 75% of the total possible points, but failed level F with 25% of the total possible points received a score of 5.5).

As noted earlier, students were tested with both the English and Spanish versions of IRAS each Spring. The Spanish version was constructed to parallel the English version and was not simply a translation. Rather, the same principles used to ground the English version were followed in

building the Spanish version (e.g., Spanish word counts were employed to select appropriate vocabulary items for the decoding, definition, and comprehension tasks). The scales within the Spanish version were scored in the same fashion as those just described for the English version.

The design of IRAS, together with the technique for determining a student's level of competence on the test, led to the postulation of an extremely simple model of growth over time. The model, shown in Figure 1, represents student growth over the years of schooling as a straightline function. The correspondence with the grade level of the basal reading series is also displayed on the graph, along with boundary limits for progress that are one year above or below the expected level. The "typical" student, based on the instructional materials, should have trouble with the lowest level of IRAS in kindergarten, but should meet criterion on the second, fourth, and sixth levels of the test when exiting from the first, second, and third grades, respectively.

The normative model shown in Figure 1 implies linear growth, as does the IRAS "levels" model. A linear model of progress has much to recommend it, and to the extent that the design of the materials for IRAS has achieved this goal, the growth track for this test should yield data that are extremely easy to interpret.

Analysis of average performance. This section describes the longitudinal average data for the subsample of 63 students drawn from Cohorts I and II on several (but not all) components of the Interactive Reading Assessment System in English and Spanish.

Let us first examine performance on the Definitions task, the data for which are shown in Figure 2. The averages have been laid out according to the growth track model presented earlier. The English and Spanish results

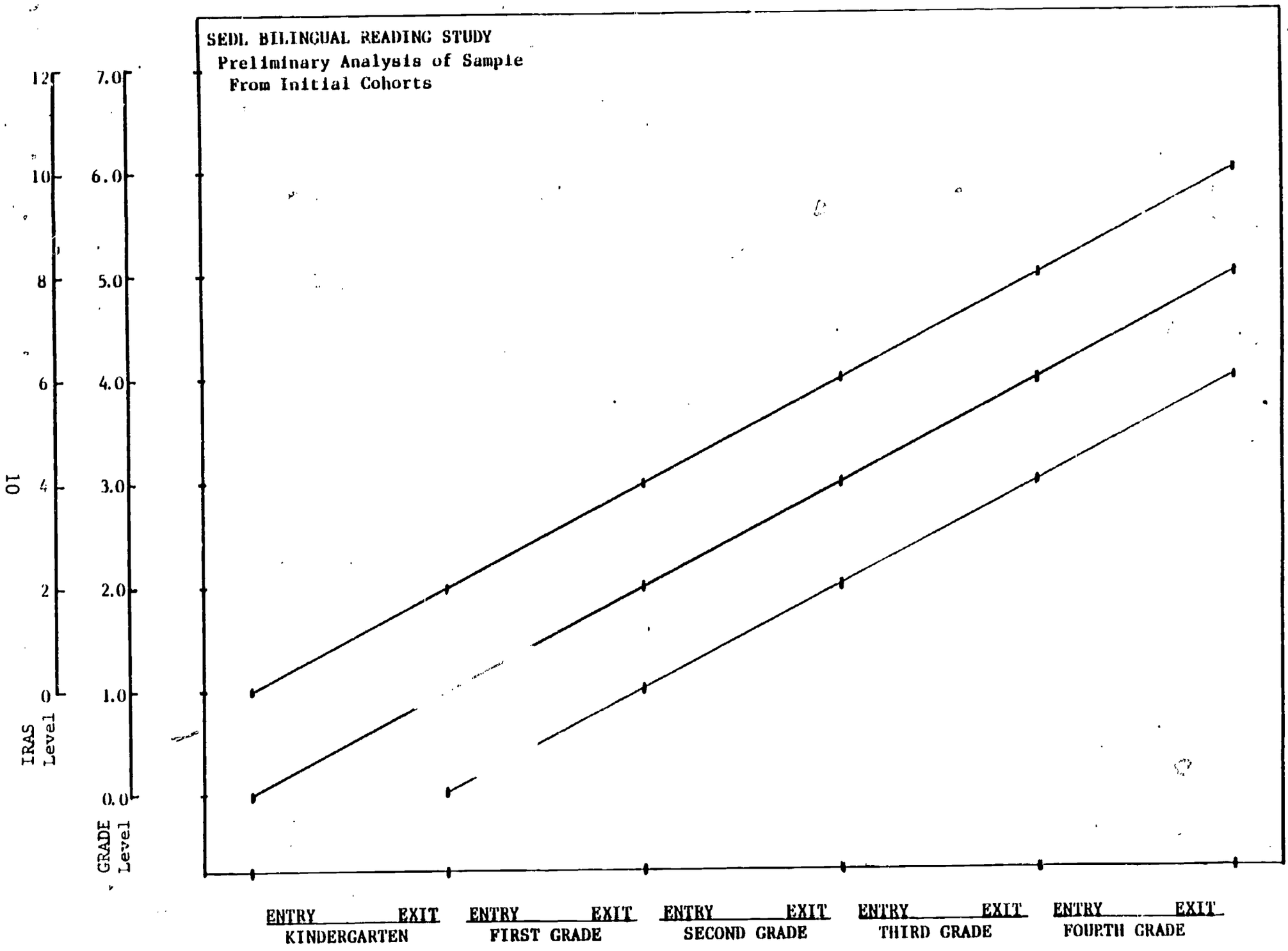


Figure 1. Growth track model based on design of IRAS levels.

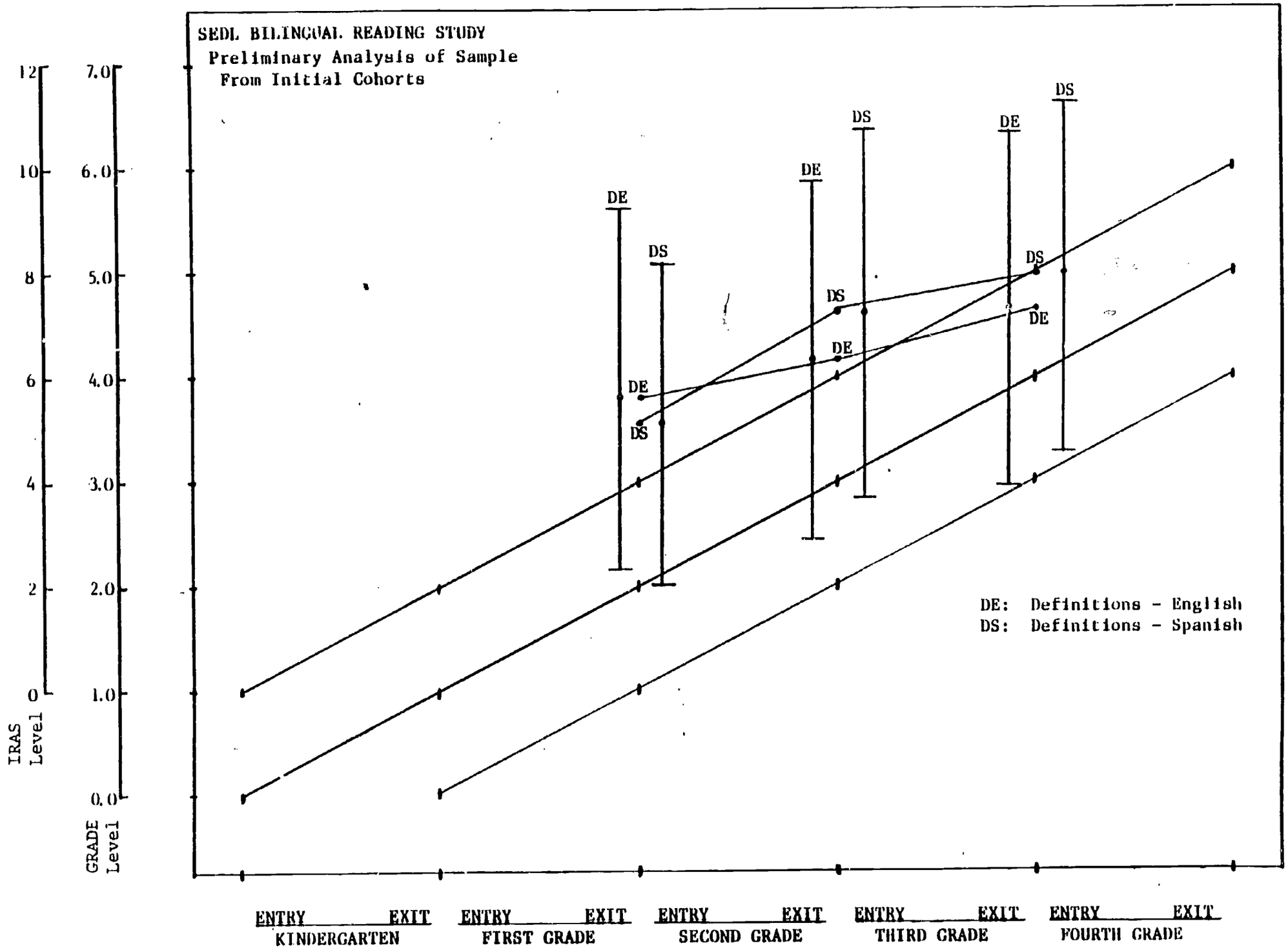


Figure 2. Average growth track data for Vocabulary Definitions as measured by IRAS in English and Spanish.

have been juxtaposed in the graph to permit an assessment of the degree of bilingualism in this sample of students. The Definitions task is based entirely on students' oral language skills--the target word is pronounced for the youngster, who is first asked to explain what the word means, and who is then prompted with a set of three alternatives if a suitable explanation is not forthcoming. This task, therefore, seems appropriate as one index of the level of competence in the spoken language and, at the upper levels of IRAS, of formal knowledge of the more complex concepts from each of the languages.

The layout of the aggregate data on the growth track follows the same pattern for each of the graphs that follow, and thus a detailed discussion of Figure 2 is called for. Students were tested in the Spring of each school year, so that the data points in the figure represent performance on exit from the grade indicated on the abscissa. Additional information is available for precursor skills and language performance on entry to kindergarten and at several other points along the growth track. These sources of data will be "added to the track" during subsequent analyses to provide a more complete representation of growth patterns in reading. For the present, the focus is on IRAS data at the end of first through third grades. The boldface symbols in the figure mark the averages for Definition of English words (DE) and of Spanish words (DS). The vertical bars beside each of the averages indicate the extent of variability (one standard deviation on either side of the mean).

As cautioned earlier, the averages in Figure 2, and elsewhere in this paper, are based upon a subsample of the first two cohorts only, a small and unrepresentative sample of the overall target student sample. With

this caution in mind, a few comments are nonetheless deserving of mention. First, as we suspected, students are able to define words that are normatively quite a bit more difficult than the readability limits in typical basal series, especially in the early primary grades, where the level of word knowledge is almost two grade levels ahead of the nominal limit.

The second point to notice is that the averages for English and Spanish are virtually identical for this sample. The reader should bear in mind that averages are seldom typical of individual profiles. The means in the figure could reflect a situation in which most of the children are equally competent in Spanish and English at all grades, or where half of the students are quite proficient in one of the languages and virtually deficient in the other--or a number of combinations of other patterns.

Some idea of the behavior of individual students in first grade can be gained from the scatterplot in Figure 3 of English versus Spanish definition performance. A preponderance of the students demonstrated substantial competence in both languages; 27 of the 63 scored above the third-grade level in both languages at the end of first grade. Another 13 children were quite capable in English, though not in Spanish, and the remaining students did poorly on the English test, with varying levels of capability on the Spanish test. Only two of the students were below the first-grade level in both languages.

