# How good is your test?

*Funda Küçük and JoDee Walters*



*This article reports on a study of the validity and reliability of tests administered in an EFL university setting. The study addresses the question of how well face validity reflects more objective measures of the quality of a test, such as predictive validity and reliability. According to some researchers, face validity, defined as the surface credibility or public acceptability of a test, has no theoretical basis since it is based on the subjective perceptions of stakeholders such as teachers and students. However, due to lack of time or resources, or due to a perceived lack of competence, practitioners tend to rely on the 'appeal' of language tests, rather than seek empirical evidence. This article describes several ways of evaluating achievement tests, comparing their results in order to shed light on what measures can and should be taken to ensure that achievement tests accomplish their purposes.*

## Background

In large educational institutions, achievement tests are designed by testing offices, rather than individual teachers, to ensure standardization. Unfortunately, instructors and test takers may not trust these tests or the testers (Hughes 2003: 1). Even if teachers design their own achievement tests, such tests may not accurately reflect the students' language knowledge and skills. For these reasons, the assessment of achievement tests is crucial.

One way to assess such tests is to examine the qualities that determine their effectiveness. Bachman and Palmer (1996: 38) define these 'good qualities' as reliability, validity, authenticity, interactiveness, wash-back impact, and practicality. Bachman (1990: 289) identifies validity as crucially important. In general, validating a test involves gathering empirical data and other relevant information to show that the test is indeed measuring what it intends to measure. There are several validity types, including predictive validity and face validity, each of which entails collecting data in different ways.

To investigate predictive validity, which indicates that the test accurately predicts the possible future success or failure of the test takers (Hughes 2003), test scores are correlated with scores on tests taken some time later. Face validity, another way of looking at the validity of a test, refers to the degree to which the test seems valid, in the eyes of those involved in taking or administering it, in terms of testing what it has to test (ibid.). Research into face validity requires investigation of the subjective judgements and perceptions of the test's stakeholders (preferably both instructors and students).

While validity is a fundamental quality of tests, reliability is a precondition for validity, because unreliable test scores cannot provide suitable



*ELT Journal Volume 63/4 October 2009; doi:10.1093/elt/ccp001*

Advance Access publication February 23, 2009

grounds for valid interpretation and use (Bachman 1990). Two essential concepts are involved in reliability: 'scorers' reliability' and 'reliability in terms of the test takers' performance' (Hughes op.cit.). Scorers' reliability refers to the degree to which test scores are free from measurement error (Rudner 1994: 3). Reliability in terms of the test takers' performance refers to the extent to which test scores of a group of test takers are consistent over repeated test applications (Hughes op.cit.).

Several researchers have conducted studies in an attempt to assess the reliability (Cardoso 1998; Nakamura 2006) or validity (Ösken 1999; Serpil 2000; Yeğin 2003) of tests given in university English preparatory programmes. However, none of these studies has explicitly compared face validity with more objective measures of tests. Therefore, this study aimed at both measuring the validity and reliability of achievement tests administered in a particular EFL setting and exploring how well face validity reflects relatively more objective measures of tests: reliability and predictive validity.

## The study

This study addresses the following research questions:

1. To what extent do the achievement tests possess face validity in the eyes of the instructors and students?
2. To what extent do the achievement tests possess reliability, in terms of both (scorers' perceptions of) scorer reliability and the test takers' (perceptions of) performance?
3. To what extent do the achievement tests possess predictive validity?
4. How closely does the face validity of the achievement tests reflect the reliability and predictive validity of these tests?

The study was conducted at the Zonguldak Karaelmas University (ZKU) Foreign Languages Compulsory Preparatory School, in Zonguldak, Turkey. While ZKU is not an English medium university, the preparatory school (PS) aims to prepare students for those courses in their future university departments that will include some teaching and materials in English. In the PS, students attend speaking, writing, and reading courses in addition to a main course of integrated skills, which includes grammar instruction. There are also video and language laboratory courses which include listening skills. Two midterm achievement tests are given in each term, and a final exam is administered at the end of the course. On these tests, while grammar, writing, vocabulary, and reading are represented, listening is not tested at all, and speaking is represented only on the final exam, with its weight relatively insignificant. Neither the video nor the laboratory course is separately represented on the tests. After PS, the students move on to a General English (GE) course, addressing the four language skills, taken alongside their regular university courses.

