A PEER ASSISTED APPROACH TO ASSESSMENT AND EVALUATION IN A KOREAN UNIVERSITY

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE MASTER OF ARTS IN TEACHING DERGREE AT THE SCHOOL FOR INTERNATIONAL TRAINING, BRATTLEBORO, VERMONT

BY

JOHN STEERE WENDEL

APRIL 2009

IPP ADVISOR: KATHLEEN GRAVES

Author’s Signature ________________________

This project by John Wendel is accepted in its present form.

Date: ____________________

Project Advisor: _______________________

Project Reader: ______________________

Acknowledgements: I thank Kathleen Grave's for her clear advice and counsel.
I gratefully acknowledge the A&E committee members with whom I worked during the fall of 2007. I'm so thankful for their time, patience, and honesty. They gave me so much to report about. I would also like to acknowledge Anne Katz's part in helping me see a whole lot in a short time.

Abstract

Author: John Wendel

Title: A Peer Assisted Approach to Assessment and Evaluation in a Korean University

Current Address: Dongguk University
707 Seokjang-Dong Won Hyo Kwan Kyo Yang Kyo Yook Won
Gyeongju, Gyeongbuk, 780-714
Republic of Korea

Phone: (054) 743-8730

Institution: School for International Training

Program: Master of Arts in Teaching Program

Degree: Master of Arts in Teaching (M.A.T.)

SMAT 25

Thesis Advisor: Kathleen Graves

Abstract: This paper details collaborative professional development with a focus on classroom assessment and evaluation of Korean University students. The observations, opinions, and conclusions reported herein are based on committee work carried out by the author at Dongguk University during the fall semester of 2007. Members of the committee were first asked to define areas of assessment and evaluation of students that either interested them, or that they found problematic. Through methods described in this paper (journaling, committee meetings, materials sharing, observation and feedback), each team member had opportunities to actively reflect on their own practices, to both articulate and challenge beliefs about assessment and evaluation, and to use their team members as immediate resources for professional development. The context in which this work was done is explained in detail. Key areas of this reports content include homework, notions of communicative competence, the teacher's impact on students in the classroom testing environment, and developing analytic rubrics. Detailed reports are given about the relationship between the aforementioned methods and content areas that were examined, and how having a framework for working with others greatly aided in reflection, articulation, and action.

Educational Resources Information Center (ERIC) Descriptors

Test Construction

Test Reliability

Teacher Developed Materials

Teacher Improvement

Communicative Competence

Action Research

Assessment

Table of Contents

Chapters

Chapter 1

Introduction

The purpose of this report is to detail collaborative work I did with four of my English conversation colleagues at Dongguk University (Gyeongju City, South Korea), during the fall semester of 2007. Our university supports research and workshop committees to explore issues relevant to our field, or to engage in some form of professional development. We called ourselves the “Assessment and Evaluation Team”, from here referred to as “A and E”. I was the committee chair and we had three other full time members, and one adjunct. I will use pseudonyms and refer to them as Stella, Martha, and Ingrid, and our adjunct Benway. Benway was not able to participate full-time due to a more pressing committee commitment, though he provided vital contributions.

Each of us had grown weary of working either in isolation or in a slipshod fashion with our peers. We desired a more reliable framework, with regular meetings, journaling, and observations of each others’ oral exams followed by feedback sessions— all described in section on methodology— where we could work collaboratively. We believed that through collaborative engagement we could prompt deep reflections about our practices, and facilitate awareness in ways we cannot when working alone. We were interested in adding to each other’s repertoire of assessment techniques, believing ourselves to be essential and valuable resources for one another. Finally, we felt strongly that while we gain a great deal from reading literature about assessment and evaluation, our combined knowledge of our specific context can lead to a more informed and relevant dialog about the subject.

I shall account for two broad threads: my working collaboratively with peers and insights I gained about assessment and evaluation along the way. Before further elaboration on the structure of my report, I find it necessary to explain how these two threads became important to me, and how I came to work with this group of distinct individuals.

The terms "assessment and evaluation" rarely, if ever, entered my mind after I began teaching English in Korea over seven years ago. I don't think the terms came up very much in conversations with colleagues during my first four years of teaching. While working in a language center, I was certainly involved in tasks that could be characterized as forms of assessment or evaluation. However the stakes were never very high and I was not terrifically conscious of issues surrounding student evaluation, or concerned with making it stand at the center of my teaching life.

I came to Dongguk University to teach credit classes without ever having to consider what goes into writing testing specifications, rubric descriptors, or creating tasks to effectively gauge oral proficiency. Meaningful discussions with fellow teachers about these matters were few and far between during my first couple of semesters here. If conversations about basic issues surrounding our trade moved beyond the surface level, colleagues tended to put up a certain amount of resistance.   We didn't follow much of a curriculum, and many considered administrative laissez-faire to be a good thing. "Just do whatever, man," was the mantra of the day.

About the best advice I might have received during my first semester was, "Just give them projects to do. Let them be creative." Beyond this advice to throw activities with hazy learning outcomes at students, the conversation would go no further. As to how to actually sort students according to letter grades, a typical response would be, “Well, you just pretty much know an A grade when you see it.”

The real pangs of frustration I suffered during my first semesters teaching in a university mostly stemmed from both my own ineptness with methods of assessment and means of critically and fairly evaluating student performance, and the reluctance of the majority of my colleagues to discuss these matters except on the most superficial terms.

Initially my explorations into the worlds of assessment and evaluation were, for the most part, personal and based upon what I valued at the time. Eventually the least competent teachers in our department were cut loose, and we received an infusion of new blood, folks who showed more outward concern for these issues. Finally we got a new dean, and a big part of his initial agenda was to find out just how we arrived at the grades we administered at the end of the term.

He noticed that participation was a grading criterion for most teachers, sometimes counting for up to thirty percent or more of students’ final scores. This terribly concerned him, and he asked us all to send him individual reports detailing how we justify participation scores, along with a general grade break down for our freshman conversation classes. Cries of fascism were heard by some who felt his inquiry was an infringement upon our sovereignty. I count this episode as the biggest wake up call of my teaching career to date.

I had to confront the fact that I couldn’t adequately justify where the final numbers were coming from, and that this problem extended well beyond participation scores. I had no systematic way of keeping track of participation, and I had not carefully considered what participation really entails. During the semester, I’d often make marks in my role book next to students' names (HP for high participation, LP for low participation). These marks were based mostly on my gut feeling from my weekly encounters with the students, and they were not systematically applied. By the end of the semester I often simply looked at the picture cards I have students make, and marked down a number between one and thirty. Considering that I had up to two-hundred and fifty students in those early semesters, I was not engaging in an outstanding feat of memory recall. In all too many cases, I was simply throwing out a hunch.

Though my memo to the dean was full of bold assertions about my ability to discriminate between high and low participation, I honestly had no way of making myself accountable for how those embarrassingly arbitrary numbers were thrown down. Students either "really seemed to try" once a week in class, or they didn't. This problem extended beyond arriving at participation grades. If I'd been called upon to offer examples of grading criteria for oral exams, I would have had a difficult time articulating exactly what the numbers ultimately meant. While I had basic criteria—grammar, pronunciation, clarity, etc.—I had only begun to consider what I really expected from student performance. With the possible exception of scoring grammatical accuracy, I was mostly throwing a lot of numbers around. My marks were based more on my feelings about how the students were doing than how students' performance compared against reliable methods of testing coupled with measures of performance.

I don't want to suggest that I was completely incompetent. I was searching. The grades students received weren't totally without justification, they just weren't well thought out. It was only when I was called to make myself accountable for my practices that I truly began to consider my actual role as a Dongguk University English conversation teacher, and the degree to which assessment and evaluation are at the core of being accountable to my students, myself, and our department.

I first had to wake up and acknowledge that giving students final letter grades is not only a core function of what I do, but a necessary function of teaching in this context. I once thought that this role ran counter to the facilitator role, that too much emphasis on grading corrupted a vaguely considered progressive educational agenda. This could be true in certain contexts, but the fact remains that we score on a grading curve. The truth is that once I started to accept my role as a sorter of students, the more conscious I became of the gap between the grades I was marking down and the means I had used to arrive at those grades. I have become more conscious of what I expect students to have relative mastery over, methods of gauging that mastery, and means of evaluating student performance.

My first summer at SIT, with an emphasis on reflective teaching habits, and an online semester with Anne Katz's course in Curriculum Design and Assessment were naturally beneficial to this journey I'd embarked upon. A simple reading she had assigned by D. Blaz (2001) concerning writing rubrics instantly led me to try to better describe what I mean by good and bad when evaluating students' grammar, fluency, or other criteria. Other works like Luoma's *Assessing Speaking* (2004), Cohen's *Assessing Language Ability in the Classroom* (1994), and Brown's *Language Assessment: Principles and Classroom Practices* (2004) helped me to explore testing language learners from planning, to writing, and finally scoring. Just as important, though, I no longer felt that I was working in isolation. Suddenly I was surrounded by work-place colleagues who were more anxious to share materials, and eager to discuss these issues in a proactive manner— discussing what is possible in this context, rather than simply focusing on how our hands are tied by an impossible administration and unmotivated students.

In the semester prior to our work on the A and E committee, I had worked with the same team members—minus our adjunct Benway, and with Martha chairing—focusing on collaborative professional development. We devoted the spring 2007 semester to investigating daily classroom practices. We kept journals, observed each others' classes, gave each other feedback, and held regular meetings. We shifted our focus to assessment and evaluation for fall 2007.

Our work together was essentially a form of action research, but carried out in a rather loose fashion. Michael J. Wallace (1998) sums up what for me is the most attractive feature of engaging in action research in his book *Action Research for Language Teachers.* He ties it to the reflective cycle and differentiates it from other forms research:

> Action research involves the collection and analysis of data related to some aspect of our professional practice. This is done so that we can reflect on what we have discovered and apply it to our professional action. This is where it differs from other more traditional kinds of research, which are much more concerned with what is universally true, or at least generalisable to other contexts. (p. 16-17)

Our team was not interested in discovering anything definitive about methods of assessment and evaluation. The data we used for reflection came in the form of what we observed from others and what others reported based on their observations. Our conclusions tended toward the illuminative or heuristic. Throughout the process I found I was better able to either discover things about my practices I had not realized before, or to challenge, affirm, or reject pre-held notions about my practices. I do not then assume that conclusions I made affecting this particular context will apply to, say, a private ESL program in the United States where learners come from all over and the class sizes are small and intimate.

