NOTICE

The quality of this microform is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

Reproduction in full or in part of this microform is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30, and subsequent amendments.

AVIS

La qualité de cette microforme dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

La reproduction, même partielle, de cette microforme est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30, et ses amendements subséquents.

An Information Processing Approach to Judges' Agreement and Disagreement Patterns When Encoding Verbal Protocols

by

Jocelyne G. Schael

Thesis submitted to the School of Graduate Studies
of the University of Ottawa
in partial fulfillment of the requirements
for the degree of Master of Arts in Education

ISBN 0-315-60070-5

UNIVERSITÉ D'OTTAWA
UNIVERSITY OF OTTAWA

Acknowledgements

This research could not have been completed without the assistance of many people. The support of family in pursuing my studies was invaluable. Especially my husband who helped with graphs and editing of the text.

The excitement, challenge, and diversity permeating the field of cognitive science were revealed by my thesis advisor, Dr. Jean-Paul Dionne. His capacity to motivate students to pursue research is exceptional. His advice and guidance were essential in the design of this study.

I would like to thank Manjari Gopal for her patience in proofreading the manuscript. Special appreciation is also extended to the participants in this study.

Abstract

An Information Processing Approach

to Judges Agreement and Disagreement Patterns

When Encoding Verbal Protocols

The technique of protocol analysis, in exploring cognitive structures and processes, evolved with the information processing model of the human mind. The reliability of encoding the concurrent verbal protocols remains a major concern which has received little attention. The present research investigates the basis of agreement or disagreement among judges when applying a coding scheme to protocols.

An attempt is made at answering the following questions: How can reliability be assessed? Is the final decision reached congruent with the reasons given to arrive at that decision? Are there any particular qualities in judges which may have an effect on the encoding procedure? Does the type of task (well-defined or ill-defined) have an effect on the encoding procedure? It was predicted that judges who are more skilled at categorizing protocols would use certain strategies that less skilled judges would not. The sample consists of 20 university graduates, 10 with experience in protocol analysis and 10 without, each balanced for sex. Judges were asked to "think-aloud" while encoding protocols with a grid of categories provided to them. The level of agreement on the final decisions was analyzed under several criteria. When calculated from the codes written by the judges on the protocols the overall

agreement among judges was 61.4% but when calculated from the verbalized responses collected during the execution of the encoding task, the agreement was 85.1%. The marked difference indicates that the codes selected by the judges are not always coherent with the reasons invoqued or that the grid categories are ambiguous. The results of the study showed no perceived differences between experienced and inexperienced judges. A gender difference was observed reflecting a richer knowledge base relative to the content of the protocols, which required the planning of a three course meal. Criteria to improve the quality of a grid and the reliability of coders are suggested.

Table of Content

Appendices

List of Tables

List of Figures

Chapter 1

Introduction

The increased emphasis on the use of verbal reports as data to study cognitive processes has produced a powerful research tool - protocol analysis. Enormous amount of data are gathered which must be reduced for analysis. First, the data are transcribed, then categorized and interpreted. There are basically two goals in protocol analysis:

- the protocol information is taken explicitly as it is in many problem-solving protocols and a step-by-step rationale can be reconstructed and even computer simulated in some cases;
- the verbalizations are analyzed to generate a model or a theory of the underlying cognitive processes.

Verbal reports are collected while subjects are performing a task. The task presented can be of two different types:

- One where all the information needed to execute the task is available and the solution is known, similar to geometry and physics problems encountered in high school or at university level (well-defined problems);
- and one where the subject must contribute to the explicit definition of the task and where the goals could be multiple, similar to what may be encountered in real life (ill-defined problems). Solving problems representative of every day life may not be as clear as solving a logical problem (Newell & Simon, 1972), and developing a replicable model is no small

feat. Trying to trace underlying cognitive processes used while performing activities requires inferences. The present research focuses on protocols derived from ill-defined problems.

The main concern is with investigating the reliability of assigning categories, or their representative codes, to verbal protocols. Consistency is found when independent observers are able to analyze data from the same task and reach the same conclusions. Ericsson and Simon (1984) have elaborated on the technique of protocol analysis, giving many details on the procedure involved. However, little has been developed for testing the reliability of the procedure.

The general objective is to study the response agreement and disagreement among judges when encoding verbal protocols. When reliability is mentioned in research reports it is usually obtained with 2-5 judges, which is insufficient to detect patterns of agreement and disagreement. In the present study responses from 20 judges with varied backgrounds provide data for pattern identification. The key questions addressed are: How reliable is the encoding procedure? Is the final encoding congruent with the reasons given to arrive at a decision? Which characteristics have an effect on the encoding procedure?

The evolution of the protocol analysis technique and its application to the field of education are reviewed in chapter 2. The research rationale and design are also introduced. In chapter 3, the procedure used to obtain concurrent verbal reports is developed and the environmental conditions, training, and data analyses plan are described. This is followed by the

results interpreted and summarized in the light of the research questions in chapter 4. The results indicate that agreement in interpretations of verbal protocols is possible with minimum bias from judges' backgrounds as long as the grid is non ambiguous, the codes complete and discriminating, and practice sessions are provided. Recommendations to improve the reliability of the coding technique in protocol analysis is presented in Chapter 5. Criteria for developing a reliable grid, choosing judges, and designing an encoding task were suggested from the findings. Suggestions on the direction of future research in the area conclude the report.

Chapter 2

Review of the Literature

Questioning the nature of thought is as old as history itself. Considering the extensive literature on the topic, it was necessary to focus the review on relevant research which provided insights into the inquiry.

Approaches to the Study of Cognitive Abilities

Studies in the identification of mental abilities repeatedly lead to the definition of the abilities being observed. Different conceptions of the mind have led to different priorities concerning the aspects relevant to observe. During the opening decades of the century, when scientific methods became the accepted way of studying human behavior, introspection was discarded as not being objective; researchers' main concern became the analysis of overt behavio (S-R theory) with minimal description of the mind. It was believed by scholars such as Pavlov, Skinner, Thorndike, and Watson that individuals were passive reflectors of various forces and factors in their environment. The psychometricians of that era considered their main goal to be the "measurement" of intelligence. Measuring mental abilities can be traced to the work of Binet in 1905 which gave rise to many series of standardized achievement tests. Internal mental processing was limited to the observation of its product or resultant behavior such as answers to test

questions. The objective was to predict as accurately as possible the level of "performance" the subject could reach. In the sixties, with the advent of computers, renewed interest in the workings of the mind was observed (Anzai, 1979; Greeno, 1974; Newell & Simon, 1972; Norman, 1970), and a new approach to cognition was born. The mind was then viewed as being capable of processing information instead of only storing information; not only the amount of information had to be measured but the processes used to accumulate this information had to be identified. Moreover, intensive research efforts in neuroscience and neuropsychology joined the rekindled interest in cognition in providing physiological evidence to validate cognitive theories (Arbib, 1982; Churchland, 1988; Gardner, 1985; Martindale, 1981; Vocate, 1987), while artificial intelligence contributed in studying the architectures and processes of thinking and knowledge (Anderson, 1982; Minsky, 1975; Rumelhart & Norman, 1985; Shank & Abelson, 1977; Simon, 1976).

There is now increased emphasis on identifying the processes underlying thinking, be it reasoning, learning, memorizing, or intellectual abilities. Sternberg (1984) has developed a theory of intelligence in which he proposes that assessment of mental abilities be done in the light of cultural background, and individual differences existing in knowledge structures. Messick (1984) describes two reasons why measurement of intellectual abilities should not be stripped from prior knowledge, culture, and individual differences. These are: measurement is a dynamic process, and personal and situational circumstances delimit score interpretations. In

support to the first reason, J. R. Anderson (1982), Ericsson & Simon (1984), Messick (1984) and, Resnick and Ford (1981), consider comprehensive assessment of learning as a "dynamic process" which is subjected to continuous changes as the individual knows more about his world. Many measurement specialists (Cronbach, 1984, 1986; Glaser & Pellegrino, 1978; Thorndike, 1986) also agree with a "dynamic view" of testing. A recent tendency in testing concentrates on the "functional dimensions of developing competence and expertise" (Messick, 1984) and analyzes the type of instruction which can enhance individual's performance qualitatively. Messick (1984) considers that educational measurement would be enhanced by including developmental differences in subject-matter learning and performance.

