Scripta Neophilologica Posnaniensia, Tom VII, strony: 159 – 179
Wydział Neofilologii, UAM Poznań, 2005

THE EXPLORATION OF THE FUNCTION OF ENGLISH LANGUAGE SONGS IN VOCABULARY ACQUISITION: RESEARCH FINDINGS

ALEKSANDRA WACH

Introduction

Various fields, such as the psychology of music, cognitive psychology, sociology or even anthropology provide solid premises for considering music and songs as conducive to the acquisition of knowledge in general, and the acquisition of a foreign language in particular (For an account of theoretical arguments for using music and songs in learning and teaching, see Siek-Piskozub 2002, Wach 2003).

Contemporary language didactics in Poland seems to witness a paradox: despite the numerous arguments for the wide application of songs in the didactic process, teachers do not seem to give this material adequate consideration, treating it as an end-of-the-day activity and turning to it only occasionally, if at all (Majchrzycka 2000:44-9). One of the possible reasons contributing to this situation may be the fact that there has been very little research confirming the actual usefulness of songs for specific learning outcomes.

The main consideration behind the study which will be described in the present paper was to see whether in a teaching situation, if a song is employed as an alternative context material for vocabulary acquisition, it will serve the purpose well enough. The main objective of the study was thus to measure and evaluate the effectiveness of using English language songs as input for vocabulary introduction and practice, and consequently, their potential for vocabulary acquisition.

1. Research organization and conduct

The research was conducted at Wyższa Szkoła Zarządzania i Bankowości in Poznań and involved two groups, one experimental and the other one – control, of first-year students. The subjects were thus young adults, aged 18-22. There were 18 subjects in the experimental group and 17-19 (depending on the particular stage of the study) in the control group. It lasted for the whole academic year (two terms), with measurements taken at periodic intervals throughout the year.

Due to the fact that the study was incorporated into a regular elementary English course, and although the research program was a systematic and integral element of the course, only part of the syllabus and of the materials were directly connected with the demands of the research. It was assumed, however, that the research and the non-research teaching materials would complement each other and together contribute to the course objectives.

As far as materials are concerned, the experimental group worked on four songs within the research cycle. These were: *"Eternal Flame"* by The Bangles (Study 1), *"Mad About the Boy"* by Dinah Washington (Study 2), *"Cherish"* by Kool And The Gang (Study 3), and *"Drive"* by The Cars (Study 4).

Whenever the experimental group worked on the songs' lyrics, the control group was concurrently exposed to sets of sentences, composed by the teacher, containing the same target vocabulary items. Therefore, four sets of sentences were used in the control group during the study, each corresponding to one song introduced at the same time in the experimental group.

2. Research hypotheses

The main null hypothesis which motivated the study was: **There will not be a considerable difference in the mean scores on a translation test between the subjects who learned vocabulary via the input in the form of songs and the subjects who received more "traditional" instruction.**

Following this, a set of more detailed null hypotheses was formulated:

Hypothesis 1: There will be no significant discrepancies between the results obtained by both groups. The ability to translate selected items will be comparable for both groups, which will prove songs to be as effective input for vocabulary acquisition as the other kind of material used in the study.

Hypothesis 2: Since the vocabulary in the translation tests comes from sets of items previously practiced, to a large extent, on the same tasks and activities (with the exception of the competing materials – songs for the experimental group and sets of sentences for the control group), there will be no difference

between the groups' performance on the English/Polish and Polish/English translation tests. Generally, there may be a tendency for both groups to do better at the English/Polish translation, as the "receptive" direction is generally considered easier for learners.

Hypothesis 3: Since both groups are going to be exposed to the lexical items with identical intensity, only through different materials, which are considered equally effective for vocabulary acquisition, they are both expected to acquire the correct spelling of the items at a comparable level. Thus, they are expected to translate a similar number of items from Polish to English with incorrect spelling.

Hypothesis 4: There will be no observable difference between the groups' performance in the follow-up test, which implies that long-term retention levels will be comparable for both groups. Language gains at all stages (that is, the post-test and the follow-up test) will not be significantly higher for any of the groups.

Hypothesis 5: Since the competing experimental and control materials are supposed to be equally effective, there will be no significant difference in the groups' scores when the design of the study is reversed.

Hypothesis 6: The period at which the material is introduced and the measurements taken will not show a noticeable difference between the groups' levels of achievement. Toward the end of the year the feeling of novelty disappears and the subjects are expected to become generally less enthusiastic about various teaching procedures, including the investigated treatment. Thus, a general decline in achievement is expected, along the same lines for both groups.

3. Quasi-experiment*

Organization

To enhance the reliability of the results, data collection was repeated four times at periodic intervals throughout the academic year – twice a semester, each time using a different song and a different corresponding vocabulary set. In the present description, each vocabulary-development treatment within the quasi-experiment will be called a "Study"; thus, there was Study 1, Study 2, Study 3 and Study 4.

---

* In a quasi-experiment, there is no randomization in the assignment of subjects to groups. The other conditions required in a true experiment, i.e. the existence of two comparable groups, and a pre- and post-test design, are fulfilled (Larsen-Freeman 1991:15, Nunan 1992:230, Cohen and Manion 1994:169).

The subsequent stages were the same at each of the four Studies: there was a pre-test administered to both groups prior to the onset of experimental instruction, the instruction itself, a post-test administered immediately after its completion, and finally, a delayed follow-up test.

