Evaluation of Language Proficiency for Iraqi EFL Learners of English as a Foreign Language

By

(.Raed Abdul-Ilah Mohammed Hussein (M.A

/Department of English/College of Education

ABSTRACT

This paper introduces and discusses issues related to the challenge of obtaining more valid and reliable proficiency evaluation. Seven subscales of English language proficiency were administered to 63 Iraqi students at the first- year university stage. The purpose was to identify best combinations and weights of these subscales for predicting EFL proficiency while following constraints imposed by the nature of the syllabus. Error identification, grammar accuracy, vocabulary, and composition subscales were indicated as best predictors, but error recognition was ultimately replaced by reading comprehension to better evaluate course objectives. Error identification was indicated as an indirect measure of composition writing ability. The study seeks to fulfill the following :important purposes

1. To identify an optimal combination of subtests for the prediction of general English Language proficiency at the end of the first- year university stage (College of Education-Dept of English) to make sure that these students are liable to continue their study in this discipline successfully. Optimal is here interpreted as that combination which explains the greatest variance in general language proficiency.

2. To indicate appropriate weighting for each of the selected subtests. This weighting entails both the numbers of items to be allocated to each subtest .and the application of a constant factor to the scores of each subtest

3. To identify possible indirect measures of writing and Listening Comprehension skills, since direct measures often suffer from inadequate .facilities for their administration or subjectivity in their scoring

4. To propose a method for the identification of appropriate items for use within each subscale desired.

1. Introduction

The term 'language proficiency' has been traditionally used in the context of language testing to refer in general to knowledge, competence, or ability in the use of a language irrespective of how, where, or under what conditions it has been acquired (Bachman, 1990:16).Reform of traditional language proficiency examinations was planned and conducted not merely as a means of ensuring greater examination quality but also with the specific aim of improving the quality of English instruction. Textbook, curricular, and teacher training reforms were found to be of limited utility in a system where teachers ultimately prepared students to pass the traditional examinations. So long as the examination itself encouraged rote recall of memorized compositions, rules of grammar, or narrowly restricted sets of facts from the syllabus, achievement on the examination and thus in the course of instruction was not consonant with attainment of language proficiency. It was possible for the students to succeed on the examination without their having demonstrated proficiency in the language and without the teachers' having conformed to the prescribed instructional practices.While concurrent studies are underway to assess the impact of these examination reforms on teacher practices and student achievement, the present study reports results of an attempt to devise an examination that is proficiency measurement oriented for learners in the present stage.English component has contained several subtests tapping abilities in grammar, reading comprehension, composition writing and vocabulary inference. As the exact nature of the college exams changes from year to year, measures of test reliability and validity have not been routinely available (Schleifer and Upshur 1979). Until recently, no scientific basis was established for the selection and weighting of appropriate subtests to be included in the exams. Pervasive problems of subjectivity and low reliability in the scoring of compositions and absence of adequate facilities for assessing listening and speaking skills have remained. Although directed to the

resolution of the specific difficulties mentioned, the present study is of general interest in that a procedure is advanced for the development of language .examinations that serve proficiency measurement functions

Method .2

2.1 **Sample**

In statistics and testing, a sample is a group of individuals which is selected to represent a population (Richards et al, 1992:321). A representative sample is the one which contains a good representation of the population from which it is selected .The present study drew as its representative sample 63 students at the first-year university stage at university of Babylon-College of Education-Dept. of English. These students had .received eight years of instruction in English as a foreign language

.Instrumentation 2.2

An English Language proficiency test battery was developed for the purposes of the study. Test battery means a group of tests which are given together to a student or group of students (Richards et al, 1992:34).This test consisted of seven :subtests with varying numbers of items and reliabilities as reported in Table 1

TABLE 1

Reliabilities for All Subscales at Present Length and Adjusted to 50-Item Length

Subscale	Type of Reliability	Items	Coefficient of Reliability	item-50 Reliability
Grammar accuracy	KR-20	20	0.708	0.858
Reading comprehension	KR-20	12	0.544	0.833
Vocabulary	KR-20	20	0.722	0.867
Composition	Split Half	X 10 *3	0.834	0.962

Appropriate Response	KR-20	20	0.734	0.873
Error identification	KR-20	20	0.496	0.711
Listening Comprehension	KR-20	20	0.846	0.932
Grand Total	KR-20	122	0.929	0.843

