College Board Report No. 99-2

Developing a Portfolio Assessment: Pacesetter® Spanish

ANDREA FERCSEY, CARMEN LUNA,
EVA PONTE, and PABLO ALIAGA

The College Board

College Board Report No. 99-2

Developing a Portfolio Assessment: Pacesetter® Spanish

ANDREA FERCSEY, CARMEN LUNA,
EVA PONTE, and PABLO ALIAGA

College Entrance Examination Board, New York, 1999

Andrea Fercsey is an assessment specialist at ETS.
Carmen Luna is a group leader at ETS.
Eva Ponte was a 1996 summer intern at ETS and a graduate student at the University of California at Berkeley.
Pablo Aliaga was a 1997 summer intern at ETS and is a graduate student at the University of Michigan.

Founded in 1900, the College Board is a not-for-profit educational association that supports academic preparation and transition to higher education for students around the world through the ongoing collaboration of its member schools, colleges, universities, educational systems and organizations.

In all of its activities, the Board promotes equity through universal access to high standards of teaching and learning and sufficient financial resources so that every student has the opportunity to succeed in college and work.

The College Board champions—by means of superior research; curricular development; assessment; guidance, placement, and admission information; professional development; forums; policy analysis; and public outreach—educational excellence for all students.

Additional copies of this report (item #200272) may be obtained from College Board Publications, Box 886, New York, New York 10101-0886, (800) 323-7155. The price is $15. Please include $4 for postage and handling.

Printed in the United States of America.

Acknowledgments

We would like to acknowledge the six Pacesetter® Spanish teachers who participated in Phase I of this research project: Eva Chapman and Ruth Ann Kerkoc, Leland High School, San Jose, California; Gilda Nissenberg, North Miami Beach High School, Florida; Dianne Michel Stenroos, Rutland High School, Vermont; Carol Thorp, East Mecklenburg High School, Charlotte, North Carolina; and María Elena Villalba, Miami Palmetto Senior High School, Florida.

We would also like to acknowledge the following Pacesetter Spanish teachers who participated in Phase II of this research project: Esperanza Cobián-Bashara, Silver Creek High School, San Jose, California; Sharon Dyrland, Surrattsville High School, Clinton, Maryland; Rosalba García, Bellaire High School, Bellaire, Texas; Ruth Ann Kerkoc, Leland High School, San Jose, California; Dennis Lavoie, Fairport High School, Fairport, New York; Gilda Nissenberg, North Miami Beach High School, Miami, Florida; Nereida Samuda Zimic, William B. Travis High School, Austin, Texas; Carol Thorp, East Mecklenburg High School, Charlotte, North Carolina; and María Elena Villalba, Miami Palmetto Senior High School, Miami, Florida.

The contributions and efforts of these individuals were essential throughout all stages of the project.

The authors would like to thank Carol Myford, Rick Morgan, and Gita Wilder for their help in interpreting and analyzing data, as well as for their review of an earlier draft of this paper.

Contents

Abstract 1

PHASE I REPORT 1

Introduction 1

Background 1
- *The Pacesetter® Program* 1
- *Pacesetter Spanish* 2
- *Portfolio Assessment* 2

Pacesetter Spanish Portfolio Research Project (Phase I) 4
- *Method* 4
- *Portfolio Analyses* 7
- *Description of Elements* 7
- *Distribution of Ratings* 7
- *Time Used to Read Portfolios* 8
- *Culminating Assessment* 9
- *Portfolio Assessment Matrix Correlations* 10
- *Portfolios Read Twice: Reliability Study* 11
- *Third Rater Analysis* 15
- *Selection of Benchmarks* 15

Summary of Findings 16

Recommendations 16

References 16

Phase I Appendix 37

Tables
1. Distribution of Ratings 7
2. Distribution of Ratings by Aspects and Dimensions 8
3. Time Used to Read Each Portfolio 9

3A. Summary 9
4. Average Time Employed to Read Portfolios From Different Teachers 9
5. Average Time Employed by Each Rater 10
6. Time Employed to Read Portfolios at Different Sessions 10
7. Culminating Assessment and Dimension 2 Portfolio Ratings 10
8. Portfolio Assessment Matrix Correlations 11
9. Summary of Portfolio Score Discrepancies 13
10. Summary of Portfolio Score Discrepancies for Each Aspect in Each Dimension 13
11. FACETS OUTPUT (7.1.1) Students Measurement Report (Arranged by FN) 14
12. Distributions of Raters' Judgments 15
13. Portfolios Read Three Times 15

Figure
1. FACETS map for the whole scale (21 portfolios, two readers) 12

PHASE II REPORT 17

Rationale and Expected Outcomes 17

Pacesetter Spanish Portfolio Research Project (Phase II) 18
- *Method* 18
- *Summary of Ratings* 19
- *Time Used to Read Portfolios* 21
- *Distribution of Ratings for Each Rater* 21
- *Portfolios Read Twice* 22
- *Correlation Among the Ratings of Portfolios* 27
- *Reliability Study* 27
- *Analyses* 29
  - *Students* 29
  - *Raters* 33
  - *Aspects and Dimensions* 33

Summary of Findings 34

Recommendations 34

References 35

Phase II Appendix 67

Tables
1. Frequency of Ratings in Each Rating Category 19
2. Frequency of Ratings in Each Category for the Aspects of Dimension 1 20
3. Frequency of Ratings in Each Category for the Aspects of Dimension 2 21
4. Time Employed by Each Rater in Minutes 22
5. Total Count and Relative Frequency by Category and Rater 23
6. Comparison of Portfolios and Raters 24
7. Discrepancies Between Raters 26
8. Specific Discrepancies Between Raters 26
9. Bivariate Correlation Matrix 27
10. Rater Severity 29
11. Percentage in Each Scale Category 33

Figures
1. FACETS: Partial credit model for raters 28
2. FACETS: Partial credit model for aspects 30
3. FACETS: Rating scale model 31

3A. FACETS: Student adjusted scores 32

Abstract

Portfolios are one of the assessment tools used in Pacesetter® Spanish. In Phase I of this study, conducted in 1995–96, a first attempt was made to develop a standardized portfolio assessment system. As part of this system, a set of guidelines and an assessment matrix were prepared. The assessment matrix was piloted during a portfolio reading in which six teachers participated. At the end of the reading, some portfolios were identified as benchmarks to be used for training purposes.

After the reading, results were compiled and analyzed. Data were collected to investigate several aspects of the portfolio assessment system (i.e., scores, the reliability of the raters, correlation with the Culminating Assessment, and time needed to read portfolios). Our purpose was to determine which components of the portfolio assessment system were working as intended and which might need to be changed. To this end, we explored ways of gathering preliminary evidence of the reliability and validity of the portfolio assessment system. Our findings suggested that refinement of the matrix was needed as well as clearer guidelines for the selection of artifacts to include in the portfolios.

In Phase II of our research, in consultation with participating teachers, a revised set of guidelines and a revised assessment matrix were distributed to all schools implementing the Pacesetter Spanish course in 1997. Seven teachers (three of whom had also been a part of Phase I of this study) participated in a portfolio reading in 1997, at which both the guidelines and the matrix were further refined for immediate dissemination to the schools. A new set of benchmark portfolios was identified, and a set of sample portfolios was prepared for use in subsequent professional development sessions for Pacesetter Spanish teachers.

In general, the results of our findings are encouraging. Initiating national and/or regional portfolio scoring sessions would greatly enhance the validity and reliability of the portfolio component as an integral part of the Pacesetter Spanish program.

Phase I Report

Introduction

The Pacesetter® program developed by the College Board and Educational Testing Service (ETS) is an integrated set of learning outcomes, course materials, and instructional experiences, including various approaches to assessment, and is currently being offered in three subject areas: English, Spanish, and Mathematics.

The Pacesetter Spanish program has provisions for student self-assessment, peer assessment, local assessments, and end-of-course standardized assessment, as well as portfolio assessment. All Pacesetter Spanish students are expected to assemble portfolios with products from their work throughout the course in order to show evidence of what they have learned.

From March to April 1996, the Pacesetter Spanish program conducted a research project on the assessment portfolio component, with the following goals: (a) to prepare a working definition of Pacesetter Spanish portfolios as assessment tools, (b) to develop a standardized portfolio assessment system, and (c) to utilize the research findings for upcoming professional development sessions for teachers. This paper describes the latest developments in these three areas.

Background

The Pacesetter® Program

The Pacesetter program is part of the College Board's efforts to promote achievement of educational excellence for all students. The College Board defines Pacesetter as a school-based instructional assessment program that integrates high academic standards for all students, teacher preparation to help achieve those standards, and a broad range of assessments measuring student achievement and facilitating instructional decisions.

The Pacesetter program is structured around three essential components. First, it includes course frameworks specifying content curriculum, which detail the knowledge and skills students are expected to master. Teachers and students are encouraged to have a shared understanding of what content areas are to be covered throughout the course. Second, it provides professional development sessions for teachers, which include a varied range of activities. Third, the program incorporates several types of assessment, ranging from embedded assessments to standardized assessments.

The College Board will conduct research to examine the program outcomes: how the program prepares students for future careers as well as how it prepares students for other College Board assessments, namely SAT® I: Reasoning Test, SAT II: Subject Tests, and Advanced Placement Program® courses and examinations. As part of program implementation, research will focus on how teachers adopt Pacesetter classroom strategies. Finally, the whole assessment process will be monitored to provide feedback for further revisions to instructional materials and methodology.

Pacesetter Spanish

The aim of Pacesetter Spanish is to provide third-level high school students with a contextualized approach to language mastery. The program integrates learning outcomes, course materials and learning experiences, and various types of assessment.

Because the program embraces a contextualized approach to learning, course objectives go beyond those established by traditional instruction. The three main expected course outcomes are: (a) to use the Spanish language to acquire knowledge, (b) to understand Hispanic cultures, and (c) to use Spanish effectively to communicate with others. These outcomes support the five goals of the National Standards for Foreign Language Learning. They focus as much on the process of acquiring new information and cultural skills as they do on acquiring specific knowledge in specific domains of the language.

Pacesetter Spanish covers a broad range of content. Students learn about the contributions of various Hispanic figures in different fields. The curriculum materials facilitate students' ability to make connections between different aspects of Spanish-speaking cultures, as well as to gain interdisciplinary knowledge. The program has a holistic approach to language skills.

The materials of the course framework are thematically linked. Students read authentic texts, engage in discussions, develop research projects, and experience aesthetic works (music, arts, etc.). Teachers may choose to supplement course materials in different ways, and to use additional sources of information, the resulting products presented as evidence of students' learning.

The role of students in the program varies somewhat from their role in traditional courses. Students are encouraged to work both independently and in groups, to conduct research projects using different resources, and to approach learning from an interdisciplinary perspective (i.e., they use knowledge and inquiry skills from other areas of the curriculum to help them learn Spanish). They also take an active role in the evaluation process by monitoring and evaluating their own performance. Pacesetter Spanish empowers students to be autonomous, self-directed language learners.

Finally, assessment aims at both the processes and the products of instruction. Assessment is integrated into the course, taking alternative approaches beyond traditional tests. Students are active participants in the process, conducting self-assessment through learning logs and journal activities, and through peer assessments. Local assessments are designed by teachers to evaluate student performance and teacher effectiveness. The Culminating Assessment, a standardized component provided at the end of the course, may serve accountability purposes. Students are expected to maintain a portfolio with samples of their work to show evidence of achievement in the various aspects of the dimensions assessed through this component.

These new approaches to instruction demand not only a change in the curricula and the role of students, but also in the role of teachers. Teachers are expected (1) to teach a language course not centered in the language itself, (2) to teach an integrated and interdisciplinary course with a holistic approach, (3) to implement cooperative learning approaches in their classrooms, (4) to focus on process as much as on products, (5) to be able to provide different sources of information, (6) to be able to search for different sources of evidence of students' performance, (7) to teach students different learning strategies, (8) to use the scoring rubrics appropriately, (9) to adequately manage portfolios, and (10) to promote student self-assessment and self-learning.

To attain this competency, participating teachers help define (a) the knowledge and skills necessary for students' success and (b) ways to find evidence of students' learning. These teachers also help create scoring rubrics. To teach Pacesetter Spanish, all participating teachers must attend professional development sessions before the beginning of the academic year and are also expected to attend midyear meetings. Both are designed to give ample opportunities for the participants to exchange suggestions on pedagogical approaches and techniques, as well as to give feedback on the current approach to the content of course materials. It is to be noted that the leaders and trainers at these sessions are themselves teachers implementing the Pacesetter Spanish program in their classrooms.

In the 1996–97 school year, the Pacesetter Spanish program was taught in approximately 25 districts, 60 schools, by 80 teachers, to 4,000 students.

Portfolio Assessment

The United States educational system has faced several problems in the past. Resnick and Resnick (1992) have pointed out the link between testing efforts and educational reforms in this country. This link has focused attention on school assessment practices and its problems. Some assessment alternatives have been formulated by the educational community in an attempt to overcome these problems, including performance-based assessment.

Performance-based assessment involves both instructional processes and products. One of its main features is that students create or construct complex responses and products. Answers can no longer be considered as simply right or wrong. Alternative answers given by

students and teachers use evidence from different sources and media. Because this process generates complex responses, human judgment is required to evaluate students' responses along a continuum of achievement. Performance-based assessment can be linked to and embedded in instructional practice, and it is expected that this link will allow assessment to inform and guide instructional practice.

Portfolio assessment is a special type of performance-based assessment. A portfolio is a selected collection of students' work over time evaluated by one or more raters using clear criteria. In some cases, portfolios may also include students' reflections about their own work and progress. Portfolios, like most types of performance-based assessments, are considered tools to guide instructional practice. If a portfolio assessment system is correctly implemented, instruction should be expected to: (1) pay more attention to the process of learning, (2) lead teachers to think, evaluate, and modify their own practices, (3) change students' traditional role from a passive one to one in which they take charge of their own learning and the evaluation of that learning, (4) be guided by current theories of development and the process of learning, and (5) make explicit what content is valuable for both teachers and students.

To promote this change, portfolios should be more than a pile of schoolwork. Gitomer and Duschl (1995) indicate that good portfolio practice requires fundamental changes in conceptions of (scientific) knowledge, teaching, learning, and assessment. They explain in detail what these changes mean in the area of science education. Some of their claims can be applied to other educational domains. First, curricula should reflect the nature of the discipline being taught; students should use procedures, methods, and practices similar to those used by professionals in the discipline. The dynamics and interactions found in the field should also be an integral part of instruction. Second, it is necessary to take a constructivist approach to learning: to have conceptual change as the goal of learning. Third, assessment should serve the needs of students and teachers as well as others who have a stake in what happens in the classroom. Criteria used to consider student performance should be explicit, and assessment must meet three conditions: (a) it should focus on knowledge and skills considered important within the discipline, (b) it should contribute to instruction and learning, and (c) it should serve accountability purposes.

Arter and Spandel (1992) state that portfolios will only have the desired effects if they are carefully planned. They give a definition of portfolios that includes multiple elements and lists requisites for a good portfolio system. First, it is necessary for users to have a clear idea of the purpose of the portfolio, since different objectives will lead to quite different portfolios. Second, students should participate in the selection of the portfolio content. This selection process demands self-reflection from students, forcing them to analyze their own work and what elements in that work best serve to demonstrate their knowledge and skills. To do this, students should have clear guidelines of what artifacts to select as part of their portfolios. Third, assessment criteria must be fully and carefully defined and open to all. Finally, assessment and instruction should form an integrated whole. Portfolios can monitor students' accomplishments as well as provide for new and better ways for students to improve on their achievements.

Portfolios define the teacher's role in a "new" way: Teachers are considered professional researchers studying their own practice. (This view of teachers was already proposed by Elliot [1985].) This new definition of teaching requires teachers to revise and change their beliefs and approaches to instruction, learning, and assessment, and to rethink instructional goals. Without such changes it would be difficult to attain the goals set by portfolio assessment. Sheingold, Heller, and Paulukonis (1995) identified the following five categories in which teachers reported changes: (1) using new sources of evidence, (2) sharing responsibility for learning and assessment, (3) changing goals of instruction, (4) using new ways of evaluating evidence, and (5) changing the general view of the relationship between assessment and instruction.

With respect to students, portfolios are beneficial in several ways. First, students tend to find them motivating and engaging. Second, through portfolios students become more aware of and responsible for their own learning and are able to make more informed choices about their work and how to approach it. Third, these students—more than students in traditional assessment settings—tend to collaborate with other students and with teachers. Fourth, different approaches and levels of learning are allowed within the portfolio framework; students are given a broader range of opportunities to learn and to demonstrate their learning. Also, because portfolios pay more attention to how student work compares to established standards instead of to the work of other students, a less competitive environment is created. Finally, because portfolio contents tend to focus on what is important to learn, and for what purposes, students are better prepared for work life and life in general.

Despite their promising outlook, portfolios offer no panacea. Several problems may preclude their being as beneficial for education as could be expected, the most important problem one of implementation. In this respect, there is a big gap between theory and practice. Teachers, the principal catalysts of portfolios, must be prepared to perform their new assigned role. They must

be able to (a) clearly define the purpose of instruction and undergo all the above-mentioned changes, (b) clarify complex tasks and provide students with additional sources and media of information, (c) be able to find evidence of students' learning in different ways, making continuous and accurate judgments, and (d) continuously revise and adapt the instructional process. Some teachers find this an enormous responsibility in terms of investment of classroom time, professional development, and scoring. This may explain the reluctance of some teachers to implement portfolio assessment in their classrooms.

However, teachers are not the only ones who may find it difficult to pursue a portfolio assessment model. The general public is also concerned about how reliable, valid, and fair portfolio scores are for accountability purposes. These issues may be even more critical when portfolios are used in large-scale assessments than when they are used as an assessment tool in the classroom. Portfolio scoring becomes an essential part of the portfolio system in terms of reliability, validity, and fairness. The first step in portfolio scoring is to develop rubrics and construct an assessment matrix. The rubrics should include both the content and the processes students are expected to master. Rubrics should also describe what kind of evidence we should expect for the different levels on the performance scale. Ideally, teachers should be involved in the process of defining and refining the rubrics. Some studies have been conducted to study how consistent and accurate readers are in their judgment (see Bridgeman, Chittenden, and Cline (1995); Myford and Mislevy (1995); and Koretz (1992, 1994)). Reliability has usually been reported in terms of inter-rater agreement (i.e., the correlation between the scores two different raters give to the same student work). When discrepancies between these raters are large (the criteria for "large" depends on the scale used), usually a third reading is done. The use of a third rater reduces the impact of rater severity on a student's scores (i.e., how strict a rater is when evaluating a student's work), but this procedure does not resolve the problem of rater interchangeability. For example, if both raters are extremely severe or lenient, this procedure will not be able to detect that, and the final score given to the student may be erroneous. Myford, Marr, and Linacre (1996) propose to calibrate raters accurately and precisely, adjusting examinees' scores for rater severity differences, hence improving reliability calculations by removing the errors in those scores associated with rater severity.

Statistics can point out the problems; however, to find out the nature of these problems, a naturalistic study is needed. Mislevy and Myford (1995) conducted both a statistical and a naturalistic analysis. They identified some challenges that raters faced, as well as specific types of portfolios that were difficult to evaluate. Among the challenges faced by raters, Mislevy and Myford included: (1) the "empathy mode," a tendency to evaluate students in terms of potentials; (2) the "bounce effect," the effect the last portfolio read has on the reading of the next one (i.e., comparison to another student is used instead of analyzing the student's work by defined rubrics); (3) the overuse of irrelevant background knowledge for decision making; (4) the use of the middle levels of the scale, which the authors call "play it safe"; and (5) having many or few experiences with the medium or style in which the student is working. With respect to portfolios, challenging portfolios are those in which there is a lack of consistency (i.e., good ideas but not enough technique to develop them), different levels of abilities, different sources of evidence, and unique or very special portfolios.

Pacesetter Spanish Portfolio Research Project (Phase I)

Method

Portfolios are one among several assessment modes included in Pacesetter Spanish, and they provide teachers and students with a tool of great instructional and motivational value. In Pacesetter Spanish, a portfolio is a collection of significant samples of student work over time, accompanied by clearly stated evaluation criteria and students' reflections on their own learning progress. Most of the activities and final projects of the units in the Pacesetter Spanish course materials are considered acceptable choices for inclusion in students' portfolios. Because of the variety of these activities, portfolios may contain a rich array of samples, including videos of student reports and role playing, audio excerpts of informal and formal interviews and presentations, drafts and final versions of written work, collages, and other descriptions of projects.

Portfolios can yield valuable data about students' progress and lead them effectively through their language learning. They function as guides by allowing students to make choices, helping them to both understand and demonstrate how they reason, create, and use strategies. Also, portfolios promote students' reflection on their work and learning. Thus, portfolios help students offer evidence of their progress toward meeting established outcomes, and enable them to take on the responsibility for their own learning process.

The Pacesetter Spanish portfolio research project goals are: (1) to prepare a working definition of Pace-

setter Spanish portfolios as assessment tools, (2) to develop a standardized portfolio assessment system, and (3) to use the project findings for training teachers in upcoming professional development sessions.

To achieve these goals, several issues were addressed: (1) the development of an assessment matrix to take into account course outcomes, performance indicators, and scoring rubrics for the Culminating Assessment, (2) piloting and refining the resulting assessment matrix, and (3) identifying benchmarks and developing a set of sample portfolios to be used for training purposes.

Project personnel included two ETS project staff members from the Pacesetter Spanish program and a summer intern. Andrea Fercsey and Carmen Luna of the Assessment Division coordinated this phase of the research with the help of Eva Ponte, graduate student at the University of California at Berkeley, who was the summer intern assigned to work on this project. Six Pacesetter Spanish teachers participated in the research project, representing four different districts in which the program is being implemented. It should be noted that these participating teachers had a very short period of time (two months) to devote to the implementation of portfolios in their classes. Such a limitation posed various constraints, and it is expected that a full school year devoted to portfolios would ease these extraneous problems.

The first meeting was held in April 1996. Colleagues who had worked on the Pacesetter English portfolio project presented their experience with their program. The goals, issues, and design of the Pacesetter Spanish Portfolio Research Project were discussed and revised as needed.

The role of portfolios as an assessment tool was clarified: Pacesetter Spanish portfolios would be used to evaluate classroom products and not processes. However, since portfolios are a collection of products over time, the information gathered through their evaluation may be used to enlighten instructional materials and methods.

Through discussions with participants, it became apparent that teachers needed much guidance to be better prepared to help their students select samples for portfolios. It was emphasized that the materials needed to be extremely teacher friendly, and issues of portfolio management (especially logistics and storage) were also discussed. It was determined that the content of Spanish portfolios for this phase of the research project should be two writing samples, two speaking samples, one reading comprehension sample, one listening comprehension sample, and the final projects for the units. It was made clear that this list of artifacts was only a suggestion open to revision.

A consensus was reached to work with the same terminology used in the five strands of the Pacesetter Spanish Culminating Assessment and end-of-year report. These five strands are: "Beginning," "Developing," "Promising," "Accomplished," and "Advanced." Using course outcomes as the basis for the elaboration of the assessment matrix, two dimensions were defined. The first dimension was called "Demonstrating Knowledge of Hispanic cultures," and the second dimension was named "Using Spanish to Communicate Effectively." The three aspects of Dimension 1 are not evaluated explicitly in the Culminating Assessment. A decision was made to include Dimension 1 in portfolio assessment to better address the course outcomes. A preliminary assessment matrix was then developed, and several aspects were included within each dimension. A preliminary description of each of the five strands of the scale was developed. A cover sheet and basic guidelines for students were reviewed and approved. An example of these forms can be found in Appendix 1.

The assessment matrix included three aspects under Dimension 1, "Demonstrating Knowledge of Hispanic Cultures": *Showing awareness of the diversity of Hispanic cultures, Identifying contributions of Hispanic figures,* and *Making connections with other disciplines and own culture(s)*. The second dimension, "Using Spanish to Communicate Effectively," included two aspects: *Deriving meaning from texts and personal interactions* (receptive skills: listening and reading) and *Expressing meaning in oral and written form* (productive skills: speaking and writing). This version of the assessment matrix is included in Appendix 2. Concerning the use of the assessment matrix, it was noted that some artifacts may be used to evaluate different dimensions—and even different aspects within those dimensions—so raters have to be prepared to do multiple readings of the same work as needed (e.g., a videotape may need to be seen twice: the first time to evaluate the student's knowledge of Hispanic cultures, and the second time to evaluate the student's ability to communicate in Spanish). For all aspects of both dimensions, the matrix also included a category called "Not Enough Evidence to Judge." The raters were to use this category when they felt they were unable to give a rating because of a lack of sufficient artifacts.

During April and May 1996, participating teachers carried out a more concerted implementation of portfolio assessment in their own classrooms. Students were informed of the draft assessment matrix, cover sheet, and guidelines. It was pointed out to both students and their parents that this was a research project, and that the findings would facilitate new approaches and revisions to the Pacesetter Spanish program. Prior to the second meeting, a scoring sheet was also developed. (This form is shown in Appendix 3.)

At the second meeting in June 1996, the assessment matrix was piloted, and sample portfolios were read

and scored. Participating teachers were asked to provide 25 portfolios each. It should be noted that one participant was unable to attend. However, she did send the corresponding portfolios for her students, and these were integrated as part of the pool to evaluate. During the reading sessions, raters were asked to identify the artifacts and contents of all portfolios in the scoring sheets. The transcription of comments written in the scoring sheets is included in Appendix 5.

This second meeting started with a session devoted to feedback. Participants discussed the implementation of portfolios in their classrooms and noted problems they encountered. First, they mentioned they had too little time to give their students a clear understanding of how portfolios work and enable the students to gather the necessary information. These difficulties seem inherent in this research project. The particular time constraints of the project should not be problematic when portfolios are implemented in Pacesetter Spanish at the beginning of the school year. Second, participants related some problems with portfolio management throughout the school year. ETS project staff suggested the following stages to help manage portfolios. During the first marking period of the academic year, students should concentrate on the organizational aspect of portfolios. During the second marking period, students should become fully cognizant of the rubrics. During the third marking period, a conference should be held with individual students to ensure that they can fully justify each artifact chosen at that point and that they become aware of any aspects not duly addressed. Then, students should finish gathering all necessary evidence for their portfolios. At the end of the process, there should be no glaring omissions or lack of clarity. A final conference should then be conducted.