I mentioned that our research was carried out in a loose fashion, because during the course of the year we did not collectively focus on any one particular issue, nor did we develop formal means for observing specific behaviors during oral exams, like check sheets or tables specifying areas of focus. Rather, we each identified issues or problems that interested us, and used the team as a support group for further exploration of the issues, or to help each other better identify problem areas. Attempts to quantify results were not made. We were five teachers on separate but intertwined journeys. Throughout this report I will describe salient features of my work with the A and E team. For the most part, I will confine my reporting to my own experiences, observations, and illuminations.

Chapter 2 examines the context under which our work took place. Chapter 3 explains our methodology. Many of the conclusions I have drawn from our work might not make sense to language teachers working in a different location under different circumstances, so I find it necessary to give an account of how our department operates. While everyone on the team was free to pursue their own area of concern, actually working on our disparate interests in a meaningful way would have been either frustrating or impossible without a framework for coming together and communicating with one another. Having a sketch of our methods up front also underscores the fact that the illuminations and conclusions I report throughout this paper were borne out of a process for collaborating with colleagues.

The following chapters, 4 through 7, are organized around basic target areas associated with assessment and evaluation. They represent an arc that begins with areas that affect weekly classroom life and notions of forming students' competencies, then move into the actual testing zone where teachers observed each other during student oral exams and engaged in feedback, and I finally explore the tools teachers use to evaluate student performance during oral exams. In each area I identify new awareness I had about my beliefs and practices, problem areas I wanted to work on (often as a result of dialogue and feedback), and how my peers helped facilitate this process. Here is a basic sketch of the latter chapters:

- **Homework as a vital formative tool:** I had grown frustrated with homework because I assigned it more as a means of having yet another criterion on which to evaluate students at the end of a semester. Martha, Stella, and Benway helped me understand how weekly homework assignments can become a vital instrument in forming competencies, and helping students with formative assessment of what they are learning throughout a term.
- **Communicative competence and what we assess:** Each team member felt that the task-based focus of our curriculum was helpful in getting students to speak more in class, but that limitations arose regarding practical learning outcomes. Students could do information gap exercises, but generally had a hard time getting through basic points in every day conversation. My peers helped me expand my ideas of what communicative competence can (or should) mean in this context. I began to use a conversation flow chart to move students from the beginning to the

end of a conversation, and this would impact students' assessment, by adding to the repertoire of speech acts they are expected to perform.

- **The teacher's effect on testers:** Before my teammates came to observe my oral exams, I considered the testing situation to be not only a time to evaluate student performance, but a time to offer instruction. I tended to approach oral exams as an extension of time spent during weekly lessons. The feedback I received about my interactions with students had me reconsider how and why I was interacting with teams of students as they worked through oral tasks, and caused me to become a more calm and distant evaluator during test time.
- **Reflections on rubrics:** I have found that one of the most frustrating tasks I face when it comes to oral exams is justifying the grade I give each student. Actually describing how we differentiate a good performance from a poor performance on any particular criterion, to suit the purpose of gauging students' performance during an exam, takes time and never fully accounts for everything we think we are listening for. Using others' rubrics during their exams and discussing individual struggles with rubrics helped me to better articulate my own beliefs about writing and revising analytic rubrics.

Chapter 2

Context Overview

Class Dynamic

The majority of our classes are required English conversation courses. These are compulsory courses, and almost exclusively attended by freshman. The classes are divided into four levels, from elementary to high-intermediate. During our work together our students were grouped according to their major. Average entrance exam scores within each major determined their English conversation level.

Classes meet only once a week for two hours, sixteen sessions a semester. Classes are often crowded, with up to thirty-two students in a class. For the last three semesters, all of the English teachers have agreed to team seating arrangements, as opposed to having students sit in rows. This seems to maximize participation, and make it easier for teachers to maneuver through the classroom and give feedback to students. Students can more easily engage in pair work, while making eye contact with the partner across from them, or work teams (usually of four).

Our Curriculum

Our curriculum, levels one through three, has been based on the *Fifty-Fifty* series (Wilson and Barnard, third edition, Pearson-Longman, 2007) since spring 2007. *Fifty-Fifty* focuses exclusively on speaking and listening tasks, and our staff mostly likes how it works with larger groups. We do not meet often enough, and have too many students, to have an effective four skills curriculum. Our various investigations into assessment and evaluation focused almost exclusively on these first three levels of required English conversation.

A general syllabus is set by the academic coordinator of the department, me, with general agreement from the teachers as to how much content will be covered per semester. After using *Fifty-Fifty* for a year, we decided to split the book in half for each semester, seven chapters, while requiring teachers to cover a minimum of five. Our Level Two (high elementary) course objectives for semester one are the following:

Students will be able to:

- Discuss personal abilities
- Inquire about and give personal information
- Use dates and times with reference to specific occasions
- Discuss locations and where items belong
- Describe locations and give directions

The syllabus breaks down the component target forms that go with each task. For example, two weeks of discussing personal information are outlined in the syllabus thus:

4	Chapter 2	• **Personal Information** • Listening to people discuss what they do and where they are from • Using the simple present to state facts
5	Chapter 2	• Exchanging personal and private information about other people and each other • Basic wh-questions in the simple present

The books provide structured tasks that require students to use relevant sentence patterns and vocabulary to complete each task. The first two levels offer the most in the way of target grammatical structures (basic tenses, modals, prepositions of place, using ordinal numbers, adverbs of frequency, etc.). The third book, for the most part, assumes that students have those skills, and much of the language focus deals with indirect questions, levels of formality in speech, and uses a lot of fixed expressions for responding to questions (*That'll be great, No problem, Afraid I can't*, *I'd love to,* etc.). While our department is pleased with the series, certain issues arose for our team because of worries that students might not be able to independently utilize the language to go beyond the tasks. This issue will be further explored in the section devoted to beliefs and practices.

I find it important to state here that I strongly believe choosing a mostly task centered curriculum, with a focus on learner-learner interaction, is appropriate for elementary and intermediate learners in Korea who are recently out of high school. Our department had struggled for years to find books that everyone on staff was comfortable with. When a publishing company representative introduced me to the series I was also under the influence of an essay by Haemoon Lee (2003), *Developing the Oral Proficiency of Korean University Students through Communicative Interaction,* which several other teachers had read, and it influenced me to push for the choice.

According to Lee, despite government efforts to encourage more oral proficiency from an early age, students arrive to university from teaching environments where grammar focus is usually devoid of communicative context. Students often suffer a serious imbalance between reading skills, often high, and the other three skills (speaking, listening, and writing), which are usually rather low. Korean teachers often have low communicative skills and rely on audio and visual tapes which have no interactive component (p. 30).

Lee grounds her argument in interaction theory, starting with *focus on form,* "the process by which learners attend to the form of the language within a context of communication and which leads to successful communication" (p. 31). Negotiation rich interaction is considered a vital component to acquisition. According to Lee, "Studies of interaction, in sum, point to the conclusion that in order to acquire a language, learners should be responsible interaction partners of equal status with the native interlocutors, rather than be passively given modified comprehensible input from an instructor" (p. 32). Lee also touches on cognitive theory for its notions of hypothesis testing, and nods toward the importance of communicative interaction (p. 32-33).

Lee stresses the importance of focusing on tasks where interaction is required for completion (information exchange and information gap activities), instead of more open-ended activities (general discussion, opinion exchange, debates, and non-terminal problem solving). Emphasis on learner-learner interaction is also ideal. While there may be some risk of speech fossilization from excess of non-grammatical, or non-target input, a broader variety of interaction is usually witnessed, when peers on equal footing work through tasks that prioritize form-meaning mapping. Feedback from peers has been seen to produce more interactional modifications of target forms, than uneven correctional feedback from teachers (p. 36-37).

Assignments and testing

Assignments, quizzes, and major exams are not standardized. Teachers generally tend to administer oral mid-terms and finals. When there was no curricular oversight whatsoever, teachers felt free to go about these matters in any given direction, with or without any target objectives in mind. Also, it was not uncommon for teachers to assign discrete point fill-in-the-blank and multiple choice tests that did not reflect the purported communicative nature of our classes. This was a tricky business when moving twothousand five-hundred freshmen between interlinking semesters. Now that basic parameters have been established, we can allow for a variety of approaches with some reassurance that students will meet certain basic requirements from semester to semester. While our team explored the same curricular terrain, there was enough variation in our methods, means, and habits to generate a lifetime of reflection, dialogue, and debate.

Chapter 3

Methodology

Journaling

Everyone, except for Benway, kept a journal throughout the semester. I assigned three rounds of journal questions to myself and our regular team members, but I also made entries throughout the semester. They provided both a means of personal reflection of relevant issues, and a springboard for group discussions.

The first journal entries were meant to help each member define an area of focus, and to consider how each member wanted to use the group. The second round of questions involved our thoughts on our first round of quizzes, and inquired into thoughts about ongoing assessment. The third involved thoughts on our midterm exams and being observed. Instead of assigning a final entry, we were each responsible for discussing our role in the committee for our end of term workshop. The journal entries are described in Table 1.

Journal Entries: A&E Academic Committee, Fall 2007
Journal One: 1. Is there any area of assessment that you are particularly interested in exploring? 2. What were some methods of assessing what students have learned, and methods of evaluating performance you used last term? What worked and what didn't? 3. What do you want to gain from working with others this term?

