We report an empirical study of elementary algebra errors, conducted in three separate schools. The errors are diagnosed using mal-rules, as proposed by Sleeman (1984, 1985). Our analysis uncovers the following properties of algebra mal-rules: The frequency of mal-rules is severely skewed, there are many infrequent mal-rules and few frequent ones; mal-rules are very unstable, students typically use mal-rules very irregularly; different mal-rules have explanatory power in different schools (many of our most powerful mal-rules are previously unreported); mal-rule diagnosis is more successful with more skilled students; students' confidence ratings do not partition the total set of mal-rules, every mal-rule we find is associated with high confidence ratings by at least one student. The implications of our data for cognitive theories of error generation are discussed. Contrary to commonplace assumptions, we argue that it is impossible to make a clear distinction between slips and mistakes; most errors depend on properties of the knowledge base and the cognitive architecture. Errors in a procedural skill cannot be assumed to be purely syntactic in origin.

1. INTRODUCTION

Theoreticians have long recognized that important insights into the nature of cognitive skill and its acquisition can be gained by examining errors. This approach also has potential practical implications for instruction, because achieving a "cognitive diagnosis" of a learner's errors may be an important step towards meaningful individualized tutoring.

The general goal of this article is to extend, through an empirical study, the recent work that has offered cognitive diagnoses of errors made in basic mathematical skills. We focus on the domain of high school algebra; specif-
ically, the solution of linear equations involving a single unknown. In this respect the current research is directly descended from that of Matz (1982) and, especially, Sleeman (1984, 1985); but also in its immediate lineage is the work on repair theory (Brown & Van Lehn, 1980), and its descendant the SIERRA model (VanLehn, 1983), which has concentrated on multicolumn subtraction, but which offers explanations of error patterns that have a much broader potential scope.

A central claim of all these researchers is that many errors can be explained by the student’s mental representation and application of a faulty procedure, called either a “bug” or a “mal-rule”.

The force of this claim relies on two characteristics of mal-rules: individually and as a set, they explain a considerable proportion of the observed errors, and they can be generated by some plausible cognitive theory.

The first criterion, empirical adequacy, is self-evidently important, yet in the algebra studies has never been supported. VanLehn (1981) provides a wealth of statistical information about the diagnosis of subtraction errors, but Matz (1982) and Sleeman (1984, 1985) rely entirely on the analysis of examples, with frequency data being treated only qualitatively, if at all. So, although Resnick, Cauznille-Marmeche, and Mathieu (1987, p. 174) state that “there is great systematicity in the appearance of [algebra] errors; out of all the logically possible algebra errors only a small number are made with regularity,” we can find little firm evidence for this in the archived literature.

The second criterion, that observed mal-rules can be generated by a cognitive theory of skill, is a major challenge for cognitive theorists. It is uninteresting to assert merely that an error can be “explained” by a mal-rule, for it is trivially possible to describe a mal-rule as a condition-action pair mapping the problem state onto the observed behavior. To escape from this trivialization a computational account of mal-rule generation is required. Three separate classes of computational mechanism have been proposed in the literature: repair (Brown & VanLehn, 1980), deletion (Young & O'Shea, 1981), and misgeneralization (Matz, 1982; Sleeman, 1984). VanLehn (1983) has argued that for multicolumn subtraction all these mechanisms are necessary; his computational model SIERRA integrates versions of all three mechanisms.

We will offer some assessments of the extent to which each of these criteria is satisfied in the case of basic algebra. We report an analysis of the algebra problem-solving protocols of children, 13–14 years old, from three different Lancaster secondary schools. The specific goals of the study are as follows:

---

1 In this article we will use Sleeman’s term “mal-rule” as it unambiguously refers to a mental object, whereas manifest surface errors in written solution are also conventionally called bugs.
1. To itemize the mal-rules needed to explain the observed errors, thereby extending, where necessary, the catalogue of mal-rules assembled by Matz and Sleeman, among others.

2. To report the diagnostic adequacy of the mal-rules. How many mal-rules are needed to account for what proportion of the observed errors? How many errors are completely undiagnosable? Which mal-rules are the most common? How stable are mal-rules? Do they drift in and out of use by individual students? Is it easier to diagnose the errors of some students than others, and what explains this variance?

3. To question the meta-cognitive status of mal-rules. A priori it seems possible that observed mal-rules could be generated by psychological processes with very different meta-cognitive underpinnings. For example, students may be inducing or applying a problem solution that they earnestly believe to be correct or they may be knowingly guessing about a problem which they have no real idea how to solve. Such distinctions are likely to be important in their own right for the instructional implications of the mal-rule paradigm; beyond this, they may provide us with clues about the mechanisms that generate mal-rules, which is the topic with which our final goal is concerned.

4. To discuss the implications of the mal-rules we discover for theories of cognitive skill and error. Are algebra mal-rules symptomatic of slips or mistakes? Does the distinction between slips and mistakes help us to understand algebra errors? Is algebra a purely procedural skill, in which errors arise from syntactic perturbations to learned methods, or do students attempt to make sense of their tasks?

2. EMPIRICAL STUDIES

2.1 Overview

The exact form of our studies was influenced by the nature and extent of the access to pupils we were able to negotiate with the mathematics departments of local schools. We conducted studies in three separate schools, using two basic experimental paradigms, one in School 1, the second in Schools 2 and 3. In each school we used the same set of algebra problems.

The test items were based upon the task types used by Sleeman (1985). All were equations with a single unknown, which was always denoted by a letter x. Two of Sleeman's task types did not appear in the course textbook used at our schools so they were replaced by two new types which appeared frequently in the text. Table 1 shows the 14 task types used. Each type generated two problem schemas, one using the operands given by Sleeman (+ and *) the other using the inverse operands.

Two separate tests were constructed, each containing 28 problems, 1 generated from each problem schema. Problems were constructed so as
TABLE 1
Algebra Task Types Used in the Experiments

1. \( Mx = P \)
2. \( Mx = N + P \)
3. \( Mx = N \times P \)
4. \( Mx + Nx = P \)
5. \( Mx + N = P \)
6. \( M + Nx = P \)
7. \( Mx = Nx + P \)
8. \( Mx = N (P \times Q) \)
9. \( Mx = N (Px + Q) \)
10. \( Mx = N + P \cdot Q \)
11. \( M + Nx + Px = Q \)
12. \( Mx = N + P (Qx + R) \)
13. \( Mx = N (x + P) = Q \)
14. \( Mx + N = Px + Q \)

Note. Types 1–12 come from Sleeman (1985).

always to give an integer solution for \( x \), to guarantee consonance with our subjects' experience. The two tests contained problems of matching form but with different numeric contents. In each test the questions were arranged in a separate random but fixed order. The complete tests are shown in the Appendix.

All the subjects in the three studies were 13–14 years old. At each school we asked the Head of Mathematics to choose children who would be expected to answer correctly at least 50% of the algebra problems, but not 100%. However, as on all occasions, the tests were administered in a timetabled school lesson to a complete class of students ability did vary substantially, even within groups. Further details on the subjects are given separately for each school below.

2.2 Study 1

Subjects. Two separate classes were used, both of around the average ability for the school, but one class being rated higher in overall academic achievement than the other. All 58 subjects had recently revised the algebra they had learned earlier in the school year, much of which was more advanced than that contained in the test (involving negative numbers and non-integer answers, for example).

Procedure. Subjects were given the two tests in separate lessons one week apart. No feedback or instruction in algebra were given between the two lessons. For Test A, subjects were instructed to provide an answer for every question. For Test B, subjects were instructed to provide an answer only when they were sure it was correct. In both classes half the class did Test A
in the first lesson and Test B in the second and half had the order reversed. Within each lesson there was sufficient time for all students to complete the test. Tests were administered by the class teacher.

Eighteen students who answered fewer than 20 questions out of 28 on Test A, or who answered more questions on Test B than on Test A were diagnosed, but, because they obviously had not obeyed the instructions about which questions to attempt on which test, they were excluded from all analyses presented here. Of the 40 remaining student, 7 failed to write down any intermediate steps, and so could not be diagnosed accurately. The final total of 33 students comprised 18 boys and 15 girls.

2.3 Study 2

Subjects. Twenty-nine subjects were taken from a single class at this school, which was an all-girls school at which students had been selected as academically high achievers at age 11. These students were very familiar with algebra and had recently (within the last month) been instructed in more advanced equation solving than the stuff of our tests.

Procedure. The two tests were administered by the experimenter (second author) in two separate lessons on the same day, one in the morning and one in the afternoon. No feedback or instruction in algebra was given between the tests. The instructions for both tests were the same: to answer all questions, showing work (i.e., writing down intermediate steps), and to rate confidence in each answer on a five-point scale (provided on the question/answer sheet) from 1, "unsure" to 5, "certain." At the head of each test paper a brief paragraph explained how to use the confidence scales, and before each session the class teacher ensured that students understood these instructions. Subjects had sufficient time to complete all the problems and to go back to check their answers and confidence ratings if they wished (although we found very little evidence of any editing).

2.4 Study 3

Subjects. This school was a mixed-ability school with mixed-ability classes. Forty-six subjects received the test.

Procedure. The two tests were concatenated into a single test of 56 problems. Subjects were again asked to provide confidence ratings for each answer. However, the single 50-minute lesson was not enough time for most students to finish the complete test, so subjects were asked to get as far through the test as they could. Many subjects interpreted this instruction as a licence to select questions from throughout the test, presumably on an easiest-first basis.
A disappointingly high number of students (21) declined to show any intermediate steps in their solutions, and so were excluded from the mainstream of subsequent analyses. A further subject apparently did so much problem solving internally that she made only a brief start on all incorrectly answered problems. The 24 remaining students comprised 13 boys and 11 girls.

3. RESULTS 1: STUDENTS' PERFORMANCE, AND THE EXPLANATORY POWER OF MAL-RULES

3.1 General Performance
Students in School 1 can be considered in two groups, those who performed Test A first, and those who performed Test B first. On Test A all students attempted 27 or 28 of the 28 questions ($M = 27.8$; students answering fewer than 20 had been discarded). On Test B, students attempted an average of 23.7 questions; there was no significant difference between groups in the number of questions attempted, $t(31) = 1.25, p > .1$. The percentage-correct scores were analyzed in a two-way ANOVA with group as a between-subjects variable and test as a within-subjects variable. There was no significant effect for group ($F(1,31) = 3.9, p > .05$) and no Group x Test interaction ($F(1,31) = 2.0, p > .1$). The effect of test was significant ($F(1,31) = 9.7, p < .01$). The mean percentage-correct score on Test A was 52%, and on Test B was 60%. In all following analyses, the results for the two groups are pooled.

