EVALUATING THE TESTING COURSE IN AN MA IN ELT

Mike Orr

Introduction

The MA in ELT at University of Balamand, North Lebanon, started in 2005 in response to the realization that students graduating from the MA in English literature were largely headed for jobs teaching English as a Foreign or Second Language. The students, overwhelmingly graduates from the local branch of the Lebanese University, with a degree in English language and literature, apply to the MA in ELT program as a way to gain the professional development unavailable elsewhere. While this means the present paper reports on a specific context, the research is relevant to the wider contexts of assessment courses and professional development in language teacher education.

New educational projects may be planned and described in terms of the desired *outcomes*, which will be achieved by using specific *mechanisms* in a given *context* (Pawson and Tilley 1997). However, even such explicit description is insufficient to make useful evaluation into a practical proposition. There is a need "to identify criteria such that findings and judgments are grounded in both the experience of stakeholders and the rationale of the program" (Kiely and Rea-Dickins 2005: 14). In my experience in Lebanon, there is often a "wait and see" approach to teacher education, whether at tertiary level or when provided by one of several training organizations which operate in the region. An idea can be put into action, develop and become established practice without its success ever being formally evaluated. Where this happens, evaluation is a concern because success is being assumed. At the University of Balamand, such confidence may be due, in part, to the recruitment of mainly native speakers of English to teach on the course, and the pervasive, yet unsubstantiated, belief in Lebanon in the superior quality of teachers from the English speaking "centre" as opposed to the "periphery" where English is not the first language (Annous, forthcoming).

Evaluation of the MA in ELT program is being approached, globally, in terms of its practical relevance, and course by course, in terms specific to the topic. This paper reports on the approach taken to evaluate the "Testing and Assessment" course which runs 48 hours during one semester. The course is one of eight taken by all students and which each consist of three hours per week (see appendix A for the MA ELT syllabus). Although there are national tests prepared by a committee of university based experts, writing classroom tests is a regular part of many Lebanese school teachers' responsibilities and there is a need for trainee teachers to have a principled approach to formal assessment so that they can make informed decisions, for example, about whether or not to model their tests on those they experienced themselves as school students. I initially based the course objectives on the learning content drawn from recommended texts. Soon, however, the program developed a focus on classroom based, reflective practice, seen as a way to prepare inexperienced students for the multiple realities of school teaching in as varied a country as Lebanon (18 official religions, a large private education sector and dramatic intra and inter-regional differences in income and employment prospects). Such an approach would also serve as a way to empower the students once faced with the contradictions of teaching "on the ground" and the idealistic western vision of ELT presented in so many texts. This development happened organically and I felt it necessary to carry out the evaluation to help make a formal decision about the way we orient our courses.

Following Kiely and Rea-Dickins (2005: 15) the evaluation was made in order to learn the process of evaluation and use the information to make recommendations concerning the existing course and establish "a platform for the design and implementation of similar programs in related contexts." Thus, evaluation becomes more than just quality control, being seen rather as enhancing the quality of education. Likewise, Kelly et al. (2004) argue for a critical, but not judgmental approach which renders quality a process, not a state.

The course

The course input consisted of three principal activities. In the first, the students studied a set text, (Brown 2004), chosen for its coverage of the basic concepts, its numerous examples and its ready availability in Lebanon. The second involved studying the criteria for "test usefulness" (Bachman and Palmer 1996) and regular practice in using them to evaluate tests or test items. "Test usefulness" provides busy teachers with a practical tool, simplifying discussion of validity to the relevant construct. This, and the similarity to an evaluation checklist used on the technology in ELT course (Chapelle 2004), made Bachman and Palmer's approach preferable to the more recent, and time consuming, Weir (2005). The third activity was a collaborative, month long project, during which groups of 3 students had to choose a class of learners familiar to one of the students and write an appropriate test. Each group was set up to include at least one full-time teacher.

The students were assessed on course work, in which they presented summaries and led discussions of chapters in Brown (2004), a mid term reaction essay to two articles from the web based Language Testing Resources archive, the collaborative test writing project and a final 2 hour exam in which they were given a test from a local school (appendix B) and asked to write an evaluation report. Eight out of nine students successfully completed the course. The ninth student dropped out early in the semester to transfer to the literature programme.