Journal Two: 1. What are your thoughts on your first round of quizzes? Consider the following: Why did you make the choices for this quiz that you did? Were there any problems? Would you make any changes? 2. We all seem fond of *Fifty-Fifty*, but we also seem to all acknowledge its limitations. How are you planning to deal with broader notions of communicative competence on your midterm? 3. What are your thoughts on homework and ongoing assessment? How do you find it helpful?
Journal Three: 1. Discuss some highlights of observing others' exams. Did you get any ideas or gain any insights from any observation, or the following feedback sessions? What interesting issues arose? 2. Discuss some highlights of being observed. Did you get any ideas or gain any insights from the observation, or the following feedback session? What interesting issues arose? 3. Now that we've actually done midterms, how do you feel about your means of assessment and the way you chose to evaluate student performance? What went well? Did you experience any persistent issues? Is there anything you would change?

Table 1: Journal Entry: A&E Academic Committee, Fall 2007

Committee Meetings

We met as a team several times throughout the semester, but not on a set schedule. We first met briefly to discuss what we would be doing throughout the semester, and I assigned the first journal entry. Three meetings were then held to discuss journal entries, and those meetings averaged about two hours in length each. We had two more meetings to prepare for our end of term workshop. We discussed issues we each wanted to report back, we planned our presentation, and rehearsed.

Our meetings gave me an opportunity to offer my own ideas about issues to other people, and helped me explore an array of possibilities and consider new plans of action for various modes of assessment. They were a forum for people to articulate beliefs and often to clarify either what they actually believe, or what they mean. In this way, I found that we had opportunities to broaden our understanding of issues or terminology that buzz around our field. Most profoundly for me, the meetings helped me to identify gaps between my own beliefs and my practices.

All of our meetings were recorded onto mp3 players, except for our workshop rehearsals. Because of our lack of a single group goal, I found recordings an essential part of the process. Through listening to meetings I was better able to capture the emerging themes of our work together. Listening again to what my teammates had to say in our first meeting also gave me a better sense of areas I want to work on. In other words, just listening to other people discussing their own practices helped open up new directions. In the initial meeting, though, it's hard to get a good grasp of everything being said. Listening back has been a vital reflective tool.

Materials Sharing

Throughout the semester, we sent copies of all of our assessment materials to one another. These included quizzes, mid-term and finals specification, and grading rubrics. This practice helped serve the greater goal of broadening my repertoire. I was given multiple opportunities to essentially steal from my colleagues' various bags of tricks. I found that it also helped me to refine my own work by comparing it to others. Finally, we became more familiar with each others' practices. Before going to observe a team member's exam, I had some idea of what to expect in the process.

Observation and Feedback

Each team member and our adjunct signed up to observe at least one of each others' midterm exams. Observations were followed by recorded feedback sessions. I also observed two regular classes, one of Stella's and one of Martha's, during the semester to consider different directions for my quizzing practices, and to gather ideas for using a conversation flow chart as a formative tool for my students. I will discuss the flow chart in detail in the section on broadening my repertoire.

Final exam observations were disappointing, because I wasn't able to see as many exams as I wanted. A coworker, my fellow coordinator, had to leave on an emergency, which made it incredibly difficult to make time for observations, as I was dealing with the job of both personnel and academic coordinators. I was able to observe Ingrid's end of term projects and I took part in grading two sets of Martha's students.

The following charts show whose classes I observed and vice versa.

Midterms I Observed	**My Midterms Observed by Others**
• Martha, Level 3 • Stella, Level 1 and Level 2 • Ingrid, Level 2 • Benway, Level 2	• Level 3, Martha • Level 3 and Level 2, Stella • Level 2, Ingrid • Level 2, Benway
Finals I Observed	**My Finals Observed by Others**
• Ingrid, Level 2 final group project • Martha, two Level 4 classes (Martha delivered the exams and I marked rubrics.)	• Level 2, Stella

Observers were given assessment specifications and grading rubrics, so we came to observations with an awareness of how testing was to take place. We actually attempted to use each test giver's grading tools. Observers took notes according to what they noticed with regard to testing tasks, marking students, student-teacher interaction, or any agenda the test giver may have had. I was trying tasks that were new to me and was looking for some general impressions from observers, and I needed feedback reflecting how efficiently and effectively I carried out tasks.

It is one thing to share materials with coworkers, or to listen to explanations of how tasks work, yet another to actually experience how materials are used and tasks are carried out in real time. There is a certain art in moving from planning and into the actual execution of a task, whether it’s a classroom activity, or a high stakes oral exams. While I was especially interested in receiving feedback from others, I was equally interested in exploring others’ rhythms, pacing, interactions, and materials.

Even if I had problems with others’ techniques, I found that reflecting upon those problems allowed me to better consider my own practices. Sometimes I had problems either with the overall effectiveness of an exam, specific areas the teacher chose to focus on, or we differed on broad notions of communicative competence—are we more concerned that students say a few things, with grammatical accuracy as a primary focus, or do we expect a more elaborate exchange between students, while making more room for other criteria (fluency, clarity, preparation, etc.)? In one case I felt that another teacher’s students seemed consistently confused about the objectives of the tasks they’d come to perform, causing me to better consider my own test specifications and how I prepare students for exams. In another case I felt that grammatical accuracy was too dominant a criterion for determining test scores. Reliability was a serious issue for the teacher, and I realized that I needed to more carefully consider how my point descriptors actually reflect the types of output I expect from students.

Feedback sessions usually occurred directly after an observation. Sessions were always recorded onto mp3 players. Scoring results were compared. Feedback did not follow a strict format. I was influenced by my superficial understanding of Cooperative Development (Edge 2000), but felt that adhering to the principles of this method might be a bit limiting.

The principal aim of Cooperative Development, as laid out by Julian Edge, is to encourage independent self-development. Instead of having an observer give evaluative feedback to the teacher who has been observed, the teachers' roles are divided into Understander and Speaker. In this situation, "…the Understander deliberately sets out to make as much space as possible for the Speaker while, at the same time, actively working to help the Speaker use that space creatively" (p. 4). A major component of aiding the Speaker is in helping them to articulate their thoughts based on both intellectual and experiential knowledge. Edge sees speaking not merely as a medium for showing what we know, but actually for learning and potentially developing bases for an action plan (p. 8).

Respective views are not debated, nor are Understanders expected to offer prescriptions to the Speaker. Edge summarized the non-evaluative method in the following passage:

> One essential attitude for the Understander to have, then, is non-judgmental respect for the Speaker's views. Colleagues have every right to their views on teaching and students: they come out of their own experience and understanding. Development can only take place when Speakers recognize their own real views, and then see something in there which they wish to investigate, or to take further, or to change. Mutual, non-evaluative respect is fundamental to Cooperative Development (p. 5).

There were times when I was looking for rather blunt feedback about the messier areas of my exams. I didn't want to spend a lot of time being asked questions to facilitate my ability to articulate what was happening from moment to moment. This doesn't mean that I was blindly deferring to the better judgment of my peers. I simply found it appropriate to give the green light on frank impressions of what actually took place.

The most important rule to follow was that critical analysis was welcome, but that it needed to be as non-prescriptive as possible. It was important to say what was happening, and what affect this seemed to have in the testing situation, however we tried to avoid prescribing each other lists of do's, don'ts. This could sometimes be a fine line to toe. However, by being as descriptive as possible and asking questions where clarification was needed, feelings were, more often than not, spared.

Having another set of eyes in my classroom, and hearing feedback was instructive to me in several ways. Even when I could identify an area I needed to work on, I was often stumped when it came to the particulars. I find that when I'm too close to a particular process, I have a hard time being a truly discerning critic. I may know that I'm having efficiency issues, but feel somehow personally attached to the various steps I'm taking. Getting others to describe what's going on during my exams helped me actually see where I was being superfluous, confusing, or simply taking unnecessary steps.

I shall now turn attention to target areas I worked on with my team.

Chapter 4

Homework as a Formative Tool

One of my biggest broad-based goals for the semester's work was to broaden my repertoire of assessment techniques. I was especially interested in exploring homework as a means of ongoing, formative assessment. Later I wanted to work on striking a balance in oral exams between overly controlled tasks and more open-ended prepared conversations.

These issues were discussed during our teams first meeting together, and I realized the large degree to which I'd be able to work on each of these goals. Everyone shared their own experiences with various forms of assessment, and even elaborated on things they were planning to try out during the semester. As I listened to my teammates describe various techniques they were using (or were going to use), I was most profoundly stuck by the degree to which I began noticing gaps between my own beliefs and how my practices actually reflect those beliefs. This would be a reoccurring theme all term.

I realized, for instance, that I often spoke of the importance of formative assessment, but that my actions didn't always reflect an adequate follow up on my beliefs. H. Douglas Brown (2004) defines formative assessment as, "…evaluating students in the process of 'forming' their competencies and skills with the goal of helping them to continue that growth process." (p. 6). Arthur Hughes (1989) states that, "Assessment is formative when teachers use it to check on the progress of their students, to see how far they have mastered what they should have learned, and then using this information to modify their future teaching" (p. 5).

Homework was an area where I felt I was falling far behind the curve. I was actually right on the verge of giving up on assigning homework altogether. The problem was in both what I assigned for homework and how I checked it. I thought it was an inherent task of the teacher to faithfully collect assignments and mark them up with red ink. In past semesters I assigned written reports for homework. I found that with two hundred or more students, the task of grading even one page reports to be incredibly time consuming, while the pay-off for the students seemed minimal at best. I finally concluded there was little point in spending hours marking papers only to have the vast majority of students either fold or wad them up, and stick them someplace, never to be looked at again. On the one hand, I was able see how students' skills stood after looking at a writing assignment, but I wasn't utilizing the assignments in a way that helped students notice mistakes and errors, and learn from them. Nor was I able to really notice skills *forming.*

Because of the limited times we meet with students, with up to four weeks devoted to oral exams, engaging in anything resembling process writing would be possible only at great expense to conversation. With a greater focus being placed on speaking and listening in our class, I simply can't justify devoting more time to writing (with the exception of journaling). I finally realized, too, that it simply isn't fair to count as twenty percent of students' grades something they turn in once, and then are evaluated upon. There's no time, or reason, for the students to notice the gaps in their learning and then to apply that knowledge in any meaningful way—two assignments, two grades, little or no follow up. I don't think these assignments had much effect on forming competencies and skills.