The final matter to be noted is that the level of vocabulary proficiency increases in both languages from first through third grade, but at a rate that is substantially less than "a-year-for-a-year," and at a somewhat slower rate for English than for Spanish. The students possess a relatively extensive vocabulary at the end of first grade (the typical student can define English words like company, crowd, and electric). After thr

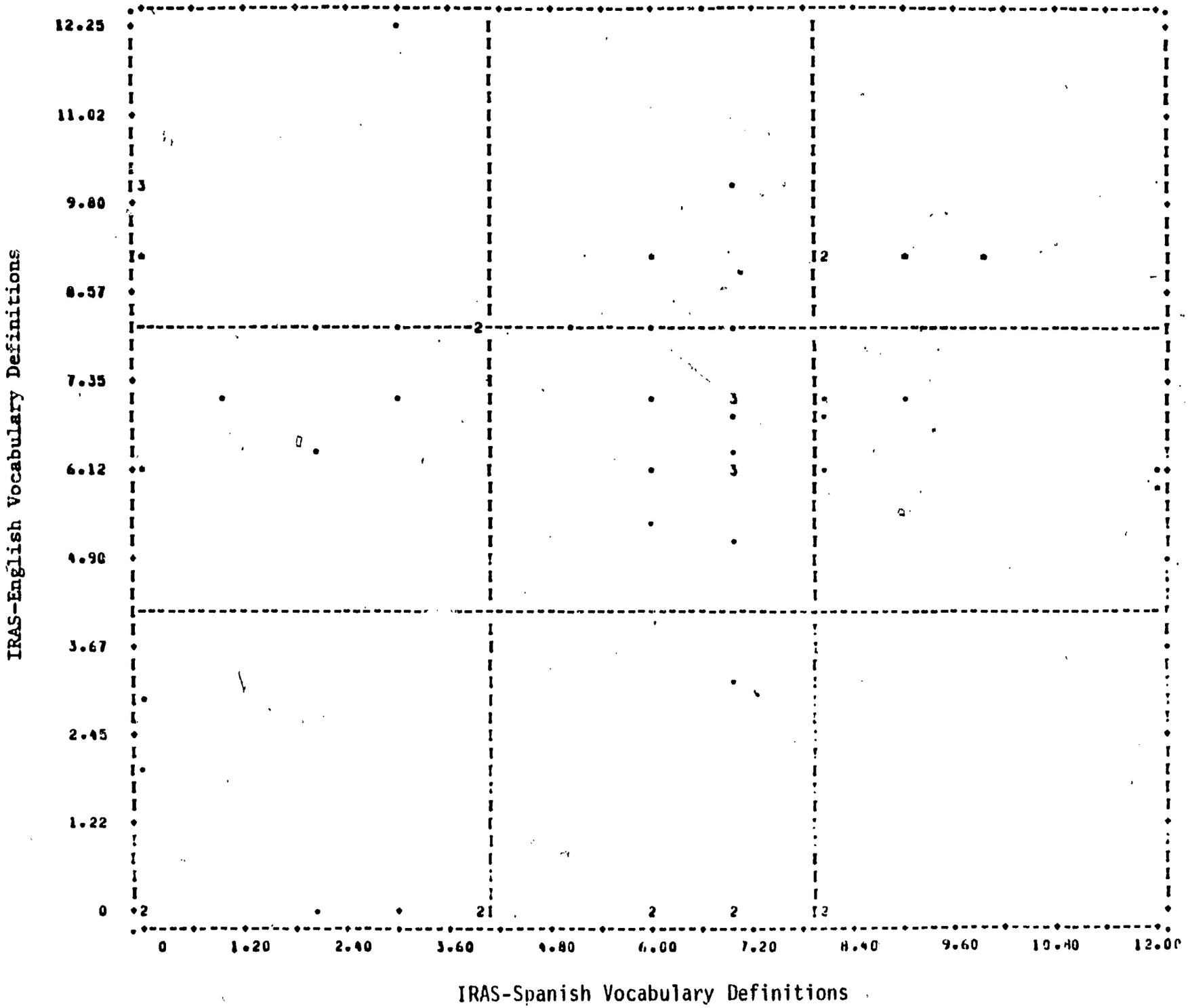


Figure 3. Scatterplot of IRAS Vocabulary Definition scores in English and Spanish for Grade 1.

years of schooling they have made relatively little additional progress through the Carroll frequency list (Carroll, Davies, & Richman, 1971); they have reached words like committee, invisible, and mission, but have not learned to handle permanent, annual, and literature.

One of the more significant features of Figure 2, for purposes of the methodology of this study, is the tendency for the average scores to increase in a straightline fashion over grades. The entry level is higher than the grade-equivalent model projects, but the reasons for this departure from the model have been noted. The rate of growth over time is less than the simplest version of the model predicts; we suspect that this result may be a relatively accurate reflection of the actual effectiveness of vocabulary instruction as presented in the typical basal series. Nonetheless, to the extent that performance does undergo systematic change over the years of schooling, the data in Figure 2 suggest that aggregate changes for Vocabulary Definitions in both English and Spanish can be accounted for reasonably well by postulating a linear growth model--students tend by and large to move across the levels of IRAS at a constant rate over time.

Decoding is an important component in beginning reading according to many scholars (e.g., Chall, 1979). There is considerable controversy about when and how phonics instruction should be introduced, but most basal systems have resolved this controversy by providing materials that permit an eclectic approach--phonics materials are made available for the teacher to use even in those series that stress comprehension from the earliest grades.

Figure 4 displays the average level of performance and the amount of between-student variability for the two tasks designed to assess decoding skill in the English version of IRAS. The two measures tend to cross-

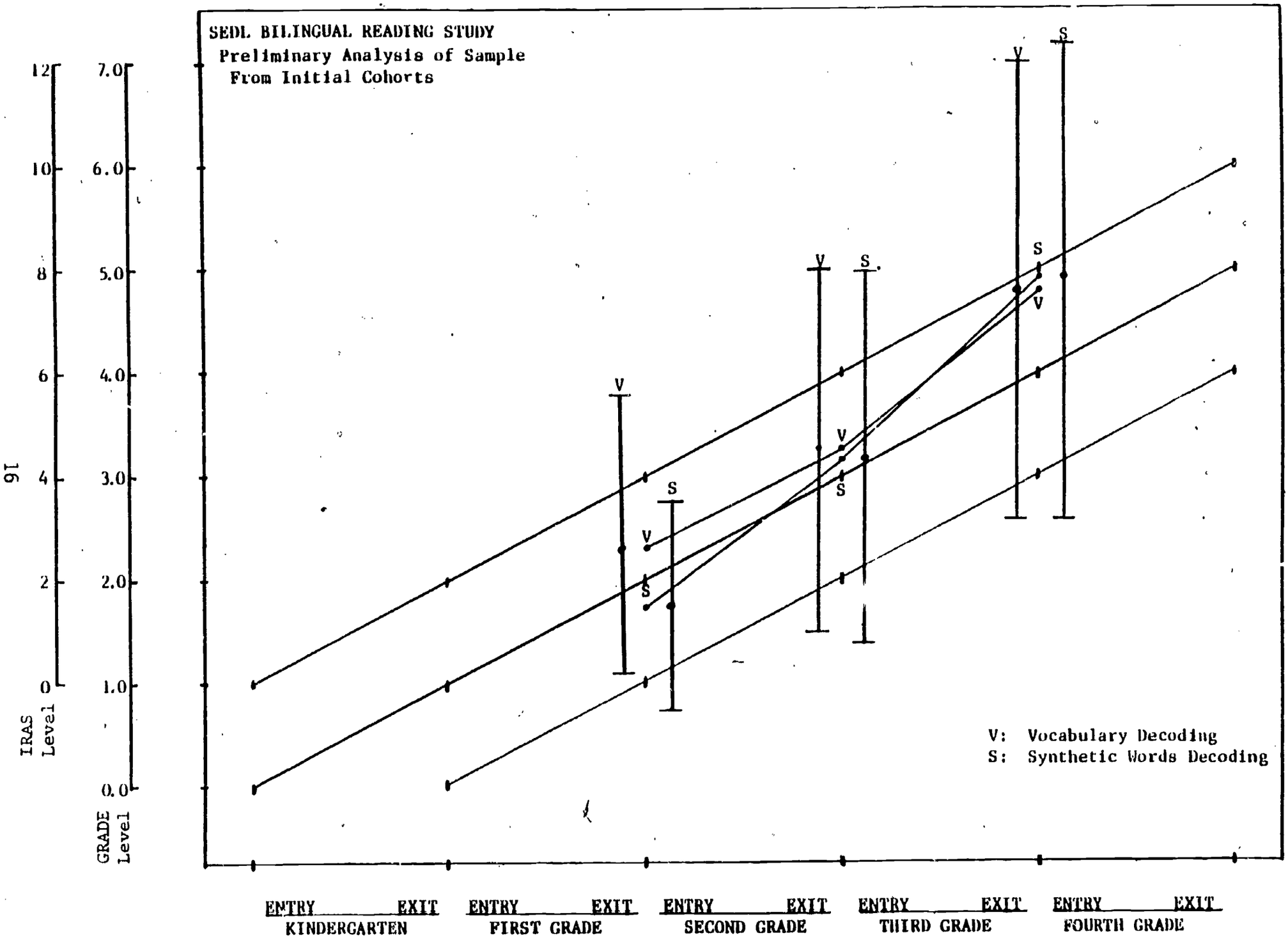


Figure 4. Average growth track data for Vocabulary Decoding and Synthetic Word Decoding as measured by IRAS in English.

validate each other in the aggregate--performance is at about the level expected from the design of the IRAS materials at the end of first grade, and increases at somewhat more than a year-per-year over the following two years. By the end of third grade, students can pronounce real words at about the same level that they can define them, on the average. By the end of first grade, synthetic words can be pronounced if they conform to relatively simple consonant-vowel-consonant patterns. By second grade, the typical student can handle a variety of more complex patterns (consonant digraphs like SH and vowel-plus-R combinations), along with polysyllabic words based on familiar combining forms (-ED, -ING, -FUL, UN-, and IM-). Third graders can manage the most complex Anglo-Saxon spellings on the list (KNOP, WRUDGE) and relatively simple polymorphic words (DACTURE, BEFADE), but are not able at the end of third grade to handle the complex Romance spellings (AFFREMIATION).

It is important to note that the variability in decoding skills is quite substantial, especially at third grade. Some students are doing extraordinarily well, while others remain at a very low level.