All students in School 2 attempted all 28 questions on each test. There was a significant improvement from morning to afternoon, $t(28) = 3.06, p < 0.01$. The mean percentage-correct score for the morning test was 91%; for the afternoon test it was 94%.

Students in School 3 attempted a mean number of 27.6 questions. The mean percentage-correct score was 56%. This is not significantly different from the performance of School 1 students on Test A, $t(55) = 1.23, p > 0.1$ or Test B, $t(55) = 0.77, p > .1$.

3.2 From Errors to Mal-rules
In all, we diagnosed the protocols of 96 schoolchildren who produced 1,097 error protocols. Such hand diagnosis is an extremely time-consuming task, and occasionally difficult. The inherent difficulties, and our strategies for overcoming them, are described in detail in Payne and Squibb (1988). The goal is to provide an explanation for each observed error in terms of mal-rules like those reported by Sleeman (1984) and Matz (1982) which are listed in Table 2.

Two additional categories of diagnosis were used to supplement mal-rules. First, we identified some errors as "arithmetic slips," when the flawed manipulation involved the calculation of a simple and explicitly written-down arithmetic expression. We justify this category not in terms of any theoretical distinction, but simply because our current concern is with algebra.
### TABLE 2

<table>
<thead>
<tr>
<th>No.</th>
<th>Rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>( M \times X = N \rightarrow X = M/N )</td>
</tr>
<tr>
<td>S2</td>
<td>( \text{pat1} +</td>
</tr>
<tr>
<td>S3</td>
<td>( \text{pat1} +</td>
</tr>
<tr>
<td>S4</td>
<td>( \text{pat1} = \text{pat2} +</td>
</tr>
<tr>
<td>S5</td>
<td>( M (Nx +</td>
</tr>
<tr>
<td>S6</td>
<td>( M (Nx +</td>
</tr>
<tr>
<td>S7</td>
<td>( M<em>X + N</em>X \rightarrow M*X + N )</td>
</tr>
<tr>
<td>S8</td>
<td>( M<em>X + N</em>X \rightarrow M*X + N )</td>
</tr>
<tr>
<td>S9</td>
<td>( M<em>X + N</em>X - M + X + N + X )</td>
</tr>
<tr>
<td>S10</td>
<td>( M*X - M + X )</td>
</tr>
<tr>
<td>S11</td>
<td>( M + N<em>X = - M</em>N + X = )</td>
</tr>
<tr>
<td>S12</td>
<td>( M + N*X = - M + N + X = )</td>
</tr>
<tr>
<td>S13</td>
<td>( M*X + N = - M + X + N = )</td>
</tr>
<tr>
<td>S14</td>
<td>( M<em>X + N = - M</em>X+N = )</td>
</tr>
<tr>
<td>S15</td>
<td>( M*X + N = - (M + N)*X = )</td>
</tr>
<tr>
<td>S16</td>
<td>( M*X = N+P \rightarrow X = M )</td>
</tr>
<tr>
<td>S17</td>
<td>( M*X = N+X + P \rightarrow X + X = M + N + P )</td>
</tr>
<tr>
<td>S18</td>
<td>( M<em>X = N + P \rightarrow M</em>X = N )</td>
</tr>
<tr>
<td>S19</td>
<td>( M<em>X = N + P \rightarrow M</em>X = N )</td>
</tr>
<tr>
<td>S20</td>
<td>( M*X = N \rightarrow X = N )</td>
</tr>
<tr>
<td>S21</td>
<td>( M*X = N \rightarrow X = &lt;N/F&gt;/M )</td>
</tr>
<tr>
<td>S22</td>
<td>( M*X = N \rightarrow X = N/&lt;M/F&gt; )</td>
</tr>
<tr>
<td>S23</td>
<td>( M*(N<em>X + P) \rightarrow M</em>X + M*P )</td>
</tr>
<tr>
<td>S24</td>
<td>( 2X/2X = 0 )</td>
</tr>
<tr>
<td>S25</td>
<td>( A*A/1 - 0 )</td>
</tr>
<tr>
<td>S26</td>
<td>( O*A - A )</td>
</tr>
</tbody>
</table>

**Note.** Many of Matz's mal-rules apply to more complex algebraic expressions than we have studied; we list only the ones that are appropriate, and that extend Sleeman's set. The mal-rules are presented as schematic condition-action pairs, some (e.g., the movement of a term to the other side of the equation without changing its sign) are surprisingly resistant to such an expression. We use A, B to stand for integers or simple expressions like 3x. We use M, N, P to stand for integers; pat 1 stands for any legal (in context) pattern of algebraic symbols, including nothing at all. +|− should be read "plus or minus" and must take the same value wherever it appears in a rule. These conventions differ slightly from Sleeman's own, which we have found hard to interpret. For example, M3NTO<RH (Sleeman, 1984, p. 390) appears to work correctly on some occasions (we have omitted this rule from our list, as we do not understand what it does).

rather than arithmetic per se. Our final diagnostic category was "undiagnosable." This category was used when the student had supplied sufficient working to afford a diagnosis, but we were unable to extract the bridge between adjacent steps.

It must be noted that a single error can sometimes be explained by alternative mal-rules. In general, ambiguous diagnoses can only be avoided if full account is taken of intermediate work. This demonstrates that students' intermediate steps towards a solution are not merely a convenience for cognitive diagnosis: They are necessary in principle. Furthermore, even when
intermediate steps are provided and used, it may be impossible to guarantee single diagnoses in some cases.

In practice, however, ambiguous cases were rare in our data, and in those few cases of potential ambiguity, the conflict could be avoided by using the following heuristic guidelines:

- Prefer single mal-rules to combinations
- Prefer no hidden do/undo steps
- Insist on concrete matching of the expression $Mx = N$ for rules M19–M22.

The first two heuristics are described by Ohlsson and Langley (1988), and included in their diagnostic program DPF. Our experience offers some support for the usefulness of these general heuristics, but cannot speak to their robustness with different data sets.

We identified no less than 99 separate mal-rules, although this figure might be reduced by grouping several sets of mal-rules into more abstract mal-rule types. This figure is in line with the reported bugs in multicolumn subtraction: VanLehn (1981) lists 104 separate mal-rules for this domain.

The frequency distribution of the mal-rules, as for the subtraction mal-rules, is markedly skewed, with a few common ones and a great many very rare ones. Across our entire sample from the three separate schools, 43 of the 99 mal-rules appeared only once, and a further 14 no more than three times. It is natural to question whether such low-frequency mal-rules would not be better described as action slips, a distinction which we will postpone to our discussion of the implications of our data in Section 6.1.

Tables 3, 4, and 5 list the most frequent mal-rules for each of the three schools, along with the number of arithmetic slips and undiagnosable errors. Mal-rules are listed if they were used by more than one student across all three schools and were observed more than three times within that school.

Table 3 (School 1 mal-rules), excludes the information from 1 subject whom we diagnosed as having very idiosyncratic difficulties. In School 3, a further 5 students were diagnosed as having idiosyncratic difficulties and were excluded from the analysis in Table 5. These subjects will be discussed individually below.

3.3 How Diagnostic are the Mal-rules?
A critical question concerns the diagnosticity of these sets of mal-rules: What proportion of the errors do they allow us to diagnose? Allowing arithmetic slips as a diagnosis, the following pattern emerges:

**School 1.** Using all the mal-rules, including those that only occurred once, we were able to diagnose 87% of the errors. If singleton mal-rules are discounted, then we diagnosed 84%. The 10 most frequent mal rules diagnosed 67%, and the 5 most frequent diagnosed 37%.
### Table 3
Mal-rules for School 1

<table>
<thead>
<tr>
<th></th>
<th>Test A</th>
<th>Test B</th>
<th>Total (N)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>M1</strong></td>
<td>$M \cdot \left( pat \right) - M + pat$</td>
<td>33</td>
<td>18</td>
</tr>
<tr>
<td><strong>M2/S3</strong></td>
<td>$M \cdot \left( Nx + P \right) - M \cdot Nx + P$</td>
<td>15</td>
<td>9</td>
</tr>
<tr>
<td><strong>M3</strong></td>
<td>$M + \left( pat \right) - \left( M + \left( - N \right) \right)$</td>
<td>6</td>
<td>13</td>
</tr>
<tr>
<td><strong>M4</strong></td>
<td>$M \cdot \left( N \cdot P \right) - M \cdot N \cdot M \cdot P$</td>
<td>9</td>
<td>8</td>
</tr>
<tr>
<td><strong>M5</strong></td>
<td>$M \cdot \left( Nx + N \right) - M \cdot \left( M \cdot M \right)$</td>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td><strong>M6</strong></td>
<td>$- M \cdot \left( Nx - P \right)$</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td><strong>M7</strong></td>
<td>$\left( pat \right) \cdot \left( pat \right) - \left( pat \right) \cdot \left( pat \right)$</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td><strong>M8</strong></td>
<td>$M \cdot \left( Nx + P \right) - Nx + P - M \cdot P$</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td><strong>M9/S2</strong></td>
<td>$\left( pat \right) + \left( M \right) \cdot \left( pat \right) = \left( pat \right) - M$</td>
<td>39</td>
<td>26</td>
</tr>
<tr>
<td><strong>M10/S4</strong></td>
<td>$\left( pat \right) + \left( Mx \right) \cdot \left( pat \right) = \left( Mx \right) \cdot \left( pat \right)$</td>
<td>29</td>
<td>23</td>
</tr>
<tr>
<td><strong>M11</strong></td>
<td>$Mx + N = Px + Q$</td>
<td>20</td>
<td>15</td>
</tr>
<tr>
<td><strong>M12/S15</strong></td>
<td>$Mx + N = \left( M \right) - \left( N \right)$</td>
<td>42</td>
<td>38</td>
</tr>
<tr>
<td><strong>M13</strong></td>
<td>$Mx + N = \left( M \right) - \left( N \right)$</td>
<td>25</td>
<td>12</td>
</tr>
<tr>
<td><strong>M14</strong></td>
<td>$\left( pat \right) + \left( pat \right) - \left( pat \right) - \left( pat \right)$</td>
<td>28</td>
<td>6</td>
</tr>
<tr>
<td><strong>M15</strong></td>
<td>$\left( pat \right) - \left( pat \right) - \left( pat \right) + \left( pat \right)$</td>
<td>20</td>
<td>9</td>
</tr>
<tr>
<td><strong>M16</strong></td>
<td>$\left( pat \right) - \left( pat \right) - \left( pat \right) + \left( pat \right)$</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td><strong>M17</strong></td>
<td>$A + B - B - A$</td>
<td>9</td>
<td>8</td>
</tr>
<tr>
<td><strong>M18</strong></td>
<td>$A - B - B - A$</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td><strong>M19</strong></td>
<td>$Mx = N - x = M + N$</td>
<td>10</td>
<td>6</td>
</tr>
<tr>
<td><strong>M20/S20</strong></td>
<td>$Mx = N - x = N$</td>
<td>1</td>
<td>11</td>
</tr>
<tr>
<td><strong>M21</strong></td>
<td>$Mx = N - x = M - N$</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td><strong>M22/S1</strong></td>
<td>$Mx = N - x = M/N$</td>
<td>3</td>
<td>1</td>
</tr>
</tbody>
</table>