This study, conducted in 2007, was focused on the tests administered by the PS during the 2005–2006 academic year. Two different groups of participants were included in this study. Fifty-two students (four from each university department) who had been enrolled during the 2005–2006 academic year formed the first group, and 29 instructors who had been

teaching at the school in the same year formed the second group. At ZKU, instructors rotate through the testing office, and all instructors participate in scoring the tests. Therefore, the instructors had all been involved in the preparation and scoring of tests.

Two questionnaires were employed in this study, one for instructors and one for students. Common sources of invalidity and unreliability suggested by the literature (for example Brown 1996, 2004; Genesee and Upshur 1996; Hughes op.cit.) provided a basis for the Likert scale items. The first sections of both questionnaires contained the same questions, concerning participants' perceptions of the face validity of the achievement tests. The second section of the instructors' questionnaire investigated scorer reliability, and the second section of the students' questionnaire concerned issues related to test takers' performance. It should be noted that these sections, while purportedly measuring the reliability of the tests, are relying on the participants' *perceptions* of reliability. However, many of the questions address very specific and observable aspects of the testing experience and so are thought to be sufficiently objective. The students' questionnaire was administered in Turkish, while the instructors' questionnaire was given in English.

In addition to the questionnaires, the 2005–2006 midterm and final exam scores for 365 students were collected, as well as scores for the 2006–2007 GE classes. These scores were used to investigate predictive validity.

## Data analysis and results

The scales in Table 1 were used in interpreting the means of the Likert scale items.

TABLE 1
Interpreting Likert scale responses

| Mean | Positively oriented questions | | Negatively oriented questions | |
|---|---|---|---|---|
| | Degree | Opinion | Degree | Opinion |
| 4.5–5 | Very high | Strongly agree | Very low | Strongly disagree |
| 3.5–4.4 | High | Agree | Low | Disagree |
| 2.5–3.4 | Moderate | Undecided | Moderate | Undecided |
| 1.5–2.4 | Low | Disagree | High | Agree |
| 1.0–1.4 | Very low | Strongly disagree | Very high | Strongly agree |

## Face validity

Face validity was determined by asking both instructors and students about how well they felt the contents of the courses were represented on the achievement tests (as in Ösken (op.cit.) and Serpil (op.cit.)). The means of both groups' responses to these questions can be seen in Table 2. The means for Q12 indicate that both the instructors and the students agree that the content of the courses is adequately represented on the tests, and, looking at the individual questions, we can see that, for many of the questions, both teachers and students respond positively. However, the groups appear to disagree on Q5 and Q6, an expected result, since, as noted previously, speaking and video courses are essentially unrepresented on the exams. Listening and laboratory courses are also not represented on the tests, and for the instructors, the means for the relevant questions, Q9 and Q10, fall into the 'undecided' range. It is unclear why the

mean for Q9 is not lower, as listening is not represented on the test at all, but it may be that some instructors believe that testing listening is not necessary in this setting. The students, on the other hand, do appear to believe that listening should be tested, given their negative response to Q9. Like the teachers, they are undecided about whether the content of the laboratory courses is sufficiently represented. Both teachers and students are also undecided about Q4, which relates to the content of the reading courses. Overall, even though it appears that there are some areas of the course that are under-represented on the tests, it can be concluded that the instructors and students see the tests as possessing a high degree of face validity.

| Questions | Instructors mean | Students mean |
|---|---|---|
| Q1 The content of the main course book 'Quartet' was represented in the exams sufficiently. | 4.310 | 4.135 |
| Q2 The content of the grammar book 'Milestones' was represented in the exams sufficiently. | 4.000 | 4.039 |
| Q3 The content of the writing courses was represented in the exams sufficiently. | 3.862 | 3.557 |
| Q4 The content of the reading courses was represented in the exams sufficiently. | 3.310 | 3.019 |
| Q5 The content of the speaking courses was represented in the exams sufficiently. | 2.276 | 2.000 |
| Q6 The content of the video courses was represented in the exams sufficiently. | 2.138 | 1.942 |
| Q7 Grammar taught in the courses was represented in the exams sufficiently. | 4.483 | 4.654 |
| Q8 The vocabulary taught in the courses was represented in the exams sufficiently. | 4.069 | 4.365 |
| Q9 The listening practices focused on in the courses were represented in the exams sufficiently. | 2.517 | 2.077 |
| Q10 The content of the laboratory courses was represented in the exams sufficiently. | 2.586 | 2.328 |
| Q11 The exercises made in the courses were represented in the exams sufficiently. | 3.965 | 3.596 |
| Q12 In general, the contents of the courses were represented in the exams sufficiently. | 3.655 | 4.135 |

