Development and Validation of an English Language Teacher Competency Test Using Item Response Theory

Reza Pishghadam
Ferdowsi University of Mashhad, Iran

Purya Baghaei
Islamic Azad University, Mashhad Branch

Hesamoddin Shahriari Ahmadi
Ferdowsi University of Mashhad, Iran

Abstract

The major aim of this study was to design and validate a competency test measuring English language teachers` knowledge for the purpose of teaching in language institutes. To this end, initially based on the guidelines laid down by eminent scholars in the field of second language teaching, a test was designed. The test was then administered among 103 EFL teachers in language institutes. And finally, Rasch measurement was utilized to substantiate the construct validity of the test. The results of the Rasch analysis exhibited that all of the items met the criteria for fit to the Rasch model and no potential multidimensionality was pinpointed. Therefore, the scale constructed can be considered as a Rasch unidimensional scale.

Keywords: Rasch, Teacher competency test, Teacher education, Validation

Introduction

As agents of social change and cohesion, teachers are considered to be at the heart of any educational system. In fact, the efficiency of any educational system is indebted to its knowledgeable and effective teachers. Sanders and Rivers (1996) regard teachers as the single most important factors affecting student achievement. Suwandee (1995) considers teaching as a two-way interaction between teachers and students in which students' knowledge depends largely on the degree of commitment to the teaching task on the part of the teacher.

It goes without saying that in the field of education, there is an urgent need for effective teachers. Research has shown that students taught by effective teachers achieve more than their peers who are taught by less effective instructors (Sanders & Horn, 1998). Hanushek (1992) discovered that students whose teachers are at the peak point of the curve of effectiveness gain 1.5 years of academic growth, compared to those who are taught by teachers at the bottom of the range. The latter groups only achieve 0.5 years of growth within a single year.

Certifications and degrees can be seen as only one index of teacher effectiveness. Teacher efficacy is in fact controlled by multiple variables, such as teaching methods, behavior towards student learning, mastery of competencies, professional decision-making and the interaction between pedagogical and subject area knowledge (Lederman & Niess, 2001).

These days, it is quite common for employers to review employment applications and resumes and then conduct an interview with the purpose of final screening. The decisions made based on these interview sessions may be influenced by various factors. These include affective characteristics, anticipated responses, environmental conditions, interviewer's research knowledge and the like (Stronge & Hindman, 2006). Variables of this kind could undermine the validity and fairness of interviews as a means of teacher recruitment.

Given the shortcomings of screening interviews, competency tests in the field of education could be used as an alternative route to selection. Teacher Competency Tests (TCT) are generally used whenever a recruiter wishes to screen candidates for a job opening. Through such tests, candidates must prove that they possess the necessary qualifications for the job. In other words, TCTs help to identify the best human resources for the most suitable positions.

In Iran, in order to meet the demands of the large body of clients, language schools, chiefly functioning within the private sector, tend to recruit teachers with adequate levels of English language proficiency, regardless of their teaching credentials. Four possible explanations could be proposed for this lack of professionalism in the field. First of all, given the considerable number of students referring to language institutes, managers feel the need to hire teachers in as short a time and with as little monetary investment as possible. Therefore, the only prerequisite considered for potential applicants is their language proficiency. Second, due to the shortage of specialized human resources in English language teaching, the mangers in the aforementioned institutes try to recruit individuals who come from a variety of backgrounds, holding degrees in English translation, English literature or even other non-English-related disciplines. Third, some applicants who already hold a degree in English language teaching do not necessarily have the minimum level of proficiency to measure up to their professional training. And finally, there is no standard comprehensive exam on English language teaching criteria based on which the managers can employ the qualified teachers. Thus, it seems that there is a dire need for designing a TCT.

To the best knowledge of the researchers, no test has been designed to measure the competency knowledge of English language teachers to date. Accordingly, the present study aims to design and validate an English Language Teacher Competency Test (ELT-CT).

Competency Testing

TCTs provide the opportunity to improve the quality of education. Such tests screen out incompetent teachers and ensure that higher quality candidates ultimately obtain teaching positions. This, in turn, leads to the heightening of standards within the educational system. According to Haefele (1993), competency tests can provide constructive feedback for individual educators, as well as providing direction for staff development practices.

Reports have also revealed that the public has reacted positively to the prospect of teacher competency testing. For instance, the American public viewed such tests as positive strategies fighting against the problems of the educational system (Anderson, 1987). Gallup polls (1984), for instance, have shown that 89 percent of the public and 63 percent of teachers hold the opinion that it should be necessary for teachers to prove their knowledge through teacher competency tests. As a result of such tests, students will also adopt a more positive view towards their teachers and the whole instructional program.

Teachers who have undergone mandatory competency tests also appreciate the emphasis that is placed on the quality of education and professionalism as a whole. They will also be aware that their work is scrutinized by principals and educational managers. Tucker and Stronge (2005) refer to this as the accountability function, which reflects a commitment to the important professional goals of competence and the quality of performance.

Qualities of Effective Teachers

Defining quality is inextricably intertwined with value judgments, and such judgments vary from one particular context and culture to another (Alexander, 2000). Despite the difficulty involved in determining a finite set of characteristics for an effective teacher, a set of variables have been identified, and are often included in competency tests. These variables can be divided into prerequisites, personality and classroom management variables.

The most prominent prerequisites include verbal ability, content knowledge, education coursework, teacher certification and teaching experience (Stronge & Hiundman, 2006). Each of the prerequisites will be briefly discussed below:

Since teachers use language to establish connections with their students, a teacher's verbal ability is believed to have an effect on student achievement. Verbal ability is of great importance since the ability to communicate content knowledge and belief in students is crucial to teaching and learning (Darling-Hammond, 2000; Hanushek, 1971).

