COGNITIVE SCIENCE **13**, 551–568 (1989)

# Studies of Diagnosis and Remediation with High School Algebra Students

D. SLEEMAN
A.E. KELLY
R. MARTINAK
R.D. WARD
J.L. MOORE

*University of Aberdeen*
*Aberdeen Scotland*

Despite extensive discussion in the literature about the diagnosis and subsequent remediation of students' errors, few studies have compared the effects of different *styles* of error-based remediation. Swan (1983) found that a conflict approach (pointing out errors made by students and demonstrating their consequences) was more effective than simple reteaching, but Bunderson & Olsen (1983) found no difference between error-specific remediation and reteaching. More studies are needed in order to understand the factors which lead to successful remediation. The three studies discussed in this article compared error-specific or model-based remediation (MBR) with reteaching in algebra. MBR bases its remediation on the model inferred for an individual student before reteaching the correct procedure. Reteaching simply shows students the correct procedure without addressing specific errors. The results show that MBR and reteaching are both more effective than no tutoring; however, MBR is *not* clearly more effective than reteaching. The results are discussed in terms of stability of errors, their relevance to educational practice and to intelligent tutoring systems (ITS). Although the studies were carried out using human tutors, the results suggest that for the purpose of remediation in the algebra domain, when taught procedurally, "classical" computer-assisted instruction (CAI) would be as effective as an ITS.

## 1. INTRODUCTION

Intelligent tutoring systems (ITS) researchers have generally made two basic assumptions. First, much ITS research arose out of the belief that to be most effective, instruction and remediation need to be highly adaptive.

This research was sponsored by ARI/ONR under contract number MDA-903-84-k-0279. Professor R.D. Hess (Stanford) and Professor K. Lovell (consultant, Dundee) provided helpful comments during the course of the project, and upon earlier versions of the material contained in this article.

Correspondence and requests for reprints should be sent to D. Sleeman, Dept. of Computing Science, University of Aberdeen, Aberdeen, AB 92 UB, Scotland.

"Classical" computer-assisted instruction (CAI) systems were considered to be intrinsically limited because of their inability to provide highly adaptive instruction (Hartley & Sleeman, 1973; Sleeman & Brown, 1982). Second, in developing ITS to remediate students' errors on the basis of a detailed prior diagnosis of those errors, it has often been assumed that diagnosis is more difficult than remediation; that is, having once inferred an accurate model of a student's error, it is then relatively straightforward to use that model to *direct* a remedial dialogue.

This article presents an empirical examination of these assumptions. Remediation based upon a student model will be referred to as model-based remediation (MBR). MBR provides procedurally orientated remediation of specific errors found in a student's solutions before reteaching a correct strategy. Its counterpart, reteaching, simply reteaches the correct method.

The studies reported here were carried out as part of the research associated with the PIXIE system. PIXIE is a data-driven ITS shell (Sleeman, 1987) that attempts to diagnose and then remediate student errors within a particular domain of knowledge. At present the knowledge base for linear algebra is the most complete.

## 2. REVIEW OF THE LITERATURE ON ALGEBRA LEARNING

Algebra errors were investigated quite extensively by Buckingham (1933) in order to develop teaching materials for drilling students in tasks likely to evoke frequent or persistent errors, and also to develop diagnostic tests for identifying individual errors requiring specific remediation.

More recent reviews imply that error-specific remediation is *known* to be superior to reteaching. Brown & Burton (1978) forsee a role for diagnostic specialists working with children having difficulty in mathematics. The specialist would conduct in-depth interviews in conjunction with specific computational tasks to detect possible procedural errors. Instruction could then be matched to each child's specific difficulties. Resnick (1984) makes a similar point but believes that greater benefit may be gained by diagnosing global misunderstandings rather than remediating each error. Macnab & Cummine (1986, p. 125), applying the work of Case (1975, 1978) to the algebra domain, also discuss the importance of showing the student both the correct procedure and pointing out the incorrect steps:

> ....demonstrating that there is a flaw in a pupil's method that can be useful in situations where the pupil is aware of a correct procedure but prefers his alternative either because he thought it up himself, or it seems easier, or for some other reason. In such cases the unsound nature of the pupil's own method may have to be demonstrated before he will adopt a correct method.

However, experimental data concerning the *effectiveness* of error-specific remediation is ambivalent. Swan (1983) concluded that error-specific remedi-

ation is more effective than reteaching over a period of 8 one-hour lessons in which there was extensive class discussion of students' errors. On the other hand, Bunderson & Olsen (1983) suggest that pointing out an incorrect procedure is *not* more effective than merely reteaching. In remediating subtraction errors, both error-specific and general remediation led to considerable improvements. This casts doubt upon the need to precede remediation with detailed diagnosis.

