〈経営研究　第4巻　第1号　平成2年9月〉

AN INVESTIGATION OF THE EFFECTS ON LEARNING OF "TRACKING"

J. Morita, H. Iida, and H. Slovic

KEYWORDS:

correlation coefficient, class organization, divided ability groups, F—test, mixed ability groups, Spearman's rank order correlation coefficient, streaming, t—test, tracking, track system

I. INTRODUCTION

The following paper summarizes the work done to date on an on—going research project whose purpose is to investigate the effects on learning of "tracking" (also referred to as "streaming"), i. e., the deliberate separation of and assignment to learning groups (in this case class sections of an English conversation course) of students according to their differing levels of English language mastery and competence.

The notion that the learning process develops in stages or sequences is expressed in the common proverb, "One must learn to walk before one can run". Educator Jerome S. Bruner expands on this idea as follows:

> Instruction consists of leading the learner through a sequence of statements and restatements of a problem or body of knowledge that increases the learner's ability to grasp, transform, and transfer what he is learning. In short, the sequence in which a learner encounters materials within a domain of knowledge affects the difficulty he will have in achieving mastery.
>
> There are usually various sequences that are equivalent in their ease and difficulty for learners. There is no unique sequence for all learners, and the optimum in any particular case will depend upon a variety of factors, including past learning, stage of development, nature of the material, and individual differences.[1)]

Given the sequential nature of the learning process, students are routinely and traditionally grouped according to their chronological age. While this has definite advantages, strictly from considerations of management of the learning situation, the underlying assumption that students of the same age have all attained the same degree of mastery of a given subject is often a mistaken one which denies the reality of the range of differences among them. Thus, division of students according to criteria which recognize their place in the learning sequence seems, intuitively, to make more sense. Hughes–d'Aeth refers to two such ways of grouping students as "mixed ability groups and divided ability groups"[2)]. He goes on to list the advantages and disadvantages of "mixed ability groups" as follows:

1. Children are not labelled: no pupil feels superior or inferior.
2. There is an improved class atmosphere because of the first

advantage.

3. Discipline problems are fewer since there are no areas where problem students can get together.
4. Pupils learn to work co–operatively.
5. There are more opportunities for teacher–pupil contacts.
6. There are more and more meaningful pupil–pupil contacts.
7. Late developers are given improved chances.
8. Pupils may, more readily, work at their own level.
9. A levelling up of attainment occurs (slower pupils improve their performance).
10. There is improved language development.
11. Brighter pupils can help less able ones.
12. There is more time given to individual pupils.
13. There is more time available before pupils' abilities need to be assessed.
14. All pupils appear more confident.
15. There is less stress or emotional tension than in a streamed situation.

There are also problems of mixed ability teaching:

1. Appropriate teaching materials are sometimes difficult to think of.
2. It is hard to teach a whole–class lesson at the correct level.
3. It is difficult to keep track of all pupils' progress.
4. Teachers need to be committed to the idea of mixed ability teaching.
5. Teachers need to spend a lot of time in preparation and re-

source—making.

6. Bright pupils may waste a lot of time.
7. Teachers spend time disproportionately on the slow learners.
8. Slow learners may feel they always fail.

Pupils who are either very good or very weak have the most need of attention in a mixed ability group. Pupils of low ability often do better than expected in a mixed ability group. They are able to take part in work which would be too difficult for them to do alone and their self—confidence increases. However, a teacher must not make too great a demand, otherwise, the pupil will lose confidence. Pupils who are very good at their subject may also have problems. The material and work may be too easy and they may not be stretched enough to do their best. These pupils become bored and frustrated when time is taken up by work which is obvious to them, or as they wait their turn for the teacher.[3)]

As Hughes—d'Aeth's comments are not aimed at "mixed ability classes", but at "mixed ability groups" within a single class of students, some of his comments appear to be irrelevent to the present situation of the classes under investigation. Nevertheless, others, such as 1, 4 and 11 in the "advantages" list, and 2, 6, 7 and 8 in the "disadvantages" list appear, to these authors, to have considerable relevance, even when generalized to the case of entire classes. Regarding "divided ability groups", Hughes—d'Aeth writes:

