Shiken: JALT Testing & Evaluation SIG Newsletter, 4 (1) April 2000 (p. 9 - 21)

On becoming a testing teacher: Preliminary notes (Part 2)

Greta J. Gorsuch

In Part One of this article, I recounted the procedures I undertook to address my needs preparatory to teaching a second language testing course at the M.A. level at a university in the U.S. As a novice testing teacher, I was mainly concerned with bridging two gaps: that between testing practitioner and testing teacher, and that between writing testing articles and talking a bout testing. I then articulated my experiences as a learner of testing and described the influence those experiences had on the initial design of my testing course in the U.S. This second and last installment will document the formative course evaluation I conducted. I will outline the types of data I gathered, when and how I gathered them, and for what purpose. Finally, I will comment on whether, and how, these different data helped me with short- and long-term teaching and course planning issues.

The Course and the Students

The course was an M.A. level course intended for students of a variety of ages and levels and types of teaching experience. There were ten students, nine of whom were Applied Linguistics majors, and one of whom was an Interdisciplinary Studies major. Two students were public elementary and junior high school teachers of Spanish as a Foreign Language, ESL, and deaf education. Two students were Japanese, teaching Japanese as a Second Language at Texas Tech. One student was Brazilian with EFL teaching experience. The remaining students had one to fifteen years teaching experience in ESL, EFL, Spanish as a Foreign Language, and American Sign Language. The course met 80 minutes twice a week for one semester for a total of 37 hours of instruction. To peruse the syllabus, goals, and objectives of the course, please refer to Part One of this article.

The Course Evaluation Program

I gathered three general types of information: First, information that would help guide my efforts while I was teaching the course (short term formative evaluation); second, information that would suggest content directions I might wish to go in or changes I might wish to make in terms of developing my knowledge and teaching skills for future testing courses (long term formative evaluation); and third, estimating the extent of my students' achievement (summative assessment).

I will focus on short- and long-term formative evaluation only due to space limitations. I planned data collection procedures based on a variety of sources, including past experience, course evaluation research (Griffee, 1999) and a newly published teacher research education book, Doing Teacher Research: From Inquiry to Understanding (Freeman, 1998).

Let us now consider the data collection procedure descriptions, course timing, purposes, selected results, and an evaluative report of each procedure. The procedure descriptions, timing, and purposes were written with the intention of being complete enough that interested readers could judge for themselves the extent to which the procedures were likely to afford valuable data in their own situations. I wrote the evaluative reports with three categories in mind including: ease of applicability of the resulting data, student load in completing the procedure, and trustworthiness of the resulting data. Through the evaluative reports, I sought to provide information on what procedures were worthwhile to use, given the purpose of the course evaluation (guidance on short- and long-term teaching and course planning issues) and realities of the author's situation (i.e., limited time and energy).

"Ease of applicability" referred to how accessible the data was to me. This included considerations of how much time and trouble it took me to gather, process, and interpret the data, and how quickly I was able to apply the data interpretations to my teaching and planning. "Student load" referred to how much time and trouble students experienced in producing the data. Finally, "trustworthiness" referred to how strong, moderate, or weak the data was in light of the assumptions of general research methodology. According to Miles and Huberman (1994, pp. 267-268), data should be evaluated in terms of its objectivity, reliability, credibility, and generalizability. In other words, how good was the data and the procedures used to gather it? Can conclusions be drawn from the data with confidence? Why? Why not?

Eight Formative Course Evaluation Procedures

(1) *Index card questionnaire*

Once during the first third of the course, students were asked to anonymously rate their level of agreement to seven statements. The statements were read aloud twice to the students, and students circled numbers from 1 (strongly disagree) to 5 (strongly agree) on an index card. Students were also asked to provide written comments and suggestions. The purpose of this short-term formative evaluation procedure was to get feedback from the students on the mode of teaching (lecture plus small group and pair work), the success of a class on spreadsheet use taught in the computer lab, and student comprehension of testing concept explanations given by me. The results are in Table 1.

Table 1. *Results of an index card questionnaire*. (N = 10, all students in the class).