Figure 5 shows the results for English Reading and Listening Comprehension (narrative texts only--analyses of the expository texts are forthcoming). In general, the ability of this sample of youngsters to handle connected text is much below the level that one might expect, judging from their performance on either the Definitions or the Decoding components. Performance is close to the grade-equivalent level at the end of first grade, but increases by only about one grade level between that time and the end of third grade. The students are better at listening comprehension than reading comprehension, as one might expect (problems with 'fluent' decoding would tend to cause difficulty for some students when reading on

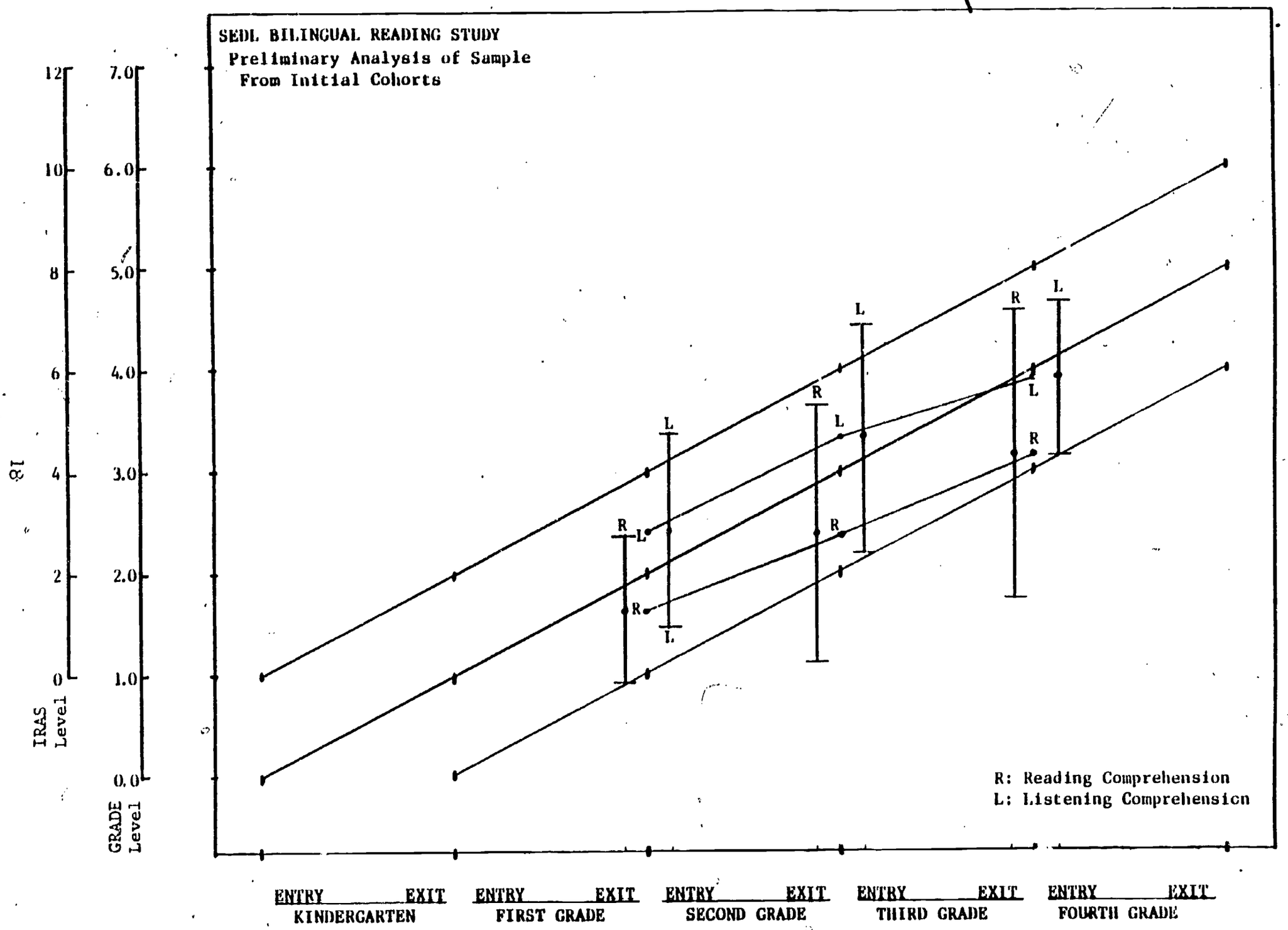
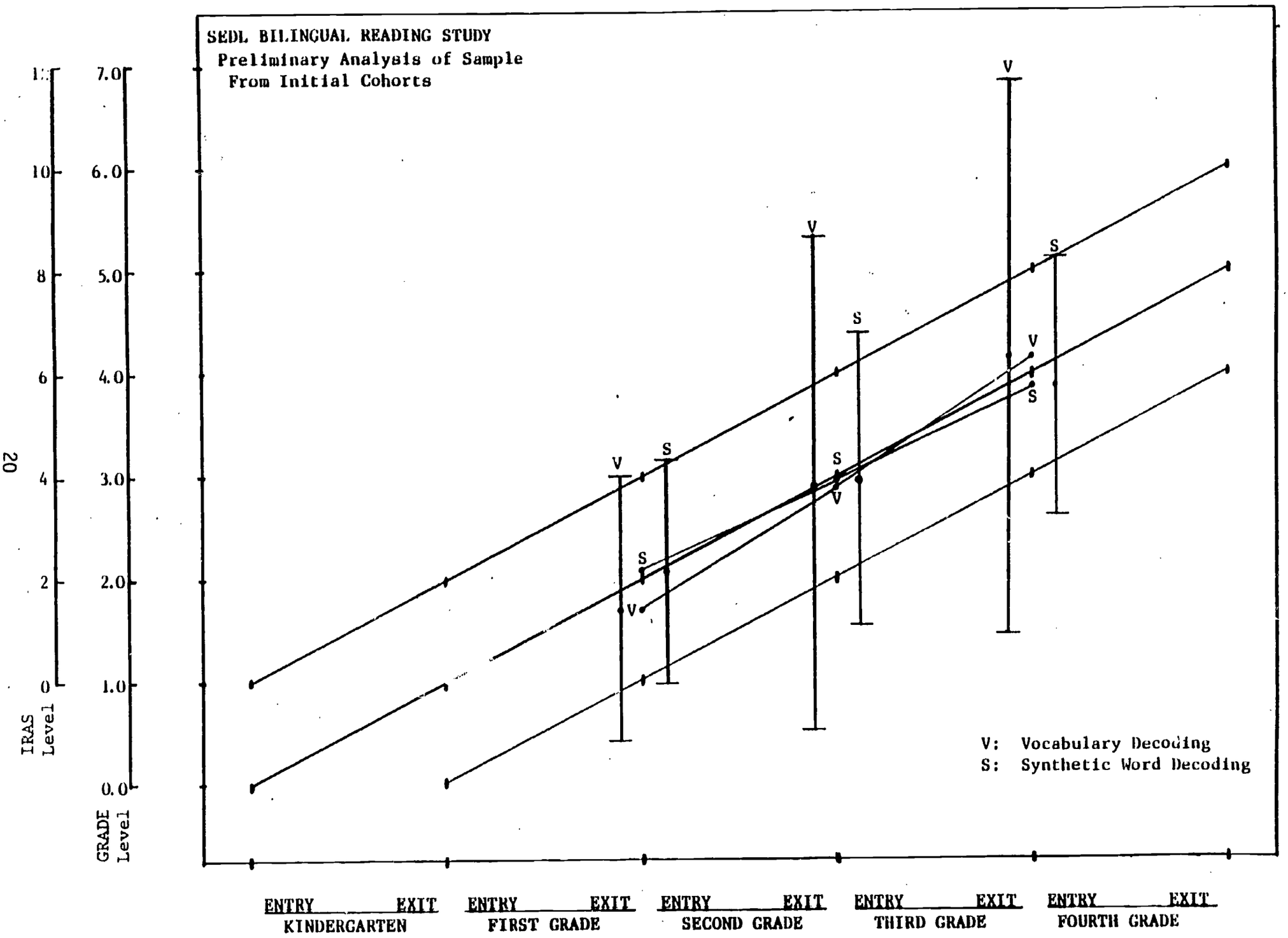


Figure 5. Average growth track data for Reading Comprehension and Listening Comprehension as measured by IRAS in English.

their own), but both forms of comprehension exhibit roughly equivalent progress from first through third grade. The Listening Comprehension index is bounded--the upper limit is at level 7.0--which accounts for the relatively small amount of variability in this measure during third grade, and which may have unduly restricted the sensitivity of this performance measure for the more capable third graders. While the average in the figure is more than a full level below the limit, examination of the frequency distribution revealed that half of the third-grade students had reached the top-most level in listening comprehension. The Reading Comprehension index is well below the upper limit, and the observed level--a year below the expected level at third grade--is probably a trustworthy indication that the students are not performing at grade level. It also appears that they are not performing up to their potential in comprehension given the level of their skills in vocabulary and decoding.

We turn next to performance on the Spanish version of the IRAS. Decoding skill levels are shown in Figure 6. Both the Vocabulary and the Synthetic Word tasks complement one another, as was true for the English version. Performance is at the levels determined by the design of the test as appropriate to the students' grade level. The rate of change is not as rapid for Spanish as for English (Figure 4).

Comprehension in Spanish is displayed in Figure 7. The patterns are similar in some respects to those for English--Listening superior to Reading, performance in both areas below the levels for decoding and definitional skills, and a slower rate of progress than expected from the design. There is one noticeable difference between the English and Spanish plots: Reading Comprehension in Spanish is virtually negligible at the end



20

e 6. Average growth track data for Vocabulary Decoding and Synthetic Word Decoding as measured by IRAS in Spanish.

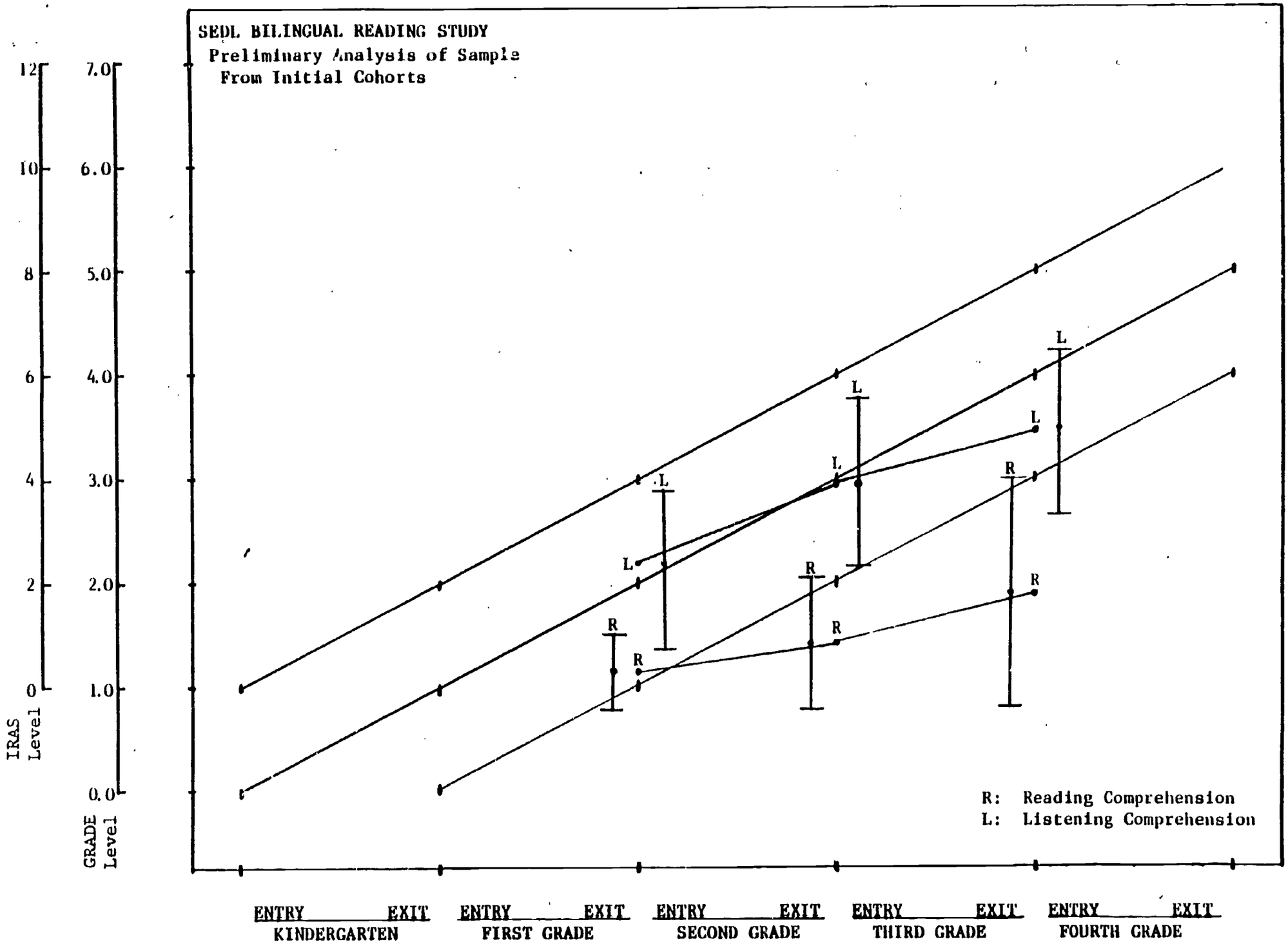


Figure 7. Average growth track data for Reading Comprehension and Listening Comprehension as measured by IRAS in Spanish.

of first grade, and increases only slightly over the next two years. The students in this sample are able to define Spanish vocabulary, and have learned to decode words when presented one at a time in isolation. They cannot handle connected text; fluent reading and the automaticity required for comprehension have not been acquired by this sample of students--not in Spanish.