I'd tried assigning dialogs for students to read during the spring 2006 semester. There was a comprehension check section on quizzes I gave the week after I assigned them. I thought that by engaging them in reading related to the language they were learning that I might encourage a little extra needed input. Two glaring problems arose here. They weren't assigned regular enough readings, and they weren't really tested on the patterns, or vocabulary, I was trying to help them notice. They had to remember details of the conversation to score high, which was not terrifically challenging. I think the quizzes did nothing of communicative importance that the readings were intended to help students notice. Where I set out to help form competencies, I mostly wound up creating a pretty useless assessment component, multiple choice quizzes, to be graded on.

This brings me to my first A&E team meeting, and the lights that went off in my head as I simply listened to others describe how they were dealing with homework. By this time, I was right on the verge of becoming a homework apostate—no hyperbole intended. I figured that since we only teach our students once a week, and with too much time devoted to oral assessment, I could justify not giving students homework. I further reasoned that students will have to engage in the material enough if they are quizzed semi-regularly, and that their oral tests force them to use the language outside of class if they want a decent grade.

I expressed my concern to the team, and was ready to stand by my recent conviction—based on my bad experiences—that my students and I simply couldn't be bothered by homework. Stella and Martha quickly showed me how homework can be given frequently and checked efficiently. I realized, too, what a mistake I'd have made had I abandoned homework altogether. By giving homework frequently, but weighting it more lightly on the grade scale, students had ample opportunities to assess gaps in their learning throughout the semester. Students' grades were neither unnecessarily inflated, nor harmed by simply doing the assignments.

What Stella, Martha, and Benway had done was to create homework books. The books consist of supplements, often extra grammar and vocabulary supplements, to weekly lessons. Benway added regular journal entries to his books. Credit is given for simply finishing each assignment, and points deducted for incomplete assignments. It is perfectly all right for students to make mistakes, because they are not counted off for them. This creates a win-win situation. Simply doing the assignment has no punitive consequence, but in fact has a positive consequence. Reviewing homework in class allows every student an opportunity to check their understanding of the material. Even students who fail to do the assignment then have an opportunity to learn something. I realized too that students could benefit from checking each others' work before we look at it as a class. This gives them opportunities to negotiate amongst themselves what they think the best answers or responses are. I learned from Benway how I could use journals as a means of generating simple discussions based on various points in our curriculum.

By making homework a smaller percent of final grades, final scores aren't overly-inflated from low-stakes assignments. Due to the nature of the homework, students aren't unnecessarily penalized as they had been before by what amounted to one-off assignments, poorly integrated into an overall assessment scheme. By simply opening myself up to other voices and fresh views, my entire perspective on homework was changed. Furthermore, I realized that this method of homework application jibed with my ideal belief in formative assessment, and I was able to narrow that gap between belief and practice.

Due to a slight misunderstanding, Stella had an extra set of Level 3 homework books. I was teaching one Level 3 course, and simply used her books. This helped me to get used to checking and reviewing homework in class. It also gave me ideas for types of assignments I could hand out regularly. Because we had set an open-door policy with regard to our classrooms, I was welcomed into both Stella and Martha's rooms to observe how they handled checking and reviewing assignments. I simply had no idea of how I was going to do this in an efficient manner, with plenty of precious time left for in-class speaking and listening tasks, or other group assignments. After a couple of observations, I had plenty of ideas, and felt no more cause for concern. I also saw great examples of how homework review can easily tie into the presentation portion of a class, and how homework materials can be utilized for speaking activities.

By the end of the semester I found that I had piled on more homework than I ever had in the past, and the process was both productive and practically painless for all. I was able to shortcut a potentially arduous process of discovery by simply entering when the door was open for me.

Chapter 5

Communicative Competence and what we Assess

Having a forum for articulating and discussing beliefs and practices can have significant consequences on classroom decisions we make that ultimately affect the content of our tests. From our team's first meeting, we almost immediately launched into a discussion about objectives and learning outcomes, which had broad implications for what gets assessed. I will discuss here what our team found to be limitations of the task-focused book, and how our team helped me broaden my notion of communicative competence and expand learning outcomes in my classes, which have had positive consequences in taking formative steps throughout the semester.

In our first team meeting, Ingrid stated her concern that by simply moving through the tasks in the book, students might feel bound by the tasks, and not be able to use the language more freely. We discussed the danger of students learning the language as formulaic chunks that they can use to complete an information gap, but they cannot then apply to novel situations.

In Ingrid's case she generally breezed through tasks, without much focus on form and its relationship to function. The book itself emphasizes the tasks, while sneaking the language focus onto the last page of each chapter. The teacher's book doesn't give adequate tips on helping students make that link— though links are there to be made. This may be based on an assumption that students will intuit grammatical meanings as they work through the tasks. At any rate, using the book can be deceptively simple, and it is easy to put the task in front of the language.

Even when we focus more on form, students aren't necessarily going to transfer their command of the simple present on a gap-fill exercise, to a more spontaneous situation. We all expressed frustration at an identical experience we've had with various students. They may do a wonderful job in class exchanging information about Bob, Carol, Ted, and Alice's daily routines in the book. We may even see beautiful examples of meaning negotiation taking place as they work through the task, and we really can't be more pleased with the learner-learner communicative activity we see in any given class. However we may later see a couple of our better elementary level students next to the coffee machine, or in the stairwell, and try on a little basic conversation with them. The conversation rapidly breaks down after "Hi". Perhaps giggles ensue, and the student buries their face in their hands. Or, maybe we are lucky and elicit a "Fine thanks, and you?" We try our luck with "What's up?" If we are very lucky, we'll hear "fine" again, mixed with a look of bewilderment, but most likely the conversation will come to a screeching halt. No small talk, but perhaps a "bye-bye". Students' anxiety about freely speaking with teachers, among other factors, plays a role in the awkwardness of such exchanges, but we all agreed that this all too familiar situation spoke to certain failures in learning outcomes of our classes.

Martha and Stella took the lead here. While our book does, in fact, help foster a high level of communicative interaction, Martha reckoned that a more expansive notion of communicative competence should be taken to heart. And what our students were lacking was a most basic sense of strategic competency: how to move a conversation from a basic beginning, middle, and to a logical end, with maybe some transitions between. She and Stella were piloting a conversation flow chart: boxes and arrows. Each box represented a functional point in a common conversation. For example: salutation, greetings, enquiry, stating a reason to say goodbye, and saying goodbye. The arrows indicate the flow of conversation between the speakers.

I would begin to use a modified version of their flow chart (described later in this section) and I cannot stress what a difference using this tool has made in getting students to more nimbly move through basic conservations. Items like greetings and farewells are easy to take for granted. Perhaps we assume that because saying hello and goodbye are the first things students learn, we forget that it can be tricky to move from one point to the next in a conversation. Using flow charts has now become a formative tool for my students, something I'd needed in semesters past to help them prepare for oral exams. It gives students a basic map whereby book tasks like discussing abilities or likes and dislikes can easily be inserted as an organic part of a conversation, instead of making tasks conversations in and of themselves. It has also helped students generate more natural sounding output during exams.

In the couple of semesters prior to our teams work together, I'd begun to evaluate students primarily on prepared conversations during mid-terms and finals. Like Ingrid, I wanted students to take the language and bend it a bit for their own purposes. I was also trying to get away from exams where student output is limited to stiff, binary question-response, statement-response format, and grading them mostly on discrete points of grammar (accuracy of production gauged on limited output). I would come to see pros and cons to this format while observing one of Martha's task based exams. I would then come to see a place for the conversation flowchart to reconcile my mixed feelings about both methods of assessment.

For their prepared conversations, students were given two weeks to prepare. Specifications for the exam were clearly laid out according to task, contents, and grammar, along with references to pages in the book they could consult, as in the table below:

English Conversation Level 2 (Spring 2007)
About our Midterm Test

Our midterm test is in two weeks! So, we need to start preparing for it.

The Test

- This is a *speaking* test.
- You will prepare a conversation with a team.
- Each team will have three members.
- You will not use notes.
- You will speak freely for about eight minutes.

Preparation

- Look on the back side of this paper.
- You are going to be a new person for this test.
- Fill in information. It will help you to talk about yourself.
- You can prepare the paper in your free time.

Conversation

- Ask each other questions, answer the questions, and *follow up* answers with more questions and statements.
- Example:
  A: Question
  B: Answer
  C: Follow up question/statement.
  A: Answer/response
  B: Follow up question, or statement
  Etc…

Conversation Contents

- **Greeting:** You will first greet each other. Introduce yourselves. Say “hello” and ask how each other are doing.
- **Personal Information:** Ask questions and discuss personal information (where/from, job, studying, etc.)
- **Family:** Briefly discuss your families (where/from, job, studying, etc.)
- **Abilities:** Talk about things you *can do*, or *know how to do.* You might also discuss a family member’s abilities.
- **Time and Date:** Discuss the time, and talk about important dates.

Grading

You will be graded on these points:

- **Grammar**
- **Fluency**
- **Contents**
- **Vocabulary**
- **Directions**

	You **Name:** **Age:** **Married:**	**Family Member 1** **Name:** **Age:**	**Family Member 2** **Name:** **Age:**
Where/from? Which part?			
Job? Who for?			
School? What studying?			
Children?			
Abilities (can do/know how to do)			
Birthday and important dates			
Hobbies			
Any other interesting information			

Table 2: English Conversation Midterm, Spring 2007

I was generally happy with running exams this way. Students were given some latitude to be creative with the material. If the students are to do well on the exam, they have to practice on their own time, thus out of class involvement with English conversation is a necessary component.

Several problems arise from prepared conversations. There is a tendency to memorize a script, and all too often students sound stilted and monotone. It is also possible for lower level students to lean on higher level students to do all the work. It's difficult to know if the language students have learned is genuine output that they can use for other novel purposes. Furthermore, instead of coaching students through points in a typical conversation, I took it for granted that I didn't need to waste much class time on saying hello and saying goodbye. So, it wasn't uncommon for teams to rather awkwardly launch into the heart of the contents, without paying attention to other constituent parts that flesh out everyday speech. Or, there might have been a little padding missing between the joints, as in the following:

Student A: My boyfriend has straight black hair. He is very handsome.