Some participants also indicated that several students viewed portfolios just as a collection of work and had problems understanding the assessment matrix. To address this issue, it was suggested that in the future students could be presented with several portfolios showing evidence of performance for each level in the scale (i.e., from Beginning to Advanced).

Participants also mentioned some positive aspects of implementing portfolios. Less able students felt they "had a shot at this." Students with average abilities said they felt very comfortable with this approach to learning. Also, as students were selecting pieces for their portfolios, they were given the opportunity to improve their products, motivating them to go over their work. This provided extremely beneficial effects in terms of both their learning experience and outcomes. More important, students felt they were learning to teach themselves and were becoming more aware of their own processes.

Both teachers and ETS project staff were extremely pleased with the quantity and quality of the portfolios. Initially some of the participants had doubted the possibility of using a portfolio system in their classroom. While they faced difficulties during the initial stage of implementation, they felt a great sense of accomplishment at the end of the project. ETS project staff were delighted to see that both teachers and students were truly becoming fully involved in the development and implementation of portfolios in the Pacesetter Spanish program.

In the next step of the meeting, the group as a whole selected portfolios that were thought to be representative of different levels of the scale. These portfolios were discussed with the guidance of ETS project staff. During these discussions, some teachers reported problems they faced when portfolios contained pieces at different levels of performance. ETS project staff indicated that since it is usual for students' learning to register different "peaks and valleys" throughout the school year, their artifacts would consequently show evidence of such differences.

During this part of the meeting, teachers also reported that poorly organized portfolios were very difficult for teachers to read. ETS project staff indicated that students should be strongly encouraged to properly organize their portfolios. (For instance, when students working as a group submit a video with a presentation or a tape with an interview, they should at all times identify themselves. Otherwise, raters who are not their teachers will find it impossible to identify students in order to evaluate their individual work.)

The next planned activity at the meeting was the reading and scoring of portfolios. Because of their lack of experience in similar endeavors, participants decided to start by working in pairs. Participants and ETS project staff then started by reading two portfolios in pairs, afterward moving on to read and score portfolios individually. Twenty-one portfolios were selected to be read twice. ETS project staff were used as third raters in case of large discrepancies. Originally, it was expected that each rater would read six portfolios; however, because some raters evaluated portfolios faster than other raters, two raters judged five portfolios, two raters read the expected number of portfolios (six), and another two raters read seven portfolios.

In all, a total of 30 portfolios were read: 8 portfolios were read in pairs, 21 portfolios were read individually by two different raters (double reading), and 1 was read by only one rater. The flow of portfolios for this reading was organized so that teachers did not read their own students' portfolios. Raters rated each aspect and then gave a total score for each dimension.

During the final part of the meeting, the assessment matrix was discussed, taking into account knowledge gained from the reading sessions. A discussion took place on whether Aspect 1 (*Showing awareness of the*

diversity of Hispanic cultures) and Aspect 2 (*Identifying contributions of Hispanic figures*) could really be considered different. The conclusion was that they were indeed different, but that teachers did not have a clear enough idea of what they were looking for in a portfolio. Consequently, they were not very effective in communicating this aspect to their students. Also, raters often indicated that there was "not enough evidence" to evaluate students' work on Aspect 3 of Dimension 1 (*Making connections with other disciplines and own culture[s]*). This may have occurred because this aspect was inadvertently omitted from the cover sheet that students were told to use to write their justification for each selected piece. It was also noted that in most cases there was little or no evidence to judge Dimension 1; there seemed to be a problem with both the quantity and quality of artifacts used to show evidence of learning for Aspect 1. There was usually plenty of evidence to judge Dimension 2, and this may be due to both students and teachers having experience evaluating productive language skills in their courses.

At the end of the process, the parameters initially indicated for selecting the required number of pieces in a student's portfolio were found to be inadequate, resulting in portfolios that seemed to be lacking in the necessary evidence to judge all aspects of both dimensions. A decision was made: There would no longer be a set number of pieces to be included, and more emphasis would be placed on the process of selecting enough artifacts to show evidence of a student's accomplishments.

Portfolio Analyses

After the second meeting, all data and information were analyzed. Certain limitations of this portfolio assessment should be considered. First, guidelines to select materials were just tentative. Second, the assessment matrix was treated as a pilot. Third, certain aspects were quite new to both students and teachers, and the time constraints prevented them from becoming familiar with those new aspects.

The group of raters was composed of five Pacesetter Spanish teachers and two ETS project staff members. None of the raters had any prior experience scoring a Spanish portfolio, and only one ETS staff had prior experience scoring portfolios—but in another content area (English as a Second Language).

Training was done first as a whole group, then in pairs. While reading portfolios as a group, raters were asked to have the assessment matrix present at all times, and to think of ways in which they were using it to make decisions. Raters were also asked to reflect on the use of specific evidence to justify the ratings they were assigning. Finally, they were asked to note all the problematic aspects they found when scoring any portfolio. When raters were reading in pairs, the two of them discussed and decided on the ratings for each aspect and holistically for the two dimensions. Next, the group as a whole discussed and shared their experiences, and then the raters started reading and scoring portfolios individually.

Description of Elements

As mentioned earlier, 30 portfolios were read. There was a total of 59 portfolio readings: Eight portfolios were read in pairs, 21 were read individually with two raters per portfolio (42 readings), 8 portfolios were read by a third rater when discrepancies of 2 points or more were found between the first two raters, and 1 portfolio was read by only one rater.

Distribution of Ratings

To analyze the distribution of ratings, we constructed two tables, one showing the frequencies of ratings given across portfolio aspects and the other showing the distribution of ratings for each portfolio aspect and dimension. The rating scale was composed of five categories: "Beginning," "Developing," "Promising," "Accomplished," and "Advanced." There is another category in the analysis, named "Not Enough Evidence," not included in the rating scale. (See page 8.)

Table 1 shows that the categories with the highest frequencies are 3 and 4 (i.e., "Promising" and "Accomplished"), which, when combined, accounted for about two-thirds of all ratings given. Overall, students' performances tend to be somewhat above average (i.e., 68 percent of the ratings were 3 or higher while fewer than 9 percent were 2 or lower), with more students performing in the middle and upper half of the scale. The high percentage of judgments falling in the category "Not Enough Evidence" is also noticeable.

TABLE 1

Distribution of Ratings

Category Number	*Category Name*	*Frequency*	*Percentage*
1	Beginning	12	2.91
2	Developing	23	5.57
3	Promising	158	38.25
4	Accomplished	114	27.60
5	Advanced	10	2.42
0	No Evidence	96	23.24
	TOTAL	413	100.00
Distribution	N = 317	Mean: 3.27	SD: .80

To study student performance and the incidence of the category "Not Enough Evidence" in more detail, Table 2 displays the distribution of frequencies in each rating category for each aspect and dimension.

As expected, the rating category "Promising" has the highest frequencies. There are only two aspects in which this does not occur: Aspect 3, in Dimension 1, and Aspect 1, in Dimension 2. In the first case, *Making connections with other disciplines and own culture(s)*, the category with the highest frequency is "Not Enough Evidence" (71 percent). The reason for this may be that this aspect was not included in the "cover sheet" form given to students to select their work and justify their selection. In the second case, *Deriving meaning from texts and personal interactions,* the rating category with the highest frequency is "Accomplished." Not surprisingly, students' performance in this aspect (receptive skills) tended to be better than in other aspects. This conforms to known patterns of language acquisition. In general, teachers had more experience evaluating both aspects of Dimension 2 throughout the Pacesetter Spanish course. Also, students had less difficulty identifying artifacts to show evidence of their achievements in Dimension 2. (They tend to show their command of a language in terms of their listening, speaking, reading, and writing abilities.)

Included in Table 2 are the means and standard deviation of the score distributions for each aspect and for the total in each dimension. The statistics describing these distributions were calculated for each aspect and dimension. When we calculated the means and standard deviations, we did not include the category "Not Enough Evidence." The means ranged from 2.8 (total Dimension 1) to 3.54 (Aspect 1, Dimension 2). The overall mean, without making distinctions between aspects and dimensions, is 3.27. This indicates, as we have seen before, that overall students' performance is medium-high. When we review the means for the aspects, we see that students' performance tends to be similar across the different aspects and dimensions. Variability within dimension, measured by the standard deviation, is also similar across different aspects and dimensions: It ranges from .61 (Aspect 1, Dimension 2) to .94 (Aspect 3, Dimension 1).

In general, we see that most ratings fell in the medium-high categories of the scale ("Promising" to "Accomplished"). A much smaller number fell in the other categories ("Beginning," "Developing," and "Advanced"). Aspect 1, Dimension 1 *(Showing awareness of the diversity of Hispanic cultures)*, was the aspect having the highest percentages in categories "Beginning" and "Developing."

TABLE 2

Distribution of Ratings by Aspects and Dimensions

Dimension 1

Category	*Category Name*	*Aspect 1 Freq.*	*Aspect 1 %*	*Aspect 2 Freq.*	*Aspect 2 %*	*Aspect 3 Freq.*	*Aspect 3 %*	*Total Freq.*	*Total %*
1	Beginning	3	5	2	3	1	2	4	7
2	Developing	8	14	4	7	0	0	4	7
3	Promising	26	44	24	41	9	15	25	42
4	Accomplished	11	19	12	20	5	9	9	15
5	Advanced	2	3	1	2	2	3	2	3
0	No evidence	9	15	16	27	42	71	15	25
TOTAL		59	100	59	100	59	100	59	100

	Aspect 1	*Aspect 2*	*Aspect 3*	*Total*
N	50	43	17	44
Mean	3.02	3.14	3.4	2.80
SD	0.89	0.80	0.94	0.84

Note: These statistics have been calculated without taking into account the category "No Evidence."

Dimension 2

Category	*Category Name*	*Aspect 1 Freq.*	*Aspect 1 %*	*Aspect 2 Freq.*	*Aspect 2 %*	*Total Freq.*	*Total %*
1	Beginning	0	0	0	0	2	3
2	Developing	2	3	2	3	3	5
3	Promising	20	34	28	48	26	44
4	Accomplished	27	46	27	46	23	39
5	Advanced	1	2	1	2	1	2
0	No evidence	9	15	1	2	4	7
TOTAL		59	100	59	100	59	100

	Aspect 1	*Aspect 2*	*Total*
N	50	58	55
Mean	3.54	3.24	3.33
SD	0.61	0.86	0.77

Note: These statistics have been calculated without taking into account the category "No Evidence."

Time Used to Read Portfolios

Because this was the first time that portfolios were read and scored, we were interested in the amount of time used to read them (as shown in Table 3). Raters were asked to keep track of this. Some forgot, hence the missing data in Table 3.

As indicated in Table 3A, the average time employed to read a portfolio was 46.74 minutes. The shortest amount of time needed was 20 minutes while the longest was 120 minutes (Table 3). Several tables were constructed to examine whether the amount of time needed to read a portfolio was different for portfolios coming from different Pacesetter Spanish teachers, for different raters, and for readings in different sessions.

As indicated in Table 4, there is some variation in terms of the time raters needed to rate portfolios from different teachers. On average, portfolios from the class of Teacher 2 took the shortest amount of time to read (41.25 minutes

TABLE 3

Time Used to Read Each Portfolio

Student	*Student's Teacher*	*Rater*	*Date*	*Minutes*
1	3	7	St/Ev	20
1	3	6	Su/M	35
2	1	6	St/Ev	20
2	1	4	St/Ev	30
3	6	4	St/Ev	20
3	6	3	Su/M	55
4	3	8	St/Ev	65
4	3	2	Su/M	40
5	3	1	St/Ev	60
5	3	7	St/Ev	35
6	4	1	St/Ev	30
6	4	7	St/Ev	30
7	4	2	St/Ev	
7	4	8	Su/M	50
8	1	4	St/Ev	
8	1	3	Su/M	35
9	2	7	St/Ev	20
9	2	6	St/Ev	25
10	6	3	St/Ev	50
10	6	1	Su/M	20
11	5	3	St/Ev	
11	5	1	Su/M	40
12	1	7	St/Ev	40
12	1	6	Su/M	45
13	2	6	St/Ev	45
13	2	4	St/Ev	35
14	2	4	St/Ev	40
14	2	6		40
15	5	1	St/Ev	30
15	5	7	Su/M	40
16	5	2	St/Ev	60
17	5	8	St/Ev	70
17	5	4	Su/M	55
18	3	6	St/Ev	40
18	3	4	Su/M	30
19	6	2	St/Ev	45
20	4	3	St/Ev	60
20	4	1	Su/M	35
21	6	8	Su/M	50
21	6	2		
22	1	2 & 3	St/M	
23	1	8	St/M	90
24	2	4 & 7	St/M	95
25	2	7	Su/M	30
26	3	4 & 7	St/M-Ev	
27	3	4 & 7	St/M	55
28	6	2 & 3	St/M	120
29	5	1 & 6	St/M	100
30	4	1 & 6	St/M	80

TABLE 3A

Summary

Mean time used to read portfolios:	46.74 minutes
Standard Deviation:	22.83

per portfolio); portfolios from the class of Teacher 5 took the longest to read (56.42 minutes per portfolio).

Table 5 displays the average time each rater needed to read a portfolio. Raters worked at different paces; consequently, we combined different pairs of raters, so that all 21 portfolios were read twice. In Table 5, Raters 1 through 7 are individual raters, number 8 refers to Raters 4 and 6, number 9 to Raters 2 and 3, and number 10 to Raters 1 and 5. Although the sample is small, it took more time to read portfolios in pairs than individually. This was expected because discussion took place during the reading. Within individual raters, Rater 6 was the fastest rater, and Rater 7 was the slowest.

Table 6 shows the average time needed to read portfolios in each session. Session 1 ran from Saturday morning through the afternoon; Session 2 took place the afternoon and early evening of Saturday; and Session 3 took place Sunday morning. Raters took more time on average to rate portfolios during the first session. This was expected since it was the first Pacesetter Spanish portfolio evaluation, and because the readings in this first session were done mostly in pairs. There is almost no difference in the average time needed to rate portfolios during Sessions 2 and 3.

Culminating Assessment

As previously indicated, Pacesetter Spanish includes a standardized assessment component called the Culminating Assessment. This assessment was developed for the areas of reading, speaking, listening, and writing. Reading and speaking were locally scored by individual teachers, and listening and writing were centrally scored

TABLE 4

Average Time Employed to Read Portfolios From Different Teachers

Teacher	*N*	*Mean*	*St. Dev.*	*Minimum*	*Maximum*
1	6	43.33	24.43	20	90
2	8	41.25	23.26	20	95
3	9	42.22	14.81	20	65
4	6	47.50	19.94	30	80
5	7	56.42	23.58	30	100
6	7	51.42	33.50	20	20

N = Number of portfolios read
Minimum = Minimum time needed
Maximum = Maximum time needed

TABLE 5

Average Time Employed by Each Rater

Rater	*N*	*Mean*	*St. Dev.*	*Minimum*	*Maximum*
1	6	35.83	13.57	20	60
2	3	48.33	10.41	40	60
3	4	50.00	10.80	35	60
4	6	35.00	11.83	20	55
5	7	35.71	9.76	20	45
6	7	30.71	8.38	20	40
7	5	65.00	16.58	50	90
8	2	75.00	28.28	55	95
9	1	120.00		120	120
10	2	90.00	14.14	80	100

N = Number of portfolios read
Minimum = Minimum time needed
Maximum = Maximum time needed

at a national reading. Dimension 2 of the portfolio scoring sheets measures language, and it is divided into two aspects: receptive skills (listening and reading), and productive skills (writing and speaking). Although we can not directly compare these aspects with the areas of the Culminating Assessment, we wanted to examine whether the patterns of these two modes of assessments were similar (i.e., if a student got a high score in the Culminating Assessment, we expected him or her to get a high score in the respective aspects of the portfolio). Both the Culminating Assessment and portfolio scales contain five strands, and there was an attempt to maintain parallelism in the descriptors for each strand, especially for Dimension 2 (where the constructs are similar).

Table 7 displays the scores obtained by students both in the Culminating Assessment and in the portfolio. The information displayed in this table indicates that in general portfolio scores tended to be lower than Culminating Assessment scores. This may have happened because those skills not included in the Culminating Assessment are usually harder for students to master and for teachers to assess.

The overall correlation coefficients between the Culminating Assessment and the portfolio scores were low (r = .15 between Culminating Assessment listening grade and portfolio Dimension 2, Aspect 1 *Deriving meaning from texts and personal interactions;* and r = .07 between Culminating Assessment writing grade and portfolio Dimension 2, Aspect 2, *Expressing meaning in oral and written forms*). These results suggest that the Culminating Assessment and the portfolio assessment may be tapping different sets of knowledge and skills, even though the constructs were assumed to be similar.

TABLE 6

Time Employed to Read Portfolios at Different Sessions

Session	*N*	*Mean*	*St. Dev.*	*Minimum*	*Maximum*
1	6	90.00	21.68	55	120
2	22	39.54	15.65	20	70
3	14	40.00	10.19	20	55

N = Number of portfolios read
Minimum = Minimum time needed
Maximum = Maximum time needed

TABLE 7

Culminating Assessment and Dimension 2 Portfolio Ratings

Student	*Culminat. Assess. Listening Grade*	*Portfolio Dim. 2 Aspect 1 Rating*	*Culminat. Assess. Writing Grade*	*Portfolio Dim. 2 Aspect 2 Rating*
1	3	3	3	3
2	4	2.5	3	3
3	3	2.5	3	4
4	2	4	3	3.5
5	5	4	5	3
6	3	4	3	3
7	4	4	3	3.5
8	5	3.5	4	4
9	5	4	5	4
10	5	4	4	3
11	4	4	5	4
12	5	3.5	5	3
13	4	2.5	4	3
14	5	3	4	3
15	4	3.5	3	3.5
16	5	4	4	4
17	5	4	4	5
18	5	3.5	5	3
19	5	2.5	3	3.5
20	4	4	4	3.5
21	5	4	5	4
22	5	4	4	4
23	5	3	5	3
24	4	3	5	3
25	5	3	5	3
26	4	4	4	4
27	4	3	3	3
28	5	3	4	4
29	2	3	2	3
30	5	4	4	4
Mean	4.30	3.47	3.93	3.48

Portfolio Assessment Matrix Correlations

The portfolio's assessment matrix included seven ratings: one rating for each of the three aspects of

TABLE 8

Portfolio Assessment Matrix Correlations

	Dimension 1 Aspect 1	*Dimension 1 Aspect 2*	*Dimension 1 Aspect 3*	*Dimension 1 Total*	*Dimension 2 Aspect 1*	*Dimension 2 Aspect 2*	*Dimension 2 Total*
Dimension 1 Aspect 1							
Dimension 1 Aspect 2	.37						
Dimension 1 Aspect 3	.51	.30					
Total Dimension 1	.72	.69	.68				
Dimension 2 Aspect 1	.26	.13	.36	.35			
Dimension 2 Aspect 2	.30	.46	.54	.43	.29		
Total Dimension 2	.34	.25	.34	.43	.61	.52	

Dimension 1, one rating for each of the two aspects of Dimension 2, and two ratings as totals for both dimensions. We include in Table 8 the correlations among all these ratings.

All correlations between aspects and dimensions are significant at the .06 level (except for the correlation between Aspect 1, Dimension 2 and Aspect 2, Dimension 1), and range from moderate (.25) to high (.72). Correlations among the three aspects under Dimension 1 range from .30 to .51, while the correlation between the two aspects of dimension 2 is .29. Two correlations are moderately high: the correlation between Aspect 2, Dimension 1 and Aspect 2, dimension 2 (.46), and the correlation between Aspect 3, Dimension 1 and Aspect 2, Dimension 2 (.54). The correlations of Aspects 1 and 2 of Dimension 2 with the total of Dimension 2 are .61 and .52.

Portfolios Read Twice: Reliability Study

Portfolio assessment is based on personal judgments. One common approach to studying the generalizability and fairness of the judgments is to ask a second person to read and score the portfolios. Due to the nature of the project and the stage in which the Pacesetter Spanish program was at that moment, and also due to time limitations, it was only possible to obtain a small sample of portfolios read twice.

FACETS (Linacre, 1989), a computer program based on Rasch partial credit models, was used to analyze the data. FACETS analyzes data from assessments that have several facets (e.g., raters, students, tasks) by assigning parameters to each facet in the model. FACETS has the form of a log-linear model for main effects and estimates those effects in logits (logits are the logarithmic odds of a given rating compared to the next lower one). Linear measures are then constructed from ordered qualitative data, allowing the separation of the contribution of each facet. We have three facets: students, readers, and portfolio aspects and dimensions. Some students were more proficient than others, some readers rated more severely than others, and some portfolio aspects were harder to get high ratings on than others. This variation among students, raters, and aspects is expected; however, unexpected variations may signal attention to unresolved or unnoticed problems; i.e., those parts of the data that seem not to follow usual patterns of variability.

FACETS produces a map in which all facets of the analysis are shown in one figure, providing the reader with general information for each facet. We include in Figure 1 a reproduction of the FACETS map.

FACETS calibrates all facets of the analysis so that they can be positioned on the same equal interval scale. This scale is shown in the first column of the map.

Students, the first facet of our study, occupies the second column of the FACETS map. This column displays the estimates of students' proficiency on the portfolio assessment. Student measures are ordered with more proficient students at the top of the column and less proficient students at the bottom of the column. Student proficiency measures range from –3.2 logits to 3.2 logits.

The third column in Figure 1 represents raters in terms of the harshness or leniency they exercised when rating portfolios. Harsher raters appear at the top of the column, and more lenient raters appear at the bottom. Rater 6 was the harshest, and Rater 3 was the most lenient. Rater harshness measures range from –2.3 logits to 2.3 logits. As expected, this variation is smaller than the variation found in student measures.

The fourth column of the map displays the portfolio in terms of its aspects. The hardest aspects are at the top of the column and the easiest at the bottom. This indicates that Aspect 3, Dimension 1 was the hardest to get high ratings on, and Aspect 2, Dimension 1 was the easiest one to get high ratings on.

Columns 5 through 11 show the most probable rating for a student at a given level on the logit scale, as

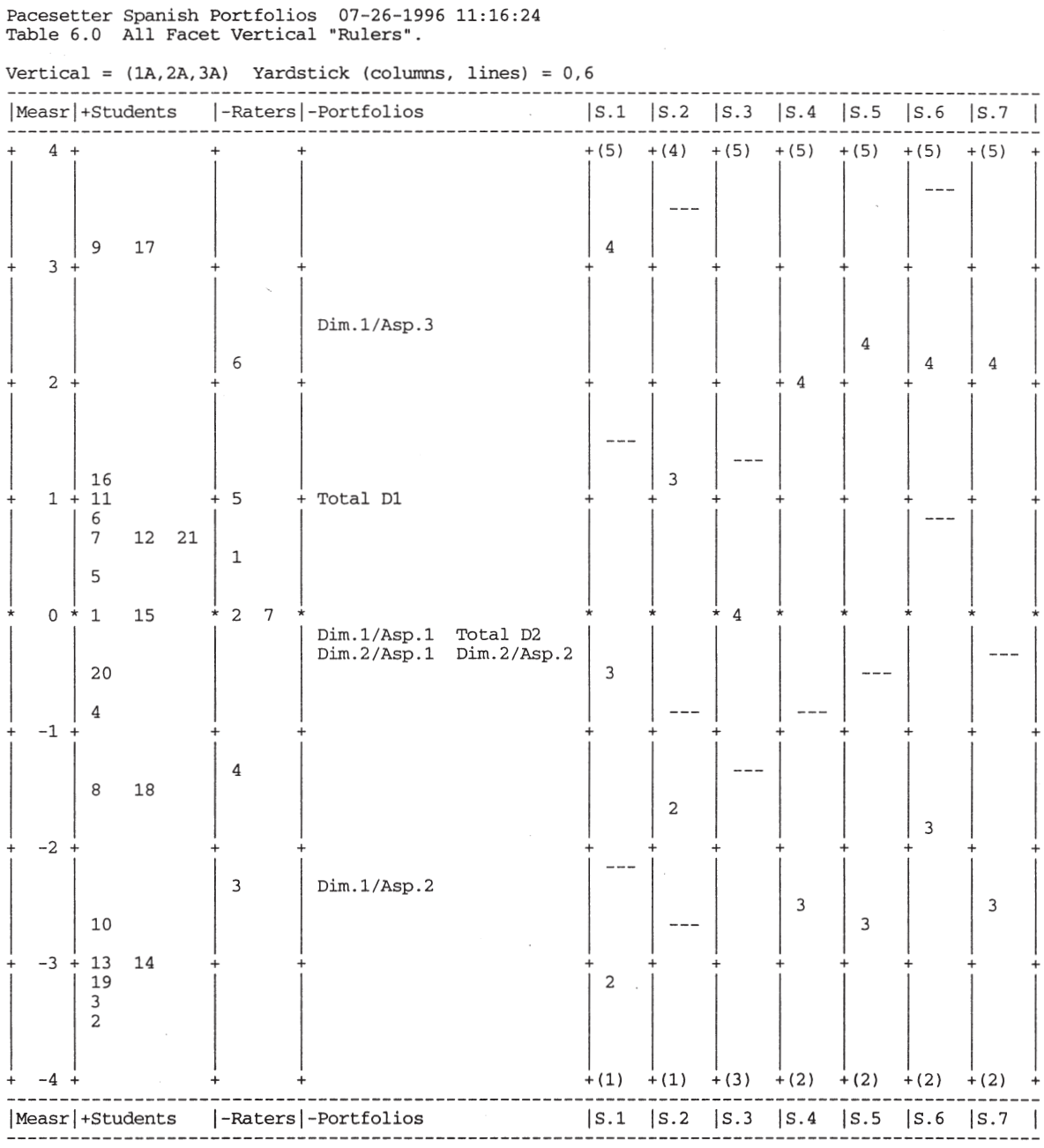


Figure 1. FACETS map for the whole scale (21 portfolios, two readers).

expected from a rater with average harshness, in each of the portfolio aspects and dimensions. The horizontal lines across a column indicate the point at which the likelihood of getting the next higher rating begins to exceed the likelihood of getting the next lower rating. For instance, if we look at scale 1 (Aspect 1, Dimension 1), we see that students with measures from –2.1 to 1.5 logits are more likely to receive a 3 ("Promising") than any other rating on that aspect.