> On the other hand, a teacher may decide to stream the pupils so that all the good pupils are together, the intermediate ability groups are ranged in other groups and the poor pupils of low ability are concentrated in a separate group. The teacher can only decide which pupils to place in the respective groups after comprehensive information has been collected about those pupils. A divided ability group allows the teacher to set different work to the various groups, according to each group's ability, at the same time. In this way, bright pupils can be set challenging and demanding work without being held up by the rest of the class whilst the teacher is able to concentrate on remedial language work with pupils who are struggling with English.[4)]

While "tracking" is widely employed in Japan in English conversation classes held at language schools and companies (one major corporation had, at one time, twenty–three "tracks" based on students' differing levels of language ability), the system is not generally employed at Japanese universities. The fact that students may be granted entrance into university via a number of different channels, some of which do not require them to demonstrate their level of English language ability, coupled with the lack of a tracking system, results in English conversation classes composed of students with widely ranging abilities and prior accomplishments. The present investigation grew out of comments from English conversation instructors at this university, who expressed their feelings that such "mixed ability classes" were both harder to teach from the teacher's standpoint, and also, less effective learning environments from the students' standpoint, than were classes to which students had

been assigned on the basis of some prior screening method designed to insure a narrower range of "competence", "ability", or "achievement", viz., the so-called "tracking" system.

After a number of somewhat emotional discussions in which faculty members expressed their opposing views of either the supposed benefits or the possible detrimental or negative effects of the tracking system, it was decided to do an experiment to demonstrate, if possible, that the use of tracking in the assigning of students to class sections of the 1st-year English Conversation course would result in greater learning.

II. METHODOLOGY

The methodology of the experiment mentioned above is briefly summarized below:

Step 1: All in-coming 1st-year students in the 1988-1989 academic year were tested during the first days of the new school year using a test designed to measure their general reading comprehension ability and overall knowledge of English sentence structure, and pronunciation.

Step 2: Students who scored in the upper one-third of the ranked scores on this test were to be randomly assigned to one of six different class sections of the English conversation course. All other students, i.e., those whose scores on the general reading test fell in the lower two-thirds of the ranked scores, were to be randomly assigned to one of twelve other class sections. These two groups constituted a "higher level track" and a "lower level track", with the only distinguishing feature of the two tracks assumed to be a real difference

in the students' "competence, ability, or achievement". Students were not told that their class assignment was based on this screening process.

Step 3: All students were then given a "listening comprehension ability test" prior to their actually beginning formal study in their English conversation classes (HAC pre–test), and were given a comparable test at the end of the academic year (HAC post–test). The scores on these two tests were then submitted to both descriptive and inferential statistical analyses.

III. STATISTICAL ANALYSES

To date, several statistical tests have been done on the HAC pre– and post–test scores, viz., for 1st–year students in the 1988–89 academic year, HAC pre– and post–test scores were first paired to determine the degree of correlation between them, rank–ordered and again tested for correlation, submitted to Student's t–test, and to the ANOVA F–test. Finally, a comparative inspection was made between these scores and those obtained from the previous academic year's 1st–year students. These various analyses are discussed below.

A. Analyses performed on 1988–89 academic year students' scores

1) Correlation coefficient—the two sets of scores were compared by computing the correlation coefficent according to the following formula:

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}}$$

(See Appendix A)

The correlation cofficient in this case is an indication of the degree of the ability to predict a student's HAC post–test score given his or her score on the HAC pre–test. The computed value of r = .548 in the given case indicates a fair degree of predictabilty.

2) Spearman's rank order correlation coefficient—to compute this statistic, the HAC pre– and post–test raw scores were first converted to rank order scores, and the correlation coefficient computed by means of the following formula:

$$r_s = 1.0 - \frac{6\Sigma D_i^2}{n(n^2-1)}$$

(See Appendix B)

This statistic is another indicator of the degree of relatedness between the two sets of scores. The computed value of r = .513 in the present case likewise indicates a fair degree of relatedness between the two sets of test scores. Had either of these correlations indicated a low degree of relatedness between the two sets of HAC test scores, further statistical analysis would most likely not have been justified.