> "This does not mean simply making achievement tests progressively more difficult at intermediate and advanced levels, but rather constructing items and tasks that tie the sources of difficulty or facilitation to the cognitive processes and structures operative at each level." (p.221)

He refers to cognitive psychology for the identification of cognitive processes and structures because cognitive psychology conceptualizes the acquisition of knowledge and cognitive skills in developmental terms.

Relative to the second reason invoked by Messick (1984) in favour of studying abilities within context, not many would dispute the fact that measurement of educational ability and achievement is affected by intrapersonal and environmental factors. Cognitive constructs or traits extracted from factor analysis and correlational approaches would benefit from detailed analysis of the underlying processes and structures

(Anderson, 1986; Glaser & Pellegrino, 1978; Simons, 1971). Product measurement provides little insight into the actual thinking processes used. Cognitive traits derived from more comprehensive data, including individual differences, may exhibit clear regularities without any current theory to describe or explain them (Simon, Langley, & Bradshaw, 1981). Such regularities may lead to the induction of new paradigms (Ericsson & Simon, 1984).

To identify cognitive processes, performance of simple and complex tasks have been analyzed. Based on the information processing architecture of the mind, many researchers in the areas of memory and problem solving have discussed the organization of knowledge and its implementation when performing a task.

J. R. Anderson's (1982) framework has had a significant impact on research. Basically, he asserts that mental activity can be described by productions being executed. Productions are "if – then" propositions (rules). When the condition part of a production (the "If ...") matches a memory pattern, the production can execute its action part (the "then ..."). The action part often produces changes in memory, which have the effect of changing or combining memory patterns. His position on skill acquisition is that learning begins as the application of declarative knowledge which is represented as a propositional network. Declarative knowledge is memorizing or knowing <u>what</u> is related to a particular topic: it can be represented as productions. Procedural knowledge is the knowledge for carrying out mental

activities: knowing how to use a knowledge. This procedural form undergoes a process of continual refinement to make the productions more selective in their range of applications. This conception of thinking and learning has been adopted in many other research (Ashcraft, 1982, 1983; Ashcraft, Fierman, & Bartolotta, 1984; Baroody, 1983; Chaiklin, 1984; Lesgold, 1988).

Rumelhart and Norman (1985) reviewed existing representational theories and concluded that knowledge may be directly represented, represented in the form of propositions or may be indirectly coded, inferred or otherwise generated in real time using schematas or scripts. Theorists have coined the term schemata or script to refer to the memory structures that incorporate clusters of information relevant to comprehension (Minsky, 1975; Rumelhart & Ortony, 1977). A primary insight of schema theories is that people do not just have isolated facts in memory. Information is gathered together into meaningful functional units. For example, the knowledge an individual has about "banks" is stored in memory so that when someone mentions going to the bank one is readily prepared to interpret comments about wicket, safe, or cheque. A script is a type of schema that specifies a list of actions that people carry out in stereotypical situations. Results from this type of research support the evidence that the more experiences one has in forming schemata, the more accurately they will guide comprehension. With more experience one will better understand the information, which suggests that "high knowledge" subjects have more useful memory structures than "low-knowledge" subjects. Thus, expertise

in a domain enables a person to distinguish what is important from what is trivial and to commit to memory more easily the information that is important.

Accumulating studies on the functional differences between low-knowledge and high-knowledge learners, or between novices and experts in a field (Kaye, Post, Hall, & Dineen, 1986; Russel & Ginsburg, 1984; to name only a few) show how restructuring becomes more dominant as experience increases within a particular knowledge domain. Studies within the expert-novice paradigm show how some tasks involve a number of subprocesses whose presence, organization, and degree of automaticity vary with expertise.

J. R. Anderson (1982), Rumelhart & Norman (1985), and the mentioned studies on functional differences provide explanations of human thinking in terms of basic information processes. These explanations can be tested for validity by comparing them with the course of thought of individuals, as revealed, by verbal reports. Ericsson and Simon (1984) have published a comprehensive report on how to use verbal reports as data. The present research focuses on this particular methodology, called protocol analysis, for identifying cognitive processes and structures.

Protocol Analysis as a Research Method for Identifying Cognitive Processes

Origin and nature of protocol analysis

How can one identify cognitive processes and structures in a person's mind? This question has been addressed by many researchers in the areas of memory and problem solving; in particular, Newell and Simon (1972) used extensively the technique called protocol analysis. Their objective was to develop computer programs which would emulate the problem solving processes of the human mind; therefore, observation of people solving problems became a major concern. They simply asked people to "think-aloud" as they attempted to perform a task. Concurrent recordings of the sessions were later transcribed to form the basis for analysis.

The Newell and Simon (1972) approach to information processing rests on the following premises: (1) a cognitive process consists in processing information: a sequence of internal states transformed by processing new information; (2) information is processed in different units: sensory register (SR), short term memory (STM), long term memory (LTM); (3) information processing consists in relating information acquired in SR with information present in LTM. Information processing is managed by two types of processes: automatic and controlled (Shiffrin & Schneider, 1977).

To these premises, Caverni (1988), added the following characteristics: (1) STM is put into use only with controlled processes, when automatic

processes are used there is a direct link between SR and LTM; (2) subjects pay attention to the information in STM; (3) one has direct access to information in STM and since its capacity is limited (Miller, 1956) only recently acquired information is accessible; and, (4) information in LTM may be recovered for processing but must first be transferred to STM.

Though the methodology is described in detail in Chapter 3, a few summarizing remarks are necessary at this stage. The recordings of people "thinking aloud" while performing a task provide the researcher with an enormous amount of data. How can the data be reduced without threatening its validity? Ericsson and Simon (1980, 1984) provide guidelines on data-gathering and data-analysis methods using verbal reports. First, tapes are transcribed verbatim. Second, the transcripts are segmented into individual statements or ideas. This simple partitioning is seldom problematic (Ericsson & Simon, 1984). The third step requires that the segments be categorized. Protocols may be categorized in two different ways: in terms of "low-level of inference" where the process is manifested by clear, explicit verbalizations; and in terms of "high-level of inference" as more or less task-independent theoretical entities.

Despite the success of the verbal report technique, three major objections keep resurfacing which have been addressed thoroughly by Ericsson and Simon (1984). These are: (1) incompleteness of reports; (2) reporting might alter performance; (3) verbalizations are epiphenomenal, idiosyncratic, and encoding is not strong and valid.

In debating whether a verbal report is complete or not, protocol analysts do not claim protocols describe completely the cognitive processes under study. Rather, they claim "that verbal reports are based on the information currently in STM or on information previously in STM that has been fixated in, and can be retrieved from LTM" (Ericsson & Simon, 1980, p.236). Thinking-aloud protocols give glimpses of a particular process (Dobrin, 1986; Ericsson & Simon, 1984; Randall, Fairbanks & Kennedy, 1986; Steinberg, 1986) the one cognitively controlled or focally attended while performing a particular task.