Statement	Mean	S.D.	Min/Max
The concepts in class are clearly explained.	4.7	.44	4/5
My questions are being sufficiently answered.	4.9	.27	4/5
The lectures help me.	4.8	.38	4/5
I want to do other activities besides lecture.	3.7	1.13	1/5
I want to learn more in class using spreadsheets and statistics programs, etc.	4.0	.85	2/5
The spreadsheet session in the computer lab helped me.	4.1	1.16	1/5
The pace of the class is (3 = about right)	3.1	.32	3/4

Students' written comments were mixed. Two out of ten students wrote no comments. Two other students commented that they thought that class was going well. Two students made comments that seemed unconnected to the course (e.g. "I worry too much."). Two students commented on the mode of the class. One student liked the computer spreadsheet class, but wanted to work on that aspect of the course individually. The other student urged that more group work be done. Finally, two students said they wanted to write or see examples of real test items (e.g. "Is it possible sometime to make some questions for an exam?" or "I'd like to see more examples of real items used on tests.")

According to Table 1, students seemed to feel that the concept explanations were clear and that their questions were being adequately answered. They also felt that the lectures helped them. Most wanted to continue the lectures, although one student wanted to also do non-lecture activities. This student later volunteered that she preferred to process the same in formation recursively, and that she felt shy asking questions in front of the whole class. Students agreed less forcefully that they wanted to study using computer programs and that the spreadsheet class had helped them. In response to this, I conducted more group activities in class, and scrapped plans for more classes devoted to computer programs. Instead, in-class time was used for demonstrations of computer software using a computer projector, responding only to specific questions of the students. Finally, more class time was devoted to looking at actual tests and test items along with item analysis results to hone students' experience working with parallel items and data. I also made the decision to ask students to actually write and pilot test items as part of their student presentation at the end of the semester. I had originally planned that students could create their own test if they wanted to, but that I would accept a description of an existing test with an analysis of how well the test seemed to work with a given group of test takers.

"The Index Card Questionnaire was an effective tool in the course evaluation, but would have been more so had it been administered on a regular basis. "

The Index Card Questionnaire data had a high ease of applicability, in that it took little effort to construct, administer, interpret, and apply to the questions for which the questionnaire was developed. Gathering this data imposed low load on the students. All they needed to do was circle some numbers and write some additional comments. Finally, the data had moderate trustworthiness. On the positive side, data was gathered from all students in relative privacy (guaranteed by anonymity). The questionnaire items were focused, yet still allowed students to make additional comments. The data, recorded on index cards, were easy to present to another researcher for confirmation of my interpretations. On the negative side, the procedure was done only once. Later administrations of the same or similar questions may have been revealing, particularly when the course moved onto the complex concepts of correlation, statistical probability, shared variance, reliability, and validity. The Index Card Questionnaire was an effective tool in the course evaluation, but would have been more so had it been administered on a regular basis.

(2) *Student interviews*

Seven out of ten students were asked to participate in a 10-minute interview outside of class during the last one third of the course. All of the students had indicated by ballot they were willing to participate in an interview, but not all the students could be interviewed before the end of the semester. The interviews took place in the teacher's office, and were tape-recorded. Due to limited time, the teacher took notes after the interviews and did not transcribe the tapes until after the end of the semester. The purpose of the procedure was to get student feedback on the final one third of the textbook. Students had reported that the concepts in this section of the book, including correlation, statistical probability, shared variance, reliability, and validity, were very difficult for them. This data was thought to be helpful in planning the coverage of these concepts in the final exam, and for review lesson planning for the concepts. Thus, this procedure was a short-term formative evaluation procedure.

After being told they could withdraw from the interview at any time without penalty, and that their comments had no relationship with their course grade, the students were asked the following questions:

1. How is the course going for you?
2. Are there any concepts with which you are having particular trouble?
3. How can I help you with that?
4. How do you feel about the mode of the course?
5. Can you suggest any changes?

