EFFECTS OF A FLUENCY DRILL COMPONENT IN MATHEMATICS INSTRUCTION

By Robert F. Orgel

ABSTRACT

This study, conducted in an introductory-level math course at the University of Kansas, compares the performance of students using the Fluency II Learning System with traditional course presentation and study methods. To simulate "real world" conditions the Fluency II Learning System was implemented under less than optimal conditions. Even so, students using Fluency II scored an average of 29 points higher than the control groups on three separate examiations. Not only was the variability significantly reduced, but the size of the difference increased on each successive examination. The group exposed to the Fluency II Learning System performed more than two times better on generalization probes consisting of untrained word problems. Differences of this magnitude appear nowhere else in the experimental, educational or training literature.

Theoretical Background. Earlier studies by the investigator suggest that terminal performances in what are generally regarded as "cognitive" tasks might be significantly improved by strengthening each "task" to both an accuracy criterion (conventionally measured by percent correct) and a fluency criterion (measured by number correct per minute). In a previous study, the investigator demonstrated that students trained to perform at high terminal frequencies retained 3 to 5 times more than low frequency students, and two times more than middle frequency students on a task where they were equally accurate, and that these differences were durable and significant. For an introductory-level math course, Math 000, an attempt was made to incorporate the generalization technology suggested by Stokes and Baer and the frame and task analysis suggested by Markle. Important concepts were trained for reflexivity, symmetry, and transitivity as suggested by Sidman, and trained both serially and concurrently as suggested by Schroeder and Baer, and Lindsley, et. al.

Method. An introductory-level Mathematics course at the University of Kansas was used as a test site for this procedure for generating and presenting curriculum. The Fall 1981 section became a control condition for all subsequent measurements and procedures. All components of the course, the examinations, the instructor, the book, etc. remained constant between the Fall and Spring semesters with the exception of the Fluency II drill component, which was added to the Spring semester presentation of the course. The following analysis will display an important difference between the two semesters that can confidently be attributed to the added Fluency II drill component. While a confound may exist due to potential differences between the students for the Spring and Fall semesters, previous

analysis by others suggest that Spring semester students tend, on the average, to score less well than the Fall semester students in remedial math courses. In addition, the group of students who independently elected not to use the drills provided an extra dimension of comparison, and thus help account for this potential source of variation. (Note: The number of students (N) in the Figures discussed below decreased in all groups across tests due to dropouts and students choosing not to use Fluency II Drill cards.)

Results for Test 1. Figure 1 shows the results for the first hour examination in Math 000. The test counted 100 points out of a possible 700 (14%) and covered Chapters 1 and 2 (out of seven chapters). The histograms on the left side of Figures 1, 2, and 3 provide raw frequency distributions, and the histograms on the right side of the figures show relative, or percentage distributions, which permit clear comparisons in the shapes of the distributions regardless of the absolute numbers. Figures 1A and 1B show the distribution of scores for the control group, Fall 1981 (N = 65, Mean = 77.64, Standard Deviation = 16.0). The other graphs in Fig. 1 are from groups in the experimental Spring 1982 class. Graphs 1B & 1G show the scores for those who elected not to use the additional drill component provided (N = 29, Mean = 77.27, Standard Deviation = 14.95). The similarity of these distributions to 1A & 1B show that the groups are originally well matched, and (2) that treatment effects other than the drill component must have been administered with approximately equal strength to the two groups.

This group of non-users (Figures 1B and 1G), although self-selected, serves as a functional control group in the experiment. If all other factors are roughly equivalent, this pattern of similarity should continue to hold between nonusers in both Spring and Fall semesters. When the graphs are normalized for differences in number by using percentage on the ordinate, the case that these two groups (1F and 1G) represent identical underlying populations prior to treatment becomes more convincing. The continued similarity in the distributions of nonusers for both semesters for all examinations improves our confidence that the differences noted are a function of the experimental manipulation, not some random variation.

Graphs 1C and 1H represent a group that used only one set of Fluency II Drill cards, typically from Chapter 1 (N = 16, Mean = 80.7, Standard Deviation = 11.98). This group showed a useful improvement when compared with the nonusers (1B): it eliminates the lower tail of the distribution.

Graphs 1D and 1I represent the group that made and reached criterion on both sets of Fluency II drill cards (Chapters 1 and 2). This group did, on average, 16 points or two letter grades better (N = 23, Mean = 93.34, Standard Deviation = 3.55) than either of the control groups or the group that used only one set of cards. A T-Test for independent groups reports a difference significant at the .001 level when compared with 1A, 1B, or 1C.

Graphs 1E and 1J represents scores for the Spring 1982 experimental class (N = 70, Mean = 83.24, Standard Deviation = 13.49). Even on this gross level of comparison, there is a difference between the Fall and Spring semesters of 6 points (from a grade of C to a B) which is significant at the .02 level. In addition, 42% of the Spring semester class scored 90% or better, compared with 21.5% for the Fall group. When the sources of variation are designated by sub-groups, however, it is clear that the majority of the difference is attributable to group 1D, the group with maximum exposure to the experimental procedure.

Results for Test 2 (Figure 2). The second hour exam replicates the patterns noted for Test 1. Covering Chapters 3 and 4, and, like Test 1 worth 100 points our of 700 (14%), graphs 2A and 2E represent the scores for the Fall (control) group (N = 57, Mean = 70.4, Standard Deviation = 26.56). The nonusers in the Spring (functional control) group graphs 2B and 2F (N = 37, Mean = 72.85, Standard Deviation = 22.65) have distributions similar in shape, central tendency, and dispersion. Graphs 2C and 2G represent the group that used and demonstrated fluent mastery of the Fluency II drill cards. Their scores (N = 25, Mean = 100.28, Standard Deviation of 10.38) demonstrate a powerful experimental effect, similar to that seen on Test 1, but stronger.

Graphs 2D and 2H show the distribution for the entire Spring class (N = 62, Mean = 85.15, Standard Deviation = 21.15). Even without the information that most of the improvement came from the treatment group (2D) the comparison of the aggregated Spring and Fall groups yields a difference significant at the .005 level on an independent group T-Test. In addition, all group members scoring in the 110-120 range came from the experimental group, and this group was five times larger in the Spring than in the Fall. Due to the magnitude of the experimental effect, 56% of the Spring class scored 90% or better, while only 30% of the Fall group scored in this range.

