THE EFFECT OF LEARNING MATERIALS ON STUDENTS' LANGUAGE GAIN

Bakil ALWALSS[*1]
Majid GHARAWI[2]
English Language Institute, Jazan University, Saudi Arabia

[1]balwalss@jazanu.edu.sa
[2]mgharawi@jazanu.edu.sa

Manuscript received 20 November 2018
Manuscript accepted 23 May 2019
**Corresponding author*
https://doi.org/10.33736/ils.1280.2019

ABSTRACT

This study investigated the effectiveness of the English programme for year one students, at Baish Community College (Males, BCCM), Jazan Community Colleges (Males, JCCM and Females, JCCF), College of Engineering (Males), and College of Design and Architecture (Females). Research tools were a programme evaluation form and two short placement tests to determine students' progress of learning English. Two placement tests were used: one at the beginning of the semester, and the other in the second half of the semester with a two-month gap. The average of all groups on the first placement test was 18.5 out of 50. ANOVA analysis showed no significant differences between group averages at $p > 0.05$ (p-value was 0.26). The level at the start is similar to all groups (homogeneous students). The second placement test showed a slight learning progress. The average of all groups was 21 out of 50, but with a high variation in percentages of gain amongst groups. Therefore, the second ANOVA analysis showed significant differences between the groups' averages at $p < 0.05$ (p-value was 0.0079). A third analysis was conducted on both tests to ensure further validity of the results; t-test for paired samples was used. All groups were positive except for Jazan Girls Community College which showed no progress at all.

Keywords: language gain, instructional materials, language programme

Introduction

In recent years, Jazan University founded a new polytechnic college in Baish Governorate, Jazan Region, Saudi Arabia. This college is called Baish Community College (BCC). Facilities and partnerships with bigger corporations, such as Saudi Electricity and Saudi Water Authorities, made decision makers at Jazan University think of generalising the experience of the BCC to all Jazan Colleges. Some of the changes done in BCC included: 1) three semesters for English while other colleges have only two; 2) 20 contact hours per week while other colleges have only 15; and 3) having English native speakers while other colleges have their majority of teachers as non-native English speakers.

The English Language Centre (ELC) (now English Language Institute ELI) decided to conduct several studies if such a "context" is to be generalised since Jazan University needed a huge budget to generalise this excellent "programme structure" to more than 25 colleges, where more than 18,000 students will be directly affected—positively or negatively. The current study is the first one in these series, where Baish Community Male College (BCCM) remains central. It is compared to two colleges of similar status, mainly Jazan Community College (Males, or JCCM) and Jazan Community College (Females, or JCCF), and two colleges of higher status, *viz*. Engineering (Males) and Architecture (Females). The instructional materials used at the BCCM match the admission language level of the students (post-beginners), whereas they (materials) are higher for the other four colleges. Therefore, the authors were highly interested to investigate the effect of the instructional materials on language gain. As pointed out by Larsen-Freeman and Long (2014), apart from studying effects of instructions on processes, "a further major question remains unresolved and in need of serious attention: How does instruction affect SLA?" (p. 536). It is the aim of this study to evaluate the English language programmes (specifically the instructional materials) offered at Jazan University across its colleges and obtain insights on its effects on students' language gain. Specifically, it will answer the following research questions:

1) Will the BCC students show language improvement better than the rest of other colleges since they have better facilities, as it is evident in the 'Programme Context'?
2) Is the progress, if any, for every college significant and tangible?
3) How can we direct future research for the English language programme at the ELC?

Review of Literature

Assessment versus Evaluation

Assessment is a fact-finding activity that describes conditions that exist at a particular time. No hypotheses are proposed or tested, no variable relationships are examined, and no recommendations for action are normally suggested. It is more or less related to exams. Normally it is directed to measure students' progress through various means and methods. Collins and O'Brien (2011, p. 42) commented that "assessment may affect decisions about grades, advancement, placement, instructional needs, and curriculum."

