English Language Proficiency Levels in Two Programs COMPARED OVER FOUR SEMESTERS

Thomas M. PENDERGAST

Abstract

This paper details, analyzes, and comments on the results of the CEEL[1] Test Battery of four components (listening, spoken fluency, correctness, and expression), administered to students at the ends of semesters 1-4 in two separate English language major programs. One program was integrated and followed a set, department-wide, self-access curriculum[2]. The other was eclectic, with each instructor choosing his own materials and procedures. The aim of the study was to identify beginning college students of similar, non-functional benchmark levels in two different programs, and to assess, analyze, and compare their subsequent levels at the end of each of four semesters.

Listening comprehension was tested with a channel capacity[3] instrument. Spoken expression testing was done with the help of a dedicated, hand-held computer nicknamed OLAF, for Oral Language Analysis and Feedback system.

Key Words: language learning, self-access, SAPL, proficiency testing, t-test, N73, channel capacity, OLAF

Introduction

A paper last year (Pendergast, 2009) presented proficiency gains data for a language program, echoing the refrain of Robert Pear (Pear, 2009) of the New York Times, known for his efforts on behalf of “comparative effectiveness research” in the area of medical treatment. What works? What’s efficient? Where’s the evidence?

Recently Bob Wachter posted a blog in the series “Wachter’s Watch,” in which he asked on February 21, 2010, whether (most people) can politically/personally/financially accept the TRUTH about various (medical) treatments.

What about language learning? Can we as instructors and administrators with vested interests stand the harsh light of investigation?

This paper looks at five stages of two college English language programs.

Program A has an integrated core in which all instructors are trained in a student-centered learning system systematically applied over four semesters. Students are rigorously tested at the end of each semester with a proficiency test battery which

provides a common yardstick for all students.

Program B allows each instructor to choose his own learning material and procedures. There is no department-wide evaluation system common to all students.

In the earlier paper mentioned above, I insisted that a valid comparison required upholding certain conditions. One was that students be of a similar beginning level, preferably non-functional, false beginners. Another was that there be no selection of students or scores, in order to avoid skewed reporting … in other words, that ALL subjects in a given program be tested and reported on.

This paper looks at things from a slightly different angle. The emphasis is not on pairing up pre- and post-test results, but on establishing approximate levels in two programs at five given times (pre-program and at the end of each of four semesters), and comparing them. The comparison is to some extent *pro forma* as the number of Program B subjects is insufficient for firm conclusions. A further interest however is in seeing if partial samples of randomly chosen subjects taken from one program over several years vary significantly … or not … and whether combining such samples into one larger cohort provides stronger evidence for level identification. This will be looked at more thoroughly in the future.

Program A is here represented by the students reported on in the earlier paper, when the emphasis was on proficiency gains from pre-test to post-test rather than on end-of-semester levels. Program B is a random group of subjects from a totally different program, chosen for their availability as representatives of the program. As can be seen from the numbers of students in Program B at each level, the situation is not ideal. In submitting this report, I accept that no firm conclusions can be drawn about the difference in efficiency of the two programs but suggest that the approach to evaluation be considered on its own merits or demerits, as a framework. As will be seen, the paper is heavy on statistics and perhaps overly dense. It represents much of what I would like to know about a program.

Channel Capacity Test N73 (Listening Comprehension)

Practical aspects of testing

Subjects (Ss) listen to 15 sentences recorded on tape, with pauses between each sentence. The first sentences are short, but they become progressively longer. Ss repeat and record what they can in spaces on the tape (N.B., immediately after the end of each sentence, not simultaneously). Each word repeated receives one point. The total number of points is scaled to a norm with the help of a template (Ferguson, 1982). There are 30 sentences and 404 words on the entire test, but there are cut-off points at 10, 15, and 25 sentences if a specified level (Ferguson, 1978) has not been reached. For practical purposes, 15

sentences are given initially, as no beginning college student in our experience of the programs under investigation has ever required more at the pre-test evaluation. During evaluation, some subjects are not evaluated beyond sentence 10 for lack of performance up to that point. Inter- and intra-evaluator reliability is very high (95%+, forthcoming). It should be noted that occasional (not in the present report) subjects require 25 sentences at the end of the fourth semester to accurately reflect their growth in the language.

