Developing learners self-assessment accuracy

Arshad Abd Samad
Developing Learners’ Self-Assessment Accuracy of Reading Skills and Strategies Using CEFR Benchmarks

Conference Paper · October 2009

CITATIONS
0

READS
43

3 authors, including:

Arshad Abd Samad
Universiti Putra Malaysia
34 PUBLICATIONS 56 CITATIONS

Masoumeh Akhondi
Hamadan University of Medical Sciences
4 PUBLICATIONS 21 CITATIONS
2nd INTERNATIONAL CONFERENCE
ILANNS 2009
(VOL. 1)
ISSUES IN LANGUAGE TEACHING AND LEARNING AMONGST NON-NATIVE SPEAKERS
LANGUAGE COMPETENCE AND PERFORMANCE
# Contents

The Classification, the Lexicalization Degree and Teaching thoughts of Chinese Chunks  
*Wu Yongyi, He Suosi & Wu Xieyao*  
1

Teaching Malay: A New Global Approach  
*Laurent Metzger*  
31

The Effects of Perception on Production in Chinese Second Language Learners of English: Is Misperception a Precursor for Variable Production?  
*Timothy E. Bonner*  
41

An Evaluation of an English Coursebook Used in an EFL Context  
*Leila Dobakhti & Norizah Hasan*  
59

TESL Trainees’ Views towards Teaching of Grammar in relation to Language Skills  
*Dr Ting Su Hie & Mahanita Mahadhir*  
73

Categories of Perceptual Learning Style Preference of Engineering Students at a Polytechnic  
*Yeoh Poh Chin*  
85

Oral Communicative Activities in the Saudi Third Year Secondary EFL Textbooks  
*Zaky H. Abu Ellif & Nooreiny Maarof*  
97

The Effectiveness of Using Text-based Messages on Learning  
*Ali Jahani & Mozhgan Azimi*  
109

Social Presence and Motivation to Engage in Online Forums among Distance Learning Students  
*Alice Shanthi & Xavier Thayalan*  
115

The effect of narrative and expository modes of discourse on reading comprehension and recall of Iranian advanced EFL learners  
*Azar Hosseini Fatemi & S.M. Reza Adel*  
127

Developing Learners’ Self - Assessment Accuracy of Reading Skills and Strategies Using CEFR Benchmarks  
*Faramarz Aziz Malayeri, Arshad Abd. Samad, Masoumeh Akhondi*  
147

Outcome-based education (OBE) in English for Specific Purposes course: Evaluation on Students’ Feedback and Performance  
*Farrah Diebaa Rashid Ali*  
159
The Issue of Plagiarism and Patchwriting among Non-Native Graduate Students
Kobra Mansourizadeh & Ummul Khair Ahmad

Significance of Invited Speeches to Language Learning for NNS
Li-ying Wu

Preferred Strategies of the Speech Act of Complaint by Malaysian ESL Learners
Maryam Farnia, Lyn Buchheit & Sahida Banu binti Salim

An Evaluation of a Second Language Instruction Management System
Mohamed Amin Embi & Afendi Hamat

Cultural Diversity: An Overview
Mozhgan Azimi

Developing the Skills of Young ESL Learners through Games
Sharifah Amani Syed Abdul Rahman, D.Rohayu Mohd Yunos, & Zafirah Ab Rahim

The Effects of Think Aloud Strategy Training on Reading Comprehension
Sharifah Shahnaz bt Syed Husain Al-Gilani

Literature as a Single Subject in Primary Schools: Perceptions, Challenges and Pedagogical Knowledge
Siti Norliana Ghazali, Razita Mohamad & Faizah Mohamad

Task-based Approach: Selection and Gradation, Type, Complexity, Implementation and Participation
Mohammad Zohrabi, Dr.Parilah Mohd Shah & Dr.Melor Yunus

The Teaching Practices Used by a Science Teacher in Developing Scientific Literacy in a Malaysian Science Classroom
Assoc. Prof. Dr. Koo Yew Lie & Puziah Abdul Wahab & Kamariah Hassan