Evaluation, on the other hand, is concerned with the application of assessment findings. It implies some judgment of the effectiveness, social utility, or desirability of a product, process, or programme in terms of carefully defined and agreed-upon objectives or values. It may involve recommendations for action. Collins and O'Brien (2011) view evaluation as:

> The systematic investigation into the process or outcomes of the implementation of a particular educational programme, also synonymous with "programme evaluation": such investigations answer calls for accountability, assist in decision making, aid programme development and planning, and serve research. (p. 143)

They, later, narrow down the definition of programme evaluation to "[a] process in which academic programmes are appraised in terms of criteria chosen to judge effectiveness or the rate of efficiency (Collins & O'Brien, 2011, p. 143). However, according to Ahmed (2008), evaluation is not concerned with generalisations that may be extended to other settings. As Norris (2009, p. 7) puts it, "evaluation can contribute to understanding and improving language teaching practices and programmes". The current evaluation is not a full review of the language programmes investigated in this study. It is directed to one component of a language programme; that is, the effect of material selection on students' learning, and thus the effectiveness of the programme itself. Desheng and Varghese (2013) briefly write about the purpose of evaluation as determining the quality of a programme through judgment on its merits. There are also studies conducted on the need to evaluate the effectiveness of learning programmes (Alobaid, 2016; Kiely & Rea-Dickins, 2005; Kunnan, 2014). This introduction is for a broad understanding about the context of evaluation in language programme. Subsequently, the following paragraphs provide the general picture of what type of tests are considered in this study.

Criterion-referenced testing is widely understood to measure "knowledge, skill or ability in a specific domain" where performance is measured against certain "existing criterion level of performance" (Fulcher & Davidson, 2007, p. 370). On the other hand, norm-referenced testing is defined as tests where "the score of any individual is interpreted in relation to the scores of other individuals in the population" (Fulcher & Davidson, 2007, p. 373). The study adopted the criterion approach so as to objectively measure the true level of English for each college. However, the comparison among groups is obviously norm-referenced.

English Language Programmes at Jazan University

Jazan University attained its institutional accreditation early in 2018 and is planning for a wide-scale and comprehensive programme accreditation. Therefore, the reference of this study's programme evaluation must not be confused with the programme evaluation of Saudi National Commission for Academic Accreditation and Assessment (NCAAA) (Now it is replaced with the Education Evaluation Commission, EEC). This study is concerned with a specific component within a non-awarding certificate programme, whereas the latter focuses on details directly related to the academic and administrative domains for purposes of accreditation and the like. The NCAAA programme evaluation is a lengthy process. It starts with individual course reports, field studies, annual programme reports, and regular programme self-studies. The reader is advised to take this into consideration so as to differentiate this type of evaluation from the EEC/NCAAA evaluation. The general NCAAA standards are summarised below (NCAAA, 2015, pp. 27-28). This type of programme evaluation is based on a wider scale reporting on “the eleven specified standards and each of the sub-standards” (NCAAA, 2015, p. 30).

1) Mission and Objectives
2) Programme Administration
3) Learning and Teaching
4) Student Learning Outcomes
5) Student Administration and Support Services
6) Learning Resources
7) Facilities and Equipment
8) Financial Planning and Management
9) Employment Processes
10) Research
11) Relationships with the Community

Methodology

In this study, both quantitative and qualitative measures were used in order to answer the research questions. The instruments used in the study were an evaluation form and two placement tests.

Instruments

An evaluation form containing 25 descriptors was used to collect qualitative data for the five different language contexts at Baish Community College (Males, henceforth BCCM), Jazan Community Colleges (Males, henceforth JCCM and Females JCCF), College of Engineering (Males), and College of Design and Architecture (Females). This form contained the seven main areas related to the English programme from admission to practice through to final summative assessment as follows:

1) Intake and Placement
2) Materials
3) Curriculum
4) Assessment
5) Teaching and Learning Environment
6) Instructors
7) Programme Environment

These parameters were selected from the review of several references which focused generally on the success of language programmes, such as Weir and Roberts (1994) and Lynch (2003).