Ericsson and Simon (1984) investigated the effects of concurrent verbalization on performance. They have reviewed empirical studies of verbalization and concluded that when the assumptions of the information processing model are not violated, the studies showed no evidence that verbalization changed the course or structure of the thought processes. In most cases, the same problem is solved in the same amount of time whether talking aloud or not (Karpf, 1973). Other studies (Bower & King, 1967; Fidler, 1983; Gagné & Smith, 1962) show that talking to oneself while working may even improve decision-making in some cases. Brehmer (1974) found no differences between task performances that required subjects to state their rules and task performances that did not. Fidler (1983) investigated concurrent, retrospective, and what he calls "interpretative" verbal protocols. The results indicated that concurrent verbalizations did not change the internal consistency of the processes but instead delayed the process so that it required more time in comparison with the same task

without verbalizations.

The effects on performance, present in other studies, were often due to characteristics of the task (Ericsson & Simon, 1984). Here, distinction must be made between automatic and controlled cognitive processes (Shiffrin & Schneider, 1977). Subjects who are expert performers on the assigned task may not provide enough details for the objective of a particular study; as processes become automated, less and less information becomes available about them. As Lesgold (1988) explained, experts have difficulty verbalizing directly a set of production rules. Novice performers use more cognitively controlled processes which require their attention while performing. These processes have access to the very limited capacity of STM for its inputs and outputs (Caverni, 1988; Gilmartin & Newell, 1976; Simon, 1976) thus making verbalizations more comprehensive.

The objection related to cognitive processes being inaccessible, is not so strong since most studies in which inconsistency has been observed or claimed, were using retrospective, not concurrent verbal reports (Ericsson & Simon, 1980). In retrospective reports, data are not descriptions of a trace in STM; identifying the steps followed may be contaminated by inferences or restructuration of the solution. Fidler's study (1983) shows that concurrent verbalizations are a more reliable and desirable form of data than retrospective reports. Since a pervading notion has existed that verbal reports are at best informal procedures, investigators have tended to overlook the methodological questions related to the collection of data and data analysis techniques. This has resulted in legitimate and illegitimate

forms of the technique (Anderson, 1986). Instructions to verbalize are very important as to what kind of report will be collected. Ericsson and Simon (1984) provide guidelines for experimenters in the utilization of the method in a reliable and valid manner.

Nisbett and Wilson (1977) tried to discredit verbal reports by arguing that people have no ability to observe directly the workings of their own minds - no direct access to higher order mental processes. Much evidence to the contrary followed (Caverni, 1988; Ericsson & Simon 1980; Smith & Miller, 1978; Steinberg, 1986; White, 1980). The main criticism directed at Nisbett and Wilson was that no distinctions were made between thinking-aloud protocols, retrospective responses to probes, and the classical retrospective report of the psychoanalyst. Ericsson and Simon (1980, 1984) provided substantial empirical support for using verbal thinking aloud protocols as data, as long as the verbalization is concurrent with the task performance, and where information is mentally represented in a language (verbal) code, or verbal recording is possible without additional processing.

Also, more recently, model-building based upon protocol data has been attacked by Cooper and Holzman (1983) as a poor method of understanding underlying cognitive skills in writing. Their essay contains most of the serious (and recurrent) criticisms levelled at protocol methodology. To restate them: (1) Using protocols to study ill-defined processes is impossible until the precise nature of the processes has been defined. In

defending protocol analysis, the methodology is not to test theories but to build theories or enrich existing ones. When exploring a research question without a predefined set of hypotheses, it may lead to the identification of new research questions. (2) Writing processes are too complex to be captured in protocols. This charge has no empirical basis - not one study was found to support this claim. Higher-level activities can be made-up of numerous subprocesses that use STM for input and output. (3) Cooper and Holzman (1983) talk more about the futility of comparing experts and novices and its educational value. As previously remarked, it is easier to observe the development of knowledge representation and their flexible restructuring in the gradual transition from novice to expert performance. When comparing expert-novice protocols some tasks involve a number of subprocesses whose presence can be the source of difficulty for novices and each of these subprocesses may constitute a skill worthy of separate instructional effort. Steinberg (1986) says that in the last few years writing theorists have learned many things about the process of writing by process-tracing techniques like protocol analysis. One has to look at Scardemalia and Bereiter (cited in Steinberg, 1986), Hayes and Flower (1981, 1983) to see how profitable protocol analysis is in describing "knowledge telling".

The Need to Assess the Reliability of the Encoding Procedure

Objective data analysis, with inter-coder reliabilities of .8 to .9 has been reported with several problem tasks in mathematical logic. These studies (Newell & Simon, 1972) and studies on puzzle solving (Tower of Hanoï Protocol, cryptarithmetic protocols, the Eight Puzzle, all cited in Ericsson & Simon, 1984) refer to coding when the information is very descriptive and represents low-level inference. Categories using low-level inference on the processes involved should be more reliable than categories referring to high-level inference (Ericsson & Simon, 1984).

Protocols analyzed in terms of high-level inference raise an important methodological issue, the encoding of processes. Categories are usually derived from a theoretical framework and assigned to protocol statements with as much objectivity as possible. Care should be taken to avoid subjectivity since complex processes are inferred from brief traces in STM. As Bereiter and Scardemalia (1983) caution, "Ultimately it is the investigator's description, not the verbal report, that must be judged true or false" (p.13). The categories or processes must be designed to be complete and as mutually exclusive as possible and they must be described very explicitly to enable independent judges to identify them accurately (Ericsson & Simon, 1984). Computer-based systems have been developed for automatic and semi-automatic encoding (Bashkar & Simon, 1977; Waterman &

Newell 1971, 1973) which are highly promising but more developmental effort is required before they can be used to encode high-level processes from ill-defined problems.

There exist few studies where inter-coder reliability has been assessed systematically for the encoding of ill-defined problems. Typically, in encoding verbal protocols, only two to five judges are used to encode protocols and disagreements are resolved by invoking the decision of a third rater. Review of studies using protocol analysis does not provide much information as to how the inter-coder reliabilities are measured and no mention is made of intra-coder reliability. Ericsson and Simon (1984) mention four studies in which assessment of codings is reported although no more is said on how it was done. In these studies, categories on the decision process have been developed. They were among the first studies using protocol analysis (collected while performing an ill-defined task) to derive task independent, complex categories, based on a theoretical framework. Assessment of reliability for that type of data was not yet established.

The first attempts at assessing reliability of encoding were very brief reports as can be seen in the following studies. Goor and Sommerfeld (1974) compared problem-solving processes between creative students and non-creative students. Two trained judges were used to read the protocols and to classify the segments into one of seven categories. To obtain an index of consistency for the two judges, the responses on a sample of 12 protocols were independently categorized by each judges. There was agreement on 94% of the 2400 responses. No more is given on the index of consistency.

In Klinger's report (1974) two student judges who were uninformed concerning the hypotheses, marked those utterances on verbal protocols that indicated evaluation and control of attention. Written instructions included examples as well as description of the two thought categories. Interscorer reliabilities comparing the judges' codings with the coding of a third judge calculated for the 27 protocols, were .82 for evaluation scores and .96 for attention scores. Klinger has provided no information on how inter-coder reliability was assessed.

Haines (1974) has attempted to evaluate the reliability of process descriptions derived from protocols. He used a decision tree model to describe the processes. The decision tree model assumes that the subjects make ordered series of tests on the alternatives and depending on the test outcome (yes or no) the subject performs further tests and finally reaches a decision. Haines asked judges to derive decision trees from the same protocols independently. There were considerable differences between the trees and Haines concluded that the tree models could not be derived reliably from the protocols. Haines' assumption that a single model underlies a subject's processes is questionable, the protocol rather than the model should be used in the evaluation.