Enhancing Students’ Learning through Cooperative Learning
Khairunisa Nikman & Ngo Kea Leng

Social Context in Stimulating Thinking in Virtual Space
Anuratha Kanniah
Developing Learners’ Self
- Assessment Accuracy of Reading Skills and Strategies Using CEFR Benchmarks

Faramarz Aziz Malayeri, Arshad Abd. Samad & Masoumeh Akhondi
Department of Language and humanity,
Faculty of Educational Studies,
Universiti Putra Malaysia
f_azizmalayeri@yahoo.com

Abstract
The purpose of this study is to empirically investigate the effects of training students taking CEFR-based self-assessment checklist of reading skills and strategies and reflecting to the reading proficiency benchmarks. The learners’ accuracy with self-assessment is investigated across an intensive period of training. There are two specific objectives for this study: 1) to investigate the accuracy of self-assessment results, 2) to determine the optimal training period. The applied research design is quasi-experimental design. A class of Malaysian semester 7 TESL students, as a whole intact group attended a semester training period. They were tested for their current reading proficiency using the TOEFL reading test battery (paper and pencil version). Meanwhile, they attempted the CEFR-based self-assessment checklist of reading skills three times across the training period. These two measures were used for the purpose of the statistical analyses. Repeated Measures Two-Way ANOVA was used to test the hypotheses. It was found out that the provided training has been effective as the student’s results were more accurate in the second and the third administrations of the self-assessment compared to the first attempt (Effect Size, $\eta^2 = .185$, $p<.05$). The Correlation Coefficients in the Correlation Matrix depicts that the third administration of the self-assessment instrument is highly correlated with the external proficiency measure (TOEFL) ($r = .88$, $p<.05$). Therefore, as the result of training students’ reflections to self-assessment of reading are more accurate and dependable as compared with the results of an established test of reading proficiency.

1. INTRODUCTION
Assessment practices now seem to have started assuming learners’ responsibility for the assessment of their learning processes and products (Little, 1991; Benson and Voller, 1997; Benson, 2001; Luoma and Taranen, 2003; Gardner, 1999). This was called the Alternative Paradigm (Hamayan, 1995). The Alternative Paradigm refers to the procedures and techniques which can be used within the context of instruction and can be easily incorporated into the daily activities of school or classroom (Hamayan, 1995, p. 213). It is particularly useful with English as a Second Language students because it employs strategies that ask students to show what they can do, in contrast to the traditional testing in which “students are evaluated on what they integrate and produce rather than on what they are able to recall and produce” (Huerta-Macias, 1995, p. 9). The alternative paradigm relies on the student’s performance while attending to the goals and
achieving the objectives of the test. Huerta-Macias (1995, p. 90) states that the main goal for alternative assessment is to “gather evidence about how students are approaching, processing, and completing real life tasks in a particular domain.” This performance-based assessment is believed to have attended the concepts of process-based views of language learning and language teaching. Baron (1991, p. 190) states that “when students internalize a definition of what quality means and can learn to recognize it, they have developed a very valuable critical ability. They can talk with their teacher about the quality of their work and take steps to acquire the knowledge and skills required to improve it”. The ongoing nature of alternative assessment involves students and teachers in making judgments about the students’ progress in language using non-conventional strategies (Hancock, 1994, p. 7). Therefore, learner’s autonomy and critical learning theories play significant roles in justifications for utilizing alternative assessment. Recent discussions have apparently emphasized that students tend to reflect and provide feedback on their personal satisfaction with their learning (Gardner and Miller, 1999). Learner’s self-reflection is the very primary step of autonomy. Hence, autonomy is defined as “situations in which the learner is totally responsible for the decisions concerned with his leaning and the implementations of those decisions” (Gardner and Miller, 1999, p. 6).