The second instrument was a placement test taken from Penguin Readers Placement Tests (Fowler, 2005) which is designed to place language learners according to their levels from A1 (level 1) to C2 (level 6) according to the Common European Framework of Reference (CEFR) (Verhelst, Van Avermaet, Takala, Figueras, & North, 2009). In this context we will refer to the learners' language proficiency by numbering (CEFR) language levels for direct correspondence with the CEFR coding. These levels should not be confused with the levels of the Association of Language Testers in Europe (ALTE, 1998), where, for example, level 3 corresponds with the ALTE level 2 (but it exactly matches CEFR level B1).

The placement test is composed of two parts, each of which carries 25 items. The first part is a mixture of levels 1 and 2, whereas the difficulty of part 2 is based on level 3. Random items were selected from Penguin Readers Test. Conducting longer Placement Tests with large samples can greatly assist in recommending the level of admission or selection of the instructional (teaching) materials. However, for purposes of quick analysis, totals of both parts were taken collectively. Score interpretation is reported below, which applies to individual learners. For the purpose of analysis, the overall average score of each college is treated as an individual learner.

- If the score is less than 15, then the level is 1.
- If the score is less than 30, then the level is 2 (which is acceptable for community colleges)
- If the score is between 31 and 49, then the level is 3 (which is required for engineering colleges)
- If the score is 40 and above, then the instructional material for that group (college) should be replaced with a higher level.

The second placement test was a replication of the first test. A period of two months was the time gap between these two tests in order to allow learning to take place. The students who sat for the first test also sat for the second test so as to control the variable of language progress (gain). The placement tests were used to measure language learning progress by comparing the group means in the total test grades for the first and second tests.

Methods of Analysis

The methods of analysis consist of the following successive steps: 1) calculating the points for the language programme contexts; 2) running analysis of variance for each test results; and 3) conducting the t-test to measure the significance of the language learning gain. We hoped that these analyses collectively would provide insights on the role of the instructional materials in the effectiveness of language programmes. The instructional materials used in BCCM are matching the admission language level of the students, whereas they were higher for the other four colleges.

To answer the first research question (language gain measured via placement tests), ANOVA and t-tests were used along with the necessary descriptive statistics to ensure the validity of the conclusion (Randolph & Myers, 2013, p. 85). Soh (2016) seems not to advocate the use of ANOVA and t-tests in the educational context, unless certain assumptions are met. However, Soh (2016, p. 31) rightly reiterates that "ANOVA answers the question "Is there at least one significant difference among the … classes?" and the t-test answers the question "Which pair of classes has a significant difference?"'' (Emphasis in the original). Moreover, since the main tool of discrimination is a language test administered twice in five different environments/conditions (programme contexts), our hypotheses were statistically tested. Thus, we assumed in both situations that:

- Null Hypothesis (H_0): $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$
- Alternative Hypothesis (H_1): $\mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4 \neq \mu_5$

µ = mean, and subscript numbers indicate colleges (treated as groups, since we have samples only from these colleges). The stages of analysis were as follows:

1) Analysis of variance (one-way ANOVA) to test the significance of the difference between group means in the first placement test.
2) Analysis of variance to test the significance of differences between group means in the second placement test. Therefore, the second ANOVA analysis should lead to three possibilities:
    a) At least one (or more) group will score a higher average (than the rest), and therefore there will be a significant difference at alpha probability level of 5% ($p < 0.05$) so that differences in averages become meaningful.
    b) All groups score higher averages in the second placement test (than the first placement test) but they show no significant differences at $p < 0.05$. This means that there is an increase in learning among the five groups, and this leads to the question: Is this change in learning (the increase of language gain) significant for each group?
    c) One (or more) groups will show no progress so that its (their) second average remains (plus or minus) near its (their) first test average, and therefore there is no significant difference at p. < 0.05. If such a finding is obtained, then a dilemma is present and a wider scale study must be conducted to find causes and recommend solutions.
3) In order to look into these three scenarios above, and to test the difference for each group (independently), placement test scores were examined using the t-test for paired groups (matched-pair t-test). Thus, error types I and II are eliminated. Everitt (2006, p. 414) defines type I error as "the error that results when the null hypothesis is falsely rejected'', and type II error as "the error that results when the null hypothesis is falsely accepted". For further explanations with examples see, Jupp (2004, p. 307) and Woods, Fletcher,

and Hughes (1986, pp. 115-7). See Appendix 1 for explanation of statistical analyses.