Results for Test 3 - Final Examination (Figure 3). The Final Examination for Math 000 counted 200 points (twice the other two tests) and was non-cumulative, covering Chapters 4, 5, and 6. As in our previous examples, the upper left hand graph (3A) represents the raw frequencies on the final exam for the Fall 1981 (control) group. (N = 45, Mean = 68.5, Standard Deviation = 18.07). Again, these scores are closely replicated by the scores for non-users in the Spring 1982 (functional control) group 3B (N = 34, Mean = 67.8, Standard Deviation 18.20). As in our two previous examples, the discriptions are almost identical to each other in shape, central tendency, and dispersion. The experimental group, represented by graphs 3C and 3G, was presented Fluency II drills for only two of the three chapters, and scored, on the average, almost 30 points, or three letter grades higher (N = 14, Mean = 96.5, Standard Deviation = 9.82) than either of the control groups. A T-Test for independent groups reports a difference significant at the .001 level when compared with 3A or 3B.

Graph 3D shows all the scores for the Spring 1982 class (N = 48, Mean = 75.45, Standard Deviation = 20.13). Once again, even this relatively gross level of comparison provides a difference of 7.65 which is significant at the .04 level. Almost all the improvement came from less than one-third of the class.

The Fall group (control) had 12 students out of 45 (26.6%) in the 80 and above category, while the Spring group (experimental) had 25 out of 48 (52%) in this group. Twelve of the top 15 scores for the Spring group came from the Fluency II-drill condition. Of the 25 above 80%, 12 were from the control group, showing that self-selection accounted for little or no variance compared with the Fluency II drill cards. The experimental group not only did better for the aggregate and the 80 and above scores, they had twice as many scores in the 100 + and the 90-99 categories as well. The improvement was systematic and stable.

These results are even more impressive when examined against the background of a sub-experiment which was performed on the Final Examination. The differences found for Tests 1 and 2 between experimental and control conditions

were so large that they posed no threat to the believability of the study. It could still be true, however, that other variables, either that the Spring 1982 class was in some way superior, or that an artifact of self-selection threaten the validity of the study. These threats to external validity always exist in cases in which random selection is not possible. Applied Behavior Analysis, which deals with human subjects in quasi-experimental designs, addresses this question with multiple-baseline designs in which a matched behavior is measured but no intervention is applied until control has been established, or reversal designs, in which an experimental treatment which has proved successful, is removed.

The curriculum of the introductory level Mathematics class lent itself uniquely to the use of a reversal design to demonstrate that the underlying populations for both Spring and Fall groups were the same. For a reversal to be appropriate, the behavior should not be able to maintain itself, and should not be a cumulative task where former practice might hinder interpretation. Chapter 10, however, consisted of Geometry, and part of Chapter 6 consisted of a subset of Analytic Geometry. Both of these curricula were discontinuous with the Algebra presented in Chapters 1-5, and would thus provide a unique opportunity to perform a reversal.

It was decided that no Fluency II Drill cards would be provided for these items, and that the performance of the Fall and Spring groups on a controlled task could be observed. If these in fact represented the same populations, then test scores for these items should be relatively similar.

The last histogram in Figure 5 shows the results for this Study. The Fall 1982 (A - control) group had a score of 69% on these items while the Spring 1982 (D - Experimental) group scored 68%. The group of problems from which these data come from represent 70.5 out of 121 possible points (58.25) while the Algebra component of the Final for which Fluency II drill cards were provided represented 50.5 out of 121 points, or 41.75%. Thus, this comparison was made on a substantial enough subset of Test 3 to reject the hypothesis of different underlying populations with confidence. In addition, the fact that the two groups of Fluency II drills distributed to the class covered only 42% of the material covered on the examination makes the differences already noted between the experimental and control conditions for this test that much more impressive.

Generalization Probe - Test 2 (Figure 4). Another important question is whether Fluency II drills of the specific kind developed by the investigator facilitate generalization to more difficult skill tasks. The task designated here as a generalization probe is the ability to successfully solve mathematical problems that come in the form of English sentences, instead of explicit mathematical notation. These are commonly denoted as "word or story problems." All the students were presented with the same material from the book, and the same homework assignments covering this material. The Fluency II drills for Test 2 contained a unique task analysis that breaks word problems into three component parts: (1) translation of the unknown elements into variable notation; (2) use of the variable notation from 1 to provide a suitable equation; and (3) the solution of the equation. Because the scores for the Fall semester already established a subpar performance on these problems relative to solving equations where the equations are given to the student, it was

decided to emphasize training of the first and second skills, with emphasis placed on the second. A wide variety of problem types were presented which the student would have to recognize almost instantaneously to meet the criterion of one correct answer every 3.25 seconds (35 correct in two minutes).

Results for this portion of the experiment are consistent and convincing. In every case, for all six problems, the Fall and Spring groups of non-users (A and B) were significantly lower than the Spring 1982 Group C which used Fluency II Drill cards and demonstrated criterion-level mastery in both fluency and accuracy. Taken as a group, the Fall 1981 Group A, who were non-users, average 41.47% for these six problems. The Spring 1982 non-users, Group B, scored 47.83% as a group. And the Spring 1982 group, which used and mastered the Fluency II drill scored an average of 83.83%, better than twice the score of the control group. Again there is a nearly complete functional matching between Groups A and B, with Group B just slightly higher in every case. The difference between the control groups (A and B) is so small when compared with the Experimental Group C, that the conclusion that we are dealing with a powerful experimental intervention that significantly improves the ability of student to successfully do word problems in mathematics is extremely clear and convincing.

Generalization Probe - Test 3, Final Examination (Figure 5). The experiment initiated on Test 2 was continued on the Final Examination (Test 3). An attempt was made to extend the reliability of the findings by replicating with four additional similar word problems, problems 37 through 40. The mean performance for the Fall 1981 Control Group (Group A) was 43.75. Scores for Group B, Spring 1982, Functional Control Group for the same problems was 49.25 %. Scores for Group C, the Spring 1982 Experimental Group, which demonstrated mastery of the Fluency II drill presented at a rate of 17.5 per minute averaged 83.5%.