2. REVIEW OF THE RELATED LITERATURE

In order for achieving a self-generated estimation of one’s own general or specific cumulative language proficiency, language learners are recommended to choose among the many multi-purpose instruments which have been experimentally tested so far. Among these instruments, some are made popular due to practitioners’ and researchers’ results achieved in a number of qualitative and quantitative studies (Bachman and Palmer, 1989; Oscarson, 1997; Ross, 1998). Self-assessment appeared to come of age in the 1980s with the publication of a Council of Europe text on the topic by Oscarson (Todd, 2002). It is now widely accepted that self-assessment is a successful attempt for assessing learning process and locating personal profile matched or miss-matched stance (McNamara, 2000). Since its early introduction, self-assessment has been attempted frequently around the world. Among others, the Council of Europe has remarkably conducted researches trying to test the usefulness of self-assessment in language learning and assessment (North, 2000).

The Common European Framework of Reference (CEFR) was introduced after almost a period of almost forty years of researches and inquiries into the establishing of a “Common Reference” of criterion for practicing teaching, assessing, and evaluating languages in the European context. The Council of Europe in 2001 released the revised edition of these references which contained six levels, A1, A2, B1, B2, C1, and C2. Attempts are made to describe language proficiency through a group of scales composed of ascending levels of descriptors couched in terms of outcomes (Weir, 2005).

The self-assessment grid, in the CEFR self-assessment checklist, consists of all can-do statements benchmarking the proficiency descriptors from the CEFR. As Little (2005, p. 324) states: “the CEFR scales do not claim to model progression…they present a hierarchy of communicative tasks whose successful performance depends on underling linguistic competence.” There are two kinds of self-assessment based on the CEFR descriptors, *Formative* and *Summative* (Council of Europe, 2001, p.26-27). The Formative
in *The Language Biography*, which uses goal setting and self-assessment checklists and is derived from the CEFR’s illustrative scales, and the Summative in *The Dossier*, which contains the language samples that justify the judgments recorded in the biography or checklists (Little, 2005, p. 325-326). Therefore, the two activities almost coincide in employing the CEFR criterion, that is, preparing formative language skills (ability) biography and saving samples of longitudinal ability processes in Dossiers. The contents of the checklists contribute to the attempted educational objectives, goals and targets, with respect to the differences, and they are devised to be used individually while the integration is addressed as well.

### 2.1 The Critical Issues of Validity and Training

Many attempts that have been done in order for establishing validity for self-assessment are concurrent validity correlation designs in which an already established test of language proficiency is used to establish this concurrency. Self-assessment has proven useful in a variety of contexts. There is a substantial theoretical as well as empirical support for the belief that learners’ self-assessments are beneficial (Roever and Powers, 2005). Some researchers have expressed concerns over the reliability and validity of self-assessment for language pedagogy (Davidson and Henning, 1985; LeBlanch and Painchaud, 1985). For the latest available report (Haahar & Hansen, 2008) in a comprehensive report to The Council of Europe have clearly stated that, the question of the validity of self-report techniques in general remains unanswered as the evidences from the current practices come to the discussion panels. Generally, there are three types of reports on the validity of the self-assessment instrument in the literature since the 1970s. The very basic researches were done around the 1970s to 1980s when the concept was evolving, among whom the Blanche and Merino (1989) is monumental, while others had attempted the horizons opened so far by that research. The second most important set of studies followed the Heilenman (1990). This was the most basic self-assessment research which focused on the possibility of presenting a captured concept of the role of over-/under-estimation phenomena which is threatening the self-assessment reliability; hence the validity of self-assessment would be affected. Her study has almost proven that both these factors are attempting to trade in sources of invalidity to self-report techniques. The most current practices are done only after publishing the final release of the CEFR, amongst which, the Council of Europe’s Annual Conferences Reports and the final manual for the application of the new CEFR are noticeable (The CE, 2000-2007).

