The Impact of Teaching Knowledge Assessment Tools on the Writing Skill of Iranian EFL Teachers

Hamid Reza Shahidy

Department of Foreign languages and literature
PhD Student Science and Research Branch Islamic Azad University Tehran-Iran

Hamid Reza Shahidy is a PhD student in ELT at Islamic Azad University (science and research branch, department of Foreign Languages and literature). He has been teaching as a member of faculty at Islamic Azad University of Garmsar (Iran) for 17 years. His main areas of interests are educational research and language testing.

Corresponding Author email: hr.shahidy@gmail.com.

ABSTRACT: The present study was carried out to investigate the washback effects of teaching knowledge tests, i.e. TKT, DELTA, and alternative assessment on the writing of Iranian EFL teachers. To fulfil the requirements of the study, the researcher selected the research participants at random from language teachers at Simin Language Institute, and then divided them into three groups in a random fashion. The researcher adopted a pretest-posttest comparison group design to investigate the washback effect of each teaching assessment tool and compared the three groups with respect to the writing skill. To test the null hypotheses in the study, the researcher ran parametric statistical tests, including one-way ANOVA, paired-samples t-tests. Having analyzed the statistical data, the researcher discovered that the writing skill was affected only by DELTA and alternative assessment.

Key words: Washback; DELTA; TKT; Alternative Assessment

INTRODUCTION

Assessment in second language teacher education is a rich, complex, and shifting enterprise. Added to this complexity is the more general challenge of assessing teaching as an activity: whether to document its processes (what the teacher is doing), or its outcomes (what the students appear to have learned). Freeman and Johnson (2004) believe that in second language teacher education, it is important to position the discussion of the individual teachers who are being assessed in context, since those judgments are, at least in part, a function of the individual teacher's position within the broader social setting and workforce.

Nowadays, however, many criteria are used in language teacher evaluation. Pennington (1989) categorizes language teacher evaluation tools as being either fluid-response instruments or fixed-response instruments.

One of the simple facts of life in the present time is that the English language skills of a good proportion of our people are seen as vital if our country is to have access to the information and knowledge that provide the basis for both social and economic development. There is consequently increasing demand for competent English language teachers and for more effective approaches to their preparation and professional development. Central to this enterprise are English teaching and English language teachers. Knowledge for-teaching tends to be defined exclusively as content knowledge among most language teachers in our public education system. Pedagogical knowledge is rarely focused on in general assessments; although it must be recognized as part of what teachers need to teach English as a foreign language. Knowledge-for-teaching is equated to knowing the content knowledge.

Teachers at language schools in Tehran are prepared differently, and often with different degrees of exposure to training in the knowledge and practices they need to teach effectively. Considering the varying ways that teachers are deemed qualified internationally, Barduhn and Johnson (2009) call for “fairer and more rigorous assessments ”. Further they note that, in comparison to the standardized assessments of teaching as observable behavior used conventionally, portfolios and other reflective documents may be fairer in documenting the contextual and idiosyncratic aspects that make teaching practice effective.

The primary purpose of this study is to investigate if Teaching Knowledge Test (TKT), Delta Modules, and the alternative assessment have any washback effects on the reflection of in-service teachers at Simin Language Institute. Moreover, the researcher aims to discover if these assessment tools affect the teacher
variables differently. Beliefs about testing according to Cheng (2005) reflect beliefs about teaching and learning.

METHOD

Participants
To achieve the goals of this study, 102 research participants were selected from 150 language teachers, and then were incorporated into three groups in a random fashion. The 102 research participants were all teaching English in different branches of Simin institute, and scored between one standard deviation below and one standard deviation above the mean on an actual paper – based test of English as a foreign language (TOEFL-PBT). Almost all had college degrees such as bachelor's or Master's Degrees in English. They all started teaching English after they had passed a standard entrance exam (actual TOEFL - PBT) and a teacher training course held by the institute. Our subjects had, on average, five years of teaching experience at Simin institute and other language schools in Tehran. They taught English at different levels of language proficiency ranging from beginning to advanced levels. Our subjects were both male and female. They were non-native teachers who were not statistically different in terms of language proficiency.