It is also vital for teachers to be equipped with the content knowledge of the area within which they are teaching. A study carried out in the state of California found that mathematics teachers who majored or minored in mathematics raised students with higher test scores on the Standford 9 Achievement Test (Fetler, 1999).

A third set of teacher prerequisites include teacher certification. Research has shown that teachers working in the area in which they are certified wield more influence over student learning than their uncertified counterparts (Darling-Hammond, 2000; Darling-Hammond, Berry, & Thoreson, 2001; Goldhaber & Brewer, 2000; Hawk, Coble & Swanson, 1985).

Finally, teaching experience, as another prerequisite, is believed to have a profound effect on the effectiveness of instruction. Experienced teachers possess an increased depth of perception with regards to the content and how it is presented it in class (Covino & Iwanicki, 1996). Experienced teachers also make use of a wider array of strategies, which renders them more effective with students (Glass, 2001). It should be noted, however, that it is by no means true that teachers with more years of teaching experience are necessarily better. Sanders and Rivers (1996) have discovered that a teacher's effectiveness increases within the first seven years of teaching and reaches a plateau by the 10th year.

The personality variables of a teacher are important in making students feel comfortable in the teaching environment. The personal connection established between a teacher and her students fosters the creation of a trusting and respectful relationship (Marzano, Pickering & McTighe, 1993). Effective teachers have been shown to be caring, enthusiastic, motivated, fair, respectful, reflective, dedicated, and with a good sense of humor (Black & Howard-Jones, 2000; National Association of School Principals, 1997; Peart & Campbell, 1999).

Classroom management and organization constitute the last set of variables possessed by effective teachers. There are generally fewer disruptions and off-task behavior in classes taught by effective teachers (Stronge & Hindman, 2006). When problems associated with discipline arise, effective teachers respond in a predictable manner. Effective teachers also recapture instructional time that is often lost in administrative activities, discipline, and transitions (Hoy & Hoy, 2003).

Method

Participants

Two groups of participants cooperated towards the completion of this study. The first group consisted of 10 teachers of English as a Foreign Language (EFL) with more than 5 years of teaching experience in private language institutes. Members of this first group, who participated in the piloting of the test, were all males and had at least six years of language teaching experience.

The second group of participants, who took part in the administration of the test, consisted of 103 individuals, all of whom were practicing language teachers at various language institutes in the city of Mashhad, Iran. Their level of experience ranged from one year to 15 years of teaching practice. The youngest member of the target population was 17 and the oldest was 48 years old. Most of the respondents held a bachelor's degree (63), 32 had a master's degree, and only two held a PhD. With regards to their major, 44 of the participants majored in TEFL, 29 were educated in the field of English literature, and 17 had a degree in translation.

The participants were asked to respond to a test of teacher competency with 61 items. They were provided with clear and specific instructions as to how to complete the test. The test was administered under standardized conditions.

Instrument

The instrument includes a TCT which was developed and validated for the purpose of the study. The test was supposed to measure the English language teaching competency. The test consists of 61 items including:

a. Items corresponding to the teaching of skills
b. Items corresponding to the process of assessment and testing
c. Items related to the theories of first and second language acquisition
d. Items related to teacher behavior within the classroom

In the following section we have discussed the development and validation of the test.

Procedure

The present study involved the designing and administration of a test for evaluating competence in language instruction. The test was designed over the course of a year, starting from October 2008 to November 2009. Following the guidelines laid down by the experts in the field of second language learning and teaching, a number of items were designed in multiple-choice format. The items aimed to be functional in nature and present respondents with clearly-defined situations which they had possibly encountered while teaching in their own classes. Measures were taken to avoid the inclusion of items dealing with the theoretical knowledge or beliefs of the respondents.

Questions for the test were continually developed and revised until eventually a set of approximately 70 items were achieved. Following this stage, the approved items were given to 10 experienced language teachers who were asked to specify their responses while thinking aloud. This was done in the presence of the researchers who recorded the comments for further revision. Comments which were provided by more than one of the teachers were then considered more carefully. Those items which were regarded to be unclear or ambiguous were either dropped or revised. This piloting stage resulted in a refined version of the test which included 61 items.

The second-draft version of the test was then administered to 103 participants who were given clear instructions and ample time to complete the test. Test takers were told to select the choice which they either believed to be the best course of action or that they actually practiced in their classes. All test takers were reminded that the context in mind for the items was a typical class held at a private language institute.