The students

The class consisted of nine students, all graduates from the Lebanese University. Four had jobs as teachers and between two and eight years of experience. Of the other five, one had some part time teaching experience and the rest anticipated getting jobs in local schools in the coming months but had never actually taught.

Evaluation criteria

The criteria were drawn up retrospectively once it became clear that the course focus had shifted towards classroom based reflective practice, involving many of the features of a "normative-re-educative" process (Lamie 2005: 22). A review of the literature felt to be relevant to the context was carried out and certain criteria established which provided the basis for the evaluation. Student and instructor testimony, as well as coursework were studied for evidence of the criteria. Edwards and Owen (2002) recommend that trainee teacher employers and pupils should also be approached for their input in such an evaluation study, but as with their research, this was not done. I believe this could be useful in the future, but any impact of the course on the MA students' performance as language testers will take some time to become apparent.

It has been observed that because language teacher education (LTE) occurs in diverse contexts (Johnstone 2004), there are grounds for concern about the relevance of a Western ELT model (Skutnab-Kangas 2000; Holiday 2006; Orr 2008) for teachers working in other regions. However, this has not prevented a certain consensus about desirable features in teacher education courses.

Nyikos and Hashimoto (1997) look at a graduate level course for evidence of student teachers engaging in collaborative cognitive apprenticeships, "with the emphasis on metacognitive, reflective thinking" (p.508), critical thinking and self-regulation through language. Horowitz et al (1997) also stress collaboration, and distinguish it from group work in a similar way to Dillenbourgh et al (1995) who argue that collaboration is marked by being "a continued attempt to construct and maintain a shared conception of a problem" (p.189). Similarly, Vieira and Marques (2002) suggest that LTE should be evaluated with reference to the critical attributes that make them more or less reflective and empowering for the teachers and, consequently they argue, for the learners. While their concern for democratic activism and student involvement in decision making might appear to be motivated by political correctness (Waters 2007), they are actually emphasizing the importance of learning, not learner-centredness, through context-sensitive decision making.

Egbert (2006: 167) suggests that "teacher change and growth occur through learning that is *situated* in classrooms." Rather than simply providing teachers with occasional micro-teaching sessions, the education course needs to be situated as much as possible in authentic contexts that the teachers recognize, yet which allow them to "reflect on practices different from their own" (p. 177). (cf. Lave and Wenger (1991) on situation based learning to be part of a community of practice). Egbert describes LTE and technology, but the idea is generalisable, as it is with Slaouti and Motteram (2006) who discuss "reconstructing practice", promoted by the existence of "linkages between the learning outcomes, the learning and teaching processes, and the assessment" of the

course (p. 89). Principled reflection is part of this as it involves encouraging "metacognitive processes", "conscious articulation over time" and a "reflection on apprenticeship" (pp. 90-93). This approach addresses Freeman's (1996) argument that learning how to teach is more than mastering certain behaviours in class. There is "a cognitive dimension that links thought with activity, centering on the context-embedded, interpretative process of knowing what to do" (as cited in Slaouti and Motteram 2006). This is relevant to a language testing course because of the potential for principled classroom practice to influence the teacher as test writer, and for a thoughtfully designed test to influence classroom teaching.

Kohonen (2007) emphasizes doing LTE within a reflective, experiential framework involving concrete classroom situations. He describes "transformative teacher growth" as being a process in which the teacher, and by implication the teacher as tester, "needs to works on his/her educational values, beliefs and assumptions and carry out conscious work on his/her professional identity as a language educator" (p.6).

With all these authors, there is a concern for principled, empowering and grounded teacher education, suggestive of how Bloom's (1956) taxonomy might be operationalised in an LTE context. This involves providing teachers with practice, theory, and the skills required to reflect critically on their teaching by articulating their decision making process, relating it to their professional knowledge base in such a way as to formulate the theory of teaching that they adhere to. Bringing this theory to the surface allows it to be evaluated, and enables the teacher to plan for and implement change as s/he sees fit (Farrell, 2007).