.(Where KR-20 means Kuder-Richardson formula 20 (Bachman, 1990:176

.sentences awarded up to three points per sentence 10*

In Table1, the final column represents a predictable estimate of reliability assuming each subscale consisted of 50 items instead of the actual number. This estimate is included because of the known relationship between reliability and :test length. The nature of the subscale tasks is as follows

Grammar accuracy 2.2.1

It is essential that students master the grammatical system of the language they are learning, therefore, classroom tests of grammar can play a useful role in a (language programme (Heaton, 1988:34

Grammar items required selection of one of four multiple-choice options as in the :following example

1- ……you mind checking these figures? (a. does b. would c. shall d. can)

Note: The Grammar textbooks for the first- year of the study is Murphy, R. .(1987) English Grammar in Use. Cambridge: Cambridge University Press

Reading comprehension 2.2.2

These items required reading two passages, each 200 words long, and responding to six items per passage by selecting one of four multiple-choice options to .complete a sentence stem correctly

:The lorry stopped suddenly

;A. in order not to hit the car in front

;B. in order to draw up alongside the car

;C. because the driver didn't want to run over the dog

.D. because the driver hasn't seen a cake in the air

Note: The textbooks for the first- year of the study is Alexander, L. G. (1967) .Developing Skills. London: Longman

Vocabulary 2.2.3

Vocabulary is often recognized as necessary in effective communications in a foreign language. Vocabulary knowledge is therefore considered a critical factor in the assessment of learners' language proficiency. The vocabulary items aim to investigate lexical richness in subjects' output. The test involved instructing the students to explain the meanings of a given words and phrases as they are used in the passage. An example follows:Explain the meanings of a given words and phrases as they are used in the passage: seemingly; concealed; vivid saying; .reputation; ruined; fiction; to varying degrees

Composition 2.2.4

Here the examinees were requested to write 10 guided sentences, five for each of two topics. Students were to construct appropriate sentences from the guide words (e.g., friendship / the countryside in spring). In principle, no points for errors of grammar or spelling [are deducted], provided that it is clear" that the correct response was intended" (Hughes 2003: 170). Nevertheless sentences were allowed a maximum of three points from which one point was deducted for spelling and grammar errors and half a point for punctuation and capitalization errors. Negative scores were not computed (i.e., zero was the .(lowest score

Note: The textbook for the first- year of the study is Razzak, F. A and Al-Hassan, H. (1981) College Composition. A publication of the institute for developing of .English language teaching in Iraq. Baghdad

Appropriate Response 2.2.5

Here the examinees responded to a statement or question by selecting an ,.Appropriate Response from among four options, e.g

?What did they eat

.A. This morning. C. Yes, they did

.B. I don't know. D. At the Hilton

Error identification 2.2.6

In this case the task was to identify a grammatical error in a sentence with four ,.possible erroneous segments; e.g

A B C

/ In my way to school, / I met a man / who told me that

D

.the school was on fire

Listening Comprehension 2.2.7

The cloze-recall format of this test required the examinee to fill in missing content words in a cloze passage following the aural presentation of the entire passage. Only the exact lexical item was accepted, but spelling errors were not .counted

The passage consisted of a 100-word dialog, with 20 deletions which were not ,.spaced at equal intervals; e.g

.----------Ali: Let's ----------it. We're

In each subscale, an attempt was made to match items with course content. Only grammatical structures and lexical items previously encountered in the students' .syllabus were permitted

Procedures 2.3

Students were permitted three class periods, up to two and a half hours, to complete the test-including instructions and distribution and collection of papers. All students received the test subscales in the same,

above-listed sequence. Tests were administered simultaneously in two classrooms by two trained examiners. All testing was completed in one session on one day. .Cheating was minimized by careful administration and observation

Results .3

Table 2 reports the means and standard deviations of all subscales and total test battery. This provides comparative indication of subscale difficulty and sample spread. Note that Appropriate Response was the easiest subscale, while .Composition proved to be the most difficult

TABLE 2

Means and Standard Deviations of All Subscales and Total Test Battery

Subscale	No. of Items	Mean	S.D
Grammar accuracy	20	10.22	3.63
Reading comprehension	12	4.88	2.27
Vocabulary	20	6.11	3.57
Composition	*X 3 10	4.01	4.31
Appropriate Response	20	13.39	3.52
Error identification	20	6.32	2.71
Listening Comprehension	20	10.87	4.40
Grand Total	122	55.76	17.41