Data Analysis

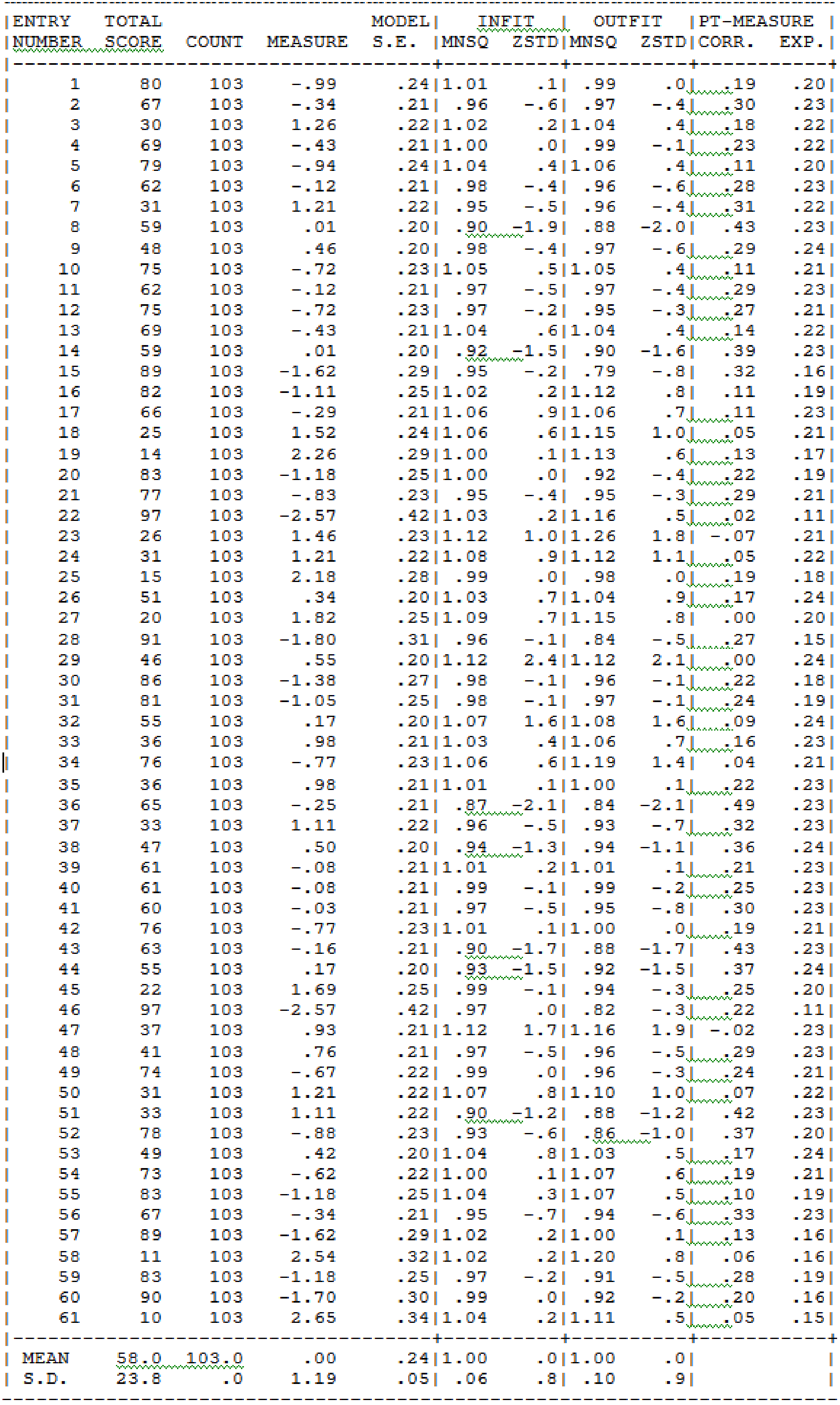
Rasch measurement was utilized to substantiate the construct validity of the ELT-CT. Rasch analysis was conducted using Winsteps version 3.66. The entire dataset with 61 items and 103 persons was subjected to Rasch analysis to evaluate the fit of data to the model and assess the unidimensionality of the instrument. If these tests are satisfied and the assumptions hold, the scale is a unidimensional Rasch scale and persons and items can be located on an interval scale.

Results

As the results of fit statistics show, all items fit the Rasch model following the criteria suggested by Bond and Fox (2007). Items which do not fit the Rasch model have outfit and infit mean square (MNSQ) indices outside the acceptable range of 0.70-1.30. Misfitting items are signs of multidimensionality and model deviance. As Table 1 shows, none of the items have infit and outfit MNSQ indices outside the acceptable boundary.