Results for the generalization probe on Test 3 validates the generalization probe for Test 2. In all four cases the ordering effect seen on Test 2 is seen on Test 3. Nonusers (no exposure) scored the lowest; users with a limited exposure to Fluency II training scored slightly but not significantly better. Those demonstrating fluent mastery were approximately twice as accurate on standard examinations as either of the other two groups. The probability of seeing both this regularity in the order and the magnitude of the results occurring by change is less than .0001. In only one of the ten problems (Problem 46-Test 2) did the average score for the Experimental Group fall beneath 80% correct. This problem, however, was apparently the most difficult in the set, with scores for Group A of 21%, and scores for Group B of 22%. Contrasted with these scores, Group C with an average of 56%, represents a 167% improvement, a factor of x 2.67, compared with Groups A and B, the largest improvement index recorded for the 10 problem set. The mean improvement factor for the generalization probes taken together was 109%, or x 2.09, with a standard deviation of .40, indicating not only a highly successful, but an very stable, reliable, and lawful experimental result.

DISCUSSION. There can be no doubt that Fluency II Training significantly improved the performance of the students who used it. It must be pointed out that the experimental procedure was utilized at the lowest boundary of expected effectiveness. For this study a criterion of 35 correct responses per two minutes (17.5 per minute) was specified

fluent mastery. Other studies we have done demonstrate definitively that a criteria of 40-60 correct responses per minute produced much more durable learning as measured by retention and generalization. Another unfavorable experimental condition was that the contingencies for mastery of the Fluency II Drill cards represented only 15% of the grade, which was too weak a contingency to adopt and perfect a totally new repertoire of study behavior. Is it possible, then, that the higher exam scores can be accounted for by self-selection? Could it be merely that better students were made better? Or that by adding an additional task, another way to predict the more "motivated" students was found?

The experiment was designed to control for these possibilities in a number of ways. First, there is the strong relationship between the Fall Group (A) and the Spring Nonusers (B) mentioned earlier (see Table 1). Table 1 summarizes the data from this study. Groups A and B are remarkably similar on all these examinations in shape, dispersion, and central tendency, and are uniformly different from Group C (the experimental condition). When the distributions are normalized for size by comparing percentages instead of raw frequencies (Figures 1, 2, and 3 - 1F and 1G, 2E and 2F, 3E and 3F), the similarity is clear across important characteristics.

To get the Functional Control Groups (1B, 2B, 3B) to match the True Control Groups (1A, 2A, 3A) and still have a significant effect of the kind noted for the Experimental Group we would have had to pre-select a functional control group of students who were one grade point higher initially. Then, with the lower end of the distribution removed, the remainder would appear much like the control group. This is an unlikely explanation because of the close matching for shape, as well as central tendency and dispersion on all three distributions. The probability of getting repeatable similarities of the type noted is highly unlikely if underlying populations are different. In addition, the "good student" alternative hypothesis is refuted because the same relative percentage of A's and B's were obtained in the Spring Nonusers. There are two other areas of comparison, however, which allow us to refute this threat to interpretation with impunity.

The generalization probes provide another view of the relative performance of groups A, B, and C on matched tasks. The stability of the relationship noted across all ten probes for the A and B groups significantly increases our confidence that they are from the same populations. The fact that these data show the B Group had slightly higher scores than the A Group represents no threat to the validity of the study because of the small relative magnitude when compared with the Experimental Procedure (C). Thus, if there existed an a priori difference in the underlying populations it is of slight importance when compared with the amount of improvement in performance accounted for by the Fluency II Drill component.

There is another hypothesis which, in light of the summative results, is more tenable - that there was some Fluency II Drill usage by B group, and that the improvement noted is a function of amount of drill. In addition, the mere exposure to the sheets would tend to serve the same function as distribution of a note sheet which highlighted the important items to be studied for a given section. This would tend to upgrade the study time of those who used ordinary techniques, and would account adequately for the slight improvement recorded. Additionally, the fact that the number of Fluency II Drill users decreased with the passage of time means that, for Test 2, there were a number

of students who had previously used the drills for Test 1, but did not for Test 2. The same is true to an even greater extent for Test 3, where only 14 students demonstrated fluent mastery on the two chapter drill sets compared with 39 for the first test, and 25 for the second. If, as previous studies indicate, the improvement provided by the usage of this procedure is durable, we would expect it to show up in terms of a slight improvement on the generalization tasks for these students.

A final comparison makes this last hypothesis the more plausible of the two. The comparison for both semesters on the Geometry section of the Final Examination suggests that if anything, the Spring group had a slightly less mature mathematical ability. Not only did the Spring group score less, but 5 to 10 of their number made their own cards and drilled themselves. Thus, a true controlled comparison would probably have yielded a difference of from 4-8% less for the Spring group, enough to yield a statistically significant difference confirming that the Spring group was slightly inferior to begin with.

The evidence clearly suggests, even without this comparison, that the underlying populations were not different in any important way prior to the introduction of the experimental procedures used, and that the differences noted are certainly a function of those procedures. The data and analysis presented provide incontrovertible support for the conclusions reached.

CONCLUSIONS

1. Even given a weakened version of Fluency II Drill procedure, a mean improvement of 27.8, or almost 3 letter grades was noted in this study.

2. In all cases the variability of scores in the experimental group was significantly lower than the control groups for the areas Fluency II was used, showing a high degree of experimental control.

3. On the most difficult group of problems for this population, the experimental procedure reliably improved performance by a factor of more than two.

4. Without the benefit of the Fluency II Drill procedure, the performance of the Spring group would have been identical to the control group. All differences noted in the study are clearly a function of the procedures used.

IMPLICATIONS

1. The Fluency II Learning System operating at only partial efficiency provides significant and important improvements in learning technical and factual information.

2. The significant improvements in generalization noted suggest that the Fluency II Learning System is extremely well suited for training complex conceptual skills. This makes the Fluency II an ideal choice for developing managerial, supervisory, and sales training programs.

3. The high positive correlation between increases in fluency and increases in generalization suggests that optimal levels of fluency increase generalization by 3 to 5 times that of traditional training technology.