Results

Language Context of Five Colleges based on Programme Evaluation Scores

The major areas consisted of several descriptors each of which was rated on a scale of four points. The highest possible "Programme Context" score is 100 points, which is the multiplication of the highest score of a descriptor (4) by the total number of descriptors in the evaluation form (25). A descriptor would be rated 4 if it fully fulfilled the actual indication of the descriptor and would be rated 1 if it only met the least requirements. This form, admittedly, was neither comprehensive nor was it completely objective; it was merely a logical subjective evaluation. It remained, however, practical for the purpose of abstracting the language programme for each college, and to assign a qualitative weight for each programme. The descriptors were related to the language programme itself, but not to the type of learners, though.

It is evident that the BCCM was, and is still, well catered for and well equipped (refer to Table 1). Other programmes needed some slight improvements. The major difference was the instructional material because it matched the level of the students at the BCCM, whereas it was higher in the rest of the colleges. One may rightly point to another difference which was the class streaming according to the admission placement test but this is not practical for the BCCM since all students at that college study the same instructional material. The instructional materials adopted for each college during the conduct of this study is presented in Table 2.

Table 1
Results of the five language programmes, Jazan University

Criterion and its Descriptors	BCCM	ENGNG	DESIGN	JCCM	JCCF
Intake and Placement					
A. Admission Policy	4	3	3	3	3
B. Classes streamed based on Placement Testing	4	1	1	1	1
Materials					
A. Technology (smart boards, overhead projectors, etc)	4	4	3	3	2
B. Internet accessibility 24hours	4	4	2	3	2
C. Standardised tests	4	2	2	2	2
D. Supplementary material	4	2	2	1	1
E. All materials graded to students' level	4	1	1	1	1

Curriculum					
A. Number of contact hours per week	3	3	3	3	3
B. Specified outcomes	4	4	4	4	4
C. Comprehensive approach to teaching all skills	4	4	4	4	4
D. Daily Homework	4	4	4	3	3
Assessment					
A. Frequency of standardised quizzes	4	3	2	2	2
B. Varied Assessments	4	4	4	4	4
C. Student surveys	4	4	4	4	4
D. Student Interviews	4	2	2	2	2
Teaching and Learning Environment					
A. Teacher to Student ratio	4	3	2	2	2
B. Classroom management policy	4	4	4	4	4
C. Student-centreed, not lecture style	4	3	3	2	2
D. Classroom environment standardization	4	3	3	2	2
E. Class size	4	3	2	2	2
Instructors					
A. Native English speakers	4	2	2	2	2
B. Experienced qualified teachers	4	4	4	3	3
C. Ongoing professional development	4	4	4	4	4
Programme Environment					
A. Quality Assurance system in place	4	4	4	4	4
B. Disciplinary and Absence policies	4	4	4	4	4
Total	99	79	73	67	65

Table 2

Textbooks adopted for Year One, semester one

COLLEGE	Hours/ semester	INSTRUCTIONAL MATERIALS	CEFR LEVEL
BCCM	300	• The New Headway Beginner (2006) • Supplementary material, too	A1
JCCF JCCM	225	• The New Headway Plus Pre-Intermediate (2010)	A2

DESIGN ENGINEE-RING	• Access Interactions, Diamond Edition: Reading and Writing (2009) • Access Interactions, Diamond Edition: Listening and Speaking (2009) • Basic English Grammar, Third Edition, (2006)	A2+

One should note that not all units of these students' books are covered due to the limited amount of contact hours per week, except for the BCCM. A year later of conducting this study, the instructional materials were replaced with more appropriate textbooks, but with one more level above. Contact hours were increased, too.

Language Progress based on Results of First and Second Placement Tests

The ANOVA results for the first placement test in Tables 3 and 4 show that the averages of the colleges are almost identical for the first test. This is an advantage for the second ANOVA in examining the hypotheses of this study. The level "starting" point is identical to the running start line.