The students' knowledge base, as far as the language testing course is concerned, came from two sources. First, the basic introduction provided by Brown (2004) was adopted, particularly sections on the relation between testing, assessing and teaching, and designing different types of classroom tests. Second, and in order to give the students a practical tool kit for their work on tests in schools, Bachman and Palmer's (1996) criteria for evaluating "test usefulness" were included. The criteria are reliability, construct validity, authenticity, interactiveness, impact and practicality. These are preferred to Brown's own chapter on evaluating tests because of the single presentation of validity, with aspects other than construct validity being subsumed within the other criteria. It is felt that in describing "test usefulness" the authors start from the user's needs and provide the theory as necessary; exactly what teachers appreciate (Reid 1999). A further reason includes the students' familiarity with Bachman and Palmer's Communicative Language Ability (CLA) from previous courses in the MA program. Additional input was provided from the Cambridge Assessment series when describing constructs, for example Buck (2001) for listening.

On the basis of the above, certain criteria were established, against which the language testing course was evaluated after it had finished. These were formulated as the following desirable outcomes:

1. The students *express themselves* knowledgably about the purposes of assessment and about different techniques for assessing second language ability in the four skills, grammar and vocabulary;
2. The students *express themselves* knowledgeably about Bachman and Palmer's (1996) criteria for evaluating test usefulness;
3. The students *apply* the criteria for test usefulness to ready made language tests and use the criteria to inform their own writing of tests for use in school;
4. The students will *reflect* critically in, on and for action, in the process of collaborating on tasks.
5. The students *perceive* the course as of direct relevance to their work situation and of practical use to them as teachers.

Having originally seen the learning outcomes in terms of content knowledge, practical application and my own methodology, I had not considered what I would need to carry out an evaluation. Thus, one concern was whether or not the data available was sufficient. This will be considered along with the results.

The evaluation

A principle source of data was the student diary kept during the month long test construction project. This was ostensibly kept in order to develop the habit in preparation for the following semester's Practicum course. I suggested the diaries be a record of the project work and a place for questions which could be discussed later, as well as any other reactions they had to what they were doing. The students knew the diary work would not affect their grade and this will hopefully have limited the possibility of the students writing "what they think we want to hear" (O' Rourke

1998: 410). Lee (2007) discusses how pre-service teachers learn to reflect and argues that they need not wait for a practicum course, but can begin during other courses. She examined journal entries for evidence of learning additional perspectives on teaching; relating one's own experiences to the course content; and an understanding of the significance of a "broader, social, cultural and pedagogical context" (p. 325). Student diaries were found useful by Halbach (1999) for evaluating a methodology course and Yatasuka and Higashibara (2005) used electronic portfolios with reflection sheets to evaluate a preservice course at university, although the evaluation was only based on the student teachers' responses to a Likert scale checklist.

Given the scope of the criteria established for the present study, I also decided to use the project tests produced by the groups of students, the final exam evaluation reports of a local school test and my own diary. I wrote this up once a week after class, but occasionally made notes during class or preparation. Furthermore, six months after the course, I interviewed the students and asked what they remembered of it, and about its usefulness for their work as teachers. This qualitative data was examined for evidence of the evaluation criteria.

The evaluation process of extracting data from their diaries, projects and test reports was discussed with the students. They approved the plan and gave formal consent for their work to be used and the results of the evaluation presented for publication.

In what follows, students' real names have been replaced with:

Angie – experienced teacher; Mariam – experienced teacher; Rim – experienced teacher; Farah - little experience; Ziad – no experience; Sara – no experience; Eliane – no experience; Souad – no experience.

Results

1. The students express themselves knowledgably about the purposes of assessment and about different techniques for assessing second language ability in the four skills, grammar and vocabulary.

The students' diaries, test evaluation reports and interview transcripts were examined for evidence. In the diaries there were many references to the term *achievement* while the term *formative* is only common in the interview transcripts. For example, Farah remembers, "We learnt about cloze tests and rewriting sentences and summarizing…about knowing if it's for helping the students with their problems or giving grades." In general, all the students mentioned a wide variety of techniques, both familiar from school (e.g. reading aloud; paraphrasing parts of a text) and new to them (e.g. paired oral interviews; listening tasks; identifying pronominal reference in reading texts). Below are some illustrative examples.

Rim wonders about assessing grammar with an error correction task and concludes, "I'm thinking of replacing it with multiple choices, but that's got weaknesses too." Sara comments, "It's a problem to find tasks that assess achievement, rather than just proficiency." Souad decides "compare/contrast 2 photos is better than just describing one picture. It makes them use more complicated grammar." In Angie's group, however, the diaries do not show the same discussion about task type, "we'll use multiple choice questions because the kids are familiar with these."

