COGNITIVE SCIENCE 8, 413–437 (1984)

Error Detection Processes in Statistical Problem Solving

CARL MARTIN ALLWOOD
University of Göteborg

Problem solvers' error detection processes were studied by instructing 16 subjects to think aloud when solving two statistical problems. The evaluative episodes occurring in subjects' protocols were analyzed into Affirmative evaluation, Direct error-hypotheses, Error suspicion, and Standard check episodes, the last three of which are assumed to cover all main types of error detection processes. Most errors (78%) were found to have contributed to a solution part that triggered some evaluative episode. However, only one-third of the undetected errors had contributed to such a solution part. The Standard check episodes, seen as centrally-invoked, only led to the detection of few errors in proportion to the number of times they were performed. Evidence was found for two types of spontaneous error detections, one occurring abruptly and the other as a result of a more elaborated error detection process, initiated by the solver perceiving the solution as dissatisfying or strange. The perception of a symptom was a fairly reliable source of information about errors. However, subjects often did not manage to detect the error after having noticed a symptom. The closer a relevant Error suspicion episode followed an error, the greater was the probability of detecting the error. Good problem solvers detected a higher proportion of their errors compared to poor problem solvers, probably due to differences in the processes leading up to the triggering of an evaluative episode rather than to differences after the episode had been triggered.

Whenever a person tries to solve a problem, he or she has different types of knowledge resources available. In order to be as successful as possible, the problem solver should utilize these resources in a way that is advantageous for solving the problem. The problem solver's way of utilizing the knowledge resources can be thought of as the technique factor of his or her problem

Acknowledgement: I wish to thank Henry Montgomery and two anonymous reviewers for constructive criticism of the manuscript and Anders Oden for statistical consultation. Donald Norman provided many helpful suggestions as to the improvement of the manuscript.

Correspondence and requests for reprints should be sent to the author at the Department of Psychology, University of Göteborg, Box 14158, S-40020 Göteborg, Sweden.

solving. Correspondingly, we can think of the knowledge resources available to a particular problem solver as the knowledge factor of his or her problem solving (Allwood, 1976). Thus, the end result of a person's problem solving is determined by the combination of these two factors in interaction with the problem. The same combination of factors will also determine the problem solver's success in detecting any errors that have been made in the solution process.

This paper explores how people detect their own errors. The approach taken is to see the detection of errors as part of the larger context of problem solving processes. More specifically, error detection is seen as being a possible end result of certain kinds of evaluative processes.

From a general perspective, problem solvers' activities are assumed to be of two types: namely, progressive activities and evaluative activities. A problem solver is in a progressive phase when he or she works directly toward the goal state of the problem. By contrast, the problem solver is in an evaluative phase when he or she evaluates some already performed part of the problem solution. The evaluations may be either affirmative, i.e., assuming that the solution is correct, or negative, i.e., for some reason heeding the possibility that the solution is erroneous. Negative evaluations are assumed to make up the set of error detection processes. By definition, error detecting always occurs in negative evaluation episodes.

Any error detection process may be divided into two parts, namely:

1. triggering of the error detection mechanism, i.e., the initiation of a negative evaluation process; and
2. later steps taken in the evaluation process. These steps may eventually include discovery and elimination of an error.

Obviously, an error detection process may be triggered even if an error has not been made. Thus, it is assumed that the triggering of the error detection mechanism is independent of whether an error actually exists or not.

Only a few studies on the detection of one's own errors have been performed. These studies were done with regard to specific activities or types of errors such as writing (Hayes & Flower, 1980), reading (Carpenter & Daneman, 1981), slips (actions deviating from a held intention) (Norman, 1981), speech (Nooteboom, 1980), statistical problem-solving (Allwood & Montgomery, 1981, 1982). Still, the results of these studies can be compared with respect to different phases of the error detection process. Such a comparison is made in this paper. It should be noted that the cited authors do not describe error detection processes in the way proposed here.

The internal rejection of possible erroneous outcomes has also been discussed in the literature (e.g., for speech production Baars, 1980; Laver, 1980). However, since the present paper deals only with error detection pro-

cess in relation to overtly occurring errors these studies will only be briefly mentioned.

Triggering of error detection processes. There are two basic ways that error detection processes may be triggered. First, they may be invoked centrally irrespective of the specific content of the produced solution. Such processes have been suggested by Hayes and Flower (1980), Sussman (1975), and Allwood and Montgomery (1982). Second, error detection processes may be invoked by spontaneous triggering due to some feature of the performed activity or produced result (Allwood & Montgomery, 1982; Carpenter & Daneman, 1981; Hayes & Flower, 1980; Nooteboom, 1980; Norman 1981; Sussman, 1975).

It is of interest to consider more specifically why the spontaneously triggered error detection processes are invoked. Hayes and Flower (1980) have suggested that error detection is initiated when a match occurs in short-term memory between stored representations of specific errors and the same error made in the emitted performance. Sussman (1975) argues along the same line, suggesting that a few generally described specific situations will trigger error detection processes. An example of such a situation is when a mental operator is called upon to act without its application conditions being fulfilled. The idea of one or many monitors detecting certain classes of errors, which has been suggested by Laver (1980) and Norman (1981) also seems to belong in this category.

In this study we argue that error detection processes are also triggered when the problem solver perceives a discrepancy between the activity or result he or she has produced and one or more of his or her expectations about the activity or result. This notion can be seen as a generalization of results presented by Carpenter and Daneman (1981) and by Baars (1980). Carpenter and Daneman presented evidence that in reading, error detection processes are triggered because of a perceived inconsistency between an earlier interpretation of the text and new information read. In a study of covert errors in speech production, Baars (1980) argued that "...certain error outcomes will be suppressed on the basis of contextual expectations" (p. 313).

Diagnosis in the error detection process. When an error detection process has been triggered, the problem solver has the possibility of performing a diagnosis as to the exact nature of the possible error. In Hayes and Flower's (1980) model, the diagnosis is completed and the error is fully specified through the activation of the error detection process, i.e., through the triggering of the error production. In Sussman's system (1975), the error is only vaguely described by the error signal. Later in the process, further diagnosis always occurs. Carpenter and Daneman's (1981) results show that in read-

ing, further error diagnosis is not obligatory but is simply one of several error recovery heuristics for which these authors found evidence.