TABLE 2
Perceptions of face validity

## Reliability

Previous studies have looked at scoring reliability issues either by observing scoring practices (Brown 2003) or by examining the scores themselves (Manalo and Wolfe 2000), and no studies have examined reliability from the test takers' perspective. Reliability in this study was measured by asking specific questions of the instructors and students regarding the administration and scoring of the tests. The means of the instructors' responses to the questions about scorers' reliability are presented in Table 3. The mean for Q13, the instructors' overall perception of scorers' reliability, falls into the range of 'agree'. This indicates that, in the eyes of the instructors, the degree of scorer's reliability is relatively high. However, despite this overall impression, the means

for several questions indicate some potential problems in scorers' reliability. The instructors do not agree that students are identified by number rather than name for subjective scoring, and they are undecided as to whether more than one scorer was used for such scoring, whether all scorers are trained, and whether their colleagues scored the exams reliably.

| Questions | Mean |
|---|---|
| **Q1** The questions included in the exams permitted objective scoring. | 3.724 |
| **Q2** Testing office provided a detailed answer key. | 4.138 |
| **Q3** The scorers who marked the exam papers were trained. | 3.344 |
| **Q4** Students were identified by number, not name when scoring was subjective (e.g. in writing sections) to provide objectivity. | 2.448 |
| **Q5**$^{+}$ Only one instructor scored each exam paper when scoring was subjective. | 3.448$^{+}$ |
| **Q6** The rating scales included in the key helped me while I was scoring the exam papers | 3.828 |
| **Q7** We had meetings to agree with acceptable answers after the exams. | 4.448 |
| **Q8** The class which I instructed as the main course teacher and the class which I invigilated during the exams were two different classes. | 4.670 |
| **Q9** The class which I instructed as the main course teacher and the class whose papers I scored were two different classes. | 4.670 |
| **Q10**$^{+}$ The deadline for scoring and returning the exam papers to the main course instructors affected my scoring practices negatively. | 3.757$^{+}$ |
| **Q11** I scored the exam papers in a reliable manner. | 4.621 |
| **Q12** All my colleagues scored the exam papers in a reliable manner. | 3.445 |
| **Q13** In general, the scoring system was reliable. | 3.586 |

TABLE 3
Scorers' perceptions of reliability

+ Indicates negatively oriented items.

In Table 4, the means of the students' responses to questions about test structure are given. The mean for Q21, which asks whether, overall, the test structure had a negative impact on performance, is 3.596, indicating disagreement (on the negative-orientation rating scale). In other words, in the eyes of the students, it appears that the degree of reliability in terms of the test structure is relatively high. The students' responses to individual questions give greater insight into specific aspects of test structure. Of the four negatively oriented questions (Q1, Q2, Q3, and Q8), the means of Q3 and Q8 indicate disagreement, pointing towards test structure reliability in terms of the number of questions and difficulty level of the questions. The means of Q1 and Q2, however, indicate that the students are undecided as to the independent nature of the test questions, as well as whether there are too many questions. The means of the remaining, positively oriented questions all indicate agreement. Thus, it can be concluded that, according to the students, the test structure contributes to reliability.

| Questions | Mean |
|---|---|
| **Q1**$^{+}$ Sometimes, two (or more) questions in the test seemed to be closely related, so that if I could not answer one question, I could not answer the other question either. | 2.846$^{+}$ |
| **Q2**$^{+}$ The exams included too many questions. | 3.385$^{+}$ |
| **Q3**$^{+}$ The exams included an insufficient number of questions. | 3.904$^{+}$ |
| **Q4** The instructions explaining what to do in each section in the exams were explicit and clear. | 3.769 |
| **Q5** The points allotted for each section of the exam were always stated in the exam papers. | 4.692 |
| **Q6** Time given to the students to complete the exam was always stated in the exam papers. | 4.556 |
| **Q8**$^{+}$ All the questions in the exams had the same difficulty level. | 3.654$^{+}$ |
| **Q9** The exam questions were explicit and clear. | 3.692 |
| **Q10** The layout of the exam papers was fine. | 4.365 |
| **Q11** The exam papers were legible. | 4.556 |
| **Q12** The tables which were employed in the exams were clear and easy to interpret. | 4.231 |
| **Q21**$^{+}$ In general, the structure of the tests hindered my ability to display my best performance in the exams. | 3.596$^{+}$ |

TABLE 4
Students' perceptions of test structure

+ Indicates negatively oriented items.