Recent PIXIE studies have also cast doubt upon the need for detailed diagnosis (Kelly & Sleeman, 1986; Martinak, Schneider, & Sleeman, 1987; Putnam, 1987). It was found that teachers generally do not adopt the role of a diagnostician, even when in a tutorial situation. Diagnosing errors tends not to be the teacher's primary goal. Putnam suggested that teachers in their remediation followed a "curriculum-based script." However, no comparison of the effectiveness of tutoring based on a curriculum-based script as opposed to error-specific remediation was made. In light of the conflicting data, a systematic comparison of error-specific remediation (or MBR) with script-based reteaching is needed.

## 3. A SERIES OF FORMAL EXPERIMENTS TO COMPARE MBR WITH RETEACHING

In an initial study using PIXIE as the tutor for the task of solving single algebra equations containing one unknown (Appendix A), no differences were found between the MBR and Reteaching conditions (Martinak et al., 1987). However, as only 30% of students' errors were diagnosed, the two conditions provided the same treatment for 70% of the errors as the system simply retaught a task when unable to make a diagnosis. In addition, students made few errors, an overall mean of 3.64 errors in 17 tasks. Given the small number of errors and the low rate of diagnosis, it is not surprising that no significant difference between the two groups was noted. This study did however lead the authors to believe that the issues of diagnosis and remediation were much more subtle than initially suspected, therefore it was decided to replicate the study using *human* tutors. This became the first of three studies.

## 4. A STUDY TO COMPARE MBR AND RETEACHING BY HUMAN TUTORS

The first experiment was carried out in Scotland in Autumn 1986 at School L. It was hypothesised that MBR would be superior to Reteaching.

### 4.1 Materials

Materials were a 20-item pretest, a 20-item posttest, a MBR tutoring script, a Reteaching tutoring script, and a list of tasks to be worked during tutor-

ing. The pretest and posttest were matched, item by item, for form and difficulty (Appendix A).

The tutoring scripts (Appendix B) were based on PIXIE's approach to remediation. The MBR script directs the tutor to point out to the student each error made, and to explain it before the correct procedure is given. The Reteaching script merely directs the tutor to reteach the procedure.

The tasks to be worked during tutoring consisted of 20 sets of three items, each set containing items of equivalent form and difficulty to those of the pretest and posttest. The first item of each set was taken from the pretest.

### 4.2 Subjects

Students from two 2nd- and two 3rd-year mathematics classes in School L participated in the study. Average ages of the students were 13 years 4 months, and 14 years 6 months, respectively. On the basis of the pretest, 44 students who had scored below 80% were identified as requiring tutoring.

These subjects (as with the subjects of the two following studies reported in this article) had received algebra instruction that was largely procedural; that is, algebra was treated as a series of transformations without extensive reference to possible meaning. Subjects received approximately 90 minutes per week of algebra instruction starting at the beginning of Year 1. At the time of this, and the two following studies, subjects had received their full complement of initial teaching in the task types listed in Appendix A.

### 4.3 Procedure

In the week after the pretest, 44 students were randomly assigned to MBR or Reteaching. Each student was individually tutored for approximately 35 minutes. All tutoring sessions were audiotaped. Tutoring consisted of having the student first rework an item marked as incorrect on the pretest. If the item was again worked incorrectly, the student received remediation appropriate to the condition and worked at most two further practice items of the same format. This procedure was repeated for each item scored as incorrect on the pretest. In the Reteach condition, in order to discourage students from making their own comparisons, all earlier work was removed before teaching the correct procedure.

The week after tutoring, a group posttest was given to the students during a normal mathematics class. A delayed posttest, identical to the immediate posttest, was given approximately two months after the first posttest.

### 4.4 Results

Analyses of the data are based upon scores from 38 of these students (6 were absent from school for either the tutoring or the posttest). Posttest scores were taken as a measurement of the effectiveness of tutoring. Table 1 presents the mean scores and standard deviations by condition.

TABLE 1
Mean Scores* by Condition

| Condition | Pretest | Posttest 1 | Posttest 2 |
|---|---|---|---|
| MBR | 12.53 | 14.32 | 12.24 |
| (N=19) | (3.86) | (3.37) | (5.62) |
| Reteaching | 11.63 | 14.42 | 12.76 |
| (N=19) | (3.10) | (3.64) | (4.44) |

* Maximum=20. Standard deviations in parentheses.

Although there was a significant overall mean difference between the pre- and posttests for both groups $t(37) = 4.20, p < .001$, there was so significant difference by condition: $t(36) = .09, p > .92$. The overall mean scores for the delayed posttest (two months after the first posttest) had reverted to the pretest levels[1] and were also not significant by condition $t(32) = .30, p > .70$.

Further analyses of the data using only students who scored low marks on the pretest (i.e., those scoring 13 or less) showed no significant differences by condition, $^{H}(19) = 0.43, p > .66$. Nor were there any differences to be found when students' errors were reclassified into several levels of severity using criteria developed in Sleeman (1983).

### 4.5 Discussion

This experiment confirmed the results of the previous computer study, and showed that MBR and Reteaching are comparable, even with low-scoring students. This was interpreted here to imply that the form of MBR used was not effectively communicating with the student. The next study attempted to make MBR more effective.