Table 3
Summary of ANOVA single factor (test one) for first placement test

	Count	Sum	Average	Variance		
BCCM	17	310	18.24	59.19		
JCCF	25	438	17.52	8.51		
JCCM	32	557	17.41	38.31		
DESIGN	20	413	20.65	39.19		
Engineering	27	525	19.44	30.95		
Source of Variation	SS	Df	MS	F-ratio	P-value	F critical Value
Between Groups	179.85	4	44.96	1.34	0.26	2.45
Within Groups	3888.23	116	33.52			
Total	4068.08	120				

The following conclusions were drawn from the first ANOVA analysis.

1) The Null Hypothesis is not rejected, which states that there are no significant differences among the averages of the five groups at $p<0.05$ (F 4, 116 = 1.34, $p = 0.26$). We can be sure 95% that the minor differences are due to chance. Note that the F-critical value is 2.45, which is higher than 1.34, i.e. the F-ratio.
2) Variance within groups is accepted because they are not grouped according to a pre-set criterion, e.g. placement test. They were selected randomly with varied language levels. Only JCCF shows a small variance.
3) All groups scored lower than the expected level, which we set to be 25 out of 50.

4) The limitation of this test, though reliable, is the small size of the samples.
5) Since ANOVA is designed to examine the significance of differences within and between groups, there remain some data hidden. Therefore, further descriptive statistical analyses were conducted for each group separately. Summaries are tabulated in the following section.

The conclusion for the first ANOVA is that there are no major differences in group averages, and therefore the null hypothesis is accepted. Majority of the students in these colleges almost have similar language levels. They are at the beginning of level 2 (A2) on the Scale of the Common European Framework of Reference (CEFR).

In order to find the true meaning of the means of these groups, we performed some basic descriptive analysis. The conclusions were as follows:

1) All groups were positively skewed to the right, where a majority of the observations laid in the first half of the normal distribution. This means that the majority scored below the average in general.
2) No major differences were observed between the means, modes, and the medians. In addition to this, their standard errors were small, which indicates that these samples can be true representatives of their populations (Hinton, 2004).

Table 4
Placement test one: Results summary for first placement test

	BCCM	JCCF	JCCM	Design	Engineering
Mean	18.24	17.52	17.41	20.65	19.44
Standard Error	1.87	0.58	1.09	1.40	1.07
Median	17	17	16	19.5	18
Mode	18	17	14	19	17
Standard Deviation	7.69	2.92	6.19	6.26	5.56
Sample Variance	59.19	8.51	38.31	39.19	30.95
Skewness	3.17	1.28	1.68	1.85	1.35
Range	37	12	30	26	27
Minimum	9	14	10	14	10
Maximum	46	26	40	40	37
Confidence Level (95%)	3.96	1.20	2.23	2.93	2.20

The confidence limits can easily be affected by single extreme results (Peers, 2006). Therefore, we looked into the interquartile range (IQR) which should clearly show the true range because the medians are not influenced by extreme cases. The Interquartile Range for all groups is given in the table below. The first quartile (Q1) is the first 25% of the sample who scored below the Q1 value, whereas the third quartile (Q3) is the upper 25% who scored above the Q3 value. Table 5 shows that the true range is not high.

Table 5
IQR for first placement test

	1st Quartile (Q1)	3rd Quartile (Q3)	IQR
BCCM	16	18	2
JCCF	15	18	3
JCCM	13.75	20	6.25
DESIGN	16.75	22	5.25
Engineering	16	21	5

Though the first ANOVA analysis for the second placement test results showed identical language levels for the students during admission, certainly there are variations within the students' individual abilities. Because we were interested to measure students' means, a second language test (of the same difficulty as the first one) was administered to the same groups, with a time gap of two months: 120 contact hours for the BCCM and 90 contact hours for the rest of the colleges.

Table 6
Summary of ANOVA single factor (second placement test)

	Count	Sum	Average	Variance		
BCCM	11	238	21.64	96.46		
JCCF	19	317	16.68	9.23		
JCCM	24	485	20.21	40.61		
DESIGN	20	489	24.45	56.47		
Engineering	21	424	20.19	27.86		
Source of Variation	SS	Df	MS	F-ratio	P-value	F critical value
Between Groups	606.63	4	151.66	3.69	0.0079	2.47
Within Groups	3694.80	90	41.05			
Total	4301.43	94				

The first conclusion is that the null hypothesis is rejected in favour of the alternative hypothesis, which states that there is a significant difference among the averages of the five groups at $p > 0.05$ (F 4, 90 = 3.69, p = 0.0079). Because the p-value is so small, we conducted ANOVA at 0.01% level, and it proved to be less than 0.01% (with slight increase for the F-critical value but remained smaller than the F-ratio). Therefore, we can be sure 99% that the differences between these averages are not due to chance, and that they are meaningful.