Analysis of individual protocols. The aggregate data presented above provide a characterization of the general trends in the data, but give little insight into the performance of individual students--these will be examined in this section. Four groups of students will be presented below, each group is distinctive from the other because of the sequence of teachers who taught them from first through third grade, and the large variations in the instruction received from these teachers.

The first two protocols, Figures 8 and 9, are for students in Group A. These two students excel on the Definitions task in both English and Spanish; they also do extremely well on both of the Decoding tasks (the low level for Student 0007 on the Synthetic Decoding Task in second grade appears to be an unexplainable outlier). Comprehension in English is at or above grade level for both students; the levels for Reading and Listening are fairly close. Comprehension in Spanish is lower than in English, varies over the two students, shows a tendency to increase from first to third grade, but is much below the levels of Definition and Decoding.

The next protocol, Figure 10, is for a single student selected as representative of Group B. The level of performance on the Definition task appears relatively high at the end of first grade (actually, the student is at the average for all students), and remains constant after that. This pattern holds for both English and Spanish. Decoding skills and Comprehen-

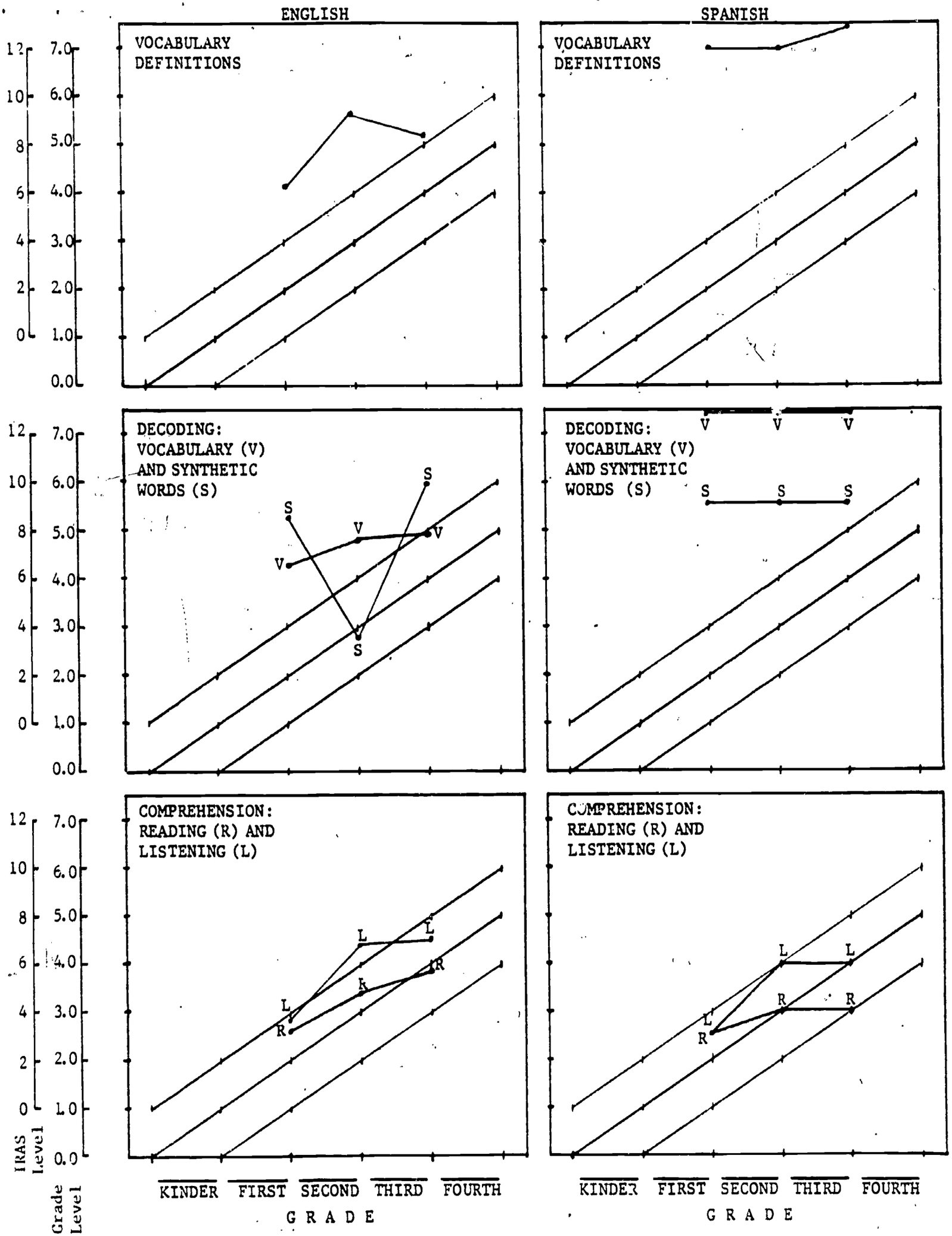


Figure 8. IRAS performance profiles in English and Spanish for student 0007 (Group A).

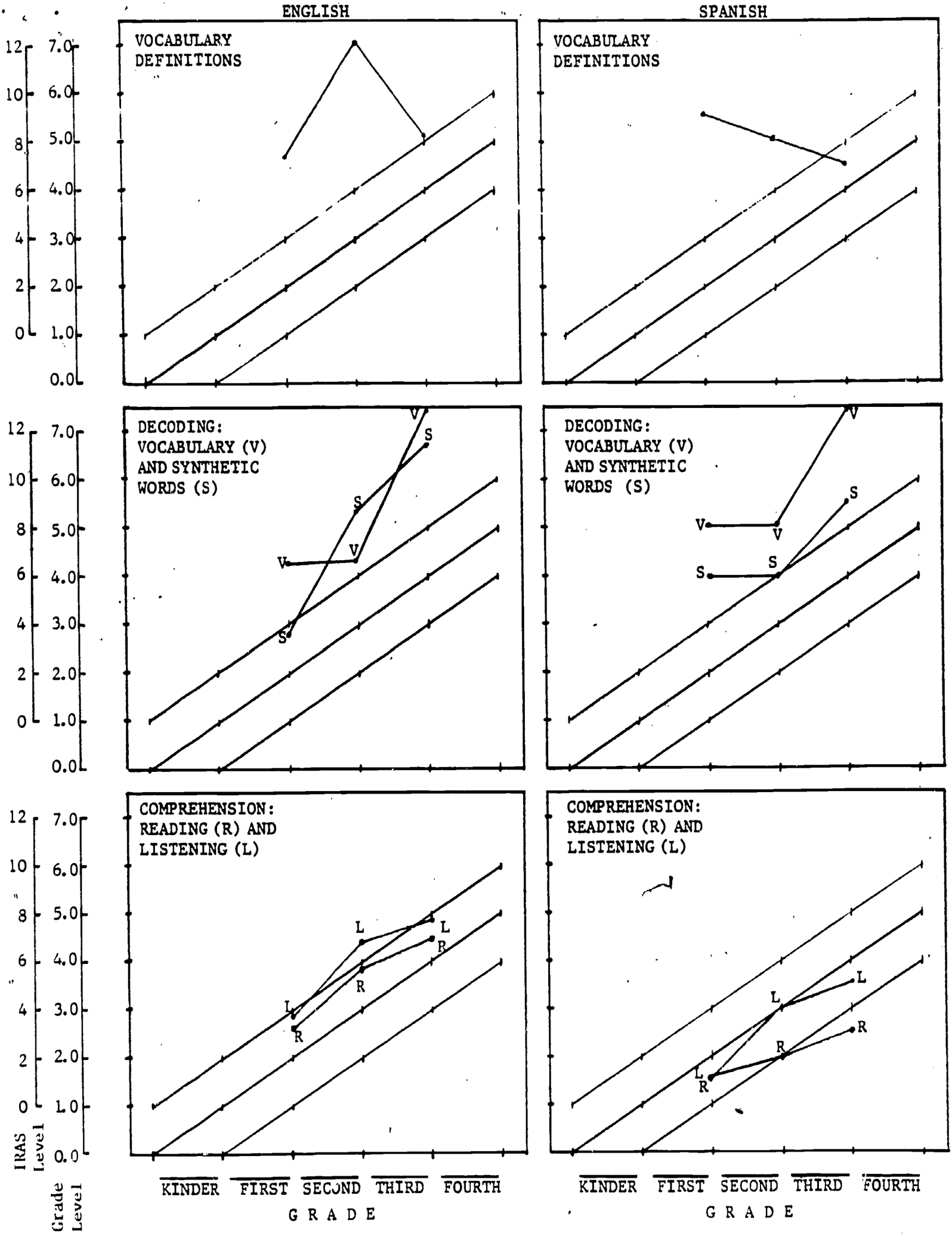


Figure 9. IRAS performance profiles in English and Spanish for student 0052 (Group A).