Instrumentation
To collect reliable data for the purpose of testing the null hypotheses, the researcher applied the following assessment tools in this study.

Instruments used to measure the dependent variables are as follows.

TOEFL (ETS, 2004)
Pre-Professional Skills Test: Writing (ETS: Praxis Series, 2009)
The instruments used in the treatment phase.
Teaching Knowledge test: TKT (UCLES 2010)
DELTA (UCLES 2010)
Alternative Assessment
Teacher self-observation form (Christison & Bassano, 1984)
Teacher observation by others' form (Brown, 2007)
Teacher portfolio assessment (Doolittle, 1994)
Evaluation of teachers by their students' questionnaire (National Schools of Character: Award Winning Practices, 2005).

Procedure
To begin with, the researcher did a survey on a number of language schools in Tehran to find teachers who were willing to cooperate in this study. Head teachers at Simin language institute were willing to have their teachers exposed to such washback project. In the first phase of the study, the researcher selected 102 language teachers who scored one standard deviation above and below the mean on an actual paper-based TOEFL. Afterwards, the researcher divided them into three groups in a random fashion.

In the second phase of the study, the researcher did a survey on the teacher variable, i.e. including the writing skill of all the 102 teachers who were randomly assigned into the three groups. To do so, the researcher gave all the participants a Pre-Professional Skills Test: Writing (ETS: Praxis Series, 2009) intended for non-native language teachers.

The actual samples of TKT, DELTA, and alternative assessment tools along with the instructional and coaching materials related to all these three assessment tools were randomly assigned to the three groups respectively. The researcher coached the language teachers in the three research groups for the content and requirements of TKT, DELTA, and alternative assessment respectively. The purpose of coaching each group for one of these three assessment packages at random was to familiarize the teachers in each group with the assigned assessment tool and to use each assessment tool as a lever to urge the research participants to study the resources related to the content of these assessment tools and live up to the teaching standards set by these teaching assessment tools. These assessment packages came in three different colors. The orange package contained an actual version of DELTA and its relevant coaching materials. The yellow package contained an actual sample of TKT and its relevant coaching materials. The violet package had alternative assessment materials related to Teacher self-observation, teacher observation by others, teacher portfolio assessment, evaluation of teachers by their students, and the instructions on how to meet the requirements of the alternative assessment. In order to brief the research participants, i.e. the teachers in the three groups, the researcher met each teacher in person to coach him or her on how to prepare for the assigned teaching assessment tool.
In the last phase of the study, a Praxis series posttest on the writing skill was administered to the three groups to determine if the teaching assessment tools, for which the three groups were coached had any significant impact on the writing of the EFL teachers in TKT, DELTA, and AA groups respectively.

RESULTS

Since there is a random assignment of the research participants into three groups (A, B, C), the researcher adopted a pretest-posttest comparison group design. The research involved a comparison in which the independent variables differ among three groups (A, B, and C). Therefore, the researcher adopted an pretest-postest comparison group design to test the hypotheses in question.

Four assumptions of interval data, independence of subjects, normality and homogeneity of variances should be met before one decides to run parametric techniques to analyze his or her data (Field, 2009). The present data are measured on an interval scale. The subjects are independent, i.e. the performance of any of the subjects on the tests is not affected by the performance of other subjects. The assumptions of normality and homogeneity of variances require that the population - not the samples - from which the samples have been selected, should have a normal distribution and should show homogeneous variances (Field, 2009; Pallant, 2005). These authors not only doubt the appropriateness of testing the normality of the samples hoping that it would be an indication of the normality of population, but also believe that the assumption of normality which is usually tested through the skewness and kurtosis needs not be checked for one-way ANOVA because of its robustness against the violation of this assumption. As considered by Bachman when discussing the assumption of normality (2005) "... ANOVA is quite robust to violations of this assumption, except where sample sizes are quite small". Despite the above mentioned considerations the normality of the present data was tested. As displayed in Table 1, the ratios of skewness and kurtosis over their respective standard errors are within the ranges of +/- 1.96 (Field, 2009; Pallant, 2005).