Summary

The students show they know how to talk about tests for use in school although I would have preferred to get more idea about each student's ability to rationalize the choice of one task type rather than another. While all the students were clear about the purposes of assessment, the inexperienced teachers wrote less about how the different task types chosen could be used to give them the desired information about the learners. The data gathered did not give a full picture. Direct questions on the topic would likely be more successful.

2. The students express themselves knowledgeably about Bachman and Palmer's (1996) criteria for evaluating test usefulness.

The students' diaries, interview transcripts and evaluation reports of a local school test were examined. In their diaries, the students made comments about usefulness in general and about each of the six criteria. The evaluation reports and interviews did not always refer to all six criteria. Below are some illustrative examples. In the interests of space, not all criteria are covered.

Reliability

Rim's diary mentions that: "a holistic rating scale for writing will be more practical but an analytical scale more reliable," while Sara's diary wonders about finding reading texts that will "appeal to boys and girls so that some pupils' lack of background knowledge does not affect the reliability of the scores." In the interviews, all students mentioned reliability as being about consistency and trying to eliminate other factors from affecting the scores. They all mentioned the

threat posed by the subjectivity of the rater when faced with comprehension question answers containing spelling and syntax error. They made no mention of using statistical procedures to check reliability. Although covered in one session, the test writing project and evaluation report did not provide the opportunity to practice on test results. The point will be noted in the conclusion.

Construct validity

Mariam reflects on construct validity in her diary, **"**we cannot test everything, but what we can do is provide a variety of tasks to be able to make good inferences." She goes on to write "with these multiple choice comprehension questions, we can make inferences on their understanding of details." She criticizes another item because there are "no relevant inferences we can make from this exercise." Souad, in her report, seems unclear about construct validity when discussing reading ability but gives a good critique of an editing task, pointing out how underrepresented this leaves the construct of L2 writing. Eliane also criticizes this task because "it only tests the knowledge of morphology and grammar, but it doesn't really test the writing skill." However, in her interview, Eliane referred to evaluating construct validity as a question of checking the relevance of a task, rather than asking if the construct itself is sufficiently represented.

Authenticity

In the evaluation report, Farah considers authenticity and writes, "this task is like school work, ok, school is the real world for these students." In the interview, Sara said that authenticity of test task had stuck in her mind as the key principle because the test should be based on "useful language", although she said nothing about authenticity of the test task itself.

Practicality

As for practicality, Eliane's report summarizes what all the students say, describing it as having the resources to make the test "easy to administer, easy to do and easy to mark." As such, students saw a relation between practicality and scorer reliability, although there was no mention of validity in their discussions of practicality.

Summary

In their diaries, all the students discuss Bachman and Palmer's (1996) criteria of construct validity, authenticity and practicality. Reliability is discussed in terms of item design, but not of later analysis, logical in this case given they did not have test results to work on. Some of them make little reference to interactiveness and impact, particularly the washback effect a test might have on teacher practices before the test. Impact and interactivity are noticeably absent from the interview transcripts. There are, however, comments about all the criteria in the diaries of at least one member of each group, allowing for the supposition that there was, at least, discussion of all six criteria during the test construction project. It appears that some criteria are understood more easily in relation to certain skills than others, for example the construct validity of listening tasks is hardly mentioned. This may be because they take it for granted that authentic looking listening tasks have construct validity. Performance on the test evaluation task indicates that some students who could write a diary entry about the impact and/or interactiveness of their group designed task, were not so clear about these criteria when faced with a task that had to be done alone.

3. The students will be able to apply the criteria for test usefulness to ready made language tests and use the criteria to inform their own writing of tests for use in school.

The test evaluation reports include some recommended changes clearly based on applying the criteria, while the tests produced during the project show the students applying them somewhat unevenly. Below are some illustrative examples. In the interests of space, not all criteria are covered.

Construct validity

Eliane's evaluation report recommends eliminating a reading comprehension question which asks for an opinion unrelated to the text, on the grounds of being irrelevant to the construct of reading. She also recommends including a vocabulary from context task "so we can know more how they read." Ziad's evaluation report proposes making "the students use the words, not just choose the right one so the construct validity will be improved for vocabulary." Angie's group produced a potentially useful achievement test for 7 and 8 year olds. The group carried out an evaluation of usefulness and made changes when they found the criterion of construct validity caused problems: copying the format from her school, Angie had included a reading aloud test with scoring criteria of fluency and pronunciation, but described the construct in terms of grammar competence and knowledge of vocabulary.

