Shiken: JALT Testing & Evaluation SIG Newsletter, 3 (2) Jan 2000 (p. 8 - 17)

On becoming a testing teacher: Preliminary notes (Part 1)

Greta J. Gorsuch

This article describes the self-evaluation of a novice-testing teacher in a second language testing course at a U.S. university. Many readers of this newsletter may wonder what an article on evaluation of a testing teacher in the U.S. has to do with them. If you are like me when I taught at Japanese junior colleges and universities, many of you are probably relegated to teaching undergraduate general oral English courses, and may feel you never have to think about testing outside the bounds of testing your own students or writing entrance exams. What call would you have for teaching second language testing? To this, I would argue that many of the students in your general English courses are likely enrolled in junior and senior high school English teacher certification programs (*kyoushoku menkyou*). They have a good deal to learn from the examples you introduce in testing students in your own classes. What they especially need is for you to make your implicit beliefs about testing explicit in terms they can understand.

In my last two years in Japan, I found myself explaining to students how and why I would be testing them. I began to use testing field specific terminology in my explanations and figured out simple, recursive ways of explaining key concepts and terms. Soon teaching certificate students were knocking on my door asking how to write tests for their upcoming teaching practica.

"as social and demographic pressures push secondary and tertiary EFL teachers in Japan towards more diverse ways of teaching, they will also be forced to learn more varied ways of testing"

Testing classes and test writing components of teacher preparation courses are becoming increasingly important features of both graduate and undergraduate schools in Japan. One piece of evidence of this increased interest in testing in EFL is a translation of J.D. Brown's Testing in Second Language Programs (1996) by Minoru Wada. Further, as social and demographic pressures push secondary and tertiary EFL teachers in Japan towards more diverse ways of teaching, they will also be forced to learn more varied ways of testing. Not even hardened reading-translation (*yakudoku*) high school teachers now try to test students' conversation ability merely by multiple choice or fill-in-the-blank tests. This increasing interest in testing is influencing in-service teacher education programs, which are offered by nearly all municipal and prefectural boards of education, budgets allowing. Therein may be opportunities to teach testing.

Article Map

The first part of this article has two sections. In the first section, some of the preparatory student needs and ways they were met are outlined. After this, my experiences as a learner of testing are described, as well as the way it influenced my testing course. The second installment of this article also has two sections. The first outlines how a testing course was evaluated. The second discusses some short- and long-term teaching and planning issues.

Teaching Testing in West Texas

When I was assigned to teach a second language testing seminar for a M.A. in Applied Linguistics program in West Texas, fear was the first reaction. Despite a love of testing, I had never tried to explain testing principles to anyone. Knowing about testing principles and teaching them seemed to be two entirely different things.

To prepare for the course, I got help with verbalizing testing concepts since I would be lecturing the students at least part of the time, and also answering student questions. I had written testing articles before, but sensed a barrier between my literate self and verbal self. Many things which I could explain on paper with ease came out quite unclearly when I tried to talk, even given planning time. Imagining impromptu questions from students made fears more acute. I needed to find examples of testing concepts being explained in the oral/aural mode. One source came along almost immediately. One of my original testing teachers was teaching a second language testing course at a university in Tokyo. I asked to audit and record every class I attended. To my surprise, I found that the teacher actually lectured only a moderate amount of time. He spent more time discussing real life examples that illustrated the concepts he thought important. He also spent a large chunk of time simply asking for and answering questions. Often his next lecture would follow the theme of students' questions.

Listening to the teacher's class tapes before teaching my own testing class was helpful at least for the first two-thirds of the course. Towards the end of the semester, I needed the tapes less. After listening to the tape, I would review the textbook selections I had asked the students to read for class and identify the key concepts. I tried to recall from the tape what the teacher had said in class about those concepts, then jotted them down in my course planner. I worked from memory to avoid repeating what my teacher had said. This also helped me to process the concepts more deeply. I then delved into my own teaching and testing experiences and wrote up short examples that I thought would allow students to redigest the concepts in terms more attuned to their experiences as teachers. I also noted questions that students asked

and noted my answers.