4. The size and reliability of gains recorded with use of the Fluency II Learning System suggest that it currently represents the most advanced training technology available for developing a broad range of skills including complex concept formation and problem solving skills. It will provide impressive results for any training program to which it is applied.

GROUPS SUMMARIZED BY CONDITION AND SEMESTER

TABLE 1

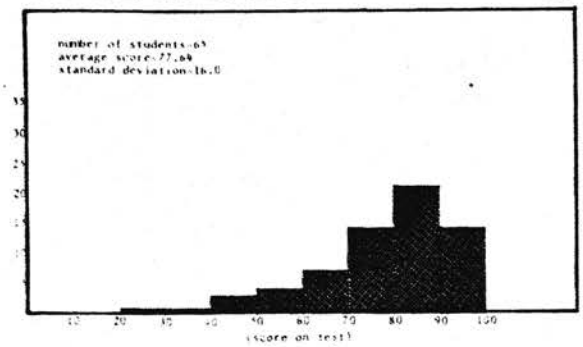
	FALL 1981 (A) NON-USER (CONTROL)	SPRING 1982 (B) NON-USER (FUNC. CONTROL)	SPRING 1982 (C) EXPERIMENTAL
TEST 1	N = 65 M = 77.64 SD = 16.0	N = 29 M = 77.27 SD = 14.95	N = 23 M = 93.34 SD = 3.55
TEST 2	N = 57 M = 70.4 SD = 26.56	N = 37 M = 72.85 SD = 22.65	N = 25 M =100.38 SD = 10.38
TEST 3	N = 45 M = 68.5 SD = 18.07	N = 34 M = 67.8 SD = 18.20	N = 14 M = 96.5* SD = 9.82

* Received Fluency Drill Training for only 2 out of 3 Units Tested.

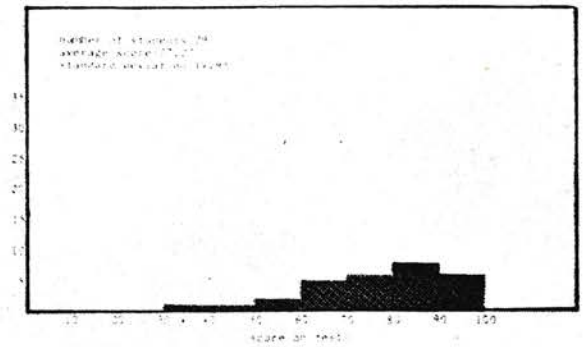
FIGURE #1

TEST #1 (by number of students)

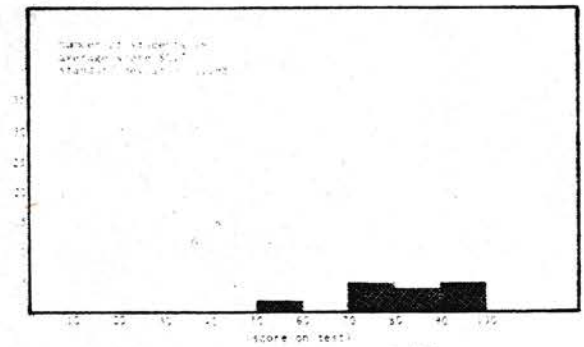
Fall 1981: all non-users 1A



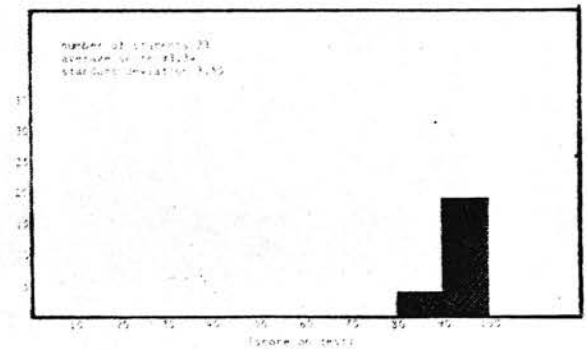
Spring 1982: all non-users 1B



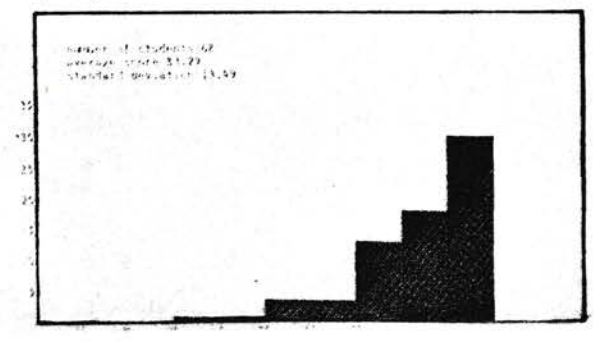
Spring 1982: 1 set of Fluency Drill 1C



Spring 1982: 2 sets of Fluency Drill 1D

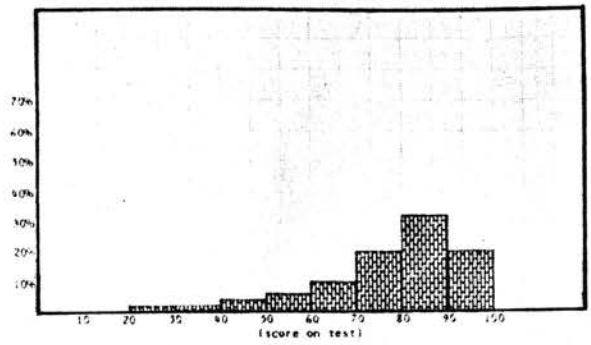


Spring 1982: entire class 1E

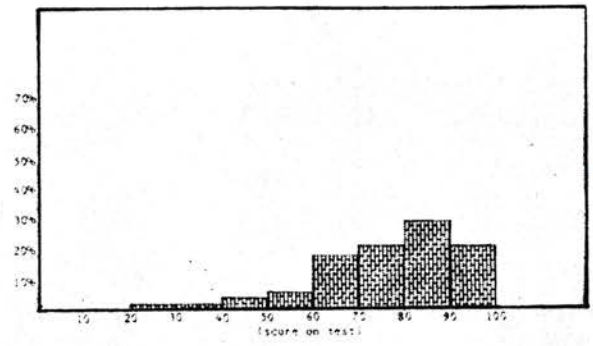


TEST#1 (by percentage)