The pre-, post- and follow-up tests were identical for both groups. In each test, the following four variables were measured:

1. the number of items correctly translated from English to Polish (Test 1),
2. the number of items correctly translated from Polish to English (Test 2),
3. the number of items translated into English but misspelled (Test 3),
4. the number of items not translated at all (Test 4) (cf. Table 1).

Table 1. briefly summarizes the general scheme of the quasi-experimental design.

Study 1		**Study 2**		**Study 3**		**Study 4**	
Experimental group: **B** Control group: **A**		Experimental group: **A** Control group: **B**		Experimental group: **B** Control group: **A**		Experimental group: **B** Control group: **A**	
Pre-test	TEST 1 TEST 2 TEST 3 TEST 4	Pre-test	TEST 1 TEST 2 TEST 3 TEST 4	Pre-test	TEST 1 TEST 2 TEST 3 TEST 4	Pre-test	TEST 1 TEST 2 TEST 3 TEST 4
Instruction		Instruction		Instruction		Instruction	
Post-test	TEST 1 TEST 2 TEST 3 TEST 4	Post-test	TEST 1 TEST 2 TEST 3 TEST 4	Post-test	TEST 1 TEST 2 TEST 3 TEST 4	Post-test	TEST 1 TEST 2 TEST 3 TEST 4
Follow-up test	TEST 1 TEST 2 TEST 3 TEST 4	Follow-up test	TEST 1 TEST 2 TEST 3 TEST 4	Follow-up test	TEST 1 TEST 2 TEST 3 TEST 4	Follow-up test	TEST 1 TEST 2 TEST 3 TEST 4

Table 1. The scheme of the quasi-experiment design employed in the study

On the basis of the test scores, a statistical inference was drawn. The statistical analysis fell within the General Linear Model (GLM). Analysis of variance techniques (**ANOVA**) of repeated measures was applied, which tested within-subjects effects (that is, the pre-, post- and follow-up test results) as well as between-subjects effects (that is, differences between the experimental and control groups). The ANOVA repeated measures design is considered appropriate to overcome the disadvantages of a small sample (Henning 1986:706).

The general scheme of the hypotheses in the present study is based on the effect of a given variable, that is, on the verification whether there are between-group differences in scores depending on belonging to a given group as well as on the time of measurement. In the study, the significance level for hypotheses testing has been set at a level of α=0.05.

The descriptive statistics (mean, standard deviation) and the tests of within-subjects and between-subjects effects provided by calculations within ANOVA were further enriched by the application of the **t-test**, an inferential statistics tool which helped determine the statistical significance between the groups' scores at particular measures.

Results

Test 1: measuring the number of items correctly translated from English to Polish

TEST 1	Source of variance	SS	df	Ms	F	Sig.
STUDY 1	TIME	283.101	2	141.551	107.179	0.000
	TIME/GROUP	58.187	2	29.093	22.029	0.000
	GROUP	138.960	1	138.960	13.871	0.001
STUDY 2	TIME	95.743	2	47.871	26.127	0.000
	TIME/GROUP	12.283	2	6.142	3.352	0.041
	GROUP	115.175	1	115.175	7.949	0.008
STUDY 3	TIME	93.167	2	46.583	55.595	0.000
	TIME/GROUP	8.662	2	4.331	5.169	0.008
	GROUP	9.170E-02	1	9.170E-02	0.018	0.895
STUDY 4	TIME	101.343	2	50.671	82.872	0.000
	TIME/GROUP	7.181	2	3.590	5.872	0.004
	GROUP	0.137	1	0.137	0.037	0.848

Table 2. Test 1: Analysis of variance for t-level questions (Key: SS = sum of squares, df = degrees of freedom, Ms = mean squares, F = the F-ratio, Sig = the significance level of the F-ratio)

Table 2 presents the levels of significance for both the within-subjects (**time** effect, which refers to pre-, post- and follow-up tests within each group) and between-subjects (**group** and **time/group**) effects. As can be seen, the within-subjects effects have proven significant: both groups' achievement levels are significantly different at each measurement. The between-subjects effects display significant differences as well, apart from the **group** effect in Studies 3 and 4, where the differences appear to be statistically insignificant. This indicates that, as far as Test 1 scores are concerned, in Studies 3 and 4 both groups' achievement was comparable.

Variable	Group A		Group B		t-test		
	M	SD	M	SD	t	df	sig
S1T1pre	3.5882	1.6225	3.0556	2.0714	0.843	33	0.405
S1T1post	9.1765	1.4246	5.0000	2.8901	5.468	25.106	0.000
S1T1follow	7.5294	1.7363	5.3889	2.1458	3.232	33	0.003
S2T1pre	3.5789	2.5888	2.3889	1.7197	1.637	35	0.111
S2T1post	6.0526	2.4827	3.9444	2.6003	2.523	35	0.016
S2T1follow	6.3158	2.5397	3.5000	2.6844	3.279	35	0.002
S3T1pre	1.1579	1.5371	0.3333	0.7670	2.081	26.748	0.047
S3T1post	2.8947	2.2827	3.0556	1.0556	–0.277	25.660	0.784
S3T1follow	1.8421	1.8934	2.3333	0.8402	–1.029	25.112	0.313
S4T1pre	1.4737	1.3068	0.8889	0.9634	1.542	35	0.132
S4T1post	3.2105	1.6859	3.8333	0.9235	–1.403	28.213	0.171
S4T1follow	2.5263	1.5765	2.2778	0.8948	0.594	28.797	0.557

Table 3. Test 1: Descriptive statistics and critical values of t (Key: M = mean, SD = standard deviation, t = the t-value, df = degrees of freedom, sig. = significance)

Statistically significant differences between group A's and group B's scores occur in Study 1 on the post-test and the follow-up test, in Study 2 on the post-test and the follow-up test, and in Study 3 on the pre-test. The mean scores on the remaining measures are not significantly different. It is interesting to note that Studies 1 and 2 (which were designed differently, as in Study 1 the experimental group was group B, while in Study 2, exceptionally, it was group A), bore significant differences in scores, while Studies 3 and 4 did not.