Benchmark

Figure 1

N73 Channel Capacity **(Listening Comprehension)**	Program A	Program B
	Benchmark	**Benchmark**
Date	Apr., 2003	Apr., 2009
Number of subjects	57	10
Test time	3.5mins.	3.5mins.
Number of sentences given	15	15
Number of items	141	141
Maximum possible scaled score	1000	1000
Mean scaled score	32	33
Standard Deviation (StD)	18.800	22.321
K-R21 Reliability	0.913	0.937
Standard Error of Measurement	2.490	7.059
Unpaired t test two-tailed P value	0.8805	
"Significantly" different?	"No"[4]	
Median	32	40
High	75	62
Low	0	0

Comment on Figure 1: The testing dates are at variance due to the fact that Program A no longer exists. All Program A data represent the last class with a sufficient number of subjects to report on. Both samples were initially evaluated (pre-tested) in their first week of matriculation, as a benchmark. The testing instrument was the N73, described in the note and in Pendergast, 2009. The high reliability is due to "a very refined choice of items, and partly due to the large number of items presented in a very short time: 141 in 3.5 minutes; 483 in 8 minutes." (Ferguson, 1978). The present subjects were given 15 (out of a total 30) sentences to repeat on tape. There are 141 words (items) altogether in this selection. If 42 or fewer items are successfully repeated, the test is cut off at ten sentences

and the score is scaled. If the testee is able to repeat only 33 or fewer items, the score is scaled to zero. The t test suggests no significant difference between the groups in Figure 1, with mean scaled scores of 32 and 33, respectively. At this level, the average subject is essentially non-functional in English, as the minimum functional, scaled score is 50 (tourist survival) out of a maximum of 1000 (native equivalent).

First and Second Semesters

Figure 2

N73 Channel Capacity **(Listening Comprehension)**	Program A **End 1st Sem**	Program B **End 1st Sem**	Program A **End 2nd Sem**	Program B **End 2nd Sem**
Dates	July, 2004	July, 2009	Dec., 2004	Jan., 2010
Number of subjects	57	8	57	25
Mean scaled score	60	42	70	35
StD	16.997	23.250	19.722	23.612
K-R21	0.806	0.926	0.833	0.940
Standard Error of Measurement	2.251	8.220	2.612	4.722
Unpaired t test two-tailed P value	0.0094		< 0.0001	
"Significantly" different?	"Very"[4]		"Extremely"[4]	
Median	62	40	66	35
High	95	70	128	85
Low	20	0	20	0

Comment on Figure 2: This figure represents what has happened in listening comprehension in the two programs by the end of the first semester (the first two columns) and the second semester (the second two columns). The mean scores show an increasing gap between programs A and B. Program B actually declines during the second semester. It should be noted that the number of testees in the B group triples, due to the availability of more subjects at this point. Reliability is firm in all the data presented, never falling below KR21 0.777. KR21 is reputed to be a conservative estimate of reliability (Brown, 1996). The t test "significant difference" comments "very"/"extremely", etc., are provided gratuitously by the *Graphpad* (cf. *References*) statistical software. Note that "extremely significant difference" is applied to a P value of < 0.0001 in the present case and that this figure defines the N73 relationship between Programs A and B from the end of the second semester onwards. The "low" score of "0" indicates that at least one student was unable to adequately repeat more than 33 words of the 15 sentences/141 words given on the recording. In Program B, this continued to be the case for at least one student on through to the end.