<table>
<thead>
<tr>
<th>GROUPS</th>
<th>N</th>
<th>Skewness Statistic</th>
<th>Std. Error</th>
<th>Kurtosis Statistic</th>
<th>Std. Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>TKT</td>
<td>30</td>
<td>-0.608</td>
<td>.427</td>
<td>-1.424</td>
<td>.638</td>
</tr>
<tr>
<td>DELTA</td>
<td>30</td>
<td>.796</td>
<td>.427</td>
<td>1.864</td>
<td>-.573</td>
</tr>
<tr>
<td>AA</td>
<td>30</td>
<td>-0.028</td>
<td>.427</td>
<td>-0.066</td>
<td>.059</td>
</tr>
</tbody>
</table>

The assumption of homogeneity of variances needs not be checked either particularly when the sample sizes are equal (Bachman, 2005) as is the case in this study. However the assumption of homogeneity of variances is also checked through the Levene's tests whose results will be discussed when reporting the one-way ANOVA results.

A one-way ANOVA is run to compare the TKT, DELTA and Alternative Assessment (AA) groups on the TOEFL in order to prove that they were homogenous in terms of their general language proficiency prior to the administration of the treatment. As displayed in Table 2, the mean scores for the TKT, DELTA and AA groups on the TOEFL are 561.60, 568.33 and 573.67 respectively.

<table>
<thead>
<tr>
<th>TOEFL</th>
<th>N</th>
<th>Mean</th>
<th>Std. Deviation</th>
<th>Std. Error</th>
<th>95% Confidence Interval for Mean</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Lower Bound</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TKT</td>
<td>30</td>
<td>561.60</td>
<td>22.777</td>
<td>4.159</td>
<td>553.09</td>
<td>570.11</td>
<td>500</td>
</tr>
<tr>
<td>DELTA</td>
<td>30</td>
<td>568.33</td>
<td>16.626</td>
<td>3.036</td>
<td>562.12</td>
<td>574.54</td>
<td>550</td>
</tr>
<tr>
<td>AA</td>
<td>30</td>
<td>573.67</td>
<td>20.535</td>
<td>3.749</td>
<td>566.00</td>
<td>581.33</td>
<td>527</td>
</tr>
<tr>
<td>Total</td>
<td>90</td>
<td>567.87</td>
<td>20.523</td>
<td>2.163</td>
<td>563.57</td>
<td>572.17</td>
<td>500</td>
</tr>
</tbody>
</table>

The results of the one-way ANOVA (F(2, 87) = 2.70, P > .05; ω² = .036, a weak effect size) indicate that there were not any significant differences between the means of the three groups on the TOEFL (table 3). Thus it can be concluded that the groups were homogenous in terms of their general language proficiency prior to the administration of the treatments.

<table>
<thead>
<tr>
<th>TOEFL</th>
<th>Sum of Squares</th>
<th>d.f.</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Between Groups</td>
<td>2193.867</td>
<td>2</td>
<td>1096.933</td>
<td>2.704</td>
</tr>
<tr>
<td></td>
<td>Within Groups</td>
<td>35290.533</td>
<td>87</td>
<td>405.638</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>37484.400</td>
<td>89</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
It should be noted that the assumption of homogeneity of variances was met (Levene’s F = .70, P > .05).

Table 4. Homogeneity of Variances for the TOEFL

<table>
<thead>
<tr>
<th>Levene Statistic</th>
<th>df1</th>
<th>df2</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOEFL</td>
<td>.701</td>
<td>2</td>
<td>87</td>
</tr>
</tbody>
</table>

![Figure 1. TOEFL Pretest of Writing](image)

**Pretest of Writing**

A one-way ANOVA was run to compare the TKT, DELTA and Alternative Assessment (AA) groups on the pretest of writing in order to prove that they were homogenous in terms of their knowledge of writing prior to the administration of the treatment. As displayed in Table 5, the mean scores for the TKT, DELTA and AA groups on the pretest of writing are 22.16, 23.94 and 20.28 respectively.