Table 1

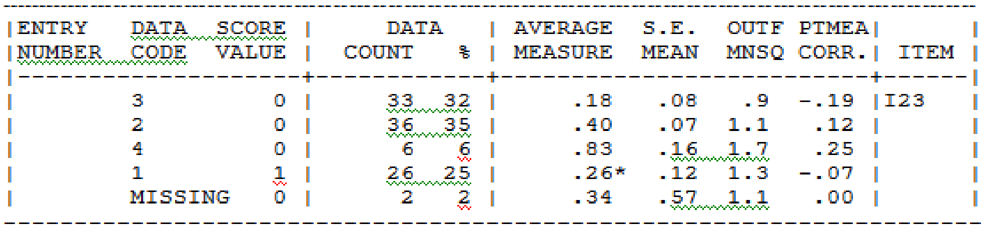
Item Statistics in Descending Order of Difficulty



```
-------------------------------------------------------------------------------------
|ENTRY    TOTAL                      MODEL|    INFIT   |   OUTFIT   |PT-MEASURE  |
|NUMBER    SCORE   COUNT   MEASURE     S.E. |MNSQ    ZSTD|MNSQ    ZSTD|CORR.   EXP.|
|------------------------------------------+------------+------------+------------+
|      1      80     103      -.99      .24|1.01      .1| .99      .0|  .19    .20|
|      2      67     103      -.34      .21| .96     -.6| .97     -.4|  .30    .23|
|      3      30     103      1.26      .22|1.02      .2|1.04      .4|  .18    .22|
|      4      69     103      -.43      .21|1.00      .0| .99     -.1|  .23    .22|
|      5      79     103      -.94      .24|1.04      .4|1.06      .4|  .11    .20|
|      6      62     103      -.12      .21| .98     -.4| .96     -.6|  .28    .23|
|      7      31     103      1.21      .22| .95     -.5| .96     -.4|  .31    .22|
|      8      59     103       .01      .20| .90    -1.9| .88    -2.0|  .43    .23|
|      9      48     103       .46      .20| .98     -.4| .97     -.6|  .29    .24|
|     10      75     103      -.72      .23|1.05      .5|1.05      .4|  .11    .21|
|     11      62     103      -.12      .21| .97     -.5| .97     -.4|  .29    .23|
|     12      75     103      -.72      .23| .97     -.2| .95     -.3|  .27    .21|
|     13      69     103      -.43      .21|1.04      .6|1.04      .4|  .14    .22|
|     14      59     103       .01      .20| .92    -1.5| .90    -1.6|  .39    .23|
|     15      89     103     -1.62      .29| .95     -.2| .79     -.8|  .32    .16|
|     16      82     103     -1.11      .25|1.02      .2|1.12      .8|  .11    .19|
|     17      66     103      -.29      .21|1.06      .9|1.06      .7|  .11    .23|
|     18      25     103      1.52      .24|1.06      .6|1.15     1.0|  .05    .21|
|     19      14     103      2.26      .29|1.00      .1|1.13      .6|  .13    .17|
|     20      83     103     -1.18      .25|1.00      .0| .92     -.4|  .22    .19|
|     21      77     103      -.83      .23| .95     -.4| .95     -.3|  .29    .21|
|     22      97     103     -2.57      .42|1.03      .2|1.16      .5|  .02    .11|
|     23      26     103      1.46      .23|1.12     1.0|1.26     1.8| -.07    .21|
|     24      31     103      1.21      .22|1.08      .9|1.12     1.1|  .05    .22|
|     25      15     103      2.18      .28| .99      .0| .98      .0|  .19    .18|
|     26      51     103       .34      .20|1.03      .7|1.04      .9|  .17    .24|
|     27      20     103      1.82      .25|1.09      .7|1.15      .8|  .00    .20|
|     28      91     103     -1.80      .31| .96     -.1| .84     -.5|  .27    .15|
|     29      46     103       .55      .20|1.12     2.4|1.12     2.1|  .00    .24|
|     30      86     103     -1.38      .27| .98     -.1| .96     -.1|  .22    .18|
|     31      81     103     -1.05      .25| .98     -.1| .97     -.1|  .24    .19|
|     32      55     103       .17      .20|1.07     1.6|1.08     1.6|  .09    .24|
|     33      36     103       .98      .21|1.03      .4|1.06      .7|  .16    .23|
|     34      76     103      -.77      .23|1.06      .6|1.19     1.4|  .04    .21|
|     35      36     103       .98      .21|1.01      .1|1.00      .1|  .22    .23|
|     36      65     103      -.25      .21| .87    -2.1| .84    -2.1|  .49    .23|
|     37      33     103      1.11      .22| .96     -.5| .93     -.7|  .32    .23|
|     38      47     103       .50      .20| .94    -1.3| .94    -1.1|  .36    .24|
|     39      61     103      -.08      .21|1.01      .2|1.01      .1|  .21    .23|
|     40      61     103      -.08      .21| .99     -.1| .99     -.2|  .25    .23|
|     41      60     103      -.03      .21| .97     -.5| .95     -.8|  .30    .23|
|     42      76     103      -.77      .23|1.01      .1|1.00      .0|  .19    .21|
|     43      63     103      -.16      .21| .90    -1.7| .88    -1.7|  .43    .23|
|     44      55     103       .17      .20| .93    -1.5| .92    -1.5|  .37    .24|
|     45      22     103      1.69      .25| .99     -.1| .94     -.3|  .25    .20|
|     46      97     103     -2.57      .42| .97      .0| .82     -.3|  .22    .11|
|     47      37     103       .93      .21|1.12     1.7|1.16     1.9| -.02    .23|
|     48      41     103       .76      .21| .97     -.5| .96     -.5|  .29    .23|
|     49      74     103      -.67      .22| .99      .0| .96     -.3|  .24    .21|
|     50      31     103      1.21      .22|1.07      .8|1.10     1.0|  .07    .22|
|     51      33     103      1.11      .22| .90    -1.2| .88    -1.2|  .42    .23|
|     52      78     103      -.88      .23| .93     -.6| .86    -1.0|  .37    .20|
|     53      49     103       .42      .20|1.04      .8|1.03      .5|  .17    .24|
|     54      73     103      -.62      .22|1.00      .1|1.07      .6|  .19    .21|
|     55      83     103     -1.18      .25|1.04      .3|1.07      .5|  .10    .19|
|     56      67     103      -.34      .21| .95     -.7| .94     -.6|  .33    .23|
|     57      89     103     -1.62      .29|1.02      .2|1.00      .1|  .13    .16|
|     58      11     103      2.54      .32|1.02      .2|1.20      .8|  .06    .16|
|     59      83     103     -1.18      .25| .97     -.2| .91     -.5|  .28    .19|
|     60      90     103     -1.70      .30| .99      .0| .92     -.2|  .20    .16|
|     61      10     103      2.65      .34|1.04      .2|1.11      .5|  .05    .15|
|------------------------------------------+------------+------------+------------+
| MEAN      58.0   103.0       .00      .24|1.00      .0|1.00      .0|            |
| S.D.      23.8      .0      1.19      .05| .06      .8| .10      .9|            |
-------------------------------------------------------------------------------------
```

Distracter analysis of items showed that item 23 has badly-written distracters. As shown in Table 2, the correct response for this item is option 1. The average measure of the respondents who have chosen this item is 0.26 logit. While the average measure of those who have chosen the incorrect options, 2 and 4 are 0.40 and 0.83, respectively. The average measure of those who have skipped this item is 0.34. These are all larger than the average measure of those who have answered the item correctly, which is an unexpected event. The distracter- measure correlation between wrong options and measures are expected to be negative and for the correct response to be positive which is violated here. Distracters for such items need to be modified and rewritten.