Student B: Oh, very handsome. OK, goodbye.

Student A: Goodbye.

Before being introduced to, and considering the full potential of, the conversation flow chart it seemed that my choices for exams were limited to either a prepared but memorized conversation, or herding students through some simple tasks, but receiving limited output. I'd observed a couple of Martha's exams, prior to our official work on the A and E team. I wanted to get an idea of how to score more fairly than I felt I'd done before. Martha had pairs of students working through specific tasks. For example, one task was to ask for students to ask the other a favor and for their partner to either accept or refuse to do that favor—things they had practiced in class. Students were given prompt cards describing their roles. They were expected to use indirect questions and stock phrases commonly used when accepting or refusing to do something. I scored a few with Martha then found myself journaling. I then made a table of the relative advantages and disadvantages I saw in each method:

Task Driven

Advantages	Disadvantages
• Puts students on the spot. • Memorization might be less of an issue. • Students have to use patterns and structures spontaneously. • They (mostly) either know it or they don't.	• Tends to focus overtly on accuracy of utterances. • Limited to very specific tasks. • Communicative aspect is sometimes limited to specific situations. • Scoring also rather limited to accuracy of response.

Prepared Conversations

Advantages	Disadvantages
• More opportunity to assess fluency. • Students have opportunities to use language for their own purposes (applicable to their lives). • Allows for some creativity. • Ensures collective participation.	• Memorization is an issue (authenticity of output could be seriously questionable). • Scoring can be problematic (deciding what to check). • Marking might become more subjective. • Defining scoring descriptors can be difficult.

Table 3: Advantages and Disadvantages of Task Driven Versus Prepared Conversations in Exams

I observed one of Stella's and one of Martha's classes to get an idea of how I might incorporate a conversation flow chart into my class. Martha was doing a lesson on describing people. During the last segment of class, she reviewed greetings, and then instructed the class that they would be working through the conversation flow chart. For their "actual conversation" students had to describe their boyfriend/girlfriend, and then discuss family members, and use the question, "What does __________ look like?"

Since our books—and most conversation books I'm familiar with— don't recycle strategic means of working through conversations from beginning to end, I immediately saw great opportunities to plug and play, and build as we go along throughout the semester. I decided to use a more linear format and I wanted to incorporate small talk into the model, seen in the table below:

Conversation Model
1. *Greetings*
A: **B:** **A:** **B:** **Etc…**
2. *Basic Things (Small Talk)*
A: **B:** **A:** **B:** **Etc…**
3. *Big Conversation (Main Topic)*
A: **B:** **A:** **B:** **Etc…**
4. *Reason to Say Goodbye*
A: **B:** **A:** **B:**
5. *Say Goodbye*
A: **B:** **A:** **B:**

Table 4: Conversation Model

I immediately began using the Conversation model with students during class, and have students incorporate exam tasks into its broader framework. If students are practicing asking for and giving directions, they may first have to start with a greeting and make a little small talk before beginning the task, which will fall under the "Big Conversation" heading. Now when it comes time for exams students have had plenty of formative coaching and practice on moving from one conversational point to the next. As an evaluator, it helped me to find testing means that strike a balance between herding students through a series of brief and simple tasks (and what sometimes amounted to substitution drills), and openness of prepared conversations. Instead of having students prepare an eight to ten minute conversation, tasks can be integrated into the broader framework of the conversation model.

Chapter 6

The Teacher's Effect on Testers: Reconsidering My Role During Exams

Simply having another set of eyes observe my exams and honestly report back, was a critical factor in helping me to evaluate my own practices, and to consider possibilities for refining my approach. I became primarily interested in better understanding (seeing) how my actions and those of my other examiners during an exam hugely impact reliability, or how precisely a test measures specific learning outcomes.

Andrew D. Cohen (1994) considers three important factors that affect test reliability: test factors, situational factors, and individual factors. Test factors include the explicitness of instructions, appropriateness of tasks, level of ambiguity of items, and the reliability of ratings. Situational factors might include the actual conditions of the room, and how the examiner presents the materials. Cohen divides individual factors into both transient factors and stable factors. Transient factors could include health and psychological state of the examinee, motivation, relationship with the examiner, etc. Stable factors concern the examinee's actual intelligence, present English speaking ability, and their familiarity with the testing procedures (p. 36-38).

Cohen states, "To improve reliability, the teacher/test constructor would want to have clearer instructions, use more and better items, give the test in a more comfortable and less anxiety-provoking environment, and have the respondents be motivated to complete the tasks involved" (p. 37).

Both feedback and observation of others helped me to reflect on these various factors. I would add teacher intervention (or rater interference) as a critical situational factor impacting reliability. My most profound insights came from my peers' comments regarding my interactions with examinees and how my actions created a testing situation affecting the individuals being tested. Then noticing when and how my peers interacted with students during exams helped me determine when it is appropriate for the rater to step in and when the rater is simply interfering with the task at hand.

Before discussing feedback and observations about exams, I will discuss how I organized midterm exams, with reference to a basic task I had all my Level 2 students do, and where I often found myself susceptible to interfering with students' performance.

Students worked in pairs. They were required to know how to ask for and give directions to places on a map, to exchange basic information about people, to describe how people look and dress, and to work through points in the aforementioned conversation flowchart. I chose to have students work though these tasks with each other, while I made notes and marked them according to criteria and point descriptors on an analytic rubric. Students were given the following instructions two weeks before the exam:

English Conversation Level 2 Midterm Exam (Spring 2008)

The Test

- This is a *speaking* test, and you will work in pairs.
- You must arrive to the exam with your partner.

Conversation Tasks

I. Discussing Abilities and Personal Information (Chapters One and Two):

- I will give you and your partner cards with missing information.
- You will ask and answer questions about a man or woman, a couple, and your partner.
- Practice a lot! Your questions and answers should be grammatically correct and smooth sounding.
- Answer questions with full sentences.
- Practice the speaking task on page 13 **and** the question-answer activity I gave you.

II. Giving Directions and Your Conversation Model ("Sexy Conversation"):

- You and your partner must be prepared to begin and finish a conversation.
- Follow your conversation model*:*

  - Greetings
  - Basic things/Small talk: each others' appearance, weather, etc.
  - Big conversation: **Giving Directions**
  - Give a reason to say goodbye.
  - Say goodbye.

- I will give you and your partner a map
- Ask about locations on the map.
- Give directions to that location.
- Practice asking for directions and giving directions (pages 38 and 40).

Grading

You will be graded on these points:

Individual Score	Team Score
• **Grammar** • **Fluency** • **Listening and Comprehension** • **Vocabulary**	• **Content and Organization** • **Timing** • **Preparation**

Table 5: Spring 2008 English Conversation Midterm, Level 2

Students had practiced the five steps in the conversation model (flowchart) from the second week of the semester. I showed them an example of the character cards I would give them on the test day, and referred them to tasks in the book and handouts to help them practice.

Rater interference was an issue that came up in at least every feedback session where my exams were the subject, and was further discussed in group meetings. Martha was especially articulate on this issue. She was concerned with the role we take when we step in to evaluate. Is it a good idea to be less a teacher at this time than a calm, listening, observing evaluator? For Martha, it was a simple choice between putting on one's "teacher's hat", or taking it off and replacing it with a "tester's hat". Both she and Stella noticed that I was rather whimsically changing those hats, and sending confusing signals to my students as to what they were doing there in the first place.

Most of the students talked about weather when making small talk. Whenever one of the partners used the question "How's the weather today?" I would almost immediately disrupt and ask, "Hey, why are you using that question?" I had mentioned to them during class times to simply comment on the weather (A: *Hey it's really nice, gloomy, hot, cold today*. B: *Yeah, it is*.) Both Martha and Stella noticed that students sometimes grew less comfortable with following up where they had left off, and generally responded with more hesitancy. For Martha, it seemed that by putting on that teacher hat, the students became more confused about whether they were being taught or evaluated.

In subsequent exams I was better able to notice that when I did jump into students' zone of performance, that I compromised both efficiency and rater-reliability. The tasks became a lot more stilted and I was less focused on listening and making notations to complement the criteria on my evaluation rubric. If something needs to be clarified so that students stay on the right track, it may be appropriate to step in. However, I came to realize that how and when I step in greatly affects what follows. Sometimes it's hard to remain patient when evaluating students for hours on end, back to back. I find it’s easy to show aggravation, even when students are going off course just a bit.

For final exams I had students prepare conversations that summarized themes, language, and conversation points we’d practiced throughout the year. Again, they worked in pairs, while I marked a grading rubric. A portion of their conversation had to focus on asking and answering questions about a family member. They had to ask for basic information (age, job title, where they work), what they look like, abilities/things they can do, and inquire about what they do in their free time. During an early exam, I noticed myself getting testy with students because they were bouncing questions and responses back and forth like a tennis ball, sounding something like the following:

“What does your father do?”

“My father is lawyer. What does your father do?”

“My father is a police. How old is the father?”

“Father is fifty. How old is your father?” Etc.

Instead of letting them go right ahead, in the manner they had practiced, I interrupted their flow with questions and feedback. I asked them about making follow up statements (“Oh, really?”, “That’s interesting”, etc.). While I’d mentioned strategies for sounding interested in class, I had not given students materials and coaching on the subject. I then demonstrated how they sounded and looked to me: stiff, robotic, and with eyes darting between the ceiling and floor. It simply seemed “less authentic” to go back and forth like that. I’d broken whatever flow these students had.

Martha’s teacher hat/tester hat categories really began to resonate with me. My interference sent students a signal that something was wrong. Giving on the spot feedback is even challenging in a normal classroom context, when students are engaged in a task and reasonably relaxed. They often react to on the spot criticism as if they are simply being told, “bad English”—and they may go so far as to apologize or knock themselves on the head with their knuckles. During a test, when more is at stake, the anxiety is necessarily higher.

