Memory & Cognition
1989, 17 (6), 759-769

The validity of verbal protocols

J. EDWARD RUSSO
Cornell University, Ithaca, New York

ERIC J. JOHNSON
University of Pennsylvania, Philadelphia, Pennsylvania

and

DEBRA L. STEPHENS
University of Maryland, College Park, Maryland

The reactivity of a "think aloud" verbal protocol and the veridicality of different retrospective protocols were tested over four dissimilar tasks. Generating a concurrent protocol altered accuracy in two tasks, simple addition and a choice between two gambles, and generally prolonged response times. Such reactivity partially qualifies the dominant theory of protocol generation (Ericsson & Simon, 1984). Retrospective protocols yielded substantial forgetting or fabrication in all tasks, supporting the consensus on the nonveridicality of these methods. It is concluded that protocol validity should be based on an empirical check rather than on theory-based assurances.

The ascendancy of the information processing paradigm has increased the demand for data that trace cognitive processes. One source of process data is a "think aloud" verbal protocol in which subjects report their thoughts during the performance of a primary task. The increased use of these data has prompted an inquiry into their validity. Nisbett and Wilson (1977) addressed several fundamental issues, especially the intrusion of fabricated mental events into verbal reports. Their broad indictment elicited many clarifying responses (e.g., Quattrone, 1985; Sabini & Silver, 1981; Smith & Miller, 1978; Turner, 1986, 1988; Wright & Rip, 1981).

Ericsson and Simon (1980, 1984) responded with an analysis of protocol validity based on a theory of protocol generation. To reduce invalidity, they warn against all retrospectively collected protocols as subject to forgetting and fabrication. They also suggest shunning concurrent (think aloud) protocols where instructions invite self-theorizing or other introspective explanations. That is, proper protocols ask subjects to report their thoughts, not to explain them.

Ericsson and Simon's theoretical position implies that proper instructions and task selection can achieve independence between verbalization and the primary process. Neither process interferes with the other so long as (1) subjects report only the contents of short-term memory (STM) and (2) those contents are in oral form, that is, coded as a string of phonemes. This excludes tasks that require a recoding from a nonoral (e.g., pictorial) representation to an oral one and cognitive processes that leave little or no trace in STM, especially those that are automated.

One qualification to this theory-based assurance of validity involves the competition for processing resources between the primary task and verbalization. Especially for tasks in which rehearsal of partial results places a heavy burden on STM, Ericsson and Simon (1984) warn that "interruption and suppression of rehearsal lead to a rapid loss of information from STM. Hence, we would expect that prolonged attention to items in STM to allow verbalization will be disruptive of tasks that impose high loads on STM" (p. 249). It is worth noting that the competition for processing resources is eased by multiple resource pools (Navon & Gopher, 1979; Wickens, 1980, 1984). For instance, Wickens (1984) specifically suggests at least partially distinct resources for encoding/processing and for responding, a distinction corresponding roughly to our primary task and concurrent verbalization. However, the multiple-resource view is not uniformly accepted (e.g., Kantowitz, 1987), nor are the different pools of resources necessarily independent. For instance, Wickens (1987) suggests that processing and responding may draw on some common resources. Furthermore, although there is considerable evidence for noninterference between spatial and verbal processing (e.g., Brooks, 1968), the translation of a pictorial to an oral code may require some of both resources. Thus, limited processing resources may qualify any assurance of validity, even if multiple-resource theories are accepted.

The authors thank Gerhard Deffner, K. Anders Ericsson, Donald Fiske, John Payne, Herbert Simon, and Peter Wright for helpful comments on an earlier draft. This acknowledgement should not be taken to imply their agreement with our conclusions. We also gratefully recognize the assistance of Ann Thibodeau in data collection. Major support was provided by NSF Grant DAR76-81806, with additional support from a bequest by C. N. Newman to the Graduate School of Business, University of Chicago. Eric J. Johnson is affiliated with the Wharton School at University of Pennsylvania; Debra L. Stephens is affiliated with the College of Business and Management, University of Maryland. Requests for reprints may be sent to J. Edward Russo, Johnson Graduate School of Management, Cornell University, Ithaca, NY 14853.

TWO FORMS OF INVALIDITY

Protocol invalidity can take at least two forms: reactivity and nonveridicality. A verbal protocol is *reactive* if verbalization changes the primary process. Reactivity can occur either as a change in the primary process that may alter the outcome of that process or as a simple prolongation of response time (RT). Process changes are usually considered consequential in that they can invalidate the theoretical conclusions the data were designed to provide. In contrast, lengthening RT is seldom consequential in that few theories make predictions at that level of detail for tasks longer than a few seconds in duration (for an exception, see Just & Carpenter, 1980).

A protocol is *nonveridical* if it does not accurately reflect the underlying primary process. Nonveridicalities include errors of omission (e.g., not reporting some thoughts) and errors of commission (e.g., reporting mental events that did not occur). Although omission is consequential, fabricated intrusions are usually more serious, because these data enter into the protocol's analysis as if they were veridical.

TESTING THE VALIDITY OF VERBAL PROTOCOLS

What constitutes a definitive test of the validity of a think aloud protocol? Tests of reactivity usually compare a silent control to the concurrent protocol condition. Ideally, these two groups could be compared on the basis of *criterion data*, that is, some perfectly valid and detailed measure of the underlying process. Effects of verbalization would appear as differences in the criterion data between the verbalizing and silent conditions.

Of course, criterion data do not exist. Indeed, there are few other process-tracing methods to turn to. Among these are eye fixations and process markers, such as the naturally occurring "moves" in such problem-solving tasks as the Tower of Hanoi (Simon, 1975). Unfortunately, these overt behaviors do not reveal the level of detail provided by verbalization. Thus, in many tasks, the existence of even a close approximation to ideal criterion data is problematic.

Fortunately, tests of reactivity can also be based on such ordinary output measures as accuracy and RT. Significant differences between a silent control group and a verbalizing group can be attributed to the concurrent protocol. Differences in accuracy are particularly important, because they reflect fundamental changes in the primary process that most investigators would regard as consequential. One can imagine circumstances in which the primary process is altered but overall accuracy and RT are not significantly changed, but, in the absence of criterion data, the systematic examination of accuracy and RT can provide a meaningful if incomplete test of reactivity.

Testing the veridicality of a concurrent verbal protocol is much more difficult. Indeed, it is nearly impossible without another simultaneous source of process data.[1] In addition, concerns about reactivity naturally take precedence over lack of veridicality because there is little point to testing whether or not a report is veridical if verbalization has already changed the primary process being reported. Thus, our main focus will be reactivity.