Students varied in the length of their answers. According to the teacher's interview notes, several students reiterated their wish to continue working with actual data alongside test items, and to speculate why some items "worked" and some did not. All students reported that the textbook seemed to get suddenly very difficult as it moved into correlation and related testing concepts. At the same time, students reported that repeated lectures with recursive explanations of testing concept s helped them understand the concepts gradually. Students reported being particularly troubled by the variety of reliability and dependability formulae being presented them in the textbook. They were not sure in what situations they should use one formula, and not another.

There was a sharp division between students as to their preferences concerning whole class, small group, and pair work activities. Two students stated they liked small group and pair work activities, while the remaining students claimed to prefer whole class instruction with extended question and answer sessions. One student speculated this was the case as the time for the final exam moved closer. I also speculated that students were at a point in their learning where they sensed gaps in their knowledge. In any event, students claimed in the interviews that they wanted answers to specific questions t hey had formed, and that whole class instruction with question and answer sessions afforded them that. Several students said they liked the feeling of class community and felt comfortable working with the whole class.

The result of this data was that I extended question and answer periods while downplaying small group activities. I decided not to test students on the different reliability formulae themselves but rather to give students opportunities in the final examination to comment on the importance of aiming for reliability in tests and to make specific recommendations for increasing reliability. I believed that matching reliability formulae to different tests and testing situations was a skill that needed time and actual testing experience to develop. I also anticipated that students would work with reliability while piloting and then presenting their own tests at the end of the course.

The student interviews had a low to moderate ease of applicability. On one hand, the interview questions were easily created and the interviews themselves took only ten minutes. On the other hand, I found that it was too time consuming to transcribe the interviews. I had to rely on data expressed as impressionistic notes. The notes revealed a wide variety of issues that were initially difficult to interpret and apply to my planning and teaching. The procedure imposed a low to moderate load on students. While students were inconvenienced by having to meet with me outside of class, they only had to answer the questions as they saw fit. Unfortunately, the procedure had only weak to moderate trustworthiness. On the negative side, only seven of ten students were interviewed, and they were interviewed only once. The interviews we re not transcribed, although they were available for replaying on a tape. However, this potentially time consuming task made it difficult to ask an independent researcher to confirm my interpretations. On the positive side, the questions were focused, with provision for students to comment freely on other topics if they wished to. The interviews were conducted in privacy, although they were not anonymous. The student interviews were moderately worthwhile, although difficult to do. The procedure could be improved by developing a written method of noting or categorizing students' responses that would adequately capture their thoughts and still make the data easily accessible for planning purposes.

(3) *Informal teacher questions*

During breaks in lectures or small group working every class meeting, I would ask students to restate the concepts I had just explained in their own words. I would either confirm or would add to their articulations (no student was ever completely wrong). At first, only a few students would answer but gradually, every student began to answer, sometimes dually (one student would begin an answer and the other would complete it) . The purpose of this procedure was to check students' comprehension of testing concepts and to give students opportunities to verbalize the concepts. I could tell which concepts needed more explanation or exemplification from students' responses. Therefore, this was a short-term formative evaluation procedure. Student's seemed to have the most difficulty articulating the concepts of correlation, construct validity, and internal consistency reliability, and the relationship between the three.

The informal teacher questions had a high degree of ease of applicability. I was able to immediately use the data and apply it to my teaching plans, either immediately or in the next class meeting (I would note which concepts students had trouble restating in my personal log - see below). The procedure implied a low to moderate load on students. At first, students seemed to struggle with two issues: (1) saving face while struggling to answer in front of the class; and (2) marshaling their ability to articulate the concepts with what limited vocabulary and experience they had. Later, students seemed to worry about the first issue less. As they learned testing vocabulary and gained experience with the concepts, they seemed to struggle less with the procedure. The trustworthiness of the procedure was weak, however, because my restatement requests and students' responses were not transcribed or written down in any systematic fashion, thus making validation of my interpretations difficult. While student restatements were likely captured on the audiotape recordings of the classes (see below), it would take too much time to locate these brief instances on the tape for corroboration. Also, no records were kept concerning which students responded. It would be difficult to check my anecdotal observations that every student tried to restate concepts at least once. On the positive side, this procedure was done at least once for every class meeting, thus providing a steady, longitudinal flow of information. This procedure was extremely useful, but could be made methodologically stronger by consistently noting instances and content of student restatements.