Table 3 reports the intercorrelations of all subscales and test total. From Table 3, it is clear that Composition, Grammar accuracy, and Vocabulary appear to bear

the highest initial relationship to the Grand Total for the entire battery(0.78, 0.76, and 0.74, respectively). Thus we might decide on these subscales as good .measures of general proficiency

There are, however, at least two problems associated with using the Grand Total with subscale correlations as indicators of general proficiency. First, subscales have an additive, consistent, diverse contribution to total score, which is a .function in part of the variability of the subscales

TABLE 3

(Intercorrelations of All Subscales and Test Total (N=63

Subscale	GA	RC	VC	CM	AR	EI	LC	GT
Grammar accuracy	1.00							
Reading comprehension	0.52	1.00						
Vocabulary	0.41	0.50	1.00					
Composition	0.52	0.39	0.50	1.00				
Appropriate Response	0.54	0.46	0.38	0.45	1.00			
Error identification	0.42	0.43	0.49	0.49	0.39	1.00		
Listening Comprehension	0.37	0.29	0.41	0.41	0.25	0.33	1.00	
Grand Total	0.76	0.66	0.74	0.78	0.69	0.68	0.66	1.00

We can eliminate the resulting distortions by correcting for part-whole overlap. Following this procedure we obtain the following corrected correlations of subscales with Grand Total: GA, .65; RC, .58; VC, .62; CM, .65; AR, .56; EI,

.58; LC, .48. Here the contributions of subscales themselves to total score are systematically removed from subscale-total score correlations.

A second problem requiring solution is the distortion due to varying levels of reliability for subscales. The magnitude of subscale-total score correlation is in part a function of the reliability of the subscale. We can hold the effect of varying reliabilities constant by correction for attenuation. Table 4 reports the intercorrelations of Table 3 following correction for attenuation.

TABLE 4

Intercorrelationos of All Subscales and Test

(Total, Corrected for Attenuation(N=63

Subscale	GA	RC	VC	CM	AR	EI	LC	GT
Grammar accuracy	1.00							
Reading Comprehension	0.84	1.00						
Vocabulary	0.57	0.80	1.00					
Composition	0.68	0.58	0.60	1.00				
Appropriate Response	0.75	0.73	0.53	0.58	1.00			
Error identification	0.70	0.82	0.82	0.76	0.65	1.00		
Listening Comprehension	0.48	0.42	0.53	0.49	0.39	0.52	1.00	
Grand Total	0.80 *	*0.82	0.76 *	* 0.74	0.68 *	0.85 *	0.54 *	1.00 *

* The coefficients in this row have been both corrected for attenuation and adjusted for part-whole overlap.

From the disattenuated correlations of Table 4 we can begin to reconstruct an optimal combination of subtests for the prediction of general language proficiency. We can also attend to our forestated purpose of identifying indirect measures of composition and Listening Comprehension. In this latter connection, we note that the highest observed correlation with Composition (CM) was 0.76, with Error identification (EI). This leads us to conclude that Error identification may serve as an indirect measure of composition writing ability. Results for (Listening Comprehension (LC

were less conclusive since the highest observed correlations (Vocabulary, 0.53, and Error identification,0 .52) were comparatively low. Madsen's (1977)results correlating Appropriate Response type tasks with traditional Listening Comprehension as measured in the Michigan Test of Aural Comprehension had led us to expect that Appropriate Response (AR) format would prove to be an indirect measure of Listening Comprehension (LC). Such expectations were not borne out by our results, perhaps because our test was not speeded. We also used a different Listening Comprehension measure, although Henning's (1978) results showed a strong relationship between our listening recall format and traditional measures of Listening Comprehension employed in the UCLAPE. We chose to reject cloze format as a correlate of Listening Comprehension because previous research asserting a relationship has failed to disattenuate correlations (Oller 1973), and because of the uncertain pedagogical value of the traditional cloze procedure. While the traditional cloze procedure appears to provide a reliable measurement of general proficiency, it is not clear that it is a sufficiently useful or valid measure of EFL achievement. It fails to provide specific diagnostic information to the students, it inadequately reflects course content, and, when reliability is held constant, the kinds of abilities it does tap appear to be more .adequately covered by other examination components

For purposes of assembling an optimal combination of subtests for the prediction of general language proficiency, we relied on the correlation matrix of Table 4 in .the generation of the multiple correlation and regression data of Table 5