It was, likewise, illuminating to watch others interact with students during exam time. I’m now convinced that less interaction between teacher and students is better. There’s certainly a time and a place to step in, but I generally found that the more calm, or poker-faced, teachers were, the more smoothly the exams went.

In one case I found that Ingrid’s obvious displays of displeasure and frustration caused a student to be totally reluctant to speak up at all. This student came in seemingly prepared, and spoke with a high level of fluency relative to the tasks. But, constant interruptions followed by signs of disapproval by the rater caused the student to become more hesitant and cautious. The student didn’t really understand what she was being asked to do, and was being led to believe that whatever English she produced was the wrong English. Some of this frustration may also have stemmed from a lack of pre-exam scaffolding, and testing specs that came in the form of oral instructions.

Everyone at certain points was willing to offer students a little guidance, but some were more selective than others about when they would step in. Neither Benway nor Martha ever interrupted to correct grammar, or to change the direction of a dialogue. Benway stated that sitting quietly and listening for the discrete points students were hitting (or not) was absolutely essential to rater reliability. If there was a moment of uncertainty, he would offer a point of clarification, and would certainly answer any questions students had, even if the question occurred in the middle of a task.

I had a couple of false-starts during my final exams, like my last example. However, I tried to come in more relaxed and ready to remove myself as much as possible from the students' zone of performance, once the tasks were set up. Tests ran a lot more efficiently. If prompted, I tried to interact in a relaxed, almost burr of a voice. Students were far less apt to get the deer in the headlights look. With my mind more focused on actually rating, I feel that more valid scores were given.

Part of my tendency to step in during oral tests had something to do with the lack of intimate learning opportunities that takes place between teacher and student during regular classes, and a desire to highlight points that may or may not have been effectively covered during class time. The classes are crowded, and while plenty of opportunities arise to help individuals out, more time is devoted during speaking activities to circulating and making sure students are on task. So, I found it hard to resist trying to take advantage of teaching and learning opportunities with only two or three students in the class, and the absence of twenty-five to thirty-five other voices buzzing all around.

Martha's concern about choosing a hat to wear during oral examinations begs the question that testing time and teaching time are essentially distinct. While I think it is important to acknowledge that teaching/learning opportunities can arise during a test, it is still important to consider where the act of teaching takes leave, and the act of testing enters.

In his essay *Testing to Learn: a Personal View of Language Testing,* Brian Tomlinson (2005) argues that "…the main purpose of language testing is to provide opportunities for learning, both for students who are being tested, and for the professionals who are administering the tests" (p. 39). Intuitively I sympathize with Tomlinson's thesis, and was influenced by much of it. It helped me better consider missing links between how students practice English in the classroom and how exams are given. However, I have come to think that Tomlinson's focus on testing as learning can be carried too far in the testing situation. I find it hard to argue with the following point:

> …if a test provides useful experience to the students, then preparation for it will be useful too. On the other hand, if the tests assess knowledge or skills of little relevance to real-world communication (e.g. through questions on the grammar of discrete bits of a text, or on the correct word to fill in a blank in a context free sentence) then the preparation for the test could take up time which would be more usefully spent providing learning opportunities of real world relevance. (p. 44)

In Korea we can easily see the negative effects of preparation for high-stakes proficiency exams like TOEIC, the dearth of communicative learning opportunities that go along with both the administration and preparation for those tests, and the low payoff in terms of actual communicative competence. Dongguk foreign teachers are hired to teach conversation classes. Mimicking standardized testing practices runs counter to the broad goal of improving our students' communicative competence. I was not always conscientious about connecting what students practice in class and what they are tested on. During my first two or three semesters I administered written midterms that tested vocabulary and discrete points of grammar, and didn't resemble speaking tasks performed in class.

Tomlinson argues in favor of learning from preparation for exams. He states, "… the most effective way to prevent assessment tasks from inhibiting the learning process is to make the promotion of learning their primary purpose (whilst making sure they achieve their assessment purposes too).” (p. 42) To Tomlinson, class time and test time should mirror one another:

> …it is possible to replicate classroom task situations in the examination. In that way we can ensure that class preparation for performance examinations is useful, and that if classroom tasks typically replicate features of real-world communication, the washback effect of the examinations will be positive. (p. 43)

Eventually I began using only oral midterms and final exams. Reading Tomlinson caused me to reflect on a discrepancy between relative fluency students showed in class, and opportunities to demonstrate and be assessed on their fluency during exams. Where students were asked to carry out a variety of speaking tasks or elaborate on topic based material in class, my exams generally had teams of two or three students doing substitution drills. For example, the tests might have had me checking that they were using *be* and *do* properly in the simple present, by having them ask each other a couple of questions based on a prompt on a card (often/watch movies), with no regard to broader conversational context.

Tomlinson then goes on to advocate using the testing situation as time for learning. He states that, “…it is perfectly possible for learners to gain new knowledge, and to develop new awareness and skills whilst actually taking a test.” (p. 44) My initial reaction to feedback I received about my interactions with students during test time was that I thought it valid to use oral exams to either offer a little instruction, or make students aware of what they are doing. Tomlinson actually suggests giving feedback and advice during tests (p. 45). I began to see, through other teachers’ descriptions of my exams and observing others, that students can have certain expectations about exam time, and blurring lines between learning, teaching, and exam taking can jar those expectations, and create a frustrating experience for the exam takers.

In the end I think that what is most important, at least in my context, is that testing specifications reflect the areas taught, evaluation reflects the degree to which students have mastered those areas, that students are made aware of what is expected, and grading is made as transparent as possible. Martha championed these very points throughout our work together. I now believe it is a stretch to assert that the lines between teaching, learning, and testing must always converge, and to insist on this stretches the limits of the functions of testing, at least in this context.

In Neus Figueras' counterpoint to Tomlinson's essay, Figueras (2005) sympathizes with the desire to counteract negative washback that tests can have on real life learning. However he argues persuasively for clarifying our purpose in both teaching and testing. And he calls for clarity in terminology, especially with regard to the term test. He points out that for Tomlinson, a test is seen as "sometimes meaning formal examination, sometimes meaning assessment activity, and in any case seeming to cover any activity, instrument, or resource to be used in the classroom to foster learning." (Figueras, 47-48). He goes on to state:

> Tomlinson does not seem to consider that testing and teaching, albeit related, may have very specific functions with very specific quality criteria which are not completely identical. He provides quotes which point at similarities, even parallelisms, between teaching and testing, but these are taken out of context, to suit his own argument, with no effort to analyze implications, and one is left to wonder whether good testing should be the same as good teaching. (48)

It may be that Tomlinson has found ways to organically meld teaching/learning and testing. If I had more time with students, and fewer of them, perhaps I could take more of his suggestions to heart. I feel sure that teachers should give students descriptive feedback about areas where they have done well and where they need work. Feedback can and should allow for learning opportunities, but I've come to realize that in my testing situation it is best placed after speaking tasks are finished. As discussed earlier, intervention during tests compromises my ability to reliably assess performance. I wouldn't have quickly seen this had certain peers not described what was actually taking place in my classroom.

Giving testing teams post-exam feedback is also a tricky area, and I hadn't always taken it for granted that I should discuss exam results with students immediately following test performance. I had previously thought if my rubrics were sufficiently explanatory, I could rustle students out the door, and give them their rubrics the following week. If they were interested they could read for themselves what it said. The problem is that while a good rubric can describe point values reasonably well, they don't recall what transpired during the exam— and only I can decipher my additional notes. I'd actually observed some of Martha's exams the semester before we began exam observations as a committee. She would take a little time to discuss with teams specific areas where that caused them to lose points, but also to tell them if, when, and how they had done well. I noticed that more than the potential for learning opportunities, feedback simply provided closure after the exam. Hearing directly from their teacher just following the test, students rarely seemed bewildered by the testing experience, as was all too often the case when I'd say goodbye without an explanation, usually telling students they would see their scores on the following week. From then on, I decided that I wanted to be more consistent about post-exam feedback simply out of fairness to my students.

During our team's official work together, my delivery of post-exam feedback was also discussed, especially the quantity and focus. Benway and Ingrid inquired about what I was doing. Ingrid wondered about the value of even talking to students after the test, instead of handing them their grades and scooting them out. Her point is not invalid, especially since we deal with a lot of students and efficiency is certainly an issue. She felt, though, that the feedback wasn't really doing the students any good.

Benway's descriptions gave me a little more to reflect upon. He noticed that feedback took at least five minutes per team, and that I had a tendency to launch into little mini-lessons, without always directly pointing students attention to tasks they had performed. He pointed out that students often seemed pretty flustered after the exam, and he wondered if there was only so much they would retain from what he considered rather excessive instruction. Like Tomlinson, I would like students to go away from the testing situation having learned something, or that the test resemble just another phase in their formative development, but I have realized that I must also accept limitations, some of which are out of my control: fatigue, motivation, concentration, basic institutional constraints, etc. However, observing others give intelligible explanations, digestible advice, and how they discriminate between more and less important points, I've worked on making post-exam feedback briefer and hopefully more worthwhile.

Chapter 7

Reflections on Rubrics (An Inexhaustible Journey)

Our work together gave us the opportunity to use each other’s evaluation rubrics while sitting in on exams, compare results, and discuss choices we made both during feedback and in larger meetings. This was one of the most interesting aspects of our cycle together. It helped expand my thinking about the exercise of creating rubrics as a means of measuring oral output and team dynamics, and the vexing act of writing and revising scale descriptors. It also helped me think more deeply about notions of supposed objectivity in scoring.

Martha suggested that we not only use each others’ rubrics while observing each others’ exams, but that we compare results as a means of validation. We were only able to observe one of each other’s exams, and we didn’t use a formal means of tabulating our results. We simply compared scores, which made it onto audio recordings. Except for Ingrid’s exams, our scores showed a tendency to merge, but sometimes not until we had scored a couple rounds of exams. We sometimes needed to adjust to the language on the rubrics. Martha and I had been sharing rubrics and used a lot of similar phrasing, so our scores only showed marginal variation. Benway’s rubrics were hyper-focused on accuracy of production, and our scores never varied.