(4) *Five homework assignments*

This procedure was based on application exercises and review questions in the textbook (Brown, 1996) on pages 89 (NRT and CRT item analyses calculations), 147-149 (calculating and converting standardized scores, interpreting descriptive statistics and histograms), 182-183 (calculating Pearson Product, Spearman rho, and point biserial correlations), 2 27-230 (calculating Cronbach's alpha, K-R20, standard error of measurement, agreement coefficient, kappa coefficient, phi lambda, phi dependability), and some additional CRT item analyses (item facility, B-index, difference index) work based on a dataset I had studied in a research project. The homework was done by the students during the first two thirds of the course. Students had at least five days to complete each homework assignment. Answers were checked in class and then examined by me privately. In the rare event that students got any items wrong, they were given a second chance to complete the homework after a conference with me. This happened five times, with three different students. Some errors seemed to come from students' inexperience with spreadsheets. When asked to explain an item analysis formula, for example, they could do so but then seemed to have trouble getting the formula expressed correctly on the spreadsheet.

The results of this data gathering procedure caused me to schedule additional in-class spreadsheet and statistical program demonstrations using a computer projector, and to invite students to come to my office for tutoring. I also urged students to work with their classmates on the homework, which two students indicated they did. Students had such trouble with the NRT reliability and CRT dependability formulae homework in the textbook that I spent an additional two class meetings talking about them and demonstrating the calculations. Students told me the sheer number of formulae presented in the textbook was simply too much for them to process conceptually and learn to calculate at the same time. Thus, this formative evaluation procedure afforded me input for both short- and long-term course planning. On the long-term, I plan to cover reliability and dependability more slowly, using additional data sets and testing scenarios that I will either gather from other sources, or create myself.

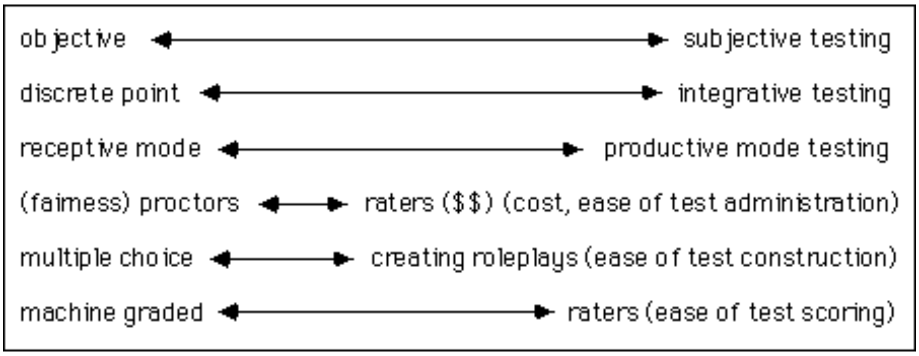
The five homework assignments had high ease of applicability. I could see immediately from students' completed assignments, never more than two pages long, what concepts and procedures they were having trouble with, and which I needed to review in class. Student load was high for this procedure. Students remarked that the assignments increased in complexity over time. One student told me she spent two to three hours on one assignment. This procedure created data with strong trustworthiness. The data was collected over the greater portion of the course, showing students' abilities and development over a range of tasks. Also, the data was rendered into a form that could be easily reviewed by a colleague to confirm my conclusions about students' achievement. Finally, the procedure consisted of tasks which I believed covered a full range of the statistical concepts and procedures presented in the textbook. The homework assignments were an invaluable part of the course evaluation.

(5) *Teacher's personal log and class planner*

This data was generated before and after each class meeting in a spiral bound notebook. On the left hand side of the page, I would write my goals for, concerns about, and plans for the next class meeting. On the opposite side, I would write my reactions and thoughts about the class after it had met. At least half of each page was filled with handwritten notes. Using the log/class planner helped me in developing review strategies for difficult concepts, and to identify areas of my teaching I thought I needed improvement:

September 7 - Overloaded Ss with trade offs presented in the textbook. Too many things for Ss to absorb.
Saw many frowns of incomprehension. A visual would help?