In an effort to study how reliable raters' judgments were, and also with the aim of identifying problematic portfolios, we analyzed the discrepancies in ratings between raters of the same portfolio. Table 9 includes the number and percentage of scoring discrepancies for each aspect of the portfolio, and in total.

As we have indicated before, the scale covers five levels: beginning, developing, promising, accomplished, and advanced. We also had a category for cases where

TABLE 9

Summary of Portfolio Score Discrepancies

	Dimension 1 Aspect 1	*Dimension 1 Aspect 2*	*Dimension 1 Aspect 3*	*Dimension 2 Aspect 1*	*Dimension 2 Aspect 2*	*D 1*	*D2*	*Total*	*Percentage*
Same Score	8	7	14	4	11	6	8	59	40
1–Point Disc.	6	7		10	10	7	9	49	33
2–Point Disc.	2	1		1		1		5	3
Evid. vs. No evidence	5	6	7	6		7	3	34	23

there was not enough evidence to judge. In the last row of the table, we have included those situations in which one rater said there was no evidence, and the other considered there was enough evidence and gave a judgment.

The number of agreements is moderate; it constitutes 40 percent of all paired judgments. (However, it should be noted that cases where both raters considered there was no evidence to judge have been included here as "agreements.") The number of portfolios having only a 1-point discrepancy is also moderate (33 percent of all paired judgments). Finally, the number of discrepancies of 2 points was small (3 percent).

An important finding is that the number of discrepancies between raters who judged student work and raters who said there was not enough evidence to judge was considerable (23 percent of all paired responses). Most discrepancies of this type occurred in Dimension 1, especially in Aspect 3.

We were also interested in seeing at which specific levels of the scale these discrepancies occurred. Table 10 displays those discrepancies for each portfolio aspect and for the portfolio as a whole.

The type of discrepancy with the highest occurrence (41 percent) is that in which one rater considered a student's performance promising while the other considered it accomplished (discrepancy 3-4). Another type of discrepancy with a high percentage of occurrence (22 percent) was that in which one rater thought there was

TABLE 10

Summary of Portfolio Score Discrepancies for Each Aspect in Each Dimension

Discrepancy	*Dimension 1 Aspect 1*	*Dimension 1 Aspect 2*	*Dimension 1 Aspect 3*	*Dimension 2 Aspect 1*	*Dimension 2 Aspect 2*	*Dimension 1: Total*	*Dimension 2: Total*	*Total Frequency*	*Total Percentage*
No Evid.–1	1							1	1
No Evid.–2	1	2				2		5	6
No Evid.–3	3	4	3	3		4	2	19	22
No Evid.–4			3	2		1	1	7	8
No Evid.–5			1	1				2	2
1–2	1							1	1
1–3		1						1	1
2–3	2	1		1	2	2	2	10	11
2–4	1			1				2	2
3–4	3	6		9	7	5	6	36	41
3–5	1					1		2	2
4–5					1		1	2	2
TOTAL (Discr.)	13	14	7	17	10	15	12	88	
Agreement (full scores)	8	6	2	4	11	5	9	45	
Agreement (No Evid., 0-0)		1	12			1		14	
TOTAL (Agreement)	8	7	14	4	11	6	9	60	

no evidence to judge, and the other considered the work showed evidence of being promising (difference "No Evidence" – 3). This latter type of discrepancy is more important and indicates that raters' internalization of the assessment matrix and search of evidence was problematic. One possible explanation is that some raters judged only those things students had indicated as evidence of their learning in a specific area, and other raters used materials not marked by the student for that specific area. The only aspect in which there are no discrepancies of this type is Aspect 2, Dimension 2.

Aspect 3, Dimension 1 registered the fewest discrepancies. It should be noted, though, that most agreements were of the type "No Evidence"–"No Evidence," meaning that both raters judged there was not enough evidence to evaluate a student's performance.

Usually a 1-point discrepancy is considered small enough to categorize the rating as an acceptable judgment. However, problems may arise when there is a cut point. FACETS provides a "fair" measure, which indicates the score a student would have received had his or her ratings been adjusted for rater effects. Even small adjustments can be important for those students whose scores lie in critical cut-score regions. Table 11 shows each student's observed average score and his or her "fair" average score (i.e., adjusted for severity or leniency of the raters providing the ratings for that student).

The reported scores of six students would have changed after results were adjusted for rater severity or leniency. Four students (students 5, 6, 15, and 12) would have changed their reported score from "Promising" to "Accomplished," and two students (students 20 and 8) would have changed their scores from "Accomplished" to "Promising."

We also studied the behavior of each rater separately. Table 12 indicates how many times each rater used the judgment "No Evidence" and the other levels of the scale.

Raters 4 and 6 used the category "Not Enough Evidence to Judge" very frequently. In each case over 35 percent of their judgments were "Not Enough Evidence."

TABLE 11

FACETS OUTPUT (7.1.1) Students Measurement Report (Arranged by FN).

CATEGORY "NOT ENOUGH EVIDENCE TO JUDGE" = MISSING

Pacesetter Spanish Portfolios 07-26-1996 11:16:24
Table 7.1.1 Students Measurement Report (arranged by FN).

Obsvd Score	Obsvd Count	Obsvd Average	Fair Avrge	Model Measure	Model S.E.	Infit MnSq	Infit Std	Outfit MnSq	Outfit Std	PtBis	Nu Students
33	10	3.3	3.5	.40	.65	0.3	-2	0.3	-2	.39	5 5
41	11	3.7	4.1	3.23	.65	0.4	-1	0.3	-1	-.06	9 9
49	13	3.8	3.6	.94	.61	0.4	-1	0.3	-1	.26	11 11
43	12	3.6	3.5	.71	.64	0.4	-1	0.4	-1	.55	21 21
40	12	3.3	3.5	.86	.62	0.5	-1	0.4	-1	.35	6 6
46	13	3.5	3.3	-.47	.59	0.5	-1	0.4	-1	.44	20 20
15	5	3.0	2.7	-3.03	.86	0.4	-1	0.4	-1	.00	14 14
41	13	3.2	3.5	-.01	.60	0.6	0	0.5	0	.18	15 15
22	9	2.4	2.4	-3.54	.60	0.6	-1	0.6	-1	.52	2 2
51	14	3.6	3.7	1.12	.57	0.7	0	0.6	-1	.06	16 16
36	10	3.6	3.1	-1.53	.67	0.7	0	0.7	0	.30	8 8
24	9	2.7	2.6	-3.02	.59	0.9	0	0.8	0	.36	13 13
34	11	3.1	3.3	.04	.62	0.8	0	0.8	0	-.30	1 1
32	12	2.7	2.5	-3.13	.51	1.3	0	1.5	0	.18	19 19
24	8	3.0	2.9	-1.53	.69	1.6	0	1.1	0	.39	18 18
51	12	4.3	4.1	3.16	.61	1.7	1	1.7	0	.34	17 17
33	10	3.3	3.6	.64	.65	1.5	0	1.8	1	-.23	12 12
35	11	3.2	2.5	-3.26	.65	1.8	1	2.0	1	.14	3 3
43	12	3.6	3.5	.71	.64	2.0	1	1.8	1	-.02	7 7
36	12	3.0	2.6	-2.70	.55	1.9	1	2.4	2	.37	10 10
42	13	3.2	3.2	-.91	.61	2.2	2	2.4	1	.26	4 4
Obsvd Score	Obsvd Count	Obsvd Average	Fair Avrge	Model Measure	Model S.E.	Infit MnSq	Infit Std	Outfit MnSq	Outfit Std	PtBis	Nu Students
36.7	11.0	3.3	3.2	-.54	.63	1.0	-0.2	1.0	-0.3	.21	Mean (Count: 21)
9.4	2.0	0.4	0.5	1.99	.07	0.6	1.4	0.7	1.4	.23	S.D.

RMSE .63 Adj S.D. 1.89 Separation 2.98 Reliability .90
Fixed (all same) chi-square: 219.0 d.f.: 20 significance: .00
Random (normal) chi-square: 20.0 d.f.: 19 significance: .39

TABLE 12

Distributions of Raters' Judgments

Rater	*Beginning 1*	*Developing 2*	*Promising 3*	*Accomplished 4*	*Advanced 5*	*Not Enough Evidence*	*Total Number of Judgments*
1		2 (5%)	23 (55%)	13 (31%)		4 (9%)	42
2	5 (14%)		11 (31%)	15 (43%)		4 (11%)	35
3	1 (3%)	1 (3%)	7 (20%)	24 (69%)		2 (6%)	35
4		1 (2%)	15 (31%)	4 (8%)	6	23 (47%)	49
5	1 (2%)	8 (19%)	17 (40%)	10 (24%)		6 (14%)	42
6			27 (55%)	4 (8%)		18 (37%)	49
7	1 (2%)	1 (2%)	18 (43%)	17 (40%)		5 (12%)	42

As we can see, these raters also read more portfolios than the others. In future analysis, it may be interesting to study the relationship between time used to read portfolios, search for evidence, and use of the "No Evidence" category.

We can intuitively see why FACETS indicated rater 3 as the most lenient: 69 percent of the ratings that this rater gave were 4's. No other rater gave that high a percentage of 4's. Rater 6 was the harshest; only 8 percent of the ratings this rater gave were 4's, and the rater gave no 5's.

Third Rater Analysis

Using only the portfolio ratings for total Dimension 1 and total Dimension 2, we selected some portfolios with extreme discrepancies. We defined "extreme" as portfolios with dimension ratings 2 or more points apart, and also portfolios with ratings of evidence (any evidence) as opposed to "Not Enough Evidence to Judge." Due to time constraints, not all portfolios were read twice when this selection was made. Eight portfolios were selected for a third reading.

We analyzed the ratings each rater gave to each portfolio aspect (see Appendix 5). Table 13 summarizes the results of this analysis.

As we can see, portfolios 14 and 17 produced a lot of discrepant ratings, indicating that these were problematic portfolios. Discrepancies mostly occurred in Aspect 1, Dimension 1 (4 portfolios) and Aspect 1, Dimension 2 (3 portfolios).

Selection of Benchmarks

After the June 1996 meeting in Miami, ETS project staff selected some portfolios to be used as benchmarks. Staff took into account the discrepancies between raters as well as other factors such as content and variety of artifacts when selecting the benchmarks. The selection was made using a table in which portfolio scores were presented by matched pairs of raters (this table is included in Appendix 4; those portfolios selected as benchmarks are indicated in bold). Besides the portfolios selected using the information given in the mentioned table, another portfolio that was read by a group was also selected. Thus, a packet of portfolio materials was prepared and used at the three professional development sessions carried out in summer 1996 for new Pacesetter Spanish teachers. The midyear meetings at the end of the fall semester of 1996 also included sessions on portfolios, and the samples from the training packets served to disseminate the Pacesetter Spanish portfolio model.

Also, as an ancillary result, the rubrics for the four linguistic skills were revised, and a decision was made that Pacesetter Spanish would have a unified set of rubrics to be used throughout the course. It is expected that both teachers and students will become thoroughly familiar with these rubrics and should encounter little difficulty applying them to their work during the school year.

In order to analyze the information gathered from the portfolios, and to provide some useful feedback to teachers, all scoring sheets with all written comments were transcribed. A copy of these transcriptions can be found in Appendix 5.

TABLE 13

Portfolios Read Three Times

Portfolio	*Ratings in Agreement*	*Ratings 2 points apart*
3	6	1
5	5	2
8	5	2
12	5	2
13	6	1
14	3	3
17	1	6
18	7	0

Summary of Findings

- This research showed that some elements of the delineated Pacesetter Spanish system of portfolio assessment worked as intended, while others need further refinement.
- Pacesetter Spanish portfolio raters very often used the medium category of the scale ("Promising"), and not the extreme categories ("Developing" or "Advanced"). This may have been due in part to the small sample of portfolios read. In addition, students in a level three language class tend to cluster in the middle of the scale, which confirms actual classroom experience.
- Often the length of time needed to read a Pacesetter Spanish portfolio depended on the variety and complexity of artifacts included (i.e., a portfolio with many different types of media took longer to read than a portfolio with only paper-based samples). Although such variety and complexity are encouraged throughout the course framework, this element of time needs to be taken into account when preparing to read portfolios.
- The correlation between Dimension 2, "Using Spanish to Communicate Effectively," and the Pacesetter Spanish Culminating Assessment was low. These findings suggest that both assessment instruments serve to evaluate different sets of knowledge and skills, although the construct may appear similar. Whereas the Culminating Assessment evaluates four separate linguistic skills (reading, speaking, listening, and writing), Dimension 2 of the Pacesetter Spanish portfolio assessment system groups them in two aspects: *Deriving meaning from texts and personal interactions* (receptive skills) and *Expressing meaning in oral and written form* (productive skills).
- All correlations between aspects and dimensions of the Pacesetter Spanish portfolio assessment system were found to be significant, except for the correlation between Aspect 1, Dimension 2 and Aspect 2, Dimension 1.
- The number of rating agreements was moderate. However, there was a considerable number of discrepancies between raters regarding the presence or lack of enough artifacts to judge students' performance within the parameters of the Pacesetter Spanish portfolio assessment system.

Recommendations

- As shown by Phase I results, further research is needed to refine the assessment matrix and produce a final version to implement as part of the formal components of the Pacesetter Spanish program.
- A set of more comprehensive and clearer instructions should be given to teachers and students about the Pacesetter Spanish portfolio goals, processes, and logistics. Some suggestions included requesting students to correctly copy and identify all samples they present for evidence of individual work, as well as reducing all large artifacts such as posters to a manageable size.
- It is necessary to discuss in depth the process of selection of students' work (i.e., the number of products necessary to demonstrate a certain level of performance). Guidelines for the selection of material for Pacesetter Spanish portfolios need further clarification.
- More support should be given to teachers to better understand and teach Dimension 1 as defined in the Pacesetter Spanish portfolio assessment system. It is necessary to define what constitutes evidence of achievement in this area, how and where to find evidence, and how it can be collected by their students.
- Further discussion is warranted regarding the appropriateness of grouping the four linguistic skills in the two aspects of Dimension 2 of the Pacesetter Spanish portfolio assessment system.
- In order to apply the new revised version of the Pacesetter Spanish portfolio assessment matrix, a new research project should include an expanded portfolio reading with more raters and a larger number of portfolios. It would be beneficial for the five original teachers who participated in the Phase I reading to be incorporated in this expanded reading as experienced raters to help train the new teachers.
- The expanded portfolio reading should be carried out following the model delineated in the corresponding section of the Pacesetter Scoring Handbook.
- In addition to quantitative data, other qualitative data pertaining to discussion and debriefing sessions with participants should be collected during the Pacesetter Spanish portfolio reading.
- The new assessment matrix and guidelines for portfolios should be disseminated to all schools implementing the Pacesetter Spanish program so that teachers and students can use them.
- New benchmark portfolios from the expanded Pacesetter Spanish portfolio reading should be identified, and a set of sample portfolios should be prepared for use in future training sessions.

References

Arter, J.A. and Spandel, V. (1992). Using portfolios of student work in instruction and assessment. *Educational measurement: Issues and practice, Spring 92*, 36–44.

Bridgeman, B., Chittenden, E., and Cline, F. (1995). *Characteristics of a portfolio scale for rating early literacy*. Princeton, NJ: Educational Testing Service.
Educational Testing Service. (1995). Capturing the power of classroom assessment. *Focus 28*. Princeton, NJ.
Elliot, J. (1985). Facilitating action research in schools: Some dilemmas. In R. Burgess (Ed.), *Field methods in the study of education* (pp. 235–242). Lewes, England: Falmer Press.
ETS Trustees' Colloquy. (1995). *Performance Assessment: Different needs, difficult answers*. Princeton, NJ: Educational Testing Service.
Hardy, R.A. (1995). Examining the costs of performance assessment. *Applied measurement in education, 8*(2), 121–134.
Gifford, B.R. and O'Connor, M.C. (Eds.) (1992). *Changing assessments. Alternative views of aptitude, achievement and instruction*. Boston: Kluwer Academic Publishers.
Gitomer, D.H. and Duschl, R.A. (1995). *Moving toward a portfolio culture in science education*. Princeton, NJ: Educational Testing Service.
Koretz, Daniel. (1994). *The evolution of a portfolio program: The impact and quality of the Vermont portfolio program in its second year (1992–93)*. Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.
Koretz, Daniel. (1994). *Lessons from an evolving system. Interim report: The reliability of Vermont portfolio scores in the 1992–93 school year*. Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.
Koretz, Daniel. (1992). *Can portfolios assess student performance and influence instruction? The 1991–92 Vermont experience*. Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.
Koretz, Daniel. (1992). *The reliability of scores from the 1992 Vermont portfolio assessment program*. Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.
Mislevy, R.J. (1995). *On inferential issues arising in the California learning assessment system*. Princeton, NJ: Educational Testing Service.
Myford and Mislevy. (1995). *Monitoring and improving a portfolio assessment system*. Princeton, NJ: Educational Testing Service.
Myford, C.M., Marr, D.B., and Linacre, M. (1996). *Reader calibration and its potential role in equating for the test of written English*. Princeton, NJ: Educational Testing Service.
Resnick, L.B. and Resnick, D.P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B.R. Gifford, and M.C. O'Connor (Eds.), *Changing assessments. Alternative views of aptitude, achievement and instruction*. Boston: Kluwer Academic Publishers.
Sheingold, K. and Fredericksen, J. (1995). *Linking assessment with reform: Technologies that support conversations about student work*. Princeton, NJ: Educational Testing Service.
Sheingold, K., Heller, J.I., and Paulukonis, S.T. (1995): *Actively seeking evidence: Teacher change through assessment development*. Princeton, NJ: Educational Testing Service.
Shepard, L. (1989) Why we need better assessments. *Educational Leadership, (46 7),* 4–9.
Wolf, D., Bixby, J., Glenn, J., and Gardner, H. (1991). To use their minds well: Investigating new forms of student assessment. In G. Grant (Ed.), *Review of Educational Research 17,* 31–74. Washington, DC: American Educational Research Association.

Phase II Report

Rationale and Expected Outcomes

Pacesetter Spanish is a new and challenging third-level high school Spanish course developed by the College Board and Educational Testing Service (ETS) that combines a carefully integrated set of learning outcomes, course materials, and instructional experiences, including various approaches to assessment. Throughout the course, there is a provision for student self-assessment through learning logs and related journal activities, for peer assessment through work in pairs and small groups, for ongoing teacher-developed assessments, and for a standardized end-of-course assessment, as well as for students to assemble portfolios with products from their work during the year.

A portfolio is a collection of significant samples of student work over time, accompanied by clear criteria for evaluation, as well as the student's own reflections on his or her progress. Portfolios, although not replacing other types of assessment, give teachers and students another tool of great instructional and motivational value. The portfolios for Pacesetter Spanish may contain a rich array of samples, including videos of student reports and role playing, audio excerpts of informal and formal interviews, drafts and final versions of written work, and collages and descriptions of projects. In fact, many of the activities and final projects for the units are acceptable choices for inclusion in student portfolios.

In the case of Pacesetter Spanish, portfolios are to be used for assessment purposes, not to look at progress but to evaluate results, thus also serving to inform and give feedback on instructional materials and methods. Portfolios yield valuable data about students' achievement and effectively lead students through the process of acquiring a new language. They serve as a guide for students as they make choices and demonstrate how they reason, create, strategize, and reflect. Ultimately, portfolios allow students to take on responsibility for their own learning process and offer evidence of their

progress toward meeting the course outcomes. These portfolios help teachers and students be reflective of their work, assess their effectiveness as teachers and learners, and reshape their approach to the course materials. Sessions on assessing these student portfolios have become an integral part of the professional development institutes for teachers taking place each summer in different areas of the country.

Research is needed to ensure that it is appropriate to include portfolio assessment as an important part of the integrated instruction-based assessment component. This study is designed to facilitate the practice of developing and assessing portfolios and to ensure that the procedures being taught at teacher institutes will lead to accurate, fair assessment of portfolios. The specific goals are (1) to prepare a working definition of Pacesetter Spanish portfolios as assessment tools; (2) to develop a standardized portfolio assessment system; and (3) to use project results as a training tool for future professional development sessions for teachers. In summary, the ultimate goal is to enhance classroom practice and assessment.

In Phase I of the Pacesetter Spanish Portfolio Research Project, carried out in 1995–96, the following issues were considered: (a) the role of the portfolio as an assessment tool, its working definition and its appropriateness within the course framework; (b) the development of scoring rubrics and an assessment matrix, taking into account course outcomes, performance indicators, and scoring rubrics for the four language proficiency skills (reading, writing, listening, and speaking); (c) the student's perspective (cover sheet and guidelines); (d) discussion and revision of the assessment matrix; (e) reading portfolios; (f) debriefing session; and (g) preparation of training packets for 1996 Professional Development Summer Institutes for Teachers.

Both the 1996 Professional Development Summer Institutes for Teachers and midyear meetings included sessions on portfolios, and the samples chosen for the training packets served to disseminate the Pacesetter Spanish portfolio model.

The final draft of the corresponding report was produced, and the assessment matrix needed to be refined accordingly.

For Phase II of the Pacesetter Spanish Portfolio Research project the following tasks were planned:

- In spring 1997, in consultation with the teachers who participated in Phase I of this project, and taking into account Phase I results, the assessment matrix was refined.
- In summer 1997, during the three-day refresher session at the Delaware Summer Institute for Teachers, a portfolio reading was held. The new matrix was piloted by the participating teachers. Since it is essential that readers develop a shared understanding of the scoring criteria and learn to apply the criteria consistently, benchmark portfolios from Phase I were selected for training purposes. After the initial training, the reading was carried out following the model delineated in the portfolio section of the Pacesetter Scoring Handbook (see Appendix A). In addition to tabulation of scores, other data (for example, notes from discussion with teachers, minutes from the debriefing after reading, etc.) were collected during the reading. The final version of the revised portfolio scoring matrix and guidelines was refined for immediate dissemination to participating schools in order for Pacesetter Spanish teachers to start applying them in the 1997–98 school year.
- In fall 1997, the compilation of the results of this reading was analyzed. Benchmark portfolios from Phase II were identified, and a set of sample portfolios was developed to be used in the 1997 midyear professional development training sessions for teachers.

After Phase I, initiated in 1995–96, a new assessment matrix was prepared, ready to be further refined. After Phase II, this new and revised matrix was applied and disseminated for further use by Pacesetter Spanish classroom teachers.

Through discussions with participants to determine whether the portfolios are a valid reflection of student achievement, and by calculating how reliably the portfolios can be scored, we expect to have a basis for expanding the use of portfolio assessment as an integral part of the Pacesetter Spanish program.

Pacesetter Spanish Portfolio Research Project (PHASE II)

Method

Andrea Fercsey and Carmen Luna, of the Assessment Division, coordinated this phase of the research with the help of Pablo Aliaga, a graduate student at the University of Michigan, who was the ETS summer intern assigned to work on this project. David Baum of the Assessment Division also assisted at the portfolio reading.

Portfolios were evaluated on the campus of the University of Delaware from July 10 through 13, 1997. Six schools sent in work by students enrolled in Pacesetter Spanish during the 1996–97 school year. Seven teachers participated in assessing these portfolios; two of them had prior experience scoring Pacesetter Spanish portfolios. Not every school that sent in portfolios had teachers participating in the assessment of these portfolios. Three others assisted in the assessment of the

portfolios. All three worked for ETS; two with prior experience scoring Pacesetter Spanish portfolios.

Ninety-five portfolios were designated to be assessed at the beginning of the grading session. Of these, a total of 82 were read. Thirteen portfolios were rendered unreadable due to missing data (i.e., failure to incorporate cover sheets). Cover sheets (see Appendix B) are essential to the evaluation of portfolios because they contain a rationale of the choice of a particular document as evidence of a specific dimension and aspect. Of the 82 portfolios that were read, 51 were read once, and 31 were reread by a different rater. Hence, a total of 113 readings were completed. Ratings were provided for the three aspects of Dimension 1, "Demonstrating Knowledge of Hispanic Cultures" (*Showing awareness of the diversity of Hispanic cultures, Identifying contributions of Hispanic figures, and Making connections with other disciplines and own culture(s)*) and the two aspects of Dimension 2, "Using Spanish to Communicate Effectively" (*Deriving meaning from texts and personal interactions*; i.e., receptive skills such as reading and listening, and *Expressing meaning in oral and written form*, i.e., productive skills such as speaking and writing). Raters used the Pacesetter proficiency categories ("Beginning," "Developing," "Promising," "Accomplished," and "Advanced") as well as "Not Enough Evidence to Judge" and "Missing" to rate the portfolios. Appendix C provides the scoring rubrics for each aspect. Appendix D provides rater comments about rubrics and portfolios.

Summary of Ratings

Table 1 shows the frequency of ratings in each category. Of the 882 ratings, 31.4 percent were "Promising." The majority of ratings (58.2 percent) were either "Developing" or "Promising." Last year 66 percent of the ratings fell within the "Promising" and "Accomplished" categories and 2.42 percent were "Advanced." However, this year less than 1 percent of the students received a rating of "Advanced." Finally, 5.4 percent of the ratings assigned were "No Evidence" due to lack of artifacts.

On average, this year's students received lower ratings than last year's students. In 1996, 68 percent of the ratings were 3 or higher. In 1997, only 45 percent of the ratings were 3 or higher. This year's students may have been less able compared to last year's students, the raters could have been more severe in their grading, or the raters may have been better instructed in scoring this year because of the previous year's study.