In the end I don’t believe that simply arriving at similar scores necessarily provided absolute validation, nor do I think we had a large enough pool of evidence.

However, the practice was revealing, and inevitably led to reflective dialogue and insights into each teacher's beliefs about rationalizing test scores.

Rubrics were Stella's major focus throughout the term. She had been shifting gears with her grading procedure, and was ready to start using analytic rubrics that made her feel more unbiased in her grading, and to make her marks "feel more justified". She had used rubrics designed by Benway, which were very logical and mathematical, but felt that they didn't quite suit her needs. She wanted to consider student performance in a more dynamic fashion, and a way of judging performance according to what types of language she expects students to produce.

Stella didn't feel ready to design her own rubric, and stitched together rubrics from two different sources (See Appendix: Table 6). She used an individual scoring rubric that I had designed several semesters ago, and no longer had any use for, and a team scoring rubric of Martha's. Her experience, especially during her first round of exams, was a cautionary tale to all—a fact she gladly shared with everyone at our end of term workshop. Stella's biggest frustration was that she didn't feel that she had ownership over the rubric(s) she was using. She experienced actual panic at first because she found it difficult to figure out how the descriptors actually jibed with the performances she was listening to. And because she wasn't intimate with the phrasing in the descriptors, she had a hard time using them for feedback.

Sari Luoma (2004) points out that,

> …the validity of scores depends equally much on the rating criteria and the relationship between the criteria and the tasks. If the criteria are developed in isolation long after the tasks have been finalized, as they often are, there is a danger of a mismatch, which leads to a loss of information about the quality of the performance (p. 171).

I would argue that there is simply a danger when the criteria are either considered in isolation, or generic criteria are over-applied to any given speaking task. While I tend to agree with Blaz (2001) about the usefulness of generic rubrics that can be applied multiple times (p. 31), it is a valuable exercise for individuals to use rubrics as a reflective exercise to articulate what expected behaviors point values on a scale for various criteria are meant to represent. Simply stating that a score of three represents an excellent job tells us almost nothing about how students are using language or interacting with one another. Stating that grammatical excellence means that students used specific patterns covered in a class, and that any errors were slight and had little or no effect on meaning, gives the rater information to anchor their scores on. If students have this information before the exam, they can have a better idea of what is expected of them.

Stella encountered further problems from my individual score rubric due to my own ill-conceived considerations at the time, and because of its inappropriateness for this particular assessment. First off, each criterion is set on a five point scale. I had since made the switch to using even numbers. I had consistently had problems when a middle category divides high and low, and Stella said that this was consistently a problem for her. We concur with the following assertion by Blaz:

> When using a rating scale, try to use an even number of points… because then there is no middle. Many people are tempted, if the product or behavior is not clearly outstanding, to assign a score in the middle of a range. This is called *central tendency.* If there is no middle score, they must then make a decision whether the product falls to the better-quality or lower-quality side of the center (p. 30).

Another problem was that I had made categories for both *grammar* and *accuracy and use.* Accuracy and use seemed redundant here. I had originally used the rubric for a business English class, where I must have thought that there were certain matters of pragmatic use that needed to be considered. Here it was both useless and confusing.

I don't think it's possible to ever sum up with perfect precision the phenomena we are tying to account for, but having basic descriptions that serve as a guide, and that raters feel some command over, makes more sense than trusting that we simply know the difference between either raw numbers on a scale, or adjectives like good and bad.

I mostly felt at a loss when scoring in Ingrid's class. Her rubrics only listed criteria (grammar, vocabulary, presentation, pronunciation), with no descriptors for each point value, and each criterion ran on a five point scale (See appendix: Table 7). I thought perhaps since we covered much of the same material that we might have similar inner-criteria, but this didn't prove to be the case. I was especially at a loss for scoring pronunciation, because I had no idea what areas of pronunciation I was looking for.

I noticed that our scores didn't match at all even before our post-exam discussion. I never noticed her make any notes. Students finished, numbers were circled, and that was that. Further confusion followed because Ingrid was operating on a sliding scale. Students whom she felt were already of higher ability, but could have tried harder, were simply marked lower than students of lower ability who seemed to make an effort. This was done irrespective of actual expected learning outcomes.

Even though Stella had trouble adjusting to evaluation tools that weren't hers, she was able to adjust. Our scores also tended merge, with mild variation. She was in the very early stages of using the rubrics, and I happened to be more aware of the criteria. Even though the tools weren't ideal, they *did* manage to describe, and I think contrasting Ingrid and Stella's situations speaks volumes for working with descriptors to hang onto point values. Stella later reported that the more familiar she became with the language on her rubric, the more consistent her scoring became. With descriptors in place the scorer could be more likely to match the speaking events taking place in an exam with the language meant to describe expectations of higher and lower performance. I think descriptors impose some discipline on the scorer. In Ingrid's case, it was too easy to simply mark according to how she felt at any given moment, or according to personal feelings about various students. If she felt a student "needed improvement" for any given category, she would mark a student two, without having to pay respect to a more concrete description of what a two was meant to actually signify.

As far as objectivity is concerned, Benway came as close as one can come to being "purely objective". His rubrics are designed for each utterance occurring for each small task students are asked to carry out (See Appendix Table 8). Our test scores matched up almost perfectly. While I liked how each point value specifically accounts for a type of utterance related to the task, found the rubric almost user-proof, and extremely transparent, it also brought up certain concerns. It mostly only works for speech acts following a binary question-response, statement-response format. While I'm a big believer in helping students work on grammatical accuracy in conversation— I spend a lot of feedback time on it—it is but one important part of a much broader dynamic.

However, Benway wondered how analytic rubrics of the type Stella, Martha, and I were working with could account for the speech acts delivered in each task. How do we mathematically account for the student who gives good directions, but doesn't do so well on discussing appearance, if it's not spelled out blow by blow? He has a point. Even though we each took a lot of notes, the rubrics themselves do not account for each utterance the students make.

In my experience, even carefully considered criteria and carefully worded descriptors do not guarantee objectivity, or that grading will be one-hundred percent consistent across the board. However unless we are only going to allow for limited output, I think we have to consider moving beyond only looking at accuracy of production, and notions of fluency that only look at quick responses that contain no self-correction. Fluency and grammar may be related but can be scored as separate issues. Since students are expected to prepare for exams, and hopefully learn from the process, it also makes sense to try to account for the how well students are working on the tasks together.

I originally tried using only individual performance rubrics, but decided to incorporate a team rubric as well (See Appendix: Table 9). Martha had added a team scoring rubric to better account for pair and team dynamics. I find that the addition helps more broadly define quality of interaction between two or more people. Martha had begun her exams the week before I did, so I was able to attempt grading this way before I tried it out on my own students.

Dialoging and observation helped me consider how creating analytic rubrics can be a vital reflective task. It helps teachers explore how scoring reflects expected learning outcomes. It helps us look at evaluation as a dynamic framework, with great potential for providing feedback to learners. It helps us consider how we intend to make ourselves accountable to assigning value to student output, and likewise helps students understand what they are being held accountable for. Finally, though we can't guarantee perfect objectivity, we can ensure our students that we aren't being arbitrary.

Chapter 8

Conclusion

In the end staying committed to a collaborative process aided my exploration of various strands of students' formative development and their relationship to methods of assessment and evaluation. I'll end this paper with a summary of three features that ran through our work together and how they impacted me in terms of awareness, action, and change.

First having a forum for reflection and articulation of beliefs and practices gave me an outlet to give voice to ideas that buzz around unformed in my head. Then by reflecting on my actual practices in journals, meetings, and feedback, I could actually begin to see where gaps existed between what I thought I believed and what I practiced. Notions like communicative competence and formative assessment also came to signify something more concrete when I listened to others discuss either what they meant by these terms or how they were working on these areas. I realized that I needed to do something about how I made homework work for students, instead of giving assignments just so I'd have something to grade. I came to see that our curriculum had its positive points, but then recognized that our notions of communicative competence were inadequate when students who ace our final oral exams can't bear to strike up the simplest conversation in English with their teachers outside of class. Finally, I was given an opportunity to validate my belief that analytic rubrics, created by individual teachers for specific assessments, provide teacher and student alike with a fairer method of evaluating performance than using rubrics with numbers but no descriptors; however, I also had to consider more fully the limitations posed when dealing with several criteria at once.

Secondly, working with resourceful colleagues who were willing to share materials allowed me to more easily broaden my teaching and testing repertoire. My teammates were all too willing to show me new possibilities for homework assignments that help students develop competence in areas like discussing abilities, exchanging information about habits and routines, or asking follow up questions about things that have happened to people in the past.. What now seems rather obvious to me was a bit of a mystery then, and my teammates spared me a great deal of time and head scratching by simply saying, “Here, give this a look.”

Finally, observations and feedback sessions lent me new perspectives that greatly aided in helping me *see* my practices more clearly. I also learned from simply watching others. I cannot overstate the impact of having trusting colleagues honestly describe for me what they observed as I carried out oral exams. My peers helped me reflect upon the role of the rater in a classroom testing environment, resulting in fundamental shifts in how I behave with examinees. Watching others also helped me see what happens in the testing situation in ways that I cannot see when I am stuck in the middle of that situation.

During my journey with the A and E team, two critical forces played in my favor. They were my own naiveté and luck. Naiveté led me to believe that I could just throw myself into this group of people with an inquiry or two about assessment and evaluation, but that my real sense of purpose would reveal itself in the process. I was lucky in that the people I worked with were, for the most part, willing to spend a lot of their leisure time participating in this project, sharing ideas and resources, and that they honestly described for me what they observed about my practices in real time. I initially thought that I would focus mostly on testing students, but during our first team meeting found myself drunk on the possibilities that my colleagues presented for me. I just couldn’t help but dive right in and try to explore several directions at once. I’m not sure that I would actually recommend to people to move into collaborative inquiry in this manner, but in my case it turned out pretty well. It is, though, but one point of departure, and I realize that I have a lifelong of reflection ahead of me.