### Note
The same conventions have been used to describe these rules as were used for the Sleeman rules of Table 2. Expressions enclosed in square brackets denote the result of the expression. Expressions of the form $Mx + N$ will also match $N + Mx$. We could think of no way of representing this order independence, and we do not know whether or not this convention holds for Sleeman's rules.

### School 2
The many fewer errors made by this group were better diagnosed: 97% could be diagnosed using all mal-rules, dropping to 82% using only the best 10 mal-rules, and 72% using the best 5.

### School 3
Using all mal-rules, 82%; best 10 mal-rules, 52%; best 5 mal-rules, 42%.

These figures seem quite encouraging (although it should be remembered that they would be slightly depressed if "idiosyncratic students" were in-


## TABLE 4

**Mal-rules for School 2**

<table>
<thead>
<tr>
<th>Mal-rule</th>
<th>Total</th>
<th>Students (N)</th>
</tr>
</thead>
<tbody>
<tr>
<td>M4 $M-(N+P) - M+N-M+P$</td>
<td>17</td>
<td>9</td>
</tr>
<tr>
<td>M6 $- M (Nx - P) - M+Nx - M+P$</td>
<td>12</td>
<td>5</td>
</tr>
<tr>
<td>M3 $M +</td>
<td>- N (pat) - [M +</td>
<td>- N] (pat)$</td>
</tr>
<tr>
<td>M5 $M (Nx +</td>
<td>- P) - M (M+Nx +</td>
<td>- M+P)$</td>
</tr>
<tr>
<td>M15 $pat1 - pat2 - pat1 + pat2$</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>M9/S2 $pat1 +</td>
<td>- M pat2 = pat3 - pat1 pat2 = pat3 +</td>
<td>- M$</td>
</tr>
<tr>
<td><strong>Arithmetic slips</strong></td>
<td>23</td>
<td>12</td>
</tr>
<tr>
<td><strong>Undiagnosable errors</strong></td>
<td>4</td>
<td>3</td>
</tr>
</tbody>
</table>

## TABLE 5

**Mal-rules for School 3**

<table>
<thead>
<tr>
<th>Mal-rule</th>
<th>Total</th>
<th>Students (N)</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1 $M (pat) - M + pat$</td>
<td>29</td>
<td>13</td>
</tr>
<tr>
<td>M3 $M +</td>
<td>- N (pat) - [M +</td>
<td>- N] (pat)$</td>
</tr>
<tr>
<td>M7 $pat1 (pat2) pat3 - pat1 pat3$</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>M9/S2 $pat1 +</td>
<td>- M pat2 = pat3 - pat1 pat2 = pat3 +</td>
<td>- M$</td>
</tr>
<tr>
<td>M10/S4 $pat1 = pat2 +</td>
<td>- Mx pat3 - pat1 +</td>
<td>- Mx = pat2 pat3$</td>
</tr>
<tr>
<td>M12/S15 $Mx +</td>
<td>- N - [M +</td>
<td>- N]x$</td>
</tr>
<tr>
<td>M13 $Mx +</td>
<td>- N - [M +</td>
<td>- N]$</td>
</tr>
<tr>
<td>M14 $pat1 + pat2 - pat1 - pat2$</td>
<td>11</td>
<td>5</td>
</tr>
<tr>
<td><strong>Arithmetic slips</strong></td>
<td>11</td>
<td>5</td>
</tr>
<tr>
<td><strong>Undiagnosable errors</strong></td>
<td>39</td>
<td>13</td>
</tr>
</tbody>
</table>

eluded in the statistics). Despite our large initial number of mal-rules, it seems that a small number of mal-rules can diagnose a substantial number of the errors in any one group. There are, however, two limitations inherent in these diagnosticity data.

First, it is obvious that mal-rules are more powerfully diagnostic in one of the schools than in the other two. It is interesting to note that they are more diagnostic in the highest-performing school. This suggests the possibility that mal-rules may be a more powerful construct for nearly-learned skills than for ill-learned ones. To test this idea further, we examined the correlation between performance, as measured by number of test items correct, and diagnosticity as measured by the percentage of protocols that were successfully diagnosed, for each student. Because of the slightly different test conditions in each school, we computed three separate correlation coefficients. In all three schools the correlation was significant:
These figures lend some support to our conjecture that regular mal-rules are a property of nearly-learned skills. This may limit their application in tutoring situations, but at least it provides some positive indications about their scope.

Second, the rules that do the most explanatory work in the three separate groups have surprisingly little overlap. One index of overlap is gained by assessing how well the 10 most powerful mal-rules from School 1 diagnose the errors of Schools 2 and 3. In School 2 they diagnose fewer than 50% of the errors, and in School 3 fewer than 10%. The overlap is even worse if we compare our three samples with the mal-rules reported by Sleeman (1984) and Matz (1982) (Table 2). Only 6 of the 22 most diagnostic rules in our samples were reported by Sleeman.

The practical upshot of this finding is that a large number of prespecified mal-rules is clearly going to be necessary for catalogue-based diagnosis. Furthermore, there is no guarantee that a new sample of schoolchildren will not unearth a whole new set of mal-rules. After all, Sleeman did not observe (or does not report) many of our mal-rules, and we did not observe many of his.

One theoretical implication is that specifics of educational experience seem to have a heavy influence on acquired mal-rules. The most likely explanation to us for this seems to be that the exact form and content of students' knowledge determines the trajectory of their inductive learning; this relationship is in good accord with the theories of mal-rule generation discussed below. A second possible explanation is that mal-rules are communicated among students in some way, perhaps in collaborative problem-solving sessions. Although we do not favor this explanation in the particular context of high school algebra where the current social context encourages competition more than collaboration, we would expect the communication of understandings in a subculture to be an important explanatory notion in more informal learning situations.

Finally, we cannot discard the possibility that the different procedures in the three schools may have had some impact on the observed mal-rules. However, the similarities in the number of questions attempted and the proportion of correct answers, reported in Section 3.1, do not encourage such an explanation.

3.4 How Stable are Mal-rules?
The overall impression in our data is that very few mal-rules are used regularly by students. It is very difficult to offer precise, numeric summaries of this impression, because different mal-rules have widely different “opport-
Table 6
Stability of Mal-rules.

<table>
<thead>
<tr>
<th>Mal-rule</th>
<th>Opportunities</th>
<th>$0 &lt; p &lt; 0.25$</th>
<th>$p &lt; 0.5$</th>
<th>$p &lt; 0.75$</th>
<th>$p &lt; 1$</th>
<th>$p = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$S_1$</td>
<td>$S_2$</td>
<td>$S_3$</td>
<td>$S_1$</td>
<td>$S_2$</td>
</tr>
<tr>
<td>M1</td>
<td>14</td>
<td>10</td>
<td>11</td>
<td>5</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>M2</td>
<td>6</td>
<td>10</td>
<td></td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M3</td>
<td>4</td>
<td>12</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>M4</td>
<td>2</td>
<td>5</td>
<td></td>
<td>1</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>M5</td>
<td>12</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M6</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>M7</td>
<td>16</td>
<td>7</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M8</td>
<td>12</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M9</td>
<td>20</td>
<td>20</td>
<td>6</td>
<td>5</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>M10</td>
<td>16</td>
<td>18</td>
<td>3</td>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M11</td>
<td>4</td>
<td></td>
<td>12</td>
<td>5</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>M12</td>
<td>32</td>
<td>12</td>
<td>11</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M13</td>
<td>32</td>
<td>17</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M14</td>
<td>22</td>
<td>18</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M15</td>
<td>22</td>
<td>15</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M16</td>
<td>18</td>
<td>3</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M17</td>
<td>22</td>
<td></td>
<td>13</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M18</td>
<td>22</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M19</td>
<td>56</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M20</td>
<td>56</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M21</td>
<td>56</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. This table shows the number of students in each school using each mal-rule on a given proportion of opportunities for use.

Stabilities of Mal-rules. The opportunities of occurring, and in some cases the number of opportunities is impossible to quantify without looking closely at individual students' protocols. Our approach to this problem was to estimate the number of opportunities for a mal-rule conservatively, so that the number of uses by each student could be quantified as a maximum proportion of potential uses, providing an overestimate of mal-rule stability. For mal-rules M1–M18 we assumed opportunities were limited to one per problem, and to those problems for which the initial state matched the left-hand side of the mal-rule. For mal-rules M19–M21, which only match penultimate problem states, we itemized opportunities for each relevant student. For students in School 3, who attempted different questions, the number of opportunities had to be computed individually.