A more recent study (Svenson, 1983), used scaling with verbal protocols. The primary aim was to study two evaluative components of the information treated in a decision process, whether they were negative-positive and general-specific. Protocols were first coded and the judges selected (no mention of how many) evaluative statements from the protocols

to be rated in two subsequent experiments. There is no mention of how the reliability assessment of this first coding was done. Svenson used 16 judges in one subsequent experiment and 30 in another, to judge these evaluative statements sampled from verbal protocols. The second set of judges had to rate the specificity of the evaluative statement on a scale 0-10 where 0 meant that the statement could not at all be used for the attribute in question. The number 10 indicating that the statement was perfectly fitted for characterizing the attribute. A similar procedure was used to rate positive and negative statements. Svenson concluded that conventional interpretations of verbal protocols (encoding) can be verified in additional studies using several judges to provide data which can be treated statistically. There is no evaluation of the categorisation of protocol statements.

Reliability of encoding has been assessed with well-defined problems and proven to be objective (Ericsson & Simon, 1984). Reliability of encoding complex processes from performances on ill-defined tasks has received little attention to date. The assessments made are few and non systematic; they are usually done using two to five raters; minimum information is provided as to how the inter-coder reliabilities are measured (Goor & Sommerfeld, 1975; Haines, 1974; Klinger, 1974; Svenson, 1983).

Statement of the Research Problem and Research Questions

Studies of mental abilities and knowledge representations (Anderson, 1982; Messick, 1973, 1984; Minsky, 1975; Rumelhart & Ortony, 1977; Shiffrin & Schneider, 1977; Sternberg, 1984; to mention only a few) have shown the dynamic of thinking. This type of research has also emphasized the importance of individual and environmental factors when studying mental abilities. Based on the assumption that human thinking is information processing, protocol analysis was developed (Ericsson & Simon, 1984; Newell & Simon, 1972). The technique has proven successful (Bereiter & Scardamalia, 1983; Dobrin, 1986; Fidler, 1983; Hayes & Flower, 1981, 1983; Steinberg, 1986) and observations collected from concurrent verbal reports reflect "dynamic" processes and individual and environmental differences. The objectivity of the analysis rests on the reliability of the encoding process.

The primary concern in the present research is to investigate the basis of agreement or disagreement among many judges when encoding protocols. The analysis of verbal protocols is done through a grid of categories reflecting the underlying processes and structures displayed during the performance of a task. The reliability of this coding remains a major concern which is of crucial importance in the analysis of this type of data. Virtually no research to date has addressed the following issues: (a) how can reliability be assessed? (b) is the final decision congruent with the

reasons given to arrive at that decision? (c) are there any particular qualities in judges which may have an effect on the encoding procedure? (d) does the type of task (well-defined or ill-defined) have an effect on the encoding procedure? Therefore, in the present study, several judges with varied backgrounds, were asked to code independently the same protocols which were collected while performing an ill-defined task. Increasing the number of judges, relative to previous studies, allows the identification of patterns of agreement and disagreement as well as the assessment of reliability of encoding.

In order to address the first issue, reliability would be perfect if, given the same protocol, the judges' coding would always yield the same answer. The present research investigates the agreement/ disagreement among a larger group of judges than what has currently been done. Investigating patterns of agreement/disagreement through protocol analysis and percentages of agreement is likely to have implications for assessment of reliability of encoding in the future. It is expected that within high level agreement groups variables can be identified which will increase reliability of future research.

The second issue refers to the congruence between the final decision and the reasons given to arrive at that decision. The information processing leading to a final decision will be studied through "think-aloud" protocols and compared to the final choice. Montgomery (1983) argues that a decision process involves active structuring and restructuring of the decision situation to come to grip with the problem. These structures often take the

form of hypothesis testing activity or justification for a choice. Therefore, these will be extracted from the judges' protocols. A possible outcome would be that the reasons provided will have a matching code then there would be congruence between interpretation of statement and the coding scheme. Depending on what is observed, whether judges apply the same code for different reasons or different codes given the same reasons, the problem could reside with the grid or the judges' characteristics.

In regard to the third issue, are there any particular qualities in judges which may have an effect on the encoding procedure? It is believed that the experience of the judges at encoding influences the reliability of the technique. To test the influence of the judges backgrounds, judges with experience in protocol analysis and judges without experience have been chosen to perform the task. More skilled judges are expected to use certain kinds of processes that less skilled judges are not (Ericsson & Simon, 1984). Ericsson & Simon (1984) also mention that encoders may encode with a bias toward their own preferred interpretation. Therefore judges with varied backgrounds are likely to reduce the global bias in inferences made. A relevant question comes to mind: is there more agreement within the group with experience at encoding verbal protocols or within the group without experience at this task? Due to the nature of the task, the samples were also balanced for gender, and there one can raise the question: is there more agreement among judges within the same gender?

The final issue is directed at the effect of the type of task (ill-defined) on the encoding procedure. It is hypothesized that an ill-defined

task, is more likely to generate divergences in coding; the reliability is then likely to be underestimated. In this case, high reliability would provide stronger evidence in support of the technique. The task presented is more like the kind of problem found in every day life, where one has to contribute to the definition of the problem. It is a mixture of two levels of encoding: high-level encoding and low-level encoding. With high-level encoding, where it is required to retrieve extensive information stored in LTM (the correspondence is more abstract), lower reliability is expected. With low-level encoding, where all the information is available perceptually or explicitly in the text (the correspondence of reports with stimuli can usually be established), high reliability is expected.

The general objective of the study is to identify the patterns of agreement/ disagreement among judges when applying a grid of categories of information processing to the same verbal protocols. More explicitly, the objectives are:

1. To estimate the level of agreement/disagreement among judges with varied backgrounds (levels of experience, gender) performing an encoding task.

2. To study the agreement/disagreement among judges by contrasting final codes (application of categories) with reasons invoked for their decisions.

In summary, this research seeks to provide guidelines for testing the reliability of encoding verbal protocols by studying agreement/disagreement patterns among judges performing an encoding task. The variables under which higher levels of agreement are found will provide direction for increasing reliability in future research. A higher level of agreement is expected among judges with experience in protocol analysis. The gender effect is also analyzed. Congruence between interpretations of verbal statements and chosen codes is studied through protocol analysis; if congruence does not occur, it will be possible to identify where the problem lies. Another source of information for pattern identification comes from the ill-defined task to be encoded, it is expected that the level of encoding will generate different patterns of agreement/disagreement.

Chapter 3

Methodology

In the literature, when exploring the reliability of encoding verbal protocols, a wide range of variables were chosen. However, reliability studies which have been done so far have been limited mostly to well-defined problems. With ill-defined problems, high level of agreement among coders, is more difficult to obtain.

Design of the study

Population and sample

University graduates and professors represent the pool of candidates from which judges are likely to be selected for this type of research: assessing the reliability of information extracted from verbal protocols. Therefore, 20 graduates and professors from varied fields of specialization were asked to participate in the experiment. The disparity between judges' backgrounds is likely to produce more conservative estimates of agreement indices. They were divided into two groups having different experience in protocol analysis. The criterion for designating a "judge with experience" was one who had studied the technique and had used it in a research setting. Following this definition a "judge without experience" was designated as one who had neither studied nor used the methodology. The

sample consists of 10 judges with experience in encoding protocols (E) and 10 without experience (NE), each subsample balanced for sex. For the purpose of this study the term "judges" is used when referring to the subjects selected for the present study and the term "students" refers to the subjects whose protocols were being encoded.

Task

The individual sessions were held in a quiet room. Demographic information (age, sex, field of specialization, etc.) was collected from each judge. Each was presented with three protocols and the coding scheme developed in the Gillespie (1988) study. In the training phase of the experiment, judges were asked to encode Protocol 1; this gave them the opportunity to familiarize themselves with the "thinking-aloud" technique as well as the task. Throughout the practice session the experimenter answered questions reminding judges to continue thinking aloud. Judges were instructed to verbalize every detail of their thoughts, including the particular information they were looking at, how they were evaluating the different pieces of information, and the reasons leading to their decisions. Detailed instructions given to the judges are presented in Appendix A.