## 5. A STUDY TO COMPARE VARIANTS OF MBR AND RETEACHING

A study was conducted to investigate two factors that may have influenced the results from the first study: namely, presumed lack of cognitive dissonance or cognitive engagement on the part of the students. It was hypothesized that no differences between the conditions in the previous study had arisen because:

a) The students had no reason to accept the tutor's method of doing algebra as better than their own. Macnab & Cummine (1986) discuss the importance of demonstrating to the student the unsound nature of the pupil's incorrect method in order to create "cognitive dissonance" (CD). The

[1] Learning did not persist into the delayed posttest, but as Bunderson & Olsen (1983) point out, it is well documented that skills will decay if not under constant review. In the next study, learning did persist into the delayed posttest.

present study attempted to instill CD by having students check an answer by substituting it back into the equation to see if both sides of the equation balanced.

b) The students were not sufficiently involved with their learning (i.e., they were passive listeners to the tutor's instructions). By having students verbally repeat the correct procedure back to the tutor, it was hoped to engage them more in their learning.

### 5.1 Subjects and Procedure

Students from two 2nd- and one 3rd-year mathematics classes (average age 13 years 6 months, and 14 years 8 months, respectively) from Scottish School P took the 20-item pretest in algebra equation solving. Anyone scoring 80% or better on the pretest was not seen for tutoring. This left 48 students who were randomly assigned to four conditions:

a) MBR
b) Reteaching
c) MBR with cognitive dissonance (MBR + CD)
d) MBR with cognitive engagement (MBR + CE)

The MBR and Reteaching conditions in this study were designed so that the previous study could be replicated. Given the limited number of students, no pretest/posttest-only, Reteach + CD, or Reteach + CE conditions could be included. Appendix B shows the scripts used in each of the four conditions. The procedure was otherwise identical to the procedure of Study 1.

### 5.2 Results

Table 2 shows the mean scores in the pretest, posttest, and delayed posttest for each condition. Both quantitative and qualitative analyses were performed on the data. Analysis of variance indicated no significant differences among conditions on the pretest: $F(3,44) = .626$, $p > .50$.

The overall mean for the posttest score was significantly higher than the overall mean for the pretest ($t(43) = 5.83$, $p < .001$) showing a general pre-to posttest gain. An ANOVA on posttest scores showed no significant differences between the conditions: $F(3,40) = .797$, $p > .50$. An ANOVA on the delayed posttest scores also showed no differences between conditions: $F(3,37) = 11.81$, $p > .10$.

Because the above analyses showed no differences among the groups, errors were reclassified as algebraic or nonalgebraic (see Appendix C for examples). The mean number of algebraic errors for each condition on the first posttest were: Reteach: 2.00; MBR: 2.09; MBR + CD: 3.00; MBR + CE: 2.73, indicating no major differences among the conditions.

Following these results, it was hypothesized that a significant number of student errors might be unstable. This was investigated by tracing a given

TABLE 2
Pretest, Posttest, and Delayed Posttest Mean Scores* by Condition

| Condition | Mean Pretest Scores | Mean Posttest Scores | Delayed Posttest Scores |
|---|---|---|---|
| Reteaching | 12.08 | 15.20 | 15.64 |
| (*N*=12) | (1.98) | (2.92) | (3.07) |
| MBR | 11.91 | 15.80 | 14.73 |
| (*N*=12) | (3.33) | (3.68) | (2.05) |
| MBR+CD | 12.60 | 14.00 | 14.80 |
| (*N*=12) | (2.32) | (2.63) | (1.93) |
| MBR+CE | 12.91 | 15.90 | 13.00 |
| (*N*=12) | (2.81) | (3.33) | (1.80) |

* Maximum score=20. Standard deviations in parentheses.

error for a given task from pretest, to tutoring, to posttest. (In retrospect, it is believed that this is a stringent stability criterion.) This analysis showed that only approximately 20% of the errors from the pretest occurred in the same items during tutoring.

However, intermediate tutoring may have lowered this stability measure since some errors may have been corrected as a result of tutoring on previous tasks. If stable errors are reclassified to include those tutored in an earlier part of the session, the average percentage of stable errors during tutoring increases to 26%. Retrospective analysis of the data from the first study produced similar findings.

### 5.3 Discussion

Three analyses were applied to this study, each of which was logically driven by the earlier ones. The first two analyses did not find differential effects among conditions, in spite of the modifications to the basic MBR treatment. (It should be noted, however, that the MBR + CD condition could *not* be properly tested with this sample, because the process of substitution and verification was new to these students.)

The third analysis suggested that the phenomena of student errors is more complex than had been anticipated and pointed strongly to the instability of a high proportion of student errors. No further conclusions were drawn from this experiment because it was not designed to investigate stability; stability became the focus of the third study.