Test 1: Interpretation of the results

Generally, as far as Test 1 results are concerned (which were checking both groups' ability to translate the previously introduced vocabulary items from English to Polish), at the pre-test stage group B was weaker than group A in all four Studies, and in the first two Studies their scores were lower at all points of measurement, regardless of the experimental (Study 1) versus control (Study 2) treatment. As the research continued, however, in the two final Studies (3 and 4), there appeared a tendency for groups A and B to get closer to each other in their levels of scoring. While group B stayed at more or less the same level throughout the research period, there is an easily observable decrease in group A's overall scores toward the end. While in Study 1, conducted at the beginning of the year, group A's mean scores on the post-test almost reached the maximum level of 10 points, in Studies 3 and 4, conducted in the second semester, they barely reached 3 points (where the maximum score was 7). Moreover, the rising tendency at the follow-up stage displayed by the experimental group in Studies 1, 2 and 3 may suggest that they retained the vocabulary items more effectively over the three weeks between the post-test and the follow-up test when neither of the groups worked on the items any more.

To further sum up the results of Test 1, the English-Polish translation test, and to further compare both groups' performance, Figure 1 illustrates the results of the post-tests and the follow-up tests for all the four Studies in both groups.

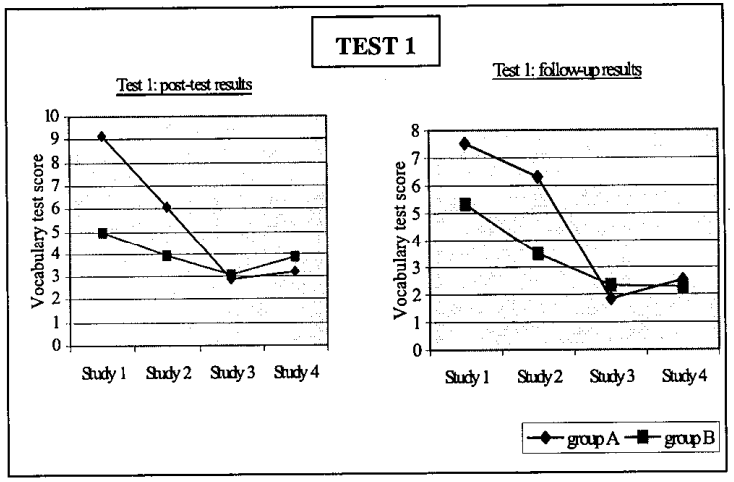

Figure 1. Test 1: Post-tests and follow-up tests results

Both on the post-tests and on the follow-up tests, group A displays a dramatic fall in results down to Study 3, from which the results rise slightly at Study 4. Interestingly, Study 3 appears to be the most critical one for both groups. There is a fall in group B's scores from Study 1 to Study 3 as well, but it is not so considerable. The rise of group B's scores in Study 4 is bigger than for group A, which makes the overall levels of achievement comparatively steady for group B, which is not the case for group A. The post-test results in Study 2, when the format of the study was reversed, do not demonstrate any considerable change.

Test 2: measuring the number of items correctly translated from Polish to English

Test 2: Results

TEST 2	Source of variance	SS	df	Ms	F	Sig.
STUDY 1	TIME	107.951	2	53.975	27.369	0.000
	TIME/GROUP	4.294	2	2.147	1.089	0.349
	GROUP	69.469	1	69.469	3.254	0.080
STUDY 2	TIME	34.619	2	17.309	13.744	0.000
	TIME/GROUP	5.141	2	2.571	2.041	0.138
	GROUP	48.722	1	48.722	9.669	0.004
STUDY 3	TIME	54.429	2	27.215	27.116	0.000
	TIME/GROUP	3.222	2	1.611	1.605	0.208
	GROUP	2.450	1	2.450	0.354	0.555
STUDY 4	TIME	26.073	2	13.037	16.979	0.000
	TIME/GROUP	1.569	2	0.784	1.022	0.365
	GROUP	0.761	1	0.761	0.201	0.657

Table 4. Test 2: Analysis of variance for t-level questions (Key: SS = sum of squares, df = degrees of freedom, Ms = mean squares, F = the F-ratio, Sig = significance of the F-ratio)

For Studies 1, 3 and 4, that is, in the designs when group A was the control and group B was the experimental one, only the within-subjects **time** effect proved to be significant. This points to apparent differences within each group at particular measurements. For Study 2, however, when the design of the quasi-experiment was reversed and group A was the experimental one, both the within-subject **time** effect and the between-subjects **group** effect are significant, which indicates that group A and group B significantly differed in achievement in this Study (Table 4).