3) Student's t–test—this test was performed on a class–by–class basis using the following formula to determine t–values for each set of data composed of the algebraic differences between each student's HAC post– and pre–test score.

$$t = \frac{\overline{D}}{S_D/\sqrt{n}}$$

(See Appendix C)

In this test, the null hypothesis was H_0: $\mu_1 = \mu_2$, namely, that the mean of the population represented by HAC pre–test scores for a given class and the mean of the population represented by HAC post–test scores for the same class were the same, implying that the two populations were, in fact, the same population. The experimental hypothesis was H_0: $\mu_1 \neq \mu_2$, namely, that the means of the two populations were significantly different, indicating that an improvement in listening comprehension ability had occurred. T–values greater than t–critical allow one to reject the null hypothesis (see TABLE 1 below)

CLASS#	n	df = n−1	t–score	› OR ‹	t–critical	n. j. p.
1	13	12	5.23868	›	2.179	+
2	9	8	4.11908	›	2.306	+
3	15	14	4.53743	›	2.145	+
4	16	15	6.22776	›	2.131	+
5	18	17	1.89100	‹	2.110	n. j. p.
6	16	15	4.44009	›	2.131	+
7	13	12	5.06989	›	2.179	+
8	17	17	1.21157	‹	2.120	n. j. p.
9	14	14	5.33772	›	2.160	+
10	18	18	5.36948	›	2.110	+
11	19	19	1.54364	‹	2.101	n. j. p.
12	17	17	6.12598	›	2.120	+
13	12	12	4.06423	›	2.201	+
14	14	14	1.77478	‹	2.160	n. j. p.
15	17	16	1.65336	‹	2.120	n. j. p.
16	16	15	2.17568	›	2.131	+
17	13	12	5.67974	›	2.179	+
18	16	15	3.08339	›	2.131	+

Fast track classes indicated by underline
+indicates "significant difference"
n. j. p. indicates "no judgement possible"
(alpha = .05)

TABLE 1: t–VALUES FOR DIFFERENCE BETWEEN HAC PRE–AND POST–TESTS

As can be seen from TABLE 1, t–scores greater than t–critical were obtained in 13 out of the 18 data samples; in the remaining 5 cases, the t–scores were less than t–critical, hence, it was not possible to state for these 5 cases that, at the given confidence level (alpha = .05), the observed differences in the HAC pre– and post–test scores were not simply due to sampling variation. In the former 13 cases, if one assumes that the HAC test is a reliable measure of student's listening comprehension ability, then it is possible to assert that the statistically "significant" differences are indicative of students' real improvement in this area for these 13 classes. But, it should be pointed out that the t–test does not distinguish between the two tracks, and does not provide any basis for making comparisons between "higher" and "lower" tracks; therefore, one cannot conclude that the improvement which occurred in 72% of the classes was due to the tracking system. This improvement might just as well have occurred had the tracking system not been employed. Finally, it should be noted that only 2 out of the 6 "higher level track" classes were among the 13 "improved" classes; the remaining 4 "higher level track" classes had t–scores lower than t–critical.

4) ANOVA *F*–test—this was done to determine if, as a result of the original screening of students, there actually existed significant differences in the composition of class groups, or if, on the contrary, despite the screening of the students, the populations represented by the two tracks were essentially one and the same. This *F*–test used the raw test scores and involved a comparison of the variance in scores existing between the class groups (S_B^2) with the variance exisiting between scores within each of the class groups (S_W^2); the

test was performed on HAC pre– and post–test scores separately, using the following formula:

$$F = \frac{S_B^2}{S_W^2}$$

(See Appendix D)

The null hypothesis for this test was $H_o : \mu_1 = \mu_2 = \cdots = \mu_k$, namely, that the means of the populations, of which the various sets of data from the 18 class sections (including HAC scores from both "higher level" and "lower level track" students) are assumed to be representative, were the same, and thus, could be considered to be one and the same population, i. e. that there was no significant difference in the English listening comprehension ability of students in either of the two tracks. The experimental hypothesis in this test was $H_1 : \mu_i \neq \mu_j$, namely, that the population means were not the same, and that students in one or more of the class sections possessed a significantly different degree of listening comprehension ability compared to their peers in other class sections.