REACTIVITY: EXISTING EVIDENCE

We are fortunate that Ericsson and Simon (1984) have thoroughly reviewed the empirical literature. To simplify our narrower review, we exclude the large number of studies that used either retrospective protocols or improper instructions. However, we retain tasks that entail some recoding into oral form, since any reactivity might still be negligible. Thus, we focus on studies where the task and instructions were most favorable to nonreactivity.

There are 12 such studies, which we divide into a primary group of 5 (Fryer, 1941; Karpf, 1973; Roth, 1966; Walker, 1982; Wegner, cited in Merz, 1969) and a secondary group of 7 (Carroll & Payne, 1977; Dansereau, 1969; Dansereau & Gregg, 1966; Feldman, 1959; Johnson & Russo, 1978; Kazdin, 1976; Smead, Wilcox, & Wilkes, 1981). To the primary group we add Fidler (1983) and Schweiger (1983), which were not available to Ericsson and Simon.

In the primary studies, the test for reactivity (and other aspects of protocol performance) was not an incidental check, but an important methodological goal. This is reflected in their designs, which tend to contain more trials and subjects than studies in the second group. Of these seven, three reported significantly longer RTs (Fidler, 1983; Karpf, 1973; Wegner, in Merz, 1969) and one reported that fewer problems were solved within a fixed time period (Fryer, 1941). The results of the last study are ambiguous because it is unclear whether each problem took longer to solve or whether the same number of problems was completed with a higher proportion of errors.

The seven remaining studies were not designed mainly as tests of the concurrent protocol methodology, but nonetheless included an empirical check for reactivity. No reactivity was found in any of these seven studies. However, this result is not conclusive for several reasons. First, some of the investigators reported either accuracy or RT, but not both. This leaves open the possibility that reactivity occurred in one measure but not the other. Second, these tests were less powerful than were those in the primary group, typically using fewer subjects and trials per subject. Finally, we worry about a publication bias in which experimenters who find protocols reactive do not pursue, much less publish, these data.

The near absence of reactivity in the empirical literature has led to a consensus that "the verbal protocol procedure slows down the process slightly but does not change it fundamentally" (Payne, Braunstein, & Carroll, 1978, p. 36). Nonetheless, the evidence is not definitive. We find it noteworthy that, although few studies systematically test reactivity, four of the seven rigorous tests found at least a weak form of it. Our first goal is a more systematic test of reactivity over a variety of tasks. By choosing tasks that both satisfy and violate Ericsson and Simon's proposed conditions for validity, we also establish boundary conditions for their theory of protocol generation.

METHOD

The experimental strategy is to contrast a concurrent verbalization with a silent (control) condition. Besides these two conditions, we also examined three types of retrospective protocols in which subjects solved a problem silently and, immediately afterward, reported their thoughts while solving it. In the first retrospective condition, subjects had only their responses (i.e., the problem solutions) before them as they recalled their thoughts (*response-cued*). In another, they saw only the original problem (*stimulus-cued*). In the third retrospective condition (*prompted*), subjects saw the original problem and, superimposed on it, a replay of the sequence of eye fixations they had made while solving it (Russo, 1979). The retrospective conditions served two purposes. First, a comparison of the concurrent and retrospective conditions allowed us to investigate the magnitude of forgetting and fabrication universally attributed to the latter (Nisbett & Wilson, 1977; Payne et al., 1978; Smith & Miller, 1978), at least to the degree that the concurrent condition serves as an accurate standard. Second, because these protocols make different demands on the subject than does a concurrent one, differences in any observed reactivity might help diagnose possible sources of that reactivity.

Primary Tasks

The selection of tasks was critical to a fair but rigorous test of protocol reactivity. It was guided by three considerations. First, although we would have liked to test many tasks, we also needed sufficient statistical power to detect small levels of reactivity. This dictated a focus on fewer tasks. Note that a rigorous test of differences in accuracy must be based on many trials because only a single categorical observation (correct or incorrect) results from each trial. Second, we sought a variety of tasks (e.g., both verbal and pictorial and with and without a heavy STM load) that would vary with respect to Ericsson and Simon's theory-based predictions of reactivity. Although this criterion runs counter to our own more agnostic view that it is hard to know in advance whether or not reactivity will occur, it accords with our goal of empirically examining their predictions about the kinds of tasks free from invalidity.

Four dissimilar tasks were used: (1) a verbal task, anagrams; (2) a numerical task, choosing between two simple gambles; (3) a pictorial task, Raven's (1958) progressive matrices; and (4) the mental addition of three three-digit numbers, a task that imposes a heavy STM load. Two of these tasks, anagrams and gambles, meet the Ericsson and Simon criteria in that they utilize orally encoded information and subjects were properly instructed to report only the contents of STM, not explanations or elaborations of those contents. The addition task might not be free of reactivity because of its dependence on the rehearsal of partial results. Raven's task risks reactivity, because verbalization requires a recoding from a pictorial to an oral code.[2]

Stimuli

From pretest data, we selected for each task 55 problems that met two criteria: an overall accuracy roughly between 70% and 75% and a range of individual problem difficulty between 50% and 90%. The final stimuli for each task are described below.

Addition. Each problem required the mental addition of three three-digit numbers displayed in a standard 3 × 3 matrix. To impose a STM load, we required the subjects to add the columns in right-to-left order without returning to any earlier column. Although this right-to-left order is the most common, other strategies can be adopted (see Hitch, 1978). We monitored the subjects' compliance by tracking their eye fixations and verbally warned them if they backed up to a previous column. This was rarely necessary.

Anagrams. Each anagram consisted of five letters (e.g., GORRI) to be rearranged to form an English word (in this case, the word Rigor). Problems were selected from Arnold and Lee (1978a, 1978b). In all cases, there was only one solution, excluding proper names.

Gambles. Each problem presented two gambles consisting of a probability of winning and a corresponding payoff given to two digits (e.g., .32/$6.90 and .54/$4.30). There was no loss associated with the complementary probability. One gamble was displayed below the other. Because the subjects were instructed to choose the gamble with the higher expected payoff, the task was essentially mental multiplication.

Raven's matrices. Each Raven's (1958) progressive matrix is a 3 × 3 array of figures, with the bottom right cell missing. Subjects must discover the pattern in the array and complete it by selecting one alternative from another set of figures. This other set contained three of the eight answers provided in Raven's standard test booklet. We used the two most frequently chosen distractors along with the correct one, randomly arranged in a fourth column to the right of the matrix. Our stimuli were chosen from Sets D and E and Advanced Sets I and II (Raven, 1958).

A difficulty occurred in pretesting the anagram and Raven's problems. Even when unable to solve a problem, the subjects sometimes persisted. This created long RTs and, in the case of determined subjects, an overly long experimental session. To prevent this, the subjects in these two tasks were given the option of requesting hints, such as the first letter of the correct anagram word. These hints were to be used only if the subject could not otherwise reach a solution. Any trial in which a hint was requested was classified as an error since the subject was unable to solve it without assistance and since the intervention arbitrarily prolonged RT.

