Br. J. educ. Psychol., 61, 240-247

EVALUATION OF THE TEACHING OF STATISTICAL CONCEPTS BY INTERACTIVE EXPERIENCE WITH MONTE CARLO SIMULATIONS

BY C. G. WEIR, I. C. McMANUS AND B. KIELY
(Department of Psychology, University College, London)

SUMMARY. Two studies evaluated the effectiveness of interactive computer demonstrations in teaching about statistical sampling distributions. Students used a Monte Carlo simulation of either standard errors (SEMDEMO) or F-distributions (FDEMO) and were subsequently tested on both concepts. In Study 1 only lower ability students using FDEMO showed improved attainment related to their specific experience. SEMDEMO was then simplified, following student feedback, and study 2 then showed higher specific attainment related to interactive experience with both SEMDEMO and FDEMO, particularly in lower ability students. Reasons for improved performance may include increased practice and deeper processing of concepts.

INTRODUCTION

Ambitious claims have been made that computer technology will enhance teaching and solve problems connected with shortage of staff (Clayden and Wilson, 1988, Hammond and Allison, 1988, Harrison *et al.*, 1988), although in practice many projects have not fulfilled their promise (Ingham, 1988). Evaluation of teaching by computer has been frequently advocated (Bork, 1985; Blease, 1986; Johnston, 1987; Gardner, 1988).

Criteria for evaluation of educational initiatives include diverse aspects, often more technological than pedagogical. For instance, Lieblum (1988) lists flexibility, transportability, graphic support, reliability, and student management as important features. While technological criteria are important, it is also valuable to explore both cognitive attainment and affective responses to computer-based learning (Kemmins *et al.*, 1987). Interacting with a program may force a student to process material more deeply which will lead to improved memory performance (Craik and Lockhart, 1972; Craik and Tulving, 1975). Some possible mechanisms for improvement are that interactive experience will make cues more distinctive or improve the organisation of memory so that retrieval is enhanced (Bower *et al.*, 1969) or simply improve memory by lengthening practice time.

In most computer-based education projects, learners have more control than is usual in a classroom situation. Wong (1987) showed that experience with a dynamic modelling system influenced pupil choice of physics problems in a national examination and reduced imprecision errors. Gay (1986) showed that biology students were able to take advantage of the freedom to sequence information if they had high aptitude but not if they had low aptitude.

The present study examined the role of active participation in enhancing attainment and positive attitudes, and considered the additional value provided by interactive experience compared with lectures in understanding statistical concepts, within the context of normal teaching. During their second year statistics course, two matched groups of psychology students were given interactive experience with one of two demonstrations concerned with sampling distributions, a core concept for understanding experimental design in psychology. Few psychology students are good at applying sampling distributions to experimental situations (Tversky and Kahnemann, 1974; Greer and Semrau, 1984).

By using two demonstrations, one on standard error of the mean (sem) and the other on the F-statistic in an analysis of variance context, we evaluated the specificity/generality of concepts attained during the interactive experience. Both demonstrations used Monte Carlo simulations, repeated samples being drawn from theoretical distributions, with histograms of sample means or F-statistics cumulated on the visual display unit on successive iterations. Subsequently questions about each of the concepts were answered by students on statistics tests.

If attainment is specific to the type of interaction then students in the standard error demonstration group (SEMDEMO) should be better at answering test questions on standard errors than the group

carrying out F demonstrations (conversely the FDEMO group should do better on questions relevant to their experience). If subjects generalise from one demonstration of sampling distributions to another, then the SEMDEMO group should have good performance on both concepts whereas the FDEMO group should only have better performance on the F questions. This asymmetry was due to sem always being taught before F-distribution.

STUDY 1

METHOD

Participants

Thirty-nine second-year university psychology students were assigned to matched groups according to their marks from the first year statistics class: SEMDEMO group (N=20, first year mean = 57 per cent) and FDEMO group (N=19, first year mean = 60 per cent).

Apparatus

A network of 12 RML Nimbus computers was used to present demonstrations which used Monte Carlo simulations. They were written originally in BBC BASIC but were later translated into PASCAL so it would execute faster, and could access an extensive graphics library and a superior random number generation function.