Table 5. Descriptive Statistics for the pretest of Writing

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean</th>
<th>Std. Deviation</th>
<th>Std. Error</th>
<th>95% Confidence Interval for Mean</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pretest</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TKT</td>
<td>30</td>
<td>22.16</td>
<td>6.998</td>
<td>1.278</td>
<td>19.55 24.77</td>
<td>12</td>
<td>39</td>
</tr>
<tr>
<td>DELTA</td>
<td>30</td>
<td>23.94</td>
<td>7.848</td>
<td>1.433</td>
<td>21.01 26.87</td>
<td>12</td>
<td>43</td>
</tr>
<tr>
<td>AA</td>
<td>30</td>
<td>20.28</td>
<td>6.431</td>
<td>1.174</td>
<td>17.88 22.68</td>
<td>10</td>
<td>36</td>
</tr>
<tr>
<td>Total</td>
<td>90</td>
<td>22.13</td>
<td>7.194</td>
<td>.758</td>
<td>20.62 23.63</td>
<td>10</td>
<td>43</td>
</tr>
</tbody>
</table>

The results of the one-way ANOVA (F (2, 87) = 1.23, P > .05; ω² = .005, a weak effect size) indicate that there were not any significant differences between the means of the three groups on the pretest of writing.

Table 6. One-way ANOVA for the Pretest of Writing by Groups

<table>
<thead>
<tr>
<th></th>
<th>Sum of Squares</th>
<th>d.f.</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pretest</td>
<td>Between Groups</td>
<td>163.005</td>
<td>2</td>
<td>81.503</td>
<td>1.238</td>
</tr>
<tr>
<td></td>
<td>Within Groups</td>
<td>5728.101</td>
<td>87</td>
<td>65.840</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>5891.106</td>
<td>89</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

It should be noted that the assumption of homogeneity of variances was met (Levene’s F = 1.26, P > .05).

Table 7. Homogeneity of Variances for the pretest of Writing

<table>
<thead>
<tr>
<th></th>
<th>Levene Statistic</th>
<th>df1</th>
<th>df2</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pretest</td>
<td>1.068</td>
<td>2</td>
<td>87</td>
<td>.348</td>
</tr>
</tbody>
</table>
A paired-samples t-test was run to compare the mean scores of the TKT group on pretest and posttest of writing grammar. As displayed in Table 8, the TKT group shows a higher mean (21.27) on the posttest of writing grammar than the pretest ($M = 21.27$).

<table>
<thead>
<tr>
<th>GROUP</th>
<th>Paired Differences Mean</th>
<th>Std. Deviation</th>
<th>Std. Error of the Difference</th>
<th>$T$</th>
<th>Df</th>
<th>Sig. (2-tailed)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TKT</td>
<td>Writing-POST writing –PRE</td>
<td>.893</td>
<td>9.899</td>
<td>-2.803</td>
<td>4.590</td>
<td>.494</td>
</tr>
</tbody>
</table>

A paired-samples t-test was run to compare the mean scores of the DELTA group on pretest and posttest of writing. As displayed in Table 9. The DELTA group shows a higher mean (29.37) on the posttest of writing than the pretest ($M = 23.94$).
The results of the paired-samples t-test indicate that the difference between the two means is statistically significant ($t(29) = 3.32$, $P < .05$, $r = .52$, a large effect size).

A paired-samples t-test was run to compare the mean scores of the AA group on the pretest and posttest of writing. As displayed in Table 13, the AA group shows a higher mean (30.32) on the posttest of writing than the pretest ($M = 20.25$).