Variable	Group A		Group B		t-test		
	M	SD	M	SD	t	df	sig
S1T2pre	3.9412	2.6094	2.8333	2.4793	1.288	33	0.207
S1T2post	6.7059	3.2933	4.6111	3.1462	1.925	33	0.063
S1T2follow	6.2353	3.1131	4.5556	2.7056	1.707	33	0.097
S2T2pre	1.0526	1.0788	0.3333	0.5941	2.492	35	0.018
S2T2post	2.7895	2.2504	1.1111	1.4507	2.710	30.950	0.011
S2T2follow	2.5789	2.1165	1.0000	1.3284	2.700	35	0.011
S3T2pre	2.0000	2.0000	1.2222	1.1144	1.471	28.491	0.152
S3T2post	3.3684	2.2903	3.2778	1.2744	0.150	28469	0.882
S3T2follow	2.5789	1.8048	2.5556	1.5038	0.43	35	0.966
S4T2pre	1.4211	1.0706	1.1111	1.3235	0.785	35	0.438
S4T2post	2.5789	1.2164	2.2222	1.5925	0.768	35	0.447
S4T2follow	2.0526	1.4710	2.2222	1.2628	–0.377	34.678	0.709

Table 5. Test 2: Descriptive statistics and critical values of t (Key: M = mean, SD = standard deviation, t = the t-value, df = degrees of freedom, sig. = significance)

The t-value proves to be statistically significant only in the case of Study 2, when the organizational format was changed and group A was the experimental group. In Study 2, the means are statistically different at all three measurements: the pre-test, the post-test and the follow-up test. At all remaining measurements within Test 2, the differences in scores obtained by groups A and B are not statistically significant (Table 5).

Test 2: Interpretation of the results

The general conclusions for Test 2 are in a few aspects similar to those concerning Test 1. The stronger group (group A) scored objectively higher, regardless of the treatment. Moreover, again, the discrepancies between the groups are greater in Studies 1 and 2, at the beginning of the research period. The results of Studies 3 and 4 (conducted in the second semester) indicate, however, that the originally weaker experimental group (group B) at the level of post-test and follow-up approximates or even outscores the stronger one. Finally, the tendency for the experimental group to retain more vocabulary items at the follow-up stage is confirmed in Test 2 as well. The differences are, however, statistically insignificant, apart from Study 2, when group B's (the control group) scoring was the lowest of all remaining Studies.

A comparative analysis of the post-tests and follow-up tests results within Test 2, measuring both groups' ability to translate the items from Polish to English (cf. Figure 2), highlights a considerable difference between Studies 1 and 2 in the case of both groups. Study 2 (where the design of the study was reversed) marks a significant fall in the results, parallel for both the control and the experimental groups. Group A's results are maintained at approximately the same level for Studies 2, 3 and 4, while group B's scores, after the sudden fall at Study 2, reach considerably higher levels at Studies 3 and 4, equaling group A's results. It is apparent from this data that the weaker group, group B, reacted significantly worse as the control group, while for group A, the initially stronger one, the change in the control versus experimental treatment did not bring about a significant difference in scoring.

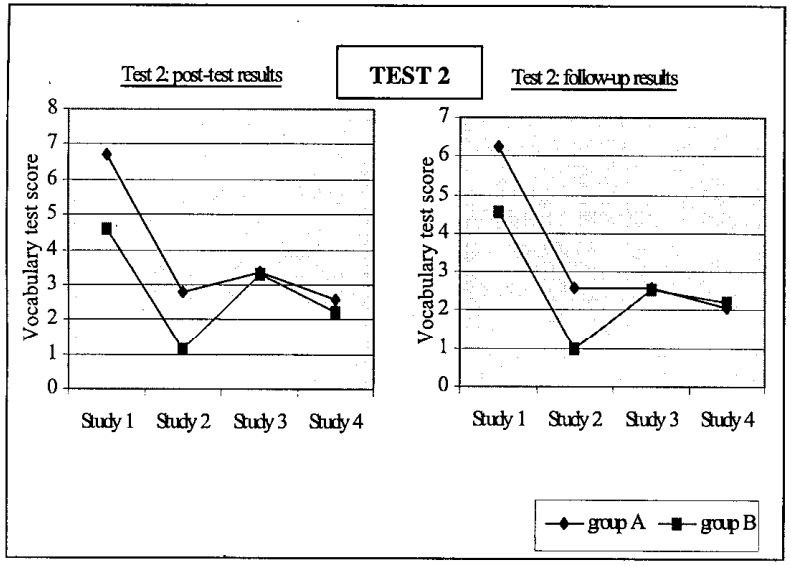

Figure 2. Test 2: post-tests and follow-up tests results

Test 3: measuring the number of items translated from Polish to English, but misspelled

Test 3: Results

TEST 3	Source of variance	SS	df	Ms	F	Sig.
STUDY 1	TIME	3.388	2	1.694	1.876	0.161
	TIME/GROUP	0.302	2	0.151	0.167	0.846
	GROUP	13.640	1	13.640	1.830	0.185
STUDY 2	TIME	3.434	2	1.717	5.793	0.005
	TIME/GROUP	1.055	2	0.528	1.780	0.176
	GROUP	7.968	1	7.968	12.795	0.001
STUDY 3	TIME	0.334	2	0.167	0.383	0.683
	TIME/GROUP	0.478	2	0.239	0.549	0.580
	GROUP	1.607	1	1.607	1.031	0.317
STUDY 4	TIME	0.933	2	0.466	1.020	0.366
	TIME/GROUP	1.797	2	0.899	1.967	0.148
	GROUP	2.387	1	2.387	1.128	0.295

Table 6. Test 3: Analysis of variance for t-level questions (Key: SS = sum of squares, df = degrees of freedom, Ms = mean squares, F = the F-ratio, Sig = significance of the F-ratio)

For Test 3, which was measuring the number of items translated into English but misspelled, again, as for Test 2, the only Study in which statistically significant levels are observed is Study 2, in which both the within-subjects **time** effect and the between-subjects **group** effect have proven significant. In all remaining Studies neither the within-subjects nor the between-subjects effects are statistically significant (see: Table 6). This indicates that the independent variable (the wrong spelling of items) proved not to be significant in most of the Studies.