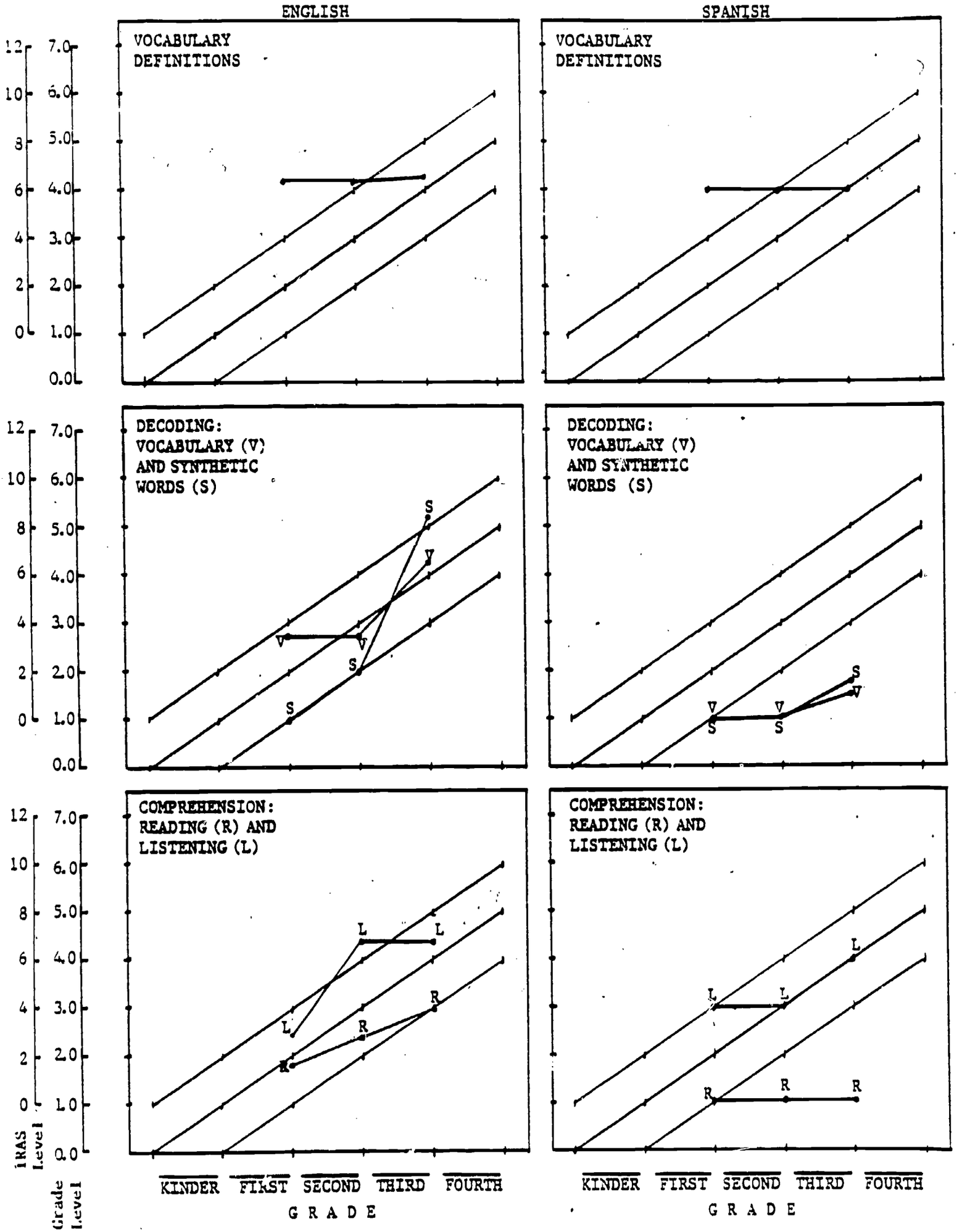


Figure 10. IRAS performance profiles in English and Spanish for student 2014 (Group B).

sion performance increase over the grades, somewhat below the expected rate, but still a noticeable amount of growth.

The protocols for the two students from Group C (Figures 11 and 12) may appear rather confusing at first glance. Student 2097 appears to have no command of English through the end of second grade, based on the performance on the Definition task, whereas Student 2082 is quite fluent in English by the end of first grade. Both students have a reasonable command of Spanish vocabulary by the end of the first grade. The data on English Decoding are quite consistent for both youngsters--neither could decode anything on IRAS, whether familiar or synthetic words--through the end of second grade. Both students showed marked improvement in their skills at decoding English words between the end of second grade and the end of third grade. Neither youngster had much success in decoding Spanish through the end of second grade, but both showed some gain during third grade, especially student 2097. Finally, both youngsters had considerable difficulty in comprehending spoken English text at the end of first grade. Reading Comprehension in both languages was negligible through the end of second grade, with the two youngsters exhibiting variable degrees of success in English Listening Comprehension, and doing reasonably well in Spanish Listening.

The patterns are complex, but a plausible theme can be constructed. The students differ markedly in their entry language capability, and one might suspect that they differed on entry to school (a fact confirmed from the language measures collected). In any event, nothing happened during the first and second grade that allowed either student to attain mastery of reading skills of any sort in either language--both children remained illiterate after the first two years of primary-grade instruction. During

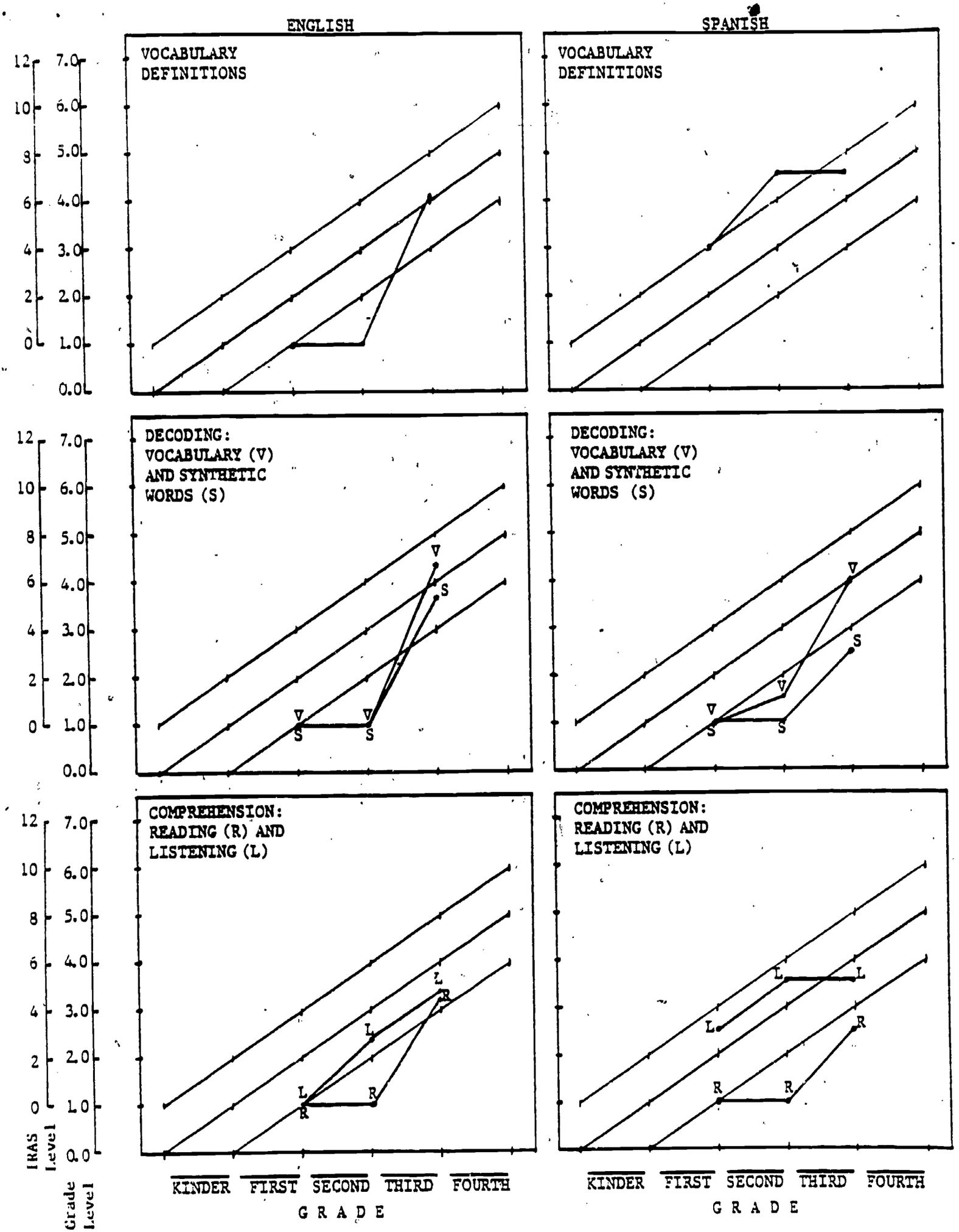


Figure 11. IRAS performance profiles in English and Spanish for student 2097 (Group C).

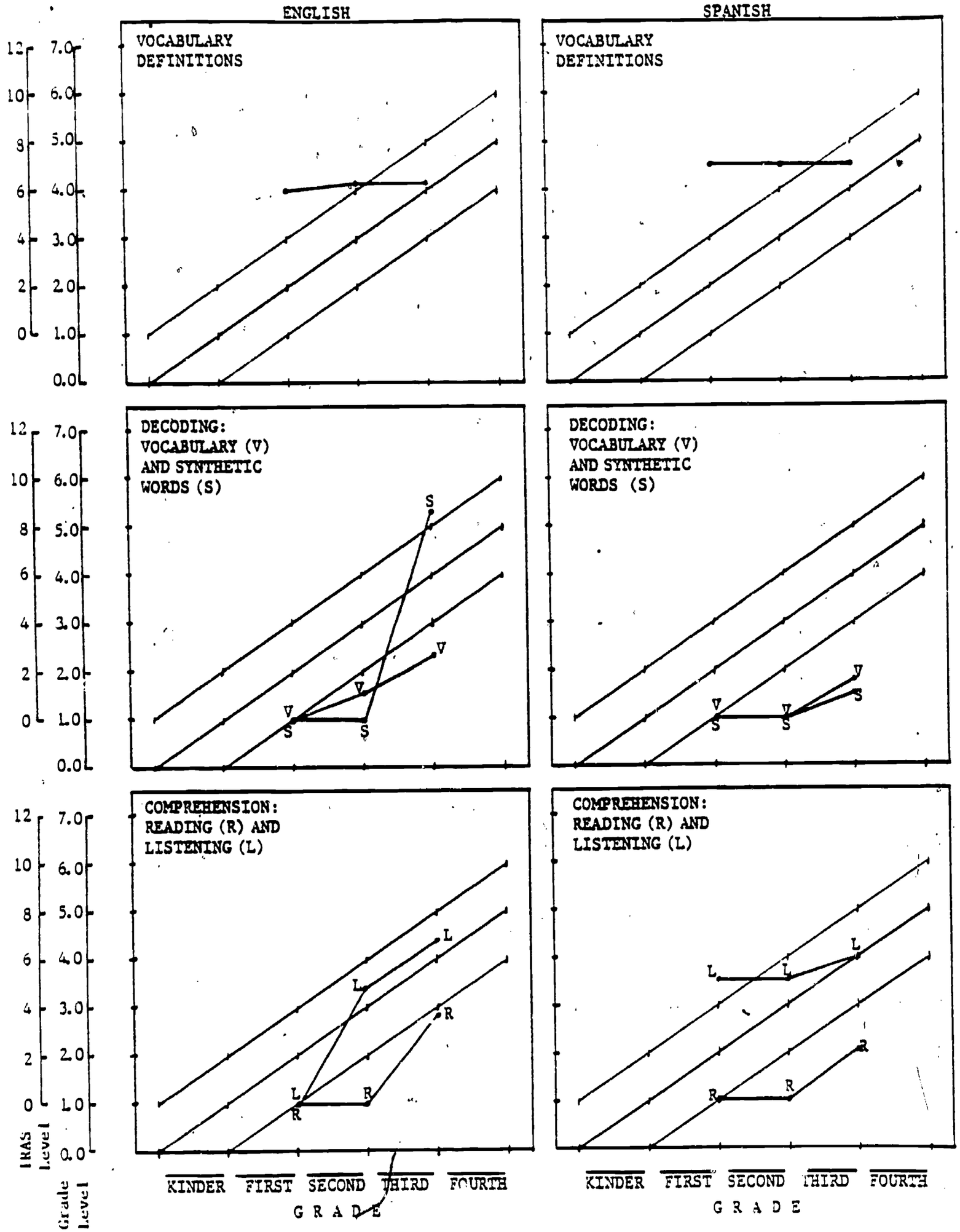


Figure 12. IRAS performance profiles in English and Spanish for student 2082 (Group C).

third grade, both students made substantial gains in English literacy, to the extent that they are reasonably close to the grade level represented by the IRAS design. There is also some evidence of progress in Spanish literacy, slightly less than that observed in English.