In the present case, F–values of 20.726 and 4.250 were found for data sets of HAC pre– and HAC post–tests. Both of these values are statistically significant, allowing rejection of the null hypothesis, and allowing the assertion that students in at least one and possibly more of the class sections possessed a significantly different level of listening comprehension ability. It should be noted, however, that this is all that this test allows one to assert; it does not furnish any information with which to "pinpoint" which of the groups was/were different, nor to what degree. This information must be obtained by other statistical tests, which, unfortunately,

were not performed.

B. Analysis comparing HAC test–scores for 1987–88 academic year students with those of 1988–89 academic year students

All in–coming 1st–year students in the 1987–88 academic year were given the same HAC pre– and post–tests under similar circumstances as were those in the 1988–89 academic year. The former students, however, were not assigned to "tracks". While it might be thought that they, therefore, could be considered to be a "control group", there is no basis for assuming that, considered as a whole, they represent the same population as the students who came after them. Thus, while no statistical tests were performed directly comparing the two groups, the observance of a somewhat curious difference between their respective scores led to "in–group" testing for statistical significance.

It was observed that for the earlier group, test–score standard deviations for almost all classes decreased between pre–and post–tests, whereas for the later group it increased in almost every case. These decreases and increases in the standard deviations were tested for statistical significance using an F–test according to the following formula:

$$F = \frac{S_1^2}{S_2^2}$$

(See Appendix E)

The null hypothesis here was that $H_o : \sigma_1^2 = \sigma_2^2$, namely, that

the variances of the populations pairs assumed to be represented by the pairs of standard deviations for each class are equal; the experimental hypothesis was $H_1 : \sigma_1^2 \neq \sigma_2^2$. Rejection of the null hypothesis by F—ratios greater than F—critical allows one to assert that the differences between the standard deviations are due to actual population differences, and reflect a real change in the students' pre— and post—test performances. (See TABLE 2 below).

CLASS#	1	2	3	4	5	6	7	8	
1987–88									
df=	16	14	15	14	13	15	15	14	
HAC Pre–test Variance	156.154	80.517	372.029	119.429	255.412	194.196	89.933	208.133	
HAC Post–tost Variance	124.515	177.029	122.729	46.124	105.912	85.896	104.463	95.729	
Decrease in Variance	YES	NO	YES	YES	YES	YES	NO	YES	
F–ratio	1.254	2.199	3.031	2.589	2.412	2.261	1.162	2.174	
F–critical	2.33	2.48	2.39<Tc<2.43	2.48	2.55<Tc<2.60	2.39<Tc<2.43	2.39<Tc<2.43	2.48	
Reject H–null [.05]	NO	NO	YES	YES	NO	NO	NO	NO	
Significat decrease			X	X					
CLASS#	9	10	11	12	13	14	15	16	
1987–88									
df=	17	14	14	16	15	15	15		
HAC Pre–test Variance	122.408	327.238	366.314	189.441	153.600	164.117	93.362		
HAC Post–tost Variance	87.987	108.314	55.267	76.596	129.863	58.800	34.663		
Decrease in Variance	YES	YES	YES	YES	YES	YES	YES		
F–ratio	1.391	3.021	6.628	2.473	1.183	2.791	2.693		
F–critical	23<Tc<2.29	2.48	2.48	2.33	2.39<Tc<2.43	2.39<Tc<2.43	2.39<Tc<2.43		
Reject H–null [.05]	NO	YES	YES	YES	NO	YES	YES		
Significat decrease		X	X	X		X	X		X=7
CLASS#	1	2	3	4	5	6	7	8	9
1988–89									
df=	12	8	15	15	12	13	17	16	11
HAC Pre–test Variance	105.410	72.000	43.717	49.896	66.192	38.533	110.761	103.654	98.818
HAC Prst–test Variance	123.910	61.694	49.667	58.629	100.910	95.258	122.941	127.993	106.447
Increase in Variance	YES	NO	YES	YES	YES	YES	YES	YES	YES
F–ratio	1.176	1.167	1.136	1.175	1.525	2.472	1.110	1.235	1.077
F–critical	2.69	3.44	2.39<Tc<2.43	2.39<Tc<2.49	2.69	2.55<Tc<2.60	2.23<Tc<2.29	2.33	2.82
Reject H–null [.05]	NO	NO	NO	NO	NO	NO	NO	NO	NO
Significat increase									
CLASS#	10	11	12	13	14	15	16	17	18
1988–89									
df=	13	15	15	14	17	16	18	16	12
HAC Pre–test Variance	104.533	58.896	47.563	22.810	21.059	17.559	25.023	31.618	18.308
HAC Post–tost Variance	259.506	148.096	122.517	48.381	64.134	61.971	67.228	136.890	29.064
Increase in Variance	YES	YES	YES	YES	YES	YES	YES	YES	YES
F–ratio	2.483	2.515	2.575	2.121	3.045	3.529	2.687	4.329	1.588
F–critical	55<Tc<2.60	39<Tc<2.43	2.39<Tc<2.43	2.48	2.23<Tc<2.29	2.33	2.19<Tc<2.25	2.33	2.69
Reject H–null [.05]	NO	YES	YES	NO	YES	YES	YES	YES	NO
Significat increase		X	X		X	X	X	X	
									X=6
Tc=T–critical									
(alpha=05)									