The main task consisted in encoding Protocols 2 and 3. The protocols were presented in their original form, judges were allowed to assign multiple codes and were asked to write their final answers next to each statement. The experimenter recorded the verbalizations along with observations of non-verbal behaviors, manipulations, interruptions and time lapses. The tapes

and observations were later transcribed and formed the data base for analysis. Total interview time was approximately 80 minutes per subject.

Stimulus material

The information presented to the judges included the following pieces of information from the Gillespie (1988) study:

-excerpts from 3 protocols,
-grid of behavioral categories to encode these verbal protocols,
-instructions given to the student about their task,
-students' work sheets: time sheet, menu sheet, food list.

The protocols to be encoded came from students in a College of Agriculture and Technology who were asked to "think-aloud" while performing a curriculum task. This task included three sub-tasks: designing a three course menu for six people, estimating the cost and amount of food needed, and estimating the time needed to prepare each dish. The protocols to be encoded by the judges represented a practical situation where the components of the problem and the varieties of solution were numerous. Since the problem solver had to contribute to the definition of the problem, it can be characterized as an ill-defined problem. Protocols from three students were encoded by the judges in the present study and can be found in Appendix B in their original format. Protocol 1 was used in a practice exercise, the other two for data. A feature of Protocol 3 is that it was taken from a freshman at the College, while Protocol 2 was collected from an expert.

The grid was generated from the analysis of the students' protocols (Gillespie, 1988). Thirty-nine mental activities performed by the students while solving the menu planning task were extracted from their protocols. These operations were grouped under 7 broad categories:

P: Problem identification and representation

S: Solution related activities

M: Memory related activities

C: Changing the conditions of the problem

R: Reasons for particular food choices

A: Assessing and monitoring activities

E: Emotional reactions.

A detailed grid can be found in Appendix C.

The instructions given to the students contained a general description of the task. A second paragraph gave specific explanations of the three sub-tasks with an example on how to fill out the time sheet. A reminder on the "thinking aloud" technique completed the directions.

Four work sheets accompanied the instructions. A blank sheet was provided for writing down the menu. Two pages of food items were presented and divided into perishable items and pantry stock items; there was a column for checking off food items they planned to use and a column for estimating the amount they needed. The perishable items also had a column for cost estimates. The last sheet was a time sheet on which the time needed to prepare each dish and the exact time when it would be prepared could be indicated.

Thinking-aloud Technique

In the present study verbal reports are used as data. However, before describing how the data were analyzed, it is important to address the goals, assumptions, research capabilities and limitations of the technique.

When collecting verbal reports the primary goal is to obtain concurrent verbal reports. Subjects are asked to "think-aloud" while they perform a task. Thus, subjects pursue two tasks concurrently: the verbalization of their thoughts and the execution of the task.

Based on the information-processing model of the mind (Hayes, 1981; Klatzky, 1980; Newell & Simon, 1972), described in Chapter 2, three assumptions are made: (1) one can identify a cognitive process by analyzing the information that is attended to (heeded) in STM and in the stimulus; (2) attending to information takes time (e.g. attending to AB takes longer than attending to A alone); and, (3) the capacity of STM is limited.

Under these assumptions: (1) the verbalizations are concurrent with the performance of the task, therefore the relation between the verbalizations and encoding protocols will be direct memories of the steps followed; (2) the content of STM is heeded in real time and the time factor can reveal how the information is structured, or the type of information that is being accessed; (3) the thought processes will run their course since judges will not be interrupted while "thinking-aloud".

Consideration must be given to the limits imposed by the task under specified conditions. Subjects may not feel at ease with the presence of the experimenter and the recording equipment, and they may display an awareness of participating in a research. These are all mediating factors which can affect the verbalizations. After the training session, which was meant to reduce the effects of these mediators, most judges concentrated on the task with very little of the aforementioned distractors being noticeable.

Difficulties with congruity of reports may arise when judges do not have a clear understanding of the researcher's purpose. Some are overly concerned with finding the correct answers although they may have been instructed that there is no right or wrong answer. Others are concerned with accomplishing a thorough job no matter how long it takes. Some consider speed the most important factor even when seemingly explicit instructions emphasize accuracy.

An obvious limit to protocol analysis is the inability of some people to verbalize what they are doing, while others may verbalize knowledge they are not really using at the moment. Some people appear to have a special ability to talk about trivial, irrelevant matters while thinking about more meaningful topics. Others follow a "think-then-summarize" procedure, verbalizing only after a sequence of thoughts (Johnson, 1972).

A last limitation to using verbal protocols as data is that only a trace of the processes is left for the experimenter to analyze. Thought proceeds much faster than speech, and although subjects are asked to verbalize every detail of their thoughts, the process can be expected to be incomplete. In

some cases expertise may have caused automatization to develop and the process may be accelerated since the final decision may appear immediately without going through all the different stages of the process. These last cases are particularly difficult to analyze, since they may be very sketchy.

Data Analysis Plan

Protocols 2 and 3 (in Appendix B) provide experimental data for the present research. The codes decided upon by the judges were written next to each statement (final decisions) and they are used to develop the first index of agreement called the written index. The justifications for each decision were collected in the form of verbal protocols which are analyzed to develop the second index of agreement called the verbal index. Index comparisons between different groups provides information on patterns of agreement/disagreement.

Level of agreement using written codes

Protocols 2 and 3 contained 59 and 19 lines respectively; the grid contained 40 categories including a "no code" category. The codes chosen by the judges were entered in a 40 X 59 matrix (Codes by Lines) for protocol 2, and a 40 X 19 matrix for protocol 3. Grouping the lines that were regrouped by all judges, the matrices were reduced to 40 X 45, and 40 X 18 respectively; frequencies were entered in each cell (see Appendix D for complete tables). Judges level of agreement is calculated by converting the

highest frequency to a ratio: the number of judges who agree on that particular code for that particular episode in proportion to the total number of judges. Percentages of agreement among judges on codes assigned to each line or group of lines are also used to report the results. These results compare the decisions made by judges since there was no criterion or no right answer. When estimating the level of agreement it was necessary to consider that: (1) data are nominal; (2) the items to be judged were not independent, since information from preceding or following statements sometimes had to be taken into account; (3) the categories were not equiprobable; and, (4) the judges could assign more than one code per line. Reliability of scoring was assessed by computing the highest proportion of cases in which the raters agree. This is referred to as written index (W). Disregarding codes which did not follow the majority does not imply disagreement; a different code is often used because judges pursued the investigation to a deeper level or from a different perspective. These will be analyzed in a second phase of the research when the decision process used by each judge is identified. The data and the published literature on indices of agreement were the guidelines for establishing high or low levels of agreement.

Preparation of verbal protocols

The tapes were transcribed verbatim. It is important that transcripts represent recordings as accurately as possible; it must be remembered that the tapes form the data base (hard data). In this preprocessing phase,

pauses, obvious exclamations, and loud intonations were also noted in the transcripts. The protocols were edited so that each segment would represent an instance of a process or a statement. In normal speech, statements are often abbreviated and are not necessarily complete clauses or sentences. Appropriate cues for segmentation were: connecting words (and, ok, so, because, or), intonation, pauses, and ideas. The lines were then numbered to facilitate reference to the protocol. Sections of protocols are presented in Appendix E. Protocols provide the researcher with large amount of data which need to be separated into episodes each reflecting a particular process targeted in the study.

Level of agreement using protocols

Inter-judge reliability was also calculated using comments from the protocols. Certain equivalents were identified to permit the protocols to be encoded in a simpler, canonical form - this is similar to what Ericsson and Simon call encoding vocabulary. Propositions were compared for semantic equivalences and the experimenter derived a prototype proposition from these. The following example illustrates how the analysis was done (excerpts are taken from judges' protocols when encoding episodes 3, 4, 5-6 in Protocol 2; the protocol line numbers are kept in their original edited format):

155 what I think she's doing is trying to get herself organized
156 clarifying a little bit in her mind what she has to do/ (judge # 16)

257 because she was just reading her material checking her material/ (judge #17).