Variable	Group A		Group B		t-test		
	M	SD	M	SD	t	df	sig
S1T3pre	2.4706	1.9078	1.6111	1.5005	1.486	33	0.147
S1T3post	2.7059	2.3121	2.0000	1.5339	1.070	33	0.292
S1T3follow	2.7647	1.9852	2.1667	1.0981	1.111	33	0.274
S2T3pre	0.3158	0.5824	0.0556	0.2357	1.798	24.003	0.085
S2T3post	0.7895	0.8550	0.1111	0.3234	3.224	23.285	0.004
S2T3follow	0.9474	0.8481	0.2778	0.6691	2.656	35	0.012
S3T3pre	0.7368	1.0457	0.3333	0.6860	1.379	35	0.177
S3T3post	0.7368	1.0457	0.5000	0.7071	0.802	35	0.428
S3T3follow	0.5263	1.1723	0.4444	0.5113	0.273	35	0.787
S4T3pre	1.0526	0.9703	0.5000	0.8575	1.832	35	0.076
S4T3post	1.1579	1.3443	0.7778	0.9428	0.991	35	0.329
S4T3follow	0.9474	0.7050	1.0000	1.0847	-0.176	35	0.861

Table 7. Test 3: Descriptive statistics and critical values of t (Key: M = mean, SD = standard deviation, t =. the t-value, df = degrees of freedom, sig. = significance)

The levels of significance displayed by the t-test appear to exactly confirm the initial assumptions revealed by the analysis of variance; the t-value is statistically significant only in the case of the post-test and the follow-up test in Study 2 (see: Table 7).

Test 3: Interpretation of the results

As far as Test 3 results are concerned, the number of items translated with incorrect spelling was generally very low. Moreover, in almost all cases neither the within-subjects nor the between-subjects factors proved to be statistically significant. An exceptional situation occurred when group A was the experimental one; then the difference in the level of knowledge was significant – group A scored significantly higher. In the remaining Studies within Test 3, all that can be observed is a certain tendency for the experimental group, group B, to remember comparatively more words without recalling their correct spelling than group A, which is reflected in their rising scores at the post-test and the follow-up in Study 1, Study 3 and Study 4.

Figure 3 illustrates how the post-tests and the follow-up tests results within Test 3 for both groups compare.

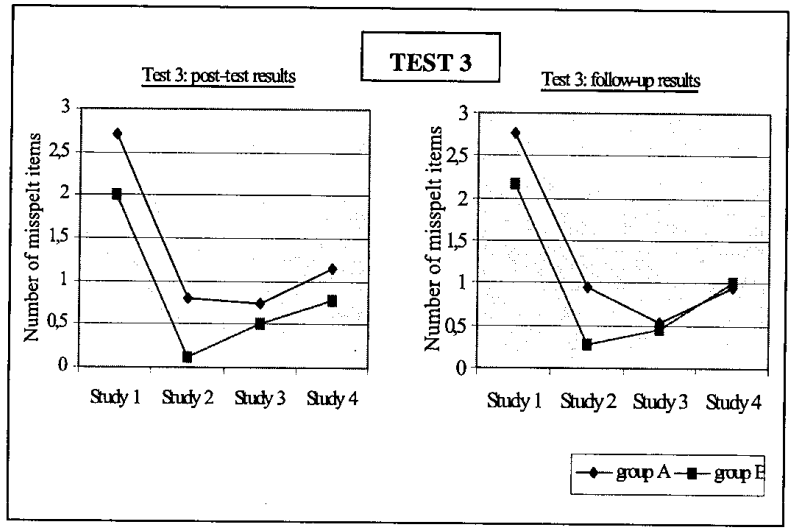

Figure 3. Test 3: post-test and follow-up tests results

It is interesting to note how the lines for both groups are surprisingly parallel on the post-tests and the follow-up tests (Figure 3). The number of items whose English equivalents with an incorrect spelling were provided was comparable for both groups, though in all models higher for group A, irrespective of the treatment. Both on post-tests and follow-up tests, the only statistically significant difference occurred in Study 2; apparently, the claim that working on a song enhances the acquisition of the phonetic forms of vocabulary items is confirmed for the stronger group, group A.

Test 4: measuring how many items were not translated at all

Test 4: Results

TEST 4	Source of variance	SS	df	Ms	F	Sig.
STUDY 1	TIME	820.629	2	410.314	102.868	0.000
	TIME/GROUP	97.086	2	48.543	12.170	0.000
	GROUP	553.998	1	553.998	11.237	0.002
STUDY 2	TIME	295.818	2	147.909	29.021	0.000
	TIME/GROUP	44.395	2	22.197	4.355	0.016
	GROUP	429.733	1	429.733	12.068	0.001
STUDY 3	TIME	300.845	2	150.423	67.559	0.000
	TIME/GROUP	28.989	2	14.495	6.510	0.003
	GROUP	9.834	1	9.834	0.492	0.488
STUDY 4	TIME	248.591	2	124.296	75.132	0.000
	TIME/GROUP	13.636	2	6.818	4.121	0.020
	GROUP	5.794	1	5.794	0.329	0.570

Table 8. Test 4: Analysis of variance for t-level questions (Key: SS = sum of squares, df = degrees of freedom, Ms = mean squares, F = the F-ratio, Sig = significance of the F-ratio)

The analysis of variance for Test 4 reveals significant differences as far as practically all effects, within-subjects and between-subjects, are concerned. In Studies 1 and 2 all effects are significant with no exceptions. In Studies 3 and 4, however, the situation looks different: here, the between-subjects **group** effect turns out to be statistically insignificant (Table 8). This clearly indicates that the situation is different in the Studies conducted in the first term of the school year and those carried out in the second term, when both groups' achievement is no longer drastically different.