Also, there was a significant decrease in the use of the "No Evidence" category. This suggests improvement in teacher and student understanding of what is needed to evaluate the portfolios: In a portfolio system like this one, the selection process is complicated and revealing. Students need to develop a deep understanding about the nature of quality in their work and how the selected pieces constitute evidence for a particular dimension or aspect. Teachers also need to be fully cognizant of all possible ways to show evidence of achievement. Consequently, the fact that portfolios in 1997 contained more complete evidence of students' work could be interpreted as a sign of better implementation in the actual classrooms.

Table 2 provides the frequency of ratings for each aspect of Dimension 1, "Demonstrating Knowledge of Hispanic Cultures."

Aspects 1 and 3 have similar rating distributions. Further, the proportion of "No Evidence" ratings in the three aspects of Dimension 1 are more uniform this year than last year. Last year "No Evidence" had quite

TABLE 1

Frequency of Ratings in Each Rating Category*

Category	Category Name	Frequency	Percentage
1	Beginning	110	12.5
2	Developing	236	26.8
3	Promising	277	31.4
4	Accomplished	112	12.7
5	Advanced	8	0.9
8	No Evidence	48	5.4
9	Missing	91	10.3
	TOTAL	882	

*Data include 13 portfolios that were not evaluated (reported as "Missing").

TABLE 2

Frequency of Ratings in Each Category for the Aspects of Dimension 1*

Category	1	2	3	4	5	8	9	
Category Name	Beginning	Developing	Promising	Accomplished	Advanced	No Evidence	Missing	TOTAL
Aspect 1								
Frequency	20	38	31	15	1	8	13	126
%	15.9	30.2	24.6	11.9	1	6.5	10.3	
Aspect 2								
Frequency	17	31	40	13	0	12	13	126
%	13.5	24.6	31.8	10.3	0	9.5	10.3	
Aspect 3								
Frequency	24	28	41	12	0	8	13	126
%	19.1	22.2	32.5	9.5	0	6.5	10.3	
GENERAL								
Frequency	18	39	39	12	0	5	13	126
%	14.3	31.0	31.0	9.5	0	4.0	10.3	
TOTAL								
Frequency	79	136	151	52	1	33	52	504
%	15.7	27.0	30.0	10.3	2.0	6.5	10.3	

*Data include 13 portfolios that were not evaluated (reported as "Missing").
Aspect 1: *Showing awareness of the diversity of Hispanic cultures*
Aspect 2: *Identifying contributions of Hispanic figures*
Aspect 3: *Making connections with other disciplines and own culture(s)*

different response rates. In 1996, over 70 percent of the ratings of Aspect 3 were "No Evidence"; by comparison, only 6 percent of the ratings of Aspect 3 were "No Evidence" this year.

Table 3 shows the frequency distribution of ratings for Dimension 2, "Using Spanish to Communicate Effectively."

Aspect 1 has a higher percentage of "No Evidence" and "Beginning" ratings, and a lower percentage of "Developing" ratings than Aspect 2 or the general dimension ratings. The distribution of ratings for "No Evidence" in the three categories is quite similar to that of last year.

A comparison of Dimension 1 and 2 ratings shows that similar percentages (57 percent to 60 percent) of the ratings were "Developing" or "Promising." For the "Beginning" rating, Dimension 1 has approximately 16 percent compared to 8 percent for Dimension 2. For the "Accomplished" rating, Dimension 1 has approx-

TABLE 3

Frequency of Ratings in Each Category for Aspects of Dimension 2*

Category	1	2	3	4	5	8	9	
Category Name	Beginning	Developing	Promising	Accomplished	Advanced	No Evidence	Missing	TOTAL
Aspect 1								
Frequency	16	26	40	19	1	11	13	126
%	12.7	20.6	31.8	15.1	1	8.7	10.3	
Aspect 2								
Frequency	6	37	43	23	3	1	13	126
%	4.7	29.4	34.1	18.3	2.4	1	10.3	
GENERAL								
Frequency	9	37	43	18	3	3	13	126
%	7.1	29.4	34.1	14.3	2.4	2.4	10.3	
TOTAL								
Frequency	31	100	126	60	7	15	39	378
%	8.2	26.5	33.3	15.9	1.9	4.0	10.3	

*Data include 13 portfolios that were not evaluated (reported as "Missing").
Aspect 1: *Deriving meaning from printed materials and oral discourse*
Aspect 2: *Expressing meaning in oral and written form*

imately 10 percent compared to 16 percent for Dimension 2. More "Advanced" ratings were given for Dimension 2 than for Dimension 1. Finally, more "No Evidence" ratings were given for Dimension 1 than for Dimension 2.

Time Used to Read Portfolios

Table 4 is a summary of the time spent on portfolios by each rater. The table provides the number of portfolios each rater read and the total time, average time, standard deviation, minimum, and maximum time it took the raters to evaluate portfolios.

On average, raters read 11 portfolios, each portfolio taking about 45 minutes to evaluate. Average time spent on portfolios ranged from 34 to 60 minutes, the range of time for each individual portfolio being 17 to 150 minutes. Average time spent per portfolio was similar to last year's average of 47 minutes.

Distribution of Ratings for Each Rater

Table 5 shows the distribution of ratings for each rater. Only two raters gave the "Advanced" proficiency rating. This may be an indication of disparate use of the

TABLE 4

Time Employed by Each Rater in Minutes

Rater	Portfolios Read During Session	Total Time	Mean	Standard Deviation	Minimum	Maximum	Median
Rater 1	11	560	51	36	17	150	40
Rater 2	10	470	47	17	30	70	43
Rater 4	10	504	50	16	32	80	46
Rater 5	14	479	34	9	20	50	35
Rater 7	13	460	35	17	18	79	30
Rater 8	9	442	49	20	25	80	52
Rater 9	16	954	60	20	34	100	56
Rater 10	9	410	46	16	25	80	40
Rater 11	16	555	35	9	25	60	30
Rater 12	5	240	48	17	25	70	50

rubrics by raters. Rater 7 gave more than one-fifth of the portfolios he/she read a "No Evidence" rating. No other rater gave more than one-sixth of the portfolios they read a rating of "No Evidence." Rater 7 never gave a rating higher than "Promising." Five raters gave approximately 50 percent of the portfolios they read a "Promising" rating. Eighty-five percent of the ratings Rater 5 gave were either "Promising" or "Developing." Consequently, it appears that rater 5 tends to rate more leniently than the other raters.

Portfolios Read Twice

Table 6 shows that of the 31 portfolios that were double read, 8 were given the same two total scores for both dimensions. After evaluating all aspects of a particular dimension, raters had to give a holistic rating to the whole dimension. Of the remaining portfolios, approximately 39 percent had a score that differed by only one point on either Dimension 1 or 2 or both. Only one portfolio had a difference in score of 3 points for a given dimension. The average absolute difference in time to score the portfolios was 15 minutes. However, this number is exaggerated by an apparent outlier of 105 minutes. The average absolute difference in time after removing the apparent outlier is 12 minutes.

Table 7 examines the discrepancies between raters within each aspect of both dimensions. Overall, 39 percent of ratings were identical. Another 40 percent of their ratings differed by a single point. Thirteen percent of the ratings differed by 2 or more points. Finally, in 10 percent of the cases, one rater gave a "No Evidence" rating and the other rater believed there was sufficient evidence to score the portfolio on that aspect. This is fewer than last year.

Table 8 shows where specific discrepancies between raters occurred. Thirty-two percent of the discrepancies involved a rating of "Developing" and a rating of "Promising." A year ago the greater number of discrepancies involved a rating of "Promising" and a rating of "Accomplished." Nineteen percent of the discrepancies involved ratings of "Beginning" and "Developing." Only 6 percent of the discrepancies involved a rating of "Beginning" and a "No Evidence" rating. Aspect 2 of Dimension 2 registered the fewest cases of discrepancies.

TABLE 5

Total Count and Relative Frequency By Category and Rater

	Rater 1	Beginning	Developing	Promising	Accomplished	Advanced	No Evidence	Total
Rater 1	Frequency	7	21	21	21	4	3	77
	%	9	27	27	27	5	4	
Rater 2	Frequency	6	16	34	14	0	0	70
	%	9	23	49	20	0	0	
Rater 4	Frequency	25	24	14	4	0	3	70
	%	36	34	20	6	0	4	
Rater 5	Frequency	8	36	47	6	0	1	98
	%	8	37	48	6	0	1	
Rater 7	Frequency	18	27	26	0	0	20	91
	%	20	30	29	0	0	22	
Rater 8	Frequency	8	28	17	5	0	5	63
	%	13	44	27	8	0	8	
Rater 9	Frequency	24	37	22	27	0	2	112
	%	22	33	20	24	0	2	
Rater 10	Frequency	10	20	14	9	4	6	63
	%	16	32	22	14	6	10	
Rater 11	Frequency	1	20	65	23	0	3	112
	%	1	18	58	21	0	3	
Rater 12	Frequency	3	7	17	3	0	5	35
	%	9	20	49	9	0	14	

TABLE 6

Comparison of Portfolios and Raters

Student	Rater	Time in Min.	Total Dim. 1	Total Dim. 2	Absolute Difference in Time	Absolute Difference* Dim 1	Absolute Difference* Dim 2
1A3	9	56	1	2	16	2	0
1A3	11	40	3	2			
1A5	7	30	1	8	5	2	Undefined**
1A5	8	25	1	2			
1A6	5	30	3	3	10	Undefined	2
1A6	7	20	8	1			
1A7	4	47	1	1	12	1	1
1A7	5	35	2	2			
1A8	7	18	8	1	7	Undefined	Undefined
1A8	12	25	2	8			
1A9	2	30	1	2	20	0	0
1A9	12	50	1	2			
2B1	8	65	2	2	15	0	0
2B1	7	50	2	2			
2B4	1	40	1	5	8	0	3
2B4	4	32	1	2			
2B6	11	30	8	3	6	Undefined	1
2B6	9	36	2	2			
2B7	7	45	3	3	5	0	0
2B7	5	50	3	3			
2B11	12	55	3	3	25	0	0
2B11	11	30	3	3			
2B12	8	60	2	3	26	0	1
2B12	7	34	2	2			
4D3	7	29	2	8	1	0	Undefined
4D3	8	30	2	2			
4D4	4	42	2	2	3	0	0
4D4	5	45	2	2			
4D7	10	35	2	3	0	Undefined	0
4D7	7	35	8	3			
4D10	1	17	2	2	13	1	1
4D10	11	30	3	3			
4D11	9	80	3	3	50	1	1
4D11	11	30	4	4			
4D12	2	35	2	3	7	1	1
4D12	1	28	3	2			
5E2	4	75	1	2	20	2	0
5E2	10	55	3	2			

TABLE 6 (*continued)*

5E6	8	80	3	3	1	1	0
5E6	7	79	2	3			
5E9	1	60	4	4	0	0	0
5E9	11	60	4	4			
5E12	2	45	4	4	105	1	1
5E12	1	150	3	5			
5E13	9	70	4	4	10	0	0
5E13	2	60	4	4			
5E14	2	60	3	3	30	1	1
5E14	9	90	4	4			
5E15	2	70	3	3	10	1	0
5E15	4	80	2	3			
6F3	5	35	3	3	16	1	1
6F3	7	51	2	2			
6F7	11	30	3	3	15	0	0
6F7	1	45	3	3			
6F9	11	25	3	3	9	1	1
6F9	9	34	2	2			
6F14	11	30	3	3	10	2	2
6F14	9	40	1	1			
6F15	10	45	2	3	5	0	2
6F15	4	50	2	1			
6F20	4	45	1	2	15	1	1
6F20	5	30	2	3			

*Absolute Difference = Number indicates the difference between the two ratings given by separate raters to the same portfolio.

**Undefined = Difference in rating could not be calculated because one rater considered that there was not enough evidence to judge a portfolio and the other rater gave the portfolio a rating (usually a rating of 2 or 3).

TABLE 7

Discrepancies Between Raters

	Dimension 1				Dimension 2				
	Aspect 1	Aspect 2	Aspect 3	Total	Aspect 1	Aspect 2	Total	TOTAL	%
Same Rating	10	12	12	13	10	14	13	84	39
1-pt Diff.	16	13	11	11	11	13	11	86	40
2-pt Diff.	3	4	4	3	5	3	3	25	12
3-pt Diff.	0	0	0	0	0	0	1	1	1
Evid. Vs No Evid.	2	2	4	4	5	1	3	21	10

Dimension 1, Demonstrating Knowledge of Hispanic Cultures
Aspect 1: *Showing awareness of the diversity of Hispanic cultures*
Aspect 2: *Identifying contributions of Hispanic figures*
Aspect 3: *Making connections with other disciplines and own culture(s)*
Dimension 2, Using Spanish to Communicate Effectively
Aspect 1: *Deriving meaning from printed materials and oral discourse*
Aspect 2: *Expressing meaning in oral and written form*

TABLE 8

Specific Discrepancies Between Raters

	Dimension 1				Dimension 2				
Discrepancy	Aspect 1	Aspect 2	Aspect 3	Total	Aspect 1	Aspect 2	Total	TOTAL	%
No Evid.- 1	1	0	0	0	2	1	1	5	.03
No Evid.- 2	1	1	1	3	0	0	2	8	.06
No Evid.- 3	0	1	3	1	3	0	0	8	.06
No Evid.- 4	0	0	0	0	0	0	0	0	.00
No Evid.- 5	0	0	0	0	0	0	0	0	.00
1 – 2	5	4	8	2	4	1	1	25	.19
1 – 3	3	2	3	3	3	2	2	18	.16
1 – 4	0	0	0	0	0	0	0	0	.00
1 – 5	0	0	0	0	0	0	0	0	.00
2 – 3	10	6	3	6	5	6	7	43	.32
2 – 4	0	2	1	0	2	0	0	5	.03
2 – 5	0	0	0	0	0	0	1	1	.01
3 – 4	1	3	0	3	2	5	3	17	.13
3 – 5	0	0	0	0	0	1	0	1	.01
4 – 5	0	0	0	0	0	1	1	2	.02
TOTAL	21	19	19	18	21	17	18		

Dimension 1, Demonstrating Knowledge of Hispanic Cultures
Aspect 1: *Showing awareness of the diversity of Hispanic cultures*
Aspect 2: *Identifying contributions of Hispanic figures*
Aspect 3: *Making connections with other disciplines and own culture(s)*
Dimension 2, Using Spanish to Communicate Effectively
Aspect 1: *Deriving meaning from printed materials and oral discourse*
Aspect 2: *Expressing meaning in oral and written form*

TABLE 9

Bivariate Correlation Matrix

	Dimension 1 Aspect 1	*Dimension 1 Aspect 2*	*Dimension 1 Aspect 3*	*Dimension 1 Total*	*Dimension 2 Aspect 1*	*Dimension 2 Aspect 2*	*Dimension 2 Total*
Dimension 1 Aspect 1	1.000	.568	.555	.702	.398	.264	.429
Dimension 1 Aspect 2		1.000	.611	.830	.606	.511	.576
Dimension 1 Aspect 3			1.000	.795	.573	.359	.525
Total Dimension 1				1.000	.642	.543	.653
Dimension 2 Aspect 1					1.000	.615	.842
Dimension 2 Aspect 2						1.000	.835
Total Dimension 2							1.000

Dimension 1, Demonstrating Knowledge of Hispanic Cultures
Aspect 1: *Showing awareness of the diversity of Hispanic cultures*
Aspect 2: *Identifying contributions of Hispanic figures*
Aspect 3: *Making connections with other disciplines and own culture(s)*
Dimension 2, Using Spanish to Communicate Effectively
Aspect 1: *Deriving meaning from printed materials and oral discourse*
Aspect 2: *Expressing meaning in oral and written form*

Correlation Among the Ratings of Portfolios

Each portfolio received seven ratings: one rating for each of the three aspects of Dimension 1, one for each of the two aspects of Dimension 2, and two total ratings. Table 9 summarizes the correlations among the various ratings. Correlations among the three aspects in Dimension 1 range from .555 to .611, and the correlation between the two aspects of Dimension 2 is .615. The correlation between the three aspects of Dimension 1 and the total for Dimension 1 range from .702 to .830. The correlation between the two aspects of Dimension 2 and the total for Dimension 2 are .842 and .835, respectively. The correlation between the two dimensions is moderately high at .653. With the exception of the correlations between aspect ratings and total dimension ratings, the correlations are all higher than last year.

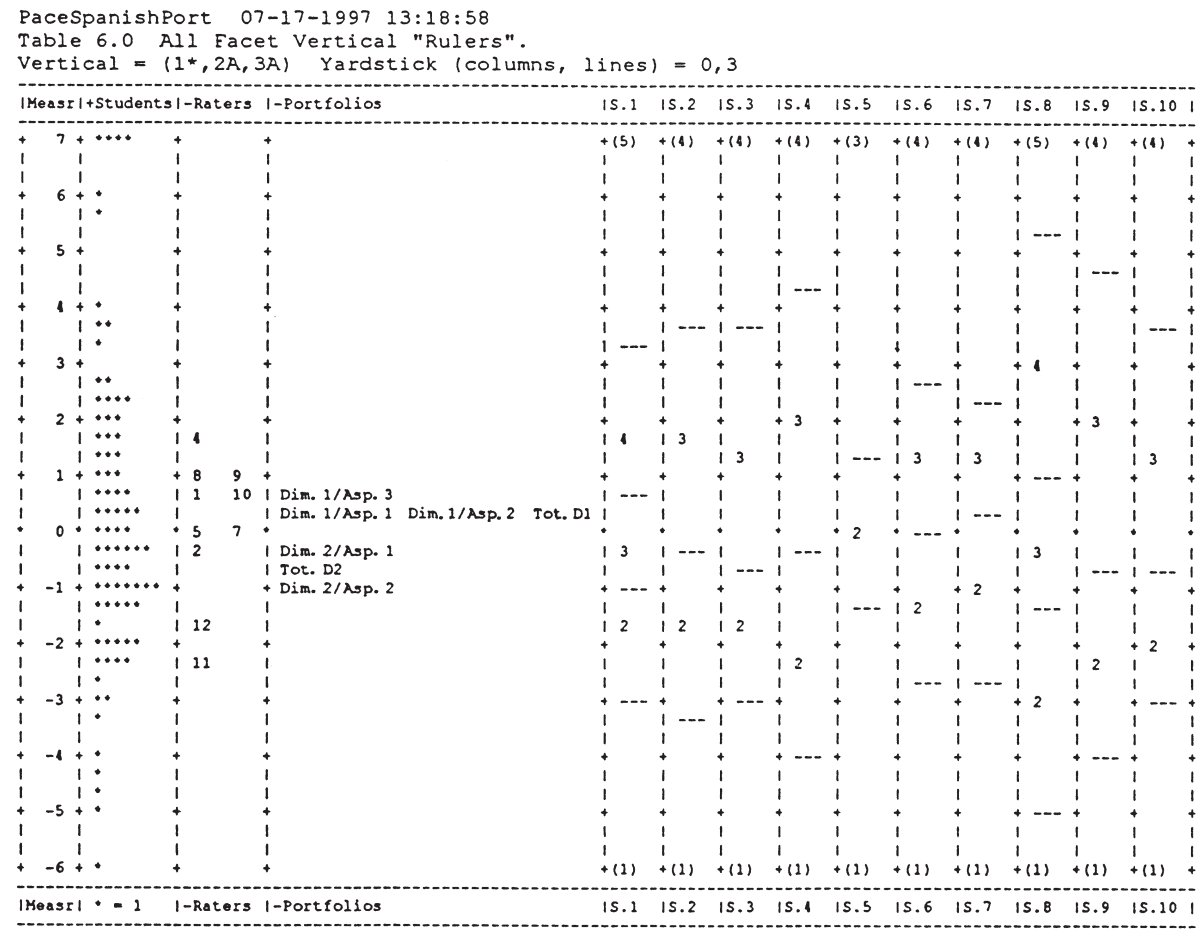
Reliability Study

As was done last year, FACETS (Linacre, 1989) was used to summarize the data. FACETS constructs linear measures from qualitatively ordered counts by means of faceted Rasch analysis. Each observation is the outcome of an interaction between elements of facets (e.g., a "student's" portfolio evaluated by a "rater" on several "aspects"). Each observation provides FACETS with information about the elements that interact to construct it. From this information, FACETS estimates a quantitative measure for each element of each facet (e.g., each student, each rater, each portfolio aspect). The measures for the elements obtained from one analysis are all in the same linear frame of reference on one common interval scale.

For this study, two partial credit models and a rating scale model were used. The partial credit models were constrained with regard to rater and aspects of the portfolio. The partial credit model for raters allows us to compare the raters. The partial credit model for aspects allows us to compare the rating scales for the aspects. Finally, the rating scale model uses all the information together to compare students, raters, and aspects of the portfolio.

The output in FACETS for these models is a set of rulers that provides a graphical description of the variable. A "+" or "–" before the facet name indicates whether the facet measures are positively or negatively oriented. The vertical axis provides a linear definition of the variable. Each element name is positioned according to its measure. Elements with extreme scores are not positioned by measure but are placed at the extreme top or bottom of the column of their facet.

Figure 1 shows the results obtained from using the partial credit model for raters. The figure reveals that not every rater is rating students in the same way. It appears that Rater 7 (R.7) is using a scale that ranges from 1 to 3. Most others are using a scale that ranges from 1 to 4. Finally, Raters 1 and 10 appear to be using the full range of ratings offered, 1 to 5. Figure 1 also indicates that Raters 11 and 12 are grading most leniently, and Rater 4 is grading most harshly. Figure 1 indicates that students tend to receive higher ratings on aspects from Dimension 1, "Demonstrating Knowledge of Hispanic Cultures," than those from Dimension 2, "Using Spanish to



Dimension 1, Demonstrating Knowledge of Hispanic Cultures

Aspect 1: *Showing awareness of the diversity of Hispanic cultures*

Aspect 2: *Identifying contributions of Hispanic figures*

Aspect 3: *Making connections with other disciplines and own culture(s)*

Dimension 2, Using Spanish to Communicate Effectively

Aspect 1: *Deriving meaning from printed materials and oral discourse*

Aspect 2: *Expressing meaning in oral and written form*

Figure 1. FACETS: Partial credit model for raters.

Communicate Effectively." This may be explained by the fact that teachers have more experience rating linguistic skills than the concepts included in the aspects for "Demonstrating Knowledge of Hispanic Cultures."

Table 10 shows each rater's observed average (i.e., average of the raw score rating given) and his or her "fair" average (the observed average adjusted for the deviation of the portfolio in each rater's sample from the overall mean of the portfolios across all raters). The rater separator index of 5.24 indicates that the ten raters can be separated into five statistically distinct severity strata. These results suggest that the raters do not use the scoring rubrics in the same manner; some tend to rate more leniently than others.

The second model used was the partial credit model for aspects. The results obtained from using that model are summarized in Figure 2. Figure 2 indicates that the rating scale for one aspect functions somewhat differently from another. The aspects generally fall into two rating scale groups. Group 1 consists of Aspects 1, 2, and 3 for Dimension 1, and total for Dimension 1 as well as Aspect 1 for Dimension 2. These aspects appear to be rated on a 1 to 4 scale. All these aspects refer either to receptive language skills (*Deriving meaning from printed*

TABLE 10

Rater Severity

PaceSpanishPort 07-17-1997 13:18:58
Table 7.2.1 Raters Measurement Report (arranged by FN)

Obsvd Score	Obsvd Count	Obsvd Average	Fair Avrge	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	PtBis	Nu	Raters
129	64	2.0	2.01	-.02	.22	0.8	-1	0.8	-1	.37	7	7
135	58	2.3	2.03	1.13	.22	0.8	-1	0.8	-1	.41	8	8
131	67	2.0	2.12	1.57	.21	0.9	0	0.9	0	.44	4	4
80	30	2.7	3.07	-1.80	.33	0.9	0	0.9	0	.46	12	12
142	51	2.8	2.86	.51	.25	0.9	0	1.0	0	.58	10	10
300	102	2.9	3.06	-2.49	.22	1.0	0	1.0	0	.42	11	11
245	97	2.5	2.65	-.14	.21	1.1	0	1.1	0	.44	5	5
168	63	2.7	2.69	-.25	.24	1.1	0	1.1	0	.51	2	2
216	96	2.3	2.08	.89	.18	1.1	0	1.0	0	.53	9	9
216	74	2.9	2.73	.60	.16	1.1	0	1.1	0	.32	1	1
176.2	70.2	2.5	2.53	.00	.22	1.0	-0.1	1.0	-0.1	.45	Mean (Count: 10)	
63.0	21.5	0.3	0.41	1.21	.04	0.1	0.8	0.1	0.7	.07	S.D.	

RMSE (Model) .23 Adj S.D. 1.19 Separation 5.24 Reliability .96
Fixed (all same) chi-square: 283.0 d.f.: 9 significance: .00
Random (normal) chi-square: 9.0 d.f.: 8 significance: .35

materials and oral discourse) or a conceptual introspection of different elements of the Spanish-speaking world and cultures ("Demonstrating Knowledge of Hispanic Cultures"). In both cases, it is harder to apply judgement or show evidence for the type of documents or behavior that demonstrate evidence of achievement of these. Group 2 consists of Aspect 2 for Dimension 2 (*Expressing meaning in oral and written form*) and total for Dimension 2, "Using Spanish to Communicate Effectively." These aspects appear to be rated on a 1 to 5 scale, although a rating of 5 is difficult to achieve. These aspects include productive skills, which are usually measured and evaluated in language classrooms. Students and teachers have more experience with the kind of evidence that is necessary to show achievement in these aspects.

The final model is the rating scale model. This model does not constrain or condition any facet. This model would be valid if the partial-credit models showed that the raters used the aspect rating scales in a similar fashion, and that the aspect scales functioned similarly. In our case they don't; consequently, the rating scale model is used with caution. The results obtained using this model are shown in Figure 3. Figure 3 indicates that students tend to get higher ratings on Dimension 1 than Dimension 2. It appears that there were three distinct strata of rater severity. Raters 4, 7, 8, and 9, who were the most harsh, form the first stratum. Raters 2, 5, 10, and 1 form a second stratum. Raters 11 and 12, the most lenient, form a third stratum. The distribution of students appears skewed, with some outliers at the top.