48 they're identifying the different parts of the problem and trying
49 to organize them/ (judge #6)

27 are all just a person talking to him or herself as they are sitting
28 down getting ready to do the task pulling together all the
29 necessary information/
33 ... sort of pre-organization more than anything/ (judge #11)

17 it looks like he's kind of setting this up in front of himself
18 organizing it in front of him/ (judge #12)

The prototype proposition extracted in this case, from the 20 protocols, was: "checking out material given and getting it organized". Propositions successfully identified as matching the prototype were counted to establish a level of agreement between judges. This is referred to as the verbal index (V). See Appendix E for a complete example. The divergent interpretations were not evaluated in this study.

Synonymity was not judged lightly and anaphoric reference was not prejudged, (e.g., pronouns, naming by description). When synonyms, anaphora, fragmentary words or propositions occurred, context was included to remove ambiguity but the range of context was kept as narrow as possible. Approximately half the protocols dealing with the particular

episode were randomly analyzed before deciding on a prototype. The entire set was then reanalyzed using the prototype as a guide. This kind of stylization rarely raises problems in coding reliability, since it essentially uses the protocol language. The goal was to extract the reasons why a particular code had been chosen. There was no need to abstract or infer since the information was taken explicitly from the protocols. A comparison of the two different indices of agreement, written (W) and verbal (V), followed.

Analysis of judge characteristics

A global picture of how each judge approached the problem was obtained from four different sources: (1) observations taken by the researcher during the recording session and/or immediately following the session; (2) the recorded training session; (3) information scattered throughout the collected protocols; and, (4) intra-judge decisions.

Protocols were searched globally to identify certain behaviors:

-had the judges read the entire protocol before starting the coding?

-was a line by line analysis conducted? or

-was the analysis effected by reading large sections of the protocol at one time?

More information was collected on

-whether the judge justified each decision,

-whether he/she checked on the material presented (students' work

sheets) or assumed that the information was there, and

–the number of words used in relation to the time they needed to accomplish the task.

This examination provided a methodological path followed by each individual when trying to encode a protocol. According to the different characteristics and approaches observed, the judges were regrouped and their patterns of agreement/disagreement were analyzed.

Chapter 4

Results and Interpretation

In order to answer the research questions, this chapter is divided into three parts. In the first section, the results are interpreted in relation to the selected variables. In the second section the results from the written indices are compared to the results from the verbal indices. The judge characteristics analysis with interpretations of differently selected agreement/disagreement patterns are presented in the third section.

Results Considering the Selected Parameters

The final decisions reached by each judge were written beside each episode. The global percentage of agreement (written index - W) on both protocols among all judges was 61.4; these results are very low compared to what is reported in the literature. A possible interpretation would be that most reliability studies published to date using protocol analysis have been with well-defined tasks, which imply a definite solution to the problem is known. In the present study, an ill-defined task is used. Another interpretation for the low level of agreement on the written indices is that judges were given a minimum amount of instruction and practice before encoding. In most reliability assessments, judges underwent a longer training session.

Fortunately the final code chosen by a judge was not the only source of data, the reasons given to arrive at a final decision were also examined. Thus, the second index of agreement was developed from judges verbal protocols. Depiction of "what has been said" by the judges into a general statement is, in Ericsson and Simon's (1984) terms, a direct semantic equivalent analysis. A full description of the methodology has been presented in Chapter 3. The global level of agreement on the two verbal protocols was 85.0 (verbal index – V) which is much higher than the level of agreement found with the written index. This result is more in accordance with published reliability reports.

When the four groups are examined separately, the levels of agreement on the written indices are higher than those for the entire group of 20 judges. Table 1 presents the percentages of agreement among the selected groups of judges; the written and verbal indices are presented separately for each protocol. In this particular case (the written indices) the low level of agreement in the entire group (54.1% and 68.8%) indicates a high degree of heterogeneity. Therefore, it is not surprising, when stratifying, to observe a high degree of variability within the smaller groups ranging from 56% to 75.6%. It is also not surprising to observe a higher level of agreement within the strata since a higher level of homogeneity is expected within stratum than between strata. It was observed that the code on which the smaller group agreed is sometimes different than the code chosen by the majority in a larger group. This could indicate that a small group of coders may decide on a code with a bias toward their own preferred interpretation

Table 1

Mean Percent Level of Agreement for Selected Groups of Judges

Groups	n	Written indices		Verbal indices	
		Protocol			
		2	3	2	3
Experienced					
Male	5	56.0	71.1	79.6	85.6
Female	5	63.1	71.1	78.7	86.7
Non-experienced					
Male	5	60.4	75.6	77.8	87.8
Female	5	70.7	74.4	88.0	88.9
Total group	20	54.1	68.6	81.0	89.0

which could be different from another group with different views. Agreement does not measure validity. Bias may come into the interpretation when a judge assumes that the person thinking aloud is thinking in the same way that he does. Encoders often compare protocol behaviors with their own behaviors, as indicated by the following example from judge #19:

99 that's exactly what she does soup or salad I wonder if there's any
100 avocados/
101 that's exactly what I would say myself/
102 that's exactly the very first thing that would go to my mind/
103 I'd like to have avocados and she did the same thing/

Evidence from judge #18 further supports this view when he infers what the student is thinking:

245 but it's not a direct substitution/
246 like in his mind he thought um avocados would be nice but they're
247 not on the list/

The encoders in these two examples attribute to the student the action or thought which they regard as most reasonable under the circumstances. Encoding requires inferences to be made, implying the assumptions that the subject made the same inferences. Several coders with similar background, although different from the background of the person performing the task, could give the same incorrect interpretation. When the protocol information is explicit and clear, the danger of bias and inference is not as great as it

is for ambiguous segments. From these results, special attention should be given to the number of judges selected for a reliability check - when a small number of judges from homogeneous backgrounds is selected, the agreement level may be inflated and biased towards the selected group. When the number of judges can not be large, one could try to select judges with distinct backgrounds.

The level of agreement on the verbal indices does not show the same discrepancy between large groups and smaller groups. A possible interpretation would be that when judges are justifying their decision, they have an intuition of what the statement means. The justifications represented in the verbal indices are therefore a mixture of implicit categories and given categories, a sort of "inter-grid".

The experience effect

Contrary to what was expected, the level of agreement was higher among judges with no experience (NE) in encoding protocols than among judges with experience (E). When averaging the agreement level between male and female over both protocols (using results from Table 1), the global percentage of agreement is 70.3 and 65.4 for the non-experienced and experienced groups respectively. Agreement obtained on the verbal indices is also higher among judges with no experience than among judges with experience, 85.7 and 82.7 respectively. An indication that can account for these results, at this point in the analysis, is the richness of the "no

experience" judges' protocols; richness in this instance means the quality as well as the quantity of the verbalizations. Quantity is identified by the number of words contained in each protocol which is generally higher for the NE group. Quality is interpreted by the details contained in the NE judge's protocols. These are similar to novices protocols which often contain every step of a solution, as opposed to the protocols of experts in which the processes would have become automatic (Shiffrin & Schneider, 1977). Experts' protocols, often indicate that they have a whole solution plan in place before they begin solving a problem (Carey, 1986).