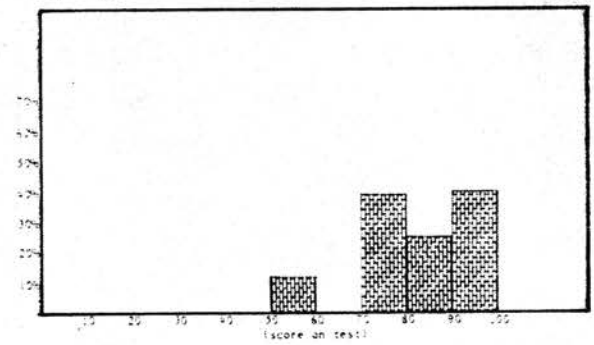
Fall 1981: all non-users 1F



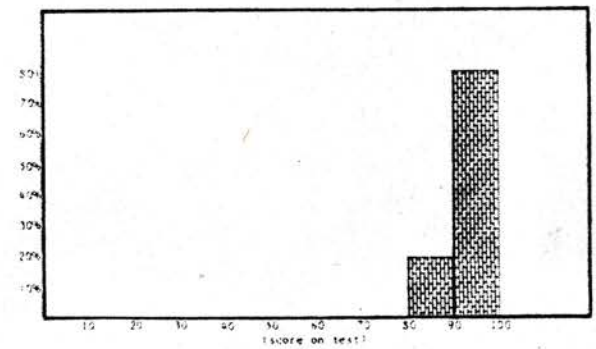
Spring 1982: all non-users 1G



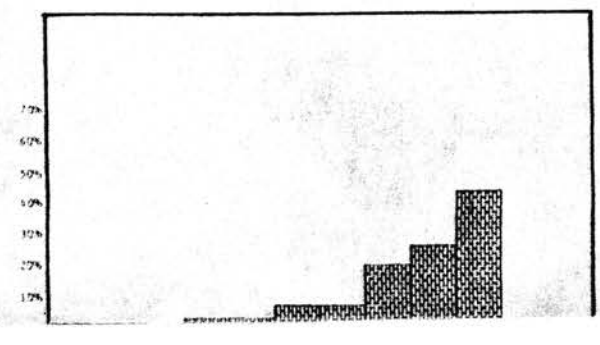
Spring 1982: 1 set of Fluency Drill 1H



Spring 1982: 2 sets of Fluency Drill 1I

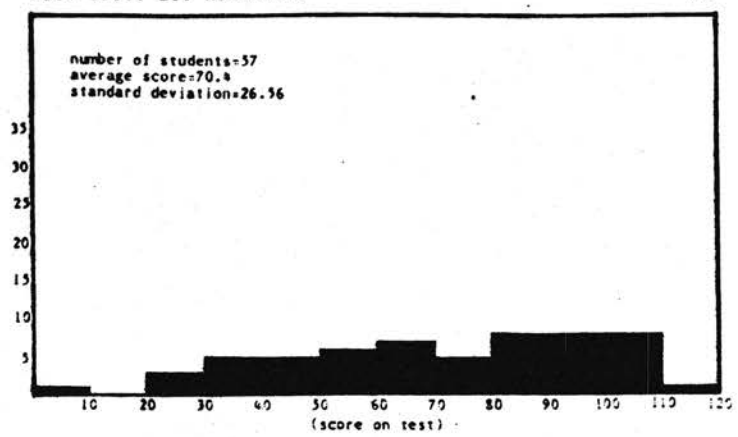


Spring 1982: entire class 1J

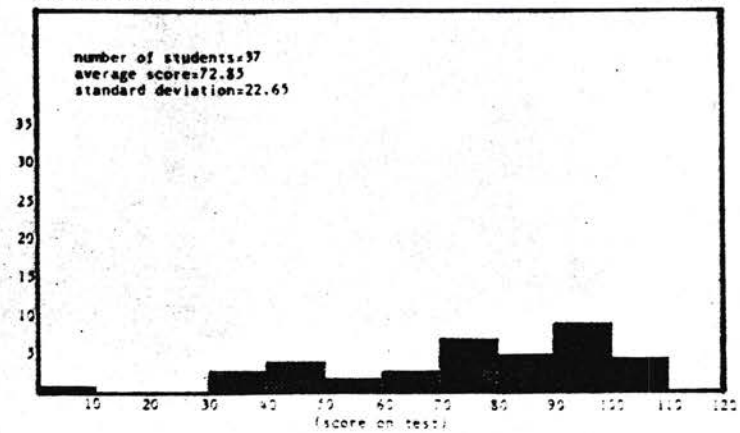


TEST #2
(by number of students)