Three students have been selected to represent Group D (Figures 13 to 15). The first two students reflect a similar pattern of growth in English reading skills--from the end of first grade on, they are making satisfactory progress in all of the components of reading that are included in the IRAS design, approximating or exceeding the levels indicated by IRAS as appropriate for their grade assignment. Their scores on the Spanish IRAS are also high, with the exception of the second grade Synthetic Decoding score for student 0044, and the typically depressed performance in Reading Comprehension.

The third student in the series shows a totally different pattern of performance. For both the English and Spanish scales, progress is apparent in the oral language skills, and in "sight word" vocabulary; the student is able to define most words in the IRAS series, can comprehend passages that are recited by the tester, and can pronounce familiar words appropriate to grade assignment. However, the youngster shows no evidence of having acquired any skills in phonic analysis by the end of third grade, and the ability to read and comprehend connected text is much below grade level. This student is obviously bilingual, given the level of success on the Definition and Listening tasks, but shows no sign of acquiring literacy in Spanish.

In a later section of this paper, we will examine the sequence of instructional programs provided to the four groups of students whose reading performance has just been described. In general, there is a fair

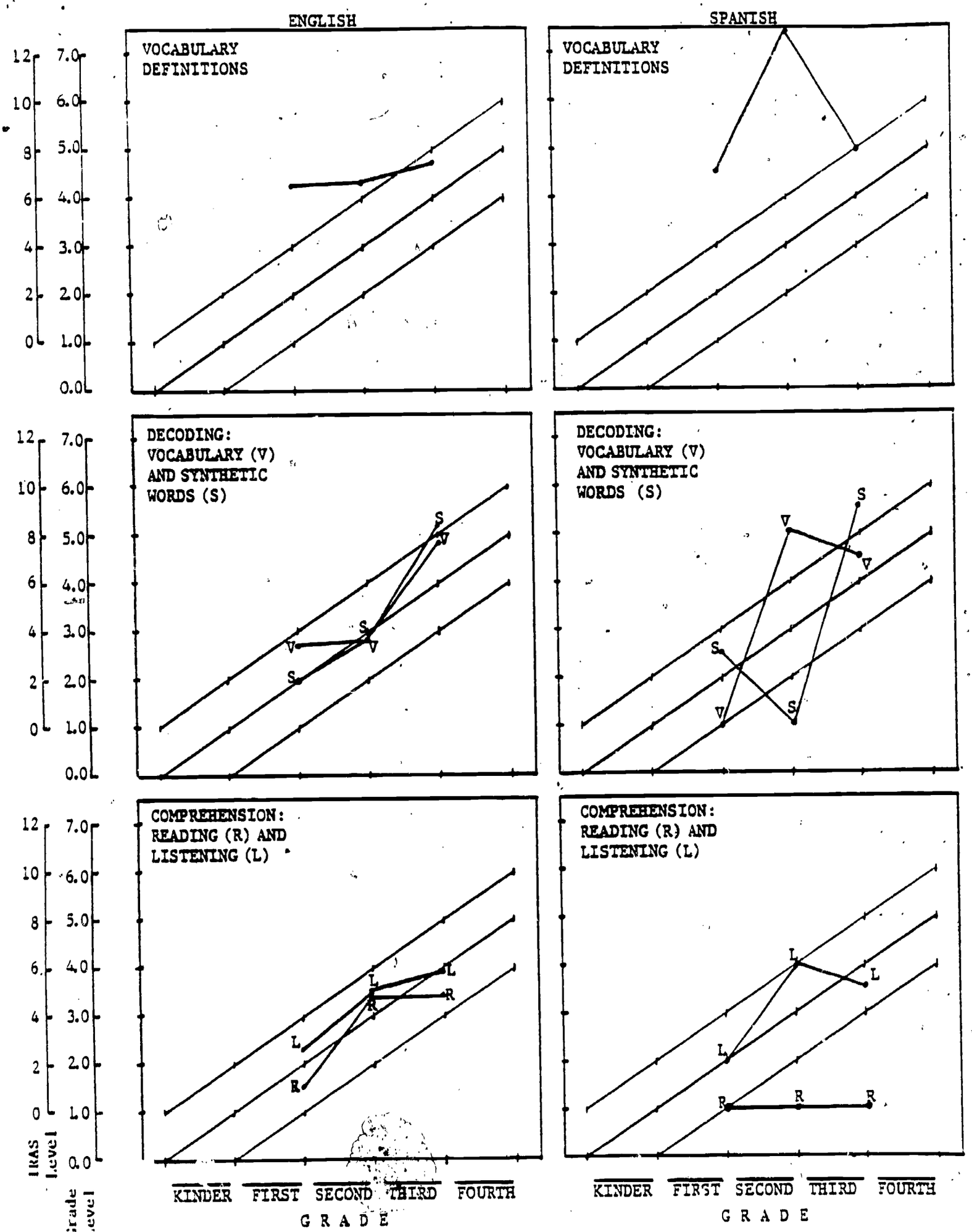


Figure 13. IRAS performance profiles in English and Spanish for student 0044 (Group D).

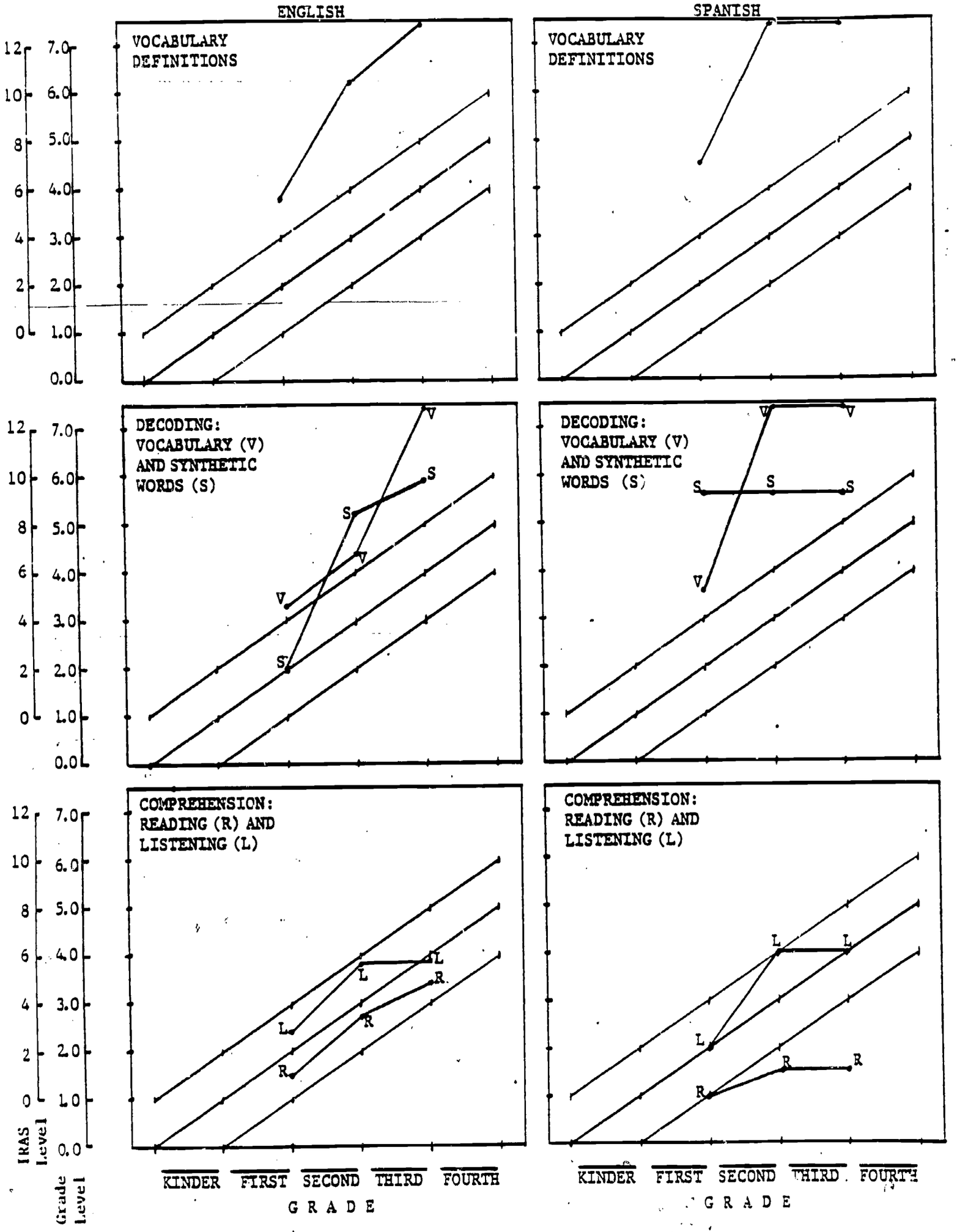


Figure 14. IRAS performance profiles in English and Spanish for student 0251 (Group D).

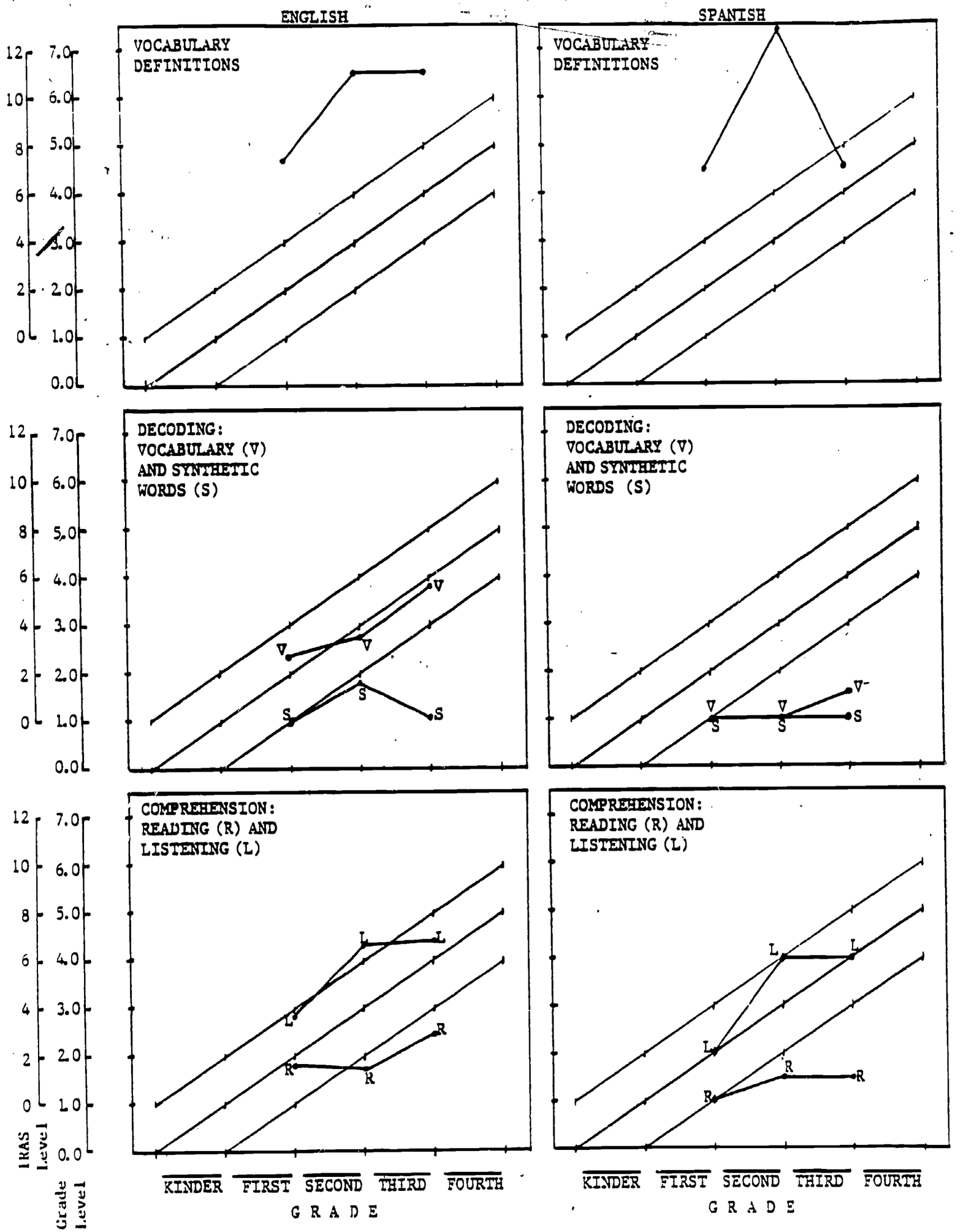


Figure 15. IRAS performance profiles in English and Spanish for student 0298 (Group D).

amount of consistency in the performance of students within a given instructional sequence--this generalization is based upon an "eyeball" analysis of the sample of students selected for assessment in each school, and will be assessed by a more formal, forthcoming analysis. The patterns are often quite distinctive, suggesting that the teacher's decision to emphasize one aspect of literacy over another can have noticeable effects on student achievement. The range of patterns in the series of figures just presented does make a point that has been confirmed by a correlational analysis--the components of reading as measured by IRAS do not exhibit the high degree of collinearity typical of most reading tests. A student may do quite poorly on one facet of IRAS, yet do quite well on other facets.