Appendix

Table 6: Stella's Oral Evaluation Rubric (Individual and Team)

	5	4	3	2	1
grammar	Extensive Proficiency Had little problem with speech patterns. Grammatical problems in no way prevented student from carrying out the task	High Proficiency Had problems with patterns, but student was pretty well understood. Arbitrary use of structures did not significantly affect meaning.	Moderate Proficiency Frequent grammar and word order problems sometimes make student difficult to understand.	Low Proficiency Grammar and sentence pattern problems often make meaning difficult to understand	Poor Proficiency Grammatical and sentence pattern errors were frequent and severe. Problems prevented clear meaning.
vocabulary	Extensive Had an excellent command of words relative to the task. Used words acquired or discussed in class.	Large Occasionally misused words, and needed to rephrase. Showed some competence with newly acquired words.	Moderate Had enough skill with words to perform the task. Frequently searched and/or misused words. Rarely used newly acquired words.	Small Had difficulty making self clear because of vocabulary limitations.	Extremely Limited Lack of vocabulary prohibited the student from carrying out the task.
accuracy and use	Excellent Target structures were used clearly and appropriately. Frequently used structures discussed in class.	Good Had occasional difficulty with word order or using key structures. Demonstrated understanding of target structures discussed in class.	Moderate Speech structures were often used arbitrarily. Poor use of structures affected meaning. Rarely used structures learned in class.	Low Had a very difficult time using patterns and structures clearly and using them with a purpose.	Poor Showed little understanding of the link between form and function.
articulation	Highly Articulate Unbroken speech. Speech was chosen carefully so that meaning was clear. Utterances sounded natural and not memorized	Mostly Articulate Speech was clear, but student frequently searched for words or correct manner of speaking.	Reasonably Articulate Speech was often disrupted due to broken speech and hesitation to find words and phrases.	Semi Articulate Student had a difficult time moving beyond formulaic phrases. Had a very difficult time making speech clear.	Inarticulate Speaking was so halting task could not be carried out.
understanding	Excellent Carried out the task and component parts with clear purpose and understanding	Good Understood the task but had small problems carrying it out.	Moderate Engaged in the role and carried out the task. Showed many signs that they weren't really following directions.	Some Understanding Had an understanding of the role, but had great difficulty working through the task. Did not demonstrate a sense of purpose.	Low Understanding Showed little indication that they could carry out the task.

Speaking exam assessment – Level 2, Semester 1, 2007

Pair work marks

	3	**2**	**1**	**0**	
Content organization	Fully covered Covered every area of specified contents.	Mostly Covered Participated in all but a small area of the specified contents.	Partly Covered Only participated in portions of the specified contents	Barely Covered Only participated in one or two portions of specified contents.	□ Start □ U7 □ U8 □ U11 □ End
Coherence	Excellent Flows naturally. Logical progression of ideas.	Good Some parts stilted, but the majority flow logically.	Moderate Occasionally natural, but largely rehearsed and stiff. Leaps from idea to idea.	Poor Stilted and unprepared.	
Timing and evidence of study	Good Task accomplished efficiently and quickly.	Reasonable Some evidence of lack of preparation but able to complete task.	Inadequate Lots of pauses and hesitations. The use of a lot of memorized chunks without evidence of thought.	Abysmal Student unable to accomplish task. Lots of giggling. Unable to understand or carry out task or its components.	

Individual marks

Student name	**Student number**	**Grammar**	**Vocabulary**	**Accuracy and use**	**Articulation**	**Understanding**	**Total**	**Comments**

Table 7: Ingrid’s Midterm Speaking Exam Rubric

Name:	Student Number:				
Criteria	**Scoring**				
Listening and Comprehension	5	4	3	2	1
Pronunciation and Clarity	5	4	3	2	1
Vocabulary	5	4	3	2	1
Grammar	5	4	3	2	1

Name:	Student Number:				
Criteria	**Scoring**				
Listening and Comprehension	5	4	3	2	1
Pronunciation and Clarity	5	4	3	2	1
Vocabulary	5	4	3	2	1
Grammar	5	4	3	2	1

5 – Excellent, 4- Very Good, 3 – Good, 2 – Needs Improvement, 1 – Fail

Table 8: Benway's Rubric for Assessing Directions

Ch. 7 Giving Directions (3/4 Value)

5	Perfect grammatical structure Fluent response Complete instructions: Exp: **Go down 2 blocks. Turn right. It is on the right, across from the library.**
4	Slow Response Problems with articles, pronouns, "-s". Exp: **Go down 2 block__. Turn right. It is on ___, across from ___ library.**
3.5	Correct, but incomplete directions (3 sentences): Exp: **Go 2 blocks. Turn right. It is on the right, ____________________**
3	Wrong, but grammatically correct directions. Problems with prepositions Exp: **It is ____ right, across ______ library.**
2	Question understood: but partial directions (1 or 2 sentences): Exp: **Go two blocks ____________________** Sentence order problems Exp: **The bank is go 2 blocks.** Verb tense construction problems: Exp: **Turning/Is turn right.** Missing Verb: Exp: ____ **down 2 blocks.** ___ **right.** _____ **on the right.**
1	One word answers Exp: **2 block. Right.** Directions jumbled to the point of unintelligibility. Comprehensible but unrelated utterances.

Figure 9: Level Two Oral Evaluation Rubric for Individual and Team

Individual Score Grading Rubric Level Two

	3	2	1	0
Grammar	Extensive Proficiency Had little problem with speech patterns and structures. Grammatical problems were insignificant and had no effect on meaning. Properly used all target structures specified for tasks.	High Proficiency Had problems with patterns, but student was well understood. Arbitrary use or omission of structures did not significantly affect meaning. Used most of the target structures.	Moderate Proficiency Frequent grammar and word order problems made student difficult to understand. Little attempt to use target structures from our lessons.	Low Proficiency Grammar and sentence pattern problems often made meaning difficult to understand. Obviously hasn't learned basic patterns appropriate to the tasks.
Vocabulary	Large Demonstrated a wide ranging vocabulary base appropriate to task. Rarely misused words, left out words, or searched for words. Attempted to use a variety of vocabulary when possible.	Moderate Occasionally misused words, searched for words, or left out words, but mostly used words appropriately. Some attempt to use new vocabulary.	Small Had enough skill to form basic statements. Little variation. No attempt to use new vocabulary.	Inadequate Had difficulty making self clear because of vocabulary limitations.
Fluency	High Unbroken speech. Utterances sounded smooth and clear. Speech was suitable to each task.	Acceptable Speech was clear, but student frequently searched for correct manner of speaking. Problems barely affected ability to work through each task.	Inadequate Speech was often broken and student hesitated to find words and phrases. May have sounded robotic due to memorization.	Poor Student had a difficult time moving beyond formulaic phrases. Had a very difficult time making speech clear.
Listening and Comprehension	Excellent Speaker obviously understood their partner. Responses logically followed from previous questions or statements.	Adequate Had occasional problem responding to questions or following up on statements. Speech mostly connected to what was said before. May have needed prompting.	Inadequate Often demonstrated a lack of comprehension. Speech often unconnected to partner's questions and statements. Needed prompting on several times.	Poor Constantly unable to perform tasks due to comprehension problems.

Pair Work Grading Rubric Level Two Class Time:______________

	3	2	1	0
Content and Organization	Fully Covered Covered **every** area of specified contents. Conversation model was fully developed.	Mostly Covered Had minor difficultly getting through areas of specified contents. Worked through the conversation model.	Partly Covered Portions of specified contents were covered. Needed too much prompting. Weak model conversation	Barely Covered Could barely work through specified contents.
Timing	Excellent Quickly asked questions and gave responses. Needed little or no prompting.	Good Occasional pauses. May have needed a little prompting.	Inadequate Constant pausing. Needed a lot of prompting	Poor So stilted and disconnected
Preparation	Good Obviously prepared for every task. Worked comfortably and naturally together. Lots of eye contact.	Adequate Some tasks were smoother than others. Some eye contact.	Inadequate Didn't prepare for each task. Didn't make much eye contact.	Poor Lack of preparation prevented students from working through tasks. No eye contact.

Name	**Student Number**	Grammar	Vocabulary	Fluency	Listening and Comp	**Total: Group and individual**	**Comments**

Works Cited

Blaz, D. (2001). Choosing and writing a performance assessment and its rubric. *A collection of performance tasks and rubrics,* 17-33. Larchmont, NY: Eye on Education.

Brown, H. Douglas (2004). *Language assessment: principles and classroom practices.* New York: Longman.

Cohen, Andrew D. (1994). *Assessing language ability in the classroom* (second edition). Boston: Heinle and Heinle.

Edge, Julian (2002). *Continuing cooperative development.* Michigan: University of Michigan Press.

Figueras, Neus (2005). Testing, testing, everywhere, and not a while to think. *ELT Journal Volume 59/1,* 47-54. Oxford University Press.

Hughes, Arthur (2006). *Testing for language teachers* (second edition). Cambridge: Cambridge University Press.

Lee, Haemoon. (2003). Developing the oral proficiency of Korean university students through communicative interaction. In Susan Oak and Verginia Virginia S. Martin (eds), *Teaching English to Koreans*. 29-49. Seoul: Hollym.

Luoma, Sari (2004). *Assessing speaking.* Cambridge: Cambridge University Press.

Tomlinson, Brian (2005). Testing to learn: a personal view of language testing. *ELT Journal Volume 59/1,* 39-46. Oxford University Press.

Wallace, Michael J. (1998). *Action research for language teachers.* Cambridge: Cambridge University Press.

Sources Consulted

Harmer, Jeremy (2001). *The practice of English language teaching* (third edition). Malaysia: Longman.

McTighe, J. and Ferrara, S (1998). Assessment approaches and methods. *Assessing learning in the classroom,* 11-20. Washington DC: NEA.

Hudson, Thom (2005). Trends in assessment scales and criterion-referenced language assessment. *Annual Review of Applied Linguistics 25,* 205-227. Cambridge University Press.

White, Ronald (1988). *The elt curriculum: design, innovation, and management.* Malden, Mass.: Blackwell Publishing