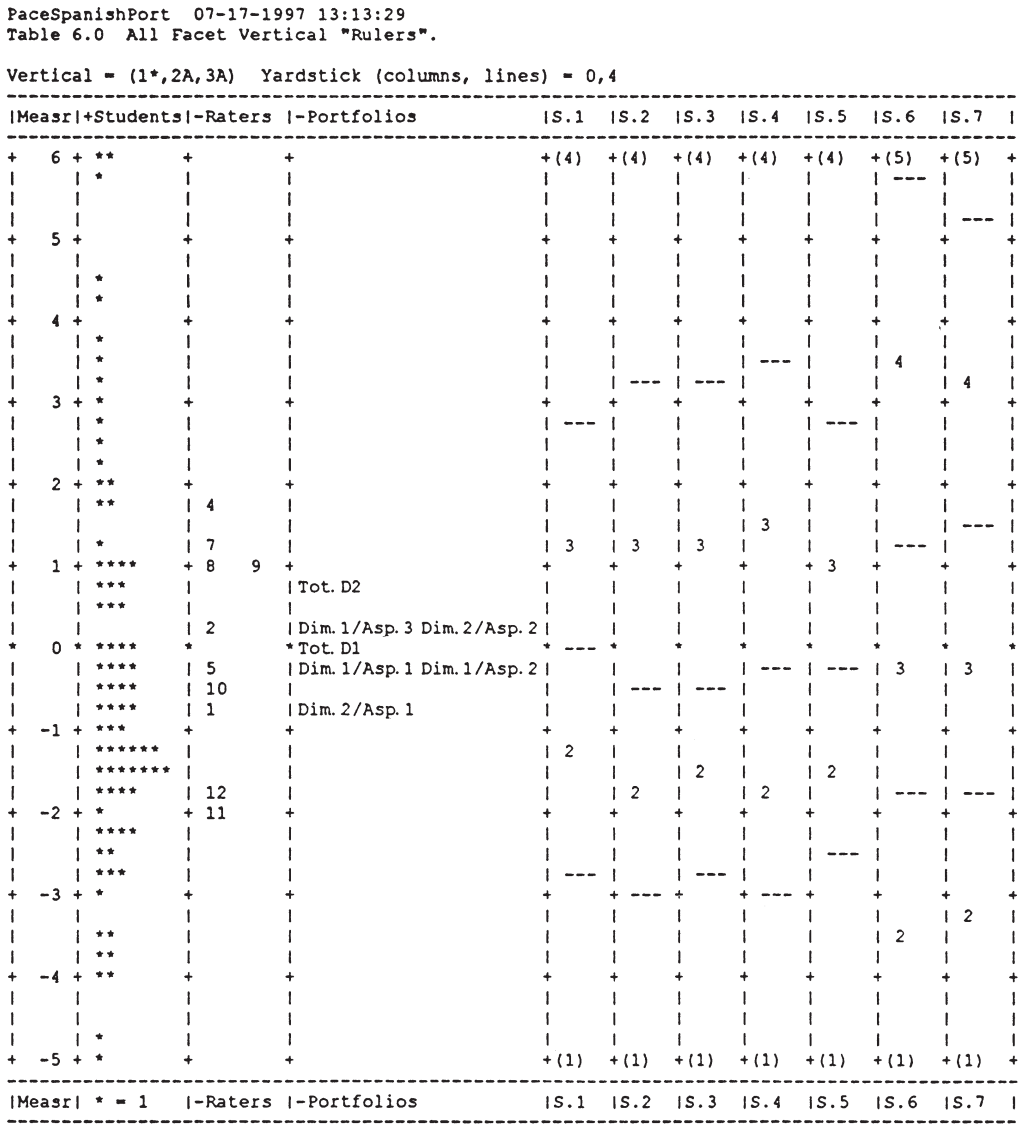
Figure 3A shows that differences between the students' observed averages and fair averages indicate that 20 students (12, 17, 18, 20, 23, 24, 25, 26, 30, 32, 36, 37, 40, 44, 48, 52, 57, 70, 73, and 79) would have had their total scores changed if we had adjusted their scores for rater severity. This rate of 20 is less than the 29 rate of change last year. It should be noted that the 14 students with asterisks in the observed score column have unusual patterns of ratings (i.e., more than typical variations in their ratings). In each case, there are one or more unexpected ratings—aberrant ratings that don't seem to "fit" with the others. A mean-square infit or outfit value greater than 1.5 indicates a student's portfolio might need another review before a score report is issued, particularly if the student's total score is near a critical decision-making point in the score distribution.

Analyses

We ran two FACETS analyses: one on the ratings of the 21 portfolios read twice in the June 1996 portfolio reading, and one on the ratings of the 31 portfolios read twice in the June 1997 portfolio reading. We then examined selected pieces of output from the two analyses to see how they compared.

Students

Each of the portfolios received 14 ratings (i.e., seven ratings each from two raters). For each portfolio, FACETS averaged the ratings to compute a total score. We then compared the range of total scores across years. The range of portfolio scores was somewhat wider in 1997 than in 1996. The total scores for the 21 portfolios rated in 1996 ranged from 2.4 to 4.3. By comparison, the total scores for the 31 portfolios read in 1997 ranged from 1.5 to 4.0.



Dimension 1, Demonstrating Knowledge of Hispanic Cultures

Aspect 1: *Showing awareness of the diversity of Hispanic cultures*

Aspect 2: *Identifying contributions of Hispanic figures*

Aspect 3: *Making connections with other disciplines and own culture(s)*

Dimension 2, Using Spanish to Communicate Effectively

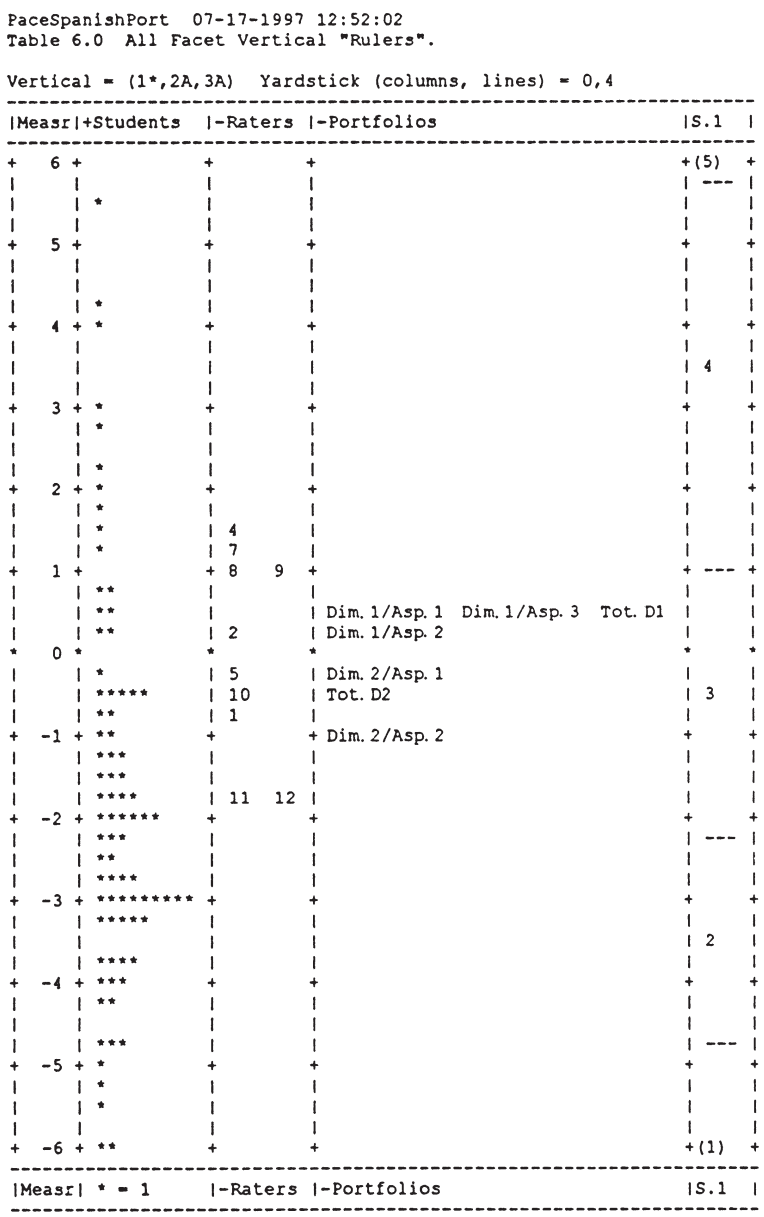
Aspect 1: *Deriving meaning from printed materials and oral discourse*

Aspect 2: *Expressing meaning in oral and written form*

Figure 2. FACETS: Partial credit model for aspects.

On average, students did better in 1996 than in 1997. The average total score for the 21 portfolios rated in 1996 was 3.3 (SD = 0.4); the average total score for the 31 portfolios rated in 1997 was 2.5 (SD = 0.6).

When FACETS calculates "fair averages" for students, it adjusts each student's total score for the severity or leniency of the two raters who rated that particular student. On average, the amount of difference between the "observed average" (i.e., the unadjusted raw score average of the raters' ratings of a given student) and the "fair average" (i.e., the observed average adjusted for rater severity or leniency differences) was 0.2 in both 1996 and 1997. The largest severity or leniency adjustment for any portfolio was 0.7 in 1997 and 0.5 in 1997.



Dimension 1, Demonstrating Knowledge of Hispanic Cultures

Aspect 1: *Showing awareness of the diversity of Hispanic cultures*

Aspect 2: *Identifying contributions of Hispanic figures*

Aspect 3: *Making connections with other disciplines and own culture(s)*

Dimension 2, Using Spanish to Communicate Effectively

Aspect 1: *Deriving meaning from printed materials and oral discourse*

Aspect 2: *Expressing meaning in oral and written form*

Figure 3. FACETS: Rating scale model.

PaceSpanishPort 07-17-1997 12:52:02
Table 7.1.1 Students Measurement Report (arranged by FN).

Obsvd Score	Obsvd Count	Obsvd Average	Fair Avrge	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	PtBis	Nu	Students
28	7	4.0	3.66	1.69	.92	0.0	-2	0.0	-2	.00	50	5E7
28	7	4.0	3.89	2.77	.92	0.0	-2	0.0	-2	.00	28	4D1
28	7	4.0	4.15	4.33	.92	0.0	-2	0.0	-2	.00	47	5E4
56	14	4.0	4.10	4.04	.65	0.1	-3	0.1	-3	.00	56	5E13
21	7	3.0	3.33	.55	.70	0.1	-2	0.1	-2	.00	61	5E18
14	7	2.0	1.87	-3.77	.61	0.1	-3	0.1	-3	.00	14	1A14
25	10	2.5	2.45	-2.28	.53	0.2	-3	0.2	-3	.64	20	2B6
16	7	2.3	2.65	-1.75	.61	0.2	-2	0.2	-2	.39	79	6F17
16	7	2.3	2.16	-3.03	.61	0.2	-2	0.2	-2	.58	43	4D16
22	7	3.1	3.05	-.44	.70	0.2	-2	0.2	-2	.46	41	4D14
41	14	2.9	2.55	-2.01	.48	0.3	-2	0.2	-2	.09	69	6F7
20	7	2.9	2.76	-1.40	.68	0.4	-1	0.4	-1	.22	81	6F19
20	7	2.9	2.26	-2.80	.68	0.4	-1	0.4	-1	.10	23	2B9
9	6	1.5	1.31	-5.48	.74	0.4	-1	0.4	-1	.61	24	2B10
18	7	2.6	2.70	-1.60	.64	0.4	-1	0.4	-1	.29	80	6F18
18	7	2.6	1.91	-3.67	.64	0.4	-1	0.4	-1	.20	1	1A1
36	14	2.6	2.92	-.88	.45	0.4	-1	0.4	-2	.22	49	5E6
16	6	2.7	2.03	-3.36	.70	0.5	-1	0.4	-1	.32	70	6F8
24	7	3.4	2.93	-.85	.71	0.5	-1	0.5	-1	.27	35	4D8
9	7	1.3	1.74	-4.12	.82	0.5	0	0.5	-1	.20	73	6F11
28	13	2.2	2.55	-2.02	.45	0.5	-1	0.5	-1	.30	26	2B12
15	7	2.1	2.52	-2.12	.61	0.5	-1	0.5	-1	-.12	36	4D9
15	7	2.1	1.50	-4.82	.61	0.5	-1	0.5	-1	-.07	13	1A13
24	14	1.7	1.91	-3.68	.47	0.5	-1	0.5	-1	.41	7	1A7
4	3	1.3	1.54	-4.70	1.18	0.6	0	0.5	0	.16	40	4D13
23	7	3.3	2.78	-1.34	.70	0.6	-1	0.6	-1	.14	22	2B8
42	14	3.0	2.46	-2.26	.49	0.6	-1	0.6	-1	-.18	25	2B11
25	7	3.6	3.41	.82	.73	0.6	-1	0.6	-1	.43	44	5E1
38	14	2.7	3.01	-.55	.46	0.6	-1	0.6	-1	.21	58	5E15
10	7	1.4	1.73	-4.15	.71	0.6	0	0.7	0	.32	12	1A12
35	14	2.5	2.05	-3.31	.45	0.6	-1	0.6	-1	-.03	37	4D10
12	7	1.7	2.07	-3.27	.63	0.6	0	0.6	0	-.05	19	2B5
25	7	3.6	3.81	2.32	.73	0.7	0	0.6	0	.39	60	5E17
24	7	3.4	3.33	.55	.71	0.7	0	0.7	0	.30	27	2B13
11	7	1.6	1.52	-4.76	.70	0.8	0	0.7	0	.25	8	1A8
21	7	3.0	2.91	-.93	.70	0.7	0	0.7	0	.44	68	6F6
27	13	2.1	2.22	-2.89	.46	0.7	0	0.7	0	.43	45	5E2
32	7	4.6	4.43	5.58	.78	0.7	-1	0.7	-1	.43	48	5E5
17	7	2.4	2.22	-2.89	.62	0.7	0	0.7	0	.04	16	2B2
12	6	2.0	2.35	-2.55	.66	0.7	0	0.7	0	.17	33	4D6
20	7	2.9	2.96	-.74	.68	0.7	0	0.8	0	-.35	63	6F1
16	7	2.3	2.83	-1.16	.61	0.7	0	0.7	0	.19	17	2B3
27	14	1.9	2.16	-3.05	.45	0.7	0	0.7	0	.45	82	6F20
36	14	2.6	2.73	-1.48	.45	0.7	0	0.8	0	.13	65	6F3
20	9	2.2	2.64	-1.78	.54	0.8	0	0.8	0	.11	30	4D3
14	7	2.0	2.13	-3.11	.61	0.8	0	0.8	0	.04	78	6F16
26	14	1.9	1.66	-4.32	.47	0.9	0	0.8	0	.41	3	1A3
9	7	1.3	1.21	-5.99	.82	0.8	0	0.8	0	.24	10	1A10
28	14	2.0	1.83	-3.89	.46	0.8	0	0.8	0	.53	76	6F14
15	10	1.5	1.86	-3.79	.58	0.9	0	0.8	0	.35	5	1A5
40	14	2.9	2.99	-.62	.47	0.9	0	0.9	0	.10	21	2B7
23	7	3.3	2.82	-1.21	.70	0.9	0	0.9	0	-.19	51	5E8
25	14	1.8	2.20	-2.94	.44	1.0	0	1.0	0	-.12	15	2B1
21	6	3.5	3.74	2.03	.77	1.0	0	1.0	0	.00	72	6F10
21	7	3.0	3.27	.34	.70	1.0	0	1.0	0	.03	64	6F2
34	14	2.4	2.32	-2.64	.46	1.0	0	1.0	0	.30	71	6F9
33	13	2.5	2.52	-2.11	.47	1.0	0	1.1	0	.25	39	4D12
12	7	1.7	2.16	-3.04	.63	1.1	0	1.1	0	.16	2	1A2
48	14	3.4	3.61	1.51	.50	1.1	0	1.1	0	-.24	57	5E14
47	14	3.4	3.25	.28	.52	1.1	0	1.0	0	.37	38	4D11
20	7	2.9	2.66	-1.72	.68	1.1	0	1.1	0	.13	54	5E11
17	7	2.4	1.76	-4.06	.62	1.2	0	1.2	0	-.27	11	1A11
28	14	2.0	2.23	-2.85	.44	1.2	0	1.2	0	-.01	31	4D4
21	7	3.0	3.10	-.26	.70	1.3	0	1.3	0	-.19	62	5E19
23	9	2.6	2.60	-1.89	.56	1.4	0	1.4	0	.31	34	4D7
10	6	1.7	1.45	-4.97	.69	1.4	0	1.4	0	.13	32	4D5
52	14	3.7	3.41	.80	.56	1.4	1	1.3	0	.06	52	5E9
25	7	3.6	3.92	2.92	.73	1.5	0	1.5	1	-.25	66	6F4

Figure 3A. FACETS: Student adjusted scores.

```
PaceSpanishPort  07-17-1997 12:52:02
Table 7.1.1  Students Measurement Report  (arranged by FN).

-------------------------------------------------------------------------------------------
| Obsvd   Obsvd  Obsvd   Fair |        Model | Infit      Outfit   |       |
| Score   Count Average Avrge |Measure  S.E. |MnSq ZStd  MnSq ZStd | PtBis | Nu Students
-------------------------------------------------------------------------------------------

|    23*     14     1.6   1.39|  -5.16   .48 | 1.7   1    1.7   1  |   .16 |  9 1A9
|    12*      7     1.7   2.07|  -3.27   .63 | 1.6   1    1.7   1  |  -.21 | 74 6F12
|    22*     11     2.0   2.11|  -3.17   .51 | 1.6   1    1.7   1  |   .26 | 77 6F15
|    20*      9     2.2   2.20|  -2.93   .54 | 1.7   1    1.7   1  |   .16 |  6 1A6
|    14*      6     2.3   2.13|  -3.12   .66 | 1.8   1    1.8   1  |  -.31 | 75 6F13
|    18*      7     2.6   2.34|  -2.58   .64 | 1.8   1    1.9   1  |   .18 | 29 4D2
|    19*      7     2.7   3.01|   -.58   .66 | 2.1   1    2.2   1  |  -.59 | 46 5E3
|    17*      7     2.4   2.31|  -2.66   .62 | 2.1   1    2.3   1  |  -.53 |  4 1A4
|    19*      7     2.7   3.01|   -.57   .66 | 2.4   1    2.4   1  |   .34 | 59 5E16
|    15*      7     2.1   2.51|  -2.14   .61 | 2.8   2    2.8   2  |  -.23 | 67 6F5
|    50*     14     3.6   3.54|   1.25   .52 | 2.7   3    2.9   3  |   .24 | 55 5E12
|    20*      7     2.9   2.66|  -1.72   .68 | 3.2   2    3.2   2  |   .27 | 53 5E10
|    30*     13     2.3   2.52|  -2.10   .47 | 3.8   4    4.2   4  |   .43 | 18 2B4
|     6*      6               |( -8.24  1.87)|Minimum              |   .00 | 42 4D15
-------------------------------------------------------------------------------------------
| Obsvd   Obsvd  Obsvd   Fair |        Model | Infit      Outfit   |       |
| Score   Count Average Avrge |Measure  S.E. |MnSq ZStd  MnSq ZStd | PtBis | Nu Students
-------------------------------------------------------------------------------------------
|    23.2    9.1    2.6   2.56|  -1.72   .62 | 0.9  -0.4  0.9  -0.4|   .14 | Mean (Count: 82)
|    10.7    3.2    0.7   0.73|   2.31   .13 | 0.7   1.5  0.8   1.5|   .25 | S.D.
-------------------------------------------------------------------------------------------
RMSE (Model)   .64  Adj S.D.  2.22  Separation  3.48  Reliability  .92
Fixed (all same) chi-square: 947.8  d.f.: 80  significance: .00
Random (normal) chi-square: 78.2  d.f.: 79  significance: .50
-------------------------------------------------------------------------------------------
```

Figure 3A *(continued).*

Raters

In 1996, in Phase I of the portfolio study, each of the seven raters rated five portfolios, but not all raters rated the same set of five portfolios. For each rater, FACETS calculated the average of the ratings that rater gave the five portfolios she or he rated. The average of the ratings each rater gave ranged from 3.0 for Rater 5 to 3.6 for Raters 3 and 4.

In 1997, in Phase II of the portfolio study, each of the 10 raters rated from two to nine portfolios. The average of the ratings each rater gave ranged from 1.8 for Rater 4 to 3.1 for Rater 11. On average, raters tended to give lower ratings to the portfolios scored in 1997 than to those scored in 1996.

Aspects and Dimensions

When raters used the scoring rubrics for the various aspects and dimensions, there were differences in 1996 and 1997 in the percentages of ratings falling in each scale category. Those differences are shown in Table 11.

TABLE 11

Percentage in Each Scale Category

SCALE CATEGORY	DIMENSION 1 ASPECT 1		DIMENSION 1 ASPECT 2		DIMENSION 1 ASPECT 3		DIMENSION 1		DIMENSION 2 ASPECT 1		DIMENSION 2 ASPECT 2		DIMENSION 2	
	1996	1997	1996	1997	1996	1997	1996	1997	1996	1997	1996	1997	1996	1997
1	5	22	3	19		24		19		15		5		7
2	14	35	9	37		27	6	35	5	28	12	28	5	36
3	57	30	59	31	64	44	36	33	40	43	67	47	46	41
4	22	13	29	13	27	5	56	12	52	15	18	17	46	12
5	3				9		3		2		3	3	3	3
Mean	3.0	2.3	3.1	2.4	3.5	2.3	3.1	2.4	3.6	2.6	3.5	2.9	3.5	2.7

Dimension 1, "Demonstrating Knowledge of Hispanic Cultures"

Aspect 1, *Showing awareness of the diversity of Hispanic cultures*

- In 1996, raters used all five scale categories; in 1997, raters used only four scale categories (i.e., 1 to 4).
- In 1996, the majority of ratings were 3 or 4; in 1997, the majority of ratings were 2 or 3.

Aspect 2, *Identifying contributions of Hispanic figures*

- In 1996 and 1997, raters used four of the five scale categories (i.e., 1 to 4).
- In 1996, the majority of ratings were 3 or 4; in 1997, the majority of ratings were 2 or 3.

Aspect 3, *Making connections with other disciplines and own culture(s)*

- In 1996, raters used only three scale categories (i.e., 3 to 5); in 1997, raters used four scale categories (i.e., 1 to 4).
- In 1996, the majority of ratings were 3 or 4; in 1997, the majority of ratings were 2 or 3.

Dimension 1

- In 1996, raters used four scale categories (i.e., 2 to 5); in 1997, raters used four scale categories (i.e., 1 to 4).
- In 1996, the majority of ratings were 3 or 4; in 1997, the majority of ratings were 2 or 3.

Dimension 2, "Using Spanish to Communicate Effectively"

Aspect 1, *Deriving meaning from texts and personal interactions*

- In 1996, raters used four scale categories (i.e., 2 to 5); in 1997, raters used four scale categories (i.e., 1 to 4).
- In 1996, the majority of ratings were 3 or 4; in 1997, the majority of ratings were 2 or 3.

Aspect 2, *Expressing meaning in oral and written form*

- In 1996, raters used four scale categories (i.e., 2 to 5); in 1997, raters used all five scale categories.
- In 1996, the majority of ratings were 3; in 1997, the majority of ratings were 2 or 3.

Dimension 2

- In 1996, raters used four scale categories (i.e., 2 to 5); in 1997, raters used all five scale categories.
- In 1996, the majority of ratings were 3 or 4; in 1997, the majority of ratings were 2 or 3.

Summary of Findings

- This research showed that the aspects of the delineated Pacesetter Spanish system of portfolio assessment worked better in Phase II than in Phase I. Students and teachers had a better understanding of what constitutes evidence of achievement as well as of the selection process and organization of artifacts included in the portfolios.
- Pacesetter Spanish portfolio raters used the categories of "Promising" and "Developing" for the majority of ratings, but less than in 1996, and not the extreme categories ("Accomplished" or "Advanced").
- The length of time to rate the portfolios in 1996 and 1997 was an average of 45 to 50 minutes. Although variety and complexity are encouraged throughout, this element of time needs to be taken into account when preparing to read portfolios.
- Although higher than in 1996, the number of rating agreements was moderate. However, there were a considerable number of discrepancies between raters regarding the presence or lack of enough artifacts to judge students' performance within the parameters of the Pacesetter Spanish portfolio assessment system.
- The proportion of ratings of "No Evidence" for Dimension 1, Aspect 3 was dramatically reduced.
- Raters differ in level of severity exercised when rating portfolios. Some grade significantly more harshly than others. About one-fourth of the students would have received different total scores if their scores had been adjusted for differences in rater severity.
- The rating scales for the individual aspects function somewhat differently from one another. Raters used all five-scale categories for some of the rating scales but only used four-scale ratings for other scales.
- The rating patterns for about 20 percent of the portfolios contained one or more surprising or unexpected ratings (i.e., ratings that don't seem to "fit" with the other ratings given to that portfolio). When such variability is present in a set of ratings, it is suggested that those in charge of monitoring quality control in the scoring session give those portfolios a second look to determine which ratings are "out of sync" with the other ratings, and whether those unexpected ratings shall stand "as is" or be changed.

Recommendations

- A final version of the matrix should be incorporated as part of the formal components of the Pacesetter Spanish program.