Interactivity

In their reading comprehension tests, Angie's group had an item based on the copying of whole

sentences and needed help to see that interactivity was very low. However, they went on to include an inference question in the reading comprehension task "to be more demanding" and increase interactivity. Mariam's group prepared a grammar test with a text where occasional lines had an extra word. This was in response to a diary comment, "for the grammar, how do we really make them think about the language?"

Impact

Farah's diary discusses using the test to motivate learners and that a letter writing task would have "positive impact because I have a struggle to teach them how to write a letter." Rim's group included a paired oral task as a way to encourage teachers to "make the effort to get the students speaking – even if it is difficult." Angie's group, however, included some vocabulary items which require memorization of definitions without any indication of understanding.

Summary

The students all showed an ability to make principled decisions. There was a high degree of tolerance for inauthentic tasks in tests for use in schools on the basis that they were typical of school work. Interactivity was generally good, probably because the students chose to use task types they had encountered in their text book. Perhaps surprisingly, there was little evidence of impact influencing decisions and some tasks unlikely to produce positive washback were left in the local school test. On the other hand, the tests produced by the students were generally such that in order to prepare for them the teacher would have to prepare lessons with good language learning potential, meaning focus and opportunities to practice and develop communicative skills.

4. The students will reflect critically in, on and for action, in the process of collaborating on tasks.

The students' diaries kept during the test writing project, and my own diary, were examined for evidence.

In my diary, the most common category of entry concerns the dynamic of the collaboration on the test construction project. For example, "Rim (the experienced teacher in the group) is quite convincing, the others gave their opinions and suggestions but once she suggested the video task they just went along with it."

My diary also contains many references to the questions the students asked and my repeated need to question them to see if they had ignored some obvious threat to reliability or if they had really thought about construct validity and the inferences they would be able to make on the basis of a task. "It's interesting how suddenly they got the point that their reading comprehension questions wouldn't tell them much about the learners' abilities beyond matching words in the questions with others in the passage…but it takes me to ask how a kid would **do** the test, before they start to ask if the results would be much use to them."

"Had to ask again why they will deduct marks for spelling in the listening comprehension note taking test. Still seem to be looking for what the kids can't do rather than what they can."

In the students' diaries, five relevant categories of comment were identified.

1. Perceptions of group work.

Two of the three experienced teachers use the pronoun "I" to refer to a number of tasks whereas the inexperienced teachers generally use the pronoun "we". For example, Mariam took charge in her group, and even felt at one point: "I think I have to finish the project by myself." In contrast, Ziad who was in the latter group writes, "We want to design…; We started thinking…." Moreover, he finishes with, "the test we wrote was a good opportunity for me as a new teacher to learn from an experienced teacher."

There are many references to the positive feelings generated by the group work. Eliane writes, **"**We searched the internet together and found some grade 2 test material – it's really exciting when this happens." Likewise, Farah notes, "We are really cooperating…really excited... group work really helped to a great extent in constructing our knowledge and become good learners." Only one student, Mariam , mentions negative feelings, "Working in groups is sometimes frustrating!" This is related to the practicality of off campus collaboration on the test writing project. Although one group was able to meet easily, the others lived quite far apart. Moreover, this took place during a period of fighting in the city and electricity was frequently cut. When it was safe to travel, these students found the only practical solution was to come to class early and use this time for the project. As Sara comments, they found themselves bringing individual contributions to the project, "at the end, we decided to search each one alone, and then bring new ideas and texts."

2. Questioning their actions.

Sara asks, "What have we chosen? Do different topics benefit boys more than girls? Interesting

idea. I'll discuss it with my friends" and Mariam wonders, "I feel I have doubts, have we missed something?"

3. Categorizing, selecting and prioritizing.

The students spent a lot of time working backwards from tasks they had chosen, categorizing them to see if a particular construct was covered. They also wrote about the need to be selective about what to include and in what order. Ziad explains, **"**We created a table for each skill and divided it into tasks and the type of answer we expect. We need to think about the suitable order for the tasks." Skills and question type seem to be the criteria for ordering rather than item difficulty.