## 6. MBR AND RETEACHING IN THE CONTEXT OF STABLE ERRORS

Given the apparent instability of errors suggested by the analyses of the two previous studies, this study was carried out to investigate error stability in

TABLE 3
Mean Pretest and Posttest Scores*by Condition

| Condition | Pretest 1 | Pretest 2 | Posttest |
|---|---|---|---|
| MBR | 27.50 | 29.70 | 41.22 |
| (N=8) | (12.79) | (14.04) | (9.09) |
| Reteach | 28.00 | 33.56 | 41.62 |
| (N=8) | (14.49) | (11.06) | (4.47) |
| Control | 23.44 | 25.44 | 26.00 |
| (N=8) | (12.35) | (13.44) | (14.36) |

* Maximum=51. Standard deviations in parentheses.

depth, and to compare the effects of MBR and Reteaching in the context of stable errors.

### 6.1 Materials

A 51-item stability measure consisting of 17 sets of three algebra tasks was developed; each task within a set was generated by the same template. The template for the first set is aX=b; for set 17 it is aX=b*c(dX+e). Using these templates, a pretest and posttest were constructed, with the requirement that the first item of each set on both tests would be identical.

### 6.2 Subjects and Procedure

Ninety-six students in the 2nd- and 3rd-year classes at School L were pretested; 21 of these were not considered to need tutoring because they had at least 88% of the items (45 of the 51) correct; a further 23 students were excluded because they were absent from school during the week of tutoring; another 15 students were present for only one of the two pretests; 37 students remained. Of these, 27 had at least one stable error. These 27 students were than randomly assigned to the conditions MBR, Reteaching, and Control. In this study a stable error was defined as one made by the same student at least twice in both pretests, not necessarily on the same task.

Students in the treatment conditions were seen individually for a 50-minute period. To put the student at ease, each student was asked to work the first six items from the pretest. After the first six items had been worked, stable errors were tutored. An error was only tutored if it occurred again when the student worked the same task during the tutoring session. After all "stable" errors had been tutored, unstable errors were tutored. Students in the Control group simply took the pretests and posttest. The week after tutoring, a posttest was given to the classes involved.

### 6.3 Results

Of the 27 students, 3 were absent from school, 1 from each condition, when the posttest was given. Table 3 gives the means and standard deviation by condition for the remaining 24 students. Significant omnibus $F$ ratios were

TABLE 4
Analysis of the 19 Stable Errors Occurring in Both Pretests

| Error Type | *N* of students who had a particular stable error | Prevalence of each error (as % of subjects, *N*=24) | *N* occurring in both pretests | Frequency of each error (as % of all errors) |
|---|---|---|---|---|
| 1. Bracket | 14 | 58 | 146 | 10.3 |
| 2. Precedence | 13 | 54 | 96 | 6.8 |
| 3. Computational | 7 | 29 | 76 | 5.3 |
| 4. Change side not sign of x-term | 5 | 21 | 68 | 4.8 |
| 5. Subtract a coeff | 4 | 17 | 188 | 13.2 |
| 6. Add coeff and constant | 3 | 13 | 55 | 3.9 |
| 7. Add a negative sign | 3 | 13 | 17 | 1.2 |
| 8. Uses a number twice | 3 | 13 | 51 | 3.6 |
| 9. Drop an X | 2 | 8 | 29 | 2.0 |
| 10. Inverted division | 2 | 8 | 123 | 8.7 |
| 11. Minus before the wrong number | 2 | 8 | 48 | 3.4 |
| 12. Subtract a multiplier | 2 | 8 | 11 | 0.8 |
| 13. Subtract a multiplied x-term | 2 | 8 | 22 | 1.5 |
| 14. $ax=b \Rightarrow x=a-b$ | 2 | 8 | 16 | 1.1 |
| 15. Drop a negative sign | 1 | 4 | 16 | 1.1 |
| 16. Multiply across by a coeff | 1 | 4 | 6 | 0.4 |
| 17. Ends task for $ax=bx$ | 1 | 4 | 12 | 0.8 |
| 18. $ax=b \Rightarrow x=-(a+b)$ | 1 | 4 | 6 | 0.4 |
| 19. $a*bx+cx=d \Rightarrow a*(-d)=-bx-cx$ | 1 | 4 | 6 | 0.4 |
| Subtotals | | | 992 | 69.7 |
| Error types 1–9 in students for whom these errors were not stable | | | 277 | 19.5 |
| Subtotals | | | 1269 | 89.2 |
| Unstable errors (i.e., errors of types 20–46) | | | 153 | 10.8 |
| Subtotals | | | 1422 | 100.0 |

found for the posttest: $F(2,22)=6.33$, $p<.008$. Post hoc analyses using the Scheffe test showed no differences ($p>.05$) between MBR and Reteaching, but both being better ($p<.05$) than the control condition on the posttest. Although the cell sizes were small, this result again bears out the findings of the first two studies.

### 6.4 Analyses of Stability

The total set of errors encountered in this experiment was classified into 46 different types. Only 19 of these (Table 4) were stable (i.e., occurred at least twice in both pretests for at least one student). The other 27 types occurred either infrequently in the pretests (i.e., were not stable) or only in the posttest.