Programs

Two demonstrations, SEM (standard error of mean) and F, comprised several sections on a display. At the top was a window where a sampling of points from a normal distribution was animated graphically. Next, values of relevant statistics for the current sample were presented, viz. mean and variance for the SEM demonstration; and mean, stand deviation (SD), sums of squares, and F for the F demonstration. At the bottom of the display were windows where results from repeated samples were cumulated in histograms and tables. Figure 1 shows examples of full screens from the demonstrations after a small number of runs have been performed. Users had a menu of choices at the outset of each program where it was possible to choose whether to run one random sample at a time or to run a specified number of samples rapidly. They were also able to select values for some parameters: (1) SEM demonstration, sample size; and (2) F demonstration, sample sizes, relative difference between the means, and SDs.

Materials

Worksheets were devised for the demonstrations directing users initially to run a few Monte Carlo samples slowly to confirm the values of each calculation; then to run many replications rapidly to observe the distribution of the relevant statistic. In the SEM demonstration, users set different values of the sample sizes in the two parallel samples, e.g., N=4 and N=25, so that noticeably different sem values would be obtained. In the F demonstration, the worksheet directed users first to observe samples with the means set equal (Ho in analysis of variance; then to set a difference of 1.0 (SD units) between the sample means (as illustrated in Figure 1), and finally to vary sample variances.

On both sheets, applied questions were interspersed throughout to help students generalise from the demonstrations to psychology experiments. These "applied" questions were part of problem assignments for the other group of students. An example is, "Why would you have more confidence in a memory span result based on an experiment with 50 subjects than one based on 10 subjects?"

A questionnaire with nine yes/no questions was given to participants after their interactive experience. some of the questions were worded so that "no" was the positive answer. Participants were asked: (1) if the demonstration was interesting, (2) gave insights into concepts, (3) clearly showed the idea of repeated sampling, (4) whether more demonstrations would be desirable, (5) were the graphics helpful, or (6) too complicated, (7) was the time sufficient for the session, and (8) to complete the worksheet, and (9) were the worksheets well-prepared.

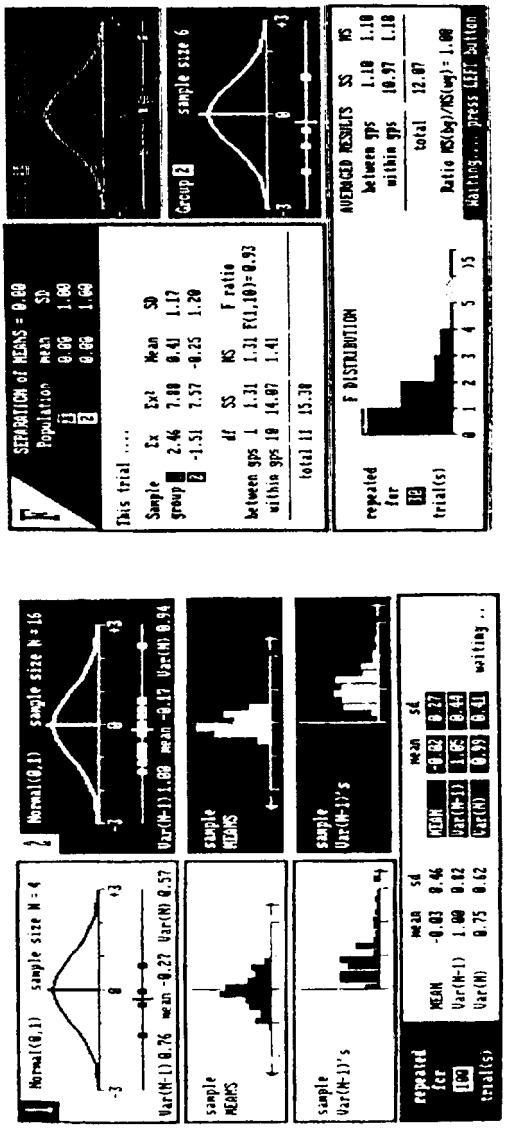
Procedure

The interactive computer session was an hour-long problems class in a statistics course following exposure to the demonstrations in a lecture. Students in the study worked independently using the appropriate worksheets. They were directed to write down some values and confirm calculations in the cumulative windows using hand-held calculators. The attitude questionnaire was administered to students following their interactive experience.