4. Referring back to their instructor.

Mariam typically comments, "we talked through the reading task with Mike… it is not enough to give us good inferences, so add another task!"

5. Critically relating project work to concepts and practices previously studied.

Sara describes how she "looked at the students' book and at some tests of mine I kept from the last year at school. I don't think they are useful for Rim's students because they are designed like the TOEFL! Today I read the summary of the speaking chapter. Not all the suggested tasks can be used in our test. Interview and discussions are the most useful." Farah explains that they "agreed that if a student was off topic we still give a grade for grammar – I feel we become more open in thinking about tests."

Summary

The students asked me, themselves and each other questions, although I also had to question them. Reflection was further seen in their referral to test usefulness, their experience and the knowledge gained from their text book. They were aware of the benefits of collaborating and while commenting on how they enjoyed this, they also described some frustration. Experienced teachers appeared to want to work faster, although this may have been a consequence of the allocation of tasks to individuals between group sessions and the consequent replacement of collaboration with the less useful, cooperative type of teamwork criticized by Dillenbrough et al (1995).

5. The students will perceive the course as of direct relevance to their work situation and of practical use to them as teachers.

Evidence for this criterion was looked for in the students' diaries and interviews. There are indications that the course broke down the separation between teaching and testing in a way beneficial for the learners. Sara writes, "Not only in test writing but actually in all my lesson planning I think of the criteria" while Souad adds, "We use the criteria whenever we're thinking about how the students are managing a task and what we can learn about the students from their work." Angie, criticizes her school's tests for not being weighted properly according to objectives. She found herself using the test evaluation experience during the observations for the Practicum course, as a guide to evaluating what the teacher was doing in class. Eliane finds the course useful for dealing with new responsibilities. "At least I can have a plan when I sit down and I know what to say when I explain things to the coordinator."

Summary

The students appreciated the course as contributing to their work as teachers, enabling them to gather "feedback on the effectiveness of the teaching program itself" (Bachman and Palmer 1996: 8) as well as being useful for the classroom quizzes and tests they are required to produce.

Conclusions

The criteria chosen for evaluating the Testing and Assessment course represent a model of language teacher education based on collaborative learning and interaction between theory and situated practice through reflection on decision making which articulates thinking about the student's knowledge base.

The use of the criteria permits a generally positive evaluation of the course, although the students' ability to apply some of Bachman and Palmer's (1996) criteria is in doubt. The following recommendations are made. Desired outcomes should be specified from the beginning to include situated, collaborative and reflective practice, rather than these being "secondary" considerations of methodology. The objectives should also include the ability to express oneself appropriately about the theoretical base that constitutes the knowledge to be accessed critically during reflection. These two recommendations will facilitate the third which is that data collection for evaluation be planned in advance to include a wide variety of data: formal observation of the students at work on the project, periodic interviews, and a more formal type of questioning come to mind. A fourth recommendation is for the project work to include opportunities to work on some test results in order to practice simple statistical procedures for checking reliability.

Finally, it is recommended that the course should allow more time, overall and in class, to the test writing project to allow for mainly collaborative as opposed to cooperative type working. This might require reorganizing the course and making the students responsible for reading more material out of class. The students and instructor could discuss on-line, material presented in the form of documents and slideshows accessed on a simple webpage such as a wiki hosted by www.wetpaint.com, and which I already use for other courses. This would replace some of the class preparation time as the instructor would be providing more scaffolding and less whole class input. It would certainly help solve the problems of arranging to collaborate off campus when many of the students have jobs and prefer to avoid unnecessary travel, especially in the evening.

Finally, evaluating this course has been a learning experience in itself and will be a reference point within the developing, critical approach to quality assurance on the MA in ELT program.

References

Annous, S. (forthcoming). Le "Nativespeakerism" et l'identite des enseignants non natifs de l'anglais au liban. In Dervin, F. and Badrinathan, V. (eds.) *L'enseignant non natif:* Identites et legitimite dans l'enseignement-apprentissage des langues etrangeres. New York: Peter Lang.

Bachman, L. and Palmer, A. 1996. *Language Testing in Practice: Designing and Developing Useful Language Tests*. Oxford: Oxford University Press.

Bloom, B. 1956. *Taxonomy of Educational Objectives: Book 1, Cognitive Domain*. New York: Longman.

Brown, H., D. 2004. *Language Assessment: Principles and Classroom Practices*. White Plains, NY: Pearson Education.