The results of the paired-samples t-test indicate that the difference between the two means is not statistically significant. The results of the paired-samples t-test indicate that the difference between the two means is not statistically significant ($t(29) = 4.55$, $p < .05$, $r = .64$, a large effect size).
A one-way ANOVA was run to compare the TKT, DELTA and Alternative Assessment (AA) groups on the posttest of writing in order to probe the effects of TKT, DELTA and the alternative Assessment on the improvement of the writing of the Iranian foreign language teachers. As displayed in Table 15, the mean scores for the TKT, DELTA and AA groups on the posttest of writing are 21.27, 29.37 and 30.32 respectively.

The results of the one-way ANOVA (F (2, 87) = 29.07, P < .05; \(\omega^2 = .38\), a large effect size) indicate that there were significant differences between the means of the three groups on the posttest of writing.

It should be noted that the assumption of homogeneity of variances was not met (Levene’s F = 4.86, P < .05) but there is no need to worry because the samples are of equal sizes, and the ANOVA is robust against the violation of the assumption.

Although the F-value of 29.07 indicated significant differences between the mean scores of the three groups on the posttest of writing, the post-hoc Scheffe’s tests were ran to compare the groups two at a time. Based on the results displayed in Table 17, it can be concluded that;

There is a significant difference between the writing of the DELTA (M = 29.37) and TKT (M = 21.27) groups on the posttest of writing (M difference = 8.10, P = .000 < .05).

B: There is no significant difference between the writing of the DELTA (M = 29.37) and AA (M = 30.32) groups on the posttest of writing (M difference = -.94, P = .768 > .05).
C: There is a significant difference between the writing of the AA (M = 30.32) and TKT (M = 21.27) groups on the posttest of writing (M difference = 9.04, P = .000 < .05).

<table>
<thead>
<tr>
<th>(I) GROUP</th>
<th>(J) GROUP</th>
<th>Mean Difference</th>
<th>Std. Error</th>
<th>Sig.</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>DELTA</td>
<td>TKT</td>
<td>8.100</td>
<td>1.304</td>
<td>.000</td>
<td>4.85 - 11.35</td>
</tr>
<tr>
<td>AA</td>
<td>TKT</td>
<td>9.049</td>
<td>1.304</td>
<td>.000</td>
<td>5.80 - 12.30</td>
</tr>
</tbody>
</table>

* The mean difference is significant at the 0.05 level.

DISCUSSION AND CONCLUSION

Based on the results of the study, a number of interpretations can be made. First, the writing of language teachers in the DELTA and AA group developed resulting from the significant influence exerted upon them through DELTA, and AA. The result is likely to be attributed to the fact that the test tasks in DELTA and alternative assessment appear to represent the teachers’ writing skills more than TKT did. "When test tasks, according to Brown & Abeywickrama (2010), are consistent with course goals and curriculum, teachers and students are likely to be more motivated to perform them, as opposed to multiple-choice questions". Brown & Abeywickrama (2010) also argue that "self-assessment (used in alternative assessment) may sometimes be an appropriate way to challenge students to discover their own mistakes, and this can be particularly effective for writing performance". The indirect multiple-choice format of TKT seem to be of not much relevance to the writing tasks which language teachers may do in real-life situations. Davies et al. (1999) illustrates that a multiple-choice test (MCT) of writing tends to put pressure on the examinees to practice the MC items rather than to practice the skill of writing itself. Hughes (2003) suggests that "direct testing will help to achieve beneficial backwash effect on the skill (writing skill for example) that we are interested in fostering". The fact that TKT is more indirect than a direct test may be an important factor which nullifies its significant washback effect on the writing skills of the teachers in this study. McNamara (2000) states that "performance assessments have better washback than multiple-choice test formats which focus on isolated elements of knowledge or skill".

This result, to some extent, supports Fulcher (2010)’s claim that ‘when test content and instructional content are closely aligned, teachers are covering the necessary material to achieve desired educational goals’. Koretz and Hamilton (2006) describe this alignment as having taken place when "the knowledge, skills and other constructs measured by the tests will be consistent with those specified in the content standards".