Correction of errors in the error detection process. As a result of the error detection process, the solver may or may not correct an error. Sussman (1975) proposed that if the diagnosis succeeds in specifying the error then the correct solution is automatically supplied from memory. This is obviously not valid for all human error detection. Most of us have probably experienced the frustration of not being able to supply the correct solution even in cases where an error has been fully specified. Furthermore, Allwood and Montgomery (1982) report data from a problem-solving study where a subject had localized an error exactly without providing a new solution. Still, Flower and Hayes (1980) appear to take the same line as Sussman (1975) when they write that the action side of their error detection processes contains "a procedure for fixing the fault" (p. 17). Neither does Norman (1981) note any difficulties in the correction of a detected error, but with slips this position seems reasonable since slips presumably only occur in connection with knowledge well-known to the subject.

The present paper deals with students' error detection when solving problems demanding specialized knowledge. The subjects were asked to verbalize aloud all their thoughts when attempting to solve two statistical problems. The paper reports on subjects' error detection processes with regard to how these processes are triggered and with regard to the steps and measures taken by the subjects in these processes. In contrast to other studies on error detection the present study analyzes all negative evaluative processes occuring in the problem-solving process, not just those leading to the detection of errors. The present paper also studies the relation between the subjects' degree of problem-solving proficiency and some aspects of their error detection behavior. An analysis is also made of variations in how different types of errors are detected.

Subjects

Seventeen subjects (13 male, 4 female) studying first year statistics at the University of Göteborg, Sweden, were paid for their participation in the study. One female subject was deleted due to the sparsity of her think-aloud report. Eleven subjects had low high school grades in mathematics and had followed a nonscientific (i.e., little mathematics) syllabus, whereas the remaining five subjects had high grades in mathematics and had pursued a scientific (i.e., a great deal of mathematics) syllabus. A few days before participation in this study, all subjects had taken an examination on a course covering material relevant to the problems given to the subjects.

Procedure

Each subject was asked to solve two statistical problems as described below. All experimental sessions were conducted on an individual basis. Each subject was requested to think aloud while attempting to solve the problems and was instructed to vocalize all thoughts including those that seemed unimportant. If a subject remained silent for more than half a minute while working on a problem, he or she was asked by the experimenter to verbalize his or her thoughts. Subjects were allowed the use of a pen, a calculator for elementary arithmetic calculations, writing paper and a booklet containing statistical formulae and tables.

Problems

The two problems given to the subjects were one time-series problem and one regression problem. The problems were worded as follows (in translation from Swedish):

Problem 1. "The following figures show the sales in million Sw.Kr. each quarter during the years 1971-1973.

	I	II	III	IV
1971	13	12	8	25
1972	24	21	17	45
1973	46	34	26	85

Use a suitable model and analyze the development of the sales. Carry out the analysis by calculating seasonal indices and the sales values adjusted for seasonal influence."

Problem 2. "In an agricultural experiment, the way in which the amount of fertilizer affected the size of the wheat harvest was studied. The following observations were collected.

Wheat harvest kg/100 M²	Fertilizer kg/100 m²
40	100
45	200
50	300
65	400
70	500
70	600

Calculate the regression line of kilo harvest on kilo fertilizer and a measure of the linear correlation. The standard deviations for kilo harvest and kilo fertilizer have been calculated to 13.20 and 187.08, respectively.

How much on the average will the yield increase for each kilo of fertilizer used? How big a harvest can one expect if one uses 350 kg of fertilizer (per 100 m^2)?''

The correct solutions to both of these problems, partitioned into six and seven main substeps for the two problems respectively, are given in the Appendix.

RESULTS

The raw data used in the present analysis consists of transcriptions of tape-recorded think-aloud reports given by subjects as they solved the problems described above. Subjects' written solutions to each of the problems were also used.

The time used by each subject for Problem 1 ranged from 18 to 77 min, with a mean of 48 min. The corresponding values for Problem 2 were 22 to 64 min, with a mean of 36 min.

Subjects' think-aloud protocols were analyzed with the help of various coding systems. The interjudge reliability for a particular coding system was computed by dividing the number of identical codings from two independent judges with the total number of codings made by one coder. In order to obtain a measure of the interjudge reliability for a particular coding system, subjects were randomly selected until about 25% of all items to be coded with that particular coding system had been picked out. The items for the selected subjects were then judged by the two coders. As the reliabilities for the coding systems used vary from .83 to .92, the median being .88, the reliabilities can be regarded as acceptable.

As a preliminary to further analyses, all problem-solving errors (except rounding errors) for which evidence existed in the data analysed, were identified. Errors detected by the subjects and errors made within larger errors were also counted. An example of the latter type of error are errors made when inserting values into an erroneous formula. All in all, the subjects made a total of 327 errors.

Next, the 327 identified errors were classified as belonging to one of the following categories.

Execution errors. The majority of the errors made, 202, were classified as belonging to this category. Execution errors are defined as low level errors such as computation errors, errors where a part of the calculation is done erroneously after it has already been performed correctly once, copying er-

rors and errors such as hitting the wrong key on the calculator (see also Matz, 1979).

Solution method errors. Sixty-seven errors consisted of an erroneous selection or specification of solution method for one of the main substeps in the problem. (Correct solution methods for the main substeps are given in the Appendix.) This class also includes errors made when using transformations in order to simplify numerical calculations. If an error could be classified both as execution error and solution method error, it was counted as an execution error. An example of an error in the present class occurred when a subject calculated the sales values adjusted for seasonal influence (Problem 1) by multiplying the observed values with the corresponding seasonal value. (The correct solution method is to divide the observed values by the sales values.)

Higher level mathematical errors. This class consists of mathematical errors involving more advanced mathematics than simple arithmetical computations. Examples of errors in this category are errors made when performing algebraic manipulations of equations and errors in the use of logarithms. Sixteen of the 327 errors were coded into this category.

Other types of errors. Five errors consisted of making a mistake in relation to the functioning of the calculator when combining operations on the calculator. Eight errors consisted of wrong labelling, such as calling $S \times I$ values S values.