Table 6 lists the total number of opportunities for each mal-rule, together with the number of students who used the mal-rule on any given proportion of the opportunities. It can be seen that most students use most mal-rules on a very low proportion of the opportunities: The vast majority of proportions are less than 50%. The only exceptions are mal-rules M4 and M6, which have so few opportunities that any usage must be quite "regular."
### 3.5 Distribution of Mal-rules Across Question Types

Table 7 enumerates the occurrences of mal-rules on each problem type. For convenience, the mal-rules have been divided into three sets: bracket manipulation mal-rules (M1–M8), which can only occur on (some of) Question Types 8, 9, 10, 12, 13; last-step-division mal-rules (M19–M21) which can occur on any question type; other mal-rules (M9–M18), which can occur on mixed subsets of the rules. Because of their limitation to the last step of problem solutions, and a particular, simple problem state, we have not shown the distribution of the last-step-division mal-rules. (For the record, these
mal-rules were very evenly distributed, with no more than four occurrences on any one question type). Some mal-rules are noted as occurring on question types which, for the sake of our stability analysis, would not have been regarded as opportunities (marked with swords). This reflects the possibility that earlier errors lead to unanticipated intermediate states, and reiterates the exaggerated nature of our stability estimates.

For each question type, it is possible to enumerate a crude index of complexity by counting the number of operations required to successful solution: Type 1 requires a single division (or multiplication), Type 14 requires two change-side operations, two additions (or subtractions) and a division (or multiplication). Intermediate types require 2, 3, 4, or 5 operations.

For each mal-rule that can occur at more than one complexity level it is possible to compute the correlation between complexity and number of occurrences, using every applicable question type as a data point. We computed the Spearman rank correlation coefficient for each mal-rule which could occur on 4 or more question-types (i.e., M1, M9, M10, M12, M13, M14, M15, M17, M18). None of these correlations approached significance; 3 out of 9 were negative. It seems that our index of complexity does not capture an interesting proportion of the variance in mal-rule distribution across questions.

Eyeballing Table 7, we can confirm that most mal-rules do seem fairly evenly distributed across the question types upon which they are applicable. One anomalous result of some interest is the extent to which Rules M12 and M13 seem likely to occur on bracket-manipulation problems. Is it possible that these rules sometimes denote a confusion about the reduction of an $Mx + N$ term, only when enclosed in brackets? In our data overall, however, we could find no evidence of such a bias; although M12 and M13 occur frequently on question types involving brackets, they do not apply reliably more often to terms enclosed in brackets. We will offer an alternative explanation for this effect below.

3.6 Distribution of Mal-rules Across Numeric Contexts

Although it appears that, in general, mal-rules are not tied to particular subsets of those equation types for which they are applicable, it remains possible that the specific numbers used in equations could exert an important influence. It must be noted that our equations were not designed to test specifically for numeric influences, so that each specific equation form is only instantiated with one set of numbers on each test. This makes it impossible to investigate the influence of numbers on the equation solving of individual students, but it is possible to look across all students for between-question imbalances.

Each cell of Table 7 records the mal-rules that occur on four separate questions (2 schemas $\times$ 2 numeric instantiations). For every schema on
which more than 10 mal-rules of a given type are noted, at least 15% occur on each question, with one exception: M13 occurs 14 times on Question 3, and only once on a question of similar form in Test B. This peculiarity will be examined in Section 6.2. In general, mal-rules are very evenly distributed among questions within question types. It is impractical to compute likelihoods for these statistics, but, with the exception we have noted, there is no strong evidence of specific numbers having reliable effects.

A more general potential influence of numeric context is that mal-rules requiring a subtraction or a division may be more available if those operations lead to natural numbers than if they lead to negative numbers or fractions. For example, Sleeman's mal-rule, S1, Mx = N - x = M/N was observed only four times across all three studies, but this may be because our equations were constrained to produce integer answers (see Section 2.1), making situations where M > N contingent on earlier errors (the same argument applies for M21). Investigating the four occurrences of S1, we find that two were in the context of earlier errors, and did produce integer answers, but the other two produced fractional answers.

None of our frequent mal-rules involve division, but the subtraction versions of M3, M9, M10, M12, M13, M14, M17, and M18 may, in principle, all suffer from this bias, depending on the numbers in the equations. To investigate this, we can inspect the specific questions, to check which ones may mitigate against which mal-rules, and test this against the data. (As Table 7 shows, M17 and M18 are too sparse, and too spread across questions for any such analysis to make sense.)

For M3, and M9, the numbers in the questions never bias against the subtraction version. For M3, M > N on both questions which match M - N(pat) (Q3 and Q41). M3 was observed on both these questions. Similarly for M9, the negative integers on the left-hand side of the equation in Q11 and 438 were both smaller than a right-hand side integer. M9 was observed on both these questions.

For M10, only Question Type 12 presents opportunities for the subtraction version of the mal-rule. However, on both the questions of form Mx = N + P (Qx + R), (i.e., Q3 and Q41) application of the mal-rule may be discouraged because P*Q > =. Nevertheless, M10 was observed on both these questions (three times on Q3, twice on Q41). This compares with eight occurrences of the addition version of M10 for the remaining two questions of Type 12.

For M12 and M13, M > N is true for 8 of the 16 patterns that match Mx - N or M - Nx. Specifically, M > N is true on Q3, Q5, Q17, Q23, Q25, Q34, Q47, and Q54, whereas N > M is true on Q4, Q11, Q22, Q30, Q31, Q38, Q41, and Q45. For M12 there are 17 occurrences on the M > N ques-

2 The importance of this potential bias in our data was brought to our attention by Carolyn Kieran.
tions, and 17 occurrences on N>M questions. For M13 there are 4 occurrences on the M>N questions and 12 occurrences on the N>M questions.

These statistics do not demonstrate that no students prefer natural number results. It is impossible to analyze such issues without deliberately crafted equations. What can be concluded is that there is no crude bias that renders the abstract descriptions of the mal-rules, or the estimates of their stability, invalid.

4. RESULTS 2: IDIOSYNCRATIC STUDENTS

As noted above, we judged 6 students (1 from School 1, 5 from School 3) to have special problems, in the sense that they each produced error protocols of a regular form that did not appear at all in the protocols of other students, and were not readily expressible in the rule notation of Tables 3, 4, and 5. To some extent, the judgment to treat these students separately is questionable, but we believe that their error patterns are sufficiently different in kind from the mal-rules that are distributed among children to justify this approach. This qualitative difference plays a part in our discussion of mal-rule generation in Section 6.2.

The "idiosyncratic" student (SH) from School 1 exhibited very peculiar behaviors. His answers were rarely given in standard form, even when the work toward them appeared to be correct, or easily diagnosed. For example:

\[
SH:
5x = 2(4x + 2)
\]
\[
| \quad M12
\]
\[
5x = 2 \times 6x
\]
\[
| \quad 5x = 12x
\]
\[
| \quad 5x/12x
\]

The last step, which was extremely common in SH’s protocols appears to suggest a belief that the aim of the exercise is to isolate a 1 on one side of the equation, and report the other side as the answer. Sometimes, this step is extended to produce a single answer of the form Mx, as shown below.

\[
SH:
6x = 90/3
\]
\[
| \quad 6x = 30
\]
\[
| \quad 6x/30 = 5x
\]
If this answer is taken as meaning "x = 5" then it is correct. A fairly high proportion (about half) of SH's answers are "correct" in this sense, and the other half can be partially diagnosed in terms of the mal-rules, like the example above. It appears that SH has some understanding of the standard algebra notation, but has invented his own variant for some manipulations. Furthermore, the above protocol illustrates another feature of our protocols: Children will sometimes use the equal sign to separate two steps in the derivation, instead of moving onto a new line. This extended use of the equal sign (to denote logical equivalence between equations) is strictly erroneous, but need not lead to incorrect problem solutions.

Two "idiosyncratic" students (MF and SF) from School 3, showed strong evidence of a misconception that is noted briefly by Sleeman (1984). For equations with two x terms, they provided two separate answers. MF provided two answers for 8 of the 9 such questions he attempted. SF provided two answers for 4 of 6.

MF:

\[
5x = 8x - 6
\]
\[
\text{ } \text{ } \text{ } \text{ } x = 10 \quad x = 7
\]

SF:

\[
9x = 2x + 49
\]
\[
9 \cdot 7 = 65
\]
\[
2 \cdot 8 = 16
\]
\[
+ 49 = 65
\]

SF's protocol also exhibits a clear arithmetic slip (9 \cdot 7 = 65), and a nonstandard way of showing working and answers.

The fourth idiosyncratic student (DR) used a strategy of adding all the integers appearing on the left and right of the equation, and dividing the bigger by the smaller. For example:

DR:

\[
9x = 6 + 3(2x + 4) \quad 3x + 2(x + 5) = 45 \quad 2x + 8 = 3x + 2
\]
\[
\text{ } \text{ } \text{ } \text{ } x = 15 \quad 10 = 45 \quad 10x = 5x
\]
\[
\text{ } \text{ } \text{ } \text{ } x = 1.5 \quad x = 4.5 \quad x = 2
\]

Each of these errors could be described by a (different) combination of mal-rules. But the regularity is hard to deny.

The fifth idiosyncratic student (LF) used a rather more obscure strategy, which appeared to involve subtracting M from Mx terms to isolate the x,
and, presumably to preserve equilibrium, subtracting it from all the other terms as well. But this strategy invariably was mixed with other errors, making it difficult to pin down. For example:

**LF:**

\[
5x = 8x - 6 \\
| \\
x = 3x - 1 \\
| \\
x = 2x \\
| \\
x = 1
\]

The sixth and final idiosyncratic student (NJ) was even harder to understand. We could find no regularity in her protocols at all. For example:

**NJ:**

\[
6x = 39 - 3 (8x - 7) \\
| \\
3 off 7, 39 and 3 \\
= 7 - 3 = 4 \\
(39 - 3) = 36 + 4 = x = 40
\]

In summary, it appears to us that the idiosyncratic students have problems understanding algebra. These problems are not easily explained by positing conventional mal-rules. Some broader implications of this finding are discussed in Section 6.2.

**5. RESULTS 3: META-COGNITIVE STATUS OF MAL-RULES**

The design of our studies allows us to question some of the meta-cognitive judgements that are associated with mal-rules, giving us a glimpse at some of their underlying causes. Of particular importance is the distinction between mal-rules in which the subject "believes," and those which have been generated by a desperate student who simply has no idea how to solve the problem. It is possible that the wide variety of mal-rules we have reported could be collapsed into a subset of genuinely-believed-in misconceptions.