Table 5 presents the means of the students' responses to questions about testing conditions. The mean for Q22, which asks whether, overall, the testing conditions had a negative effect on the students' performance, indicates disagreement (on the negative-orientation rating scale). Questions 14 and 15 asked about the amount of time given for the test; the students do not agree that the time given is too short, but they are undecided about whether too much time is given. They also do not agree that the light, temperature, or ventilation in the testing environment hinders their performance, but they are undecided about the effect of noise. The means for the remaining questions indicate agreement. Overall, it appears that, in the eyes of the students, the degree of reliability in terms of the testing conditions is relatively high.

| Questions | Mean |
|---|---|
| **Q7** Information about how much the given tests would affect the final grade was always announced. | 3.519 |
| **Q13** The instructors helped us to get used to the format of the exams. | 4.039 |
| **Q14**$^{+}$ The time given to complete the exams was too short. | 3.865$^{+}$ |
| **Q15**$^{+}$ The time given to complete the exams was too long. | 3.308$^{+}$ |
| **Q16** Equal timing was given to all classes which took the same test. | 4.558 |
| **Q17**$^{+}$ Distracting sounds and noises lowered my performance in the exams. | 3.077$^{+}$ |
| **Q18**$^{+}$ The little amount of light in the classrooms lowered my performance in the exams. | 4.096$^{+}$ |
| **Q19**$^{+}$ The degree of the temperature in the classrooms lowered my performance in the exams. | 3.750$^{+}$ |
| **Q20**$^{+}$ The little amount of air in the classrooms lowered my performance in the exams. | 3.635$^{+}$ |
| **Q22**$^{+}$ In general, the bad environmental conditions hindered my ability to display my best performance in the exams. | 3.923$^{+}$ |

TABLE 5
Students' perceptions of testing conditions

+ Indicates negatively oriented items.

## Predictive validity

Predictive validity was estimated by correlating the scores of the PS's midterm achievement tests with those of the final exam, as well as by correlating the final exam scores with grades from GE classes, taken the year following PS. The scores of 365 PS students' midterm achievement tests, two in each term, were correlated with their final exam scores (see Table 6). Because the scores were not distributed normally, Spearman's rho was calculated. Significant positive correlations can be seen in all cases, indicating that performance on the midterm achievement tests can fairly accurately predict performance on the final exam.

TABLE 6
Correlations among PS test scores

| | PS first term | | PS second term | |
|---|---|---|---|---|
| | 1st midterm | 2nd midterm | 1st midterm | 2nd midterm |
| PS final exam | .713* | .762* | .766* | .801* |

* Spearman's rho, one-tailed, $n = 365$, $p < .01$.

To determine the ability of the PS final exam to predict success after PS, the first and second term GE midterm and final exam scores and final grades of the same students were obtained and correlated with their PS final exam scores (see Table 7). Once again, Spearman's rho was used.

TABLE 7
Correlations between PS final exam and GE course grades

| | GE first term | | | GE second term | | |
|---|---|---|---|---|---|---|
| | Midterm | Final exam | Final grade | Midterm | Final exam | Final grade |
| PS final exam | .621* | .248* | .572* | .543* | .460* | .545* |

* Spearman's rho, one-tailed, $n = 365$, $p < .01$.

Table 7 shows significant positive correlations between the PS final exam and the exam scores and final grades given for the GE classes, although these correlations are not as strong as those seen in Table 5. The highest correlation is seen with the first midterm, and the lowest is seen with the first final exam. However, the correlations with the final grades for each term are somewhat similar, and they indicate a moderate degree of correlation between the PS final exam and the final grades for GE classes. The lower correlation coefficients, as well as the variation among them, may be explained by the fact that, although the same textbook is used in all GE classes, each teacher constructs and administers his/her own tests.