September 9 - Planning lessons in this notebook helps me focus on which concepts I need to cover and how.
I organized trade offs on testing issues into a visual:



objective ↔ subjective testing

discrete point ↔ integrative testing

receptive mode ↔ productive mode testing

(fairness) proctors ↔ raters ($$) (cost, ease of test administration)

multiple choice ↔ creating roleplays (ease of test construction)

machine graded ↔ raters (ease of test scoring)

I will review these and then have students relate their testing presentation proposals to these issues.
Make sure Ss have plenty of time to do small group with this!

Concerns:

1. keep answers to questions short and simple
2. allow other Ss to respond to questions
3. reiterate my office hours for Ss who want help

September 10 - Class went well. The visual helped, Ss were nodding in understanding. Group work seemed to function well. My answers to questions were still too long. But, I did have two Japanese classmates explain TOEFL to a woman who didn't know what it was.

I found that having the class plan in front of me during class meetings really focused my thoughts. Whenever I had specific concerns or goals, I would put a check next to them after I had addressed them. The log provided rich data that profoundly changed my plans for class meetings.

Another example:

> **November 14 -** The double presentation sessions really eat into class time. The presentations have been OK, but only OK. Ss can make good pilot tests but cannot clearly explain their constructs, nor can they adequately describe their students. In other words, they have not yet become full participants in the conversation of the testing profession yet. However, they're doing OK with reliability calculation, etc.

I spent the following class meeting reviewing a presentation planning sheet and giving examples of what would constitute inadequate and adequate descriptions of students and constructs. Some of the students seemed to do better after that.

"the log was an invaluable source of information for short-term formative evaluation."

The personal log and class planner had very high ease of accessibility. Data gathered in this formative evaluation procedure was immediately available for use in course planning in the short term. There was no student load. The trustworthiness of this data was moderate. On the positive side, the data was gathered continuously throughout the course. Further, it was recorded into a notebook which could be shown to a colleague or other researcher for verification of my interpretation of the data. Finally, the data could be corroborated with the tape recordings of the classes. On the negative side, the log took no other format other than to write pre-class concerns and plans on one side of the notebook and impressions of the class afterwards on the other side. This meant that no pre-selected themes were followed and that the data could appear to be weakly organized. Despite this shortcoming, the log was an invaluable source of information for short-term formative evaluation. Themes appearing in the log could be used to generate themes to focus on for the next testing course; for example, my seemingly daily concern with overloading students by offering complex answers to their questions. Thus, the log also had potential long-term formative evaluation value.

(6) *Two pre-presentation assignments*

Students completed two homework assignments describing their presentation proposal to develop, pilot, and revise a test. The first assignment was completed by students on September 27, and the second, a revision, on October 11. The purpose of the assignments was (1) to give students opportunities to articulate in writing their plans, and (2) to allow students to get focused, non-stressful feedback from me entirely in the context of their proposals. A third, unstated purpose was to gauge students' comprehension of testing issues in application to real testing situations. The two assignments, even though collected only two weeks apart, straddled the unit in the textbook covering reliability. Thus, the later homework assignment demonstrated students' developing awareness of reliability issues. A few students were even able to articulate specific strategies for increasing the reliability of their tests.

Students were able to demonstrate a general grasp of the notion that tests need to be piloted and then revised after analysis. Students seemed very clear on the notion of calculating item analyses such as item facility and the B-index. However, no student mentioned item quality analysis. In a possibly related omission, students did not discuss what constructs they wished to capture in their tests, nor how their proposed items related to effectively capturing those constructs. In response to this interpretation of the data, I arranged for a guest lecturer to conduct a workshop on defining and operationalizing constructs. The lecturer used an academic listening skills CRT he had developed for a course in Japan to illustrate his conceptualization and operationalization of the construct of course specific academic listening skills.