Table 2
Distracter analysis table

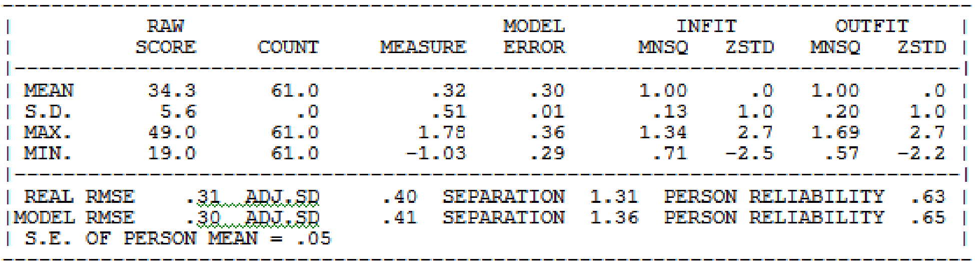


```
------------------------------------------------------------------------
|ENTRY    DATA  SCORE |     DATA     | AVERAGE  S.E.  OUTF PTMEA|      |
|NUMBER   CODE  VALUE |  COUNT     % | MEASURE  MEAN  MNSQ CORR.| ITEM |
|---------------------+--------------+--------------------------+------|
|         3        0  |    33   32   |    .18   .08    .9  -.19 |I23   |
|         2        0  |    36   35   |    .40   .07   1.1   .12 |      |
|         4        0  |     6    6   |    .83   .16   1.7   .25 |      |
|         1        1  |    26   25   |    .26*  .12   1.3  -.07 |      |
|         MISSING  0  |     2    2   |    .34   .57   1.1   .00 |      |
------------------------------------------------------------------------
```

The Cronbach's alpha reliability of the test is 0.64 which is moderate. This moderate reliability is due to the narrow spread of the persons in the analysis. As Table 3 shows the raw score standard deviation of the sample is only 5.6 out of 61, which is a very narrow spread of person abilities indeed. When the item with negative point-measure correlation indices are deleted from the test, the reliability increases to 0.68.

The separation index of the persons is 1.31, which translates to a person strata index of 2.70. Person strata index indicates the number of distinct ability levels which can be identified by the test (Stone & Wright, 1988; Wright & Stone, 1988). The minimum person strata index is 2, which means that the test is capable of distinguishing at least 2 strata of persons, namely, high-ability and low-ability persons. For a strata index of 2, a separation index of at least 1 is needed. A reliability index of at least 0.50 is required for a separation index of 1. It should be noted that the moderate reliability, separation and strata indices for this test is due to the low standard deviation of person abilities. If another sample with a wider spread of abilities were to be tested, these statistics would improve.

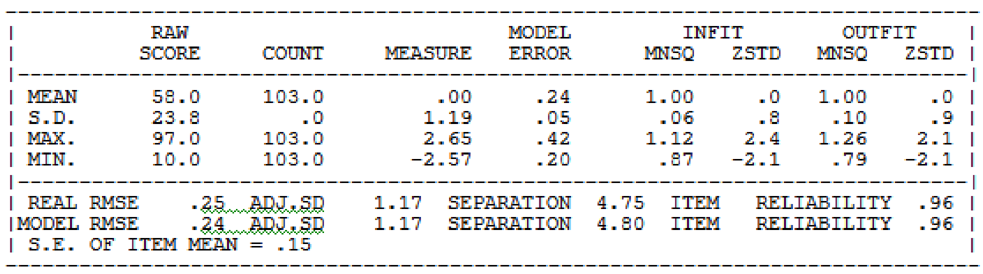
Table 3
Summary of 103 measured persons



```
-------------------------------------------------------------------------------
|          RAW                          MODEL         INFIT        OUTFIT     |
|         SCORE     COUNT     MEASURE   ERROR      MNSQ   ZSTD   MNSQ   ZSTD  |
|-----------------------------------------------------------------------------|
| MEAN     34.3      61.0         .32     .30      1.00     .0   1.00     .0  |
| S.D.      5.6        .0         .51     .01       .13    1.0    .20    1.0  |
| MAX.     49.0      61.0        1.78     .36      1.34    2.7   1.69    2.7  |
| MIN.     19.0      61.0       -1.03     .29       .71   -2.5    .57   -2.2  |
|-----------------------------------------------------------------------------|
| REAL RMSE    .31  ADJ.SD    .40  SEPARATION  1.31  PERSON RELIABILITY  .63  |
|MODEL RMSE    .30  ADJ.SD    .41  SEPARATION  1.36  PERSON RELIABILITY  .65  |
| S.E. OF PERSON MEAN = .05                                                   |
-------------------------------------------------------------------------------
```

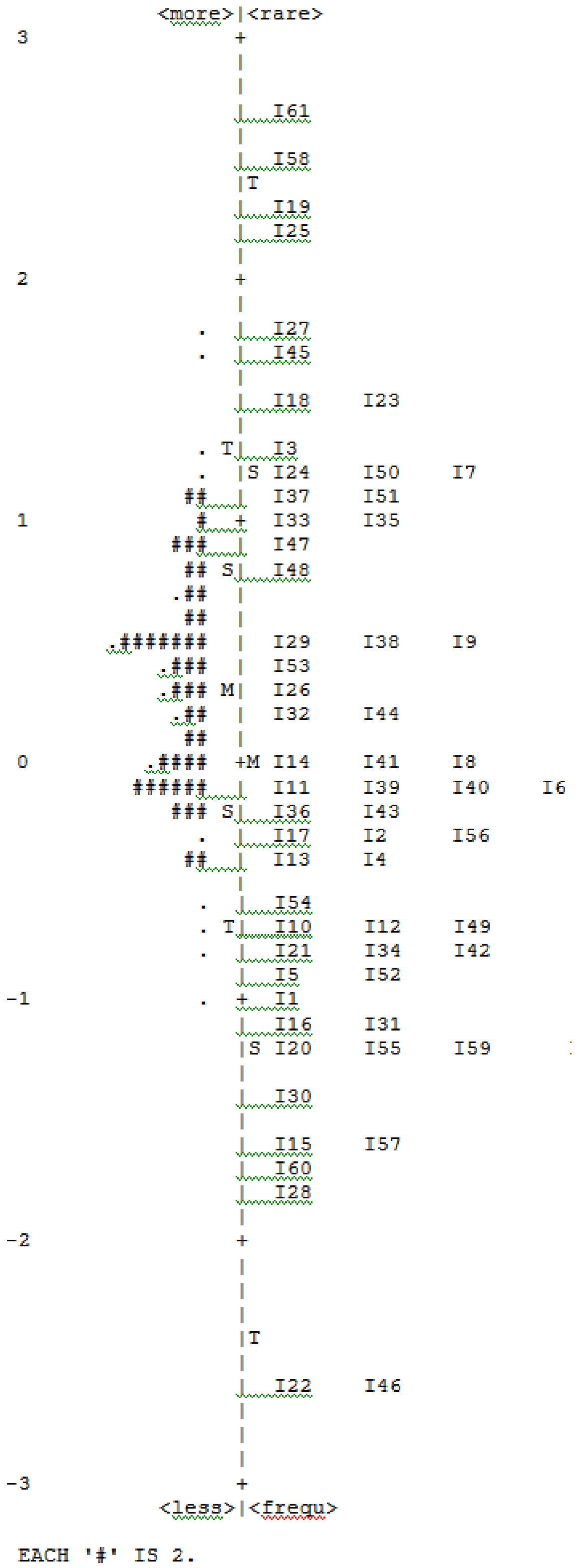
As Table 4 demonstrates, the reliability for the items is very good. That is, the chances that the difficulty ordering of the items be repeated if the test were given to another group is extremely high. This is because there is a wide spread of difficulty in the items as the standard deviation of item difficulty estimates is 1.19 logits and the separation is 4.75.