Attainment of sem and F-distribution concepts was assessed on routine course tests. All answers to "open" questions were marked independently by two judges who were blind to the group assignment.

FIGURE 1

EXAMPLES OF FULL SCREEN FROM THE DEMONSTRATIONS



The left panel shows a full screen from the sem demonstration after 100 samples have been made. Two different sets of random samples are cumulated independently: the mean of sample means for the left-hand samples (–0.03) is presented on a white background in the bottom window of the screen; the value for the right-hand samples (–0.02), appears on the dark background. The right panel shows a typical screen from the F distribution demonstration of a two-group experiment where the population means were equal. A histogram of F-values from 30 repeated samples is cumulated in the bottom window alongside the average sum of squares. Students were introduced to the demonstrations one window at a time.

Mean inter-marker reliability on these questions was 0.65 (range 0.47 to 0.80). Scores were averaged over the markers and scaled between 0 and 10, with 10 being best.

RESULTS AND DISCUSSION

The attainment of the groups will be mentioned first followed by the replies to the attitude questionnaires.

Attainment

Table 1a shows that the FDEMO group had the predicted superior performance on questions about the F-distribution relative to the SEMDEMO group. However, the prediction was not confirmed in the performance of the sem questions since both groups had similar scores. A two-way randomised blocks analysis of variance performed on these data did not show the predicted interaction effect ($F(1,37)=1.70$). The only significant effect was that the scores on the F questions were superior to those on the sem questions ($F(1,37)=16.58$, $P<0.01$). Because the questions were not matched for difficulty this may be due to the questions themselves, or some other methodological/subjects variables.

TABLE 1

MEAN SCORES IN THE TWO STUDIES

a. Study 1: Mean Scores (SD) on the sem and F-distribution questions for both SEMDEMO and FDEMO groups.

	Group	
	SEMDEMO	FDEMO
sem Question	6.18 (1.57)	6.29 (1.82)
F	7.06 (1.93)	7.99 (1.46)

b. Study 2: Mean Scores (SD) on the sem and F-distribution questions for the three groups, SEMDEMO, FDEMO and Control.

	Group		
	SEMDEMO	FDEMO	Control
sem Question	7.37 (1.30)	6.44 (1.73)	5.88 (2.46)
F	5.99 (1.36)	6.12 (1.06)	5.14 (1.62)

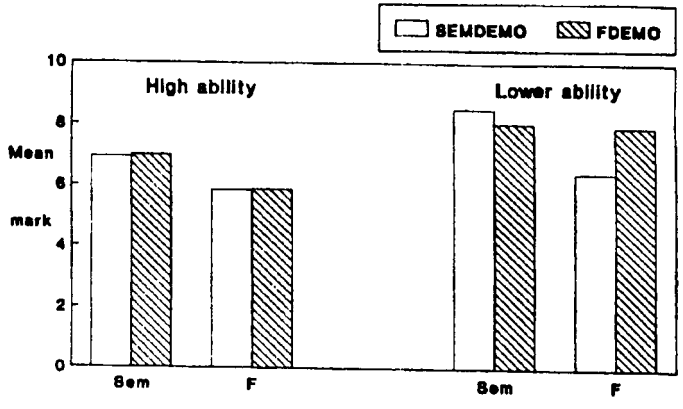
Figure 2a shows the performance of subgroups which have been divided according to their previous statistics grades. Candidates with A grades are in the "high ability" group; those with other grades in the "lower ability" group. The lower ability sets in the FDEMO group showed a greater superiority on the F questions relative to the scores of the lower ability sets of the SEMDEMO group. A three-way analysis of variance with demonstration group, ability level and type of question as factors, revealed significant main effects for ability group (higher sets better, $F(1,35)=7.18$, $P<0.01$) and type of question (as before, $F(1,35)=15.58$, $P<0.01$) but the interaction between demonstration group and question failed to reach significance ($F(1,35)=2.29$).