One other minor source of test concept explanation in the oral/aural mode was the video series *Mark My Words* (1997) put out by the Language Testing Research Center (LTRC) at the University of Melbourne. This six video series covered topics such as "Language Proficiency Assessment" and "Classroom-based Assessment," and featured short interviews with testing luminaries like Geoff Brindley and Tim McNamara. But while the videos were an excellent review of general testing concepts and current issues in performance assessment, the interviewed subjects used a highly literate discourse style which sounded as though they were reading from a book. This meant that their utterances were densely packed with ideas. Interviews with ordinary language teachers featured in the videos were more helpful in that they talked about testing in terms of their classroom experiences, which was a discourse style I was leaning towards.

According to many teacher education researchers, our own experience as learners greatly influences how we ourselves will teach. This is not to say we are carbon copies of our teachers, but that our experiences form a kind of template upon which general expectations of what a class should be like and what constitute teaching are based (see Cohen and Spillane, 1992; Freeman and Richards, 1993; Kennedy, 1989; Lortie, 1975; MacDonald and Rogan, 1990; Porter, Floden, Freeman, and Schwille, 1987). It is hardly surprising: faced with the task of conducting a university level testing course, and the only experiences I had with such courses was as a learner.

I set out to articulate how I had learned testing concepts, and what had actually been learned. In doing so, I discovered three things: First, it was easy to quickly identify the things which were subsequently relevant to me as a testing practitioner; second, concepts and activities that I had not learned or experienced in my previous courses but later needed as a testing practitioner were also identified; and third, I found that my overall goals for students in my testing course had been strongly shaped by my previous teachers. My teachers had wanted me to apply testing principles to real situations and data through hands-on activities, and I wanted this for my students, too. Table 1 outlines that basic second language testing course that I taught.

Table 1. *How a Masters Level Language Testing Course Was Taught*

Course: Testing Language Skills

Date: Summer 1995

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

Course Purpose: (from syllabus) This course will provide students with a working knowledge of the basic principles for test construction and testing procedures with an emphasis on second language settings. Students will look critically at a variety of first and second language tests including standardized tests, integrative language tests, discrete-point tests and tests of communicative competence. No previous knowledge of statistics or higher mathematics is required. Students will learn the necessary statistical procedures to use in "testing the tests." This will enable them to read test manuals with understanding and construct their own examinations.

Course Requirements: Final exam, innovative test description, homework exercises, and participation. There are four homework exercises based on small datasets (N = 20, maximum) presented in the textbook: (1) item analysis; (2) descriptive statistics; (3) correlation; (4) reliability.

Mode of Class: Each three hour class consists mainly of lectures, anecdotal example explanations, and extended question and answer periods. Students are expected to read chapters from the textbook. The teacher will go over the textbook material and demonstrate mechanical calculations on the board. The entire textbook will be covered, in the order of the chapters in the book. During the question and answer period, the teacher will relate many of the concepts to practical situations made up of his extensive testing experiences at the program level. Students are also expected to complete exercises at the end of each chapter. Going over the exercises will take up part of the Q and A sessions. The homework exercises will consist of mechanical calculations of scores, data interpretation, and self-guided examination of existing tests known to the students. Students are encouraged to use spreadsheet and statistical programs, but are not given actual instruction in doing so. The teacher will also give out many examples of analytical and holistic rating scales and talked about them.