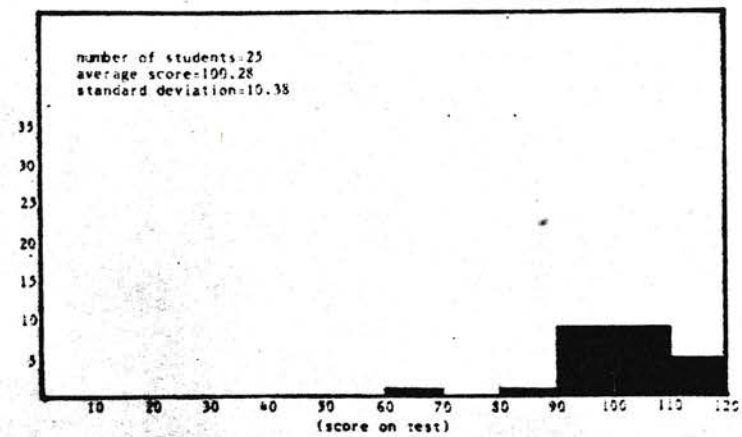
Fall 1981: all non-users 2A



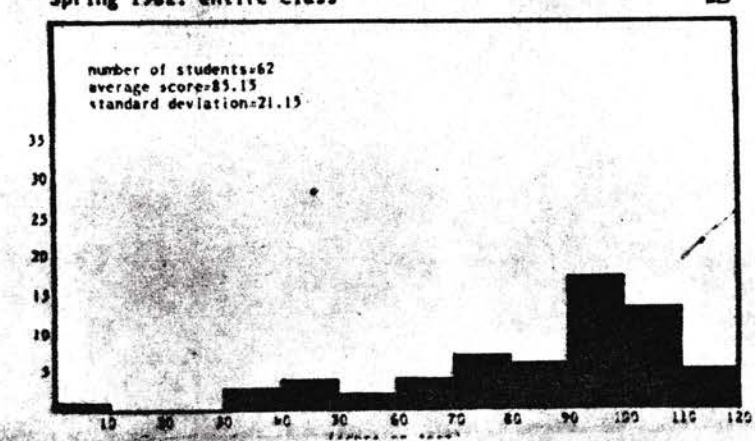
Spring 1982: non-users 2B



Spring 1982: full Fluency Training 2C

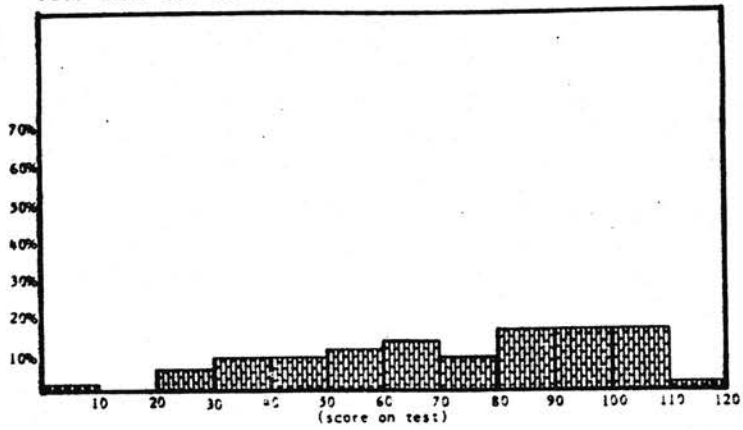


Spring 1982: entire class 2D

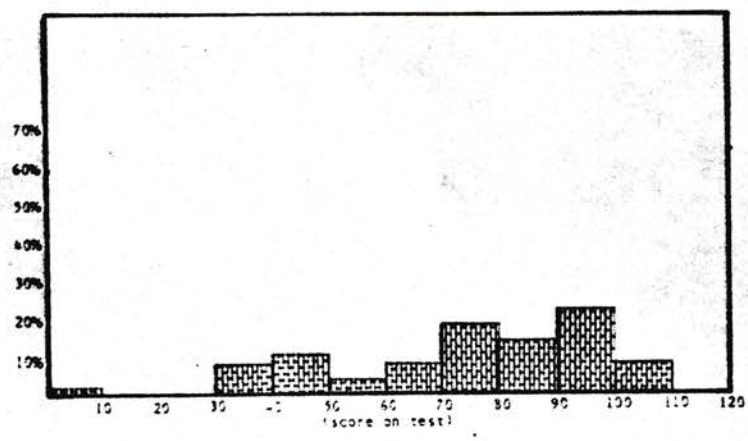


TEST#2
(by percentage)

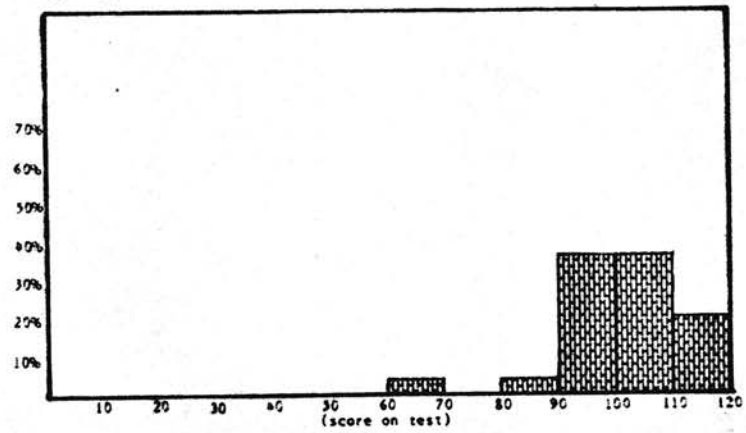
Fall 1981: all non-users 2E



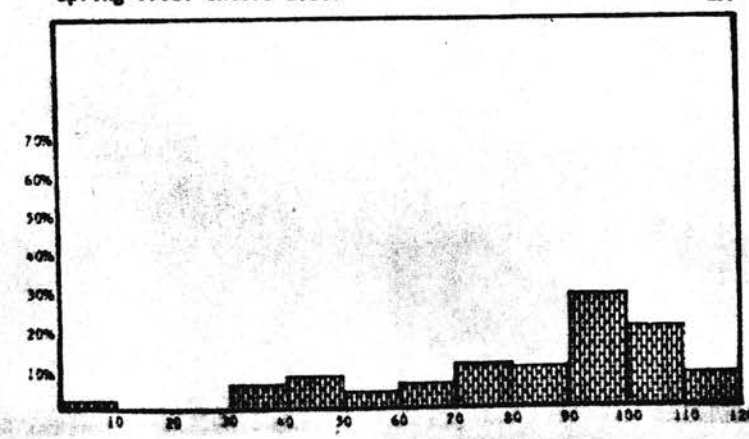
Spring 1982: non-users 2F



Spring 1982: full Fluency Training 2G



Spring 1982: entire class 2H

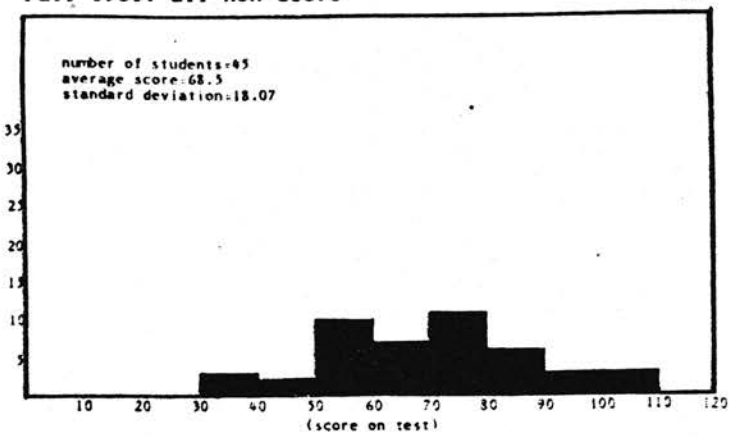


TEST #3(final)
(by number of students)