Skip errors. This class of error occurs when the problem solver forgets or does not finish one of the main substeps (see Appendix) in the problems. Twenty-nine such errors were made by the subjects.

Error Detection Processes and Subjects' Errors

Distribution of evaluative episodes. The evaluative episodes in subjects' protocols were classified into different types. This was done in order to arrive at a better understanding of the context in which subjects detected their errors. All parts of subjects' protocols that started with some type of evaluation or control statement were included in the analysis. An evaluative episode was considered to be terminated when subjects developed their solution further than it had been developed when the evaluative episode in question started. The following four coding categories were used in the analysis. The last three categories attempt to cover all negative evaluation episodes and thus by assumption all error detection processes.

1. *Affirmative evaluation.* These episodes consist of an affirmative evaluation of the correctness of the solution. For example, a subject calculated the expected harvest for 350 kg of fertilizer to 56.58 and said: "56.58 on 350, yes that seems to be reasonable, it is right in between here, yes it should be like this." When an affirmative evaluation occurred within one of the other three types of evaluative episodes, it was not coded into the present category.
2. *Standard check.* These episodes are initiated when the subject, independently of specific properties of the produced solution, decides to perform some kind of general security check (i.e., not because the subject notes some feature of the performed solution as strange). For example, a subject having added some terms, said: "356, the first time anyway. Then I will add them once more, in the other direction so that I won't do the same kind of error anyway...(adds the terms again)."
3. *Direct error-hypotheses formation* (for short, Direct error-hypotheses). These episodes are initiated by an abrupt (presumed) detection or correction of an error. For example, one subject said "Y = 50 plus...minus it should be." Another example is a subject who when dividing observed values by trend values (Problem 1), switched the two values around and said "...observed value 8, and the trend value 25, now I made a mistake, eh, 15, yes exactly, it is supposed to be the other way around." To repeat, the Direct error-hypotheses need not occur immediately after the error is made but can occur later in the solution process. Nor need they detect any actual errors.
4. *Error suspicion.* These episodes are initiated when a subject notes some feature of the performed solution as strange or unexpected in a negative sense. For example, a subject who due to earlier miscalculations had calculated the value of the correlation coefficient to be 10, initiated an episode classified as an Error suspicion episode with the statement "That is impossible...." An other example occurred when a subject in Problem 1 had divided 26 by 53,72 and multiplied it by 100. Having correctly arrived at the answer 48.4 the subject exclaimed: "Oh! A terrible oscillation, 48.4 it is. ...". After this, the subject attempted to find the presumed error by redoing the division. The difference between Direct error-hypotheses and Error suspicion episodes is that in the former, subjects' initial comment always related to a specific (possibly only presumed) error whereas in the Error suspicion episodes it concerned some property of the produced solution without directly diagnosing a specific error.

Of the 413 evaluative episodes classified, 38% were analyzed as Direct error-hypotheses episodes, 36% as Error suspicion episodes, 18% as Stan-

dard check episodes and 7% as Affirmative evaluation episodes. The bottom row of Table I shows the frequency of each type of evaluative episode. Since the Affirmative evaluation episodes are not considered to be examples of error detection processes, they will only be sparsely commented on here.

Errors and the evaluative episodes. Next we consider an analysis which determines how many of each type of evaluative episode were triggered by an erroneous solution part. A solution part is regarded as erroneous if an error has contributed to it. Table I shows that only 9 of the Standard check episodes (12%) concerned an erroneous solution part. In contrast, 99 of the Error suspicion episodes (66%) were triggered by an erroneous solution part, $p < .01$, $N = 15$, Wilcoxon matched-pairs signed-ranks test (also used for the significances reported below unless otherwise stated). An even higher percentage of the Direct error-hypotheses episodes were triggered by erroneous solution parts (95%). No errors were detected in episodes triggered by correct solution parts.

Since not all erroneous solution parts triggered evaluative episodes, it is of interest to see exactly which errors had contributed to those solution parts that did trigger an evaluative episode. The errors that had contributed to such a solution part are called relevant to the triggered episode. Table II shows how many errors of each error type were relevant to the various kinds of evaluative episodes. All in all, 224 (69%) of the errors made were detected (i.e., either discovered or corrected). The great majority of all errors, 254 (78%), were relevant to at least some evaluative episode. However, this was the case for only 30 (30%) of the undetected errors.

Comparison between types of errors. Table II shows for each error type and episode type whether the errors were discovered or corrected. Subjects detected 176 (87%) of the execution errors and 34 (52%) of the solution method errors (Table II). More specifically, 64% of the execution errors and 23% of the solution method errors were detected in the Direct error-hypotheses episodes, $p < .01$, $N = 14$. In contrast, an approximately equal

TABLE I
Number of Different Types of Evaluative Episodes

	Affirmative Evaluation Episodes	*Direct Error-Hypotheses Episodes*	*Standard Check Episodes*	*Error Suspicion Episodes*
Triggered by erroneous solution part	6	149	9	99
Triggered by correct solution part	24	8	67	51
Total	30	157	76	150

TABLE II
Number of Errors Relevant to Different Types of Evaluative Episodes Divided into Errors Corrected, Discovered and not Discovered for Each Type of Evaluative Episode

	Execution Errors	*Solution Method Errors*	*Higher Level Math. Errors*	*Other Types of Errors*	*Skip Errors*
Direct error-hypotheses					
Erroneous new alternative	6	0	0	0	0
Correct new alternative	123	15	2	3	0
Error suspicion					
Erroneous new alternative	2	2	4	0	0
Correct new alternative	42	15	2	3	0
No discovery	11	12	5	0	0
Standard check [a]					
Correct new alternative	3	2	0	0	0
No discovery	0	1	0	0	1
Not relevant to any evaluative episode					
No discovery	15	19 [b]	3	7	28
Total	202	66 [c]	16	13	29

[a] Two further Standard checks led to Error suspicion and the discovery of six errors (listed under Error suspicion).
[b] One error relevant to an Affirmative evaluation episode. The five remaining errors relevant to an Affirmative evaluation episode were in addition relevant to some other evaluative episode.
[c] One error excluded in this and the following analyses since the experimenter aided the subject in this case.

proportion of the two error types were detected in the Error suspicion episodes (22% compared to 26%). However, it also deserves attention that a large proportion (41%) of the solution method errors relevant to some Error suspicion episode were not detected. This can be compared to the corresponding proportion of execution errors (20%), $p < .10$, $N = 5$, $T = 0$.