TABLE 2: F–TEST FOR SIGNIFICANT DECREASE IN HAC TEST SCORE VARIANCES

As TABLE 2 indicates, for the 1987–88 students, decreases in standard deviation occurred in 13 out of 15 cases, of which 7 of these cases were determined to be statistically significant. Likewise, for 1988–89 students, standard deviations increased in 17 out of 18 cases, and of these, 6 were found to be significant. For both groups there were a number of statistically significant "near misses" at the 95% level of confidence. Hence, one is tempted to make the assertion that tracking of students has a tendency, for reasons not known, to widen the range of students' ability, wheras "mixed ability groups" tend to narrow the range. If real, this is clearly an interesting phenomenon, and suggests further investigation should be done, both for the sake of verification, as well as to determine, if verified, the underlying causes.

IV. DISCUSSION

Despite the best intentions of all parties concerned with this "experiment", a number of serious criticisms of the methodology and procedure should be mentioned, so that future attempts to measure the effects of tracking may be assured a greater chance of success. These are discussed below.

A. Delays in analyzing data

Unfortunately, for a variety of reasons, a considerable length of time elapsed between the original screening of students, their placement into one of the two tracks, obtaining of HAC pre– and post–test scores and the analysis of these scores, with the result that people responsible for these actions were unable to remember clearly

exactly what they had done, nor their reasons for doing so. Clearly, this is an undesirable situation; thus, if future experiments are to be performed, gathered data should be analyzed as soon as possible.

B. Lack of a control group

In the present experiment, both "higher track" and "lower track" class sections can be considered to be "experimental groups", as the separation into two tracks reduces the degree of randomness in the assignment of students into class sections, thus leaving no control group from which to determine "basal data" for the purpose of comparison. This is a serious flaw in the experimental design.

While it appears, at first, that one solution to this problem would be to use students' HAC pre–and post–test data from the previous (1987–1988) academic year, during which no tracking of students occurred, as a set of "basal data" and to consider the entire group of '87–'88 first–year students as the "control group", it is not possible to assume that these students are representative of the same population of students as the students on which the present experiment was carried out, thus ruling out the possibility of any meaningful comparison. From the viewpoint of experimental design, a better approach would be, perhaps, to separate students into three groups, viz., a mixed ability group to serve as a control group, and two divided ability groups, one consisting of students of the highest ability, and the other of students on the low end of the ability scale. Inferential statistical tests could then be performed which would, hopefully, indicate whether students in the divided ability groups

demonstrated a greater degree of improvement than those in the control group.

C. Student ability dianostic procedure

In connection with the procedure employed to separate the students into the two tracks, it is questionable whether the reading comprehension exam was the most appropriate vehicle, as it is not clear exactly what relationship exists between reading comprehension and listening comprehension ability. In future, a better method of separation might be to use HAC pre–test scores to determine which group/track students should be assigned to. This method would have the advantage that later comparison comparison of students' HAC pre–and post–test scores would reveal only the differences in the students listening comprehension ability, and differences due to varying degrees of reading comprehension ability would not exert any significant influence on the data.