## Face validity, reliability, and predictive validity

Table 8 presents a summary of all analyses conducted in this study, to facilitate the consideration of how well face validity reflects more objective measures of the worth of tests. It should be noted that questions about scoring, test structure, and testing conditions that might have reflected the scorers' or test takers' personal opinions have been eliminated from this summary.

| | Face validity | Reliability | | | Predictive validity | |
|---|---|---|---|---|---|---|
| | | Scoring | Test structure | Test conditions | PS midterms | PS final exam |
| Instructors | 3.655 (mean Q12) | 3.85 (mean Q1–10) | | | .713[a] | .621[e] |
| Students | 4.135 (mean Q12) | | 4.04 (mean Q1, 4–6, 8–12) | 3.81 (mean Q7, 13, 16–20) | .762[b] | .248[f] |
| | | | | | .766[c] .801[d] | .572[g] .543[h] .460[i] .545[j] |

TABLE 8
Summary of analyses

Scale ranges: 4.5–5 = very high; 3.5–4.4 = high; 2.5–3.4 = moderate; 1.5–2.4 = low; 1–1.4 = low. a = 1st term, 1st midterm; b = 1st term, 2nd midterm; c = 2nd term, 1st midterm; d = 2nd term, 2nd midterm, all correlated with PS final exam and significant at $p < .01$. e = 1st term GE class midterm; f = 1st term GE final exam; g = 1st term GE final grade; h = 2nd term GE midterm; i = 2nd term GE final exam; j = 2nd term GE final grade, all correlated with PS final exam and significant at $p < .01$.

We can see that the face validity of the achievement tests is high and reflects the high level of reliability indicated by the means of the questions asking about specific, observable aspects of scoring, test structure, and testing conditions, all essential ingredients of a test's reliability (Brown 2004; Genesee and Upshur op.cit.; Hughes op.cit.). In addition, the strong positive correlations obtained between the PS midterm tests and the final exam indicate a high level of predictive validity of the midterm exams. More moderate correlations are obtained between the PS final exam and GE exams and final grades, but this may be explained by the less standardized nature of the tests given in the GE exams. It appears that the PS final exam has at least some predictive ability for future performance in the following English class. Thus, we can conclude that, at least for this set of tests, the assessment of face validity fairly accurately reflects more objective measures of test quality.

## Conclusion

The most immediate pedagogical implications drawn from the study largely concern the curriculum unit and testing office of the ZKU Preparatory School. While investigating face validity and reliability, several potential weaknesses in the tests, test administration, and scoring have emerged. It appears that several of the classes are under-represented on the achievement tests. This lack of representation may result from a lack of clear, well-defined objectives, since clear objectives help test writers determine which language points to give weight to on achievement tests and help teachers decide what should be taught. The recently opened curriculum unit, responsible for overseeing the development and delivery of the curriculum, may wish to initiate the process of establishing clear goals and objectives and ensuring that they are understood by the teachers and reflected in the curriculum. In turn, the testing office should examine the extent to which these goals and

objectives are represented in the achievement tests and examine the conditions under which tests are administered and scored.

Another immediate implication for the ZKU Preparatory School to emerge from this study is that, since the predictive validity of the midterm achievement tests is relatively high, these scores can be employed to predict students' achievement on the final exam and to identify those students in danger of failing. Such students might be offered extra assistance and support, to improve their chances of success. There is also some evidence that many students who score poorly on the PS final exam will also perform poorly in GE classes the next year. Indeed, 76 per cent of those scoring below the mean on the preparatory final exam also received a final grade that was below the mean in the first semester of the GE class. Such a trend underscores the need to provide extra assistance to students who perform poorly in PS, to ensure that their poor performance does not continue in subsequent English classes.

Several implications for other institutions can be drawn from the current study. The questionnaires employed in this study might be used in other institutions to assess their own achievement tests. Furthermore, the study illustrates a process by which other institutions may examine the appropriateness of such achievement tests.

Finally, with regard to the relationship between face validity and more objective measures, it was shown that the face validity of these achievement tests accurately reflects the aspects of reliability measured in this study, as well as the predictive validity. This might imply that administrators and testing officials may rely on perceptions of face validity in determining the worth of a test. However, the face validity and reliability analyses revealed some important weaknesses in the testing system. These weaknesses would not have been revealed if the researcher had looked at only face validity, only reliability, or only predictive validity. Thus, this study has revealed the importance of looking at tests from multiple perspectives, in order to get information from a variety of sources.