In order to examine this question, we must first confirm that our students have some access to how well they are performing, and that our experimental tasks allowed them to express this meta-knowledge. Do students in School 1 answer correctly a higher proportion of questions on Test B, in which they should give answers only when sure, than they do in Test A in which they attempt all the problems? Do students in Schools 2 and 3 give higher confidence ratings for their correct answers than for their incorrect ones?
Both these questions need to be answered at two levels. First, the data for the complete sample must be examined to see whether there is any reliable evidence in the sample as a whole for the hypothesized response bias. Second individual subjects can be studied to see if they exhibit the expected behavior.

In School 1, the mean performance (correct/attempted) on Test A was 52% and on Test B, 60%. As reported in Section 3.1, this difference is highly significant, $F(1,31)=9.7, p<.01$.

Examining the individual subjects revealed that 10 of the 32 subjects scored better proportionally on Test A than they did on Test B. These subjects clearly either have not obeyed the instructions, or are unable to assess confidence in an answer meaningfully. They are therefore excluded from subsequent analyses of meta-cognition.

For Schools 2 and 3, a formal statistical test was not necessary to ascertain that some meta-knowledge was applicable to the confidence ratings. When subjects' mean confidence ratings on correct versus incorrect scores was computed, only 3 subjects out of 29 from School 2, and 1 subject out of 19 from School 3 failed to be more confident in their correct answers, and in these cases the difference between means was minimal. Again, these subjects were dropped from further analysis of meta-cognitive factors. The mean confidence ratings for correct answers was 4.77 in School 2 and 4.05 in School 3. The mean confidence in incorrect answers was 3.54 in School 2 and 2.23 in School 3.

It seems that subjects in all three samples were able to use some meta-knowledge to select the problems they attempted or to rate their confidence in their answers. Does this meta-knowledge partition the mal-rules? It may be possible to discover some mal-rules in which nobody really believes, and so fail to appear in School 1/Test B or always receive low-confidence ratings in Schools 2 and 3. Such a partition may provide practical relief from the large total number of mal-rules, as intuitively it seems that for instructional purposes "believed-in" mal-rules are the most important. Unfortunately, we can find no such partitioning in our data. We will consider School 1 first.

From Table 3 it can be seen that all the mal-rules that appear more than three times appear in both Test A and Test B. This pattern is maintained when the 10 subjects who showed no clear evidence of meta-cognitive judgment are excluded. It makes little sense to pose the same question for less frequent mal-rules, for, as we have seen, there are more total errors on Test A, and so we would expect a majority of mal-rules on that test simply by chance: 3 out of 3 in Test A would not be a convincing bias.

When we consider Schools 2 and 3 a similar pattern emerges. For each of these samples we listed the mal-rules that appeared in incorrect solutions given a confidence rating of either 4 or 5. All the mal-rules in Tables 4 and 5 appear in this list. Once again, all the reasonably frequent mal-rules appear
to be believed in by at least one person, and once again it does not make statistical sense to test whether this is true for low-frequency mal-rules.

In summary, although we have evidence that our subjects were not behaving randomly when they made confidence ratings or decided not to attempt questions, we can find no mal-rules that seem to reflect only "coping/guessing" strategies and never genuine attempts at solution.

This has two important implications. Pragmatically, it denies us the possibility of reducing our mal-rule set. Theoretically, it suggests common psychological underpinnings for low- and high-confidence errors. One possible explanation is that the psychological resources that can be applied to coping or "wild guessing" are the same resources as those that are applied to perform inductive learning. This account suggests that some "believed-in" errors may emerge initially as some kind of coping strategy, and drift into their believed-in meta-cognitive status with use. A second possibility is that some mal-rules are accompanied by confidence because they have arisen as unnoticed performance slips, but that these slips, in turn, produce internalized, weak rules that are associated with low-confidence solutions. This second account relies on theoretical ideas that are introduced in the next section.

6. DISCUSSION

Theories of cognitive error in mathematics all share two assumptions. The first assumption is that errors can be classified as "slips" or "mistakes." Indeed, this assumption has a widespread acceptance in the psychology of cognitive skill. The second assumption is that mal-rules arise in the context of purely formal manipulation of symbols, in the absence of any semantic rationalization of these manipulations. Again, this assumption, or variants of it, play an important role in cognitive theory beyond mathematical skills. Several computational accounts of cognitive skill (e.g., Anderson, 1983; Card, Moran, & Newell, 1983; Langley, 1985) characterize skill as the representation of learned methods for performing tasks. Learning is modelled by the acquisition and tuning of methods, leaving no role for understanding (e.g., elaborating initial problem states; Payne, 1988).

In this section we find that both assumptions are undermined by our data. Beyond this, we outline a theoretical account of algebra errors that answers some of the riddles posed by the breaking of the first assumption, by using mechanisms that break with the second.

6.1 Slips Versus Mistakes

The distinction between these two basic error types is currently widely accepted in the literature on cognitive errors (Anderson & Jeffries, 1985; Norman, 1981, Reason & Mycielska, 1982). In Norman's terms the distinc-
tion hinges on the intentions of the actor. If you intend to perform the appropriate action but fail to do so, then you have slipped; if you formulated the intention incorrectly then you have made a mistake. In this framework, then, slips are seen as a consequence of the performance system, of the fixed human information-processing architecture, and are heavily influenced by factors such as working-memory load, allocation of attentional resources, and so on. Mistakes, on the other hand, are evidence of shortcomings in competence. Slips are architecture based, mistakes are knowledge based.

Matz (1982) has argued that both mistakes (which Matz calls "conceptual errors"), and slips (which Matz calls "execution errors") are to blame for algebra errors, but which sort of error do our mal-rules describe? The initial aim of mal-rule theories was certainly to model mistakes, rather than slips. Brown and VanLehn (1980, p. 380) took a very strict approach, declaring that "an entire test's answers must be generated by a bug before we are willing to say the bug exists." If we were to take a similar approach, we would be left with a tiny number of mal-rules to analyze.

In fact, Brown and VanLehn have also been forced to revise this strong statistical separation of slips from mistakes in the face of the phenomenon of "bug migration" in subtraction tests, which denotes a switching back and forth between mal-rules in a family (a "migration class"). VanLehn (1981) further notes some tendency for mistakes to come and go, in a way that parallels the instability of our data, but not to the extent that we have reported. VanLehn points out that the empirical instability of mal-rules undermines the most simplistic empirical aspect of the slips/mistakes distinction: Infrequent, irregular errors may, after all, be due to mistakes.

Two error-generating mechanisms have been advanced as explanations of the phenomenon of bug migration. The first, repair (Brown & VanLehn, 1980, VanLehn, 1983) describes errors as arising from local problem solving at "impasses," problem states in which the learner's knowledge does not support a goal-directed operation. The local problem solver uses inductive shortcuts ("repairs") to bridge to a new problem state from which the knowledge base can take over (the details of the repairs need not concern us here). According to repair theory, irregular errors arise from the application of different repairs to the same impasse. The second account of mistake instability is due to Sleeman (1984), who points out that misgeneralization can account for varied errors, because several misgeneralizations of the same rule will often be possible. This may be due in part to the fact that the psychological scope of misgeneralization as a theory is underconstrained—surely all conceivable misgeneralizations of a rule would not be entertained by students—but the general point seems correct.

Despite their capacity to cope with bug migration, neither of these mechanisms readily handles the characteristic instability of our data. The first
problem is explaining the cooccurrence of mal-rules with correct versions. Both mechanisms generate errors by positing a resource that is called up in the absence of ready-made rules. In order to explain the existence of mal-rules alongside correct rules, they must posit that the correct rules have at one stage been forgotten (or never learned), but later reinvented by the repair or generalization mechanisms. Even if this is allowed, a second problem must be faced. Not only does the correct rule coexist with a mal-rule, but for almost every student, for almost every mal-rule, the correct version is preferred on the vast majority of occasions (see Table 6). The only mal-rules exempt from this observation are M4 and M6, which can occur only on few questions, and so must occur regularly, if at all. VanLehn’s (1983) SIERRA model, a descendant of repair theory, includes a mechanism—storing impasse/repair pairs as “patches”—which could conceivably be tailored to bias a student to one rule in a migration set, but this device cannot be stretched to explain the bias of all students to the same rule (i.e., the correct one). We must conclude that current versions of mistake-generating theories are unable to cope with the universal irregularity of our observed algebra errors.

In this case, it is appropriate to examine whether the algebra errors we have noted are simply slips. Anderson and Jeffries (1985), investigating novices’ Lisp errors, come to exactly that conclusion: The majority of errors are slips due to losses from working memory, rather than misconceptions. To undertake a similar examination, in the absence of any experimental manipulations, and having jettisoned the most simple minded empirical distinction between slips and mistakes, we must derive more subtle empirical predictions from the cognitive mechanisms that have been postulated to underlie slips.

Two broad classes of mechanism have been put forward to explain cognitive slips, including slips of action and speech errors. The first, mentioned above, is loss of information from working memory. The second (e.g., Norman, 1981; Reason & Mycielska, 1982) is, roughly, deployment of attention, or cognitive control: Competition among the activation levels and triggering conditions of coexisting cognitive demons or schemata. These two mechanisms make quite different empirical predictions.

When slips are due to competition in the control system we would expect slips to appear relatively infrequently, but to persevere at all levels of expertise, and even to increase with expertise in some circumstances, as performance shifts into “open loop” or unmonitored modes (Reason & Mycielska, 1982).

When slips are due to losses of information from working memory, frequencies may be expected to decrease with expertise, as the associated cognitive changes will lead to lower working-memory loads. Task complexity would also be expected to play a major role: Abstract working memory is taken to be a general resource (e.g., Broadbent, 1984), and anything that in-
creases the load on this resource will increase the propensity for slips. Using exactly this argument, Anderson and Jeffries (1985) showed that Lisp errors, which increased on problems with additional "irrelevant complexity" (e.g., extra layers of brackets), must be due to losses from working memory.

Armed with these empirical predictions, turn again to the algebra errors here. Certainly, there are some cases of errors that we diagnosed as mal-rules, but found on only one occasion, which may uncontrovertially be classified as action slips. Take, for example, the error $Mx = N \rightarrow x = N/M/M$. This error has "slip" written all over it—it satisfies all the properties of control-based slips: It is very infrequent, only one subject out of 96 makes it, on only one occasion, and she is a high performer. Intuitively, also, it appears to involve a straightforward failure of control, "failure to pop the stack" in the terms of a popular control model. However, some of the other errors that occur only once are less easy to intuit being generated by either of the slip mechanisms, for example, $M(Nx + P) \rightarrow M + P - N$. With these one-off error types, however, it is impossible to discount simple writing slips, or wild guessing. We will concentrate again on the more frequent mal-rules from Tables 1, 2, and 3.