Third and Fourth Semesters

Figure 3

N73 Channel Capacity (Listening Comprehension)	Program A	Program B	Program A	Program B
	End 3rd Sem	**End 3rd Sem**	**End 4th Sem**	**End 4th Sem**
Date	July, 2003	July, 2009	Dec., 2004	Jan., 2010
Number of subjects	57	12	57	17
Mean scaled score	96	56	112	55
StD	27.453	33.738	28.945[6]	41.514[6]
K-R21	0.886	0.946	0.882	0.971
Standard Error of Measurement	3.636	8.959	3.834	10.069
Unpaired t test two-tailed P value	< 0.0001		< 0.0001	
"Significantly" different?	"Extremely"[4]		"Extremely"[4]	
Median	95	57	107	50
High	152	114	193	128
Low	54	0	66	0

Comment on Figure 3: This figure brings us to the second year and the final two semesters. Both programs show impressive differences between the second and third semesters. The difference between third and fourth semesters is less impressive, however, with Program B actually declining once again. Keep in mind that the subjects are not all the same in Program B's figures from semester to semester.

What do the scores mean?

Ferguson in his *Listening Comprehension Test N73* (Ferguson, 1982) provides simple guidelines which are elaborated elsewhere (Pendergast, 2009), but the following is adequate for our needs at the basic levels of the subjects under discussion:

0 Practically zero

100 Asks and answers question on daily personal needs and familiar topics with very limited vocabulary. Makes frequent basic errors in structure and pronunciation.

200 Converses intelligibly within most social situations but without complete control of structure and pronunciation. Restricted vocabulary.

Elsewhere, levels are characterized in a different way:

50 Tourist survival (able to "get around," but not converse)

150 Social survival (able to "converse" with "caretaker" assistance)

250 Social autonomy (able to converse fairly freely with native speakers)

End of the First and Fourth Semesters

Spoken Expression Data: Fluency, Correctness, and Expression

Use of the Same Cartoon Elicitation: The Family (TF)

Practical aspects

Subjects (Ss) were given a 16-frame cartoon with a typed opening sentence on the back for reference and to establish the tense of the narration, in this case the historical present. For example, "This is a day in the life of the Smith Family. Every day, Billy and Mary Smith get up at 7.30 …"

The frames of the cartoon continue the story. Ss are given 2.5 minutes to look over the story and consider what they will say. With the cartoon in hand, they record what they see for 90 seconds. The recording is analyzed for fluency, correctness, and expression by a trained evaluator using a dedicated, hand-held computer called OLAF (for Oral Language Analysis and Feedback system). The digital readout gives scores for fluency (tone groups per minute), correctness (on a scale of 100%), and expression (i.e., how much and how clearly something was communicated, on a scale of 1000).

Three cartoon sequences were used in the study, the same one – for comparison – after the first and fourth semesters, and different ones after the second and third semesters.

Fluency: End of first and fourth semesters

Figure 4: The Family

N73 OLAF (TF)	Program	Program	Program	Program
Spoken Expression	A	B	A	B
Fluency	**End 1st Sem**	**End 1st Sem**	**End 4th Sem**	**End 4th Sem**
Date	July, 2003	July, 2009	Dec., 2004	Jan., 2010
Number of subjects	57	7	57	17
Maximum possible score	N.A.[5]	N.A.[5]	N.A.[5]	N.A.[5]
Test time	90 secs.	90 secs.	90 secs.	90 secs.
Mean tone groups/minute	8	6	11	9
StD	2.686	1.291	2.584	3.350
K-R21	N.A.[5]	N.A.[5]	N.A.[5]	N.A.[5]
Standard Error of Measurement	0.3558	0.4880	0.3423	0.8125
Unpaired t test two-tailed P value	0.0579		0.0110	
"Significantly" different?	"Not quite"[4]		"Yes"[4]	
Median	8	6	11	10
High	13	8	16	15
Low	3	4	6	3

Comment on Figure 4: This figure shows results of fluency after both the initial and final testing on the same sequence of pictures, for comparison. There is no maximum possible score, as even native speakers vary in their rate of speech. For reference, however, on this particular test a native speaker would have no difficulty in speaking at the rate of 18-20 tone groups per minute. A tone group, in general, corresponds to a clause with a main verb (Ferguson, 1998). There are variations, however, and training to administer this test with OLAF requires roughly a day of induction and additional practice. A "unit of information" provides a general idea of what a tone group is.