TEST#3(final)
(by percentage)

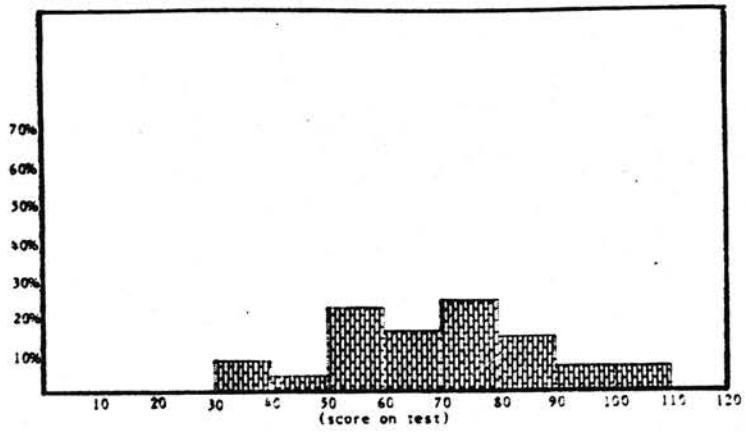
Fall 1981: all non-users

3A



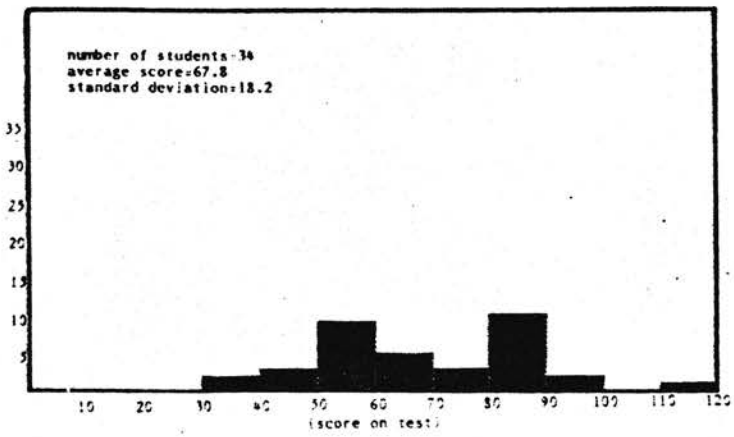
Fall 1981: all non-users

3E



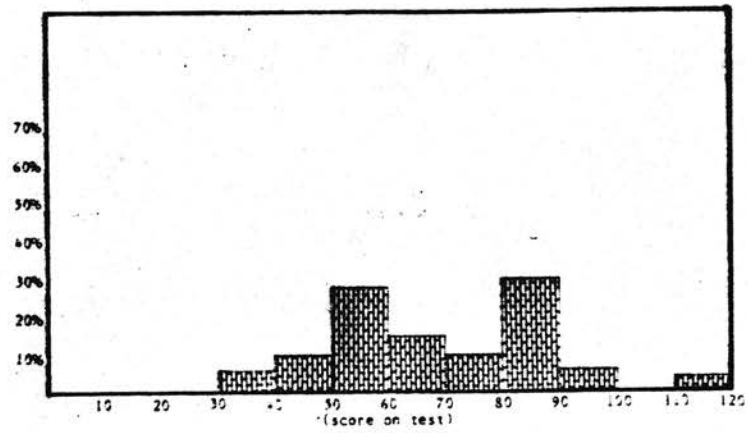
Spring 1982: non-users

3B



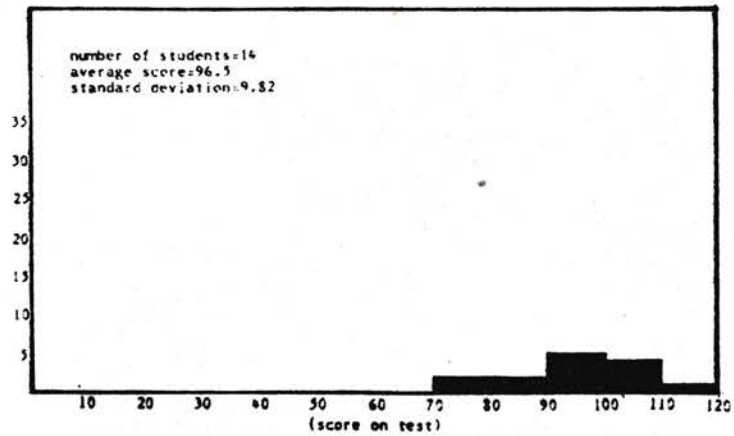
Spring 1982: non-users

3F



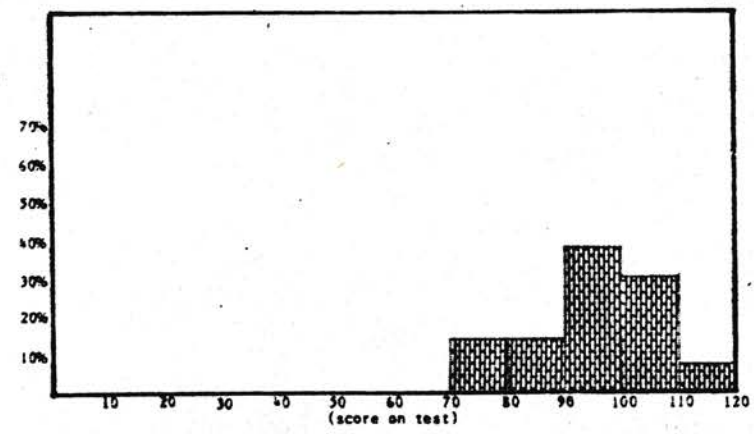
Spring 1982: full Fluency Training