A closer examination of the differences between experienced versus non-experienced judges is done by examining the frequency distributions of episodes (Total = 63) on which the judges display their agreement or disagreement. Levels of agreement for the two indices developed in this study are reported in Table 2 for the groups determined by level of experience and gender. The written index distribution shows a general pattern of higher frequencies at the lower level of agreement (2 out of 5 judges); although the female group shows a higher level of agreement. A reverse pattern is found in the verbal index distribution, with higher frequencies appearing at the higher level of agreement (5 out of 5 judges). The marginals do not add up to 63 because the statements on which there was no agreement are not included. A possible interpretation of the higher level of agreement with the verbal index may be that the agreement between judges was often determined before they reached their final decision using the given grid. As mentioned earlier, the imposed grid may create some

Table 2

Frequencies of Episodes for Various Level of Agreement on Both Indices Among Selected Groups of Judges

Number of judges agreeing	ME		FE		MNE		FNE	
	W	V	W	V	W	V	W	V
2	29	8	16	3	18	5	10	2
3	12	9	11	8	19	9	15	4
4	10	17	19	19	15	22	21	19
5	11	21	12	29	10	25	15	37

Note. The cell frequencies are the number of episodes out of a total of 63 on which judges agreed. The total number of judges per selected groups was 5. ME = males with experience in protocol analysis; FE =females with experience in protocol analysis; MNE = males with no experience; FNE = females with no experience; W = written indices; V = verbal indices.

ambiguity between what is implicit and what is imposed. Judges may have agreed on a single interpretation of a statement but justified their use of different codes.

From these data (Table 2) consisting of frequency counts and nominal categories, the chi-square technique was applied to determine whether there was a statistical association between level of experience and level of agreement. From the results of the chi-square analyses presented in Table 3, one concludes that there is no significant association between these two variables at the .05 level of significance. This conclusion applies to both indices.

The gender effect

A generally higher level of agreement was found among women in Protocol 2 except for the experienced group on the verbal indices (see Table 1). This protocol contained many details about ingredients needed in particular recipes and cooking techniques. When "expertise in the field" is considered as having knowledge of specific methods and technical terminology, women in this sample showed more "expertise in the field of cooking" than men. An example from judge #2 shows how difficult it may be to make a judgment on a topic one knows little about:

466 (READS) baked alaska now that's all/
467 is that all that's in baked alaska/
468 ice cream and egg whites and sugar/

Table 3

Chi-square Results for Experience Variable

Experience	Number of judges agreeing 2	3	4	5	χ^2	df	p
			Written index				
With	45	23	29	23			
Without	28	34	36	25	6.88	3	.0750
			Verbal index				
With	11	17	36	50			
Without	7	13	41	62	2.69	3	.5563

Judge #2 had little idea of the ingredients needed for a baked alaska, it would be difficult for him to judge whether the student estimates the exact amounts needed.

In Protocol 3, where calculations were numerous, a noticeable instance of gender difference was in the use of the S13 and S14 codes. Code S13 was defined as follows: "misjudges amounts appropriate for six people", and S14: "confuses amount of food/beverage served with amount of raw ingredients needed to prepare the dish". To realize that someone is not using the right amount requires some expertise in the field. These codes were used 18 times by women compared to only 9 times by men. It would be difficult for a judge to decide whether the student is misjudging the amounts needed for six people when he has little knowledge of the recipe.

Why is the gender difference not noticeable in Protocol 3? First the protocol did not contain so many technical terms related to cooking. Protocol 3 was taken from a beginner and this variable may have contributed to the absence of more task specific terminology. Secondly, the indices represent the majority. In cases where misjudgment or confusion were detected (an S13 or S14 code), the majority of judges chose the S10 code where "calculating amounts" was a more global definition. Codes S13 and S14 were not exclusive of code S10, and the expertise which could have been revealed with the use of S13 and S14 did not surface.

Chi-square analyses were applied to data from Table 2 to determine whether there was an association between gender and level of agreement

when encoding verbal protocols. From the chi-square results presented in Table 4, the conclusion is that there is a significant association at the .05 level of significance for both indices. From these results one may conclude that men and women in this sample show a differential basis on their level of agreement when encoding verbal protocols.

The inter-protocol effect

Short extracts from the protocols being studied are used in this section. For the complete context, the entire protocols can be found in Appendix B. Agreement and disagreement is examined in both protocols.

Agreement. An examination across protocols shows a greater level of agreement for all groups on Protocol 3 (see Table 1). There is no formal criteria for determining when reliability is high, an index over 80% is generally considered a high level of agreement. When examining the highest indices, in protocol 3 (see Appendix F, Table F-2 for levels of agreement on each episode), the episodes are of the following type:

14 um 500 ml. um
15 dinner rolls 12 dinner rolls and
16 some potatoes 6 potatoes and

Almost all judges decided on code S10: "calculates amounts/quantities of food items needed for six people". Stating numbers supplies explicit information to match the S10 code, a low level of inference being necessary, higher reliability is expected. In Protocol 2, there are only 10 out of 45

Table 4

Chi-square Results for Gender Variable

Gender	Number of judges agreeing 2	3	4	5	χ^2	df	p
			Written index				
Male	47	31	25	21			
Female	26	26	40	27	10.59	3	.0144
			Verbal index				
Male	13	18	39	46			
Female	5	12	38	66	8.24	3	.0410

episodes with a high level of agreement. The episodes and chosen codes on these are of 5 different types. To cite one of these episodes:

23 oh I better be a good dietitian here and use vegetable oil it
24 doesn't have as much saturated fat.

The R2 code was chosen by the majority and is defined as "explains that a food choice has been made for nutritional reasons". The connection between the verbalization "not as much saturated fat" and the R2 code "nutritional reasons", is quite explicit and a low level of inference is needed to encode it. In other words, when the link between the defined category and the verbalizations is direct and the level of inference is low, the coding is more reliable.

Disagreement. When the reliability indices are <u>very low</u> the context does not seem to have been captured by the judges, though it did not always appear to be difficult to grasp. For instance in Protocol 3, on lines 6 and 7, the student says:

6 I'm taking the food list and I'm going to pick out the foods
7 I'd like for this dinner for six people or six adults.

Although the student explicitly stated his intentions, only 2 judges kept it in mind when interpreting the rest of the protocol. Five more judges remembered this information half way through the task and used it in interpreting subsequent statements. At the end of the task, another judge

remarked that the student was probably doing two tasks at the same time. However, this particular judge did not review his previous decisions. In Protocol 3, to be able to do justice to the student, judges had to understand that 2 tasks were being performed concurrently: the menu planning task and estimating the required amounts of food. Therefore the student ran down the list of food items as he "thought aloud" while he made calculations and planned his menu.

The extent to which each segment is independent of the other is considered an important variable on the level of agreement in the context of fixed predetermined coding categories for well-defined problems (Ericsson & Simon, 1984). In the present situation (an ill-defined problem), if the segments are considered independently, it undermines the degree of agreement. Evidence provided in lines 6 and 7 is important for accuracy in subsequent coding decisions. Problems related to non-independence of segments is discussed below.

A more complicated context is displayed at the beginning of Protocol 2 when the student is glancing through his material before beginning the task. The judges who took time to examine the material which was identical to that presented to the students may have recognized the titles from each page as well as the instruction sheet wording. There is a better chance for a fair judgment when information about the physical environment (especially when visual stimuli are present) accompanies the protocol or is discussed during a practice session. In this particular research design, only a section of the protocol was presented and no discussion on individual context

information took place. Some judges withheld judgment on ambiguous statements where they felt "something more was going on". These statements were often related to an incomplete record of the perceptually available information.

Up to this point in the analysis, three main findings have resulted. (a) A wide range of agreement/disagreement was found. Agreement was found when the level of inference between verbal statements and codes was low, and disagreement occurred when level of inference was high. (b) The level of task dependent expertise (cooking expertise) seemed to have more of an impact on agreement among judges than experience with the technique of protocol analysis. (c) There was more agreement in encoding a beginner's protocol than when encoding an expert's protocol.