Subjects

Twenty-four students served as paid volunteer subjects. Each participated in five 2-h sessions and was paid $9.00 plus $0.05 for every problem solved correctly. For each first hint (on an anagram or a Raven's problem) $0.02 was subtracted, with $0.01 subtracted for every subsequent hint.

Instructions

The subjects in the concurrent protocol condition were instructed to think aloud while solving the problem.[3] When the subjects were silent for more than a few seconds, they were prompted by "Please tell me what you are thinking." Prompting was rarely needed; it was used roughly once every 50 trials. In the stimulus- and response-cued retrospective conditions, the subjects were asked to "tell what you were thinking as you solved the problem." The prompted instructions were to "explain why you looked where you looked and what you were thinking when you looked there." The first part of these instructions, to explain why, was improper. It invites the kind of self-theorizing that can make retrospective reports nonveridical. We retain the prompted protocols in the analyses of results for completeness and because any problem they created may be negligible, as a subsequent analysis will suggest. Nonetheless, they should be treated skeptically.

Design

The most important design decision was whether to use a repeated measures design or a between-subjects design. The former risked carryover effects when the same subject used all protocol methods. The latter introduced additional subject variance, reducing the statistical power needed for a rigorous test of accuracy differences. We opted for a repeated measures design but minimized carryover effects by familiarizing the subjects with each protocol method in an initial session (as described below).

Thus, the subjects performed all four primary tasks, each in a different verbalization condition. The order of tasks and the pairing of tasks with protocol condition was counterbalanced using a Latin square. Each group of 12 subjects comprised a block of three different Latin squares, as described in Cochran and Cox (1957; see Plan 6.12, p. 241). This design enabled us to test the interaction between protocol method and task, which amounts to the testing of method effects within each task.

We varied the type of retrospective protocol between subjects, alternating the response- and stimulus-cued instructions. Thus, all 24 subjects received the concurrent, control, and prompted instructions, whereas only half participated in the response-cued condition and the other half participated in the stimulus-cued condition. Every task-by-method cell contained 55 trials, the first 10 of which were considered practice.

Apparatus

Eye fixations were measured by an Applied Sciences' Eye View Monitor (Model 1996). A PDP 11-34 computer was used to control the display of the stimuli and to record the subjects' response latencies, answers, and eye fixations. Although eye fixations were used only to monitor the addition task, these data were recorded for all tasks and methods. The stimuli for all tasks except Raven's were displayed on a CRT screen; the Raven's matrices were shown on 8½ × 11 in. transparencies, positioned in front of the screen by the experimenter. Protocols were recorded on audio cassette tapes.

Procedure

The subjects performed one task in each of the four test sessions. An initial session was devoted to training on all four protocol methods using a fifth task, a type of number puzzle. We expected this to minimize any carryover between methods that might lead to different effects for different orders of the methods.[4] At the start of each of the next four sessions, the subjects received instructions for the task and protocol method. The 10 practice trials were presented, followed by 45 test trials. Retrospective protocol generation began immediately after the buttonpress ending a trial. The subjects were required to generate the appropriate protocol during all trials including practice.

For anagrams, a subject could receive a hint via a keypad. For Raven's matrices, the hints had to be requested verbally and were read aloud by the experimenter.

RESULTS

If the assumption of independence between the primary task and verbalization is correct, the anagram and gambles tasks should be free of reactivity except possibly for a prolonged RT. In contrast, because verbalization in the Raven's task requires a pictorial to oral recoding, reactivity is more likely. Similarly, the high STM load in the addition task may make it vulnerable to disruption by concurrent verbalization.

Accuracy

The accuracy for each protocol method and task is shown in Figure 1. Each datum is based on 270 trials (45 test trials for 6 subjects). The absolute accuracy is reported for the no-protocol control condition, and the change in accuracy relative to the control (i.e., the magnitude of any reactivity) is reported for each protocol condition.

Two findings stand out. First, some tasks show significant reactivity. A concurrent protocol significantly improved the accuracy of a choice between two gambles (+.20) and significantly decreased the accuracy of adding three-digit numbers (−.14). Second, other tasks show no effect: The a priori contrast (described below) comparing concurrent and silent control groups over the four tasks showed no significant difference. Thus, the impact of protocol generation depends strongly on the task,

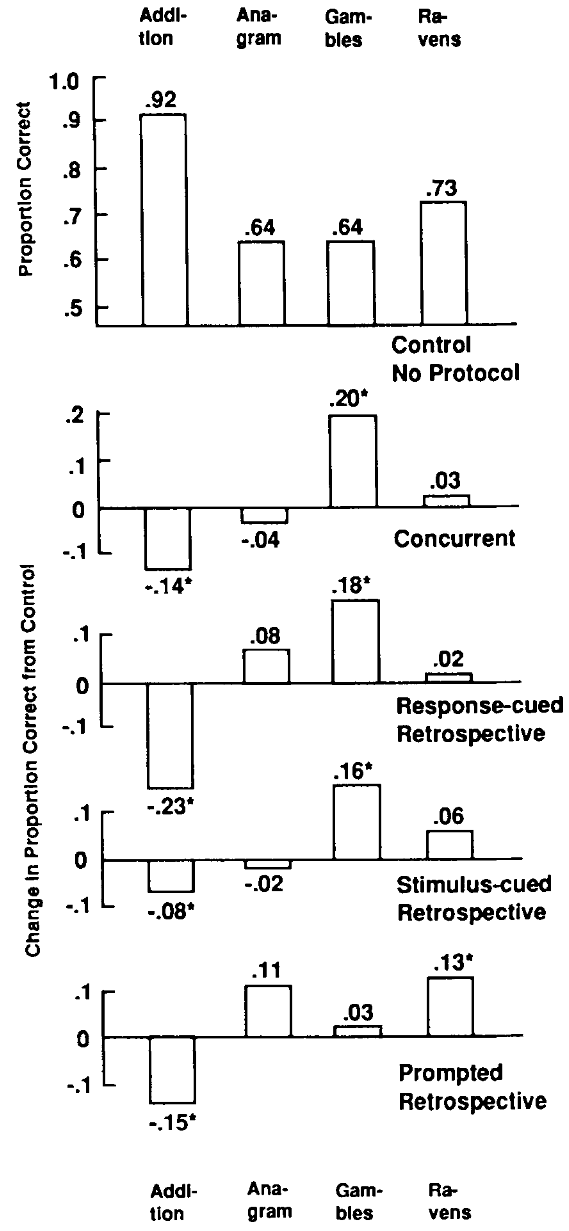

Figure 1. Accuracy of control condition and the accuracy difference between control and each of four protocol methods. An asterisk indicates a statistically reliable ($p < .05$) difference in accuracy between a protocol method and the control condition using the logistic regression analysis described in the text.

suggesting that the causes of reactivity are not general but due jointly to the demands of the task and to verbalization.