3C



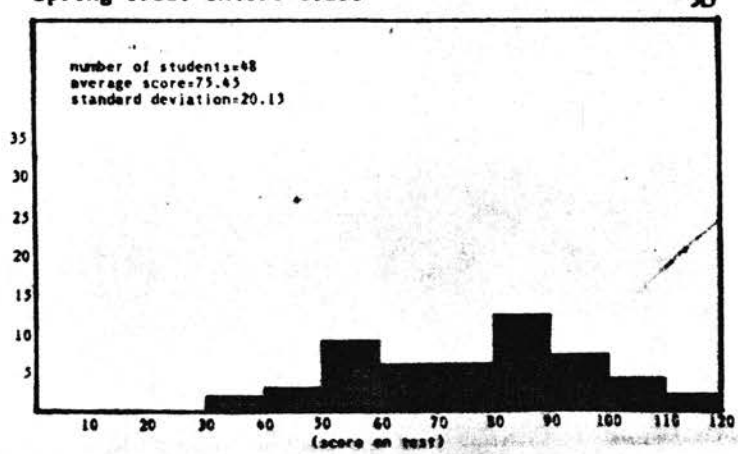
Spring 1982: full Fluency Training

3G



Spring 1982: entire class

3D



Spring 1982: entire class

3H

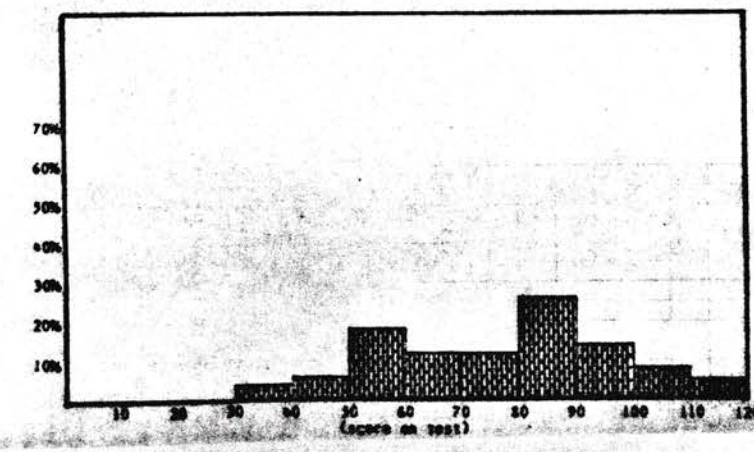


FIGURE #4: PERCENT CORRECT ON SIX WORD PROBLEMS (TEST#2)

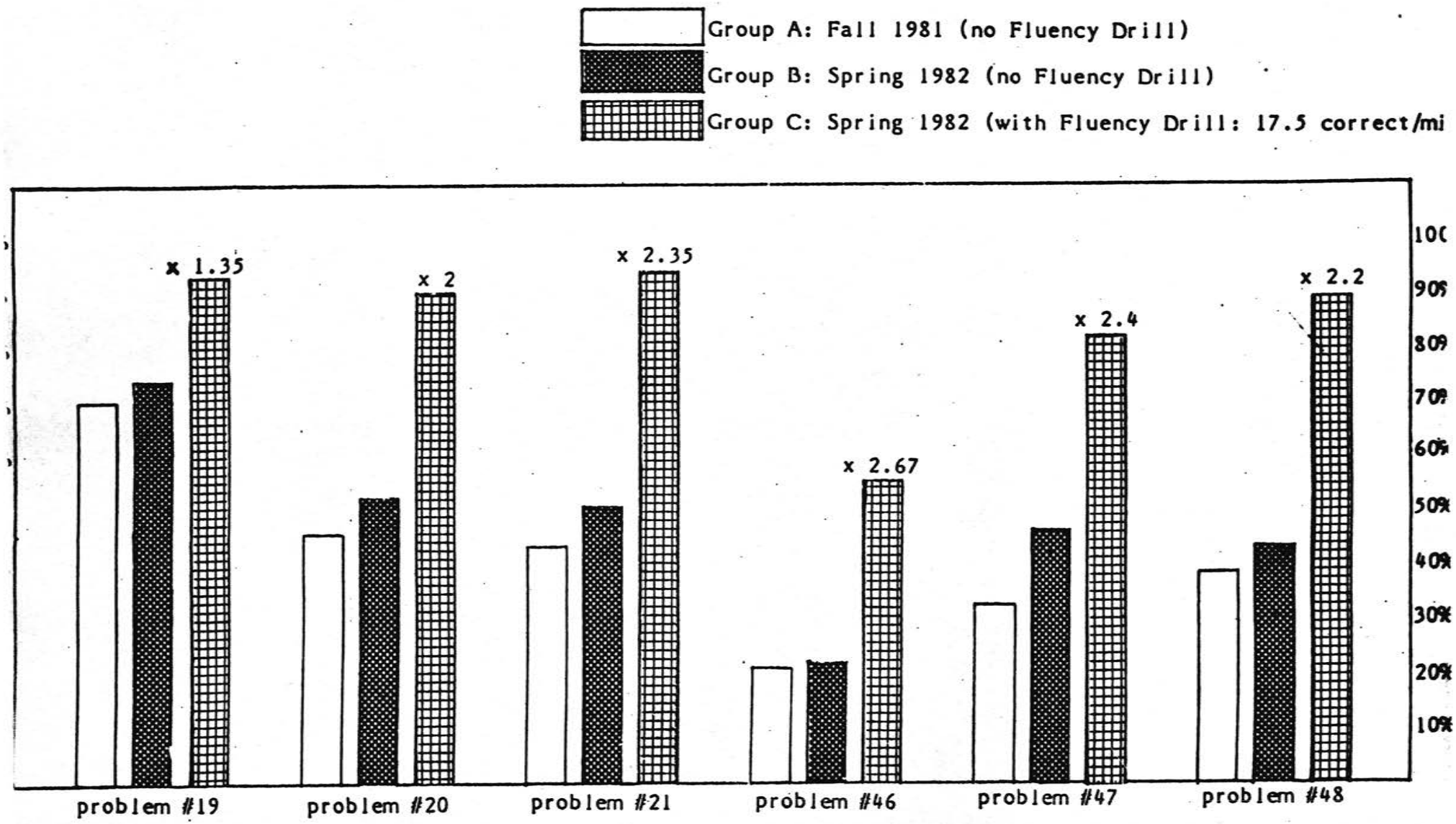


FIGURE #5: PERCENT CORRECT ON FINAL
(WORD PROBLEMS AND TOTAL GEOMETRY)