The relative independence of the IRAS components raises some problems for analysis, but the instrument also has the capability of assessing student responses to differential emphasis in the curriculum, a point that will be made more clearly as we report on the instructional program in the section which follows. Finally, though there is a fair degree of consistency in students' responses to instruction, individual differences are also observed, as seen in the three students in Group D. This point will be raised again in presentation of the instructional profiles.

Observation of Classroom Instruction

In this section we will describe the procedures used in observing classroom instruction, and the preliminary analyses of this data set for the sample of teachers providing instruction for the subsample of students in Cohorts I and II (13 first-grade teachers, 8 second-grade teachers, and 8 third-grade teachers).

Reading and Mathematics Observation System. The reading instruction period of each classroom in the Bilingual Reading Study was observed an

average of five times during the year that a target student was enrolled in the class, with each observation lasting from 45 to 60 minutes. The Reading and Mathematics Observation System (RAMOS, Ca ee & Calfee, 1976) was the instrument used to record the observations. RAMOS is a real-time, categorical system, where for each of the instructional groups in the classroom, the observer notes the significant changes that take place over time in each of several categories considered to be significant indicators of effective instruction. The categories selected for this preliminary analysis include:

- ROLE -- tracked the teacher's involvement in direct teaching; from direct instruction, to discussion, to facilitation, to non-instructional engagement such as management or the preparation of materials.
- LANGUAGE OF INSTRUCTION -- at one end of this continuum, instruction was entirely in English, while at the other extreme only Spanish was used.
- DECODING -- documented the relative curriculum focus on decoding at a given moment from emphasis on analytic phonics skills (such as letter-sound recognition, spelling pattern recognition) to integrative skills, such as whole word recognition, to non-decoding skills such as auditory discrimination, visual discrimination, and letter recognition.
- COMPREHENSION -- documented the relative curriculum focus on comprehension activities, from emphasis on major ideas and making inferences, to literal facts, to vocabulary enrichment, to non-comprehension activities.

- TECHNIQUE -- the instruction may have emphasized global features of a topic followed by analysis (whole-to-part), or begun with an analytic strategy followed by integration (part-to-whole).
- TASK -- the work assigned to the student may have entailed activities directly related to the formal treatment of language (writing, discussing, listening, or reading), or it may have had no immediate relevance to formal language (handling art materials).
- MATERIALS -- the materials for instruction could have been books or book-related media, or could have had no direct relevance to the print medium (art material, picture cards).
- PRODUCTIVITY -- throughout the observation, the observer continuously rated the productivity of each group as high, medium, low, or none.
- NOISE -- a judgment of the amount of noise for each group in the classroom was also made, again, from high to none.

This abbreviated description is intended only as a sketch of the observation system; the categories available to the observer under each of the headings listed above were quite extensive, providing the observer with relatively concrete guides to the appropriate codes.

Analysis of average scale values for RAMOS. In its original form, RAMOS resembles a narrative of the events in the classroom. A moment-to-moment classification of each event for each group is available in an abbreviated code, which can be read by an experienced observer, but which is not immediately "understandable" by the computer. Accordingly, a PASCAL program was prepared to convert the coded format into an expanded format, in which the codes for each group for every minute of observation were presented in a line-by-line record. This expanded record was then used to obtain (a) the average value for each category over time for each group in

the classroom for each observation, (b) weighted averages taking group size into account for the classroom as a whole, and finally, (c) an average for each category for every teacher (collapsed over observations). These aggregate data are subject to the same limitations discussed earlier in connection with student achievement measures.

The average scale values for each of the RAMOS categori- are shown in Figure 16 for the 29 first through third grade classrooms. As indicted in the figure, the classifications used by the observers for each category were arranged on a unidimensional numerical scale, generally ranging from 1 to 9. The scaling was based on the judgment of the Laboratory staff and consultants experienced in classroom instruction. The figure shows that some of the categories changed little or not at all over grades (e.g., Language), whereas other categories changed rather markedly (e.g., Comprehension, Task, and Materials).

These averages are presented primarily as a frame of reference for the group protocols to be described shortly, and are of limited generality because of the restrictions on the sample. However, certain trends in the data deserve mention. Teachers adopted a role of direct instruction less than two-thirds of the time in these classes--the typical role was slightly more active than "facilitation," but not much more so. English was the predominant language at all grades. Decoding was not greatly in evidence, less than 20 percent of the time. Comprehension-like activities were more common, especially in second and third grades. (The Decoding and Comprehension scales can be added together for a rough estimate of the amount of emphasis on these two components of reading.) As can be seen under Focus, only about half of the time was spent with an emphasis on text-based instruction, with a noticeable increase from first to second grade. The

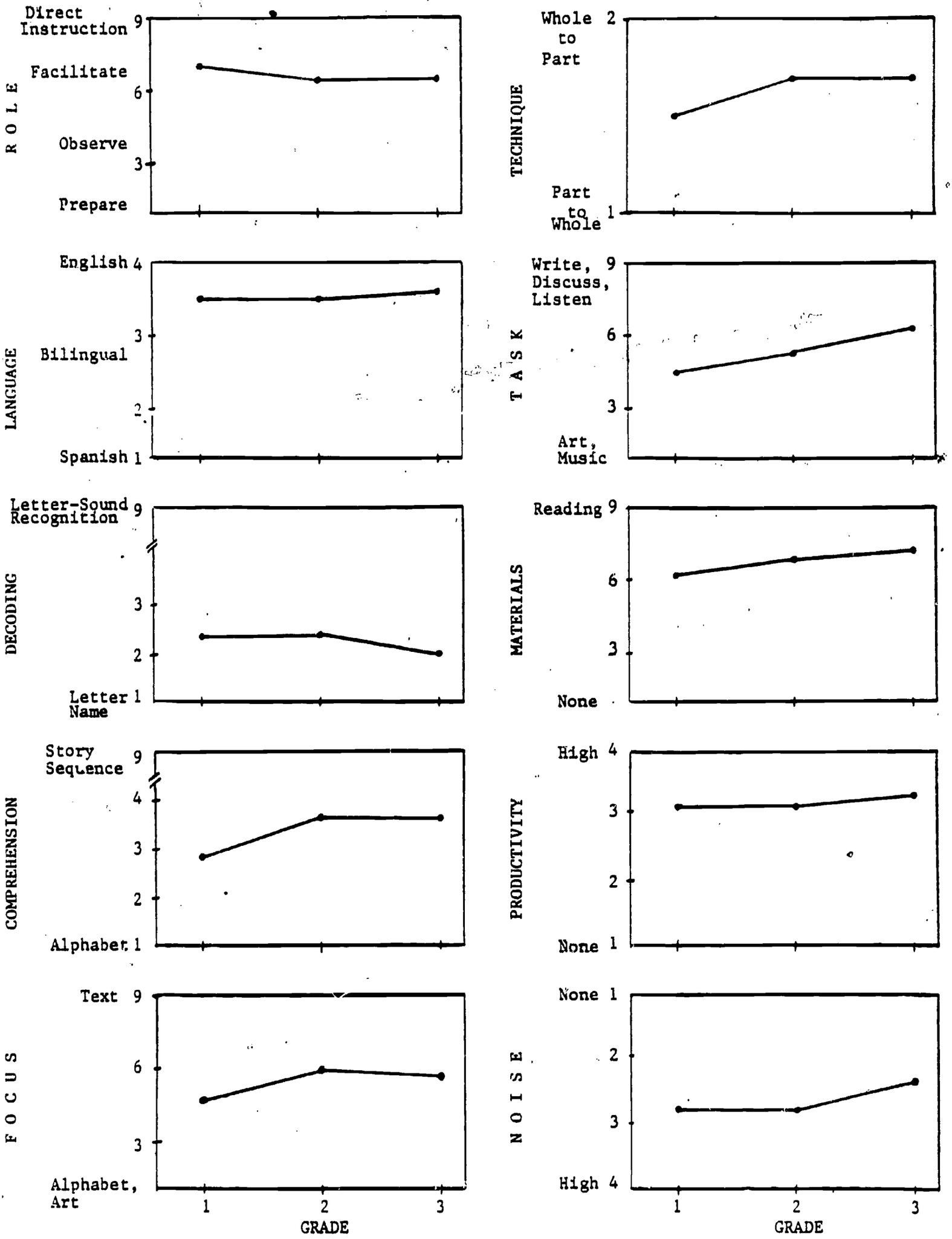


Figure 16. Average scale scores for RAMOS categories at first through third grade.

next three scales also increase from first grade onward--the trend is toward more whole-to-part instruction, toward greater formality in the tasks presented to the students, and toward greater reliance on formal textual materials. The final two panels indicate that students were reasonably productive on the average, and that the noise level was judged to be slightly high in first and second grade, becoming quieter in third grade.

As a caution, these average values reveal nothing about the trends over the school year, nor do they show anything about within-class variations from group to group--these analyses will be forthcoming. The averages also conceal the differences between classrooms, which turn out to be rather substantial in some instances, as we shall see next.

Analysis of instructional sequences. In Figure 17 are the observational profiles for the four groups of students whose achievement data were presented earlier in the report. For three of the groups, classroom data were available for all three grades; the second grade data for Group A have yet to be analyzed. Two subsets of scales--Performance/Noise and Focus/Technique/Materials/Task--were moderately correlated in this data set, and have been combined by means of average standardized scores to simplify the exposition.