D. Mis–assignment of students into class sections

As the power and responsibility to assign students to various class sections of the English conversation course rested not with the experimenters themselves, but with administrative personnel, who either lacked understanding of the experimental procedure or needed to satisfy requirements other than those consistent with the tracking experiment, the separation of students into "higher track" and "lower track" class sections on the basis of their scores on the reading comprehension ability test did not actually occur according

to the set criterion. (see TABLE 3 below).

CLASS SECTION	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
	LT	LT	HT	LT	HT	LT	LT	HT	LT	LT	HT	LT	LT	LT	HT	LT	HT	LT
RANK	9	24	1	17	3	44	114	14	89	14	17	9	49	24	4	28	2	17
	66	60	6	17	6	46		17	97	28	35		60	28	4	35	23	17
	105	72	11	89	12	72		26		28	39		78	56	6	39	37	49
	105	78	12	93	14			49		28	39		78	89	38	44	39	60
		97	24	105	28			56		60	49		97	93	66	49	49	72
LT = LOWER TRACK			24	105	39			114		97	72		105	97	66	66	56	84
			46		49			130		105	72		114	105	89	105	59	97
HT = HIGHER TRCK			46		60			147			78				93		60	114
			66		72			147			84				105		66	
			78		84			166			84				114		84	
			78		114			166			97				130		122	
			93		122			196			114				147		122	
			130		136			258			114				196		136	
			163		182			272			122				210		166	
			196		221			272			122				215		166	
			227		229			264			122				229		276	
			227		229			290			175				243		276	
			229		269			302			276				310		281	
n = 19			272		313			336			330				318		332	
CUT-OFF RANK = 114																		
PROPORTION																		
MIS-ASSIGNED	4/19	5/19	7/19	6/19	8/19	3/19	1/19	3/19	2/19	7/19	6/19	1/19	7/19	7/19	9/19	7/19	9/19	8/19
PERCENT MIS-ASSIGNED	21	26	37	32	42	16	5	68	11	37	32	5	37	37	47	37	47	42

TABLE 3: PERCENTAGES OF STUDENTS MIS-ASSIGNED TO CLASS SECTIONS

As can be seen from TABLE 3 above, non-negligible percentages of students scoring below the upper one-third of ranked scores cutoff point (114) were assigned to "higher track" class sections, and of students who scored above the cutoff point were assigned to "lower track" class section, thus confounding the data. Unfortunately, this major error in methodology was discovered only after the above mentioned statistical analyses had been done.

V. SUMMARY AND CONCLUSIONS

An experiment was done to determine the effect on learning of grouping students together based on their having similiar levels of English language knowledge and competence ("tracking"). This experiment involved an initial, broad separation of students into two "tracks", based on their performance on

a general exam of reading comprehension ability. All students were then given a test designed to measure their listening comprehension ability prior to their attending English conversation classes over the course of one academic year; at the end of the year, students were retested on the listening comprehension ability test and the test scores on the two listening comprehension ability tests were submitted to various statistical analyses. While Student's t—test revealed a significant improvement in listening comprehension ability in 72% percent of the class sections (13 out of 18), and the *F*—test revealed a significant difference existing between one or more of the 18 different class sections, both at the beginning and the end of the academic year, it was not possible to pinpoint the class section(s) which exhibited this difference, nor to state the magnitude of the difference, nor to ascribe the improvements revealed by the t—test to the effects of tracking. Only after the statistical analyses had been performed was it discovered that the initial assignment of students to either a "higher level track" or a "lower level track" had not been done according to the criterion originally set, and that non—negligible percentages of students had been mis—assigned to the "wrong" track. This constituted a serious methodological error in the experimental design. The lack of a control group, and the questionable use of the reading comprehension ability test as the vehicle for diagnosis and separation of students may be viewed as additional errors of methodology. The interesting phenomenon of scores by students in tracked classes showing a tendency towards increased variance on the successive HAC test, while scores from untracked classes tended towards decreased variance, was noted.