Buck, G. 2001. *Assessing Listening*. Cambridge: Cambridge University Press.

Chapelle (2001). *Computer Applications in Second Language Acquisition*. Cambridge: Cambridge University Press.

Edwards, C. and Owen, C. 2002. "What should go into an MA TEFL programme? Teachers' evaluations of the taught components of a sample programme". *ELTED 7*: 54-73.

Egbert, J. 2006. Learning in Context: Situating Language Teacher Learning in CALL. In Hubbard, P. and Levy, M. (eds.) *Teacher Education in CALL*. Philadelphia: John Benjamins. 167-181.

Dillenbourgh, P., Baker, M., Blaye, A. and O'Malley, C. 1995. The Evolution of Research on Collaborative Learning. In Reimann, P., and Spada, H. (eds). *Learning in Humans and Machines: Towards an Interdisciplinary Learning Science*. London: Pergamon. 189-211.

Farrell, T.S. 2007. *Reflective Language Teaching: From Research to Practice*. London: Continuum.

Halbach, A. 1999. "Using trainee diaries to evaluate a teacher training course." *ELT Journal* 53/3: 183-190.

Holliday, A. (2006). "Native-speakersim." *ELT Journal* 60/4: 385-387.

Horwitz, E., Bresslau, B., Dryden, M., McLendon, M.E., Yu, J. 1997. "A Graduate Course Focusing on the Second Language Learner. *Modern Language Journal* 81/4: 518-526.

Johnstone, R. 2007. Language Teacher Education. In Davies, A. and Elder, C. (eds.) *The Handbook of Applied Linguistics*. Malden MA: Blackwell Publishing. 649-671.

Kelly, M., Grenfell, M., Allan, R., Kriza, C. and McEvoy, W. 2004. *European Profile for Language Teacher Education – A Frame of Reference*. European Commission Directorate General for Education and Culture. Retrieved 15 December 2008 from www.ec.**europa**.eu/education/languages/pdf/doc477_en.pdf

Kiely, R. and Rea-Dickins, P. 2005. *Program Evaluation in Language Education*. Basingstoke: Palgrave Macmillan.

Kohonen, V. 2007. Developing Foreign Language Education through Transformative Teacher Growth. *Humanising Language Teaching*. 9/1. Retrieved 12 December 2008 from www.hltmag.co.uk.

Lamie, J. 2005. *Evaluating Change in English Language Teaching*. Basingstoke. Palgrave: Macmillan.

Language Testing Resources. http://www.languagetesting.info

Lave, J. and Wenger, E. (1991) *Situated Learning. Legitimate peripheral participation*. Cambridge: University of Cambridge Press.

Lee, I. 2007. Preparing Pre-service English Teachers for Reflective Practice. *ELT Journal* 61/4: 321-329.

Nyikos, M. and Hashimoto, R. 1997. Constructivist Theory Applied to Collaborative Learning in Teacher Education: In Search of ZPD. *The Modern Language Journal* 81/4: 506-517.

O'Rourke, R. 1998. "The Learning Journal: from chaos to coherence." *Assessment and Evaluation in Higher Education*. 23/4: 403-413.

Orr, M. 2008. Localising the global: ELT websites as a context for teacher development. An evaluation of the British Council NENA region ELT Networking project. Conference paper at Cultural dialogue: communication channels among nations – Minia University, Egypt. 3-5 November 2008.

Pawson, R. and Tilley, N. 1997. *Realistic Evaluation*. London: Sage.

Reid, D. 1999. Investigating Teachers' Perceptions of the Role of Theory in Initial Teacher Training through Q Methodology. *Mentoring and Tutoring* 7/3: 241-255.

Skutnabb-Kangas, T. 2000. Linguistic Human Rights and Teachers of English. In Kelly Hall, J. and Eggington, W.G. (eds.) *The Sociopolitics of English Language Teaching*. Clevedon: Multilingual Matters. 22-44.

Slaouti, D, and Motteram, G. 2006. Reconstructing practice: language teacher education and ICT. In Hubbard, P. and Levy, M. (eds.) *Teacher Education in CALL*. Philadelphia: John Benjamins. 81-97.

Vieira, F. and Marques, I. 2002. Supervising reflective teacher development practices. *ELTED* 6: 1-18.