Nearly 40% of the errors not relevant to any evaluative episode were skip errors, none of which (except one) were relevant to any evaluative episode. It is also noteworthy that 7% of the execution errors and 29% of the solution method errors were not relevant to any of the evaluative episodes, $p < .01$, $N = 15$.

Error Suspicion Episodes

The Error suspicion episodes are of special interest since they provide a general and flexible way of detecting errors. They will now be considered in more detail.

Triggering of Error suspicion episodes. Since one-third of the Error suspicion episodes were triggered by correct solution parts, it appears that the features which cause the Error suspicion episodes to trigger do not necessarily have anything to do with errors made. Instead, the properties which case the Error suspicion episodes to trigger should be described in a more general way. Accordingly, it is of interest to analyze what caused these episodes to trigger. The following types of triggering were found in the data.

1. *Reaction to the value of a result.* Such a reaction might occur because some result value differs from other calculated values in some way. A result value might also trigger an Error suspicion episode because of its general level or because it falls outside of the normal or allowed for range of a variable. A value might even trigger an Error suspicion episode because of its unusual appearance, for example, the number 1666 triggered an Error suspicion in one subject, apparently because of the three identical figures.
2. *Insecurity with the method used to solve the problem.* This category was used, for example, when a subject said he didn't recognize the method he had used to solve a particular substep.
3. *Other reasons.* This category was used when none of the other two categories applied.

Table III shows that there was no difference between Error suspicion episodes triggered by correct or incorrect solution parts with respect to how they were triggered. The triggering of the great majority of all Error suspicion episodes was due to a reaction to a numerical value.

Activities in the Error suspicion episodes. In this section we consider some aspects of what the subjects actually did in the Error suspicion episodes.

TABLE III
Number of Error Suspicion Episodes Triggered in Two Different Contexts by Different Causes

	Triggering Due to Value	*Triggering Due to Insecurity About Solution Method*	*Triggering Due to Other Causes*
Triggered by erroneous solution part	78	12	9
Triggered by correct solution part	33	10	8

Since a number of different activities usually took place within an episode, each episode was, for coding purposes, divided into smaller segments. An attempt was made to create segments to which only one coding category would be applicable, although specific segments were sometimes coded into more than one of the categories in the following coding system.

1. *Diagnosis.* The subject comments upon some unreasonable property of the solution. For example, one subject complained that the correlation coefficient had a value higher than 1, and another that the value of the regression coefficient was negative and the slope of the regression line positive (Problem 2).
2. *Error hypothesis.* The subject mentions some specific hypothesis about what the error could be. For example, one subject, when calculating the trend values in Problem 1, said: "But it is the logarithm here that I presumably have put in the wrong place." Another example is a subject saying: "There seems to be numerical errors."
3. *Discontent.* The subject utters general discontent with the solution. Examples are: "Yes, this seems odd," and "I have made an error somewhere."
4. *Checking activity.* This category was applied when a subject either performed some activity in order to check the presence or absence of some feature in the solution (for example, some subjects checked whether the values derived from the regression equation were approximately the same as the corresponding original values) or repeated parts of his or her old solution. To repeat one's solution is really a check of a property of the solution since when repeating the old solution one is usually checking for identity of results. The present category was also used when the subject explicitly mentioned that he or she tried to remember some information, or had failed to do so. Finally, the category was used when a subject read all or part of the problem instruction text.
5. *Error detection.* The subject detects an error. For example, a subject who had mixed up the x and y values in Problem 2 said (after some other attempts to localize the error): "I have done this wrong, I have mixed them up again!".
6. *Change of solution.* The subject gives a new solution to some part of the problem already solved.
7. *Exit from Error suspicion episode.* The subject develops his or her problem solution further than it had been developed when the current Error suspicion episode started. Thus, this category marked the end of an Error suspicion episode.
8. *Giving up.* The subject gives up his or her attempts to solve the problem.

It should be noted that the coding scheme used for the actual analysis contained 10 categories. However, two categories: Affirmative evaluation and Justification of solution method, will not be reported on here. Furthermore, for the sake of brevity only a summary of the main results of this analysis is given without presenting the data in a table.

Change of solution. Subjects changed their solution in 65 (44%) of the Error suspicion episodes. It is interesting to note that the change in 63 of these 65 episodes involved change of an erroneous solution part. Due to subjects' activities, the percentage of Error suspicion episodes with relevant errors, was decreased from 66% to 40%.

Lack of activity. The results show that subjects did not perform any, or only one, activity in 34% of the Error suspicion episodes. Somewhat less than half of these Error suspicion episodes (42%) had relevant errors and only in 4% of the episodes did the single activity involve a change of the solution. It seems that the subjects often ended Error suspicion episodes prematurely.

Utilization of symptom information. In 33% of the episodes which still had relevant errors at the end of the episode, the subjects either did nothing at all or ended the episode by expressing dissatisfaction with the solution (as coded by the categories Diagnosis, Error hypotheses, or Discontent). This figure does not include episodes where the subjects gave up. Errors were still relevant to the episode 11 out of the 12 times when subjects ended an episode by expressing dissatisfaction with the solution. These results suggest that subjects did not fully take advantage of the symptom information perceived in these situations.

Number of activities. As mentioned earlier, some Error suspicion episodes were triggered by erroneous solution parts and others were not. It is of interest whether subjects' problem-solving behavior differed in these two types of Error suspicion episodes. It should be noted that the original 10 category coding scheme was used for this analysis. An analysis of the number of coded activities shows that the subjects performed fewer activities in episodes triggered by correct solution parts ($M = 2.0$) compared to episodes triggered by erroneous solution parts ($M = 8.0$), $p < .01$, $N = 12$.