Since the lower ability groups were likely to gain the most from the interactive session, an analysis of variance was performed excluding the high ability students leaving 14 in SEMDEMO group and 12 in FDEMO group. This revealed significant effects of question type ($F(1,24)=10.24$), $P<0.01$) and a marginally significant interaction ($F(1,24=3.30$, $P=0.08$). The groups performed equally well on the sem questions (SEMDEMO group, 5.86; FDEMO group, 5.87) but the FDEMO group did better on questions about F-distributions relative to the SEMDEMO group (SEMDEMO group, 6.43; FDEMO group, 7.96).

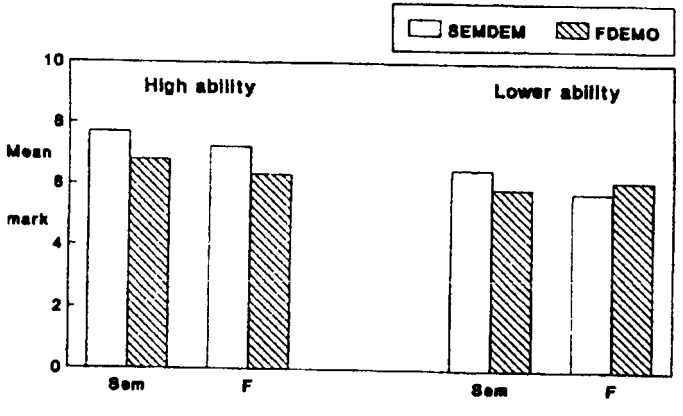
FIGURE 2

ABILITY GROUPS

STUDY 1



STUDY 2



a. (upper panel). The Scores of the SEMDEMO and FDEMO Groups on both the Sem Questions and the F-distribution Questions for the High Ability and the Lower Ability Students in Study 1.

b. (lower panel). The Scores of the SEMDEMO and FDEMO Groups on Both the Sem Questions and the F-distribution Questions for the High Ability and the Lower Ability Students in Study 2.

Attitude questionnaire

The questionnaire results were divided into three categories; liking the demonstration (interesting, wanting more), learning from it (gaining insight, finding the graphics helpful, not too complicated, and the idea clear) and finding the practical aspects helpful (useful worksheet which was not too dense, having sufficient time to understand the concepts). Table 2 shows small unreliable differences between groups on "liking" and "learning" from the interactive session (LIKE, $F(1,37)=1.16$; LEARN, $F(1,37)=3.15$, $P=0.08$). However, a large, reliable difference on the attitudes towards the practical aspects of the session was found where the SEMDEMO group was less satisfied with the arrangements than the FDEMO group ($F(1,37)=11.43$, $P<0.01$). Scrutiny of the individual questions revealed that the SEMDEMO group felt the worksheets were not well prepared (71 per cent) and the graphics were too complicated (66 per cent).

To summarise the outcome of study 1, the predicted improvement in attainment was obtained for the group which interacted with the demonstration of the F-distribution. The prediction was not

confirmed for the sem demonstration. The SEMDEMO group indicated that the graphics and worksheet were too complicated, so in a second study the demonstration was simplified and the worksheet was staged in sections.

TABLE 2
RESULTS OF THE ATTITUDE QUESTIONNAIRE SUMMARISED INTO THREE CATEGORIES: LIKING THE DEMONSTRATIONS, LEARNING FROM THEM, AND FINDING THE PRACTICAL ASPECTS OF THE TASK SATISFACTORY.

	Study 1 Group			Study 2 Group	
	SEMDEMO	FDEMO		SEMDEMO	FDEMO
LIKE	1.6	1.4	LIKE	1.6	1.9
LEARN	1.5	2.0	LEARN	2.4	2.2
PRAC	–1.4	–0.3	PRAC	0.1	–0.2

NOTE: Larger numbers are more positive but no further scaling has been attempted.

STUDY 2

METHOD

The second study was similar to the first except for the modifications in the sem demonstration and its worksheet, and the addition of "no-experience" control group. If the control group performs worse than either demonstration group, generalisation of experience from one demonstration to the other may have occurred, or increased attention to students doing the demonstrations may have improved some aspect of learning. The control group will also provide information about the relative difficulty of the two types of questions for the class and any non-specific advantages of "hands-on" computer experience.