Waters, A. 2007. ELT and 'the spirit of the times'. *ELT Journal* 61/4: 353-359.

Yatsuka, M. & Higashibara, Y. (2005). Evaluation and Development of Preservice Teacher Training Courses with Student Teachers' Electronic Teaching Portfolio in Shinshu University. In C. Crawford et al. (Eds.), *Proceedings of Society for Information Technology and Teacher Education International Conference 2005* (pp. 256-259). Chesapeake, VA: AACE. Retrieved January 12 from http://www.editlib.org/p/18994

Appendix A

Semester	**Course**	**Course**	**Thesis**
First	Second Language Acquisition	Methodology I	
Second	Testing and Assessment	Research Methods in ELT	Prepare research proposal
Third	Methodology II	Teaching Practicum	Carry out research project
Fourth	Information Communications Technology in ELT	Educational Management	Carry out research project. Present research project

The MA in ELT syllabus followed by the students referred to in this paper.

Appendix B

Grade 4 English test from a local private school. This test was one of several provided by teachers working in north Lebanon. It is copied here exactly as it appears in the original. The name of the school and the test date have been deleted.

Grade 4
Time: 2 hours

English test
Final exam

…….
25

I. Read the following text.

Milk is considered the most nearly perfect of all foods. Because it contains most of the elements the body needs, a person could live on milk alone for some time.

In addition to containing the body needs, milk contains them in a form that is easy for the body to use. There is fat in milk. We often get this fat in the form of butter or from drinking whole milk or cream. Milk also contains sugar and protein, both of which are necessary to the body.

In desert countries, people get milk from camels and some of the nomads drink reindeer milk. In some countries such as Lebanon, much milk comes from goats. In the United States, milk is ordinarily gotten from cows. Since there is one cow for every nine people in the United States, many gallons of milk are available to each man, woman or child. If each one of us would drink his or her share of milk each year, we might all be healthier.

II. – Reading comprehension: answer the following questions (5 pts)

1 – Give a suitable title for the text (1 pt)

……..

2 – Is milk considered an important type of food? Why? (1 pt)

……..

……..

3 – Does milk contain fat? If yes, in what form do we get it? (1 pt)

……..

……..

4 – What nutrients are usually found in milk? (1 pt)

……..

……..

5 – Do you like to drink milk or eat milk products? Why or why not? (1 pt)

……..

……..

III – Vocabulary (7 pts)

I – Choose the correct meanings of the underlined words (3 pts)

(1) In the United States, milk is ordinarily gotten from cows.

a- usually

b- approximately

(2) Some of the nomads drink reindeer milk.

a- wanderers

b- wild animals

(3) Milk is considered the most nearly perfect of all foods

a- is studied as

b- is thought to be

2- Choose the word that best fits the meaning (2 pts)

(a) **more healthy** means	(1) ordinarily
(b) your **part** is your	(2) healthier
(c) if it **can be gotten**, it is	(3) share
(d) **usually** means	(4) available

3- Omit the odd word (2 pts)
(a) banana-apple-milk-cucumber
(b) chair-table-sofa-dog
(c) father-daughter-friend-mother
(d) lucky-healthy-drink-brave

IV – Grammar (8 pts)

1- In the first line of the passage: define the verb "is", its tense, and why? (2 pts)

……..

……..

2- Find in the text one subject pronoun and replace it with a suitable noun. (2 pts)

……..

……..

3- Each sentence of the following contains one error, underline it and correct it. (2 pts)

(a) In desert countries, people gets milk from camels: ……..

(b) Yesterday, Tom goed to the beach with his friends: ……..

(c) Since there are approximately one cow for every nine people: …….

(d) We often gets this fat in the form of butter: ……..

4- Replace the underlined words with suitable pronouns. (2 pts)

(a) Jim likes pizza: …… (c) My parents love me: ……
(b) Gaby is a shy person: …… (d) Lions are wild animals: ……

V – Writing task (5 pts)

The following passage contains a number of errors in grammar and spelling.
Underline the errors and correct them in the space provide.

Speak without fear

The biggest problem most peopel faces in learning a new langauge is their own feer. They worry that they will not says things correctly or that they will looks stupid so they don't tolk at all. Do not do this. The fastest way to lern anything is to do it – agayn and again until you got it right. Like anything, learning English requires practise. Do not let a lettle fear stop you from gotting wat you wants.