In total, 1422 errors occurred in the two pretests. Of these, 1269 (89%) were of the 19 types. This figure however includes errors which were not

TABLE 5
Numbers of Stable Errors per Student

| Number of stable errors | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| Number of students ($N$=24) | 4 | 11 | 2 | 3 | 1 | 1 | 2 |

stable for every student who made them; that is, some students made errors of type 1–19 only once in one or both pretests. Excluding these and counting only stable errors, 992 (70%) of the 1422 errors were of the 19 types.

Table 4 shows the frequency and prevalence of each of the 19 stable error types. Frequency refers to the proportion of the total errors made. Prevalence indicates the proportion of students who made a particular error. These percentages refer only to the students who scored less than 88% in the pretests and had at least one stable error, not to the whole student population.

It can be seen that the relationship between frequency and prevalence is not necessarily linear. For example, precedence errors had high prevalence but a relatively lower frequency, whereas inverted division errors had high frequency with low prevalence. In other words, not so many students made inverted division errors, but those who did, made them often.

The number of stable errors made by individual students ranged from 0 to 7. Table 5 shows that students who had stable errors had a mode of 2 stable errors each.

### 6.5 Remediating Stable Errors by Condition

Because not all the stable errors made during the pretests were made again by the same students during the tutoring sessions, not all students' "stable" errors were tutored. Also, obviously none of the stable errors made by the control group were tutored. Table 6 therefore shows an analysis of only the tutored stable errors for the MBR and Reteach groups, together with the control group data for the same error types. Table 6 shows the numbers of students making particular errors (i.e., the prevalence data) and the numbers of errors made (i.e., the frequency data).

These data replicate the findings of the total error data from this and the previous two studies: Both the MBR and Reteach conditions were effective in remediating errors, but there was no discernible difference between them. This finding appears to hold even when the analysis is confined to stable errors alone. This result is consistent with the result reported by Sleeman (1983) for a study involving two groups: essentially MBR and Control.

The remedial effectiveness of a condition might be expressed as the percentage reduction of errors. This is also shown in Table 6. However, note that not all stable errors occurred in each condition, so the groups are only partially "matched." Also note that the frequency data gives a different measure of effectiveness to the prevalence data. The prevalence figure (the total number of student errors remediated), is probably preferable because

TABLE 6
Tutored Stable Errors: Numbers of Students Making Particular Errors*

| ERROR TYPE | MBR | | RETEACH | | CONTROL | |
|---|---|---|---|---|---|---|
| | Pretest | Posttest | Pretest | Posttest | Pretest | Posttest |
| 1. Bracket | 4(17) | 4( 8) | 4(19) | 3( 9) | 4(30) | 4(26) |
| 2. Precedence | 3(11) | 2( 7) | 4(14) | 2( 6) | 5(20) | 5(19) |
| 3. Computational | — | — | 2( 6) | 2( 3) | 4(30) | 3(17) |
| 4. Change side not sign of x-term | 1( 5) | 0 | 2(20) | 1(11) | 1( 3) | 1( 4) |
| 5. Subtract coeff | 1(34) | 0 | 1(35) | 0 | 2(25) | 2(28) |
| 7. Add neg sign | 1( 3) | 0 | 1( 3) | 1( 2) | 1( 3) | 1( 5) |
| 10. Inv Division | 1(29) | 0 | — | — | — | — |
| 11. Minus before wrong number | — | — | 1(12) | 0 | 1(12) | 1(11) |
| 12. Subtract a Multiplier | — | — | 1( 3) | 1( 3) | — | — |
| Totals: | 11(99) | 6(15) | 16(112) | 10(34) | 18(123) | 17(110) |
| % Errors Reduced: | | | | | | |
| student errors | | 45% | | 38% | | 6% |
| total errors | | (85%) | | (70%) | | (11%) |

* Numbers of errors made by students in parentheses.

frequency data is biased by the greater number of opportunities to make some errors than others. For example, a student who subtracts coefficients will have many more opportunities to make this error than a student who makes bracket errors, which can only be made in tasks that contain brackets.

## 6.6 Discussion

*6.6.1 Effects of Different Forms of Remediation.* Once again, Table 3 shows that both MBR and Reteaching are better than no treatment, but with no distinction between the two. Table 6 suggests that even when the analysis is confined to *stable* errors alone, the result appears to be the same. The authors still find this surprising, and offer the following speculations:

1. Perhaps the MBR and Reteaching treatments used in this study were still too similar. The tutoring scripts may not have differed sufficiently. (In particular, problems with the MBR + CD condition of the second study were noted because the students did not have an appropriate skill.)
2. A related possibility may have been the *length* of the exposure. Although 50 minutes of tutorial seemed a substantial period, the positive effects reported by Swan (1983) followed 8 one-hour lessons.
3. Even though the MBR and Reteaching groups were highly comparable on procedural tasks, had they been tested for conceptual understanding of algebra (rather than procedural understanding) it *might* have been

found that the MBR students would have outperformed the Reteaching students. The next point raises this issue more generally, with respect to initial teaching.