This study has several limitations. First, the reliability of the achievement tests was determined through perceptions of instructors and students, rather than by direct measurement. However, the specific and observable nature of many of the aspects addressed on the questionnaires was felt to be a relatively objective way of examining these aspects of reliability. It is also felt that relying on the observations and experiences of those directly involved in the testing process contributed to the manageable nature of the test assessment process described here. A related limitation concerns the length of time between the actual testing and the administration of the questionnaires. Due to the need for information on tests administered throughout the year, it was only possible to involve students from the previous year. It is possible that their memories of the testing experience were inaccurate. Another study of this type might follow students throughout the year, with data collected at each test administration. Another limitation might be the lack of any objective assessment of content validity. While determining content validity is a valuable although time-consuming and labour-intensive undertaking, the process described here is believed to be more manageable for teachers, testing officials, and administrators

wishing to assess the quality of their tests. Moreover, this process has the advantage of revealing the strengths and weaknesses of the testing system of an institution, which may lead to more in-depth evaluations, including assessment of content validity.

*Final revised version received July 2008*

## References

**Bachman, L. F.** 1990. *Fundamental Considerations in Language Testing*. New York: Oxford University Press.

**Bachman, L. F.** and **A. S. Palmer.** 1996. *Designing and Developing Useful Language Tests*. New York: Oxford University Press.

**Brown, A.** 2003. 'Interviewer variation and the co-construction of speaking proficiency'. *Language Testing* 20: 1–25.

**Brown, H. D.** 2004. *Language Assessment: Principles and Classroom Practices*. New York: Longman.

**Brown, J. D.** 1996. *Testing in Language Programs*. Upper Saddle River, NJ: Prentice Hall Regents.

**Cardoso, R. M. F.** 1998. 'Authentic foreign language testing in a Brazilian university entrance exam'. *Texas Papers in Foreign Language Education* 3/2: 51–70.

**Genesee, F.** and **J. A. Upshur.** 1996. *Classroom-Based Evaluation in Second Language Education*. Cambridge: Cambridge University Press.

**Hughes, A.** 2003. *Testing for Language Teachers* (2nd edn.). Cambridge: Cambridge University Press.

**Manalo, J. R.** and **E. W. Wolfe.** 2000. 'The impact of composition medium on essay raters in foreign language testing'. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, USA. ERIC document reproduction service no. ED443836.

**Nakamura, Y.** 2006. 'Analysis of a placement test: an interim report of a pilot version'. *Hiyoshi Keio University Bulletin* 37: 81–91.

**Ösken, H.** 1999. 'An assessment of the validity of the midterm and the end of course assessment tests administered at Hacettepe University Department of Basic English'. Unpublished MA thesis, Bilkent University.

**Rudner, L. M.** 1994. 'Questions to ask when evaluating tests'. *Practical Assessment, Research and Education* 4/2. Available at http://pareonline.net/getvn.asp?v=4&n=2.

**Serpil, H.** 2000. 'An assessment of the content validity of the mid-term achievement tests administered at Anadolu University Foreign Languages Department'. Unpublished MA thesis, Bilkent University.

**Yeğin, O. P.** 2003. 'The predictive validity of Başkent University proficiency exam (BUEPE) through the use of the three-parameter IRT model's ability estimates'. Unpublished MA thesis, Middle East Technical University.

## The authors

**Funda Küçük** graduated from Hacettepe University English Translation and Interpretation Department in 2000. She is also a Bilkent University MA/TEFL graduate. She has been an English instructor for eight years and is currently teaching at Zonguldak Karaelmas University in Turkey. She has published a grammar book with her colleagues from the same university, namely *Milestones of English Grammar—Perfecting and Practicing English Structure*.

**Email: fundak79@yahoo.com**

**JoDee Walters** completed her MA/TEFL at the American University in Cairo in 1993, and her PhD in Applied Linguistics at the University of Nottingham, UK, in 2006. She has taught English in university and secondary school settings in Egypt, the US, and the UK, and was involved in teacher training in the US, the UK, and Japan. She joined the MA/TEFL programme at Bilkent University in 2006.

**Email: walters@bilkent.edu.tr**