From this workshop, it became apparent that while students were clearer on how to articulate a construct, they were having trouble with matching item types to those constructs. In other words, they were unable to work through what their students had to do cognitively to answer their test items and then match this with the construct. For example, one woman, of veteran a public elementary school deaf and ESL education, created long worksheets with discrete point short answer questions in order to test her students' comprehension of a story told orally and via American Sign Language. It was clear that she had come from a tradition of assigning worksheets to students, but she was unable to articulate why or how she thought these limited production items allowed her students to adequately demonstrate comprehension of the story. After much thought, I decided that exploring the cognitive bases of test items and then matching them to constructs was a complex issue and probably beyond the parameters for the current course. These were issues that I would have to explore more thoroughly and then develop a series of materials and lectures for a future testing course. Thus, this procedure served both short- and long-term formative evaluation needs.

The pre-presentation assignments had moderate to high ease of applicability. Students created two pieces of data each, rendered into paper and ink form. Generally, data like this is easy to interpret and use in evaluation. However, the assignments were 2-3 pages each, and completed in varying formats and styles. This created a lot of data in different forms to interpret. The procedure incurred considerable student load. Two students stated the assignments had been cognitively and emotionally difficult. One student complained that she did not want to commit herself to paper, feeling she was not up to the task of writing a test, something she thought only "experts" did. The procedure generated data of moderate to strong trustworthiness. On one hand, the procedure was focused on a single task, and was collected twice, thus all owing students to demonstrate their growing abilities over time and also reveal recurring, persistent problems they were having. The data was paper and ink, which could be easily shown to a colleague for confirmation of my interpretations. Finally, all students completed this procedure, indicating adequate sampling of the class. On the other hand, the procedure was used only twice. A longitudinal procedure with three or four samplings may have afforded more information. However, the procedure provided information on what I thought was a critical aspect of the course, practical application and articulation of testing concepts. With modification, it could be an important source of data for course evaluation.

(7) *Class tape recordings*

Every class meeting was tape-recorded using a hand held tape recorder placed on the teacher's table at the front of the classroom. The purpose was to get feedback on the coherence of my explanations of testing concepts and responses to students' questions. While this procedure had the potential of informing course planning in the short-term, it was intended as a long-term formative evaluation procedure to help determine which teaching skills I needed to develop for future courses. In this case, "teaching skills" means the ability to articulate clearly testing concepts and answer questions adequately (recall from Part One of this article my concern about this). Unfortunately, this procedure had very low ease of applicability. Each class generated an 80-minute tape, which, while potentially filled with rich data, was very difficult to interpret. There was simply so much there, even given my narrow focus of interest. After the first few very time consuming sessions of listening to the tapes, I gave up. It was easier finding out from students via other means (e.g. index card questionnaire) whether my explanations and answers were getting through. The procedure implied no load on students. The trustworthiness of the data generated by the procedure could not be evaluated for reasons mentioned above. This procedure is likely most valuable in situations where recording equipment is better, and there are focused, highly delineated plans for the data, e.g. using the tape to systematically corroborate other data commenting on the teacher's performance.

(8) *Goal coverage questionnaire*

This final procedure was conducted at the end of the course in order to get feedback on the whether the state goals of the course were adequately covered. Thus, this procedure was used as long-term formative evaluation for the purpose of designing future testing courses. The questionnaire had 25 items. 23 of the items invited students to indicate their level of agreement on a 5 point Likert scale to a series of statements (1 = strongly disagree, 5 = strongly agree). The results of the questionnaire are in Table 2 below.

Table 2. *Goal coverage questionnaire results* (N = 10, all students in the class).

Item	Mean	S.D.	Min/Max
We learned the following concepts:			
criterion-referenced test	4.7	.483	4/5
norm-referenced test	4.9	.316	4/5
item statistic	4.5	.707	3/5
data type	4.4	.699	3/5
mean	4.9	.316	4/5
standard deviation	4.8	.422	4/5
performance testing	4.1	.568	3/5
"alternative assessment"	4.3	.675	3/5
test reliability	4.8	.422	4/5
test validity	4.6	.516	4/5
We learned to do the following things:			
how to work with spreadsheet datasets	4.4	.966	2/5
how to work with pencil + calculator datasets	4.8	.422	4/5
how to assess tests	4.5	.527	4/5
procedures used to administer tests	4.6	.516	4/5
how to improve an existing test	4.3	.675	3/5
how to improve an existing test procedure	4.5	.527	4/5
how to create your own test	4.7	.675	3/5
how to create your own testing procedure	4.7	.675	3/5
how to identify helpful sources for testing	4.3	.675	3/5
how to identify helpful sources in teaching journals	4.5	.707	3/5
how to identify helpful sources in research journals	4.5	.707	3/5
how to identify helpful sources on-line	4.3	.675	3/5