Content Areas: * criterion- versus norm-referenced tests; * relationship of CRTs and NRTs to different types of decisions; * history of language testing; channel versus mode; * discrete point versus integrative tests; * psychological constructs; * test fairness; issues involved in adopting, adapting, and creating tests; guidelines for giving tests and maintaining records; developing and improving test items; *item types: multiple choice, receptive response, matching, etc.; norm-referenced test item statistics; * criterion-referenced test item statistics; * nominal, ordinal, and interval scales; reading and creating histograms; * central tendency statistics; * dispersion statistics; * normal distribution; * outliers; * standardized scores; skew, kurtosis; * Pearson Product-Moment correlation (calculation and interpretation); * significance, meaningfulness (shared variance); * Spearman Rank-Order correlation; * point biserial correlation; * measurement error; * types of NRT test reliability estimates (test-retest, equivalent forms, internal consistency); * Spearman-Brown prophecy formula; * Cronbach alpha; * K-R20, K-R21; *interrater, intrarater reliability; * standard error of measurement; * CRT consistency estimation (threshold loss agreement, squared error loss agreement, domain score dependability); * agreement coefficient; kappa coefficient; phi lambda dependability; phi dependability; * confidence intervals; content validity; construct validity; * standards setting; * relationship of testing to curriculum; * developing goals and objectives.

Bottom Line: Students will cover a wide variety of content topics focusing on programmatic level CRT and NRT creation, use, and CRT and NRT score interpretation. Students will develop skills in connection with many of the content topics, particularly in completing calculations and displaying numerical data as homework. An "exploring language testing with statistics" course. Post-positivism (realism) with a strong streak of humanism.

Hidden Curriculum: CRTs should be used in the majority of educational situations. Tests have social implications and effects. Tests also have personal impact on students. Tests, and test scores, are often used irresponsibly (e.g., Japanese university entrance exams). Listening to other students' questions in the testing class is helpful to students. Practical applications to testing concepts should always be found. Students

should actively seek ways to use math and statistics by analyzing tests (data). Numerical data is valued. Math and statistics are not as hard as you think.

What Students did not get: Experience in creating criterion referenced goals and objectives. Experience comparing tests with specific curricula. Exposure to "alternative forms" of assessment, including portfolios, etc., in which the assessment generates descriptive rather than numerical data. Experience in creating and revising CRTs for specified classroom situations. Experience explaining or presenting testing content topics.

Other Sources Recommended:
Cronbach, L. (1990). *Essentials of psychological testing* (Fifth Edition). New York: Harper Collins Publishers, Inc.

American Educational Research Association/American Psychological Association/National Council on Measurement in Education. (1990). *Standards for educational and psychological testing.* Washington, D.C.: Author.

Table 2. *How an Ed.D. Level Language Testing Course Was Taught.*

Course: Doctoral Seminar-Advanced Topics in Language Testing

Date: Fall 1997

Text: Schumacker, R. & Lomax, R. (1996). *A beginner's guide to structural equation modeling.* Mahwah, NH: Lawrence Erlbaum Associates. Many handouts and additional readings.

Course Purpose: Introduce students to advanced concepts in testing, particularly linked issues of reliability/dependability/generalizability, and convergent/concurrent/content/construct validity. To give students practice using computer programs designed to explore these issues using large, authentic datasets (N = 500, minimum). Not quite right. Not all of the datasets were large. Some from Tabachnik and Fidell were not > 500, if I recall.

Course Requirements: Write a paper that demonstrates knowledge of concepts covered in class, preferably on a topic relating directly to students' dissertations. Homework assignments which involved statistical analysis and discussion of large datasets. Not a paper. Rather, it was to conduct a pilot study with special attention given to reliability and validity.

Mode of Class: Some lecture with extensive Q and A periods. Extended periods of small group, hands-on, guided use of statistical computer programs, including SPSS, EQS, GENOVA.

Content Areas Covered: *Uses of factor analysis; *factor analysis rotations; *exploratory and confirmatory factor analysis (2 homework assignments); * theta and omega reliability estimates (homework assignment); *generalizability theory (G and D study homework assignment); *path analysis (homework assignment + in-class work); structural equation modeling (homework assignment + many in-class tasks); *univariate and multivariate outliers; *issues of theory building (is theory imposed on the data, or does data make the theory?); multitrait/multimethod analyses item response theory (in-class tasks); *creating competency tasks (homework assignment); *a priori hypothesis testing versus "data snooping"

Bottom Line: The students learned the rudiments of discovering dimensionality in testing instruments, by analyzing large datasets using a variety of sophisticated statistical analyses. Post-positivist with a focus on the data itself, not the students. What is this-an assessment? Critical realist, I would say. This was said explicitly.