Amount of Solution Performed Between Error and Error Suspicion Episode

One might speculate that the sooner a symptom of an error is noticed after the error is made, the greater are the chances for its discovery. To evaluate

this hypothesis, the distance, i.e., the amount of problem solution performed between the occurrence of an error and each of the Error suspicion episodes that the error was relevant to was compared between corrected, discovered and undetected errors.

This analysis involved 98 errors, i.e., all errors relevant to some Error suspicion episode. The following coding system was used:

1. *At once.* The Error suspicion episode occurs directly after the error has been made.
2. *Before next main substep started.* A subject has carried out one or several simple operations after the error but has not yet started the following main substep (see Appendix).
3. *Between one and two main substeps.* An Error suspicion episode takes place in the next main substep after the one where the error occurred.
4. *After two main substeps.* The Error suspicion episode takes place during the second or subsequent main substep after the error has occurred.

Table IV shows that subjects detected (defined by a change to a new erroneous alternative or, in a few cases, by explicit detection) or corrected errors more often in Error suspicion episodes occurring soon after the error, compared to Error suspicion episodes occurring further away from the error. After one main substep the probability of correction decreased drastically.[1]

The same result holds when the analysis only includes the first occurring, relevant Error suspicion episode after an error, $p < 0.01$, $N = 11$ (shown within parentheses in Table IV).

An analysis of the errors detected in the Direct error-hypotheses episodes is in line with these results. This time, the amount of solution performed between the occurrence of the error and its detection was analyzed. Nearly all the error detections in Direct error-hypotheses episodes (93%) were anlayzed as taking place "at once" after the errors were made, and most (96%) involved the correction of the error.

Problem-solving Proficiency and Error Detection Behavior

This section deals with correlations between problem-solving proficiency and various aspects of the problem solvers' error detection behavior. As it is

[1] In order to evaluate this result statistically, the p-value was calculated for the result of each subject with Pitman's test (Bradley, 1968). .5 was then subtracted from these p-values. This procedure gives a number for each subject which, if it is greater than zero indicates a positive trend and if it is smaller than zero indicates a negative trend. In the next step these values were entered as differences into Wilcoxon's matched-pairs signed-ranks test which yielded a significant T-value ($p < .01$, $N = 11$).

TABLE IV
Number of Relevant Errors Corrected, Detected and Not Detected in Error Suspicion Episodes Occurring at Different Distances After the Errors. Within Parentheses, Number of Corresponding Errors for the First Error Suspicion Episode After an Error

Relevant Error Suspicion Episode Occurring:	*Correct New Solution*	*Only Detected*	*Not Detected*
Directly after error	21 (20)	4 (4)	8 (8)
Before next main substep started	31 (28)	2 (2)	29 (18)
Between one and two main substeps	1 (1)	1 (0)	8 (2)
After two or more main substeps	8 (2)	1 (0)	48 (13)
Total	61 (51)	8 (6)	93 (41)

not clear how to define problem-solving proficiency, three different definitions were used. All three definitions focused on the errors made by the problem solver. A low degree of problem solving proficiency implies, according to definition:

A) many errors made by the problem solver;
B) many errors left in the final solution; and,
C) many solution method errors and many errors in "advanced" mathematics.

Obviously, other properties of the problem-solving activities such as efficiency and creativity should be included in order to achieve a complete definition of problem-solving proficiency, but for the present purpose the three definitions given seem adequate. Table V shows correlations between low problem-solving proficiency as defined by these three definitions and various aspects of subjects' error detection behavior.

In general, it can be noted that problem-solving proficiency according to definition B most often correlates significantly with the investigated behaviors and that definition C nearly as often correlates significantly with the investigated behaviors. In this context, it may be of interest that four of the five subjects with high grades and more advanced education in mathematics were among the six most proficient problem solvers according to definition B.

As can be seen in Table V, the correlations between the three definitions of problem-solving proficiency were all positive and significant. Thus,

a lower degree of proficiency according to definitions A and B were associated with making a higher number of solution method errors (including higher level mathematical errors). (Hereafter when these two types of errors are collapsed they will be called conceptual errors.)

We now consider the correlations between problem-solving proficiency and error detection. Table V shows that problem-solving proficiency (according to definitions B and C) was significantly correlated with the proportion of the errors that was detected. As these correlations are negative the implication is that more proficient problem solvers detected a larger proportion of their errors compared to less proficient problem solvers. On a general level, this result holds for both execution errors and conceptual errors. It also holds for both Direct error-hypotheses episodes and Error suspicion episodes. On a more specific level, the result holds for both execution errors and conceptual errors in the Direct error-hypotheses episodes, whereas it only holds for the conceptual errors in the Error suspicion episodes. (For the Error suspicion episodes the result only holds according to definition B).

TABLE V
Correlations Between Problem-Solving Proficiency Defined in Three Different Ways and Various Aspects of Subjects' Error Detection Behavior

	Lack of Problem-Solving Proficiency According to:			
	Total Number of Errors	*Errors in Final Solution*	*Number of Conceptual Errors*	*df*
Errors in final solution	+.73**	—	—	14
Number of conceptual errors	+.70**	+.87**	—	14
Proportion detected errors of all errors	−.30	−.71**	−.57*	14
Proportion detected execution errors of all execution errors	−.45	−.74**	−.51*	14
Proportion detected conceptual errors of all conceptual errors	−.35	−.72**	−.53*	14
Proportion errors detected in the Direct error-hypotheses episodes	−.38	−.79**	−.77**	14
Proportion execution errors detected in the Direct error-hypotheses episodes (of all execution errors)	−.37	−.54*	−.47	14
Proportion conceptual errors detected in the Direct error hypotheses episodes (of all conceptual errors)	−.22	−.51*	−.50*	14
Proportion errors detected in the Error suspicion episodes (excl. errors detected in Direct error-hypothesis episodes)	−.24	−.56*	−.37	14

TABLE V (continued)