TABLE 5

Stepwise Multiple Regression of Four Disattenuated Subscales on Corrected Battery Total as a Dependent Measure of General (Language Proficiency, (N=63

Item .No	Source	R	R^2	R^2 adj	R^2 inc	F/R^2 inc	.d.f
.1	Grammar accuracy	0.850	0.723	0.721	0.723	**475.04	1.182
.2	Reading comprehension	0.897	0.805	0.803	0.082	**76.11	1.181
.3	Vocabulary	0.904	0.818	0.815	0.013	**12.86	1.180
.4	Composition	0.906	0.822	0.818	0.004	*4.02	1.179

B,1 = 0.343 , B2 = 0.374 , B3 = 0.204 , B4 =0 .103

p < .05, ** p < .01

Note from the results reported in Table 5 that Error identification, Grammar accuracy, Vocabulary and Composition respectively were the only four subscales to qualify for entry into the multiple regression equation. Addition of further subscales did not contribute to further prediction of our dependent measure of general proficiency. Moreover, the beta coefficients supplied on the bottom row of the table provide an indication of appropriate weighting for each subscale; thus our ideal test becomes roughly 33 per cent Error identification, 37 per cent Grammar accuracy, 20 per cent Vocabulary and 10 per cent Composition.

Although Table 5 presents a psychometric ideal, in the real world we have found certain constraints. First of all, as was reported in Table 1, Error identification exhibited the lowest reliability of all subscales, indicating that it is difficult for

the uninitiated test developer to prepare suitable items in this format that would match examinee ability. If we are preparing specifications for teachers to follow in developing tests, we cannot recommend inclusion of an Error identification subscale in the battery. In the second place, Reading comprehension is of such basic importance in the curriculum that we cannot justify its exclusion from the battery, even if we have established that it, like Listening Comprehension, has little pyschometric value in predicting general proficiency beyond the other subscales identified.In determining actual policy, we know in advance that we must exclude Error identification and include measures of reading, writing, and grammar usage in our exam in order for it to find acceptance. We therefore returned to the multiple regression procedure as reported in Table 6 to determine 1) the fourth subscale to include along with the three given subscales, 2) the magnitude of the resultant R^2, and 3) the appropriate weights to be applied to .each of the four final subscales

TABLE 6

Multiple Regression Analysis of Four Subscales

(of the Language Proficiency Battery (N=63

Item .No	Source	R	R^2	R^2 inc	F/R^2 inc
.1	Reading comprehension	0.820	672 .0	672 .0	*372.88
.2	Grammar accuracy	.0 845	714 .0	042 .0	*26.58
.3	Composition	.0 885	783 .0	069 .0	*57.24

.4	Vocabulary	.0 896	803 .0	020 .0	*18.17

B1= 0.187 , B2=0 .312 , B3 = 0.250 , B4 = 0.283

p < 0.01 *

As Table 6 indicates, we found Vocabulary to be the best remaining candidate for inclusion in the test battery. We also found the resulting battery to have the respectable R of 0.896. The beta weights suggest that our battery should be comprised of roughly 18 per cent Reading comprehension,30 per cent Grammar .accuracy, 24 per cent Composition, and 28 per cent Vocabulary

Conclusion .4

In the present study we have attempted, in an Iraqi educational context, to determine an optimal combination of subtests for the prediction of general English Language proficiency at the completion of first-year university stage. To accomplish this end we have had to be sensitive both to psychometric and prevailing curricular concerns. Our results supported the inclusion of Reading comprehension, Grammar accuracy, Composition, and Vocabulary formats. A combined multiple correlation of 0.896 was found for this battery when the disattenuated and overlap-corrected total battery score correlations were used to .represent dependent variable relationships

The standardized regression coefficients suggested comparative weightings of 0.187 (RC), 0.312 (GA), 0.250 (CM), and 0.283 (VC). Since these weights apply to z score transformations of the subscale raw scores, they are sensitive to and provide controls for the fact that our actual subscales varied in numbers of items and in total-score variability. Our findings supported the use of Error identification format as an indirect measure of Composition Writing (r = 0.76); .however, no such correlate was found for Listening Comprehension

At this point it is worthwhile considering in what ways our proposed reformed examination differs from the traditional format. Reflecting on such differences may help highlight the contrast between examination that provide optimal

measurement of both achievement and proficiency and those examinations that permit teachers and students to depart from declared goals and expectations.