As can be seen from the figure, the differences in fluency between the programs at both the end of the first and fourth semesters are "significant" but not off the chart. This is not true of Correctness, however, as seen in Figure 5.

Correctness: End of the first and fourth semesters

Figure 5: The Family

N73 OLAF (TF)	Program	Program	Program	Program
Spoken Expression	A	B	A	B
Correctness	**End 1st Sem**	**End 1st Sem**	**End 4th Sem**	**End 4th Sem**
Date	July, 2003	July, 2009	Dec., 2004	Jan., 2010
Number of subjects	57	7	57	17
Maximum possible score (%)	100	100	100	100
Test time	90 secs.	90 secs.	90 secs.	90 secs.
Mean correctness (%)	42	32	58	31
Balance of mean Flu and Cor	Low Cor	Very Low Cor	Balanced	Very Low Cor
StD	13.045	14.253	15.603	11.875
K-R21	0.866	0.902	0.909	0.857
Standard Error of Measurement	1.728	2.932	2.067	2.880
Unpaired t test two-tailed P value	0.0626		< 0.0001	
"Significantly" different?	"Not quite"[4]		"Extremely"[4]	
Median	41	27	60	32
High	63	55	86	52
Low	17	20	14	11

Comment on Figure 5: "Correctness" is a concept peculiar to OLAF testing and is explained fully in Ferguson, 1998. Simply stated, each tone group is rated at one of four levels of correctness, designated as S(yntax) 1, 2, 3, c. Each tone group is divided into natural stress groups. If a stress group is internally incorrect, the tone group is rated as S1, with "S" standing for "Syntax." If two neighboring stress groups do not harmonize, the

tone group is S2. If a tone group does not harmonize with the preceding tone group, it is S3. A conventional (totally correct) tone group is Sc (for conventional). The words "correct" and "harmonize" in this context mean that a following word, stress group, or tone group is or could be acceptable.

Examples: **I go / to church / on Sunday //Sc**

Last / week // I go / to church / on Sunday //S3. ("Last / week // as a tone group is acceptable, as is the following tone group by itself. // I go / to church / on Sunday // is, however, a *non sequitur*, given the previous tone group // Last / week //.

I go / in church / on Sunday //S2 ("in church" is by itself an acceptable stress group … which would not normally harmonize with the preceding stress group.

I goes / to church / on Sunday //S1 (/ I goes / is an unacceptable stress group.)

The evaluator presses one of four (S1-Sc) buttons on the computer while listening to the sample. Each button press registers as a tone group and counts for Fluency. The level of Correctness depends on which of the four buttons is pressed. A chip in the computer simultaneously calculates "Expression" to a maximum of 1000 (native speaker-equivalent for the sample) to indicate the amount of conventionally comprehensible production.

Figure 5 shows Correctness scores for Programs A and B at the end of the first and fourth semesters, respectively. The 16-frame cartoon elicitation was the same in each case, with a year and a half between the two administrations. As these are proficiency tests, there is naturally no prepping and the cartoons are kept secure.

There was virtually no "grammar" taught in Program A. The learning material (mostly SAPL) was, however, designed to alert Ss to mistakes and give them opportunity to self-correct. Program B was a standard, eclectic curriculum containing a normal amount of grammar translation and explanation in the native language.

Correctness of approximately 30% or lower is practically incomprehensible and understanding is heavily dependent on the intuition of a native-equivalent interlocutor.

Note that the difference between the two programs goes from "not quite significant" after one semester to "extremely significant" after four.

For comfortable communication, there is a question of "balance." Someone who speaks very quickly but with deficient pronunciation or correctness will not be understood. There is a calculation (and a template) which indicates the extent to which a sample of speech has Balance (B), Low Correctness (LC), Very Low Correctness (VLC), or Very, Very Low Correctness (VVLC). The latter is normally incomprehensible. The same is true of low Fluency, where we have B, LF, VLF, and VVLF. If you think of a Fluency of six tone groups per minute, for example, you realize that the listener audits only one unit of information (tone group) every ten seconds. Inevitably, the mind wanders and communication is

diminished or extinguished.