Heilenman (1990) has reported a low correlation between the course grades and graduate students’ self-assessment of their French language skills. Her other results have vividly emphasized the role of over-/under-estimation in establishing any validity for the self-report instrument. However, prior to this, LeBlanc and Painchaud (1985) reported a high correlation of $r=0.80$ between proficiency test scores and self-assessment. The achieved result was confirmed by other researchers, for example, Wilson (1999) found that self-rating correlated $r=0.75$ and $r=0.70$ with TOEIC (Test of English for International Communication) listening and reading skills respectively. Interestingly related to this study, Tannenbaum, *et al.* (2007) administered the TOEIC and 75 “can-do” statements to 8000 examinees. The investigators found a correlation of TOEIC reading with self-rating across the five domains to be $r=0.65$. These studies are in line with the body of available knowledge as reported by Ross’ (1998) meta-analysis.
Ross (1998) conducted a meta-analysis of studies (an analysis of experiential factors) dealing with self-assessment in second/foreign language skills. For reading skill Ross (1998) has located 23 correlations with an average correlation coefficient of $r = 0.65$. The study by Ross (1998) proved that there is a moderate to high positive correlation reported between any self-assessment of language skills and an external standardized test of English proficiency. Whereas all of these researches had utilized the ACTFL-based self-assessment of language skills, the next generation of instruments is being developed based on the proficiency descriptors in the CEFR guidelines and descriptions. However, the criticism of the new CEFR guidelines appears before they are applied in practice. Weir (2005, p. 282) in a recent review on the major threats to the validity of CEFR criteria and the “can-do statements and benchmarks”, doubts some plausible sources of inefficiently ensured validity, as for the context validity, and for the theoretical validity. Weir (2005) notes that the activities, i.e. the “can-do” statements, are seldom strongly related to the quality of actual performance and the wording of some of the descriptors are non-consistent or transparent enough for the development of the tests. Yet, there are a number of empirical studies which are focused on the plausible training with the self-assessment checklists descriptors (Little, 2005; LeBlanch and Merino, 1989; Figueras et al., 2005; Kohonen, 2001). Such training is supposed to improve the students’ performance in interacting with self-assessment checklists (LeBlanch and Merino, 1989). The more training with adapting and adopting personal language ability measures with the standardized CEFR benchmarks, the more correlated the results of such measures will be with an established test of reading ability. This improved correlation is hoped to result in supporting validity foundations for self-assessment of language ability.

3. METHOD

The current research is planned based on Quasi-Experimental design, where random access to the sample of the population is either impossible or would disarrange the current on-going group arrangement of the research participants (Campbell and Stanly, 1966). However, Kerlinger (1986, quoted in Cohen, 2000, p. 214) refers to quasi-experimental design as “compromise design”, which could be employed in most educational situations where random assignment is impracticable. One of the most common designs among the current quasi-experimental designs in practice is “the non-randomized, pretest-posttest design”. In this particular case, an intact group of semester seven TESL students at Universiti Putra Malaysia were selected and were trained for fourteen weeks, four hours per week on odd week days, two hours per session. The main research question was as follows:

Do the TESL students’ accuracy with the CEFR-based checklist of self-assessment of reading improve after a period of intensive training with the CEFR descriptors of the reading skills and reading strategies?

In order to deeply investigate other dimensions of the manipulated variable, i.e. self-assessing ability in reading comprehension, the researcher formulated other specific research questions looking into the aspects and dimensions of proficiency in reading and the ability to adopt and adapt self-assessment instruments to personal reading comprehension proficiency. These aspects are interfaces between the levels of reading proficiency (competent or proficient, intermediate, and less competent) based on the
observed measures of reading comprehension by TOEFL scores, and the duration of exposure to training in reflecting to content of a self-assessment checklist at three points of time across training.

3.1 Description of Treatment

The developed treatment/ intervention could be best characterized in term of the following components:

1. Specification of the contents
   The benchmarks are developed from the corpus of language materials (receptive as well as productive language use) adopted from the language learners around Europe. For the sake of this research since the main focus is on the reading skill, the contents corresponding to the benchmarked descriptors were pinpointed among the reading subject matters already in use for the reading classes as well as their normal daily language encounter.

2. Standardization of Judgment
   Training at this level must consists of assessing performance in relation to CEFR levels using standardized samples, benchmarking local performance samples and comparing the results to understand contents of the CEFR criteria. Textbook contents, at this stage, are compared with the available samples of “Rated Texts”. These textbooks were decided on based upon their comparability with the CEFR and availability to the researcher. The selected sources came mainly from the Mifflin’s Reading Series (2003).