1 = strongly disagree, 2 = disagree, 3 = I don't know, 4 = agree, 5 = strongly agree.

Students responded with agreement to all items. They apparently felt that the stated goals of the course had been met. However, the lower mean score items (below 4.5) indicated content areas that perhaps needed more, or more recursive coverage, including: data type (continuous, ordinal, nominal) performance testing, "alternative assessment," working with computer spreadsheets, improving existing tests, identifying other testing resources, and identifying on-line resources. It is also possible to interpret these lower average items as having actually been covered in class, but they were simply harder for students to grasp. In either case, these content areas likely deserve more attention in the next testing course I will teach.

This procedure had a high ease of accessibility. Because the data were rendered into descriptive statistics matched with testing concepts and skills, it was easy to see which areas students thought had been more, or less, covered. Student load was low. They simply had to read the items, and circle their responses. Trustworthiness of the data was moderate. On the positive side, the questionnaire used to gather the data was highly focused, and had very good content validity. Each item was carefully matched to the stated goals and objectives of the course. All ten students completed the questionnaire, ensuring complete sampling of the group. On the negative side, the questionnaire was not validated. It had not been piloted, nor revised in light of the pilot data. Finally, the questionnaire was administered only once, at the end of the course. It could be argued that asking students to accurately remember items touching on the early stages of the course was unreasonable. Perhaps the questionnaire could be broken up and administered throughout the course. With modifications, this procedure could be a valuable source of long-term formative assessment.

Conclusion

Taken together, the eight forms of short- and long-term formative evaluation were effective in providing information to me to deal with immediate and future teaching issues and course planning issues. Improvements in accessibility and trustworthiness of the data can be made in several of the procedures, particularly the index card questionnaires, student interviews, and class t ape recordings. Although space did not permit to report data from the summative evaluation procedures for this course (student presentations, student roleplay, final exam), the formative and summative evaluation data taken together indicated that many, although not all, of my goals were met. Students did learn the content I set out to cover, and students did get extensive hands-on experience working with data and test development. Students did experience their classmates as a warm, friendly learning community, and students had opportunities to articulate and apply their knowledge. Several students indicated in their class work that they could be testing teachers to colleagues and administrators. However, students were generally unable to clearly articulate which constructs they wanted to capture in their tests, and how their test items captured those constructs. Only one student, a Japanese woman, indicated an awareness that the test items she developed for her test were not tapping into the construct she had intended. Through a dictation test, she hoped to capture student comprehension of the meaning of the sentences the students transcribed:

> From the results of this piloting, the instructor noticed that even if the subjects can transcribe correctly, they do not always understand the meaning of a sentence. On the other hand, some of them [subjects] understood the meaning correctly even if they could not fill in the blanks. The complex issue of matching items and constructs is only nominally covered in the textbook. Clearly, I will need to develop a unit from other sources on this important issue.

I would like to thank the students in my testing course for their cooperation and inspiration to me. I would also like to thank all my testing teachers for helping me get to where I am now.

References

Brown, J.D. (1996). *Testing in language programs*. Upper Saddle River, NJ: Prentice-Hall Regents.

Freeman, D. (1998). *Doing teacher research: From inquiry to understanding*. Pacific Grove, CA: Heinle & Heinle Publishers.

Griffee, D.T. (1999). *Course evaluation by quantitative and qualitative methods*. Unpublished doctoral dissertation, Temple University, Tokyo, Japan.

Miles, M.B. & Huberman, A.M. (1994). *Qualitative data analysis*. Thousand Oaks, CA: Sage Publications.