Each of the mal-rules, except M4 and M6, which, as we noted above may be regular misconceptions, occurs more frequently for students in the lower-performance schools. Consider, for example, the "change-operand" errors (M14–17), and the "change sides without changing signs" (henceforth "change-sides") errors (M9–10). Of all our mal-rules, perhaps these are intuitively the most sliplike; "forgetting" to change the sign of a term when moving it to the other side of the equation is just the kind of careless blunder that even very skilled problem solvers make (Lewis, 1981). But skilled problem solvers would certainly not be expected to make these errors with the frequency that some of our subjects did. Of the 32 students at School 1, only 1 did not make any "change-sides" errors, and the mean frequency was over 5 occurrences per student (Table 3). But in the most expert group, from School 2, these errors occurred in total only 27 times, an average of less than one per subject (Table 4).

It may be argued that the relationship between mal-rule occurrence and skill is circular: Is our index of skill not, after all, the number of errors? There are three counterarguments to this critique. First, we do have an independent index of skill in the general educational level and aspirations of the three schools. School 2 students have been selected on academic ability, they would be expected a priori to make fewer mistakes, but not fewer control slips. Second, the relationship holds for most mal-rules that, nevertheless, do occur in the most expert group (we have already considered the exceptions M4, M6). Third, the relationship does not hold as strongly for those errors we labeled as arithmetic slips. In School 1, 0.6 of the students, in School 2, 0.4 of the students, and in School 3, 0.2 of the students made arithmetic slips. In each case the average was about 2 slips per student.
We conclude that our data do provide meaningful evidence of a relationship between the frequency of each mal-rule and the skill of the student. This relationship undermines the control-slip explanation of errors.

The working-memory-failure explanation fares no better. First, as we reported in Section 3.5, we can find no evidence that increases in problem complexity increase the likelihood of errors. Again, the point is well illustrated with the most "sliplike" of our mal-rules. Consider the "change-sides" errors, M9 and M10. Table 7 shows that both these two error types occur on practically all those questions on which they could occur, on both tests. Furthermore, if we look at the question types which promote most "change-sides" errors, we see that they are simple, rather than complex.

There are severe empirical difficulties then, for classifying our errors as pure slips. Even in the absence of empirical reasons, there are theoretical reasons to reject such a simple diagnosis. A bald recourse to "undetected losses from working memory" really doesn't explain very much at all. Let us consider more closely the data of Anderson and Jeffries (1985), who use this explanation for Lisp errors. They report a substantial difference between the performance of their novices on the different functions, with far more errors being made on CONS. This difference is just as important (more important in a practical sense perhaps) than that caused by the complexity manipulations, yet it is not explained by the working-memory-slip hypothesis, for there does not seem, at first glance at least, to be any differences between the functions in terms of working-memory load (Gilmore, 1986). The point this example raises goes beyond the immediate context to attack simple working-memory-load explanations generally: To explain or predict interesting properties of errors, it will always be necessary to elaborate an architecture-based account with some theory of the mental representation of the skill. In the Anderson and Jeffries case, such a theory of representation needs to explain why the CONS procedure should place more stress on working memory than does the LIST procedure. In our case the representational theory must explain why errors occur during some algebraic manipulations more than others.

Such a theory must also explain the characteristic irregularity of our mal-rules. It seems that mal-rules might be characterized as the cooccurrence of a slip and a mistake, explained by an interaction of properties of the knowledge-base and properties of the cognitive architecture. We will sketch such an explanation below after examining another important concept, the semantic rationalization of procedures.

6.2 Symbol Manipulation Versus Semantic Rationalization
Competence in any symbolic domain, including algebra, necessarily involves the internalization, in some form, of procedures for manipulating the symbols so as to transform starting problem states into goal states. Such proce-
dures may or may not be associated with some "semantic rationalization," some account of why the manipulations are valid, in terms of the meaning of the symbols on which they operate.

In an influential article, Resnick (1982), considers this question in the domain of children’s subtraction skills, and subtraction mal-rules. She concludes that most children learn to subtract by learning "the syntactic constraints of written subtraction, without connecting them to the semantic information that underlies the algorithm" (Resnick, 1982, p. 138). Resnick’s main argument is that the most common subtraction mal-rules, considered by Brown and Burton (1978) and Young and O’Shea (1981), tend to respect syntactic constraints, but violate semantic constraints. This suggests that semantic constraints are playing no role in the skill: It is purely syntactic symbol manipulation.

This argument needs some clarification, because, by definition, errors of any kind must violate both syntactic and semantic constraints (otherwise, there must exist syntactically correct procedures that generate wrong answers, or syntactically "faulty" methods that generate right answers). Resnick appears to rely on some intuitive judgement of which violation is the larger. For example, in discussing the "smaller-from-larger" mal-rule in which children compute column subtractions without regard to "bottom-from-top" constraint, Resnick (1982, p. 139) claims that "all the syntax of written subtraction without borrowing is respected," but fails to justify her assumption that the top/bottom comparison should be treated as outside the scope of the syntax.

Whatever the formal weaknesses in Resnick’s argument, its intuitive force is clear. Many subtraction errors can be modelled by small syntactic changes to correct subtraction procedures; indeed, this is perhaps the key observation underlying the entire mal-rule enterprise. VanLehn (1983, p. 72), also working in the domain of young children’s subtraction is very clear about his acceptance of this position, and its role in his theory: "The first assumption is that student’s knowledge about procedures is schematic, but not teleological [i.e., semantically rationalized]." Other theorists are less clear, but can all be interpreted as assuming a purely syntactic explanatory framework.

Turning to algebra, we discover a similar, but importantly different story. First, the received wisdom is exactly the same:

Substantial evidence exists that algebra learning is addressed by most school-children as a problem of learning to manipulate symbols in accord with certain transformation rules, without reference to the justifications or meanings of these transformations (Resnick, et al., 1987, p. 174).

In support of this claim, however, no argument is offered to parallel that for subtraction. Instead, Resnick et al. cite published work on algebra mal-rules, especially that of Matz (1982) and Sleeman (1984) as making the case
for them. We are not convinced that this case is yet resolved in the way Resnick et al. suggest. No empirical argument has been presented in detail, and the success of VanLehn's (1983) theory of subtraction errors, which embodies this assumption, is of little help. It is quite possible that a purely syntactic account of errors will be valid for primary schoolchildren's subtraction but not for secondary school students' algebra.

In fact, the core argument, that errors can be described as small syntactic changes to correct procedures, is questioned by Sleeman, whose attempts to explain algebra mal-rules lead him to distinguish "manipulative" from "parsing" mal-rules. A manipulative mal-rule is defined as "a variant on a correct rule which has one substage either omitted or replaced by an inappropriate or incorrect operation" (p. 403). In other words, Sleeman is essentially taking Young and O'Shea's (1981) deletion model as the defining property of a subset of mal-rules. If a mal-rule can be generated by the deletion operation (or, rather more broadly, some perturbation mechanism) then it is deemed a manipulative mal-rule. A parsing mal-rule is defined more vaguely, it "arises from a profound misunderstanding of algebraic notation." (Sleeman, 1984, p. 403) Rules S7-S19 (Table 2) are the only examples given of such mal-rules.

On first reading, Sleeman's mal-rule categories seem to throw doubt on the purely syntactic model of errors. However, it must be noted exactly what theoretical work is being done and not being done by the manipulative/parsing distinction. Essentially, these names simply label those subsets of Sleeman's observed mal-rules that he believes can (manipulative) or cannot (parsing) be generated by syntactic mechanisms. The distinction does not appear to be one that can be made in any other way, and so it does not allow any independent demarcation of the generative scope of these syntactic theories.

Furthermore, although the definition of parsing mal-rules as "profound misunderstanding" might suggest a break with the assumption of purely procedural/syntactic skill models, Sleeman himself is unwilling to make this break. In discussing "mis-parsings" Sleeman (1984, p. 406) states: "These latter representations suggest that the student has failed to appreciate the semantics of algebraic expressions—and sees the solution of algebraic equations as a symbol-manipulating task." Again, it is as well to be aware of the status of this remark. The "representations" in question are not ones that we can be sure the student uses, rather they are putative parses of algebraic expressions that Sleeman proposes can be used to help explain the existence of "parsing mal-rules" like S7-S19 (Table 2).

Finally, Sleeman acknowledges that he is only able to provide intuitive arguments that his example-parsing mal-rules cannot be generated without recourse to mis-parsing. In view of these many difficulties, we may begin to wonder if the distinction between parsing mal-rules and manipulative
mal-rules has any force. This worry is somewhat confounded by Sleeman's assertion that parsing mal-rules can be explained by the mechanism of misgeneralization, which can explain many manipulative mal-rules, too. Nevertheless, we will argue below that the distinction is worth making, although by a different criterion from that used by Sleeman, and that it has greater implications for models of skill acquisition than Sleeman develops.

Before we turn again to our data, we must make two simple theoretical points about the a priori possible role of semantic rationalization in algebra skill. First, algebra skill involves many procedures or operations, and it is clearly possible that students can supply semantic accounts for some, but not all. The distinction between operations that can be accounted for (by the student) in terms of known relations between numbers, and those that are accounted for purely as steps in some goal-directed procedure bears the weight of the following examination of the role of semantics in algebra mal-rules. We would argue that semantic rationalizations for some of the "symbol manipulations" must be available to the children. Surely, all the children in our sample could justify transforming a subpattern 3 + 5 to 8 in terms of the relationship between numbers, rather than merely as a necessary step in the solution of the complete equation. On the other hand, more complex operations, such as expanding M(Nx + P) to M*Nx + M*P may well be done in the way Resnick et al. (1987) suggest, without recourse to any semantic account other than as a step toward a distant goal. Second, it is perfectly conceivable that semantic rationalizations of operations may be unused in routine smooth performance, but called up, or even invented, at impasses, and thus involved in the generation of errors.