For completeness, Figure 1 reports the results for the three retrospective methods. These methods largely reflect the direction of the concurrent effect. Only the gambles and Raven's tasks in the prompted condition produced error rates different from the concurrent condition. This similarity suggests a sizable overlap in the causes of reactivity for concurrent and retrospective verbalizations.

To test the significance of these differences, we used a logistic regression (Neter & Wasserman, 1974, p. 330). Alternative tests, such as ANOVAs with a dichotomous dependent variable, are less powerful or biased in an unknown direction (Neter & Wasserman, 1974, p. 322ff). For example, an arcsin transform of observed cell proportions is commonly used with dichotomous-data-like error rates. However, because this transform is performed on cell means rather than on individual responses, it entails a substantial loss of statistical power. The logit model is analogous to a conventional ANOVA, with maximum likelihood chi-square tests replacing the *F* tests. The between-subjects factors were square and subjects nested within square, and the within-subjects factors were task, method, and their interaction. We then performed a priori comparisons of each of the four protocol methods to the control group ($p < .05$, two-tailed). The results of these 16 tests are designated by the asterisks in Figure 1.

Response Time

Figure 2 displays the geometric means of the RTs for correct trials only. Depending on the proportion of correct trials, the sample sizes for the control, concurrent, and prompted methods ranged from 150 to 249 and for response-cued and stimulus-cued from 90 to 114. Negative values indicate shorter RTs relative to control; positive values indicate that protocol generation prolonged the primary task.

There is reliable evidence that a concurrent protocol lengthens task time. RT for the choice between gambles was prolonged by 22% (from 33.8 to 41.2 sec) and for an anagram by 44% (from 14.2 to 20.4 sec). A least squares ANOVA using the same factors as the accuracy analysis confirmed the significance of these two effects. Furthermore, although the effect of a concurrent protocol varied across tasks, RTs were lengthened for all tasks. They were 7% longer for both addition (14.8 to 15.9 sec) and Raven's (35.8 to 38.4 sec).

There was uniform agreement among the three retrospective methods in the direction of the change in RT. However, in contrast with the accuracy results, the retrospective effects often exhibited a different pattern from the concurrent condition.

Since RTs are commonly skewed by a few long times, a log transform was used for all significance testing. The significant planned comparisons between control and each of the four protocol methods (within each task) are reported in Figure 2. Note that the lengthening of RT by a concurrent protocol is significant for two tasks, anagrams and gambles. Furthermore, an a priori contrast comparing the means of the concurrent and control RTs over all four tasks is significant [$F(3,3105) = 53.00$, $p < .01$].

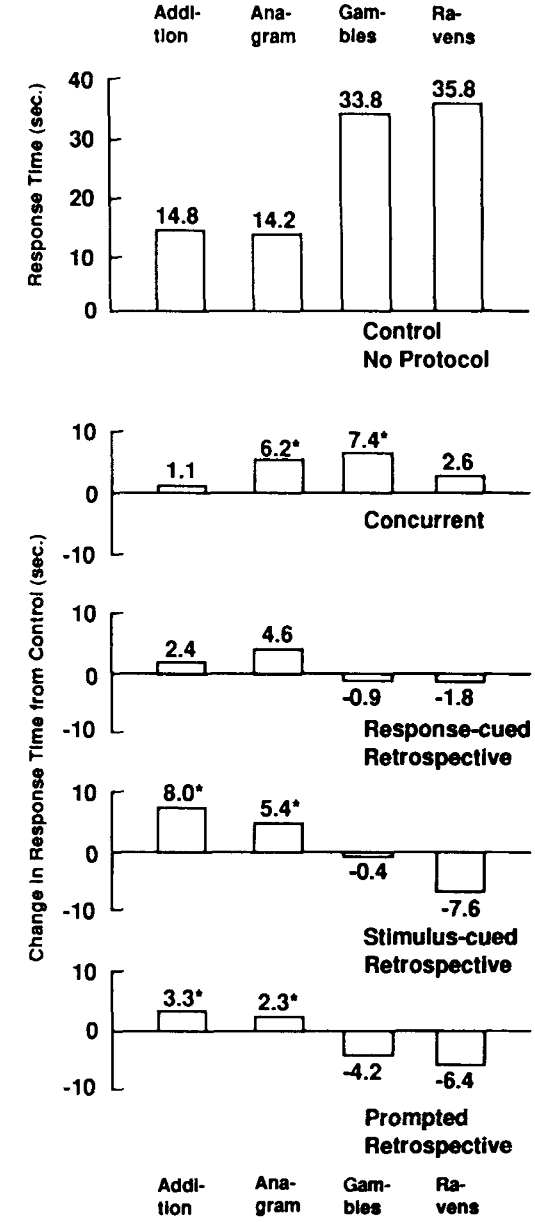

Figure 2. Geometric mean response times for all correct trials of control condition and the difference in mean response time between control and each of four protocol methods. An asterisk indicates a statistically reliable difference in response time between a protocol method and the control condition using the ANOVA described in the text.

Speed-Accuracy Tradeoff

If the accuracy rates and RTs covaried positively, we might suspect that these differences resulted from a speed-accuracy tradeoff. However, the observed changes in accuracy and RT are not compatible with such a tradeoff. As shown in Figures 1 and 2, in only 3 of the

16 cases did accuracy and RT change in the same direction, whereas 8 such patterns would be expected by chance alone. Similarly, in only 2 of the 11 cases with a significant accuracy or RT effect were the differences codirectional. Although the possibility of a speed-accuracy tradeoff cannot be excluded in the few cases with codirectional effects, this would not seem to form a general explanation for the results.

Conclusion

We set out to perform a rigorous test of the reactivity of a concurrent verbal protocol using a variety of tasks. We found a significant alteration in accuracy for two of four tasks and a general prolongation of RT. Two conclusions seem justified.

First, the generally benign view of think aloud protocols, supported until now by the virtual absence in the literature of empirical reports of significant and consequential reactivity, must be questioned. Our results provide two instances of reactivity and suggest investigating other tasks in which concurrent verbal reports are collected.