Hidden Curriculum: This implies some intention. I'm not sure what this means. Computer program copyrights must be respected. Datasets need to be screened and put in proper condition to use with computer programs. Students should be conversant with different types of computers. Students need to actually learn to use computer programs to transform their statistical knowledge and their attitudes about statistics. It's OK to force students a bit beyond their level of understanding. Not everyone learns at the same speed. Students should be asked to figure things out for themselves. Give hints, students

should do the rest. People who have doctorates should be running language programs. Did I say this? I recall telling XXXX that Ed.D.s should be able to do educational research. Program administrators should make responsible decisions based on numerical data. At least not ignore, if because of ignorance. Data can be used for many policy forming decisions, such as using path analysis to discover which students may be "at risk" in a program. Strong understanding of dimensionality is the basis of good testing. Fair testing. Students need to learn how to interpret data logically. Test and questionnaire construction are very similar. Large numerical datasets are valued. Because they are stable.

What Students did not get: In-depth experience with any one of the content topics covered. Experience working with small datasets. 48 hours ain't much. This would entail checking assumptions more rigorously I would guess. Experience explaining or presenting testing content topics.

Other Sources Recommended:
Mulaik, S. & James, L. (1995). Objectivity and reasoning in science and structural equation modeling>. In R. Hoyle (Ed.). Structural equation modeling: Concepts, issues, and applications. Thousand Oaks, CA: Sage.

Carmines, E. G. & Zeller, R. A. (1979). *Reliability and validity assessment.* Newbury Park, CA: Sage.

Note: *Testing concepts that were very relevant to the author's later work in testing.

Several issues emerged in the articulation of these experiences. One concerned the way that other instructors and I sometimes differed in our views of how testing should be taught. Another issue was my assessment of what content areas had been most relevant to me as a testing practitioner. One of my classmates, after reading my articulations, commented that she had felt that content areas that I had not chosen were relevant to her work. Clearly, my interests and concerns differed from other testing practitioners and potential testing teachers. I tend to stress certain concepts in my own testing course and not others, perhaps with or without conscious design. My preferences may become my own hidden curriculum. The term "hidden curriculum," borrowed from Zeichner, Tabachnik, and Densmore (1987) means assumptions of the teacher that are imparted to students but not mentioned in the syllabus, or other official documentation of the course. While some elements of the hidden curriculum are imparted without conscious plan, some aspects emerge intentionally as the teachers talk through the concepts of the course in a way that makes sense to them.

"While some elements of the hidden curriculum are imparted without conscious plan, some aspects emerge intentionally as the teachers talk through the concepts of the course in a way that makes sense to them."

I also found that the structure of the articulations emerged without conscious plan. As I strived to create a description of what and how I had learned testing, the categories of "mode of class," "the bottom line," etc. seemed to emerge from the act of writing down my recollections based on memory and all course documentation and homework I had retained in

my own records. As with the content areas I deemed relevant, the structure of the articulations themselves may be an artifact of my own selective interests.

Finally, it became apparent from an examination of all the data I ultimately generated from in the evaluation of my testing course that my classmates had great impact of my experiences as a learner of testing. This was not accounted for in my original articulations of my testing course experiences. In the basic testing course, one of my classmates played an essential role in getting me motivated to complete the homework assignments, and got me interested in the idea of playing around with datasets and finding other ways to analyze and interpret them. From her, I learned that doing testing was learning testing. In the advanced testing course, we classmates interacted thoroughly by reading each others' written work and moving among ourselves during classes offering and asking advice on operating computer programs, and analyzing and interpreting datasets. From these wonderful people such as Amy Yamashiro and Brent Culligan, I learned that mutual discovery and discussion were important learning tools. Evidence of the heritage of these two teachers, and my classmates, became apparent in the planning stages of the course, as will be discussed below.