It is a fact that the percentage of students at this university, who, due to low academic achievement, are required to repeat the English conversation course, is undesirably high. While some of these students may possess con-

siderable language learning ability, the greater part of them appear to be students whose general knowledge and skill level is at the lower end of the scale. Thus, by "failing the course", whereby they are assigned to a "repeat class" composed of similar students, such students are, in effect, automatically creating a "lower level track" class section. This appears to be an indication that such students might benefit from being grouped together initially and given special remedial instruction, in order to raise their general knowledge and skill level to that of the rest of the students. This fact, plus English conversation instructors' "gut feeling" that "tracking makes sense", suggests that the attempt to discover the effects of tracking should not be abandoned, and that further research is required.

Finally, it should be pointed out that this was a first attempt, and that, despite the lack of positive, compelling results, it proved to be a valuable learning experience for the experimenters.

APPENDICES

Additional information concerning the various formulae for the statistical tests employed in this project is given in the sections below:

A. Correlation coefficient

A *positive correlation* means that as one variable increases, the other likewise increases. A *negative correlation* means that as one variable increases, the other decreases. Heights and weights of humans are positively correlated, but the age of a car and its trade–in value are negatively correlated. If ρ is equal to zero, we say the

variables are uncorrelated and that there is no linear association between them.[5)]

Note: the use of the Greek letter—symbol ρ (rho) in the quote above refers to the population parameter, whereas r is used for the sample statistic.

B. Spearman's rank order correlation coefficient

In the formula shown on page 140,

$$r_s = 1.0 - \frac{6\Sigma D_i^2}{n(n^2-1)}$$

...D_i is the difference between an item's ranking in one list and its ranking in the other. Tied ranks are handled, as usual, by averaging ranks. The n value is the number of items being ranked.

When the value of n is 10 or more, the following statistic is approximately t—distributed with $n-2$ degrees of freedom:

$$t = \frac{r_s}{\sqrt{(1-r_s^2)/(n-2)}} \qquad (19.\ 14)$$

Note that r_s is a value between -1.0 and 1.0. If the two sets of rankings agree completely, then all the D values are zero, and $r_s = 1.0$. This value, in turn, produces a t value in Equation (19. 14) that is infinitely large—and therefore significant. If the two sets of rankings are in perfectly reversed order, then the sum of the D_i^2 figures divided by $n(n-1)$ and multiplied by 6 always equals 2.0. Thus... $r_s = -1.0$ and again produces a significant t value that is infinitely negative. An r_s value of zero indicates that the rankings seems to

have no positive or negative correlation.[6]

C. Student's t–test

Population Type	Hypothesis to Be Tested	Test Statistic
Two populations with continuous measures X_1 and X_2	$Ho : \mu_1 = \mu_2$	$t = \dfrac{\overline{D}}{S_D/\sqrt{n}}$ where S_D is the standard deviation of the paired differences

Comments and assumptions:

—$D_1 = X_{11} - X_{12}$, the difference of paired observations.

—n is the number of pairs.

—If the sample size is small ($n < 30$), then we must assume the populations are normally distributed.

—The t value has $n - 1$ degrees of freedom.[7]

D. ANOVA F–test for the comparison of sample means

Hypothesis to Be Tested	Test Statistic
$H_o : \mu_1 = \mu_2 = \cdots = \mu_k$	$F = S_B^2/S_W^2$ where $S_B^2 = \dfrac{\sum_{j=0}^{k} n_j (\overline{X} - \overline{\overline{X}})^2}{k-1}$ $S_W^2 = \dfrac{\sum_{j=0}^{k} (n_j - 1) S_j^2}{\sum_{j=0}^{k} n_j - k}$

Comments and assumptions:

—The F value has two degrees of freedom values associated with it: $\nu_1 = k-1$ and

$$\nu_2 = \sum_{j=0}^{k} n_j - k$$

—The populations are all assumed to be normally distributed.