Second, it would appear to be more difficult than previously thought to specify a priori whether or not a task will be altered by the generation of a concurrent protocol. In spite of Ericsson and Simon's deep analysis of the verbalization process, the predictions of reactivity drawn from their theory did not accord well with the data. The gambles task should not have been reactive, yet it was. Raven's task, involving a recoding from pictorial to oral code, might have been disqualified as susceptible to reactivity, yet it was the only one of our tasks with no significant reactivity in either accuracy or RT. We conclude that our theoretical understanding of protocol generation is not yet adequate to provide sufficient assurance about the absence of reactivity to justify foregoing an empirical check.

No single experiment is conclusive, especially one yielding results somewhat contrary to both the existing literature and the foremost theory. Awaiting a replication or extension to other tasks, the present results might be informed by the identification of mechanisms that could have caused the reactivity we observed. Although the value of post hoc theorizing is necessarily limited, some sort of explanation for the observed reactivity seems appropriate, if only as a guide to further empirical testing.

REACTIVITY IN PROTOCOL GENERATION

The empirical literature suggests at least four potential causes of reactivity: (1) the additional demand for processing resources, (2) auditory feedback, (3) enhanced learning over repeated trials, and (4) a motivational shift toward greater accuracy. These causes of reactivity are independent and task-specific in that any or all of them may be present depending on the primary task.

Additional Demand for Processing Resources

Vocalization requires that subjects (1) set up and execute the motor programs to articulate spoken words, (2) elaborate and monitor compliance with the request to think aloud, including maintaining an adequate voice level, and (3) recode idiosyncratically abbreviated oral codes so that they are intelligible to other listeners (Werner & Kaplan, 1963). A separate source of demand is the requirement to articulate strategies that become partially automated. Finally, resources are required to recode nonoral representations into oral form.

When any of these demands are present, subjects are confronted by the problem of how to allocate processing resources between the primary task and verbalization. Note that this problem remains even if multiple attentional resources are postulated (Wickens, 1987). We believe that subjects manage the demands of verbalization by using any slack resources not required by the primary task. If the verbalization demands are slight and their occurrence is compatible with the availability of slack resources, there may be no disruption of the primary process, not even a lengthening of RT. However, when the slack resources cannot easily accommodate verbalization, subjects confront a choice. They can withdraw resources from the primary process and devote them to verbalization, risking reactivity, or they can temporarily suspend verbalization, a violation of veridicality. We presume that subjects choose between these options in part by considering their relative costs in the task situation. In the present experiment, a uniform trend toward longer RTs in the verbalizing condition is compatible with an additional demand for processing resources. The significant disruption of the addition task is compatible with the extra demand on resources to verbalize a partially automated process (Zbrodoff & Logan, 1986).

Auditory Feedback

Vocalization creates additional aural stimulation that might either facilitate or interfere with performance of the primary task. In many situations, vocalizing an item facilitates recall (Penney, 1975). This effect seems to be based solely on hearing the auditory stimulus and not on the act of articulation. For example, it makes little difference whether the subject or the experimenter vocalizes the stimulus items (e.g., Crowder, 1970), and speaking aloud aids memorization more than does whispering (Tell, 1971).

Whenever tasks depend upon retention of partial results and also allow pauses during which rehearsal can occur, accuracy may increase due to auditory feedback. Merz (cited in Ericsson & Simon, 1984, p. 76) provides direct evidence that auditory feedback improves performance in simple arithmetic where the retention of partial results is critical.

Enhanced Learning

Generating a protocol may facilitate learning by giving subjects the opportunity to reflect on the primary process. This reflection may lead to the discovery of new strategies or to the improvement of old ones.

Strategy acquisition is most likely in *toolbox* tasks in which many possible strategies can be employed. For example, in anagrams, many letter pairs, positions, and pronunciations may need to be considered. Similarly, a Raven's matrix requires searching through a large space of possible rules. In contrast to toolbox tasks, addition and gambles involve applying a known algorithm. Such algorithmic tasks might be improved by the trimming of unnecessary components.

Motivational Shift

Concurrent protocols are usually generated in the presence of an experimenter and intended for subsequent transcription and analysis. Consequently, subjects can anticipate public exposure of their errors. Given such exposure, verbalizing subjects may try to shift to strategies that tend to reduce error but require more effort. For example, in our gambles task, the tedium of the required mental multiplication is easily reduced by simplifications that sacrifice accuracy (Russo & Dosher, 1983; Johnson & Payne, 1985). Evidence for a motivational shift is provided by Tetlock and Kim (1987) in a personality prediction task. They found that subjects who were publicly accountable (vs. those performing anonymously) exhibited both more complex processes and greater predictive accuracy. In general, whenever subjects' processing is publicly revealed, they may behave more in accord with the perceived preferences of the experimenter.[5]

Reactivity in Retrospective Protocols

However paradoxical it may seem, reactivity can occur in retrospective as well as concurrent protocols. The significant reactive effects in Figures 1 and 2 indicate that some cause(s) of such anticipatory reactivity must be operating.

Although feedback from articulation is ruled out, the other three causes of reactivity are not. A motivational shift can occur whenever subjects are informed that they will have to generate a subsequent verbal report; enhanced learning may be even more likely, because retrospective reports provide additional time to review the primary process. Finally, consider the possibility that subjects perceive a requirement to "have something to say" when little of the actual processing can be recalled. This might lead to the deliberate memorization of a few process components that can guide recall (or a plausible reconstruction). Such memorization would place an additional demand on processing resources.

The sources of reactivity in retrospective protocols are even less well understood than are those in concurrent verbalization. Because using a variety of protocol methods may facilitate the investigation of concurrent protocols, these retrospective effects may be worth pursuing.

Conclusion

There seem to be several possible mechanisms by which protocol generation might alter the primary task. We examined the ability of the four causes just listed to account for our results. Direct causal links could not be empirically verified post hoc. Nonetheless, the increased accuracy of the gambles task may have been aided by a motivational shift. This task in the concurrent protocol condition was the only one of the 16 task and method combinations with a significant increase in accuracy accompanied by a significant increase in RT. Additionally, the rehearsal of partial results may have been aided by overt verbalization. The other significant change in accuracy, the drop for the addition task, is compatible with the competition for processing resources in a task with a high STM level. We stress that these connections are merely speculative. Their purpose is to show that there exist specific potential causes for the reactivity we observed and, thereby, to suggest possible directions for future experimentation. Any conclusions about what caused our reactivity must await such prospective testing.

VERIDICALITY OF RETROSPECTIVE PROTOCOLS

Although our primary focus is reactivity in concurrent protocols, our experiment enables some investigation of the nonveridicality of retrospective reports. It is generally agreed that such protocols are subject to both forgetting and fabrication (Ericsson & Simon, 1984; Nisbett & Wilson, 1977; Turner, 1986). We investigate this claim by comparing the contents of the retrospective and concurrent protocols. Without perfectly valid criterion data to characterize the primary process, any conclusions cannot be definitive. Nonetheless, the observed differences among the concurrent and retrospective protocols should prove informative.