	Lack of Problem-Solving Proficiency According to:			
	Total Number of Errors	*Errors in Final Solution*	*Number of Conceptual Errors*	*df*
Proportion execution errors detected in the Error suspicion episodes (excl. execution errors detected in Direct error-hypotheses episodes)	−.20	−.25	−.09	12
Proportion conceptual errors detected in the Error suspicion episodes (excl. conceptual errors detected in Direct error-hypotheses episodes)	−.46	−.57*	−.39	12
Proportion errors relevant to Error suspicion episodes (excl. errors detected in Direct error-hypotheses episodes)	−.15	−.44	−.36	14
Proportion Error suspicion episodes triggered by erroneous solution parts.	+.07	+.37	+.56*	14
Proportion corrected or discovered errors of all errors relevant to some Error suspicion episode	−.06	−.30	−.09	14
Difference in number of errors before and after changes of the solution in the Error suspicion episodes	−.27	−.26	−.37	14
Mean length of distance from error to first relevant Error suspicion episode	.00	+.13	+.02	14
Number of Standard check episodes	−.41	−.44	−.49	14

*$p < .05$
**$p < .01$

Table V further shows that there was a nearly significant correlation between problem-solving proficiency and the proportion of errors relevant to Error suspicion episodes (by definition B). Conversely, a lower degree of problem-solving proficiency (according to definition C) was associated with a higher proportion of Error suspicion episodes triggered by erroneous solution parts. Thus, solvers with a low degree of problem-solving proficiency may be said to have less "false alarms" compared to solvers with a high degree of problem-solving proficiency.

There were no significant correlations between problem-solving proficiency and the proportion of errors (of all errors related to these episodes)

corrected or discovered in Error suspicion episodes, nor with respect to the efficiency of change in terms of the number of errors before and after changes were made. It is also of interest that problem-solving proficiency was unrelated to the "distance" between occurrence of an error and the first Error suspicion episode to which the error was relevant.

Finally, we note that there was a trend for high problem solving proficiency (according to definitions B and C) to be significantly correlated with more frequent performance of Standard checks.

DISCUSSION

The purpose of the present study was to analyze how problem solvers detect their own errors in statistical problem solving. The negative evaluation episodes (hereafter referred to as evaluative episodes) occurring in subjects' problem-solving; Standard checks, Direct error-hypotheses and Error suspicion episodes, are assumed to account for subjects' error detection processes. First, we discuss the effect of subjects' evaluative episodes on their error detection and subjects' detection of different types of errors. Then we deal with subjects' error detection processes per se, especially the Error suspicion episodes and finally we consider how error detection behavior is related to problem-solving proficiency.

Evaluative Episodes and Error Detection

An important finding is that only one-third of the undetected errors were relevant to some evaluative episode. Thus, it appears that subjects' greatest difficulty in detecting errors was reacting to the effects of their errors. This was the case especially for the skip errors, none (except one) of which were relevant to any evaluative episode. It is possible that a different problem-solving technique would improve subjects' results in this respect. More errors might be suspected if subjects paid more attention to the meaning of operations performed and results reached. However, the role of prerequisite knowledge also needs to be investigated in this connection.

Turning to the effects of the different types of evaluative episodes the results show that the Direct error-hypotheses episodes and the Error suspicion episodes were the most frequent types of episodes. Furthermore, these two types of evaluative episodes were triggered by erroneous solution parts to a much greater extent than the Standard check episodes (95% and 66% as compared to 12%). This difference is of interest with respect to error detection since no errors were detected in episodes triggered by correct solution parts. It also merits attention that all of the undetected errors that were relevant to some evaluative episode were relevant to Error suspicion episodes

(except two). It thus appears that the subjects had a partly unutilized resource in their Error suspicion episodes, whereas Standard checking by itself appears to be a less fruitful way of finding errors. One reason for the inefficiency of Standard checks may be that they are only carried out on the odd occasion. This conclusion is also supported by data presented in Allwood and Montgomery (1982).

A comparison between the two most frequent types of errors show that the proportion of detected execution errors was higher than the proportion of detected solution method errors. Two aspects of the results indicate that solution method errors cause less noticable effects than execution errors. First, a much higher proportion of the execution errors were detected in the Direct error-hypotheses episodes. This suggests that the effects of solution method errors are not well suited for the sudden type of detections occurring in the Direct error-hypotheses episodes. Second, a much higher proportion of the solution method errors were not even suspected, i.e., they did not contribute to a solution part that triggered an error detection process (29% as compared to 7% for the execution errors).

Yet another factor contributing to the low proportion of solution method errors detected appears to be subjects' difficulty in detecting and correcting those solution method errors that were relevant to some Error suspicion episode. A much higher proportion of the relevant solution method errors were undetected in these episodes as compared to the corresponding proportion of the execution errors. This result suggests that the diagnostic activities performed in the Error suspicion episodes are better tuned for detection of execution errors compared to solution method errors. We discuss hereafter, in more detail, subjects' activities in the Error suspicion episodes.

To sum up, these results point to the importance of the Error suspicion episodes especially for the detection of solution method errors.

A general factor was found to influence the chances of an error being detected in the Error suspicion episodes. The sooner after an error that a relevant Error suspicion episode occurred, the greater were the chances of detecting that error. This effect appears to be stronger for execution errors but the data show that the same trend occurs also for solution method errors.

There are at least two possible explanations for this result. First, if a problem solver once decides against the possibility of error, he or she may need a stronger reason to reconsider or change this decision in later Error suspicion episodes. Second, perceiving a negative symptom soon after the error means that the solver has less performed solution to consider when trying to locate an error.

However, it should be noted that more solution performed between the error and a corresponding Error suspicion episode is advantageous in one respect. The effects of the error have a chance of spreading to parts of

the solution where problem solvers might have clearer preconceptions as to what results are reasonable. The reason that this advantage does not outweigh the advantages of a shorter distance may be that its effect is primarily on the chances of suspecting an error rather than on detecting it.

Properties of error detection proccesses. Looking more specifically at subjects' error detection processes, the think-aloud protocols give evidence for both centrally-invoked and spontaneously-triggered error detection processes. The Standard checks occasionally performed by the subjects can be taken as evidence for centrally-invoked detection processes since they do not seem to be triggered by any detailed content. Rather they appear to occur as a result of subjects' general problem-solving technique. In contrast, the protocols show that the spontaneous error detection processes occur because the subject reacts to some specific property of the performed solution.