Variable	Group A		Group B		t-test		
	M	SD	M	SD	t	df	sig
S1T4pre	11.0000	3.9211	13.3333	4.7154	–1.587	33	0.122
S1T4post	2.3529	2.9568	9.3889	5.5108	5.468	25.106	0.000
S1T4follow	4.4706	3.7101	8.8889	4.7883	3.232	33	0.003
S2T4pre	16.0526	3.7635	18.2222	2.2375	–2.144	29.578	0.040
S2T4post	11.3684	4.5730	15.9444	3.7803	–3.307	35	0.002
S2T4follow	11.1579	4.3239	16.2222	4.2503	–3.590	35	0.001
S3T4pre	10.1053	3.2981	12.1111	1.6410	–2.361	26.706	0.026
S3T4post	7.0000	4.1231	7.1667	2.1213	–0.156	27.214	0.877
S3T4follow	6.0526	3.1530	8.6667	1.7489	0.464	28.417	0.646
S4T4pre	10.0526	2.6347	11.5000	2.7062	–1.648	35	0.108
S4T4post	7.1579	2.8139	7.1667	2.2816	–0.010	35	0.992
S4T4follow	8.4737	3.0617	8.3889	2.2000	0.096	35	0.924

Table 9. Test 4: Descriptive statistics and critical values of t (Key: M = mean, SD = standard deviation, t = the t-value, df = degrees of freedom, sig. = significance)

The t-test significance, which considers the means at particular measurements, is concordant with the previous findings. Significant differences between means are found at the post-test and follow-up tests in Studies 1 and 2, which points to significantly different levels of achievement and knowledge gains between the pre-test and the following measurements for group A and group B. In Studies 3 and 4, however, no statistical significance is found at either the post-test or the follow-up test. Achievement appears to be comparable for both groups on these measurements (see: Table 9).

Test 4: Interpretation of the results

Test 4 results are obviously closely connected with the results of the previous tests, namely Test 1, Test 2 and Test 3, which, taken all together, measured the number of items **translated**, with or without their correct spelling, from English to Polish or from Polish to English. Test 4 thus complements and further verifies the overall picture by checking how many items were not translated at all in the four Studies. The results presented above call for general comments similar to those made on the outcomes of the previous tests. Group B (the experimental group in three out of the four Studies), was not able to translate more items on the pre-, post- and follow-up test in the initial Studies (1 and 2) conducted in the first semester of the quasi-experiment. The situation looks different in the two final Studies (3 and 4), conducted in the second semester. Though still invariably weaker on the pre-tests, group B's scores on the post-tests are on almost the same levels as group A's, and on the follow-up tests group A's scores are slightly lower, which means that group B left fewer items not translated than group A.

Figure 4 presented below summarizes all post-tests and follow-up tests results for both groups.

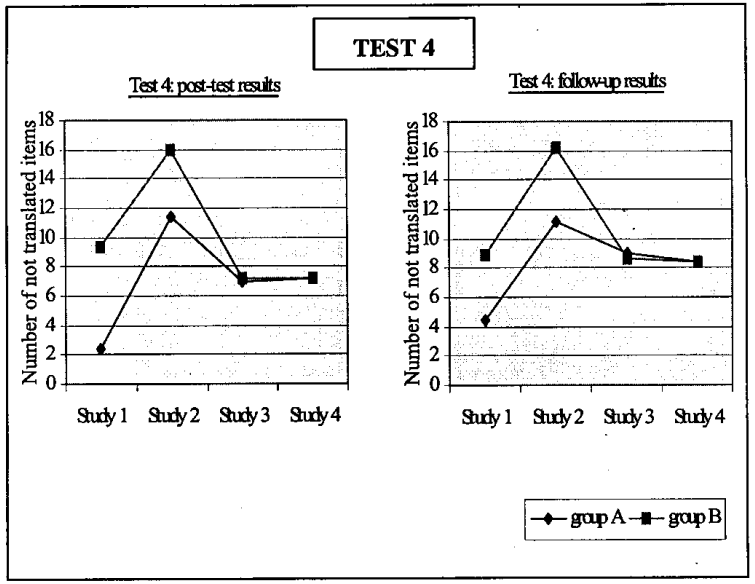

Figure 4. Test 4: post-tests and follow-up test results

The agreement with the results of the post-tests and follow-up tests within Tests 1, 2 and 3 is confirmed here. As can be seen (cf. Figure 4), the number of items which were not translated at all in group A is the lowest for Study 1 and the highest for Study 2 (where the design of the experiment was changed). Study 2 seems to be the critical point for both groups (the situation was similar in Test 2). In Studies 3 and 4 the number of words not translated at all by group A gets lower again, but not as low as at the beginning of the experiment. For group B, the number of items not translated at all is much higher in Studies 1 and 2 than for group A, Study 2 being the critical one as well. Studies 3 and 4, however, present a difference: the number of words not translated gets considerably lower for group B and than in the two previous studies, while for group A the levels in Studies 3 and 4 are higher than in Study 1, but lower than in Study 2. As a result, the scoring for both groups looks very similar. This means, however, that while for group B the number of items not translated at all got lower toward the end of the quasi-experiment, for group A it got higher in Studies 3 and 4 than in Study 1, that is, in the Studies when it was the control group.