We expect two general sources of differences, errors of omission and commission. As will be explained shortly, the kind of error should change systematically as the material to be verbalized varies along an inferential hierarchy. This hierarchy classifies individual thoughts on a continuum beginning with stimulus perception, moving through initial, middle, and final inferences, and ending with the problem solution.

The statements in the 448 available protocols were coded into five categories: perceptual, low level inferences, high level inferences (but not including the answer), strategy, and all others.[6] In all four tasks, we could identify two extreme levels of inference with frequencies well above zero. However, perceptual and strategy statements occurred only infrequently in some tasks, notably addition and gambles.

Results

Table 1 presents the mean number of statements per protocol for the four inference categories, by task and protocol method. Within each task we performed a simple ANOVA, with protocol method the only factor. This ANOVA was computed separately for the four statement

Table 1
Mean Frequency Per Protocol of Statement Types by Protocol Method and Task

Type of Statement	Task	Concurrent	Response-cued	Stimulus-cued	Prompted
Perception	Addition	0.1	0.0	0.1	0.3
	Anagrams	1.0^{a}	0.0^{b}	0.0^{b}	0.5^{c}
	Gambles	0.0^{a}	0.1^{a}	0.1^{a}	0.4^{b}
	Raven's	6.9^{a}	0.7^{b}	1.0^{b}	1.2^{b}
Initial Inferences	Addition	3.6^{a}	0.6^{b}	5.2^{c}	2.4^{d}
	Anagrams	7.6^{a}	1.2^{b}	2.2^{b}	6.0^{a}
	Gambles	4.0^{a}	2.0^{b}	3.5^{a}	3.4^{a}
	Raven's	5.1^{a}	2.1^{b}	5.1^{a}	3.9^{ab}
Final Inferences	Addition	1.2	0.9	1.2	0.9
	Anagrams	2.8^{a}	0.6^{a}	2.2^{a}	4.8^{b}
	Gambles	0.9^{a}	1.0^{a}	0.7^{ab}	1.2^{ac}
	Raven's	3.1^{a}	0.5^{b}	1.3^{b}	2.6^{a}
Strategy	Addition	0.0	0.0	0.0	0.0
	Anagrams	0.3	0.0	0.4	0.3
	Gambles	0.0^{a}	0.2^{a}	0.1^{a}	0.5^{b}
	Raven's	0.4	0.6	0.4	0.6
All	Addition	10.1^{a}	7.6^{b}	13.3^{c}	11.7
	Anagrams	13.2^{a}	4.8^{b}	6.3^{b}	15.6^{a}
	Gambles	9.7^{a}	6.7^{b}	9.3^{a}	11.5^{a}
	Raven's	17.7^{a}	4.6^{b}	10.3^{c}	15.0^{a}

Note—Within a row, means with the same letter superscripts (a, b, c, or d) are not significantly different ($p < .05$); different letter superscripts indicate significant differences by a Duncan's multiple range test. Also, note that All includes every statement. The number of All statements does not match the column sum because three statement categories are omitted, as explained in Footnote 6.

categories and for the total number of statements. The results of the contrasts among the four protocol methods ($p < .05$, two-tailed) are reported in Table 1.

Recall that we cannot assume that one protocol method is more valid than another. Thus, we cannot easily attribute observed differences across the protocol methods to omission in one method or intrusion in the other. We can, however, compare methods on the basis of the information available to the subject and examine the plausibility of forgetting and intrusion.

Forgetting. Because there is no stimulus to cue reconstruction in the response-cued condition, these protocols should be least susceptible to fabrication and thereby isolate a relatively pure effect of forgetting. All tasks had significantly fewer statements in the response-cued condition than in the concurrent. Raven's task showed the largest reduction in total statements from concurrent to response-cued, 74% (17.7 to 4.6); addition showed the smallest, 25% (10.1 to 7.6).

Because higher level processes are usually more unique and produce less confusable results, we expected more forgetting of lower level processes (e.g., perception and initial inferences) than of higher level statements (e.g., final inferences and strategies). At least this should be true of such algorithmic tasks as addition and gambles, whereas, for the anagrams and Raven's tasks, higher level inferences may not be less confusable or retained longer. Since the frequencies of perception and strategy statements are too low to provide conclusive evidence, we focused our testing on the lowest and highest levels of inference. To measure forgetting, we computed the difference between the response-cued and concurrent frequencies and then normalized by the concurrent frequency. This measure of percent forgotten was predicted to be greater for initial inferences than for final inferences, at least for the two algorithmic tasks. The respective percentages were 82% and 30% for addition and 52% and −3% for gambles, as predicted. (Negative values should be interpreted as zero forgetting.) For the two nonalgorithmic tasks, the same percentages were 85% and 80% for anagrams and 60% and 83% for Raven's. Thus, where expected, a greater forgetting percentage occurred for lower inferences. A similar result was reported by Rip (1979).

Fabrication. The most likely place for fabrications to intrude into retrospective protocols is the stimulus-cued condition, because it so easily permits reconstruction. Furthermore, algorithmic tasks, such as addition and gambles, are the most susceptible to reconstruction. Again using the concurrent method as a benchmark, we found decreases in the total number of statements for all tasks. However, as predicted, these were smaller for the two algorithmic tasks (4% for gambles and 32% for additions) than for the two toolbox tasks (41% for Raven's and 54% for anagrams). Only the drop for gambles is not statistically reliable. By pointing to reconstruction as a possible source of intrusions in retrospective protocols, these data also suggest that cuing a retrospective protocol by the original problem should be avoided. Especially for such algorithmic tasks as addition and gambles, the danger of intruded reconstructions may exceed any benefits from a richer cue to prompt the recall of the original process.

Summary

The data in Table 1 strongly suggest that retrospective protocols fail to provide a veridical reflection of the primary process. There seems to be widespread forgetting in the response-cued condition and substantial fabrication in the stimulus-cued method. Although we cannot isolate their causes, these findings seem persuasive on their own and are fully in accord with the dim view of retrospective protocols expressed in the literature. This is not to say that there are not some situations in which retrospective reports are valid (e.g., Weitz & Wright, 1979), but these data generally pose major risks to validity.

CONCLUSION

Our experiment has revealed substantial reactivity attributable to generating a concurrent protocol. Biehal and Chakravarti (1989) have recently reported similar results, specifically finding changes in a choice process induced by protocol generation.[7] This empirical evidence, coupled with the four possible causes of reactivity identified earlier, strongly suggests that the current state of theory is not yet adequate to specify tasks in which protocol generation is benign.