Expression: End of the first and fourth semesters

Figure 6: The Family

N73 OLAF (TF)	Program	Program	Program	Program
Spoken Expression	A	B	A	B
Expression	**End 1st Sem**	**End 1st Sem**	**End 4th Sem**	**End 4th Sem**
Date	July, 2003	July, 2009	Dec., 2004	Jan., 2010
Number of subjects	57	7	57	17
Maximum possible score	1000	1000	1000	1000
Test time	90 secs.	90 secs.	90 secs.	90 secs.
Mean expression	92	61	181	98
Balance of mean List and Exp	60:92	42:61	112:181	55:98
StD	42.593[6]	16.004[6]	74.696[6]	44.477[6]
K-R21	0.955	0.777	0.974	0.956
Standard Error of Measurement	5.642	6.049	9.894	10.787
Unpaired t test two-tailed P value	0.0014		< 0.0001	
"Significantly" different?	"Very"[4]		"Extremely"[4]	
Median	80	55	166	104
High	254	93	400	187
Low	35	48	48	28

Comment on Figure 6: Both programs show considerably higher levels at the end of four semesters than at the end of the first semester. Although both programs began at approximately the same level, Program A is 50% higher after one semester and 100% higher three semesters later. Keep in mind, however, that the maximum possible Expression score is 1000 for a native speaker-equivalent. It may appear that even a computerized test of only 90 seconds could not be rigorous enough to identify a native speaker, but the fact is that the native speaker almost invariably shoots the computer into a display of "HI" (for "High", or off the chart) within a few seconds, due to explosive initial fluency and correctness.

End of the second and third semesters

Spoken expression data: Fluency, Correctness, and Expression

Two cartoon sequences: The Busy Day and The Dog's Story (BD and DS)

Fluency

Figure 7	The Busy Day		The Dog's Story	
N73 OLAF (BD and DS)	Program	Program	Program	Program
Spoken Expression	A	B	A	B
Fluency	**End 2nd Sem**	**End 2nd Sem**	**End 3rd Sem**	**End 3rd Sem**
Date	Dec., 2003	Jan., 2010	July, 2003	July, 2009
Number of subjects	57	16	56	12
Maximum possible score	N.A.[5]	N.A.[5]	N.A.[5]	N.A.[5]
Test time	90 secs.	90 secs.	90 secs.	90 secs.
Mean tone groups/minute	12	9	10	8
StD	2.768	3.488	3.483	3.099
K-R21	N.A.[5]	N.A.[5]	N.A.[5]	N.A.[5]
Standard Error of Measurement	0.3666	0.8720	0.4654	0.8946
Unpaired t test two-tailed P value	0.0006		0.0707	
"Significantly" different?	"Extremely"[4]		"Not quite"[4]	
Median	11	8	10	8
High	22	16	16	14
Low	6	2	5	3

Comment on Figure 7: BD is a 16-frame cartoon sequence showing a day in the life of a man who gets up and goes to work, but doesn't do much all day. DS is a sequence showing a man taking his dog into town and leaving him to wait in the car for a while. The dog in fact jumps out of the car and has a series of exciting experiences. BD is the most fluid and uncomplicated of the three cartoon sequences, as can be seen from the higher Fluency scores. DS is the most complicated of the three. Note ("High") that it is occasionally possible for Ss to "take off" on a sequence and work up an unusually high Fluency score. On the other hand ("Low"), some Ss find themselves stunned into almost complete silence (note the lows above, indicating only two to three utterances in 90 seconds) and become totally tongue-tied.