4. As noted in the introduction, MBR essentially *assumes* that the student has a (stable) mental model, to which remedial comments are related. So an additional hypothesis created here is that students taught algebra procedurally have a (weak) mental model of the domain. Further, it has been hypothesized that students taught algebra *conceptually* should have a much stronger mental mode, and so one should observe with such students significant differences between the MBR and Reteaching treatments. Unfortunately, a local high school where algebra was (largely) taught conceptually could not be found, so this hypothesis remains untested.
5. A further speculation is that the MBR and Reteaching treatments are *very* similar because each student in the Reteaching treatment essentially notes for himself the difference between his answer, and that provided by the correct procedure, and so generates his own MBR. Experiments are being planned to probe this issue. Perhaps this mechanism provides an explanation for differences between (mathematically) able and less able students.

*6.6.2 Stability of Errors.* The assumption that student errors are completely stable has been clearly questioned by several studies, including Van Lehn (1981), Bricken (1987), as well as Sleeman (1983).[2] The following issues on error stability are raised by the results of the third study:

1. Attentional nature of some errors. Table 6 shows that only 9 of the 19 stable errors recurred during tutoring. Some of the remaining 10 errors may have been dispelled by tutorial contact alone, possibly because of the motivational effects of having a tutor work with a student.
2. Prevalence and frequency have been introduced here, and it is important to stress the two measures used to describe this phenomenon. *Prevalence* indicates the number of students in the population who have a particular error. *Frequency* indicates the proportion of the total number of errors

[2] Van Lehn (1981) investigated short-term and long-term stability of subtraction errors, and found that only 12% of the students who had errors on the first test had the same errors on the second (short-term) test. He also found that the long-term stability data are very similar to the short-term stability data. VanLehn concluded that errors in general are not stable. Further, Bricken (1987) investigated the stability of algebra errors, and found that 50% (11 of 22 students) committed at least one error on the pretest which reoccurred during their interview held two weeks later. Although the studies mentioned above all confirm that errors generally are not stable, it is important to note that each investigation measures stability using different criteria. However, lack of stability seems to be an issue regardless of how it is measured.

which are explained by the specific error. Both measures are needed to discuss the phenomenon of errors. A class teacher will essentially be concerned to know the most prevalent errors in the class, whereas a tutor will wish to have access to the actual error profile and the most frequent errors made by each student.

3. Taxonomy of errors: This study has confirmed and slightly extended the error taxonomy suggested by Sleeman (1983). From the standpoint of this study, it is feasible to talk about the following types of errors:
   - Stable errors (both remediable and "resistant");
   - Attentional errors (largely minor errors such as adding/dropping signs);
   - Classes of mal-rules, used by the same student on different occasions with the same type of task;
   - Mental slips, typing/transcription errors.

It is important for a tutor to categorize the error correctly since, for example, it may be counterproductive to tutor a student on an error which is a result of a slip; whereas it may be important to address a stable error. (These judgements are both subtle and complex.)

## 7. OVERALL CONCLUSIONS OF THE THREE STUDIES

For a more extensive discussion of the conclusions of these and related studies, including a discussion of possible additional issues to be investigated, see Sleeman, Kelly, Martinak, Ward, & Moore (1987).

- The three studies discussed in this article suggest that when initial instruction and remediation are primarily rule-based and procedural, remedial reteaching appears to be as effective as MBR. From this it follows that "classical" CAI would be as effective as an ITS. It is *vital* to investigate the range of subjects, instructional approaches, and student age-ranges for which this result holds. For example, if either or both instruction and remediation had been conceptually based, then the results might have been different.
- Despite the conclusions of the last paragraph, one should *not* conclude that reteaching by a classroom teacher will be equivalent to reteaching by a computer system. In any tutorial interaction, the nature of interaction and the nature of instruction are what count. Immediacy of feedback, which is generally lacking from the classroom situation, may well be a critical factor (Lewis & Anderson, 1985).
- It is critical that ITS receive extensive field testing.
- The subfield of ITS should *not* conclude that the task of building an ITS is impossible, but it should conclude that the task is harder than the

field initially thought[3]. Winograd & Flores (1986) make a similar point about the natural language area. Some of the initial claims made for ITS were perhaps facile and not well warranted. The "target" knowledge structures are generally far more subtle and complex than the simple models used in most current ITSs. It is possible that a more global analysis by the system (Moore & Sleeman, 1988), taking account of the student's performance at several levels, might lead to better remediation. But before such a system is built it is suggested that experiments are run to see whether human tutors are more effective when they base their remediation on such a global approach.

■ Original Submission Date: August 22, 1988

## REFERENCES

Bricken, W. M. (1987). *Analyzing errors in elementary mathematics.* Unpublished doctoral dissertation, School of Education, Stanford University.