- A set of guidelines should be included in the Teacher's and Student's Edition of the Pacesetter Spanish curriculum materials. The guidelines should include clear instructions of the selection process of work samples for the portfolio, cover sheets, rubrics, expectations, and any other pertinent information.
- Activities in the Pacesetter Spanish curriculum materials should be designed to reflect all aspects and dimensions of the assessment portfolios so students have ample opportunity to demonstrate their achievement of the course outcomes.
- Dissemination of the portfolio component of the Pacesetter Spanish program should be carried out at the different professional development sessions. Training sessions should include an overview of the steps to follow when implementing the system of portfolio assessment as well as an explanation on how to apply the matrix when evaluating portfolios.
- New benchmark portfolios from the 1997 Pacesetter Spanish reading should be identified, and a set of sample portfolios should be prepared for use in training sessions.
- Efforts should be made to organize scoring sessions at the national or regional level for Pacesetter Spanish portfolios so teachers can have an opportunity to apply the assessment matrix and have discussions on what constitutes evidence of achievement for each aspect and dimension.
- Further research is needed to ensure that the evaluation of portfolios in the Pacesetter Spanish program is done effectively and that all quality control measures are taken.
- Additional research is necessary to evaluate the different elements that comprise the Pacesetter Spanish portfolio assessment system. Results could guide further refinement of this component and clarify its connection to the goals of the overall program.
- Evaluative studies should be conducted to measure the effectiveness and impact of the Pacesetter Spanish program in general. It is particularly important to find out how each component contributes to the success of the program and its vision of integrating standards, professional development, instruction, and assessment.

References

Fercsey, A., Luna, C. and Ponte, E. (1997). *Pacesetter Spanish Portfolio Research Project: Phase I Report.* [Unpublished monograph]. Princeton, NJ: Educational Testing Service.

Linacre, J.M. and Wright, B.D. (1994). *A User's Guide to FACETS: Rasch Measurement Computer Program.* [Computer program manual]. Chicago: MESA Press.

Sheingold, K., Storms, B., Thomas, W. and Heller, J. (1997). *Pacesetter English Portfolio Assessment: 1995 Report.* Center for Performance Assessment. Princeton, NJ: Educational Testing Service.

Phase I Appendix

Appendix 1: DRAFT COVER SHEET

PACESETTER SPANISH PORTFOLIO ASSESSMENT

NAME: _____

Selection: _____ **Date work was done:** _____

Why did you select this piece of work?

What does it show or tell about you?

What did you learn from doing this assignment?

Please check the dimension(s) or aspects for which you are showing evidence with this assignment:

- **<u>Dimension 1: Demonstrating Knowledge of Hispanic Cultures</u>**

_____ **Showing Awareness of the Diversity of Hispanic Cultures**

_____ **Identifying Contributions of Hispanic Figures**

- **<u>Dimension 2: Using Spanish to Communicate Effectively</u>**

_____ **Deriving Meaning from Texts and Personal Interactions**

_____ **Expressing Meaning in Oral and Written Form**

Appendix 1: DRAFT COVER SHEET (Continuation)

PACESETTTER SPANISH PORTFOLIO ASSESSMENT

NOMBRE: _____

Título: _____ Fecha de ejecución: _____

¿Por qué seleccionaste esta obra?

¿Qué muestra o dice esta obra sobre ti?

¿Qué aprendiste de esta tarea?

He seleccionado esta obra para indicar evidencia en las siguientes dimensiones:

- **<u>Dimensión 1: Demostrar conocimiento de las culturas hispanas</u>**

____ Demostrar conciencia de la diversidad de las culturas hispanas

____ Identificar contribuciones de figuras hispanas

- **<u>Dimensión 2: Usar el español para comunicaciones eficientes</u>**

____ Comprensión de textos e interacciones personales

____ Expresión oral y escrita

Appendix 2: Assessment Matrix

PACESETTER SPANISH
DIMENSION: DEMONSTRATING KNOWLEDGE OF HISPANIC CULTURES
Portfolio Assessment: Evaluating Aspects of the Dimension

	Beginning ☐	Developing ☐	Promising ☐	Accomplished ☐	Advanced ☐	Not enough evidence to judge ☐
Showing Awareness of the Diversity of Hispanic Cultures	Vague indication fragmentary, superficial	Limited -incomplete -a few specific examples -uneven -partial undeveloped	Some examples -general insights -middle range of insights/examples	Ample/ full, wide range of examples, deeper/more profound	Consistently insightful, varied, extended	

	Beginning ☐	Developing ☐	Promising ☐	Accomplished ☐	Advanced ☐	Not enough evidence to judge ☐
Identifying Contributions of Hispanic Figures	List a few names and contributions brief, sketchy description	Some description of contributions -limited notions of contribution	Some inferences about place and impact in history	Recognition of impact in history -ample inferences about place in history	Thorough inferences -consistently insightful analysis of contributions and impact/ role in history	

	Beginning ☐	Developing ☐	Promising ☐	Accomplished ☐	Advanced ☐	Not enough evidence to judge ☐
Making connections with other disciplines and own culture(s)	Isolated -no logical connections -few, irrelevant, clueless, superficial connections	Limited -partial -not always relevant -uneven connections	Some general notions without much depth/insight	Ample examples of insightful and wide range of connections -some inferences	Thorough and insightful connections -new -beyond what they've seen in course -creativity -originality -original connections	

Appendix 2: Assessment Matrix (Continuation)

PACESETTER SPANISH
DIMENSION: USING SPANISH TO COMMUNICATE EFFECTIVELY
Portfolio Assessment: Evaluating Aspects of the Dimension

	Beginning ☐	Developing ☐	Promising ☐	Accomplished ☐	Advanced ☐	Not enough evidence to judge ☐
Deriving meaning from texts and personal interactions (Skills: Listening and Reading)	Understands few general ideas -continuous misunderstandings	Uneven comprehension, frequent/numerous misunderstandings -understands some main ideas on familiar topics	Understands main ideas and some details on somewhat familiar topics -may make some inferences depending on complexity of text	Understands main ideas and most details even on unfamiliar topics -can make inferences appropriate to the task	Thorough understanding, extrapolates information in new ways, including some social and cultural references and affective overtones	

	Beginning ☐	Developing ☐	Promising ☐	Accomplished ☐	Advanced ☐	Not enough evidence to judge ☐
Expressing meaning in oral and written form (Skills: Speaking and Writing)	Communication impeded by interference from another language -lack of fluency -inadequate vocabulary and/or pronunciation -fragmented -repetitious message forces interpretation, -pronunciation hampers communication	Limited fluency -conveys simple messages -unconnected discourse or sentences -limited range of vocabulary -some interference from another language -pronunciation may interfere with communication	Some underlying organization of thoughts/ ideas/discourse -emerging notion of adequate register -adequate range of vocabulary -self-corrections -pronunciation may be awkward but doesn't interfere with communication -level of fluency doesn't interfere with communication	Message is articulated (organized appropriately) -proper use of circumlocution -at times creative ideas, -coherent use of vocabulary -minimal interference from another language -mostly adequate use of register	Highly creative, -thorough, meaningful, -evidently organized, -well articulated and organized message, -appropriate use of register -wide range of vocabulary -little or no interference from another laguage, -high level of fluency	

Appendix 3

Student Number: Rater Number:

Artifacts Used	Dimension 1: Demonstrating Knowledge of Hispanic Cultures			Dimension 2: Using Spanish to Communicate		Notes/ Comments
	Diversity of Hispanic Cultures	Contributions of Hispanic Figures	Connections with other disciplines	Deriving meaning from texts (L&R)	Express. meaning from texts (S&W)	
Score						

Scale: Beginning, Developing, Promising, Accomplished, Advanced, Not enough evidence to judge.

Other comments/suggestions:

Appendix 4

Port. #	Rater	Date	Time Used	Dimension 1 Aspect 1	Dimension 1 Aspect 2	Dimension 1 Aspect 3	Dimension 2 Aspect 1	Dimension 2 Aspect 2	D 1	D 2
1[1]	6	St/Ev	20m	Promising	Promising	No evidence	No evidence	Promising	Promising	Promising
1	5	Su/M	35m	Accomplished	Promising	NA	Promising	Promising	Promising	Promising
2	5	St/Ev	20m	Developing	Developing	NA	Promising	Developing	Developing	Developing
2	4	St/Ev	30m	Promising	No evidence	No evidence	No evidence	Promising	No evidence	Promising
3	4	St/Ev	20m	No evidence	Promising	No evidence	Developing	Promising	No evidence	Promising
3	3	Su/M	55m	Developing	Promising	Accomplished	Accomplished	Accomplished	Promising	Accomplished
4	7	St/Ev	1h 5m	Promising	Beginning	Promising	Accomplished	Accomplished	Promising	Accomplished
4	2	Su/M	40m	Promising	Promising	No evidence	Promising	Accomplished	Promising	Accomplished
5	1	St/Ev	1h	Promising	Promising	NA	Accomplished	Accomplished	Promising	Accomplished
5	6	St/Ev	35m	Promising	No evidence	No evidence	Promising	Promising	No evidence	Promising
6	1	St/Ev	30m	Promising	Accomplished	NA	Accomplished	Accomplished	Promising	Accomplished
6	6	St/Ev	30m	Promising	Promising	No evidence	Promising	Promising	Promising	Promising
7	2	St/Ev		Accomplished	Accomplished	No evidence	Accomplished	Accomplished	Accomplished	Accomplished
7	7	Su/M	50m	Developing	Accomplished	No evidence	Accomplished	Promising	Promising	Promising
8	4	St/Ev		No evidence	No evidence	No evidence	Promising	Accomplished	No evidence	Accomplished
8	3	Su/M	35m	Promising	Promising	Accomplished	Accomplished	Accomplished	Promising	Accomplished

[1] Portfolio numbers in bold indicate the portfolio was selected as a benchmark.

Appendix 4 -Continuation

Port. #	Rater	Date	Time Used	Dimension 1 Aspect 1	Dimension 1 Aspect 2	Dimension 1 Aspect 3	Dimension 2 Aspect 1	Dimension 2 Aspect 2	D 1	D 2
9	6	St/Ev	20m	Promising	Accomplished	Promising	No evidence	Accomplished	Promising	No evidence
9	5	St/Ev	25m	Accomplished	Accomplished	NA	Accomplished	Accomplished	Accomplished	Accomplished
10	3	St/Ev	50m	Beginning	Accomplished	NA	Accomplished	Accomplished	Promising	Accomplished
10	1	Su/M	20m	Developing	Promising	NA	Promising	Promising	Developing	Promising
11	3	St/Ev		Accomplished	Accomplished	NA	Accomplished	Accomplished	Accomplished	Accomplished
11	1	Su/M	40m	Accomplished	Promising	Promising	Accomplished	Accomplished	Promising	Accomplished
12	6	St/Ev	40m	Promising	No evidence	No evidence	Accomplished	Promising	No evidence	Promising
12	5	Su/M	45m	Accomplished	Promising	NA	Promising	Promising	Accomplished	Promising
13	5	St/Ev	45m	Promising	Developing	No evidence	Developing	Promising	Developing	Promising
13	4	St/Ev	35m	No evidence	No evidence	No evidence	Promising	Promising	No evidence	Promising
14	6	--	--	No evidence	No evidence	No evidence	No evidence	Promising	No evidence	No evidence
14	4	St/Ev	40m	Promising	No evidence	No evidence	Promising	Promising	No evidence	Promising
15	1	St/Ev	30m	Promising	Promising	Promising	Promising	Accomplished	Promising	Promising
15	6	Su/M	40m	Promising	No evidence	Promising	Accomplished	Promising	Promising	Promising
16	2	St/Ev	1h	Accomplished	Promising	Promising	Accomplished	Accomplished	Promising	Accomplished
16	7	NA	1h	Accomplished	Promising	Promising	Accomplished	Accomplished	Accomplished	Accomplished
17	7	St/Ev	1h 10m	Promising	Promising	No evidence	No evidence	Accomplished	Promising	Accomplished
17	4	Su/M	55m	Advanced	Accomplished	Advanced	Advanced	Advanced	Advanced	Advanced

Appendix 4-Continuation

Port. #	Rater	Date	Time Used	Dimension 1 Aspect 1	Dimension 1 Aspect 2	Dimension 1 Aspect 3	Dimension 2 Aspect 1	Dimension 2 Aspect 2	D 1	D 2
18	5	St/Ev	40 m	Beginning	Promising	NA	Accomplished	Promising	Promising	Promising
18	4	Su/M	30m	No evidence	Promising	No evidence	No evidence	Accomplished	No evidence	No evidence
19	2	St/Ev	45m	Developing	Developing	No evidence	Promising	Developing	Developing	Developing
19	7	NA	1h	Promising	Promising	No evidence	Accomplished	Promising	Promising	Promising
20	3	St/Ev	1h	Promising	Accomplished	Accomplished	Accomplished	Accomplished	Accomplished	Accomplished
20	1	Su/M	35m	Promising	Promising	NA	Promising	Accomplished	Promising	Promising
21	7	Su/M	50m	Promising	Accomplished	No evidence	Accomplished	Accomplished	Promising	Accomplished
21	2	Su/M		Promising	Promising	NA	Accomplished	Accomplished	Promising	Accomplished

Appendix 5

PORTFOLIOS RATED BY TWO DIFFERENT PERSONS*

Student #	Agreement	Artifacts	Comments
31938	OK-Perfect	Biografía-tape (D1&2)** Speech (D2) Work Sheet (D1&2) Guía Mexico (D1&2) Work Sheet (D1&2)	Dimension 2: Promising; however, not enough evidence to judge receptive skills.
		Project (D2) Video (D2) Tape (D1&2) Itinerario (D1&2) Reading (D1&2--1NE) [?] on food (D1&2)	Dimension 1: Promising (had there been more evidence it might have been Accomplished). Dimension 2: Promising (not enough evidence!)

* This is a transcript of the original rating sheets. In trying to be faithful to the originals, we have writen between brackets (i.e.,[]) those words that were difficult to read from the handwriting.

** D1: Instrument was marked under one or more aspect of dimension 1.
D2: Instrument was marked under one or more aspect of dimension 2.
D1&2: Instrument was marked under one or more aspect of each dimension.
D1--NE or D1--NE: Instrument was marked under the respective dimension but there was not enough evidence to judge its contribution to the respective dimension.
D1&D2--1NE or D1&D2--2NE: Instrument was marked under one or more aspect of each dimension, but not enough evidence was found for one of the dimensions.

Appendix 5 - Continuation

Student #	Agreement	Artifacts	Comments
30129	Dimension 1: OK/NE Dimension 2: OK	Chart (D2) Worksheet (D2) Essay (D2) Paragraph (D1&2) Pamphlet (D1&2) Poster (D2) Essay (D1&D2) Essay (D2) Foto (D1--NE) Tape (D1&2--2NE, read text) Tape (D1&2--2NE) Invitación (D2) Fotos Labeled (D2--NE)	Dimension 1: Diversity: promising/ No evidence Dimension 2: Promising
		Reading: Popo (D2) Presentation: Invitations (D2) Mayas Project (D1&2) Page: Diario (D2) Collage (D2) Itinerario: México (D1) Collage: Cervantes(D1) Invitaciones/explic. (D2) Debate Ecológico (D1&2) Proyecto: Extinción (D2) Collage: La moda (D2--NE)	Dimension 1: Developing Dimension 2: Developing Little evidence shown in some areas yet there is A LOT in the portfolio! but it doesn't serve the purpose.

Appendix 5 - Continuation

Student #	Agreement	Artifacts	Comments
30831	Dimension 1: OK/NE Dimension 2:	Work Sheet (D2) Work Sheet (D1&2--1NE) Essay Chart (D1&2) Tape (D1&2) Paragraph (D2) Poster (D2) Essay (D1&2) Paragraph (D2) Pamphlet (D1&2) Invitación (D2) Essay (D2) Essay (D2)	Dimension 1: Not enough evidence/ Promising hispanic figures Dimension 2: Promising Note: Cover sheets not included.
		Los Chivelos[?] (D1) Fusión (D1) --evidence shown only on cover sheet comments [?] (D2) Isla Pascua (D2) Franco (D1) Martí/Juárez (D1) Biografía (D2) Pete Sampras (?) Carta [?] (D2) Pascua [?] (D2) Blasón (D2) Viaje [?] (D2) Pinturas (D1&2)	Dimension 1: Promising Dimension 2: Accomplished

Appendix 5 - Continuation

Student #	Agreement	Artifacts	Comments
31929	OK-Perfect	Quiz (D1&2) Plan de Vuelo (D1&2) Folleto Mexico (D1&D2) Banana Split (D2) Video (D2) Poema "La Muralla" (D1&D2) Tape-hero (D1&D2--1NE) Tape & Collage (D2) Tape (?) (D2)	Dimension 1: Promising Dimension 2: Accomplished
		Esta es su vida (D2) Mexico (D1&D2) Banana Split (D1&D2--2?) Hero of Year (D1&D2--1?) Video --Not included Bolívar ? (D1&D2) Hispanics in USA (D1&D2) La Muralla (NE)	Dimension 1: Promising. However, student could have included more products. Dim 2.: Accomplished. Student includes more evidence of dimension 2 than dimension 1. Note: She has more artifacts to prove evidence of writing skills. She didn't include any evidence of listening comprehension.
31296	Dimension 1: NO/NE Dimension 2: OK	Cartel: Logo de Paca (D2) Carta (D2) Tape: España (D1&2) Tape: Madre Teresa (D1&2) Letter: Isla (D1&2--1NE) Work Sheets: Colón, Cortés, Isabel, D. Marina (D1) Writing: Autobiografía (D2) Writing: Food (D1&2) Summary: (D1&2--1NE)	Dimension 1: Promising Dimension 2: Accomplished Note: Choice of dimension didn't always show evidence for making appropriate judgement. Student's writing and speaking demonstrated well articulated messages.
		Poster (D2) Carta (D2) Carta/lectura (D1&2) Worksheet (D1&2) Tape (D2) Biografía (D2) Receta (D2) Research (D1&2)	Dimension 1: Not enough evidence Dimension 2: Promising

Appendix 5 - Continuation

Student #	Agreement	Artifacts	Comments
30351	Dimension 1: OK-Perfect Dimension 2: OK	Resumen: Cajas (D2) Tape: Martha Anderson (D1&2) Writing: Goya/Picasso (D1) Writing: Popo/Ixtla (D2) Writing: Guerra Civil (D1&2) Writing: Benit./Chávez (D1&2) Poster: A la deriva (D2) Poster: La selva (D2) Poster: Immigration (D2) Tape: Malinche/Isabel (D1&2)	Dimension 1: Promising Dimension 2: Accomplished
		Poster with presentation (D2) A la deriva (D2--NE) Tape (D1) Tape (D1&D2) Resumen (D2) Work Sheet (D2) Essay (D1) Essay (D2) Essay (D1&D2) Essay (D1&D2)	Dimension 1: Promising Dimension 2: Promising
30358	Dimension 1: OK Dimension 2: OK	Comp.: Juárez/Martí (D1&2) Isabel/Maliche (D1&2) Benito Juárez (D1&2) El Greco (D1&2) Jerry García (D1&2) Collage: Popo/Ixtla (D1&2) Autobiografía (D2)	Dimension 1: Accomplished Dimension 2: Accomplished
		Poster: Quien yo ... (D2) A la Deriva (D2) Ixtla/Popo (D1&2) Jerry García (D2) Isabel/La Maliche (D1&D2) Benito Juárez (D1&2) Culminating Ass. (D2) España/Greco (D1) --student didn't indicated that she wanted to use C.A. as evidence. Los mayas (D1)	Dimension 1: Promising Dimension 2: Promising Note: Size of some of the products was very large and difficult to handle. "Interacciones personales" - Student understood it as cooperation instead of speaking.

Appendix 5 - Continuation

Student #	Agreement	Artifacts	Comments
30144	Dim 1: NO-NE Dimension 2 OK-Perfect	Essay (D1&D2) Essay/Post Card (D2) Work Sheet-paragraph (D2) Poster (D2) Picture (D2) Drawing (D1&2) Escudo (D2) Menu (D2) Párrafo (D2) Poster (D1&2--NE) Poster (D2) Paragraph (D1&D2)	Dimension 1: Not enough evidence Dimension 2: Promising.
		El [? -"ilegible"] (D1&D2) Bis (D2) Moais (D1&D2) Chicos de hoy (D2) [?-"ilegible"] (D2) Biografía (D2) Comida (D1&D2) [? - "ilegible"] (D2) Menu (D2) Extinción (D2) Cumbre (D1&D2--1NE) Moda (D2--NE) Change[?] (D1-D2)	Dimension 1: Promising Dimension 2: Accomplished
30460	Dimension 1: OK Dimension 2: OK-Perfect	Essay: Guerra Civil España (D1) Pascua postcard (D1) C.A. (D2) Debate (D1&2) Essay (D2) Worksheet (D2--NE) Escudo (D1) Chávez	Dimension 1: Promising Dimension 2: Accomplished in speaking. Not enough evidence to fully judge other aspects of this dimension. Note: Have used to score aspects not marked by student.
		Comp.: Guerra Civil, Goya (D1) Tarjeta (D1) Tape: Orales (D2) Comp.: Persona admiro (D2) Canción: Chicas de hoy (D2) Proyectos	Dimension 1: Accomplished Dimension 2: Accomplished Note: projects didn't have cover sheets, but there is plenty of evidence in all fields.

Appendix 5 - Continuation

Student #	Agreement	Artifacts	Comments
30816	Dimension 1: OK Dimension 2: OK	Comp.: Martí, Juárez, Chávez (D2) Summary: A la deriva (D2) Worksheet: Bolívar/Wash. (D2) Summary (D2) Outline: Goya/Picasso (D2) Worksheet: Hispanos (D1) Summary: (D1) Cassette: C.A. (D2)	Dimension 1: Developing Dimension 2: Promising Note: Student's portfolio needs more evidence in dimension 1. Lack of products result in an inability to arrive at an assessment other than dev. Dim 2 provides more evidence and results in assessm. of promising.
		Ancianos (D2) Comp.: Martí/Chávez (D1&2) Comp.: A la deriva (D2) Bolívar (D2) [?] (D2) Goya y Picasso (D1) [?] Hispanic (D1&2) Martí/Juárez (D1) Miami (didn't indicate dimension) Assess. (didn't indicatedimension) Mi hermana (didn't indicate dimension)	Dimension 1: Promising Dimension 2: Accomplished
30435	Dimension1: OK Dim 2: OK- Perfect	Pamphlet: autobiography (D1&2) Map: Mexico (D1&d2) Outline: Popo/Ixtla (D1&2) Cassette: A la Deriva(D1&2--1NE) Summary: El maíz (D1&2--1NE) Letter: Isla de Pascua (D1&2) Summary: El Greco (D1&2) Poem: Blasón (D1&D2) Pamphlet: España (D1&D2)	Dimension 1: Promising Dimension 2: Accomplished Note: Very good portfolio especially writing, speaking and comprehension and reading. Student does not show evidence of contribution of Hispanic figures but checks the space nevertheless. Need more information.
		Julia (D1&2--NE) Deriva (D1&2) Mapa Mexico (D1&D2--2NE) Popo (D1&2) Maíz (D1&2--1NE) Carta (D1&2) Vamisa? (D1&2) Pintura Greco (D1&2--2NE) Blasón (D1&2) España (D1&2)	Dimension1: Accomplished Dimension 2: Accomplished

Appendix 5 - Continuation

Student #	Agreement	Artifacts	Comments
30124	Dimension1 : NO-NE Dimension 2: OK-Perfect	Biografía (D2--NE) Cena (D2--NE) Cervantes (D1&2) Itinerario Mexico (D1&2) Debate Ecológico (D1&2) Animales peligro ext. (D2--NE) La moda (D2) Deriva (D2) Isla de Pascua (D2) Chicas de hoy (D2) Grupos étnicos (D1) Diario (D2)	Dimension 1: Accomplished (Should have explained more, but proved more than promising) Dimension 2: Promising (close to accomplished but no quite there: adequate register and fluency)
		Poster (D1&2--not enough) Poster (D1&2--not enough) Poster (D2) Guide (D1&2) Poster (D1&2--not enough) Cumbre (D1&2-- s not enough) La Moda (D2--not enough) A la deriva (D2) Postal (D2) Song (D2) Research (D1&2) Diario (D2)	Dimension 1: Not enough to evaluate whole dimension (on the first aspect: promising) Dimension 2: Promising

Appendix 5 - Continuation

Student #	Agreement	Artifacts	Comments
30476	Dimension 1: NO-NE	Essay (D2) Essay (D2) List (D1&2) Invitaciones (D2) Lista Comida (D1) Essay (D2) Poster (D2) Collage (D2) Work-sheet (D2) Drawing (D1&2--NE) Explanation-sketch (D1&2--1NE) Essay (D1&2) Tape (D2) Tape (D2&2) Essay (D1&2)	Dimension 1: Not enough evidence Dimension 2: Promising
		Comp.: Artista Loca (D2) Tape: Invitación banquete (D2) Tape: Debate ecológico (D2) Comp.: Goya/Picasso (D1&2--1?) Strip story: Popo (D1&2--1?) Dibujo: A la deriva (D1&2--1?) Collage: Autobiografía (D2) Name Project: (D2) Comp.: Biografía (D2) Lista comida: (D1) Invitación (D2) Eslabón (D1&2)	Dimension 1: Developing Dimension 2: Promising Note: Too much of same kinds of materials. Didn't know how to choose best works. Not enough evidence in some cases.

Appendix 5 - Continuation

Student #	Agreement	Artifacts	Comments
30457	Dimension 1: OK/NO-NE Dimension 2: NO-NE	Tape (D2) Poster/Essay (D1&2) Carta (D2) [blanco en doc.] (D1&2) Work Sheet (D2) Poster (D2) Escudo (D2--NE) Essay (D1&2) Invitaciones (D2) Video (D1) Poster (D1&2--2NE) Sketch (D2--NE)	Dimension 1: Diversity=promising/Not enough evidence. Dimension 2: Promising, but my sense is that with more evidence it would be accomplished.
		Tape (C.A.) (D2) Picture seq. (D1&2--1NE) Carta (D2) I. Pascua (D2) C. de hoy (D2) Collage (D2--NE) Escudo (D2--NE) Itinerario de Viaje (D2--NE) Lista comidas (D2--NE) Cena de honor (D2) Leopardo (logo) (D2--NE) Collage (D2--NE) Video (D1--NE)	Dimension 1: Not enough evidence Dimension 2: Very little evidence, can't judge overall dimension.
30394	OK-Perfect	Planes de Vuelo --no incluye las reflexiones (D1) Writing: Tarjeta Postal (D2) Cloze (D1&2--1NE) Pamphlet: Autobiografía (D2) Pamphlet: Viaje (D1) Writing: Toledo (D2) Tape: La Selva (D2) Writing: La Guerra (D1&D2)	Overall evidence in this portfolio indicates student has progressed to the level that shows his ability to have general insight and understanding and display examples of diff. aspects of the culture.
		Tape (D1&2) # 1 & Tape # 2 (D1&2) Video (NE)-read! Biografía (D2) Brochure (D1&D2) Resumen (D2) # 1 & Resumen # 2 (D2) Blasón (D2--NE) Plan de vuelo (D1&D2)	Dimension 1: Promising for only 2 of the aspects; not enough evidence to judge dimension as a whole, since contributions are missing. Dimension 2: Promising (great differences between receptive skills and productive skills!)