Students

Sixty-five students were divided into three groups, matched for their attainment in first year statistics. The mean marks were 57 per cent, 60 per cent and 60 per cent respectively for the SEMDEMO (N=23), FDEMO (N=21), and Control (N=21) groups.

Apparatus and procedure

The apparatus and procedure was as previously described except the SEMDEMO was modified to present one sample alone. The control group observed a lecturer operate the demonstration in class.

RESULTS AND DISCUSSION

Attainment

Table 1b shows that overall differences between groups were obtained: the SEMDEMO group scored better than the FDEMO group which was better than the Control group ($F(2, 62)=3.22$, $P<0.05$). A designed comparison showed that the Control group scored significantly less than the groups which had interactive experience ($F(1,62)=5.98$, $P<0.01$). This may be explained by a non-specific influence of interactive experience or bias in group selection since five students in the control group were absent from the demonstration session. These absent participants may have been less motivated than the others although the groups were matched on the first year performance. In subsequent hypothesis testing the control group has been omitted to make comparisons with study 1 more direct, interpretation of the data less complex, and reduce the possibility of motivational bias.

The two-way analysis of variance with demonstration group and question as factors yielded a significant interaction ($F(1,42)=4.15$, $P<0.05$). Table 1b shows the predicted superiority of the SEMDEMO group on the sem questions, and a much smaller difference between questions for the FDEMO group. Also significant was the main effect of type of questions ($F(1,42)=10.75$, $P<0.01$). In contrast to the results in study 1, the mean score for the sem questions was higher than for the F

questions. Many factors may have been responsible for the reversal, e.g., the classroom teaching, the average statistics ability of the students, or the scaling of answers. Because the questions and concepts were not matched for difficulty and better performance on sem questions was also seen in the control group, this swing may reflect random variation and is not of special interest.

Figure 2b shows the results stratified by ability levels. The lower ability sets in both demonstration groups show the predicted effects of better performance on the relevant questions. However, for the high ability sets the predicted trend is obtained only for the sem questions. A three-way analysis of variance with demonstration group, ability set and type of question as factors yielded a significant main effect of question type ($F(1,40)=9.50$, $P<0.01$) with no other significant findings.

An analysis of variance using only lower ability students (16 students remained in each demonstration group) revealed a significant interaction between demonstration group and type of question ($F(1,30)=6.61$, $P<0.01$). As predicted, on sem questions the SEMDEMO group had superior performance compared to the FDEMO group and the FDEMO group had the better performance on F questions. The only other significant finding in this analysis was a main effect of type of question which has been described elsewhere ($F(1,30)=6.61$, $P<0.02$).

Attitudes

The respondents in study 2 had similar attitudes towards "liking", "learning" and "practical aspects" of their interactive experience.

In summary, the interactive experience with a Monte Carlo simulation enhanced performance on relevant test questions, especially for the intermediate and low ability students. The opportunity provided by a "hands on" session increased scores on questions about two statistical concepts compared to a control group who only observed a lecturer operate the demonstration in class.

GENERAL DISCUSSION

This research asked if students benefit from interactive computer simulation when learning statistical concepts. For the lower ability students, whose attainment was enhanced by specific experience, the answer is yes provided the demonstrations and worksheets are tailored to the students. Study 1 showed that the FDEMO group's test scores on the F-distribution questions were better than the SEMDEMO group's. Study 2 showed improved performance resulting from both demonstrations, implying that computer interaction enhances learning concepts related to sampling distributions.

High ability students' performance was little affected by the interactive session. This finding differs from Gay (1986) who found high ability students were better able to benefit from learner control than lower ability students. Several reasons may be found for the discrepancy. In our study students did not select their own learning sequence but used a worksheet; additionally each student answered questions on two different concepts, but only had interactive experience with one. The burden of proof therefore rested on the comparison of demonstration relevant questions with demonstration irrelevant questions. Gay's study only examined students on material used in the computer-assisted learning.