Brown, J.S., & Burton, R.R. (1978). Diagnostic models for procedural bugs in basic mathematical skills. *Cognitive Science, 2,* 155–192.

Buckingham, G.E. (1933). *Diagnostic and remedial teaching in first-year algebra.* (School of Education Series, No. 11). Evanston, IL: Northwestern University.

Bunderson, V.C., & Olsen, J.B. (1983). *Mental errors in arithmetic skills: Their diagnosis in precollege students.* (Final project report, NSF SED 80–12500). Provo, UT: WICAT Education Institution.

Case, R. (1975). Gearing the demands of instruction to the developmental capacities of the learner. *Rev. Ed. Research, 45,* 59–87.

Case, R. (1978). A developmentally based theory and technology of instruction. (*Rev. Ed.*) *Research, 48,* 439–463.

Cronbach, L.J., & Snow, R.E. (1977). *Aptitudes and instructional methods.* New York: Irvington.

Hartley, J.R., & Sleeman, D.H. (1973). Towards intelligent teaching systems. *International Journal of Man-Machine Studies, 5,* 215–236.

Kelly, A.E., & Sleeman, D. (1986). *A study of diagnostic and remedial techniques used by master algebra teachers* (Tech. Rep. No. AUCS/TR8708). Aberdeen, Scotland: University of Aberdeen, Department of Computing Science.

Lewis, M.W., & Anderson, J.R. (1985). Discrimination of operator schemata in problem solving: Learning from examples. *Cognitive Psychology, 17,* 26–65.

Macnab, D.S., & Cummine, J.A. (1986). *Teaching mathematics 11–16: A difficulty-centerd approach.* London: Basil Blackwell.

Martinak, R., Schneider, B., & Sleeman, D. (1987). A comparative analysis of approaches for correcting algebra errors via an intelligent tutoring system. Proceedings of AERA, Washington, DC.

Moore, J.L., & Sleeman, D. (1988). *International Journal of Man-Machine Studies,* 605–623.

[3] Cronbach and Snow (1977) warned that producing truly individualized instruction was a demanding task, but ITS workers chose to infer their conclusions were unduly pessimistic given the complexity of analyzing diverse sets of experimental studies.

Putnam, R.T. (1987). Structuring and adjusting content for students: A study of live and simulated tutoring of addition. *American Educational Research Journal, 24,* 13–48.

Resnick, L. (1984). *Beyond error analysis: The role of understanding in elementary school arithmetic.* Pittsburgh, PA: University of Pittsburgh, Learning Research and Development Center.

Sleeman, D. (1983). Basic algebra revisited: A study with 14-year-olds (HPP Rep. No. 83-9). Stanford, CA: Stanford University, Computer Science Department. (Republished in *International Journal Man-Machine Studies, 22,* 127–149.)

Sleeman, D. (1987). PIXIE: A shell for developing intelligent tutoring systems. In R.W. Lawler & M. Yazdani (Eds.), *Artificial intelligence and education: Vol. 1* (pp. 239–265). Norwood, NJ: Ablex.

Sleeman, D., & Brown, J.S. (1982). *Intelligent tutoring systems.* London: Academic Press.

Sleeman, D., Kelly, A.E., Martinak, R., Ward, R.D., & Moore, J. (1987). *Diagnosis and remediation in the context of intelligent tutoring systems* (Tech. Rep. No. AUCS/TR8712). Aberdeen, Scotland: University of Aberdeen, Department of Computing Science.

Swan, M.B. (1983). *Teaching decimal place value. A comparative study of conflict and positively-only approaches* (Research Rep. No. 31). Nottingham, England: University of Nottingham, Sheel Center for Mathematical Education.

VanLehn, K. (1981). *Bugs are not enough: Empirical studies of bugs, impasses and repairs in procedural skills.* (Cognitive and Instructional Science Series, CIS-11 (SSL-81-2). Palo Alto, CA: XEROX, Palo Alto Research Center.

Winograd, T., & Flores, F.F. (1986). *Understanding computers and cognition: A new foundation for design.* Norwood, NJ: Ablex.

## APPENDIX A

### Pretest and Posttest items in the First Two Studies

| | |
|---|---|
| 1. $7x = 14$ | 1. $3x = 9$ |
| 2. $4x = -12$ | 2. $4x = -8$ |
| 3. $5x = 7$ | 3. $3x = 5$ |
| 4. $8x = 18$ | 4. $6x = 9$ |
| 5. $3x = 8 + 4$ | 5. $5x = 8 + 2$ |
| 6. $3x = 4*4$ | 6. $3x = 4*3$ |
| 7. $3x + 4x = 14$ | 7. $2x + 3x = 10$ |
| 8. $3x + 5 = 26$ | 8. $2x + 3 + 9$ |
| 9. $7 + 4x = 19$ | 9. $2 + 4x = 14$ |
| 10. $5x = 4x + 8$ | 10. $4x = 3x + 6$ |
| 11. $3x = 3(2 + 3)$ | 11. $2x = 3(3 + 1)$ |
| 12. $12x = 2(3x + 3)$ | 12. $24x = 3(2x + 3)$ |
| 13. $7x = 2 + 4*8$ | 13. $5x = 2 + 3*8$ |
| 14. $17x = 19x + 25$ | 14. $16x = 19x + 20$ |
| 15. $3 + 2x + 4x = 21$ | 15. $2 + 2x + 3x = 17$ |
| 16. $4 + 3x + 4x = 25$ | 16. $3 + 3x + 4x = 24$ |
| 17. $35x = 2 + 3(4x + 5)$ | 17. $37x = 2 + 3(4x + 5)$ |