Appendix 5 - Continuation

Student #	Agreement	Artifacts	Comments
30396		Blasón (D1&2) España (D1&2) Los Mayas (D1&2) Los Aztecas y los Mayas (D1&2) Plan de vuelo (D1&2) Goya & Picasso (D2) Composición: Yo soy (D2) Composición: A la Deriva (D2) Tape: Deriva (D2) Picasso & Goya (D1&2)	Dimension 1: Promising Dimension 2: Accomplished Note: Too much material.
		Cassette: A la Deriva (D2) Cassette: Goya y Picasso (D1&D2) Yo soy (D2) Essay: Ixtla y Popo (D1&D2) Letter: Querido Alejandro (D1&D2) Vea las selvas tropicales (D2) Plan de vuelo: Los mayas (D1&D2) Collage: Essay: Axtecas, mayas y toltecas (D1&D2) Essay: Tres obras de los mayas (D1&D2) Una España: Muchas culturas (D1&D2) Blasón y cognados del blasón (D1&D2)	Dimension 1: Accomplished Dimension 2: Accomplished Note: Yo soy -- I couldn't judge the product because student didn't include soruce. Tres obras de los mayas -- I couldn't find this document in the portfolio. General: Student included many documents and projects to show evidence of progress in Dimension 1. It shows that she worked on this dimension throughout the course.

Appendix 5 - Continuation

Student #	Agreement	Artifacts	Comments
30390	Dimension 1: NO Dimension 2: OK	Video (D2) Essay (D1&2) Essay (D1&2) Work Sheet (D2) Essay (D2) Video (D2) 3 Tapes (D1&2) Essay, Pamph. (D1&2)	Dimension 1: Advanced Dimension 2: Advanced Note: cover sheets were excellent source of writing and comprehension
		España Mayas Guernica/3 Mayo México Blasón Video: Me llamo Sonia Video: A la Deriva	Dimension 1: Promising Note: Although the student didn't have many artifacts, I could judge that he/she was at least Promising in demonstrating knowledge of hispanic cultures. Note: Student had one product to show evidence of the diversity of hisp. cultures. Dimension 2: Accomplished Note: Based on S/W; however, she didn't show evidence of reading comprehension
31948	Dimension 1: NO-NE Dimension 2: OK/NO-NE	Disk: Judge (D1&2) Worksheet: P[?] (D1&2) Comp.: Autobiografía (D2) Proyecto (D2) Proyecto (D1&2) Tape (lis.) (D1&2) Tape (speak.) (D2)	Dimension 1: Promising Dimension 2: Promising Note: She seems to be better than this evidence (because of the writing).
		WorK Sheet (D1&2) Essay (D1&2) Essay dibujos (D2) Essay (D2) Work Sheets (D1&2) List & Paragraph (D2) Tape (D2) Tape (D2)	Dimension 1: Not enough evidence. Dimension 2: Deriving meaning=no evidence/Express accomplished Note: I believe she may be better than promising but I don't have enough evidence.

Appendix 5 - Continuation

Student #	Agreement	Artifacts	Comments
30843		2 Newspaper, Articles (D1&2) Comp.: Chávez, Martí and Juárez (D1&2) A la Deriva (D2) Autobiografía (D2) Video: Ixtla/Popo (D2) Video: A la Deriva (D2) Tape: Test (D2) Flyer: Gracia[?] (D1&2)	Dimension 1: Developing Dimension 2: Developing
		Letter (D2) Essay: A la deriva (D2) Essay: Martí , Juárez y Chávez(D1&2) Exam (D2) Notes from presentations(D1&2) Las Artes de Madrid(D1&2) Arroz con Pollo(D1&2) [?] Video (D2) Cassette (D2)	Dimension 1: Promising Dimension 2: Promising Note: Although students seems to be better with reading and listening, he only included 1 sample of each to prove his progress while he included many products to evaluate his writing and speaking skills
30342	Dimension 1: OK Dimension 2: OK	Poster: Blasón (D1&2) Poster: [?] (D2) [?] Blasón Wash/Bolívar (D1&2) Carta (dimension?) Isabel/Maliche (D1&2) Español (D1&2) Carta (dimension?) Cajas (dimension?) Martí/Chávez (D1&2)	Dimension 1: Acomplished Dimension 2: Accomplished
		(All artif. used as measuring performance in both Dimension) Writing: MartíChávez Tape: Malinche/Isabel Writing: España Writing: Wash./Bolívar Writing: Cajas de Carton Poster: Blasón Poster: Quién soy?	Dimension 1: Promising Dimension 2: Promising Note: Student's portfolio has evidence for all aspects of each dimension. Very strong writing and speaking yet overall impression is of a very promising portfolio.

Appendix 5 - Continuation

Student #	Agreement	Artifacts	Comments
30814	OK-Perfect	Los Países (D1) Fusion de Culto (D1--NE) Tape: Juárez/Chávez (D1&D2) Lantaro (D1&D2) Un ? poeta (D2) Isabel/La Maliche (D2) Tape: Mi hermana (D2) Feliz cumpleaños Miami (D2) Culm. Ass. (D2)	Dimension 1: Promising. Student didn't include enough products to really say that she was accomplished on dimension 1. Dimension 2: Accomplished.
		Tape & Video: Sister (D2) Tape & Video: Popo (D2) Notes: Los países (D1) Composition: Juárez/Chávez/Martí (D1&D2) Collage: Lantaro Work Sheet: Immigrants (D2) Work Sheet: Isabel/ La Maliche (D1&D2 -didn't include page which requires inferences) Composition: Feliz Cumple. NE Letter of immigrant (D2) Tape: Medio Ambiente (D2) Tape: Cumbre (D2)	Dimension 1: Promising Dimension 2: Accomplished Note: put a cover on everything submitted.
31296	Dimension 1: NO/NE Dimension 2: OK	Cartel: Logo de Paca (D2) Carta (D2) Tape: España (D1&2) Tape: Madre Teresa (D1&2) Letter: Isla (D1&2--1NE) Work Sheets: Colón, Cortés, Isabel, D. Marina (D1) Writing: Autobiografía (D2) Writing: Food (D1&2) Summary: (D1&2--1NE)	Dimension 1: Promising Dimension 2: Accomplished Note: Choice of dimension didn't always show evidence for making appropriate judgement. Student's writing and speaking demonstrated well articulated messages.
		Poster (D2) Carta (D2) Carta/lectura (D1&2) Worksheet (D1&2) Tape (D2) Biografía (D2) Receta (D2) Research (D1&2)	Dimension 1: Not enough evidence Dimension 2: Promising

Appendix 5 - Continuation

Student #	Agreement	Artifacts	Comments
31948	Dimension 1: NO-NE Dimension 2: OK/NO-NE	Disk: Judge (D1&2) Worksheet: P[?] (D1&2) Comp.: Autobiografía (D2) Proyecto (D2) Proyecto (D1&2) Tape (lis.) (D1&2) Tape (speak.) (D2)	Dimension 1: Promising Dimension 2: Promising Note: She seems to be better than this evidence (because of the writing).
		WorK Sheet (D1&2) Essay (D1&2) Essay dibujos (D2) Essay (D2) Work Sheets (D1&2) List & Paragraph (D2) Tape (D2) Tape (D2)	Dimension 1: Not enough evidence. Dimension 2: Deriving meaning=no evidence/Express accomplished Note: I believe she may be better than promising but I don't have enough evidence.
30457	Dimension 1: OK/NO-NE Dimension 2: NO-NE	Tape (D2) Poster/Essay (D1&2) Carta (D2) [blanco en doc.] (D1&2) Work Sheet (D2) Poster (D2) Escudo (D2--NE) Essay (D1&2) Invitaciones (D2) Video (D1) Poster (D1&2--2NE) Sketch (D2--NE)	Dimension 1: Diversity=promising/Not enough evidence. Dimension 2: Promising, but my sense is that with more evidence it would be accomplished.
		Tape (C.A.) (D2) Picture seq. (D1&2--1NE) Carta (D2) I. Pascua (D2) C. de hoy (D2) Collage (D2--NE) Escudo (D2--NE) Itinerario de Viaje (D2--NE) Lista comidas (D2--NE) Cena de honor (D2) Leopardo (logo) (D2--NE) Collage (D2--NE) Video (D1--NE)	Dimension 1: Not enough evidence Dimension 2: Very little evidence, can't judge overall dimension.

Appendix 5 - Continuation

PORTFOLIOS RATED BY TWO READERS WORKING TOGETHER

Student #	Agreement	Artifacts	Comments
30393		Plan de vuelo: Mayas (D1&2) Autobiografía (D2--NE) Folleto (D1) Blasón: Poema (D1&2) Actividad: Paintings (D1&2) Popo/Ixtla (D1) Story: A la deriva (D2) Letter: Isla de Pascua (D2) Poem (D1&2)	Dimension 1: Promising Dimension 2: Promising Note: Reaks of promise more than [?] Note: Peaks are more numerous in the area of promising. Overall impression is of a promising portfolio.
30357		Guide: España(D1) Wr. Report: México (D1&2) Cinta: Toledo (D1&2) Summary:Ixtla/Popo (D2) Report: Martí/Juárez (D2) Interview: Profesor L. (D2) Reading: Cajas de Carton (D2) Tape: La Malinche (D2)	Dimension 1: Accomplished Dimension 2: Accoplished Note: Overall prducts demonstrate that student isaccomplished in both dimensions. Note: Overall prdocut demonstrates accomplished production; however, speaking [??]
30131		Summary: Popo/Ixtla (D1&2--1NE) Summary: Isla de Pascua (D2) Carta (D2) Diario (D2) Proyecto: Biografía (D2) Proyecto: Comp. (D1&2) Monumento (D2) Tape: C.A. (D1&2--1NE) Debate (D1&2--1NE) Proyecto: Animal (D2) Debate (D2) Cumbre (D2)	Dimension 1: Beginning Dimension 2: Accomplished

Appendix 5 - Continuation

Student #	Agreement	Artifacts	Comments
30149		Los Mayas (D1&2) Popo e Ixtla (D2) México y España (D1) Ansel Adams (D2) Nuestros Héroes (D2) C.A. (D2) Autobiografía (D2) Cena de Honor (D2) Animales en peligro (D2) La mode (D2)	Dimension 1: Promising, but I'm not so sure since there wasn't enough evidence to judge two aspects. Dimension 2: Promising, based two samples. Note: In general, I think this portfolio needs more products to have enough evidence to judge it.
31942		Tape (D2) Párrafo: Poeta (D2) Essay (D2) Speaking (D2) Receta (D1&2) Notes from heading (D1&2) Essay (D1&2) Speech [?](D2) Essay (D1&2) Note (D2--NE)	Dimension 1: Promising Dimension 2: Promising expresion only. Not enough evidence to judge receptive skills.
30819		Class Notes (D1&2) Artículo (D2) Isabel/La Malinche (D1&2) Work Sheet (D1&2) Madrid y las artes (D2) Tape: A la deriva (D2) Comp.: Chávez (D2) Tape: Popo (D2)	Dimension 1: Accomplished Dimension 2: Accomplished Note: Poorly organized. Very few evidence of inferencing [?]. Note: Some connections were made but were not checked as evidence.

Appendix 5 - Continuation

Student #	Agreement	Artifacts	Comments
30470		Poster (D2) Essay (D2) Essay (D1&2) Essay (D1&2--2NE) Lista Comida (D1&2--2NE) Invitaciones (D2) Tape (D2) Tape: Columbia (D1&2) Outline (D1&2--2NE) Foto (D1&2--NE) Sketch (D1&2--NE) Essay (D2) Carta (D2) Wrok Sheet (D2)	[Readers didn't include evaluation for each dimension]
31274		Essay (D1&2) None (D2--NE) Work Sheet (D1&2) Cover&Essay (D1&2) Work Sheet (D1&2) Essay (D1&2) Essay (D2) Essay (D2) Work Sheet (D1) Essay (D1&2) Essay (D1&2) Work Sheet & Cover Sheet (D1&2) Work Sheet & Cover Sheet (D1&2)	Dimension 1: Advanced Dimension 2: Accomplished Note: consideredaspects not noted by student

Appendix 5 - Continuation

PORTFOLIOS RATED BY ONE READER

Student #	Agreement	Artifacts	Comments
30498		Tape (D2) Tape (D2) Tape (D2) Work Sheet (D1) Essay (D2) Drawing (D2) Poster (D1&2) Essay (D2) List (D1&2) Essay (D1&2) Soup (D2)	Dimension 1: Not enough evidence Dimension 2: Promising

Appendix 5 - Continuation

Correspondence of Student # and Portfolio #

Student #	Portfolio #
31938	1
30129	2
30831	3
31929	4
31296	5
30351	6
30358	7
30144	8
30460	9
30816	10
30435	11
30124	12
30476	13
30457	14
30394	15
30396	16
30390	17
31948	18
30843	19
30342	20
30814	21
30131	22
30149	23
30470	24
30498	25
31274	26
31942	27
30819	28
30393	29
30357	30

Phase II
Appendix

Appendix A

Preparation for Portfolio Assessment Scoring Sessions

Portfolio Assessment Overview

Both the Spanish and English Pacesetter courses use portfolios as a measure of student achievement. In each subject, portfolio collections of regular classroom work are developed by students throughout the school year as a way to demonstrate their achievement in relation to the course dimensions. In both Pacesetter English and Spanish, the scoring guides used to evaluate the portfolios at the end of the semester or year are used throughout the course to clarify course expectations and serve as a guide to help students build strong collections of work.

Portfolios provide a way of looking at student achievement that is quite different from the other types of Pacesetter assessments. Whereas common tasks and even the culminating assessment look at student achievement on a particular task at a given time, the portfolio provides a view of a student's achievement across a body of various kinds of class work. Teachers who have participated as Readers at portfolio scorings often comment that looking at work from other Pacesetter classes provides models of instructional activities and strategies they can adapt for their own teaching.

Pacesetter English

As with the culminating assessment, portfolio assessment focuses on evaluating student performance in the two course dimensions of **Making Meaning from Texts** and **Creating and Presenting Texts**. Because of the design of the portfolio scoring guides, it is possible to gather performance-level information about four aspects of each of the two course dimensions as well as the two dimensions themselves. Teachers-Readers indicate that looking at student performance on various aspects of the course dimensions is especially helpful in planning the course and reflecting on their teaching.

Pacesetter Spanish

Like the English portfolio, the Spanish portfolio also focuses on evaluating student performance in two course dimensions: **Demonstrating Knowledge of Hispanic Cultures** and **Using Spanish to Communicate Effectively**. Portfolio scoring guides for Pacesetter Spanish also make it possible to evaluate student performance on several aspects of these dimensions through a single reading of a portfolio.

The Relationship Between Scoring Purpose and Scoring Design

Scoring portfolios can provide information about student achievement and provide ideas for staff development. Because of the complex nature of portfolios and the time it takes to score them (about 25 minutes per reading), different purposes can make seemingly opposing demands on the time and activities of a portfolio reading.

In order to monitor the technical quality of scores and to determine their reliability, either a sample of or all of the portfolios need to be read twice. Having extended discussions about the scoring criteria and how they are evidenced in student work supports staff development purposes.

Decisions need to be made about the use of time and how that relates to the purpose(s) of the reading. Should fewer portfolios be read so that time can be allocated for extended discussions? Should a larger sample of portfolios be read and the discussions be limited to training issues? Should there be a balance in the use of time to reflect both purposes? Are there other local purposes that might have an impact on the portfolio scoring design?

The following are some examples of the connection between scoring purpose and design:

- If the purpose of scoring portfolios is to attain scores to monitor student achievement, then a representative sample of portfolios needs to be read. Some or all of the sample needs to be read twice.
- In some cases the purpose of a portfolio reading might be to provide feedback to teachers about how they are applying the scoring criteria in their individual classrooms. In this design, teachers might be asked to prescore a representative sample of portfolios from their Pacesetter classes and submit those portfolios for rescoring at the districtwide reading. Scores from the centralized reading would be returned to teachers.
- If the primary purpose of the reading is to provide staff development, maximizing the time for teacher discussion about student performances in relation to the scoring criteria is important. Scoring a smaller representative sample of portfolios would be appropriate.

Depending on the purpose for the portfolio scoring, Readers might be asked to score each portfolio for the course dimensions and the aspects of the dimensions through a single reading of a portfolio. Some believe that this design models how teachers evaluate student performances in their classrooms where teachers consider several aspects of performance at the same time.

It is also reasonable to organize a portfolio scoring so that Readers score only one dimension and its aspects at a time. If a district is especially concerned about a particular dimension of student achievement, this design allows Readers to focus more attention on one dimension of performance. If the intent is to gain scores for all dimensions of performance within a subject area, this design requires more time because each portfolio must be read separately for each dimension.

Your district may have other purposes for holding a portfolio reading that may necessitate other designs. Determining what information is needed by whom and for what purposes should guide your district in designing a portfolio scoring to meet your needs. No matter the purpose of the reading, coming together to examine student work in portfolios is a powerful way to help Pacesetter teachers clarify and internalize the Pacesetter content and performance standards.

For more information about building and assessing portfolios, please refer to the introduction in the *Teachers' Guides* for Pacesetter English and Spanish.

The Roles and Responsibilities of the Participants

Scoring Coordinator

As with the culminating assessment, a portfolio scoring needs a coordinator within the district who can champion the decision making related to the scoring design, take responsibility for planning the scoring, involve other personnel as needed, and make necessary arrangements. In some districts, one person may serve as both the Scoring Coordinator and Chief Reader.

Chief Reader(s)

Because portfolio assessment is a relatively new practice there are few experienced Readers. An experienced Pacesetter teacher or district administrator who knows the Pacesetter course well, understands the course content and performance standards fully, and who can talk about how to find evidence of the standards in student performances, can make a fine Chief Reader. Responsibilities of the Chief Reader include selecting benchmark portfolios to use during the training, conducting the training, and facilitating discussions about evidence of achievement in student work.

Because training for portfolio scoring requires Chief Readers to know the course dimensions, scoring criteria, and benchmark portfolios well, some districts, or Chief Readers, may feel most comfortable with more than one Chief Reader. Having multiple Chief Readers may be especially helpful in districts that are scoring portfolios for the first time, since it allows Chief Readers to combine their expertise and knowledge. Chief Readers may be most comfortable focusing on only one course dimension and its aspects. In some cases, the design of a portfolio scoring may require more than one Chief Reader, especially if Readers will be scoring each dimension separately.

Table Leaders

Whether or not Table Leaders are needed at a portfolio reading is determined by the size and purpose of the reading. Because of the time it takes to read portfolios, Table Leaders do not generally monitor the scoring of Readers at their table in the same way they might at a culminating assessment reading. Instead, their primary responsibilities are to assist with the training and to answer questions that arise during the reading. Each 8 to 10 readers will need a Table Leader (or Chief Reader) they can turn to with questions.

The purpose of the reading will determine the number and role of Table Leaders.

- If gathering student achievement information is the primary purpose, Table Leaders might be expected to turn their attention first to reading portfolios and second to answering questions, because reading a larger sample of portfolios is necessary. They might be used as an additional Reader for selected portfolios or as a discussion facilitator to help Readers agree on a score.
- If staff development is the focus, Table Leaders might lead additional table-level discussions about particular issues, facilitate discussions

about portfolios between scorers with differing scores, bring questions and concerns to the attention of the Chief Reader, and only score portfolios as time permits.

Readers

Readers are usually current or future Pacesetter teachers or administrators familiar with the Pacesetter course. In some cases, teachers who do not teach Pacesetter, but who are interested in portfolio assessment or standards-based assessment, might be interested in the experience of scoring Pacesetter portfolios as a professional development opportunity. Some districts plan to involve local teacher training institute staff or community members in readings. As with the culminating assessment, compensation and other incentives or conditions are governed by local conditions.

Aides

Although Table Leaders and Chief Readers can, if necessary, distribute, collect, and organize portfolios and scoring guide forms, Aides can be a tremendous help and allow more time for Table Leaders and Chief Readers to answer scorer questions and lead discussions. Generally, one Aide is needed for each 12 Readers.

Number-Person

If scores from the portfolio reading are to be summarized at the district level, it may be helpful to involve a staff member who can set up and maintain a data base for portfolio scores and create score reports, including score distributions, as needed. There is no calculation needed to determine a proficiency score for portfolios because the "raw" score is the proficiency level as stated in the portfolio scoring guide for that subject.

Identifying the Portfolio Sample

Sample Size

As with other decisions, the purpose of the reading is important in determining the number of portfolios that can be read. If more time is used for training and discussion, a smaller sample of portfolios can be read. The following calculation can help determine the size of the sample that can be read by the number of available Readers in the allotted time.

Worksheet for Determining the Sample Size of Portfolios

Narrative Explanation	Algorithm	Calculation
A. Total hours available = Number of days for scoring x 5.0 hours (total hours a day minus lunch and breaks)	Days x 5.0 = A	
B. Scoring hours available = A minus 6 hours for training (training time may differ by district)	A - 6 = B	
C. Number of Readers available	Readers = C	
D. Total possible readings = Multiply B by C by 2.5 (portfolios read per hour)	B x C x 2.5 = D	
E. Number of portfolios to be scored twice	Portfolios to be double-scored = E	
F. Number of portfolios that can be read including those double-scored = D minus E	D - E = F	
G. Number of classes	Classes = G	
H. Number of portfolios that can be read from each class = F divided by G	F ÷ G = H	

Sample Selection

If your district determines that not all of the Pacesetter portfolios can be scored at a centralized reading, then a representative sample needs to be identified for scoring. If a sample rather than all of the portfolios is being read, the sample should be as random and representative as possible. Some possible scenarios for sampling are listed below.

- In many districts, selecting the sample may fall to the Pacesetter teachers themselves. If this is the case, teachers should be given guidelines about how to select a sample that is representative and balanced for gender or other demographic features.
- In some districts, a district-level staff member may identify a random sample of students from whom portfolios would be gathered and read.

- In other districts, teachers might send all of their portfolios to a central location where district-level staff select a random sample, balancing for gender or other demographic features.

Determining the Number of Scorers

If a large pool of scorers is available or if a decision has been made to score all of the portfolios, then the calculation below can be used to determine the number of Readers needed to score the portfolios.

Worksheet for Calculating the Number of Scorers Needed

Narrative Explanation	Algorithm	Calculation
A. Total number of portfolios to be scored	Portfolios = A	
B. Number of portfolios that will be double-scored (times each portfolio will be read)	Double Readings = B	
C. Number of total readings = Add A to B	A + B = C	
D. Number of scorer hours needed = C divided by 2.5 (portfolios per hour)	C ÷ 2.5 = D	
E. Total hours available = Number of days for scoring x 5 hours (total hours a day minus lunch and breaks)	Days x 5.0 = E	
F. Scoring hours available = E minus 6 hours for training (training time may differ by district)	E - 6 = F	
G. Number of scorers needed = D divided by F	D ÷ F = G	

Scheduling of Portfolio Scorings

Because portfolios are built throughout the school year, the timing of a portfolio reading is dependent upon the school calendar and teacher availability. Holding a scoring a few weeks before the end of the school year allows teachers to use the training they receive and any portfolio scores that are available to help them determine student grades. However, districts may choose to have teachers collect a representative sample of portfolios for use at a summer portfolio reading.

Facilities: Planning a Space for the Scoring Session(s)

As with the culminating assessment, a comfortable, well-lit, quiet location with reading tables is needed for a portfolio scoring. Generally, two portfolio Readers can comfortably read at a cafeteria table. The Chief Reader(s) is likely to need flip charts, tape, an overhead projector, and marking pens. Depending on the scoring design, number of portfolios, and security of the scoring room, a preparation room for organizing and storing portfolios may also be necessary.

Preparing Assessment Portfolios for Scoring

Coding portfolios can ease the collection of scores as well as the return of portfolios. Your district may already have a system of student numbers or some other unique code that may prove useful in portfolio identification. If not, you may want to create a simple coding system and enlist teachers and students to help with coding portfolios prior to the reading.

If a portfolio code is used, a unique number for each student is needed. A code that includes a number for the teacher, class period, and student may be the easiest. Precoded labels that students affix to their portfolios prior to the reading are helpful. With the portfolios submitted for the reading, teachers should include a list of students and portfolio code numbers.

In order to prepare portfolios for scoring, some districts may decide to have portfolios sent to a central location prior to the reading. Organizing the portfolios might include selecting a random sample to be read. It may include checking or affixing portfolio codes or creating a data base for collecting scores. It might also include creating batches of portfolios from various classes to be read at particular tables. In some cases, it may be possible to have Aides organize portfolios on the morning of the reading, if they are not involved in the training.

The scoring guide forms (pages 53-60) are meant to be copied and used at the reading. Readers mark their scores directly on these scoring guides. If Aides are available before the reading, they can precode the forms and place them in the portfolios. If Aide time is more limited, on the day of the reading place copies of the scoring guides on the tables for Readers. For organizational purposes, it is easier to copy each dimension of a subject area on a different color paper.

Assigning each Reader a unique code that can be used as an identifier on the scoring guide forms may make score and reliability analyses easier. Determining whether to give Reader codes depends on local circumstances.

Choosing Benchmark Portfolios to Guide the Training

Each district holding a portfolio scoring will need to select benchmark portfolios to use during the training. Because portfolios are complicated documents and require extended time to read and discuss, training is usually conducted using a small number (three to four) of benchmark portfolios. As with the culminating assessment, the Chief Reader(s) should work with two to three other Pacesetter teachers to select benchmark portfolios.