But the question is: Are these differences due to teaching or individual hard work and self-study? Many latent factors appear to play a role, yet the instructional materials have an influence. The colleges that have instructional materials higher than the students' level showed some progress, except for JCCF which requires further investigation.

Though there was an increase of language learning among all colleges, it was small and less than expected. Follow-up analyses would show that some colleges made better progress than others, and yet the same colleges showed variation among individuals. The question is: Why some students benefitted from the English programme better than their classmates? One may argue that it is due to individual differences (as one factor amongst many) (Dörnyei, 2005; Dornyei & Ryan, 2015). This is noted to be a possible gap that could be addressed in the future.

Descriptive analysis was conducted after the second ANOVA test**.** The high confidence level of BCCM is due to the high range (two extreme cases) combined with the small number of the sample. Therefore, their true average lies between 17.5 and 22.5 (see the IQR, Table 8). High confidence levels may render the conclusions unreliable. The use of the IQR will solve the issue of reliability of conclusions as it will be evident in the argument.

Table 7
Summary of second placement test

	BCCM	JCCF	JCCM	DESIGN	Engg
Mean	21.64	16.68	21.61	24.45	20.19
Standard Error	2.96	0.70	1.56	1.6803	1.15
Median	19	17	20	22	19
Mode	17	18	18	16	14
Standard Deviation	9.82	3.04	6.62	7.52	5.28
Sample Variance	96.46	9.23	43.78	56.47	27.86
Skewness	2.41	-0.75	2.42	0.73	0.50
Range	39	12	30	24	18
Minimum	10	9	14	16	13
Maximum	49	21	44	40	31
Confidence Level (95%)	6.60	1.46	3.30	3.52	2.40

The language ability, according to these results, of the College of Design and Architecture was better than the rest of the colleges. The JCCM showed a good percentage of progress, when compared to the BCCM, despite the fact that they had less contact hours and under-rated programme evaluation points. In short, averages of all groups showed slight improvement, except for JCCF which showed a slight decline. We would not assume that it was due to the curve of language learning, where a learner goes back at one stage, and then moves up but at a higher stage than the first point of the curve until at one stage learning becomes stable, and the curve goes up steadily. Such a process is technically known as restructuring. Restructuring, in brief, "involves knowledge changes that can be large or small, abrupt or gradual, but always qualitative and related to development or progress" (Ortega, 2014, pp. 117-118).

Table 8
Interquartile Range (IQR) for second placement test

	1st Quartile	3rd Quartile	IQR
BCCM	17.5	22.5	5
JCCF	14.5	16.5	4
JCCM	18	23.75	5.75
DESIGN	19	29.5	10.5
Engineering	16	24	8

Matched-pairs t-test was used to compare the increase, if any, for each group. Summary of the results are shown in Table 9. The hypothesised mean difference for all groups is 1 (one). The assumption was that there should be an increase of learning English between the first placement test and the second one. The time lapse between the two tests was almost two months. However, if we are assuming to examine whether there is a difference or not, then the hypothesised mean difference would be 0 (zero).

Table 9
Summary of the t-test statistics for second placement test

	BCCM	JCCF	JCCM	DESIGN	Engineering
Observations	10	19	18	20	21
Pearson Correlation	0.84	-0.27	0.83	0.63	0.77
Degrees of freedom	9	18	17	19	20
t Statistics	-2.27	0.14	-6.32	-3.55	-2.31
P(T<=t) one-tail	0.025	0.445	0.000	0.001	0.016
t Critical one-tail	1.833	1.734	1.739	1.729	1.725
P(T<=t) two-tail	0.049	0.891	0.000	0.002	0.032
t Critical two-tail	2.262	2.101	2.110	2.093	2.086

First, the p-value is less than 5% for all colleges, except for JCCF. Though it remains small (less than 9%), the small negative correlation means that there is no progress at all (in fact a decline). Second, the correlations are quite significant for colleges of BCCM, Engineering and JCCF but slightly significant for Design, though they scored better than the rest. Percentage of progress based on mean difference is given in Table 10.