The triggering of spontaneous error detection processes may occur in two ways; namely, triggering due to matching between specific errors and representations of them; and, triggering due to a mismatch between held expectations and the produced solution. The think-aloud data analyzed in the present study does not give any clear evidence concerning the two suggested trigger mechanisms. It may still be of interest to note that most Direct error-hypotheses episodes are compatible with the first type of triggering, since in these episodes the hypotheses about an error occurs abruptly, without any preceding comments. Furthermore, 95% of these episodes were related to errors. However, not all Direct error hypotheses episodes captured any errors. Accordingly, at least some of these episodes may have been triggered because of a mismatch between the expectations of the problem solver and the performed solution. Alternatively, they may have been triggered because the subject misperceived the performed solution.

The protocol data for the Error suspicion episodes are compatible with the second type of triggering mentioned above since these episodes appear to have been initiated through the subjects' taking notice of some symptom in the solution (unfulfilled expectation). This feature also makes the data for the Error suspicion episodes seemingly incompatible with the first type of triggering where, initially, an error is recognized. The analysis shows that the reported symptom often concerns some surface feature of a result value and, furthermore, that it can have various degrees of relatedness to the errors including not being related to any error at all.

Having noticed a symptom, the problem solver proceeds to deal with it. The present study does not give support to the notion that the problem solver necessarily performs any further diagnosis of the error than what occurs through the triggering of the Error suspicion episode. Rather it seems that further diagnosis is only one possibility of many. The problem solver

may also attempt to dissolve the symptom by justifying the solution with an affirmative evaluation or by explaining the symptom away. For example, the solver may argue that the symptom follows straightforwardly from the earlier solution. Furthermore, the problem solver might just disregard the symptom and continue with his or her solution, or simply change the solution to a new alternative without any further diagnosis after having perceived the symptom. These different alternatives resemble those found by Carpenter and Daneman (1981) in their study of reading.

The data show that two-thirds of subjects' Error suspicion episodes were triggered by erroneous solution parts. This result suggests that subjects' perceptions of negative symptoms are a fairly reliable source of information about errors. In this connection, it is especially noteworthy that subjects so often either left the episode immediately after having perceived the initial symptom signal or left it after having expressed dissatisfaction with their solution. This occurred in 33% of those Error suspicion episodes which still had relevant errors by the time they terminated (excluding episodes where subjects gave up). It is unclear from the data whether this occurred due to inefficient problem-solving technique on part of the subjects (for example, not recognizing the importance of the signals) or to some other reason such as lack of motivation. Yet another possible reason is that subjects experienced that they did not have sufficient knowledge to deal with the symptom. However, the fact that the data contain no comments by the subjects in that direction can be taken as an argument against this possibility.

Subjects changed their solution in less than half of all Error suspicion episodes. On some occasions, these changes dissolved the symptom experienced and at the same time created new errors. It appears that in some of these cases, subjects only aimed at dissolving the symptoms instead of using them as a means to attain a correct solution. At least five such changes made by two of the poorest problem solvers appear to have been made exclusively to get rid of a symptom perceived. For example, one subject moved the decimal comma in a row of results in order to attain more reasonable figures. Furthermore, at least two subjects abandoned a correct solution method because it led to unreasonable values. If the hypothesis is correct that Error suspicion episodes are primarily triggered by a negatively experienced symptom rather than by a concern about the teleological (explanatory) structure of the solution, then the strategy to correct the solution by dissolving the symptom without any further considerations, is particularly self defeating. This is so since the perceived symptom might have only a very indirect relation to an error made.

In brief, the results in this study argue for the existence of two types of spontaneous error detection processes. Either errors are detected in a sudden direct manner or they are detected in a more elaborated error detection

process. It may be speculated that the latter type of process is triggered by a mismatch between the solution as perceived and some held expectation. In this process the solver perceives a symptom which he or she then may attempt to dissolve. An important property of the more elaborated type of error detection process is that it is a general mechanism that can detect also those errors for which the solver has no representation. The results above support the taxonomy of error detection processes shown in Figure 1.

Problem-solving proficiency and error detection behavior. The correlations calculated between problem-solving proficiency and various types of error detection behavior show that good problem solvers (as defined by the total number of conceptual errors and by errors remaining in the final solution) detect a larger proportion of their errors than poor solvers do. This finding is valid for both conceptual errors and execution errors. The present data make it possible to further determine why this is the case. It was found that good problem solvers (as defined by number of errors remaining in the final solution) detected a larger proportion of their errors both in Direct error hypotheses episodes and in Error suspicion episodes. Whereas this effect was valid for both execution and conceptual errors in Direct error hypotheses episodes, it was only significant for conceptual errors in the Error suspicion episodes. These results show that the good problem solvers' detection of their errors cannot be explained by assuming that they had proportionally more execution errors since they detected proportionally more of their conceptual errors in both types of episodes. The data also indicate that it is not possible to explain good problem solvers' detection of their errors in terms of quicker triggering of their error suspicion episodes.

The notion that good problem solvers either are more sensitive to possible contradictions to their preconceptions or have more relevant preconceptions has some support in the data. Thus, we have already noted that good problem solvers had a greater proportion of their errors detected in the Direct error hypotheses episodes. Also there was a tendency for good problem solvers to have a greater proportion of errors that were relevant to some Error suspicion episode. Furthermore, poor problem solvers had a greater proportion in their Error suspicion episodes triggered by erroneous solution parts, i.e., poor solvers had fewer "false alarms" than good solvers. In contrast, there do not appear to be any great differences between good and poor problem solvers with respect to their efficiency in detecting and correcting relevant errors once an Error suspicion episode had been triggered. Nor are the solution changes made by good problem solvers more efficient in decreasing the number of errors compared to those made by the poor problem solvers.