3. The Empirical Stage
   At this stage, the participants practiced a variety of readings to ensure their captured ability in relating their own reading concepts and abilities with the criteria of the CEFR, assessment of reading skill. As Little (2005) states, the practice stage needs a rich text type source where learners perceive the concept of text type variation in practice to compensate their idiosyncratic language use experience.

4. RESULTS AND DISCUSSION

The main research questions attempt to investigate the possibility of improving the participants’ performance while reflecting to self-assessment checklist of reading comprehension and reading strategies. In order to examine this research question, the Repeated Measures Two-Way ANOVA was utilized. The measures were collected across three administrations of the self-assessment instrument. The participants’ performance in the three levels of Trial (the self-assessment training) was compared with their relevant achieved results from the three levels of the Treatment (the TOEFL). This procedure is supposed to represent the participants’ maturation and growth (change) in the manipulated variable (training). Meanwhile, this training interacts with their available attribute of language ability, which is the reading skill.

Table 1.1 below depicts the descriptive statistics for the three administrations of the self-assessment checklists (Admin1, admin2, and Admin3) in relation with the participants three levels of language proficiency resulted from the administration of the TOEFL test of reading proficiency.
Table 1: Descriptive Statistics for Trial (3 levels) and Treatment (3 levels)

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Std. Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Admin1</td>
<td>low achiever</td>
<td>135.33</td>
</tr>
<tr>
<td></td>
<td>intermediate achiever</td>
<td>143.83</td>
</tr>
<tr>
<td></td>
<td>high achiever</td>
<td>154.06</td>
</tr>
<tr>
<td>Admin2</td>
<td>low achiever</td>
<td>117.44</td>
</tr>
<tr>
<td></td>
<td>intermediate achiever</td>
<td>144.77</td>
</tr>
<tr>
<td></td>
<td>high achiever</td>
<td>158.17</td>
</tr>
<tr>
<td>Admin3</td>
<td>low achiever</td>
<td>107.67</td>
</tr>
<tr>
<td></td>
<td>intermediate achiever</td>
<td>138.46</td>
</tr>
<tr>
<td></td>
<td>high achiever</td>
<td>157.83</td>
</tr>
</tbody>
</table>

From the above descriptive statistics it is understood that the means for the three administrations are significantly different. Although the cumulative means for the three administrations are not too far apart from each other, the three groups of participants have achieved different means from one administration of the instrument to the next administration across the three administrations. This represents the different pace of change and maturation within the three groups based on their specific levels of language proficiency. The low achievers first administration contributes to their overestimated mean of $M_1 = 135.33$, whereas their second and third performance means were $M_2 = 117.44$, and $M_3 = 107.67$, respectively. The sharp decrease represents that this group of achievers had the most revisited estimation of their reading ability compared to the other two groups. The intermediate achievers and the high achievers had the least revisited standpoints and estimations of their reading ability. The reported mean differences for the latter two groups are still statistically significant. It can be inferred from these statistics that the pattern of change is more slant to the far least end of the continuum of estimation as the amount and duration of training has increased. For the sake of this research, the degree of this change is significant due to the nature of the developing perceptions and understandings of the reality, honesty, and the accuracy of the learners’ estimations.

Table 2: The Multivariate Tests for the Trial Main Effect [within-subject]

<table>
<thead>
<tr>
<th>Effect</th>
<th>Value</th>
<th>F</th>
<th>Hypothesis df</th>
<th>Error df</th>
<th>Sig.</th>
<th>Partial Eta Squared</th>
<th>Noncent. Parameter</th>
<th>Observed Power(a)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trial</td>
<td>Wilks' Lambda</td>
<td>.815</td>
<td>.453(b)</td>
<td>2.000</td>
<td>.000</td>
<td>.185</td>
<td>20.906</td>
<td>.986</td>
</tr>
<tr>
<td>Treatment * Trial</td>
<td>Wilks' Lambda</td>
<td>.798</td>
<td>.482(b)</td>
<td>4.000</td>
<td>.000</td>
<td>.106</td>
<td>21.928</td>
<td>.974</td>
</tr>
</tbody>
</table>

The level of significance for the trial is $p=.000<.05$. With the large F-value for both the Trial and the Treatment*Trial, and the smaller values of Wilks’ Lambda=.815, F=10.45, where $p$-value=.000<.05 and Wilk’s Lambda=.798, F=5.482, where $p$-value=.000<.05
respectively, the effect size for the main effect is considered remarkably significant. Therefore, it can be inferred that the utilized Trial (the training) has been successful in emerging changes in the participants’ accuracy across the three administrations of the trial instrument that is the self-assessment checklist.