—All the populations must have the same variance.[8]

E. ANOVA F–test for the comparison of sample variances

	Hypothesis to Be Tested	***Test Statistic***
Two populations with continuous measures X_1 and X_2	$Ho : \sigma_1^2 = \sigma_2^2$	$F = \frac{S_1^2}{S_2^2}$

Comments and assumptions:

—The populations must be normally distributed.

—The larger sample variance is always placed in the numerator; H_1 is $\sigma_1^2 \neq \sigma_2^2$.

—The F value has two degrees-of-freedom values:

$\nu_1 = n_1 - 1 \quad \nu_2 = n_2 - 1$[9]

NOTES

1) Bruner, *Theory of Instruction,* 49.
2) Hughes–d'Aeth, "Classroom Organization", 21.
3) Ibid., 22–23.
4) Ibid., 23.
5) Billingsley et al., *Statistical Inference for Management,* 452.
6) Ibid., 681–682.
7) Ibid., 324.
8) Ibid., 387.
9) Ibid., 324.

SELECT BIBLIOGRAPHY

Auerbach, Carl and Joseph L. Zinnes, *Psychological Statistics,: A Case Approach.* Philadelphia: J. B. Lippincott Company, 1978.

Billingsley, Patrick, et al., *Statistical Inference for Management and Economics,* 3rd Edi-

tion. Boston: Allyn and Bacon, Inc., 1986.

Bruner, Jerome S., *Toward a Theory of Instruction.* New York: W. W. Norton & Co., Inc., 1968.

Hughes—d'Aeth, Armand, "Classroom Organization", *Cross Currents,* Vol. XI, No. 2, 1984.

〔 抄 訳 〕

トラック・システムを採用した学習法の効果に関する調査研究

一般に大学教育では採用されることはないが，英会話学校で広く採用されている“track system”所謂能力別クラス編成による教育が，四年生大学での必修英会話の講座においても効果があるものかどうかを統計的に調査研究しようとすることが本研究の目的である。

つまり，「四年生大学での英会話の教育は，mixed ability groups での教育より，track system を採用したクラス編成による教育のほうが有効である。」という仮説を種々の統計的手法を用いて検証しようというのが本稿のテーマである。

この仮説を検証するために，1988年度に本学に入学した全学生に実施した英語読解力，構文，発音テストの得点の結果に依り，上位1/3を higher track，残り2/3と再履修者を lower track というクラスに分け，前期の初期と後期の期末にそれぞれ聴き取り能力をテストする HAC pre–test と HAC post–test を実施した。次の 5 種類の統計的手法に依り 2 回の HAC test の結果を分析した。

(1)相関係数

(2)スペアマンの順位相関係数

(3)対応のある平均差検定

(4)分散分析

(5)等分散検定

これらの統計的手法による分析結果はそれぞれ次の通りであった。

(1) $r = 0.548$. 無相関検定に依り帰無仮説は棄却され，相関があると判断できた。

(2) $r_s = 0.513$. 無相関検定に依り帰無仮説は棄却され，相関があると判断できた。

(3)18クラス中13クラスは帰無仮説が棄却され平均差があると判断できたが，6つの higher track class では 2 クラスにおいてだけ帰無仮説が棄却され，平均差があると判断できたにすぎなかった。

(4)両 HAC テストにおいて帰無仮説は棄却され，各クラスの能力差は存在すると判断できた。

(5)全18クラス中 6 クラスは帰無仮説が棄却され等分散でないと判断できた。

以上の統計的分析の過程において，クラス分けをする際に採用した読解力を中心としたテストの結果と HAC pre–test の結果を比較すると IV. DISCUSSION の D. TABLE 3 に見られるように約29％のクラス割りふりミスが発見された。このため本研究で提起した仮説が真であることを十分検証するに至らなかった。

しかし，III. STATISTICAL ANALYSIS の B. TABLE 2 に見られるように，1987～88年次の学生に実施した HAC pre–test の分散のほうが HAC post–test の分散より大であったが，1988～89年次におけるその大小は逆転してしまっており，track system の負の効果も発見された。

今後，IV. DISCUSSION にある 4 つの反省点を十分議論しつつ，Jerome S. Bruner による mixed ability groups に関する idea も概観しながら，track system の正のみならず，負の効果についても改めて検証研究したい。