Initial conception of my testing course

The department secretaries pressured me immediately for a syllabus for the testing course, which was actually doing me a great favor. In putting my thoughts into words, my course plans began to move in more definite directions. I needed to develop a course at the M.A. level for students of a variety of ages and level of teaching experience. I was told that some of the students would be public junior and senior high school teachers, and that they might be enrolled in Education, Interdisciplinary Studies, or Applied Linguistics programs. Finally, the course would meet 80 minutes twice a week for a total of 37 hours of instruction. The syllabus for that appears in Table 3 below.

Table 3. *A Graduate Testing Course Syllabus*

Course: Second Language Testing, LING 5345

Instructor: Greta Gorsuch, Ed.D.

Class Meeting Times: Monday and Wednesday, 4:30-5:50 PM

Office Hours: Tuesday 2-4 PM, Thursday 9:30-11:00 AM, Friday 11-12 noon

Welcome to the world of second language testing and assessment! In this course, I want you to get a working knowledge of basic principles of testing procedures which can applied to second language programs and

classrooms. Note what I said about "working knowledge." This means that we will be looking at actual tests and testing procedures, working with actual data, and creating tests and testing procedures that most fit your teaching situation. You might be relieved to know that no previous math or statistical courses are required for this course. We will, however, be using some basic math and statistics in the course, and I hope you will get a taste for the usefulness of statistics when looking at data of all sorts.

There are six course goals:

1. You will learn basic testing terminology and concepts, including: criterion-referenced test, norm-referenced test, item, item statistic, data type, mean, standard deviation, performance testing, "alternative" assessment, test reliability, and test validity.

2. You will learn how to work with datasets using a computer spreadsheet program, or by using the calculator + paper and pencil method.

3. You will learn how to assess tests and the procedures used to administer tests.

4. You will learn how to improve an existing test and testing procedure.

5. You will learn how to create your own test and testing procedure.

6. You will be able to identify resources that will help you with testing questions you may have in future. These may be in teaching journals, research journals, or on-line.

Class Format: There will be lectures, pair- and group-work, and student presentations. A lot of the reading you will be doing will present quite different content that what you have had in other teaching and language courses. I have two pieces of advice: First, keep up with the reading; and second, do the homework assignments I give you. The homework really does help. Your answers on the homework will also give me an idea of what topics I need to review in class, and whether I need to slow down, or speed up.

Assigned Reading: The main text is Brown, J. D. (1996). *Testing in Language Programs.* Upper Saddle River, NJ: Prentice-Hall Regents.

Other assigned readings are:

Brown, J. D., & Hudson, T. (1998). The alternatives in language assessment. *TESOL Quarterly, 32* (4), 653-675. This has been placed on electronic reserve.

Culligan, B., & Gorsuch, G. (1999). Using a commercially produced proficiency test in a one-year core EFL curriculum in Japan for placement purposes. *JALT Journal, 21* (1), 7-28. I will give you photocopies.

Murphey, T. (1994). Tests: Learning through negotiated interaction. *TESOL Journal, 4* (2), 12-16. This has been placed on electronic reserve.

Grading

Final Examination: 40% Student Presentation: 25% Homework Exercises: 25%
Participation (asking questions and expressing your viewpoints, insights, etc.): 10%

The final examination will take about one class period (80 minutes). Most of the items will be objective (only one answer is correct), but some will be open ended and I will grade the quality and comprehensiveness of your answer. I may ask you to interpret data, or do some calculations, or critique a test or testing procedure. Rest assured, however, you will not be asked to do anything we haven't covered and digested thoroughly.

The student presentation will involve a 10-minute presentation made by each student which describes a classroom test and testing procedure you would like to use, or have used, in a specific teaching situation. The format of your presentations may vary, but you should be sure to cover the following points: (1) give an adequate description of your teaching situation; (2) adequately articulate the construct you wish to capture in the test (tell us what it is you think you are testing--what skills, what knowledge, etc.);(3). give a comprehensive description of the development of your test instrument; (4) adequately describe your testing procedure with a focus on maintaining test reliability and test validity. You should also be prepared to respond to classmates' questions and comments. This should be a time of sharing and positive growth for everyone.