Equipped with these ideas, we examine our error data in order to question the proposition that algebra problem solving is a purely syntactic skill. We believe that six separate observations may be most easily explained by allowing some semantic constraints on error generation.

1. A first point relies on the integration of slips and mistakes that we have argued for above. In marking scripts we noticed many protocols that were suggestive of errors of the kind that Norman (1981) labeled "capture" errors. In our protocols, capture errors appeared as the carrying out of known arithmetic computations at the expense of algebraic reductions. For example, consider the error illustrated in the following protocol:

CE:

\[
\begin{align*}
6x &= 39 - 3 (8x - 7) \\
| \\
6x &= 36 \\
| \\
x &= 6
\end{align*}
\]
Previous analysts may have regarded such errors as clear slips, and so unworthy of further attention. We diagnosed these errors as examples of M13: \(Mx + |N - [M + |N]\). As we noted in Section 3.6, over 30% of the M13 occurrences were on this one question. What is interesting about this particular equation is that, intuitively, the intermediate calculations of \(39 - 3 = 36, 6*6 = 36\) seem highly suggested by the perceptual structure of the problem. Relatively meaningless operations may be captured by such operations that are more familiar by virtue of their semantic rationalization.

2. An aspect of error data which none of the syntactic accounts of mal-rule generation can adequately address is the relative frequency of different mal-rules. For example, VanLehn (1983) has to use some impasses and some repairs more frequently than others in order to generate appropriate subtraction bugs, but "it is not always the case that combining a common repair and a common impasse results in a common bug" (p. 120).

In general, it is very difficult to analyze relative mal-rule frequencies, because of the overall low frequencies, and the differential "opportunities" that different mal-rules have to occur. However, in certain cases it is clearly possible to perform full-blooded statistical comparisons. One such case concerns Rules M12, \(Mx + |N - [M + |N]\), and M13, \(Mx + |N - [M + |N]\), which obviously can occur in exactly the same circumstances, and are both relatively frequent in our sample. Yet M12 appears 102 times in the protocols of Schools 1 and 3, whereas M13 appears only 49 times.

One bias toward Rule M12 may occur because it can happily be carried out in equations with only a single \(x\), whereas M13 would leave no \(x\)s remaining. Consequently, we compared their frequency only on those problems with multiple \(x\)s in the initial equation, and confirmed that the large preponderance of M12 is indeed statistically reliable \(F(1,50) = 6.02, p < .01\).

How can this preference be explained? It seems to us that any syntactic mechanisms are hard-pressed to explain the fact that both M12 and M13 are common, but that M12 is more common. For example, both rules can clearly be explained by a simple misgeneralization:

M12 is derived by overgeneralizing from
\(Mx + Nx - [M + N]x\)

to \(A + B - [\text{number-part-of-A} + \text{number-part-of-B}] x\)

and M13 from overgeneralizing
\(M + N - [M + N]\)

to \(A + B - [\text{Number-part-of-A} + \text{number-part-of-B}]\)
In order to adequately explain why $M_{12}$ is more popular, we need some account that will capture the intuition that it is a “more sensible” mistake. One such account that is available in the literature is that students are using an analogy from arithmetic expressions such as $3\frac{1}{2} + 1 = 4\frac{1}{2}$ (Matz, 1982) or from natural language constructions, like “three apples plus four gives seven apples.” Note that there are two ways in which such an analogy could be constructed. It could be based on the syntactic similarity of the expressions, in which case it would be indistinguishable from the misgeneralization explanation of the mal-rule that we have just seen, or it could be based on the construction by the student of a semantics for the notation. There is no way to resolve this competition empirically from our data, but there are two reasons for admitting the meaning-based account. First, it gives us a ready explanation for the preponderance of $M_{12}$ over $M_{13}$: There are two sources for the first misgeneralization, and only one for the second. Second, if the analogy is done on purely syntactic grounds, why should we not get analogies from expressions like $3x$ to numbers like $34$? These would emerge as mal-rules of the form $Mx = M \times 10 + x$, which we did not observe, and have never been reported.

Another interesting aspect of mal-rules $M_{12}$ and $M_{13}$ is that they are not merely incorrect transformations, they are incorrect transformations of a pattern that should not be reduced. That is, they reflect inappropriate parsing of the equation into operable subpatterns. In our mal-rule set, $M_3$ is the only other clear example of a mis-parsing of this nature. It seems that these three rules represent a small, but genuine subset of parsing mal-rules, but one that is defined by a tighter criterion than that used by Sleeman. It is not the case that all Sleeman’s “parsing errors” represent transformations of inappropriate subpatterns. For example, $S_{7}-S_{9}$ transform $M \times x + N \times x$, and $S_{16}$ transforms the pattern $M \times x = N \times P$, both of which can indeed be properly reduced. By our criterion, only mal-rules $S_{10}-S_{15}$ are parsing mal-rules, and all but one of these transform the same pattern as $S_{12}$ and $S_{13}$.

That so few different mis-parsings are encountered does not sit happily with the characterization of algebra skill as purely syntactic. It seems that semantic knowledge is influencing students’ parsing. In the case of $M_{12}$, we have already seen how semantic knowledge may lead students to transform $Mx + N$, even though this involves an inappropriate parse. Furthermore, as we noted in Section 3.4, both $M_{12}$ and $M_{13}$ occur very frequently in bracket-manipulation questions. Inspecting these questions in Table 1, we can see that in each case, at the start state there is no competing semantically rationalized operation to perform. In contrast, the other questions in which the pattern $Mx + N$ exists, each allows the “change-sides” operation to be done first. The lack of com-
petition from meaningful rules may encourage the firing of parsing mal-rules.

In the case of M3, we can see that a semantically rationalized operation, \( M + | - N \rightarrow [M + | - N] \), is carried out in a context where it is not appropriate. This appears to be another example of a meaningful operation competing with and being preferred to a meaningless symbolic manipulation.

4. Mal-rule M7, ignoring brackets and their contents, appears, on the face of it to be an absurd error. It can be explained by the combination of two semantic influences we have already discussed. First, it replaces a rule for the expansion of brackets which is complex and presumably not semantically rationalized by many students. Second, it is supported by an analogy with written English in which bracketed terms can be ignored at no great cost.

5. As we argued in Section 4, some of the error patterns we observed in the "idiosyncratic" students reflect semantic processing. The clearest example comes from students who provide two answers for \( x \) in equations that have two \( x \) terms. The mistake these students are making is to attribute the wrong meaning to the equations, rather than to mis-manipulate symbols. Indeed, according to their meaning for the equations, their solutions are correct.

6. Finally, we noted a number of students using the = sign to denote the derivation of one problem state (i.e., equation) from another (as reported in Section 4). These students are extending the meaning of a symbol which has only been shown to them in a related but narrower role. This kind of misgeneralization is outside the scope of syntactic misgeneralization.

In summary, then, we believe that several observations weigh against the view that all algebraic problem solving is purely syntactic symbol manipulation. None of these observations offers unambiguous confirmation of the role of semantic processes, but together they surely suggest that the dismissal of these processes was premature.

It is a challenge for further research to establish the exact relationship between syntactic and semantic processing in algebra and in other symbolic tasks. In the next section, we sketch a model of algebra skill that casts semantic rationalization in two main roles, and in so doing helps explain the distribution of errors in our data that proved so difficult to capture with existing accounts of mistakes or slips.

6.3 Some Theoretical Implications
The first implication of our data concerns the nature of competence in symbolic domains. The slip/mistake distinction relies on a view of competence that is all-or-nothing, either students know the right rule but slip in its exe-
ution, or they do not know the right rule, and must perform local problem solving, or apply an incorrect version of the rule. The breakdown of the slip/mistake distinction we have noted suggests that this all-or-nothing view must be refined. In fact, it can readily be refined within the paradigm of rule-based models, by allowing the simultaneous representation of alternative rules (correct and incorrect) that apply in the same situation, and using some notion of rule strength to resolve conflicts. This is a common and important aspect of several recent production-rule accounts of cognitive skill and learning (e.g., Anderson, 1983; Holland, Holyoak, Nisbett, & Thagard, 1986) but has had little impact on error explanation, with the exception of Langley's (1985) SAGE system. It appears to be well-equipped to model aspects of our algebra data, allowing errors that are neither pure slips nor pure mistakes, and addressing the rampant irregularity we observed in students' behaviors. In the competing-rules model, errors are possible because faulty rules are represented, and in this sense are knowledge-based mistakes. But errors arise only when weaker, faulty rules are preferred to correct, stronger rules. Noise in the matching system could generate these misfirings, and in this sense the errors are architecture-based slips.

Under this account, the decrease in errors from novices to experts depends on the weakening of the strengths of the incorrect rules, and/or the increasing strength of correct rules. Remember that we do not want to suggest that "classical" slips or mistakes can be done away with altogether. Our framework models pure slips as misfirings of correct rules due to working memory losses. Pure mistakes, explicable entirely in terms of the knowledge base, occur when incorrect versions of rules have more strength than correct versions, or when the correct versions do not exist. The analysis of our data in Section 6.1 suggests that pure mistakes do not exist among our errors.

There are several mysteries that must be solved before the competing-rules model can become an adequate account of errors. First, we need to describe where the incorrect versions of the rules come from. Second, we need to describe some mechanisms for strength assignment and revision. Why, for example, should all students maintain alternative forms of the "change-sides" operations (i.e., M9, M10), which for every student are weaker (less likely to fire) than the correct rule?

We contend that the mechanisms existing in the literature need to be supplemented by some novel mechanisms to answer these questions. First, consider the generation, and remembering of flawed rules. As we have noted, the mistake-generating mechanisms of misgeneralization and repair have difficulty predicting the development of novel, incorrect rules in problem solvers who already know the correct versions. Yet this kind of "unlearning" seems essential to explain our data.

We suggest two mechanisms that may explain this phenomenon. First, deletion, the simple mechanism put forward by Young and O'Shea (1981) in their production-rule model of subtraction errors. As its name suggests, this
mechanism simply deletes elements of the condition and/or action specifications of a production rule. It is easy to see how M9 and M10 could be generated by this mechanism. What is much less easy to understand is why deletion should lead to an internalized mal-rule, rather than a run-time error (i.e., slip). Indeed, Young and O'Shea's article is ambiguous as to whether they view deletion as a mechanism explaining the generation of errors, or simply describing them.