Another interesting observation is that apart from Study 4, the experimental group always displays a certain rise in scoring between the post-test and the follow-up test. On the contrary, in the control group this is never the case. This clearly points not only to the greater retention of items over a longer period of time in the experimental group, but also to further rise in vocabulary levels in the period when the items were no longer practiced in the classroom.

4. Quasi-experiment findings: interpretation and conclusions

On the basis of the quasi-experiment results presented in the previous section, the six null hypotheses formulated at the onset of the study have been verified.

Hypothesis 1 (which assumed the equality between both groups' scores on vocabulary translation tests) is confirmed in the study. The ability to translate selected items, checked on the post-tests administered directly after vocabulary introduction and practice, proved to be comparable for both groups. Although group A scored higher in most instances, regardless of the experimental versus control treatment, the lines indicating both groups' scores for all tests do not indicate significant discrepancies which would make it possible to reject the hypothesis. Group A's performance on all tests within Study 1 and Study 2 was objectively significantly higher (the between-subjects variables were statistically significant for these Studies, too, while they were not for Studies 3 and 4), however, the discrepancy cannot be attributed to the use of the competing materials at the stage of instruction. Firstly, in the remaining Studies this pattern was not repeated; secondly, it was clear throughout the research that group A proved to be stronger and thus their better performance should be interpreted with additional caution. The four Studies did not yield results which would allow the rejection of Hypothesis 1 or a claim that either of the competing materials was more effective for vocabulary acquisition. Although certain tendencies emerged, as far as the general ability to translate vocabulary items is concerned, both groups progressed along similar lines. This may be attributed to the fact that the experimental group subjects, who worked on a song, due to the practice activities they did in class, were able to separate the items from the context of the song and acquire them as independent entities.

Hypothesis 2, assuming no significant difference between both groups' scores on L1-L2 and L2-L1 translation tests, was confirmed in the study as well. Although the groups worked on different materials, there was not much difference between their results in Test 1 (measuring the number of items translated from English to Polish) and Test 2 (Polish-English translation), although Test 1 results are higher

for both groups than Test 2 results, especially in Studies 1 and 2. Again, although group A's achievement was generally higher, neither of the groups was outstandingly better or substantially weaker at either the English-Polish or the Polish-English translation tests. This finding leads to the conclusion that the subjects in both groups, since they did the same consolidating tasks, could separate the items from their different contexts equally well and progressed along similar lines.

Hypothesis 3 (assuming equal levels of items misspelled on the tests in both groups) has found partial confirmation in the quasi-experiment. It was expected that the experimental group would generally do better on Test 3, which measured the number of items translated from Polish to English but misspelled. The differences between both groups' mean scores on Test 3 were, in three out of the four Studies, found to be statistically insignificant. Only in one instance, namely when the design of the study was reversed and group A was the experimental group, were the differences in the mean scores statistically significant. The overall number of misspelled items in all Studies was, in fact, very small. On the other hand, the number of items to be translated was not big at all; any patterns and tendencies could thus be observed through the four repetitions of the measurements. Therefore, even modest, though recurring, differences, can provide important insight into the results.

Hypothesis 4, which assumed no observable differences in the groups' performance on the follow-up test, has to be rejected. Naturally, the follow-up test scores must not be considered in isolation, because group A's initial level, reflected in the pre-tests for all the Studies, invariably turned out to be higher; there was not a single pre-test within the quasi-experiment on which group B's score would have been higher. Therefore, as for other results obtained in the present research, the findings need to be interpreted in ways other than a simple comparison of the mean scores.

One of the approaches that can be applied is calculating knowledge gains for both groups and comparing these, as this technique will take into account the different initial level of the groups and will consider the actual relatively measured progress. Table 10 presents the post-tests/follow-up tests as well as the pre-tests/follow-up tests differences in mean scores for both groups obtained in the English-Polish translation tests and the Polish-English translation tests (Test 2 and 3 results are calculated together) for all four Studies.

		Study 1		Study 2		Study 3		Study 4	
		exper.	control	exper.	control	exper.	control	exper.	control
TEST 1	post-test/ follow-up	**0.3**	–1.7	**0.2**	–0.4	**–0.8**	–1.1	–1.5	**–0.7**
	pre-test/ follow-up	2.2	**3.9**	**2.7**	1.1	**2.0**	1.4	**1.4**	1.0
TEST 2 and TEST 3	post-test/ follow-up	**0.2**	–0.4	–0.1	**0.1**	**–0.7**	–0.8	**0.2**	–0.8
	pre-test/ follow-up	2.4	**2.6**	**2.2**	0.9	**1.5**	0.5	**1.6**	0.5

Table 10. Post-test/follow-up and pre-test/follow-up mean score differences on Test 1 and on Tests 2 and 3 (calculated together)

In Table 10, the bigger differences in each control versus experimental pair have been highlighted. Again, though the differences are small, the gain between the post-test and the follow-up test for the experimental group was higher on 6 out of 8 tests, and similarly, in 6 out of 8 instances when the pre-test/follow up is concerned. Therefore, although group B's scores were objectively lower, its **overall knowledge gain was generally higher**. The post-test/follow/up test knowledge gain levels indicate that the vocabulary items were generally better remembered in the experimental group; the experimental group thus seems to have forgotten the meaning of fewer items in the period between the post-test and the follow-up, when the items were not practiced in the classroom any more.