Burrows (2004) point out that teachers’ responses to assessment are related to "their attitudes towards and experiences of the implementation of the assessment, their perceptions of the quality of the assessment, the extent to which the assessment represented a departure from their previous practices, and their attitudes to change itself". The findings above suggest that the nature and extent of washback, in line with Chapelle & Brindley (2010)’s perspectives, may be governed by a wide range of individual, educational, and social factors. The researcher did not find any statistically significant differences between DELTA and alternative assessment in terms of the effects they had on the writing of the research participants.
The findings of this study indicate that teacher evaluation may give great benefits to head teachers, teacher trainers, pre-service and in-service teachers who are not much familiar with the latest teaching knowledge assessment tools. This study also urges upon the probationers the need for refreshing their pedagogical knowledge, language skills (reading and writing) and reflecting on their strengths and weaknesses as they practice teaching English. This research may also enable the experienced teachers to do self-assessment and help them set professional goals based on the feedback they have received on their performance on the teaching knowledge tests such as TKT, DELTA, and alternative assessment. Moreover, in-service teachers become aware that they are always under control. For this reason, they always try their best to maintain high standards of teaching. Ongoing teacher evaluation is likely to direct the attention of practicing teachers more than before to self-assessment and self-study to the extent that every teacher would feel responsible for his or her professional development.

The results of this study can help head teachers to learn about the new means by which they might carry out a better evaluation of both pre-service and in-service teachers in order that they can recognize the professional needs of the teaching force. In addition, by looking at the findings of this research, language head teachers may be persuaded to use a teaching knowledge test to help novice teachers to pay more attention to their shortcomings in their teaching practice.

By using teaching assessment tools such as TKT, DELTA, and alternative assessment (AA) before a training program, teacher trainers might provide the pre-service teachers with the resources which add to the effectiveness of a teacher training course. Teacher trainers can also administer teaching knowledge tests to the trainees at the end of the TTC in order that the student teachers not only take their teacher training seriously but also make the great effort to fulfill the qualifying requirements of the course. Shohamy et al. (1996) state, "the power and authority of tests enable policy-makers to use them as effective tools for controlling educational systems and prescribing the behaviour of those who are affected by their results, namely administrators, teachers and students". Under those centrally controlled educational systems, tests are viewed as the primary tools through which changes in the educational system can be introduced without having to change other educational components such as teacher training or curricula (Cheng, 2005).

This study will also persuade language school principals to use teaching knowledge tests as part of a selection test to hire experienced English teachers. Nowadays, it sounds customary for most language schools to use TOEFL or IELTS as a criterion for hiring knowledgeable teaching force. The major plus point is that this research may attract the attention of textbook writers in the field of language assessment and testing to make more contributions to assessment in second and foreign language teacher education by writing about 'teaching knowledge test washback'.

There is no doubt that further research into washback may result in a better understanding of the processes involved and how different variables contribute to its effect in different situations. Washback, based on the results of this study, could change the language teachers who were the agents of change. "Teachers need to become strong advocates for change and for social justice, rather than bystanders to whom testing happens" (Fulcher, 2010). The researcher concluded that the relationship between language teachers and teaching knowledge assessment was a far more complicated issue than it might have appeared to be on the surface. A study of the effects of washback needs to use an innovation model and explore the phenomenon within a multidimensional context.

REFERENCES

English as a foreign Language. The Hague, Mouton.
319.
Pennigton MC. 1989. Directions for faculty evaluation in language education. Language, Culture and
University Press.
Shohamy E, Donista-Schmidt S, Ferman I.1996. Test impact revisited: Washback effect over time. Language
University of Cambridge ESOL Examinations.2010. DELTA Module One D031/1, retrieved from
WWW.CambridgeESOL.org on 2 JUNE 2011.
University of Cambridge ESOL Examinations .2010. Teaching Knowledge Test 001, UCLES 2010, retrieved
fromWWW.CambridgeESOL.org on 2 JUNE 2011.