For the Treatment Main Effect, i.e. the calibration of performance among the three groups of achievers based on the collected TOEFL test results, the Partial Eta Squared is $\eta^2=0.98$. Moreover, the large calculated F-value of 38.235 and the $P$-value=.000<.05, it can be inferred that the three groups of the achievers performed differently in the treatment variable (TOEFL) and their performance is associated with their performance in the trial in a significant and meaningful manner.

In order to investigate the group differences as the result of the training among the three groups of proficiency, a Pairwise Comparison is required. Such clarification is useful to explore which group did better in the three administrations of the self-assessment. Table 4.9 represents the results of the Pairwise Comparison

<table>
<thead>
<tr>
<th>Measure: MEASURE_1</th>
<th>Mean Difference e (I-J)</th>
<th>Std. Error</th>
<th>Sig. a</th>
</tr>
</thead>
<tbody>
<tr>
<td>(I) RTOEFL</td>
<td>(J) RTOEFL</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low Achievers</td>
<td>Intermediate Achievers</td>
<td>-22.20*</td>
<td>4.266</td>
</tr>
<tr>
<td></td>
<td>High Achievers</td>
<td>-36.53*</td>
<td>4.417</td>
</tr>
<tr>
<td>Intermediate Achievers</td>
<td>Low Achievers</td>
<td>22.20*</td>
<td>4.266</td>
</tr>
<tr>
<td></td>
<td>High Achievers</td>
<td>-14.33*</td>
<td>2.584</td>
</tr>
<tr>
<td>High Achievers</td>
<td>Low Achievers</td>
<td>36.53*</td>
<td>4.417</td>
</tr>
<tr>
<td></td>
<td>Intermediate Achievers</td>
<td>14.33*</td>
<td>2.584</td>
</tr>
</tbody>
</table>

Based on estimated marginal means

* The mean difference is significant at the .05 level.

a. Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).

The result of the pairwise comparison across the three levels of achievers represents the fact that these three groups of participants have kept their initial difference across the training and into the three observations. The differences in the performance of the three groups of achievers across the three administration of the instrument could best be illustrated by the graphic estimation of means of the three groups as in the figure below.
The research question is attempting the possibility of providing training and calibration on reading skills and strategies in order to improve the accuracy, reliability, and the validity of learners’ results with self-assessment of reading. In this way, it was assumed that learners follow a certain path towards achieving a working self-access ability, and will develop their self-assessment skill. The result of this maturation might be autonomy in the skill intended, i.e. the reading skill. The test of the between-subjects effect [Partial Eta Squared=.451] for the treatment and the test of the within-subjects effect [Partial Eta Squared=.185] for the trial were both found to be significantly large. These imply that the treatment was successful and the participants were affected by the training provided throughout the intervention period. The interaction pattern between the two factors [Trial and Treatment] was also informative. It represents the pattern of growth of the means achieved for the three groups from the first administration of the self-assessment checklist to the third administration. The results for the first research question confirms the reported results from a number of cornerstone researches related to self-assessment of language skills, particularly the reading skill (Janssen-van Dieten ,2000; Sasaki ,2000; Butler and Lee ,2006; Jafarpur and Yamini ,1995). However, there is a spate of research interest in developing training programs prior to introducing self-assessment into school syllabus especially in the European countries (Council of Europe, 2005). The researchers would like to call for the attention of the practitioners at local level to consider the possibility of developing such training programs and materials related. This way, it might open up the possibility to introduce alternative assessment and more importantly empower the students and benefit them from student-centered pedagogy.

Figure 1: The estimated Marginal means for the three groups of participants across the three administration of the self-assessment instrument
REFERENCES