Hypothesis 5, concerning the anticipated differences in the groups' scores when the design of the experiment was changed, was not confirmed by the results of the study, at least not to the extent expected. The reason behind the procedure of group-swapping in one of the Studies was to make the impact of the application of different materials more obvious and more easily observable, and this did not quite turn out to be the case. The results of Study 2, which illustrate the groups' results on all tests in which group A was experimental and group B was the control, do not reveal any considerable or easily noticeable differences between the two groups which could readily lead to obvious conclusions. However, on looking closer, it becomes clear that, in this Study, unlike most other cases, the knowledge gain between the pre-test and the follow-up is noticeably bigger for group A than for group B. Another difference, though not one that was predicted, becomes visible when the graphs illustrating the post-test and the follow-up test results are analyzed. For Tests 2, 3 and 4, Study 2 (when the design in the format was reversed) is a critical point for both groups, at which group B scored at the lowest levels in the whole quasi-experiment. It is certainly not easy to provide a

convincing objective explanation for this situation, and there is too little data for making generalizations. The results might have been different if the reversed design of Study 2 had been introduced or repeated at the end of the quasi-experiment period, when both groups were more used to working with "their" types of materials. Here, since this was the second measurement, the effect of novelty might have proven distracting and the subjects might not have treated the materials seriously enough. Additionally, the experimental group subjects might have not liked the song or the vocabulary might have been too difficult. These factors can never be totally excluded, especially with a song, which always appeals to the listeners' emotions. Unfortunately, this type of measurement was not repeated, so none of these claims was further validated.

Hypothesis 6 – the final hypothesis, which assumed no significant difference between particular groups' lines of achievement throughout the research period, has to be rejected in view of the results obtained in the course of the quasi-experiment. The period of time in which the materials were introduced and the tests conducted did prove to be important. The linear progression in the groups' levels of achievement reveals some interesting findings. Group A's scores got gradually lower toward the end of the school year: in the first semester, their mean scores were higher than group B's at all measurements (the pre-test, post-test and the follow-up on all the four tests, in Studies 1 and 2), whereas in the second semester, the situation looked different. Group A was still stronger on the pre-tests, however, their scores on the post-test and the follow-up test got comparatively lower, equaling group B's levels or even being outscored by group B's results. While group B stayed at more or less the same level throughout the year, group A's achievement decreased considerably in the second semester. This is reflected in the statistical significance levels for the between-subjects 'group' variable (cf. Tables 2, 4, 6 and 8). In Study 1 and 2 both kinds of variables ('time of measurement' and 'group') are statistically significant, which means that within each group there were significant differences in scores on the pre-, post- and follow-up tests as well as between the groups' scores. In Studies 3 and 4, however, the 'time' variable is still statistically significant (apart from Study 3 Test 3), while the 'group' variable is not. This indicates that both groups achieved, statistically, very similar scores.

The results may imply group A's gradual dampening of enthusiasm as the school year continued. It seems that at the beginning of the year group A students were still full of energy, conscientious about their work and interested in anything they were doing in class, especially outside-coursebook materials such as the vocabulary development handouts they were working on. The high knowledge gain between the pre-test and the post-test in Test 1 and 2 suggests that the

subjects must have been revising the vocabulary at home, probably assuming that they would be formally assessed. With time, however, in the second semester much of their initial enthusiasm seems to have disappeared, which may be partly due to the fact that they were given no grades for the tests within the quasi-experiment. The considerably lower levels of increase in Studies 3 and 4 could imply that the students stopped revising the vocabulary between sessions; most probably, in the tests they wrote what they remembered without focused attention or direct revision. The situation seems different for group B. Their progress remained at roughly the same level throughout the quasi-experiment. Although their initial enthusiasm must have worn out with time, too, it did not in the case of songs. It looks like working with songs stimulated them to the same extent at the beginning and at the end of the school year. The generally lower levels of achievement could indicate that the subjects did not consciously work on the material much, if at all, and the situation was the same throughout the time of the quasi-experiment. Consequently, group B's vocabulary learning was "incidental" throughout the study, while for Group A it seems to have been "intentional" at the beginning and "incidental" toward the end.

Therefore, in view of all the research findings described above, the main research hypothesis that foreign language songs are as effective materials for vocabulary acquisition as other, more 'traditional' materials, seems to be confirmed for the two groups which took part in the quasi-experiment. Moreover, the results of the translation tests do imply that in certain aspects songs turned out to be more effective, for example, when the overall knowledge gain is concerned or loss due to forgetting. This may be called, however, a tendency rather than an objective truth. In quasixperimental designs, in principle, some of the variables are difficult to control and the results are not as generalizable as those of a true experiment. They do, however, gather information and reveal tendencies (Larsen-Freeman and Long 1991:19-21, Nunan 1992:230). This is precisely how the findings revealed by the present study should be interpreted.

Bibliography

Cohen, L. and L. Manion. 1994. *Research Methods in Education.* London and New York: Routledge.

Henning, G. H. 1986. "Quantitative Methods in Language Acquisition Research." *TESOL Quarterly 20/4.* 701-8.

Larsen-Freeman, D. and M. H. Long. 1991. *An Introduction to Second Language Acquisition Research.* London: Longman Group UK Limited.

Majchrzycka, A. 2000. "Teachers' and Learners' Attitudes Toward Songs in EFL: Results of a Study." *Network 3/1.* 43-51.

Nunan, D, 1992. *Research Methods in Language Learning.* Cambridge: Cambridge University Press.

Siek-Piskozub, T. 2002. *Umuzykalnienie glottodydaktyki. Muzyka i piosenka na lekcji języka obcego.* Poznań: Motivex.

Wach, A. 2003. "Music and song as material conducive to foreign language acquisition: theoretical underpinnings." *Neofilolog 22.* 82-91.