Table 4
Summary of 61 measured items



```
 -------------------------------------------------------------------------------
|          RAW                              MODEL         INFIT        OUTFIT    |
|          SCORE      COUNT     MEASURE     ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|-------------------------------------------------------------------------------|
| MEAN      58.0     103.0         .00       .24      1.00     .0   1.00     .0 |
| S.D.      23.8        .0        1.19       .05       .06     .8    .10     .9 |
| MAX.      97.0     103.0        2.65       .42      1.12    2.4   1.26    2.1 |
| MIN.      10.0     103.0       -2.57       .20       .87   -2.1    .79   -2.1 |
|-------------------------------------------------------------------------------|
| REAL RMSE    .25  ADJ.SD     1.17  SEPARATION  4.75  ITEM   RELIABILITY  .96 |
|MODEL RMSE    .24  ADJ.SD     1.17  SEPARATION  4.80  ITEM   RELIABILITY  .96 |
| S.E. OF ITEM MEAN = .15                                                       |
 -------------------------------------------------------------------------------
```

The Item-person map indicates that the items are spread over the entire range of the scale; i.e., all parts of the construct are well covered by the test. Numbers on the right indicate items and # on the left indicate persons. Items and persons placed on top of the scale are more difficult and more competent, respectively. As one goes down the scale, items become easier and individuals become less able. As one can see, all individuals are clustered towards the centre of the scale and the items are spread all over the scale. The map shows that there are enough items in the region of the scale where the persons lie and this part of the scale is pretty well covered by items. Therefore, the person abilities are estimated quite precisely as is evident from the low root mean square standard error of the persons which is 0.31. Therefore, the moderate reliability of the test is due to an actual homogeneity in the persons with respect to the construct of interest.



```
                 <more>|<rare>
   3                   +
                       |
                       |
                       |  I61
                       |
                       |  I58
                       |T
                       |  I19
                       |  I25
                       |
   2                   +
                       |
                    .  |  I27
                    .  |  I45
                       |
                       |  I18     I23
                       |
                    . T|  I3
                    .  |S I24     I50     I7
                   ##  |  I37     I51
   1                #  +  I33     I35
                  ###  |  I47
                   ## S|  I48
                  .##  |
                   ##  |
             .#######  |  I29     I38     I9
                 .###  |  I53
                 .### M|  I26
                  .##  |  I32     I44
                   ##  |
   0            .####  +M I14     I41     I8
               ######  |  I11     I39     I40     I6
                  ### S|  I36     I43
                    .  |  I17     I2      I56
                   ##  |  I13     I4
                       |
                    .  |  I54
                    . T|  I10     I12     I49
                    .  |  I21     I34     I42
                       |  I5      I52
  -1                .  +  I1
                       |  I16     I31
                       |S I20     I55     I59
                       |
                       |  I30
                       |
                       |  I15     I57
                       |  I60
                       |  I28
                       |
  -2                   +
                       |
                       |
                       |
                       |T
                       |
                       |  I22     I46
                       |
                       |
                       |
  -3                   +
                 <less>|<frequ>

   EACH '#' IS 2.
```

Figure 1. Items-persons map

Discussion

Since teaching quality and teacher competence are considered to be the most powerful predictors in student success (King Rice, 2003), it is of high priority for language schools and other educational institutions to assess these qualities when attempting to recruit effective teachers. With this purpose in mind, the present study was conducted to design and validate a test which sought to identify and assess the most influential characteristics of English language teachers in private language institutes. In order to achieve this aim, a thorough analysis of content and construct validity was executed.

The researchers in this study are fully aware that quality language instruction is a multifaceted, culturally-bound concept, widely debated in different academic circles. However, the immediate need for an objective criteria based on which the common core to all effective teaching can be assessed is felt. Without an objective scale, the element of professionalism in language instruction will remain unattainable. According to Danielson and McGreal (2000), a notable problem with existing scales is that they use a qualitatively dichotomous scale (e.g., 'satisfactory' and 'needs to be improved'). Such scales do not always accomplish what they have set out to, because there is little agreement on what it means to deliver 'outstanding' performance. Besides, one teacher's 'satisfactory' rating mat be equal to another's 'outstanding'. Given the current limitations, this study has taken rigorous statistical and operational measures to improve upon existing scales designed for the same purpose.

Rasch analysis was employed to assess the psychometric properties of the ELT-CT. Analyses revealed that all the items satisfied the criteria of fit to the Rasch model. Potential multidimensionality was not detected. Therefore, the scale constructed can be considered a Rasch unidimensional scale. In other words, all the items within the developed test contribute towards the definition of a unified construct of teacher competency.