Benchmark portfolios should represent different classrooms and include a variety of types of performances. Suggested selection criteria for benchmarks include:

- one portfolio that demonstrates strong, but not necessarily exemplary, performance;
- another that demonstrates uneven performance across dimensions or aspects; and
- one or two other portfolios that are likely to bring about discussion on important issues such as evaluating English Language Learner performances, group versus individual performances, or other local concerns.

In all cases, benchmark portfolios must include work that can be easily read and copied.

Making copies of benchmark portfolios can be a time-consuming process that includes:

- making a "master copy" of each benchmark portfolio (may take up to an hour to make the first copy);
- blocking out student, teacher, and school names as well as other personal information that would identify the student;
- tracing writing and other marks that are too light to read;
- numbering each page in the "master copy";
- adding a title page (for example: Benchmark A) to the "master copy"; and
- making multiple copies for use at the training.

While some Readers find it difficult to read a portfolio with another person, it is not uncommon to have Readers share a copy of benchmark portfolios during training. Reading a portfolio in pairs tends to promote discussion by the pair about how specific evidence is viewed and the judgments that are being made. These discussions are important in helping Readers come to a common understanding of what constitutes evidence of the dimensions and how to apply the scoring criteria.

Preparing Table Leaders

At portfolio readings where Readers are scoring all dimensions of performance at the same time, most of the training is conducted by the Chief Reader. Table Leaders work with Readers at their table to help them internalize the scoring criteria and use the scoring guide forms correctly. Table Leaders answer questions and refer problematic portfolios to the Chief Reader, as appropriate.

If your design allows, Table Leaders can read and score portfolios as the other Readers do. Because it takes an average of 25 minutes to read and score a portfolio, rereading scored portfolios is not practical unless it is part of a double-scoring design or unless a Table Leader or Chief Reader has concerns about a particular Reader.

Prior to the reading, Table Leaders should be given copies of the benchmark portfolios so they can familiarize themselves with the evidence. Any notes that might have been created by the Chief Reader(s) about the assigned score for each benchmark and the evidence that supports the scoring judgment will be helpful to Table Leaders as they answer Reader questions. It is also helpful for Table Leaders to meet with the Chief Reader(s) prior to the reading to discuss the benchmarks and possible questions and issues that might need to be addressed at the reading.

Training the Readers

Portfolio-Reading Assumptions

A centralized portfolio reading, as described in this *Handbook*, is based on several underlying assumptions, including:

- scorers need to come to a common understanding of the scoring criteria;
- the criteria in the scoring guides need to be applied consistently;
- scorers need to focus on the evidence of achievement in the work;

- holistic judgments are based on the quality not the quantity of evidence; and
- scores reflect holistic judgments across the body of work in a portfolio.

Suggested Strategies for Reading Portfolios

Because portfolios often contain a large and varied collection of student work, they can seem unwieldy to read at first. Most Readers will have little or no experience looking across a body of work to make judgments. Some of the following suggestions might be useful for Readers.

- Skim the entire portfolio to see the types of performances that are included.
- If multiple drafts are included, read the "final" draft closely, then skim the other drafts.
- If several performances reflect the same type of work, such as a reading log or several poems, read some selections closely and skim the others.
- Keep notes about evidence as the portfolio is read (sample evidence charts for English and Spanish are included at the end of this document). Some Readers prefer to use "yellow stickies" to mark particular pages as they read.
- Ignore teacher grades and comments.

Assigning Scores

Often the quality of performances varies within a portfolio, making it difficult for Readers to assign a score. Readers are generally asked to assign a score that reflects the preponderance of the evidence.

Sample Training Design

This section describes training for a portfolio reading in which each Reader is scoring all dimensions and aspects of performance within a subject area from a single reading of a portfolio. A training process reflecting this design is summarized in the following chart. A narrative explanation of the activities follows the chart. Approximate times for each activity have been included for planning purposes.

Sample Training Activity	Approximate Time
Training on the first dimension with first benchmark portfolio	
• Chief Reader introduces the scoring guide for one dimension	25 min.
• Readers work in pairs reading, discussing, and scoring the first benchmark portfolio	40 min.
• Scores and issues are discussed in the whole group	25 min.
Training for additional dimension with first benchmark portfolio	
• Chief Reader introduces the scoring guide for an additional dimension	25 min.
• Readers work in the same pairs to review, discuss, and score the first benchmark portfolio for the other dimension	20 min.
• Scores and issues are discussed in the whole group	25 min.
Training on second benchmark portfolio	
• Chief Reader reviews the scoring guide(s) as needed	15 min.
• Readers work either in pairs or individually to read and score a second benchmark portfolio for all aspects of performance they will be scoring during the remainder of the reading	30 min.
• Scores and issues are discussed at each table	15 min.
• Issues and questions are discussed as a whole group	30 min.
Additional Training	
• Discussion about issues and questions related to the scoring guides takes place after each major break	
• Training on additional benchmark portfolios is done as needed	
• Revisiting previously read benchmark portfolios to discuss issues is done as needed	
• Individual help is provided as appropriate	

As with the culminating assessment reading, the Chief Reader will want to set a professional tone by welcoming Readers, outlining the scoring task, and introducing Table Leaders and any guests. It is important to acknowledge that Readers come to the reading with a sense of what strong student performance looks like. However, it is essential that Readers develop a shared understanding of the scoring criteria and learn to apply the criteria consistently.

Because portfolio scoring guides are complex, training is conducted one dimension at a time. The Chief Reader begins by reviewing the scoring guide, focusing on the definitions of the aspects and the criteria for the "accomplished" and "promising" levels of performance. Involving Readers in actively looking at and discussing the scoring criteria is important to their understanding of the differences between various aspects and performance levels. Chief Readers should then discuss the "not enough evidence to judge" category, and other issues that the Chief Reader and Table Leaders identified prior to the reading.

Readers then work in pairs, reading together and discussing the first benchmark portfolio. After a pair has reviewed the entire benchmark portfolio, they should discuss and come to agreement on scores. The Chief Reader then collects, usually by a show of hands, all of the scores for that benchmark portfolio. After the scores are tallied, the Chief Reader involves Readers in a discussion of each aspect of performance. Part of helping Readers reach a common understanding is sharing the consensus score that had been assigned before the reading and citing specific evidence that was used to arrive at that score. This process of reading the first benchmark portfolio and training on one dimension will take about 90 minutes, depending on the number of Readers and issues that arise.

After discussion is completed for the first scoring dimension, repeat the process for each additional dimension. The Chief Reader should begin by reviewing the scoring guide and involving Readers in identifying important features for the "accomplished" and "promising" performance levels. Readers should continue to work in the same pairs, reviewing and scoring the same benchmark portfolio, this time for the additional dimension. After Readers have rescored the portfolio for this dimension, the Chief Reader again collects the scores, and leads a discussion about the evidence and consensus scores for each aspect. Because Readers are assessing a portfolio they have read before, this process of scoring and discussing the second dimension will likely take about 70 minutes.

Training on the second benchmark portfolio is slightly different from training on the first. The scoring guides may need only a brief review before Readers start reading the portfolio. Depending on the purpose of the portfolio reading and the number of copies of benchmark portfolios that are available, Readers may read the second benchmark portfolio either in a pair or individually. They should, however, be reading and scoring as many dimensions and aspects as they will be expected to score during the rest of the reading. After a reasonable time for reading and scoring the benchmark (40–45 minutes), Table Leaders should collect scores and lead a discussion at their table, taking note of issues and questions. The Chief Reader can then conduct a room-wide discussion about the scoring of the benchmark, addressing specific questions on the scoring guides and commenting on issues and questions that were raised at the tables. The training on the second benchmark is likely to take about 90 minutes.

At this point, if it seems that most Readers are scoring at or within one point of the consensus score, the Chief Reader and Table Leaders may decide to move ahead with scoring rather than training on another benchmark. It is also possible to train on the next benchmark at the table level, having Readers read, score, and discuss aspects on which they disagree. If more than one reading room is available, it is possible to have some Readers who seem confident start scoring portfolios while others receive additional training.

After every significant break, such as for lunch or overnight, the Chief Reader or Table Leaders will need to answer questions that Readers have about the scoring guide. If necessary, the Chief Reader may decide to retrain Readers using a new benchmark portfolio on the morning of the second day. It is also possible to reuse one of the benchmark portfolios that has already been discussed as a way to refocus Readers on particular aspects or issues.

Controlling the Flow of Portfolios

Limiting the movement of people and portfolios will make a reading more productive. One design for distributing portfolios is to organize portfolios from different classrooms in batches (10–12) and distribute a batch to each table at the beginning of the day, after lunch, and as needed. Readers select and read a portfolio from the batch, avoiding any from their own students. Portfolios that have been read are placed in a "read once" stack. If the reading design calls for portfolios to be read a second time, Readers should select portfolios from the "read once" stack and when finished, place them in a "read twice/completed" stack. Completed portfolios should be removed from the table periodically. When the Readers have completed a batch, then a new batch should be brought to the table. This design both reduces the movement of portfolios in the room and easily allows for double readings of portfolios, if the design calls for that.

Collecting the Readers' Scores

The easiest system for collecting scores is to photocopy the completed scoring guide forms, making sure that scores and identifying information for the portfolio and Reader are clearly visible. Some districts may wish to develop their own system for collecting portfolio and Reader information and scores.

Reporting the Results

Portfolio scores are reported as performance levels (using level descriptors from the scoring guide such as "Promising," "Accomplished," etc.). However, it may be important to discuss which scores should be reported to which audiences. While teachers may find the scores for aspects of the dimensions useful in planning instruction, school board members may not need this level of detail. Deciding which scores to report is an important decision each district will need to make.

As with the culminating assessment, your district may want to produce a number of different reports using the portfolio results, including individual score reports. The easiest form of an individual report is a copy of the actual score form that includes performance-level descriptions. In addition, many teachers and administrators may find distributions to be useful. Caution should be used in creating distributions at the classroom or school level because small samples may not be representative of the student population.

APPENDIX B

COVER SHEET (MUST ACCOMPANY EACH ARTIFACT)

PACESETTER SPANISH PORTFOLIO ASSESSMENT

NAME: _____

Selection: _____ Date work was done: _____

Why did you select this piece of work?

What does it show or tell about you?

What did you learn from doing this assignment?

This piece of work shows evidence of the following dimension(s) or aspect(s):

- **<u>Dimension 1: Demonstrating knowledge of Hispanic cultures</u>**

___ Showing awareness of the diversity of Hispanic cultures

___ Identifying contributions of Hispanic figures

___ Making connections with other disciplines and own culture(s)

- **<u>Dimension 2: Using Spanish to communicate effectively</u>**

___ Deriving meaning from printed materials and oral discourse

___ Expressing meaning in oral and written form

COVER SHEET (MUST ACCOMPANY EACH ARTIFACT)

PACESETTTER SPANISH PORTFOLIO ASSESSMENT

NOMBRE: _____

Título: _____ Fecha de ejecución: _____

¿Por qué seleccionaste esta obra?

¿Qué muestra o dice esta obra sobre ti?

¿Qué aprendiste de esta tarea?

Esta obra indica evidencia en los siguientes aspectos y dimensiones:

- **<u>Dimensión 1: Demostrar conocimiento de las culturas hispanas</u>**

____ Demostrar conciencia de la diversidad de las culturas hispanas

____ Identificar contribuciones de figuras hispanas

____ Hacer conexiones con otras disciplinas y cultura(s)

- **<u>Dimensión 2: Usar el español para comunicaciones eficientes</u>**

____ Comprensión de textos y comprensión auditiva

____ Expresión oral y escrita

COVER SHEET FOR PACESETTER SPANISH PORTFOLIO

Please list all artifacts included in your portfolio. Indicate the aspect(s) and dimension(s) for which you are showing evidence with each artifact you selected.

DIMENSION 1: Demonstrating knowledge of Hispanic cultures

- **Showing awareness of the diversity of Hispanic cultures**

- **Identifying contributions of Hispanic figures**

- **Making connections with other disciplines and own culture(s)**

DIMENSION 2: Using Spanish to communicate effectively

- **Deriving meaning from printed materials and oral discourse**

- **Expressing meaning in oral and written form**

Appendix C

PACESETTER SPANISH *¿NOS CONOCEMOS?*
DIMENSION 1: DEMONSTRATING KNOWLEDGE OF HISPANIC CULTURES
Portfolio Assessment: Evaluating aspects of the dimension

Showing awareness of the diversity of Hispanic cultures	**Beginning** ☐	**Developing** ☐	**Promising** ☐	**Accomplished** ☐	**Advanced** ☐	**Not enough evidence to judge** ☐
• reflection on how diverse the Hispanic world is • examples of several cultures of the Hispanic world (more than one) • reflection on issues like current events, historical highlights, geography, literature, etc.	Student presents **superficial, vague and fragmentary evidence** of his/her awareness of the diversity of Hispanic cultures.	Student presents **limited, incomplete and uneven evidence** of his/her awareness of the diversity of Hispanic cultures. Student presents **few specific examples**.	Student presents **evidence of a general awareness** of the diversity of Hispanic cultures. Student presents **some general insights and examples**.	Student presents **ample evidence** of his/her awareness of the diversity of Hispanic cultures. Student presents **a wide range of examples, with some clear insights**.	Student presents **consistently insightful evidence** of his/her awareness of the diversity of Hispanic cultures. Student presents **varied and extended examples**.	Contains **too little evidence** of awareness of the diversity of Hispanic cultures to justify a score.

Identifying contributions of Hispanic figures	**Beginning** ☐	**Developing** ☐	**Promising** ☐	**Accomplished** ☐	**Advanced** ☐	**Not enough evidence to judge** ☐
• examples of several Hispanic figures • contributions of contemporary as well as historical figures of the Hispanic world in different fields • contributions of ethnic groups	Student presents **superficial, vague and fragmentary evidence** of contributions of Hispanic figures. Student presents **lists with a few names and descriptions**.	Student presents **limited and uneven evidence** of contributions of Hispanic figures. Student presents **few specific notions and descriptions of contributions**.	Student shows **evidence of some general** contributions of Hispanic figures. Student presents **some inferences about place and impact in history**.	Student presents **ample evidence** of contributions of Hispanic figures. Student presents **ample inferences about impact and role in history**.	Student presents **consistently thorough evidence** of contributions of Hispanic figures. Student presents **insightful inferences and analysis of contributions, as well as impact and role in history**.	Contains **too little evidence** of contributions of Hispanic figures to justify a score.

Making connections with other disciplines and own culture(s)	**Beginning** ☐	**Developing** ☐	**Promising** ☐	**Accomplished** ☐	**Advanced** ☐	**Not enough evidence to judge** ☐
• links to own culture(s) and personal experiences • comparison and contrast of Hispanic cultures with student's own culture(s) • connection of historical, economic, literary, cultural, political, geographic, and environmental information on Spanish -speaking countries to student's work in social studies, English, and other areas of the curriculum	Student presents **superficial and isolated evidence** of connections with other disciplines and own culture(s). Student presents **irrelevant examples**.	Student presents **limited, partial and uneven evidence** of connections with other disciplines and own culture(s). Student presents **examples that are not always relevant**.	Student presents **evidence of some general** connections with other disciplines and own culture(s). Student presents **examples without much depth or insight**.	Student presents **ample evidence** of connections with other disciplines and own culture(s). Student presents **some inferences and a wide range of insightful examples**.	Student presents **consistently thorough evidence** of connections with other disciplines and own culture(s). Student presents **new, creative, and original examples**.	Contains **too little evidence** of connections with other disciplines and own culture(s) to justify a score.

PACESETTER SPANISH *¿NOS CONOCEMOS?*

Portfolio Assessment: Evaluating Dimension 1

DEMONSTRATING KNOWLEDGE OF HISPANIC CULTURES	**Beginning** ☐	**Developing** ☐	**Promising** ☐	**Accomplished** ☐	**Advanced** ☐	**Not enough evidence to judge** ☐
• Consider performance on all three aspects included in this dimension.	The collection presents **superficial, vague, and fragmentary evidence** of knowledge of Hispanic cultures. Student presents few and irrelevant examples.	The collection presents **limited and uneven evidence** of knowledge of Hispanic cultures, with few specific examples, which may not always be relevant.	The collection presents **evidence of a general** knowledge of Hispanic cultures, with some general examples and insights.	The collection presents **ample evidence** of knowledge of Hispanic cultures, with a wide range of insightful examples.	The collection presents **consistently insightful and thorough evidence** of knowledge of Hispanic cultures, with varied and original analysis and examples.	The collection contains **too little evidence** of knowledge of Hispanic cultures to justify a score.

PACESETTER SPANISH *¿NOS CONOCEMOS?*
DIMENSION 2: USING SPANISH TO COMMUNICATE EFFECTIVELY
Portfolio Assessment: Evaluating aspects of the dimension

Deriving meaning from printed materials and oral discourse	**Beginning** ☐	**Developing** ☐	**Promising** ☐	**Accomplished** ☐	**Advanced** ☐	**Not enough evidence to judge** ☐
• skills included: **listening** and **reading** • comprehension and interpretation of printed materials on diverse topics of varying levels of complexity and length • comprehension and interpretation of oral discourse on diverse topics of varying levels of complexity and length	Student presents **limited evidence** of deriving meaning from printed materials and oral discourse. Student **understands a few general ideas, with continuous misunderstandings**.	Student presents **uneven evidence** of deriving meaning from printed materials and oral discourse. Student **understands some main ideas on familiar topics, with frequent misunderstandings**.	Student presents **some evidence** of deriving meaning from printed materials and oral discourse. Student **understands main ideas and some details on somewhat familiar topics and may make some inferences depending on complexity of text**.	Student presents **appropriate and meaningful evidence** of deriving meaning from printed materials and oral discourse. Student **understands main ideas and most details even on unfamiliar topics, and can make inferences appropriate to the task**.	Student presents **thorough evidence** of deriving meaning from printed materials and oral discourse. Student **extrapolates information in new ways, including some social and cultural references and affective overtones**.	Contains **too little evidence** of deriving meaning from printed materials and oral discourse to justify a score.

Expressing meaning in oral and written form	**Beginning** ☐	**Developing** ☐	**Promising** ☐	**Accomplished** ☐	**Advanced** ☐	**Not enough evidence to judge** ☐
• skills included: **speaking** and **writing** • develop and present texts orally • develop and present written texts in final drafts • demonstrate technical command of the Spanish language (language usage, structures and conventions)	Student presents **limited and fragmentary evidence** of expressing meaning in oral and written form. Student presents an **inadequate technical command of the Spanish language** (e.g., lack of fluency, isolated sentences, incomplete utterances). **Communication impeded by interference from another language.**	Student presents **uneven evidence** of expressing meaning in oral and written form. Student presents **limited technical command of the Spanish language** (e.g., limited fluency, unconnected discourse or sentences, limited range of vocabulary). **Some interference from another language may hamper communication.**	Student presents **some evidence** of expressing meaning in oral and written form. Student presents **some underlying technical command of the Spanish language** (e.g., emerging notion of adequate register, adequate range of vocabulary, self-corrections). **Minimal interference from another language does not hamper communication.**	Student presents **appropriate and meaningful evidence** of expressing meaning in oral and written form. Student presents **proper technical command of the Spanish language** (e.g., organized and articulated messages, proper use of circumlocution, at times creative ideas, coherent use of vocabulary and mostly adequate use of register). **Almost no interference with communication.**	Student presents **highly creative, thorough, and meaningful evidence** of expressing meaning in oral and written form. Student presents **clear technical command of the Spanish language** (e.g., well articulated and organized messages, appropriate use of register, wide range of vocabulary, and a high level of fluency). **Occasional errors do not interfere with communication.**	Contains **too little evidence** of expressing meaning in oral and written form to justify a score.

PACESETTER SPANISH *¿NOS CONOCEMOS?*

Portfolio Assessment: Evaluating Dimension 2

USING SPANISH TO COMMUNICATE EFFECTIVELY	**Beginning** ☐	**Developing** ☐	**Promising** ☐	**Accomplished** ☐	**Advanced** ☐
• Consider performance in the two aspects included in this dimension.	The collection presents **limited evidence** of using Spanish to communicate effectively and an inadequate technical command of the Spanish language.	The collection presents **uneven evidence** of using Spanish to communicate effectively and a limited technical command of the Spanish language.	The collection presents **some evidence** of using Spanish to communicate effectively and some underlying technical command of the Spanish language.	The collection presents **appropriate and meaningful evidence** of using Spanish to communicate effectively and proper technical command of the Spanish language.	The collection presents **creative, thorough and meaningful evidence** of using Spanish to communicate effectively and clear technical command of the Spanish language.

Not enough evidence to judge ☐
The collection contains **too little evidence** of using Spanish to communicate effectively to justify a score.

APPENDIX D

Dimension 1		Dimension 2
Comments on Artifacts		Comments on Artifacts
Diversity of Hispanic Cultures: 1		Deriving Meaning from Printed/Oral Materials: 1
-Would be useful evidence for aspect 3 -Only underlined info.; needs to use cover sheet to explain? -Only a list of foods -No ample inferences or impact -No insights included -Does not mention (Popo) -Was very well done but could have had more examples than just one. -Too much evidence; it was overwhelming		-Hard to judge; don't have a copy of the reading (H. Cortés) -Hard to tell what really knew; a lot was copied from the text -No evidence for reading -Cassette was marked (*Los Hispanos en USA*) for listening but it was really for speaking. -This student must have done more projects involving art/drawing that would have been better evidence of his abilities/knowledge -Not the best example -Just by underlining hard to know if really reads well (noticias) -Hard to judge this artifact because not part of course work (*Leyenda*) * Aspect I should refer to printed language instead of printed material
Contributions of Hispanic Figures: 2		Expressing Meaning in Oral & Written Form: 2
-No personal connection made (also applies to aspect 3) -Only mentioned one person -2 Non-Hispanics -Not in Pacesetter; outside source. -Seems to be a contradiction at the end -Same information in both artifacts -Does not show contributions -Not a good choice for aspect II (*Popo e Ixtla*) -Barely refers to the pictures -General awareness; *El Greco* -Several students conflate "Hispanic figures" and ancient/pre-Colombian civilizations -Only mentioned one person -Piece is of only 1 Hispanic, not several (*Blasón*)		-Very bad quality of recording -It is not clear if the video included was supposed to accompany this or not. This info would be helpful. -There wasn't a *prueba* in the cassette -Could not find Martin Luther King -Not a reading activity and only had 2 short questions to listen to (*Cinta-Ballena*) -Video missing -Regarding aspect 2: difference between minimal interference (accomplished) isn't clear -No oral; good writing -Not enough evidence to judge speaking -Can't hear tape -Student read his Maya essay and presented it as evidence of speaking; no evidence for listening -Not a good writing sample -I'm finding hamper to be a key word in interpreting aspect 2; first language interference/influence rarely impedes communication, nor even hamper communication, but rather only hamper the ease of communication -An assignment on grammar is not good evidence of writing -Question of multiple evidence
Connections W/Other Disciplines/Cultures: 3		
-Fails to make connection -It is not clear what this is or how it is relevant -Linking to own culture lacking as well as comparing and contrasting -No connection to own culture made -Does not demonstrate knowledge of Hispanic cultures; no personal connection (*La Biografia*) -Interview shows him interviewing applicant -No connection (*Los Hispánicos*) -Not a good example (*Yo*) -Only one sample, not enough evidence -Aspect III has at least two very distinct components; should connection to self /own culture and to other disciplines both be evident? -Unable to understand recording (Tape w/partner) -No mention of personal connection (*Blasón*) -No contribution to Hispanic figures (*Dos leyendas*) -Student is not clear on what connections w/other disciplines mean No personal connections		

Comments/Suggestions Gen.	
-No cover sheet for more than half of the artifacts presented -A lot of the artifacts were irrelevant -Heroes essay is in English -No examples for Aspect -Well organized -In general, this student did not present worksheets, but only cover sheets with reflections -Student should have chosen only the pieces strongest in each aspect -Cover sheet gave wrong dimension -In many instances, the writing on the coversheet in copied from the writing in the artifact at length and then is read on tape -This collection contains a lot. 13 artifacts for Dim 1 to be evaluated in 27 aspects. -Portfolio is enormous; did not use but at most 8 artifacts -Without clear info on coversheet, I would have not known what the artifacts were proving (The cover sheets are very important and should be part of the evaluation) -Student attached several artifacts to one cover sheet * Well organized; We should use this sample to suggest to other teachers and students how to send in their artifacts, cover sheets, etc.. (05-E-07) -This portfolio is clearly organized - Some students have deliberately chosen work from throughout the year in order to demonstrate their progress over time, which isn't treated in the rubrics. As it is weaker works submitted (maybe from early in the year) for an aspect for which stronger works also exist may lead to the impression of limited or uneven overall -Many students either count *Ixtla y Popo* as "figures" or count the Nahna creators of the legend as "figures" for aspect 2	-Shows a very positive attitude toward his learning and performance -No speaking samples -In many of the artifacts included as evidence of reading comp, it was impossible to judge what the information included; the same happened with the history documents -Need more reading sample and also more listening samples -No speaking -Not clear what printed/oral materials meaning was supposedly derived from for 3 autobio-graphical pieces -No cover sheet -Very well rounded portfolio (02-B-13) -Students indicates 3 artifacts but only 2 found. -Not a listening or reading activity -In analyzing the time used it should be considered that this simultaneous processing of the activity, i.e., writing these notes, certainly takes some time. With a reasonable number of submissions comprising a collection and disregarding technical difficulties, it doesn't take that long to read/score -Do not know which artifact to use for speaking; side A had one speaking and side B had the assessment -No speaking or listening tasks were included -Hard to tell if speaking is spoken or read -Maybe there should be a maximum # of submissions It is very difficult to cue up tapes for each recording and identify which recording corresponds to which cover sheet. -Hard to know what to listen to without title or speaking part. -No true writing sample -Needs to be more specific about what learned -Video with presentation was used to give score for speaking -Too many artifacts were submitted; limit entries to five.