Table 10
Percentage increase of learning progress during a two-month period between first and second placement tests

	Test 1	Test 2	% increase
BCCM	18.24	21.9	20.1%
JCCF	17.52	16.68	- 4.8%
JCCM	17.41	21.61	24.1%
DESIGN	20.65	24.45	18.4%
Engineering	19.44	20.19	3.9%

Implications and Conclusions

Admittedly, the drawback of this study is that it tested directly grammar and vocabulary as the underlying skills. Other skills were not directly tested, such as speaking and writing. But it was fair to a higher degree because all the five groups received the same treatment. Though all colleges showed some progress, we expected the BCCM to show a higher percentage of progress due to the structure of the programme and the higher contact hours: 120 vs 90 accompanied with some supplementary material (results of the matched-pair t-test). We attributed the language progress of other colleges to the higher level of materials adopted. Instructional materials were about “half a level” higher. However, if the instructional materials are “one full level” higher, they may have a negative effect as it was the case with JCCF. These conclusions are to some extent based on the second ANOVA results which rejected the null hypothesis p >5 (F 4, 90 = 3.69, p = 0.0079).

Placement tests can be used to determine the level of the students at the time of admission, and accordingly colleges may assign materials which are slightly higher than the test outcomes along with extra contact hours. In other words, a language policy must be reinforced upon which the selection of instructional materials and assessment are based.

The study shows a need for further investigation on the role of instructional materials on language progress. A similar study with a standardised test of 100 question items divided evenly between levels of post-elementary (A2), pre-intermediate (B1) and intermediate (B2), can be administered on four groups from different colleges, of which two groups will be treated as controlled groups and will receive instructions higher than their level. Students can be chosen from year three where they have no English classes at this level, except that of the experiment.

References

Ahmad, M. (2008). *Comprehensive dictionary of education*. New Delhi, India: Atlantic Publishers & Dist.

Alobaid, A. (2016). *Testing, assessment, and evaluation in language programmes*. Unpublished Doctoral dissertation, The University of Arizona.

Association of Language Teachers in Europe. (1998). *Multilingual glossary of language testing terms*. Cambridge: Cambridge Universrity Press.

Azar, B., & Hagen, S. (2006). *Basic English grammar* (3rd ed.). London: Pearson Education.

Collins, J., & O'Brien, N. (2011). *The Greenwood dictionary of education* (2nd ed.). California, CA: ABC-CLIO.

Coolican, H. (2014*). Research methods and statistics in psychology* (6th ed.). London: Psychology Press.

Desheng, C. & Varghese, A. (2013). Testing and evaluation of language skills. *IOSR Journal of Research and Method in Education*. *1*(2), 31-33.

Dörnyei, Z., & Ryan, S. (2015). *The psychology of the language learner revisited*. New York, NY: Routledge.

Dörnyei, Z. (2005). *The psychology of the language learner: Individual differences in second language acquisition*. Mahwah, New Jersey, NJ: Lawrence Erlbaum Associates.

Everitt, B. (2006). *The Cambridge dictionary of statistics* (3rd ed.). Cambridge: Cambridge University Press.

Fowler, W. (2005). *Penguin Readers teacher's guide: Placement tests*. Essex, UK: Pearson Education.

Fulcher, G., & Davidson, F (2007). *Language testing and assessment: An advanced resource book*. London: Routledge.

Hinton, P. (2004). *Statistics explained* (2nd ed.). London: Routledge.

Jupp, V. (Ed.) (2004). *The SAGE dictionary of social research methods*. London: Sage Publications.

Kiely, R., & Rea-Dickins, P. (2005). *Programme evaluation in language education*. Berlin, Germany: Springer.

Kirn, E., & Hartmann, P. (2009). *Access Interactions: Reading and Writing* (Diamond Edition: student's book). Maidenhead, UK: McGraw Hill Education.