The ability of the scale to discriminate among individuals at the two ends of the ability spectrum was also investigated. Results indicate a wide range of item difficulties (i.e., 2.65 to -2.57) indicating that the items on the instrument are spread wide enough to discriminate among individuals with low, middle and high levels of competency in language teaching. This means that the test can serve multiple purposes, being used both for the evaluation and selection of prospective language teachers, as well as for grading teachers currently engaged in the profession based on their level of expertise.

The results obtained as to the purpose of the study can hopefully be interpreted as having some implications for teachers and private institutions recruiting English language teachers in EFL contexts. First and foremost, as briefly mentioned above, the results of the study can be useful for both in-service and pre-service language teachers. Through the administration of this test, they become aware of the criteria which are influential in their success and effective teaching. That is to say, the test can also be employed as a tool for pedagogical purposes and raising language teacher awareness. What is more, through discussing the items and the elicited responses with their teachers, language school supervisors and planners can mark their criteria for effective instruction in a contextually-rich and dialogical manner. As a result, this awareness helps teachers to understand their students and try to meet their particular needs. Moreover, private agencies can employ the TCT to single out those individuals who have met certain specified qualifications for teaching English.

As in the case of any research, this study has its own set of limitations. First, the predictive validity of the ELT-CT has not been verified in this study. Future research needs to be done to investigate the relationship between the ELT-CT and the success of language teachers in institutes. Second, in this study the reliability was somehow moderate, which might be the result of homogenous grouping. A replication of this study with a sample of wider ability range and heterogeneity could be very informative. Third, since there is no agreement upon what competent teachers should know and which qualities are to be included into a competency test (Hyman, 1984), further studies can be done to reconsider the content validity of the test. And finally, the present study did not take a cutoff score into account. Additional research can be done to determine a cutoff score for the ELT-CT, identifying more competent individuals from less competent ones.

References

Alexander, R. (2000). *Culture and pedagogy: International comparisons in primary education.* Oxford, UK: Basil Blackwell.

Anderson, D. (1987). The pros and cons of competency testing for teachers. *Adventist Education, 50*(1), 23-38.

Black, R. S., & Howard-Jones, A. (2000). Reflections on best and worst teachers: An experiential perspective of teaching. *Journal of Research and Development in Education, 34*(1), 1–12.

Bond, T. G., & Fox, C. M. (2007) (2^{nd} ed.) *Applying the Rasch model: fundamental measurement in the human sciences.* Lawrence Erlbaum.

Covino, E. A., & Iwanicki, E. F. (1996). Experienced teachers: Their constructs of effective teaching. *Journal of Personnel Evaluation in Education, 10*(4), 325-363.

Danielson, C., & McGreal, T. L. (2000). *Teacher evaluation to enhance professional practice.* Princeton, NJ: Educational Testing Service.

Darling-Hammond, L. (2000). *Solving the dilemmas of teacher supply, demand and standards: How we can ensure a competent, caring and qualified teacher for every child.* New York: National Commission on Teaching and America's Future.

Darling-Hammonds, L., & Wise, A. E. (1983). Teaching standards or standardized teaching? *Educational leadership, 41*, 66-69.

Darling-Hammonds, L., Berry, B., & Thoreson, A. (2001). Does teacher certification matter? Evaluating the evidence. *Educational Evaluation and Policy Analysis, 23*(1), 57-77.

Educational Testing Service (ETS) (2009). Overview of PRAXIS III. Educational Testing Service, New York, *http://www.ets.org/portal/site/menuitem.1488512ecfdb8849a77b13bc3921509VgnlVCM10000022f95190RCRDdevgnextchannel=19b00ef8beb4b110VgnVCM10000022f95190RCRD,* accessed December 2009.

Felter, M. (1999). High school staff characteristics and mathematics test results. *Educational Policy Analysis Archives, 7*(9). Available: http://epaa.asu.edu/v7n9.html.

Gallup, G. H. (1984). The 16th annual Gallup poll of the public's attitudes toward the public schools. *Phi Delta Kappan 66*(1), 23-38.

Glass, C. S. (2001). Factors influencing teaching strategies used with children who display attention deficit hyperactivity disorder characteristics. *Education, 122*(1), 70-80.

Goldhaber, D. D., & Brewer, D. J. (2000). Does teacher certification matter? Highschool teacher certification status and student achievement. *Educational Evaluation and Policy Analysis, 22*(2), 129-145.

Haefele, D. L. (1993). Evaluating teachers: A call for change. *Journal of Personnel Evaluation in Education, 7,* 21-31.

Hanushek, E. (1971). Teacher characteristics and gains in student achievement: Estimation using micro data. *American Economic Review 61*(2), 280-288.

Hanushek, E. (1992). The trade-off between child quantity and quality. *Journal of Political Economy, 100*(1), 84-117.

Hawk, P. P., Coble, C. R., & Swanson, M. (1985). Certification: Does it matter? *Journal of Teacher Education, 26*(3), 13-15.

Hoy, A. W., & Hoy, W. K. (2003). *Instructional leadership: A learning-centered guide.* Boston: Allyn & Bacon.

Hyman, R. (1984). Testing for teacher competence: The logic, the law, and the implications. *Journal of teacher education, 35*, 14-18.

Lederman, N. G., & Niess, M. L. (2001). An attempt to anchor our moving targets. *School Science and Mathematics, 101*(2), 57-60.

Linacre, J.M. (2007) *A user's guide to WINSTEPS-MINISTEP: Rasch-model computer programs.* Chicago, IL: winsteps.com.

Marzano, R. J., Pickering, D., & McTighe, J. (1993). *Assessing student outcomes: Performance assessment using the dimensions of learning model.* Alexandria, VA: Association for Supervision and Curriculum Development.

Milanowski, A. (2004). The Criterion-related validity of the performance assessment system in Cincinatti. *Peabody Journal of Education, 79*(4). 33-53.