1. The reformed examination described in Table 6 is more efficient than the traditional examination. The total examination time for the four subscales is less than an hour and a half, as opposed to more than two hours. This implies further time savings in scoring or, if administration time is held constant.

2. The reformed examination does not include redundant subtests. Thus a translation component, a letter-to-a-friend component, a question-written-answer comprehension component, and a completion component have all been eliminated. Performance in these components is adequately measured already in the four subtests of the reformed examination.

3. Composition writing in the reformed examination follows a guided format. This at once permits greater reliability of scoring and prevents memorization of general composition passages.

4. The reformed examination gives greatest emphasis to objective measures of grammar accuracy. This new emphasis entails 1) greater objectivity of scoring and thus more reliable measurement, 2) more items tested and thus greater face validity as a comprehensive measure of the syllabus, 3) the possibility of more diagnostic information about discrete points of learning strength or weakness, and 4) reduced likelihood that students could neglect parts of the syllabus and succeed in the examination.

5. The reformed examination increases multiple-choice options from two or three per item to four per item. This change implies 1) reduced possibility of success due to guessing and thus more reliability of measurement, 2) greater insistence on fluency in reading and processing of information in the target language, and 3) more information being processed and provided per item.

6. The reformed examination measures reading comprehension only in objective multiple-choice format. Question-written-response items have been eliminated. This has the advantages that 1) scoring is more objective and thus less time consuming and more reliable, 2) the responses themselves are more valid in that

they entail reading and recognition rather than writing, and 3) overemphasis is not given to the writing skill, which has been adequately measured already in the composition writing subtest.

7. The reformed examination, by reducing the number of subtests while increasing the number of items within a subtest, increases the likelihood of reliable measurement at the subtest level.

REFERENCES

Bachman, L. F., & Palmer, A. S. (1996). Language testing in practice. Oxford: Oxford University Press.

Bachman, L. F. (1990). Fundamental Considerations in Language Testing. Oxford University Press.Oxford.

Beaudichon, J. (1973). Nature and instrumental functions of private speech in problem solving situations. Merrill-Palmer Quarterly, 19, 117–131.

Blanche, P., & Merino, B. (1989). Self-assessment of foreign language skills: Implications for teachers and researchers. Language Learning, 39, 313–340.

Henning, G. H. 1978. Developing English language proficiency measures for native speakers of Arabic. UCLA Workpapers in English as a Second Language XII. Los Angeles: UCLA.

Henning, G. H., Ghawaby, S. M., Saadalla, W. Z., EI-Rifai, M. A., Hannallah, R. K., and Mattar.H. M. 1981. 'Comprehensive Assessment of Language Proficiency and Achievement Among Learners of English as a Foreign Language'. TESOL Quarterly, Vol. 15, No. 4 (Dec., 1981), pp. 457-466.

Hughes, A.(1989). Testing for Language Teachers. Cambridge University Press. Cambridge.

Hughes, A. (2003). Testing for Language Teachers. 2nd ed. Cambridge Language Teaching Library. Cambridge: Cambridge University Press.

Krashen, S. D. (1987). Principles and Practice in Second Language Acquisition. Prentice-Hall International. London.

Madsen, H. S. 1977. Development of an alternate modality listening test. Unpublished report on testing research project. American University in .Cairo, Cairo, Egypt

.Techniques in Testing. Oxford University Press. Oxford .(1983)----------

Oiler, J. W. 1973. Cloze tests of second language proficiency and what they .measure. Language Learning 23, 1

Rudner, L. (2001). Reliability. College Park, MD: ERIC Clearinghouse on .Assessment and Evaluation

Schleifer, A. and J. Upshur. 1979. Analysis of the 1979 English arts examination of the Thanawiyya Amma in the Sudan. An unpublished report of the AID Test Reform and Evaluation Project, Testing Unit, Ministry of .Education, Cairo, Egypt

Ur, P. (1996). A course in Language Teaching. Cambridge: Cambridge University .Press

Weir, C. (1994). Understanding and Developing Language Tests. London: .Prentice Hall

Wright, B. D. and M. H. Stone. 1979. Best test design: Rasch measurement. .Chicago: MESA Press

Young, J. W. (2008). Ensuring valid test content tests for English language learners. R&D Connections 8. Princeton, NJ: Educational Testing .Service