Until a theory of protocol generation can fully specify the conditions of validity, the only assurance of nonreactivity is empirical. This involves adding a silent control group to the experimental design and comparing accuracy and RT between the silent and verbalizing conditions. The additional cost of such a check may be justified both by the unpredicted reactivity found in our addition and gambles tasks and by the pleasant surprise in discovering that a doubtful task like Raven's matrix is nonreactive. Furthermore, such empirical tests may eventually provide a census of reactivity that catalogs tasks in which reactivity is negligible or inconsequential, relieving future experimenters of the burden of an empirical check.

Channeling Invalidity

If invalidity is found, instructions can channel it such that the overall threat to research goals is minimized. For most tasks, there seems to be a natural hierarchy of invalidities: Disruption of the primary process is unacceptable, omissions in the verbal report are less serious, and a prolonged RT is usually inconsequential. Following this hierarchy, instructions can channel any irreducible invalidity into its least damaging form.

First, instructions can sacrifice speed for both completeness of the reports and naturalness of the primary process. Note that this may not be compatible with typical experimental instructions to minimize both errors and RT. Second, subjects can be instructed to preserve naturalness over completeness. For instance, it is common practice to recommend nondirective prompting when subjects lapse into silence. However, because prompts run counter to the emphasis on nonreactivity over completeness, they should be minimized in most situations. Instead, adequate training in verbalization prior to data collection will usually be preferred.

In spite of the substantial reactivity we have observed and the absence of a fully adequate theory of protocol generation, we do not conclude that concurrent verbal protocols are invalid and should be avoided. All methods risk some invalidity and tradeoff costs for benefits. On the basis of our own experience with verbal protocols and other process-tracing data (e.g., eye movements and manual responses), we believe that nothing can match the processing insights provided by a verbal protocol. Given their unique benefits, the challenge is to identify and reduce causes of their invalidity.

REFERENCES

Anderson, J. R. (1979). Further arguments concern representations for mental imagery: A response to Hayes-Roth and Pylyshyn. *Psychological Review*, **86**, 395-406.

Arnold, H., & Lee, B. (1978a). *Jumble – That scrambled word game #12*. New York: New American Library.

Arnold, H., & Lee, B. (1978b). *Jumble – That scrambled word game #13*. New York: New American Library.

Biehal, G., & Chakravarti, D. (1989). The effects of concurrent verbalization on choice processing. *Journal of Marketing Research*, **26**, 84-96.

Brooks, L. R. (1968). Spatial and verbal components in the act of recall. *Canadian Journal of Psychology*, **22**, 349-368.

Carroll, J. S., & Payne, J. W. (1977). Judgments about crime and the criminal: A model and a method for investigating parole decisions. In B. D. Sales (Ed.), *Perspectives in law and psychology: Vol. 1. The criminal justice system* (pp. 191-239). New York: Plenum.

Cochran, W. G., & Cox, G. M. (1957). *Experimental designs* (2nd ed.) New York: Wiley.

Crowder, R. G. (1970). The role of one's voice in immediate memory. *Cognitive Psychology*, **1**, 157-178.

Dansereau, D. F. (1969). An information processing model of mental multiplication. *Dissertation Abstracts International*, **30**, 1916-B. (University Microfilms No. 69-15746)

Dansereau, D. F., & Gregg, L. (1966). An information processing analysis of mental multiplication. *Psychonomic Science*, **6**, 71-72.

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review*, **87**, 215-251.

Ericsson, K. A., & Simon, H. A. (1984). *Protocol analysis: Verbal reports as data*. Cambridge, MA: MIT. Press.

Feldman, J. (1959). An analysis of predictive behavior in a two-choice situation. *Index to American Doctoral Dissertations* (1958-59), **19**, 149.

Fidler, E. J. (1983). The reliability and validity of concurrent, retrospective, and interpretive verbal reports: An experimental study. In P. Humphreys, O. Svenson, & A. Vari (Eds.), *Analyzing and aiding decision processes* (pp. 429-440). Amsterdam: North-Holland.

Fryer, D. H. (1941). Articulation in automatic mental work. *American Journal of Psychology*, **54**, 504-517.

Hitch, G. J. (1978). The role of short-term working memory in mental arithmetic. *Cognitive Psychology*, **10**, 302-323.

Hunt, E. (1974). Quote the raven? Nevermore! In L. W. Gregg (Ed.), *Knowledge and cognition* (pp. 129-157). Potomac, MD: Erlbaum.

Johnson, E. J., & Payne, J. W. (1985). Effort and accuracy in choice. *Management Science*, **31**, 395-414.

Johnson, E. J., & Russo, J. E. (1978). *What is remembered after a purchase decision?* (C.I.P. Working Paper No. 379). Pittsburgh, PA: Carnegie-Mellon University, Graduate School of Industrial Administration.

JUST, M. A., & CARPENTER, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, **87**, 329-354.

KANTOWITZ, B. H. (1987). Mental workload. In P. A. Hancock (Ed.), *Human factors psychology*. Amsterdam: Elsevier North-Holland.

KARPF, D. A. (1973). Thinking aloud in human description learning. *Dissertation Abstracts International*, **33**, 6111-B. (University Microfilms No. 73-13625.)

KAZDIN, A. E. (1976). Assessment of imagery during covert modeling of assertive behavior. *Journal of Behavior Therapy & Experimental Psychiatry*, **7**, 213-219.

MERZ, F. (1969). Der Einfluss des Verbalisiernens auf die Leistung bei Intelligenzaufgaben. *Zeitschrift für Experimentelle und Angewandte Psychologie*, **16**, 114-137.

NAVON, D., & GOPHER, D. (1979). On the economy of the human processing system. *Psychological Review*, **86**, 254-284.

NETER, J., & WASSERMAN, W. (1974). *Applied linear statistical models*. Homewood, IL: Irwin.

NEWELL, A., & SIMON, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

NISBETT, R. E., & WILSON, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, **84**, 231-259.

PAYNE, J. W., BRAUNSTEIN, M. L., & CARROLL, J. S. (1978). Exploring predecisional behavior: An alternative approach to decision research. *Organizational Behavior & Human Performance*, **22**, 17-44.

PENNEY, C. G. (1975). Modality effects in short-term verbal memory. *Psychological Bulletin*, **82**, 68-84.

QUATTRONE, G. A. (1985). On the congruity between internal states and action. *Psychological Bulletin*, **98**, 3-40.

RAVEN, J. C. (1958). *Standard progressive matrices*. London: H. K. Lewis.

RIP, P. D. (1979). The basis and extent of self-awareness in decision making. A study in consumer behavior. *Dissertation Abstracts International*, **40**, 4206-A. (University Microfilms No. DDJ80-01997)

ROTH, B. (1966). The effect of overt verbalization on problem solving. *Dissertation Abstracts International*, **27**, 957-B. (University Microfilms No. 65-9321)

RUSSO, J. E. (1979). A software system for the collection of retrospective protocols prompted by eye fixations. *Behavior Research Methods & Instrumentation*, **11**, 177-179.