Correctness

Figure 8	**The Busy Day**		**The Dog's Story**	
N73 OLAF (BD and DS)	Program	Program	Program	Program
Spoken Expression	A	B	A	B
Correctness	**End 2nd Sem**	**End 2nd Sem**	**End 3rd Sem**	**End 3rd Sem**
Date	Dec., 2003	Jan., 2010	July, 2003	July, 2010
Number of subjects	57	16	56	12
Maximum possible score (%)	100	100	100	100
Test time	90 secs.	90 secs.	90 secs.	90 secs.
Mean correctness (%)	58	26	48	32
Balance of mean Flu and Cor	Low Cor	Very Low Cor	Low Cor	Very Low Cor
StD	14.995	12.293	13.675	11.324
K-R21	0.901	0.881	0.875	0.839
Standard Error of Measurement	1.986	3.073	1.827	3.269
Unpaired t test two-tailed P value	< 0.0001		0.0003	
"Significantly" different?	"Extremely" [4]		"Extremely" [4]	
Median	60	23	48	33
High	84	55	86	54
Low	16	3	18	17

Comment on Figure 8: The notable aspect of these data is the "Mean correctness" figures. Keep in mind that the figure noted is a percentage (out of 100%). Even native speakers in everyday speech rate at between 90-95% due to hesitation and changes in direction. Speech where Correctness is under 30% relies greatly on the interlocutor's intuition and guesswork for comprehension. Program B's subjects are in this range *even though/because* (take your choice) they have spent about eight years working largely on an intellectual understanding of "grammar."

Program A subjects on the other hand had little exposure to grammar explanations, yet seem to have intuitively internalized significantly more "grammar" than the control group (Program B). This may be attributed to the self-correction devices built into the material used, the intervention of the partner in pair learning, and the occasional intervention of the coordinator.

Consider the case often encountered in "returnees" who have been abroad for some time. Typically, the subject speaks with relatively high fluency, but correctness under 30%! This is likely in subjects of high sociability who have lived abroad and never been corrected … for the simple reason that correction is not the thing to do in most social situations. Consequently, the subject becomes overly confident, resists correction if offered, and never

attains a high level of proficiency.

The "Balance of Flu and Cor" is taken from a template which purports to evaluate the amount of communication successfully achieved in a situation where the interlocutor is a naïf. That is to say, those native speakers of English who have lived in Japan for some time come to understand an English which the average naïve native speaker would not. OLAF evaluation attempts to reflect the objective reality.

It should be noted that the main difference in Correctness in the case of the BD elicitation was due to a consistent lack of subject-verb agreement. Program A subjects by and large accurately overcame this pitfall. Program B subjects, in several cases, ignored the convention TOTALLY. This resulted in a decisive difference.

Expression

Figure 9	**The Busy Day**		**The Dog's Story**	
N73 OLAF (BD and DS)	Program	Program	Program	Program
Spoken Expression	A	B	A	B
Expression	**End 2nd Sem**	**End 2nd Sem**	**End 3rd Sem**	**End 3rd Sem**
Date	Dec., 2003	Jan., 2010	July, 2003	July, 2010
Number of subjects	57	16	56	12
Maximum possible score	1000	1000	1000	1000
Test time	90 secs.	90 secs.	90 secs.	90 secs.
Mean expression	196	87	127	83
Balance of mean List and Exp	70:196	38:87	96:127	64:83
StD	90.156[6]	53.183[6]	49.256	42.627
K-R21	0.982	0.973	0.955	0.959
Standard Error of Measurement	11.941	13.296	6.582	12.305
Unpaired t test two-tailed P value	< 0.0001		0.0055	
"Significantly" different?	"Extremely" [4]		"Very" [4]	
Median	170	67	122	75
High	520	230	259	159
Low	69	21	48	22

Comment on Figure 9: Data are missing for one subject in Program A at the end of the 3rd semester. Hence, N = 56. This is true of all the speaking data for this semester (i.e., Flu, Cor, Exp).

Of the three series of pictures used, (as noted in the comment on Fig. 7) DS is the most complex, followed by TF. BD is the easiest, with the exception mentioned above that a lack of sensitivity to subject-verb agreement will affect the Correctness score greatly and

thereby lower the overall Expression score considerably. At lower levels, changes in Fluency influence Expression most. At higher Fluency levels, even small changes in Correctness become more important.