Thus, in conclusion, the present data indicate that the advance of good problem solvers when it comes to error detection lies more in the processes leading to the triggering of an error detection process and less in what

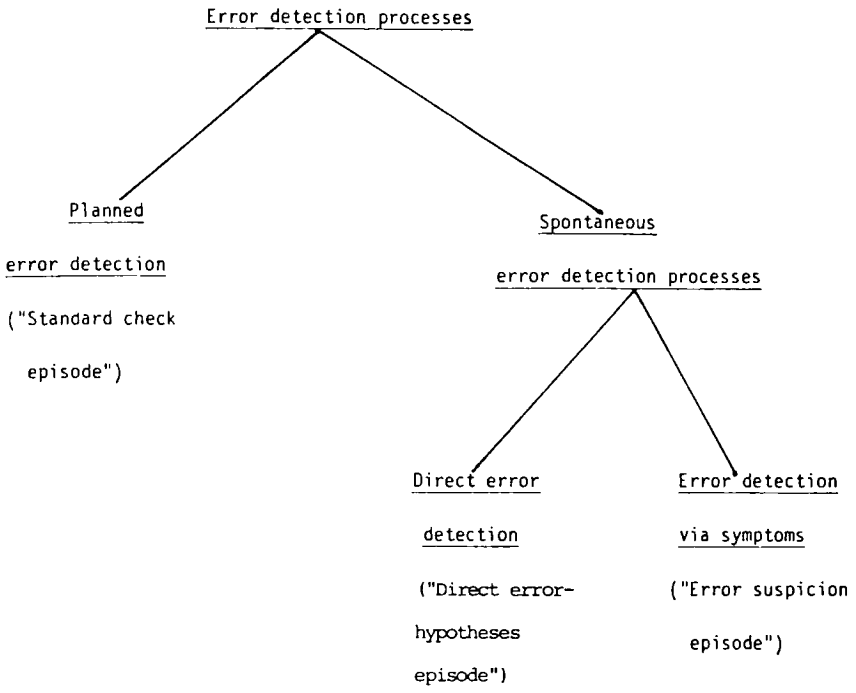

Figure 1. Taxonomy of types of error detection processes.

is done after such a process has been triggered. The tendency for good problem solvers to perform more Standard checks furthermore suggests that good problem solvers are more open to the possibility of errors.

From these results it is suggested that further research can profitably be directed towards an investigation of how to train problem solvers towards more efficient utilization of their symptom perception. This problem breaks into two: first, the question of how to improve problem solvers' perception of when a solution deviates from their preconceptions. Part of the solution here seems to be to improve the relevant knowledge structures of the problem solvers. However, there may also be a problem-solving technique aspect involved. This aspect relates to the ability to recognize and take seriously one's error symptom experiences. Second, the question of how to improve problem solvers' error detection once symptoms are recognized as such, merits attention. This task would involve more detailed analysis of subjects' exact behavior in the Error suspicion episodes.

REFERENCES

Allwood, C. M. (1976). A review of individual differences among problem solvers and attempts to improve problem solving ability. *Göteborg Psychological reports, 6,* 11.

Allwood, C. M., & Montgomery, H. (1981). Knowledge and technique in statistical problem solving. *European Journal of Science Education, 3,* 431–450.

Allwood, C. M., & Montgomery, H. (1982). Detection of errors in statistical problem solving. *Scandinavian Journal of Psychology, 23,* 131–139.

Baars, B. J. (1980). On eliciting predictable speech errors in the laboratory. In V. A. Fromkin (Ed.), *Errors in linguistic performance: Slips of the tongue, ear, pen and hand.* New York: Academic Press.

Bradley, J. W. (1968). *Distribution-free statistical tests.* London: Prentice Hall.

Carpenter, P. A., & Daneman, M. (1981). Lexical retrieval and error recovery in reading: A model based on eye fixations. *Journal of Verbal Learning and Verbal Behavior, 20,* 137–160.

Hayes, J. R., & Flower, L. S. (1980). Identifying the organization of writing processes. In L. W. Gregg & E. R. Steinberg (Eds.), *Cognitive processes in writing.* Hillsdale: Erlbaum.

Laver, J. (1980). Monitoring systems in the neurolinguistic control of speech production. In V. A. Fromkin (Ed.), *Errors in linguistic performance: Slips of the tongue, ear, pen and hand.* New York: Academic Press.

Matz, M. (1979). *Towards a process model for high school algebra errors.* (A. I. Lab. Working Paper 181). Cambridge, MA: Massachusetts Institute of Technology.

Nooteboom, S. G. (1980). Speaking and unspeaking: Detection and correction of phonological and lexical errors in spontaneous speech. In V. A. Fromkin (Ed.), *Errors in linguistic performance: Slips of the tongue, ear, pen and hand.* New York: Academic Press.

Norman, D. (1981). Categorization of action slips. *Psychological Review, 88,* 1–15.

Sussman, G. J. (1975). *A computer model of skill acquisition.* New York: American Elsevier Publishing Company, Inc..

APPENDIX

Correct Solutions to Problems 1 and 2, Divided into Main Substeps.

Problem 1

1. Choice of model
   Multiplicative model:
   $O = T \cdot S \cdot I$
   O = Observed values
   T = Trend
   S = Season
   I = Random component
2. Estimation of trend.
   The trend can be estimated by smoothed averages (or successive four-point sums) and centration. For example, T for 1971_{III} is estimated by
   $(((13 + 12 + 8 + 25)/4) +$
   $((12 + 8 + 25 + 24)/4)) / 2.$
3. Computation of O/T in order to obtain $S \cdot I$.
4. Computation of mean of O/T for each quarter to obtain S.
5. Correction of S-values by computing
   $S_i \cdot \frac{4}{\Sigma S_i}$
6. Computation of sales values adjusted for seasonal influence, by calculating O/S.

Problem 2

1. Definition of x as amount of fertilizer and y as size of harvest.
2. Choice of model.
   Linear model: $Y = a_o + bx$.
3. Computation of
   $b = \frac{n\Sigma xy - \Sigma x\ \Sigma y}{n\Sigma x^2 - (\Sigma x)^2}$
4. Computation of
   $a_o = \bar{y} - b \cdot \bar{x}$
5. Computation of $r =$
   $\frac{n\Sigma xy - \Sigma x\Sigma y}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2]\ [n\Sigma y^2 - (\Sigma y)^2]}}$
6. Identification of average increase of harvest for each kg fertilizer as the value of b, i.e. the slope of the regression line of kg harvest on kg fertilizer (slope = 0.068).
7. Computation of $Y = a + b \cdot 350$ to obtain the expected harvest for 350 kg fertilizer per $100m^2$.