RUSSO, J. E., & DOSHER, B. A. (1983). Strategies for multiattribute binary choice. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **9**, 676-696.

SABINI, J., & SILVER, M. (1981). Introspection and causal accounts. *Journal of Personality & Social Psychology*, **40**, 171-179.

SCHWEIGER, D. M. (1983). Is the simultaneous verbal protocol a viable method for studying managerial problem solving and decision making? *Academy of Management Journal*, **26**, 185-192.

SIMON, H. A. (1975). The functional equivalence of problem solving skills. *Cognitive Psychology*, **7**, 268-288.

SMEAD, R. J., WILCOX, J. B., & WILKES, R. E. (1981). How valid are product descriptions and protocols in choice experiments? *Journal of Consumer Research*, **8**, 37-42.

SMITH, E. R., & MILLER, F. M. (1978). Limits on perception of cognitive processes: A reply to Nisbett and Wilson. *Psychological Review*, **76**, 211-224.

TELL, P. M. (1971). Influence of vocalization on short-term memory. *Journal of Verbal Learning & Verbal Behavior*, **10**, 149-156.

TETLOCK, P. E., & KIM, J. I. (1987). Accountability and judgment processes in a personality prediction task. *Journal of Personality & Social Psychology*, **52**, 700-709.

TURNER, C. K. (1986). *The use of introspective reports for studying judgment processes* (Working paper). Los Angeles: University of California, Department of Psychology.

TURNER, C. K. (1988). Don't blame memory for people's faulty reports on what influences their judgments. *Personality & Social Psychology Bulletin*, **14**, 622-629.

WALKER, W. H. (1982). Retrieval of knowledge from memory: A generalization of the Raaijmakers-Shiffrin retrieval model. *Dissertation Abstracts International*, **43**, 4183-B. (University Microfilms No. DEP83-09875.)

WEITZ, B., & WRIGHT, P. (1979). Retrospective self-insight on factors considered in product evaluation. *Journal of Consumer Research*, **6**, 280-294.

WERNER, H., & KAPLAN, B. (1963). *Symbol formation*. New York: Wiley.

WICKENS, C. D. (1980). The structure of attentional resources. In R. Nickerson (Ed.), *Attention and performance VIII* (pp. 239-257). Hillsdale, NJ: Erlbaum.

WICKENS, C. D. (1984). Processing resources in attention. In R. Parasuraman and D. R. Davis (Eds.), *Varieties of attention* (pp. 63-102). New York: Academic Press.

WICKENS, C. D. (1987). Attention. In P. A. Hancock (Ed.), *Human factors psychology* (pp. 29-80). Amsterdam: Elsevier North-Holland.

WINIKOFF, A. (1967). Eye movements as an aid to protocol analysis of problem solving behavior. *Index to American Doctoral Dissertations* (1967-68), 129.

WRIGHT, P., & RIP, R. D. (1981). Retrospective reports on the causes of decisions. *Journal of Personality & Social Psychology*, **40**, 601-614.

ZBRODOFF, N. J., & LOGAN, G. D. (1986). On the autonomy of mental processes: A case study of arithmetic. *Journal of Experimental Psychology: General*, **115**, 118-130.

NOTES

[1] Rarely is a second process trace available, but for an exception see the eye fixation analysis in Winikoff (1967), also described in Newell and Simon (1972, p. 327). If a second source of process data is not available, the situation is not hopeless. One can compare the protocols from repeated trials and, using a criterion of consistency, try to identify the level of omissions and intrusions in any given trial. Also, a theory-based approach uses a task analysis to specify what information must have been used and, therefore, should appear in the protocol. This second strategy depends on the accuracy of the theory (the same theory that the protocols are testing) and on the experimenter's ability to distinguish between attended-to and automated components. Neither method is very satisfactory in that both rely critically on experimenter interpretation of the observed data. Note also that output measures alone are not sufficient. The bulk of the protocol might be highly inaccurate, yet task accuracy is unchanged because the problem solution itself is unlikely to be misreported.

[2] We acknowledge the difficulty of claiming that a task must be performed using a pictorial rather than propositional representation (Anderson, 1979). Indeed, Hunt (1974) has proposed a propositional representation and solution algorithm for the Raven task. Although both our intuition and observations of subjects' performance persuade us that this task can be safely classified as pictorial, such a determination is not critical. Our first priority is an empirical test of reactivity. Only if it is found will we worry about whether or not it can be ascribed to recoding operations.

[3] Complete instructions for the four protocol methods as well as all stimuli and task instructions are available upon request.

[4] A particularly problematical source of carryover effects is the prompted retrospective method that improperly instructed the subjects to explain why they looked at a stimulus element. Informal observation of the subjects during task performance revealed no obvious differences before and after the prompted condition. However, to test for such an effect on accuracy, the mean proportion of test problems solved was computed for all methods and tasks performed before (.756) and after (.749) the prompted method was used. This difference was not statistically significantly ($p < .05$). There also was no significant difference when practice problems were included and when the before-after difference was computed separately for each task and each method. These results argue against a carryover effect from the prompted method. However, what cannot be eliminated is the possibility of a carryover effect from the practice session in which all four methods were previewed using a fifth task. Practice with the prompted method then may have affected performance in all later sessions. This possibility can only be disconfirmed by further experimentation using a different experimental design, one between subjects or without the practice session.

[5] This possibility raises the interesting question of whether the silent or protocol condition better reflects problem-solving performance

in the natural environment. It seems plausible that the level of motivation under protocol generation may more validly reflect that of at least some real world problem solving. This possibility focuses attention on our particular definition of reactivity, namely, any change from the silent control performed under laboratory conditions. It is beyond our scope to begin to determine whether this control is itself ecologically valid.

6. Because they do not inform the tests of our hypotheses, we exclude three types of statements that were actually coded: (1) middle level inferences, (2) answers, and (3) irrelevant asides, such as "these are all hard problems." The large majority of excluded statements were middle level inferences. The protocols were coded by one judge and verified by a second judge on a 10% sample. The interrater reliability for each task was: addition, .93; anagrams, .86; gambles, .76; and Raven, .83.

7. Although they found no differences in output (the alternative chosen), they used a protocol collected from all subjects during a second choice to infer aspects of the process used to make the earlier choice. Comparing the choice processes of the verbalizing and silent groups, they found reliable evidence that the former organized their choice process more around product attributes and less around brands. Thus, they found differences in the process used to perform the primary task that could be attributed to protocol generation.

(Manuscript received February 29, 1988;
revision accepted for publication March 20, 1989.)