Kunnan, A. (Ed). (2014). *The companion to language assessment*. Hoboken, NJ: John Wiley & Sons.

Larsen-Freeman, D., & Long, M. (2014). *An introduction to second language acquisition research*. London: Routledge.

Lynch, B. (2003). *Language assessment and programmeme evaluation*. Edinburgh: Edinburgh University Press.

National Commission for Academic Accreditation and Assessment (NCAAA). (2015). *Handbook for Quality Assurance and Accreditation in Saudi Arabia: Part 2* (version 3). Retrieved from https://www.ncaaa.org.sa/en/Releases/HandBook Documents/Handbook%20Part%202.pdf

Norris, J. (2009). Understanding and improving language education through programme evaluation. *Language Teaching Research*, *13*(1), 7-13.

Ortega, L. (2014). *Understanding second language acquisition*. London: Routledge. https://doi.org/10.4324/9780203777282.

Peers, I. (2006). *Statistical analysis for education and psychology researchers*. London: Routledge.

Randolph, K., & Myers, L (2013). *Basic statistics in multivariate analysis*. Oxford: Oxford University Press.

Soars, J., & Soars, L. (2010). *The New Headway plus: Pre-intermediate* (Student's Book). Oxford: Oxford University Press.

Soars, L., & Soars, J. (2006). *The New Headway beginner* (Student's book). Oxford: Oxford University Press

Soh, K. (2016). *Understanding test and exam results statistically*. Singapore: Springer Science.

Tanka, J., & Most, P. (2009). *Access interactions: Listening and speaking*. Maidenhead, UK: McGraw Hill Education.

Verhelst, N., Van Avermaet, P., Takala, S., Figueras, N., & North, B. (2009). *Common European Framework of Reference for languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

Weir, C., & Roberts, J. (1994). *Evaluation in ELT*. Oxford: Blackwell.

Woods, A., Fletcher, P., & Hughes, A. (1986). *Statistics in language studies*. Cambridge, UK: Cambridge University Press.

Appendix 1. Explanation of statistical analyses

Analysis of Variance (ANOVA One Way Single Factor)

ANOVA essentially compares the amount of variation between groups (normally more than two groups) with the amount of variation within groups. If the average difference between groups is similar to that within groups, then the F ratio is about 1. Interpretation is as follows (Coolican, 2014; Hinton, 2004).

- First, as the average difference between groups becomes greater than that within groups, the F-ratio becomes larger than 1, and therefore should be larger than the F-critical value. If the F-critical value is higher than the F-ratio, the null hypothesis is not rejected.
- Secondly, if there is a significant difference among groups, then the P-value becomes smaller than that of the set p-value (which is defined in this study to be < 5%). In other words, the alternative hypothesis is accepted if the p-value is less than 0.05.

T-Test for Paired Groups

A matched-pair t-test (single factor) is defined by the Cambridge Dictionary of Statistics (Everitt, 2006, pp. 250-251) as:

> A Student's t-test for the equality of the means of two populations, when the observations arise as paired samples. The test is based on the differences between the observations of the matched pairs. The test statistic is given by
>
> $$t = \frac{d}{sd\sqrt{n}}$$
>
> where *n* is the sample size, *d* is the mean of the differences, and *sd* is their standard deviation.

To sum up: a *p*-value is used in hypothesis testing to help in supporting or rejecting the null hypothesis. Here we are examining the significance of the difference between means for each group separately, that is the difference between the first scores and the second scores where:

- Null Hypothesis (H_0): $\mu_1 = \mu_2$
- Alternative Hypothesis (H_1): $\mu_1 \neq \mu_2$

The interpretation of p-value of the t-critical value (in the t-test) is similar to the f-critical value of ANOVA, since the latter is essentially a combination of t-tests (Randolph & Myers, 2013, p. 133).

- A small p (≤ 0.05) will reject the null hypothesis. This is strong evidence that the null hypothesis is invalid. It means some learning progress is observed.
- A large p (> 0.05) means that the alternative hypothesis is weak, so the null hypothesis is accepted. Therefore, learning did not take place.