Milanowski, A., S. Kimball & A. Odden (2005). Teacher accountability measures and links to learning, In L. Stiefel, A. Schwartz, R. Rubenstein and J. Zabel (eds.), *Meaning school performance and efficiency: Implications for practice and research.* Larchmont, NY: Eye on education. Pp. 137-161.

Nassif, P. M. (1979). Setting standards, In *final program development resource document: A study of minimum competency testing programs.* Washington DC: National institute for education.

National Association of Secondary School Principals. (1997). Students say: What makes a good teacher? NASSP *Bulletin, 6*(5), 15–17.

National Board for Professional Teaching Standards (NBPTS). (2009). About the national board for professional teaching standards, Arlington, VA., http://www.nbpts.org/about_us, accessed December 2009.

National Research Council (2008). Assessing accomplished teaching: Advanced-level certification programs. In M. Hakel, J. Anderson Koenig and S. Elliott (eds.) *Committee on Evaluation of Teacher Certification by the National Board for Professional Teaching Standards,* NRC. Washington DC: National Academies Press.

Peart, N. A., & Campbell, F. A. (1999). At-risk students' perceptions of teacher effectiveness. *Journal for a Just and Caring Education, 5*(3), 269–284.

Pugach, M., & Raths, J. (1983). Testing teachers: Analysis and recommendations. *Journal of Teacher Education, 34,* 37-43.

Quirk, T. J., Witten, B. J., & Weinberg, S. F. (1973). Review of studies of concurrent and predictive validity of the NTE. *Review of Educational research, 43,* 89-113.

Sanders, W. L., & Horn, S. P. (1998). Research findings from the Tennessee Value-Added Assessment System (TVAAS) database: Implications for educational evaluation and research. *Journal of Personnel Evaluation in Education, 12*(3), 247-256.

Sanders, W. L., & Rivers, J. C. (1996). *Cumulative and residual effects of teachers on future students' academic achievement* (Research Progress Report). Knoxville, TN: University of Tennessee Value-Added Research and Assessment Center.

Stone, M., & Wright, B. (1988). *Separation statistics in Rasch measurement* (Research Memorandum No. 51). Chicago: MESA.

Stronge, J. H., & Hindman, J. L. (2006). *The teacher quality index: A protocol for teacher selection.* Alexandria, VA: Association for Supervision and Curriculum Development.

Suwandee, A. (1995). Students' perceptions of university instructors' effective teaching characteristics. *SLLT Journal, 5,* 6-22.

Tucker, P. D., & Stronge, J. H. (2005). *Linking teacher evaluation and student learning.* Alexandria, Virginia: Association for Supervision and Curriculum Development.

US. Department of Health, Education and Welfare. (1971). Report on licensure and related health personnel credentialing. DHEW publication 72-11. Washington D.C.: Author.

Vold, D. J. (1985). The roots of teacher training in America. *Educational Measurement: Issues and Practice, 4*(3), 5-6.

Wright, B. D., & Stone, M. H. (1988). *Reliability in Rasch measurement* (Research Memorandum, No. 53). Chicago: MESA.

A Sample of English Language Teacher Competency Test (ELT-CT)

Directions: *For each of the items below, choose the best option by circling the letters a, b, c or d. The option you choose should be based on what you practice in the classroom, or believe to be the correct course of action.*

Please note that the teaching context in mind is a typical class held at a language institute.

1. Regarding the Persian language, I...
a. never use it in my class.
b. will use it if necessary.
c. try to teach through it.
d. force my learners not to use it in the class.

3. The score I prefer to give my students is usually...
a. out of 20.
b. out of 100.
c. ranging from A+ to F.
d. in the form of a remark, such as 'good' or 'excellent'.

47. When a student does poorly on an examination, I am more likely to say....
a. better luck next time.
b. I know the exam was so difficult.
c. you should have tried harder.
d. you may not be fit for the course.

48. If a student has a kind of Persian accent while speaking, I ....
a. make him/her achieve a British accent.
b. ask him/her to mimic the American accent.
c. won't push him/her to mimic any native-like accent.
d. ask him/her to work more on his/her accent.

49. If a student drops the third-person 's' from a verb, I ...
a. correct his/her mistake immediately.
b. correct his/her mistake at the end of the session.
c. ignore it.
d. notice closely to see whether s/he makes the same mistake again.

50. If a learner has a problem with his/her learning, I ...
a. help him/her directly by giving prompts to solve the problem.
b. just facilitate the process of learning.
c. make him/her discover the solution by him/herself.
d. ask him/her to cooperate with his/her friends to solve it.

51. To enhance critical thinking in my learners, I prefer to start my class questions with terms like...
a. rate, defend
b. define, tell
c. locate, match
d. arrange, separate

57. If one of your learners says "I think you are wrong." What is the most probable feedback which you may provide?
a. That's interesting. In what way?
b. I think you will find that all of the studies show this to be true.
c. I think you need more time to understand that you are wrong.
d. No problem. This is your opinion.

Acknowledgement

We gratefully acknowledge the project reported here was supported by a grant-in-aid of research from the National Foundation of Elites (Iran) to the first author without which this research would not have been possible.

About the Authors

Reza Pishghadam is associate professor in TEFL in Ferdowsi University of Mashhad, Park Square, Mashhad, Iran. His research interests are: Psychology of language education and Sociology of language education. Email: pishghadam@um.ac.ir; rpishghadam@gmail.com

Purya Baghaei is assistant professor in the English Department of Islamic Azad University, Ostad Yusofi St., Mashhad, Iran. His major research interests are the applications of unidimensional and multidimensional Rasch models in educational research. Email: pbaghaei@mshdiau.ac.ir

Hesamoddin Shahriari Ahmadi is a Ph.D. candidate of TEFL in Ferdowsi University of Mashhad, Iran. His major research interests include: Testing and Psycholinguistics. Email: hesam.shahriari@gmail.com