Conclusion

The available data (a relatively large number of subjects for Program A, far smaller numbers for Program B) show a significant proficiency difference between the two programs at the end of each semester on all scales, ranging from "significantly different" to "extremely significantly" (< 0.0001) different, with the majority of evaluations of the latter type.

Granted that the number of subjects in Program B was quite small, the benchmark test suggests that the groups were not significantly different at the start of their programs. What then was the difference in their experience of English which led to the invariable difference in their results at each step of the way?

The integrated Program A spent most of its time in student-centered, self-access activities, using all four skills. Activity (use of the language) was very high, affectivity/involvement was high due to the individual activity, and anxiety was low, as student pairs were screened from visual and aural contact with other students and often even from the instructor/coordinator, who circulated among the groups giving advice. While students studied in pairs, Baroque music played in the background.

Program B students had a variety of teachers, materials, and class formats. It must be admitted that many students liked this variety, but the results do not seem to validate the efficiency of this non-integrated, instructor-centered approach.

Notes

01. CEEL: the Center for the Experimentation and Evaluation of Language Learning Techniques, Geneva.
02. The core of the integrated curriculum in Program A was based on a self-access learning program called SAPL, or Self-Access Pair Learning (six classes/week) and one class/week of StoryTelling/Extensive Reading.
03. The test of channel capacity (Johnson-Ferguson, 2009) requires students to listen to a small number of sentences and to repeat back what they hear. The first sentences are short but become progressively longer. The more words they can repeat, the better their score. The test used is the N73 (Ferguson, 1973).
04. Verbal descriptions of P value probability are taken from *Graphpad Software* (see References).
05. There is no practical maximum value (score) for Fluency, as this can vary from native speaker to native speaker. Therefore the KR21 calculation for reliability is inappropriate in this case.
06. In cases where Note #6 is referenced, following the *Graphpad Software* advice, the comment is:

The t test assumes that the (columns of unpaired scores) come from populations with equal SDs. A calculation is made and in Note #6 cases it has been determined that the difference is significant. Consequently, a calculation applying the "Welch correction", which does not assume equal variances, is made and those t test results reported.

References

Atherton, J.S. (2005). *Teaching and Learning: What works and what doesn't* [On-line] UK: Available: http://www.learningandteaching.info/teacing/what works.htm

Brown, J.D. (1996). *Testing in Language Programs*, pp. 197-199. Upper Saddle River, NJ: Prentice-Hall Regents.

Cleveland, H., Mangone, G., & Adams, J.C. (1960). *The Overseas Americans: A Report on Americans Abroad*. New York: McGraw-Hill.

Ferguson, N. (1973). *Listening Comprehension Test N73*. Geneva: CEEL.

Ferguson, N. (1980). *The Gordian Knot.* Geneva: CEEL.

Ferguson, N. (1998). *OLAF N73*: Objective evaluation of the ability to speak a foreign language. Geneva: CEEL.

Ferguson, N. (1999). *Language Teaching Theory: A Handbook for Professionals*. Geneva:CEEL.

GraphPad Software, San Diego California USA, www.graphpad.com".

Johnson-Ferguson, N. (2009). *My teacher is rich: How to get a job anywhere in the world with absolutely no qualifications*. Geneva: ECLT-CEEL

Pendergast, T.M. (1985). "OLAF N73: A Computerized Oral Analyser and Feedback System". In *New Directions in Language Testing*. Lee, Y.P., Fok A., Lord R., & Low, G. (Eds.). Oxford: Pergamon Press.

Pendergast, Thomas M. (2009). "English Language Proficiency Gains in an Integrated, Self-Access Program (Class of 2005): Part 1." Osaka: The Bulletin of Shitennoji University: Vol. 48, 227-244.

Scheibner-Herzig, G., Sauerbrey, H., & Kokoschka, S. (1991). "Repetition − A Means to Predict Foreign Language Oral Proficiency". IRAL XXIX/3, August, pp. 230-239.

平泉渉・渡部昇一（１９７５）『英語教育大論争』東京：文藝春秋　238pp.