We suggest that deletion-based mal-rules may become internalized in the following way. First, deletion does its work at run time, producing infrequent but regular slips. Second, the student acquires the mal-rule by induction: The input/output pattern generated by the "deleted" rule is described and remembered as a mal-rule.

This inductive redescription process is intuitively and computationally plausible, but it demands a learning mechanism that is unconsidered in the psychological or machine-learning literature. We are suggesting that learning can be based on the external product of initial encodings, and can proceed without exploiting (mutating) these initial encodings (although, of course we certainly have no wish to claim that mutative learning does not occur).

Many separate learning mechanisms have been considered in research on self-modifying productions systems, including generalization and specialization of rule conditions, composition of productions, and compilation of declarative knowledge into special-purpose productions (see the review by Neches, Langley & Klahr, 1987). All these mechanisms rely on the mutation of existing rules in order to create new rules. Chunking in the SOAR architecture (Laird, Rosenbloom, & Newell, 1986) can create completely new productions by encoding the results of problem search, and so, we believe, can Neches's (1987) heuristic procedure modification, in this case by inspecting a trace of a previous rule-based problem-solving episode. However, like the other methods, these approaches rely entirely upon internal mental structures, and are unable to learn from patterns in the external environment of the system. Machine learning has explored techniques for extracting patterns from the environment, both within the rule-based tradition (e.g., Neves, 1978) and outside it (e.g., Rumelhart & McClelland, 1986); but, to our knowledge, these processes have never been directed to the agent's own outputs.

Several workers have recently recognized that production-system models fall short of an account of interaction with the external world (e.g., Laird, Newell, & Rosenbloom, 1987; Larkin, 1989; Young & Simon, 1987). In the best developed counterproposal, Larkin (1989), using algebra as an example task, offers a model of "display-based problem solving," in which the external display aids the problem solver by obviating the need to maintain problem states and goal stacks. Our suggestion is that the display may play a role in
learning as well as in problem solving. As well as supporting the internalization of slips that seems necessary to explain our data, inductive redescription is a candidate addition to the repertoire of useful learning techniques: It can apply in any situation where an external record of problem states exists, and where more than one rule can describe the same transformations.

Rule induction from the written input and output of productions perturbed at run time, provides one plausible origin for mal-rules alongside correct versions. We don't believe it is a sufficient explanation, because several of our mal-rules do not appear to be small perturbations of correct forms. A second origin is also suggested by our data: New (mal-)rules may arise when students attempt to make sense of currently purely syntactic rules. We have already considered some general evidence for the role of semantics in error generation, and now we simply wish to note the theoretical role such processes could play in explaining the co-occurrence of correct and mal-rules; our discussion of M7, ignoring brackets and their contents, provides the clearest example of our suggestion.

The second problem a competing-rules framework must address is strength assignment and revision. Our mistake + slip explanation for mal-rules insists that the most regularly occurring mal-rules must be the strongest relative to their correct versions (although we have already argued that each mal-rule must be weaker than its correct sibling, for all subjects). It seems plausible to us that use is a major determinant of strength of algebra rules. Many systems in the literature use more refined strength revision metrics, relying not only on use but on creditworthy use (e.g., Holland et al., 1986; Langley, 1985). In the case of school algebra such a mechanism may also apply when firing mal-rules produces obviously erroneous outcomes, but in many classroom situations erroneous outcomes will not be noticed, and teacher marking may occur too late, and in too little detail to support refined credit assignment. The same will be true of any skill that is practiced in the absence of rapid feedback.

A use-based strength revision system will produce a snowball effect: The stronger a mal-rule becomes, the more it is used and so the stronger it becomes. This effect can explain the markedly skewed frequency distribution of mal-rules we have observed. However, we have also reported evidence for a secondary influence on the strength of mal-rules. In Section 6.2 we reported several examples of presumed meaningful operators being preferred in competition to presumed meaningless ones (e.g., in parsing mal-rules). This is easily accommodated in our system by allowing semantic rationalizations to increase the strength of some rules relative to others. Such a scheme dovetails with the proposed role of semantics in provoking the generation of new rules.

To summarize, in this section we have attempted to support the following theoretical framework for explaining algebra errors:
Multiple competing rules representing correct and mal-versions of "the same" transformation.

Strength of rules determines firing, but not perfectly: Weaker rules sometimes fire. This is the main cause of errors.

Strength is determined by rule use and by semantic rationalization of rules.

New rules can be generated even when correct rules already exist by (a) pattern induction from the products of existing rules, including slipped usages, and (b) trying to make sense of transformations.

7. CONCLUSIONS

We have observed several important properties of algebra mal-rules:

- The frequency of mal-rules is severely skewed. Most mal-rules occur very infrequently, a few occur frequently.
- Different mal-rules have explanatory power in different schools. Many of our most powerful mal-rules have not been previously reported.
- Diagnosis of errors as mal-rules is more successful for students who make fewer errors overall.
- Mal-rules are severely unstable. It is very rare that a student "uses" a mal rule in as many as half of its applicable conditions.
- A few "idiosyncratic" students make errors which, while exhibiting some pattern, are not readily described as mal-rules.
- Students are able to tell whether they are getting answers right or wrong, but confidence does not allow the partitioning of mal-rules into a subset that is associated with high-confidence solutions. All the most powerful mal-rules in our study appeared in the high-confidence solutions of at least one student.

The theme of these separate observations is that algebra performance is very unsystematic. This theme clashes with earlier work by Matz (1982), and Sleeman (1984, 1985) who focussed on regularities, but chimes with that of Greeno et al. (1985) who reported erratic performance of children during their first course in algebra. The current study demonstrates that aspects of algebra performance remain highly unsystematic at later stages in learning (the students in our studies had all progressed to more advanced algebra classes, and averaged higher than 50% on our tests).

The lack of systematicity in algebra errors poses challenges to current cognitive models of skills and error, as noted by Greeno et al. (1985). In Section 6 we exposed difficulties with two foundational assumptions of current theories of cognitive error: Many of our errors defy classification as slips or mistakes: some errors defy classification as purely syntactic deformations of purely syntactic procedures.
An attempt to meet these challenges was made within the framework of production system architectures in which alternative versions of rules compete for control of cognition, and in which a dynamic index of rule strength helps determine the competition. Mal-rules may be created, we argued, by the internalization of action slips as new rules, induced from the external display of algebra equations written by the student. Furthermore, our results suggest an intimate relation between procedural and conceptual knowledge, which is more complex than is recognized in current models, but which may be accommodated in the competing-rules framework by allowing sense-making procedures to generate new rules, and semantic rationalizations to increase rule strength.

APPENDIX
ALGEBRA PROBLEMS USED IN 3 STUDIES

<table>
<thead>
<tr>
<th>Test A</th>
<th>Test B</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. 2x + 5 = 9</td>
<td>29. 9x = 6 + 3(2x + 4)</td>
</tr>
<tr>
<td>2. 9x = 11 + 7</td>
<td>30. 3x = 7x - 48</td>
</tr>
<tr>
<td>3. 6x = 39 - 3 (8x - 7)</td>
<td>31. 5x = 2 (4x - 9)</td>
</tr>
<tr>
<td>4. 7x = 11 (2x - 15)</td>
<td>32. 6x = 4x + 12</td>
</tr>
<tr>
<td>5. 5x = 8x - 6</td>
<td>33. 10x - 3x = 49</td>
</tr>
<tr>
<td>6. x/4 = 3</td>
<td>34. 12 - 2x - 3x = 2</td>
</tr>
<tr>
<td>7. 3x + 2 (x + 5) = 45</td>
<td>35. 2x + 8x = 20</td>
</tr>
<tr>
<td>8. 5x = (4*10)/2</td>
<td>36. 5x = 40</td>
</tr>
<tr>
<td>9. 7x - 6 + (11*2)</td>
<td>37. 3x + 4 - 2x + 8</td>
</tr>
<tr>
<td>10. 12x = 68 - 8</td>
<td>38. 4x - 7 = 25</td>
</tr>
<tr>
<td>11. 5x - 14 = 31</td>
<td>39. 6x = 90/3</td>
</tr>
<tr>
<td>12. 9x = 2(3x + 9)</td>
<td>40. 9x + 13 = 85</td>
</tr>
<tr>
<td>13. 2x + 8 = 3x + 2</td>
<td>41. 8x = 60 - 4 (2x - 5)</td>
</tr>
<tr>
<td>14. 6x - 4x = 10</td>
<td>42. 6x = 12 + 24</td>
</tr>
<tr>
<td>15. 4x = 2*8</td>
<td>43. 4 + 3x + 5x = 60</td>
</tr>
<tr>
<td>16. 7x = 8 + 3(2x + )</td>
<td>44. 4x + 3 (x + 2) = 18</td>
</tr>
<tr>
<td>17. 72 - 5x - 2x = 23</td>
<td>45. 8x - 3 (x - 1) = 18</td>
</tr>
<tr>
<td>18. 12 - 11x = 100</td>
<td>46. 3x = 2*6</td>
</tr>
<tr>
<td>19. 6x = 7(3*4)</td>
<td>47. 30 - 2x = 16</td>
</tr>
<tr>
<td>20. 9x = 2x + 49</td>
<td>48. 6x = 3 (2*4)</td>
</tr>
<tr>
<td>21. 8x = 32</td>
<td>49. 4x = 26 - (2*5)</td>
</tr>
<tr>
<td>22. 5x - 6(x - 3) = 11</td>
<td>50. x/6 = 7</td>
</tr>
<tr>
<td>23. 21 - 6x - 3</td>
<td>51. 10x = 3 (2x + 4)</td>
</tr>
<tr>
<td>24. 5x = 37 - (3*4)</td>
<td>52. 2x = (6*8)/3</td>
</tr>
<tr>
<td>25. 9 - 3x = 15 - 4x</td>
<td>53. 3x = 1 + (2*4)</td>
</tr>
<tr>
<td>26. 7 + 10x + 8x = 97</td>
<td>54. 4 - x = 6 - 3x</td>
</tr>
<tr>
<td>27. 2x = 24/4</td>
<td>55. 4x = 18 - 6</td>
</tr>
<tr>
<td>28. 6x + 3x = 36</td>
<td>56. 6 + 4x = 34</td>
</tr>
</tbody>
</table>
REFERENCES