There were many similarities between my course syllabus and my experiences as a learner of second language testing at the basic level. For example, adopting Brown's Testing in Language Programs practically guaranteed that the content my own students would be exposed to would be same content I had experienced as a learner. Note also that in the course format lectures and homework are mentioned: this is what I experienced as a learner in the basic testing course. Finally, note that one overall goal in common between my course and my learning experiences are to have hands on experience working with data.

There were also differences between my syllabus and early testing class experiences. As stated earlier, articulating my experiences made clear to me what I had not gotten as a learner and felt that I later needed as a testing practitioner and testing teacher. One of these was experience in verbalizing testing concepts. In my course I expected students to give a presentation on a test that they had piloted, analyzed, and made revision plans for. While the requirements of the basic testing course I had included writing a description of an "innovative test," I was never actually required to pilot, analyze, and present one. In the advanced course, I was expected to pilot and analyze a test or questionnaire, but was never asked to present one orally. Further, the test or questionnaire was to be related to my dissertation, which in my case, had very little to do with my teaching situation. What I wanted in my course was to directly fit the testing concepts to students' needs (their classrooms), and then to get students to verbalize these concepts. Because of my positive experiences with my classmates-as-learning-community in both my testing class experiences, I wanted the student presentations to be an opportunity for community learning and friendly exchange of ideas. I also had hopes that the students would learn to be comfortable in presenting their testing ideas at conferences, and to colleagues and administrators at their workplaces. Even as I write this, I now realize I wanted my students to become testing practitioners and testing teachers. Even if they never taught testing formally, I wanted them to be able to share their knowledge with others, much in the ways I wrote of in the beginning paragraphs of this article.

Elements from my experience as a basic and advanced testing learner merged in the course I planned to teach. Note in the syllabus in Table 3 under the section student presentation, one of the criteria for grading that mentioned is the extent to which students can describe the construct they wish to capture in their test. The cornerstones in the advanced testing course I took were the notions of constructs and demonstrating construct validity through factor analysis. For the basic testing course, construct validity was discussed only briefly. In my own basic level course, I planned to work with the concept of construct

extensively, as a way of guiding students to writing better, more focused items.

References

Cohen, D. K. & Spillane, J. P. (1991). Policy and practice: The relations between governance and instruction. *Review of Research in Education,* 18, 3-49.

Freeman, D., & Richards, J. (1993). Conceptions of teaching and the education of second language teachers. *TESOL Quarterly, 27* (2), 193-216.

Kennedy, M. (1989). Policy issues in teacher education. (ERIC Document Reproduction Service No. ED 326 538). Available from http://www.eric.ed.gov/ERICWebPortal/custom/portlets/recordDetails/detailmini.jsp?_nfpb=true&_&ERICExtSearch_SearchValue_0=ED326538&ERICExtSearch_SearchType_0=no&accno=ED326538.

Language Testing Research Center (1997). Mark my words [video]. (Available from NLLIA Language Testing Research Center, Dept. of Linguistics and Applied Linguistics, 147-149 Barry St., University of Melbourne, Parkville, VIC 3052, Australia).

Lortie, D.C. (1975). *Schoolteacher: A sociological study*. Chicago: University of Chicago Press.

MacDonald, M. A. & Rogan, J. M. (1990). Innovation in South African science education (part 2): Factors influencing the introduction of instructional change. *Science Education, 74* (1), 119-132.

Porter, A., Floden, R., Freeman, D., Schmidt, W., & Schwille, J. (1986). *Content determinants*. Research Series No. 179. East Lansing: Michigan State University, Institute for Research on Teaching.

Zeichner, K. M., Tabachnik, B. R., & Densmore, K. (1987). Individual, institutional, and cultural influences on the development of teachers' craft knowledge. In J. Calderhead (Ed.), *Exploring teachers' thinking* (pp. 21-59). London: Cassell Educational Limited.

HTML: http://www.jalt.org/test/gor_1.htm **PDF**: http://www.jalt.org/test/PDF/Gorsuch1.pdf