We will now attempt to translate the instructional sequence for each group of students into descriptive prose. First grade for Group A entailed a high degree of direct instruction, concentration on English, an average level on the combined FTMT scale (first grade classrooms had an overall z-score of -.5 on this scale), a quiet and productive environment, an average amount of time on decoding, and more than average emphasis on comprehension. The overall picture is one of a well-managed classroom with a well defined focus on the acquisition of English literacy. Data on the

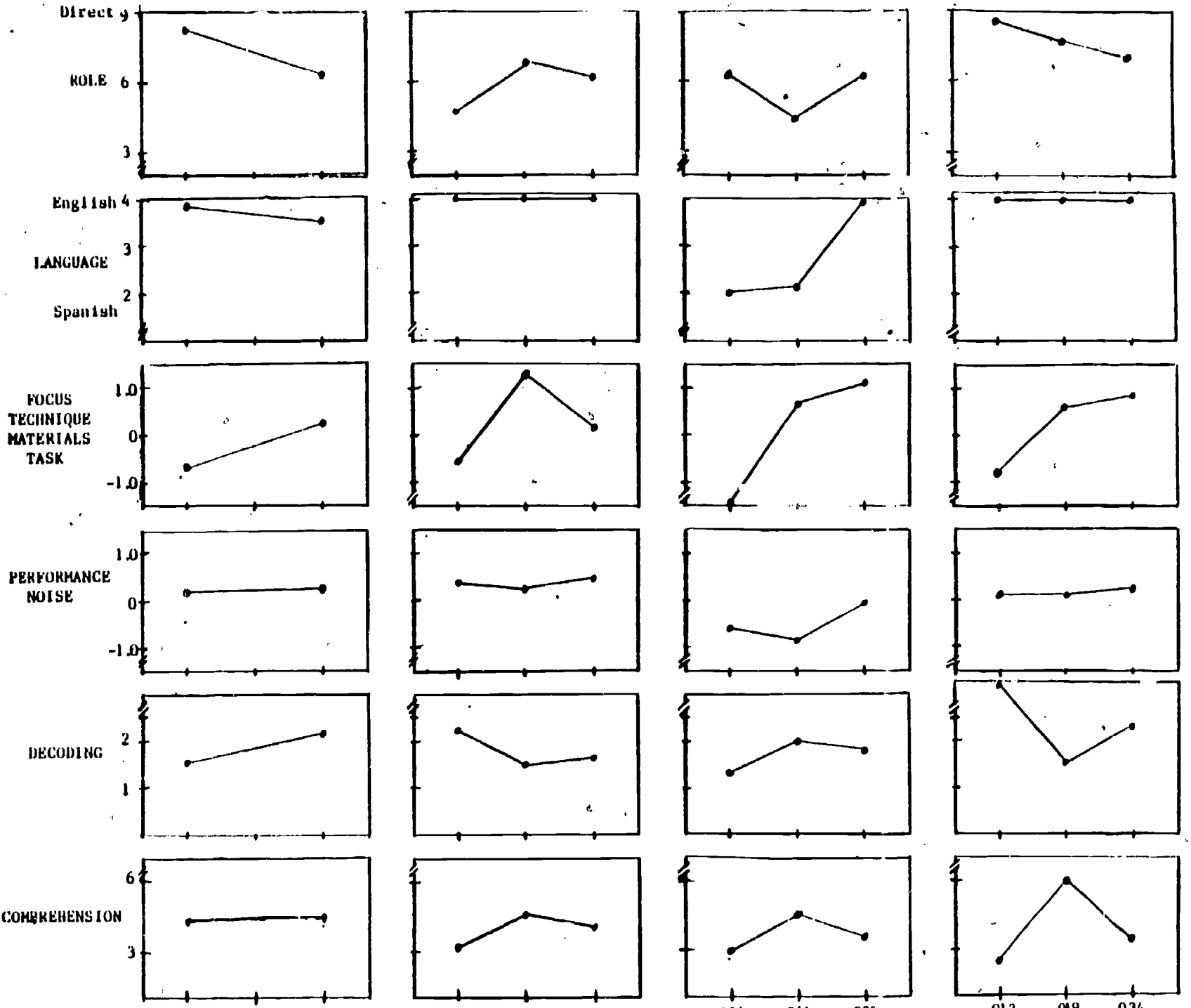


Figure 17. Individual instructional sequences for Groups A through D.

second grade program are not available for this group of students, but the third grade program tends to follow a pattern similar to the first grade class, with a few noticeable differences. The third grade teacher relied less on direct instruction, resorted somewhat more often to Spanish, and gave more than usual emphasis to decoding.

The instructional sequence for Group B reveals relatively less reliance on direct instruction in first grade, with average levels thereafter. Instruction was entirely in English. The FTMT scale, which provides an index of the extent of formality in the program, is about average in first and third grades, with special emphasis in second grade. The classrooms were relatively quiet and productive, with moderate amounts of time devoted to instruction in decoding and comprehension. Overall, the program for Group B was well-managed and focused on reading instruction--an emphasis on English literacy to the complete exclusion of instruction in Spanish.

Group C differs from the previous two groups in several respects. The instructional level varies from average (first and third grades) to low (second grade). Instruction was predominantly in Spanish during the first two grades, after which English was used in third grade. The FTMT scale was extremely low in first grade, implying little or no emphasis on formal aspects of language and text. This scale was at or above average in second and third grades. The class was noisy and unproductive in both of the first two grades, rising to an average level in third grade. The curriculum included attention to both decoding and comprehension at all grades. The overall picture shows a lack of coherence and management in the first two years, followed by a reasonably focused program in third grade. The low productivity and the high noise in first and second grades suggest poor

management. In first grade the teacher was moderately active instructionally, but gave little attention to the topic of literacy in either English or Spanish. The second grade teacher gave more emphasis to literacy, but played a relatively passive role in the classroom. Only in third grade were all the elements of an effective instructional program brought together.

Finally, let us examine the pattern for Group D. The primary years for the students in this sequence were characterized by a high level of direct instruction, exclusively in English, in classrooms that were productive and relatively quiet, and with considerable emphasis on decoding and less attention to comprehension. The level of formal language in first grade was slightly below average, but increased sharply during the second and third grades.

The patterns in Figure 17 are chiefly of interest to the degree that they can be related to student achievement, but two general reactions merit some attention. First, the sequence of instructional activities varies considerably from one group to another. Most research on teaching has focused on a single slice-of-time in the life of the student--a day, week, month or school year. The data in Figure 17, to the degree that they are valid representations, suggest that the course of instruction for the individual student may vary quite a bit from one year to the next, in ways that reflect little in the way of a coherent school-wide program.

The second reaction is to the variations in the specifics of the profile from one classroom to another. The data in Figure 17 cannot be summarized by the contrast between "good" and "bad" classroom programs. The 11 teachers represented in the figure (more precisely, the 11 teacher-year events), vary more or less independently on the set of dimensions incorpo-

rated in the RAMOS protocols. Again, this particular feature of the RAMOS instrument requires a more complicated plan of analysis, when contrasted with instruments that focus on one or two dimensions of classroom instruction (e.g., instructional time, patterns of verbal interaction, or the like). The multidimensional character of the RAMOS data structure increases our confidence in the validity of the instrument, however, because it seems to us a reasonable conjecture that the instructional programs of classroom teachers actually vary on more than a single evaluative dimension.

Measuring the Linkage Between Instruction and Achievement

Assessing the degree of correspondence between the complex patterns represented by Figures 8 to 15 and Figure 17 poses some interesting challenges. On the one hand, an eyeball approach has much to recommend it--human beings are quite capable of perceiving complicated relations in the midst of noisy environments. On the other hand, there is much to recommend procedures that are quantifiable and technically reproducible. Our approach in the Bilingual Reading Study has been to rely on experienced judgment to carry out preliminary evaluations, and to explore one or two methods for quantification. This work is still ongoing, and in this section we will only briefly describe the relations we are seeing.

The general picture that emerges from the results is fairly simple--the target students tended to show higher levels of reading achievement when instruction emphasized the more formal aspects of language, and when the classrooms were well managed. Groups A and B illustrate some variations on this theme, and the students within these classroom programs appear quite competent in all components of English reading. Group A included some time for Spanish reading (most attention given to decoding,

it would appear, though the analyses done thus far do not provide evidence on this point), and the students showed the benefits, compared to the Group B student, who received no instruction in Spanish reading, and showed no gains in Spanish literacy.

When the classroom gave less attention to formal reading, or when the teacher was unable to maintain control over the students, there was less evidence of learning. Students in Group C were illiterate in both English and Spanish at the end of second grade. Despite three years of instruction, they showed no command of print--for practical purposes, they had learned nothing about reading during the primary grades. Student 2082 showed some gain in English listening comprehension during second grade, but otherwise both students performed like entering kindergartners on exit from second grade. Both students were reasonably facile in Spanish; student 2097 was monolingual in Spanish on entry to kindergarten, and remained so through second grade. Inspection of the instructional program for these students shows little evidence of a coherent effort to teach reading in either language. While some time was allocated to decoding and comprehension, books and the other "stuff" of reading were generally present, the classroom was noisy and the students unproductive.

The data for Group C show another effect that seems to us deserving of emphasis--a formal program of reading instruction can have positive effects on reading achievement even for students who have not been exposed to such instruction in previous years. Students in Group C responded to the third-grade program of instruction as though they were "first-order Markov chains," to borrow a term from probability theory--both of the target students showed substantial gains in English decoding and comprehension during third grade, even though they had not been taught to read during the first

and second grades. To put it another way, these students showed no evidence of a cumulative deficit, nor does it appear that they were unable to learn because they had missed a "critical period" in reading acquisition.

The data from Group D reveal two trends of potential importance. These students were in a program that emphasized English decoding in first grade, followed by second-grade instruction that stressed comprehension in both languages. The achievement patterns for students 0044 and 0251 seem to reflect the instructional emphases--decoding skills are at or above expectation at the end of first grade, but comprehension is negligible; reading comprehension in English increases markedly during second grade, during which time student 0044 shows no further development of decoding skills, although student 0251 does show considerable growth in this area. In third grade, the curriculum emphasis shifts once more, from comprehension back to decoding, and once again the changes in student achievement seem to mirror this shift in relative emphasis. The eclecticism of present practice makes it difficult in most instances to draw sharp contrasts, but in this one instance the IRAS/RAMOS combination seems to be working as planned.

The second point to be noted in the data for Group D is the difference between the response of student 0298 and the performance of the other two students. The observational data in Figure 17 are averages over all the instructional groups in the classroom. While these patterns serve for an overall characterization, the actual program for individual students may depart significantly from the average for the classroom as a whole. The RAMOS data can be analyzed at two additional levels of refinement--the instructional group to which each target student is assigned, and departures of the individual target student from the profile for the group.

Departures of the latter sort were fairly uncommon, but our informal review of the observational data suggests that there are substantial differences between reading groups within a class. It is clear that student 0298 differs markedly in achievement from the other two target students in Group D; our next step in accounting for such discrepancies will be to examine the instructional program for the group to which individual target students are assigned.

References

Calfee, R., & Calfee, K. Reading and mathematics observation system - RAMOS/II (rev.). (1976). [Unpublished manuscript]. Stanford, CA: Stanford University.

Calfee, R., & Calfee, K. Interactive reading assessment system. (1979). [Unpublished manuscript]. Stanford, CA: Stanford University.

Calfee, R., & Calfee, K. Interactive reading assessment system (rev.). (1981). [Unpublished manuscript]. Stanford, CA: Stanford University.

Calfee, R., Calfee, K., & Peña, S. Interactive reading assessment system - Spanish. (1979). [Unpublished manuscript]. Stanford, CA: Stanford University.

Calfee, R., & Curley, R. Structures of prose in the content area. (1979). In J. Flood (Ed.), Understanding reading comprehension. Newark, DE: International Reading Association.

Carroll, J., Davies, P., & Richman, B. Word frequency book. (1971). Boston: Houghton Mifflin.

Chall, J. The great debate: Ten years later, with a modest proposal for reading stages. (1979). In L. Resnick and P. Weaver (Eds.), Theory and practice of early reading. Hillsdale, NJ: L. Erlbaum Associates.

Kintsch, W., & van Dijk, T. Toward a model of text comprehension and production. Psychological Review, 85, 363-394.