Since all our students had passive exposure to each demonstration during lectures, and all were given identical problems during problems classes, the differences obtained can be attributed to the students' active role in controlling the Monte Carlo simulations, which may have encouraged students to encode, remember or structure these concepts so that they were more easily applied to test problems.

Since in study 2 the control group performed less well than the groups doing the demonstrations, the motivation of the non-control students may have been elevated by the interaction, or increased practice may have caused better performance. A more cognitive explanation would be that deeper processing, or more efficient retrieval, occurred for the demonstration students.

These studies have shown that, without increasing staff contact time, attainment was improved on two important concepts in a second year statistics course for 70 students. The findings may also be applicable to teaching other difficult concepts in statistics (see Greer and Semrau, 1984).

ACKNOWLEDGMENTS. — The authors are grateful to Dr Jonckheere for helping with the marking of test questions, and to Alan Greenwood for translating the programmes into Pascal. We are also grateful to the Computers in Teaching Initiaitive for funding this study.

Correspondence and requests for reprints should be addressed to Dr C. G. Weir, Department of Psychology, Colorado College, 14 E. Cache la Poudre, Colorado Springs, Co 80903, USA.

The programs and worksheets controlling these demonstrations are available in BBC BASIC (without use of mouse control) and Prospero PASCAL implementation for RM-Nimbus microcomputers. Please contact the second author.

REFERENCES

BLEASE, D. (1986). *Evaluating Educational Software.* London: Croom Helm.

BROK, A. (1985). *Personal Computers for Education.* Cambridge: Harper and Row.

BOWER, G. H., CLARK, M. C., LESGOLD, A. M., and WINZENZ, D. (1969). Hierarchical retrieval schemes in recall of categorical word lists. *J. verb. Learn. verb. Behav.*, 8, 323-343.

CLAYDEN, G. S., and WILSON, B. (1988). Computer-assisted learning in medical education. *Med. Educ.*, 22, 456-467.

CRAIK, F. I. M., and LOCKHART, B. (1972). Levels of processing: a framework for memory research. *J. verb. Learn. verb. Behav.*, 11, 671-84.

CRAIK, F. I. M., and TULVING, E. (1975). Depth of processing and the retention of words in episodic memory. *Exp. Psychol. gen.*, 104, 268-294.

GARDNER, N. (1988). No more than a tool. *Times Higher Educ. Suppl.*, 17th June, vi.

GAY, G. (1986). Interaction of learning control and prior understanding in computer-assisted video instruction *J. educ. Psychol.*, 78, 225-227.

GREER, B., and SEMRAU, G. (1984). Investigating psychology students' conceptual problems in mathematics in relation to learning statistics. *Bull. Br. Psychol. Soc.*, 37, 123-125.

HAMMOND, N., and ALLINSON, L. (1988). Development and evaluation of a CAL system for non-formal domains: the hitch-hiker's guide to congition. *Comput. Educ.*, 12, 215-220.

HARRISON, A. J. L., BANWELL, J. K., FROST, M., and PLUMBRIDGE, W. J. (1988). An interactive teaching experiment in materials science. *Comput. Educ.*, 12, 119-124.

INGHAM, D. (1988). Medical education computing. *The Computers in Teaching Initiative Support Service File*, No. 7, 7-16.

JOHNSTON, V. M. (1987). The evaluation of microcomputer programs: an area of debate. *J. Comput. assist. Learn.*, 3, 40-50.

KEMMIS, S., ATKIN, R., AND WRIGHT, E. (1987). The evaluation of student learning. In SCANLON, E., and O'SHEA, T., (Eds.), *Educational Computing*. Chichester: Wiley.

LIEBLUM, M. D. (1988). A model describing CAL authoring systems applied to Taiga. *Comput. Educ.*, 12, 141-149.

TVERSKY, A., and KAHNEMANN, D. (1974). Judgement under uncertainty: heuristics and biases. *Science*, 185, 1124-1131.

WONG, D. (1987). Teaching A-level physics through microcomputer dynamic modelling: II. evaluation of teaching. *J. Comput. assist. Learn.*, 3, 164-175.

(*Manuscript received 1st November*, 1990)