| | |
|---|---|
| 18. $3*2x+3x=19$ | 18. $5*2x+4x=18$ |
| 19. $3(2x+4)=12(9+2x)$ | 19. $4(3x+3)=5(6+2x)$ |
| 20. $21x=3*2(2x+3)$ | 20. $24x=3*2(2x+4)$ |

## APPENDIX B

### Tutoring Scripts

**Four Rules Referred to During Tutoring**

1. Precedence—Multiply or Divide before Adding or Subtracting (Mnemonic: "My Dear Aunt Sally")
2. "Get all the Xs to one side, all the numbers to the other"
3. "To undo added things, you subtract, to undo multiplied things, you divide."
4. "Whatever you do to one side, you must do the same thing to the other side."

**Reteaching (a condition in all three studies)**

1. [FRESH PAPER]
2. Have student work the task aloud
3. If wrong, say "THIS IS WRONG".
4. [FRESH PAPER]
5. Say, "LET ME SHOW YOU *HOW* TO DO IT" (AND *WHY*—using the four rules)".
6. GIVE PRACTICE TASKS

**MBR (a condition in all three studies)**

1. [FRESH PAPER]
2. Have student work the task aloud
3. After the student has completed the task, go back to *EACH* error, say: "IT LOOKS LIKE YOU (DID).....THIS IS WRONG BECAUSE..." (Address the Four Rules)
4. Say, "LET ME SHOW YOU *HOW* TO DO IT" (AND *WHY*—Using the four rules)"
5. GIVE PRACTICE TASKS

**MBR + "cognitive engagement" (a condition in the second study only)**

1. [FRESH PAPER]
2. Have the student work the task aloud
3. "TELL ME HOW (AND WHY) YOU DID (the errors)" No need to review correct steps. [Student "targets" own errors]
4. For each error say, "THIS IS WRONG."

5. Say, "LET ME SHOW YOU *HOW* TO DO IT" (AND *WHY*—Using the four rules)" Do entire task.
6. [FRESH PAPER]
7. Re-present task.
   Say, "NOW *YOU* TELL ME HOW/WHY TO DO THIS TASK"
   You need a *how & why* for each step, if possible. If (s)he makes an error, correct it on the spot; (s)he does not have to repeat this step.
8. GIVE PRACTICE TASKS

**MBR + "cognitive dissonance" (a condition in the second study only)**

1. [FRESH PAPER]
2. Have the student work the task aloud
3. Even if the answer is correct, say: "PLEASE CHECK YOUR ANSWER" Have the student substitute their answer for X. If they don't remember substitution, remind them. Substitute in their wrong answer, and have them agree that the two sides do not balance. Ask, "HOW DO WE KNOW THAT IT IS WRONG?"---because the two sides don't balance. If they seem TOTALLY lost, you should give a very obvious example (e.g., 5X = 15).
   If their *wrong* answer DOES balance the sides, say, "EVEN THOUGH THIS VALUE FOR X IS RIGHT, YOU GOT IT FOR THE WRONG REASON.....(and explain)...
4. Say, "THIS IS WRONG"/"THIS METHOD IS WRONG."
5. Say, "LET ME SHOW YOU *HOW* TO DO IT" (And *Why*—Using the four rules)" Do entire task.
6. GIVE PRACTICE TASKS

## APPENDIX C

### Classification of Errors in the Second Study

**Algebraic Errors**

1. Changed the side but not sign of an x-term
2. Changed the side of the larger positive x-term but placed the minus sign before the smaller x-term.
3. Added negative sign(s), e.g., $3x+5=9+6 \Rightarrow 3x=-9-6-5$
4. Added an x-term to a constant, e.g., $4x+3 \Rightarrow 7x$
5. Ended with the step $ax=bx$.
6. Inverted division. $ax=b \Rightarrow x=a/b$
7. Subtracted the coefficient.
8. Changed the side of a multiplier, e.g., $5*3x+7=9 \Rightarrow 3x+7=9*5$
9. Dropped an x, e.g., $3x+3=6x \Rightarrow 3x+3=6$
10. Other low frequency algebraic errors.

**Non Algebraic Errors**

1. Precedence errors
2. Distributive property errors
3. Misreading a sign/number
4. Incomplete working
5. Item not attempted
6. Arithmetic errors
7. Dropping a negative sign
8. Other non-algebraic errors.