Comparison of Written and Verbal Indices

Why was the written index lower than the verbal index? The first reason would be that the wording in the definitions of the codes may have many meanings, thus permitting two or more interpretations to be coded differently. For example, in episodes 1 and 3 to 9, of protocol 2, the prototype statement extracted was "organizing all presented material by reading it or summarizing it" – this prevailing interpretation among judges corresponded to 5 different categories on the grid depending on how the word "material" was defined. The material was the sheets presented to the

students which were defined as synonymous with: the instructions, for the judges who chose the P1 code (Reads instructions); part of the instructions, for the P2 choices (Summarizes all or part of instructions); the problem, for the P3 choices, (Identifies or defines the problem); or an element in the problem statement, for the P5 choices (Refers to an element in the problem statement). Locating position with respect to the entire task, the A1 choices were more universal judgments of the situation. The nature of language is a central issue in cognitive science. Language is an instrument of thought and the language in terms of which a model is constructed constrains what kind of realities can be stipulated (Bruner, 1984). Semantic problems can be reduced when the training session is more extensive and includes open discussion between judges and experimenter.

A second reason why the judges agreed more on the interpretations of the segments (verbal indices) than in their final decisions (written indices) is that judges had very little time to familiarize themselves with the numerous categories. As an example, from episode 10 in Protocol 2, 16 judges out of 20 explained that the student was planning a strategy but 3 of them did not find the code P4 which says "chooses a strategy for attacking the problem"; they used a different code, one that would be a close match.

A third reason would be that some categories seem to be missing, incomplete, or too general. A category for: "checking the list of food items to see what was available" should have been included. Since many "checking" behaviors are present in the second protocol, the statements

relating to those behaviors became ambiguous when no matching code was found. There were actually so many statements relating to "checking the list of food items available" that a deeper examination of the judges' protocols was needed to find out how each judge had decided to encode those particular statements. When reasons were initially given for selecting a particular code, that code was often used in subsequent statements without restating the reason(s) for the choice. For instance, many expressions in the following style were found: " he's obviously gone back to the category that I feel is not there ... it will have to be an S1" (judge #19). From this example the selected code (S1) took on multiple meanings, the explicit meaning from the grid and an extended meaning that fits the "checks the list" behavior. For these reasons Table 5 was constructed from the analysis of the first episodes encoded and related to "checks the food list". The particular code chosen and the reasons for the choice are condensed in Table 5. When the particular code was used in subsequent statements without justifications, it was not necessary to back track and read large sections of protocols to find the reasons behind the choice. Information from Table 5 made it faster and more accurate (not assuming the code implied the definition from the grid) to continue the protocol analysis. When an episode relating to "checks the list of food items" was analyzed, the chosen code was compared to codes in column 2 and the reasons were immediately available in column 3.

Table 5

Various Codes Chosen by Each Judge to Define "checks the food list" Behaviors

Group & judge no.	Selected Codes for "checks the food list"	Justifications
ME		
1	P5	food list is an element in the problem statement
2	S1,S2,S3	looking for food items is designing the courses
	S5	listing is synonym with checking food items off as they go
3	P5	list is an element in the problem statement
4	A3	checking to see if a food item is available is an assessment of a portion of the task because students were told to check and see if food item was available
5	P5	food list is an element in the problem, referring is synonym with checking
FE		
6	S5	when checking food items to go in a dish
	S2	when checking dish within a course
7	S5	checking the list as well as designing courses and listing ingredients in a recipe
8	S1,S2,S3	all encompassing codes: checking list, choosing strategies, listing ingredients in a recipe
9	S5	student is listing when checking the list
	S8	groups episodes together, uses "substituting one food item for another" when the food item checked for is not on the list
10	S5	checking and listing can be synonym

(table continues)

Table 5 (continued)

Group & judge no.	Selected Codes for "checks the food list"	Justifications
MNE		
11	S5	student starts listing ingredients and realizes it's not on the list
12	S1,S2,S3	designing courses and checking for food items is part of it
13	S5	while a student checks the list he/she names or lists the items at the same time
	P3	since student is complying to restrictions he/she identifies the problem
	P5	list is an element in the problem
	P2	list is part of the instructions
14	A3	when student decides on a food item it means it's on the list, an assessment has been done
	A4	when student realizes a food item is not available he/she makes changes or additions to the menu
15	P2	having to check list, student rereads his/her instructions
	P4,S5	identifies a need (chooses a strategy), may go on to listing more specific needs (ingredients in a recipe)
FNE		
16	S1,S2,S3	since no code for checking the list, designing a course is a general definition
	S5	deciding on a food item, after checking the list
17	S1,S2,S3	checking list is looking at different possibilities, he/she designs menu
18	P2	food list is part of the instructions
19	S1,S2,S3	no checking list code, becomes part of designing
20	S1,S5	checking list is designing and listing food items

Another category which, if included, might have eliminated ambiguity was a code for cooking techniques such as "sauté", "grate", and "broil". When such terms were found in the protocols, a search was on for an accurate and specific definition. Only very general definitions, such as "designs a course" (S1, S2, or S3), or "lists ingredients" (S5), could be found to encode this type of statement.

As a last remark on missing categories, the grid was often surveyed for another code which would describe comments such as "I will have a cup of coffee now". Verbal protocols often contain such statements which are not directly related to the task. Since many judges left such statements with no codes, including a category for these would facilitate encoding.

Categories S1 (designs the appetizer), S2 (designs the main course), and S3 (designs the dessert), caused some confusion; they are incomplete categories. With these three codes a very specific enumeration of what the meal "may consist of" was mentioned, except for bread. When bread was added to the meal, ambiguity occurred resulting in less reliability. A beverage was included in the definitions of the three codes, S1, S2, and S3. When a beverage such as wine was chosen as part of the meal, it was sometimes unclear whether it should be with any particular course. Some judges solved the problem by assigning S1, or S2, or S3; in other words, not making a decision. Others decided on one course according to their own eating habits, for example judge #11 who says: "dessert wines aren't very fashionable these days, that's probably a dinner wine" and he chose the main course code, S2.

Certain categories were too general; an example would be categories S1, S2, and S3. Judges # 8, 10, 11, 12, 16, 17, and 20 used this approach, adding other codes when the link between verbalizations and category was very explicit. Otherwise they would revert to the more general code.

A second example would be with the A category (assessing and monitoring activities), where 17 judges out of 20 interpreted episode 62 in Protocol 2 as "an indication that the task had ended". Their choices are presented in Table 6, with their reasons for these choices extracted from the protocols. It was observed that only 8 agreed on the same A2 code. In the A category, it seemed very difficult for the judges to differentiate between A1, A2, A3, and A5. Describing a short statement such as "that's my menu" and having to decide whether this is an evaluation, an assessment, a summary, or locating ones' position becomes relative, since these processes may be either very short or very lengthy. In A1, A2, A3, and A5, the second portion of the definitions referring to the task were also too general: the entire task, what has been done, a portion of the task, and the end of the task. Since the task included designing three courses, estimating the amount of food needed, and planning the time needed to prepare the meal, a more specific definition of each stage of the task would have eliminated some confusion. It is a complex task to define categories in general terms, so that they include all behaviors relating to it and that at the same time, discriminate well. Code A2 was chosen by the majority but it is not necessarily the best description of the behavior. "That's my menu" seems too short to be what one would normally call a summary, an assessment, or

Table 6

Problems with Generality of Group A Activities

Group & judge no.	Selected Codes for "that's my menu"	Reasons (protocol extracts)
ME		
2	A2	Summarizing what has been done is a global thing, because she is recapitulating in her own mind I think that that is done
3	P5	End of task statement, strictly speaking he's referring back to an element of the problem statement and it's finished but I'm not comfortable with that
4	A2	(Recording missing for that episode these are the reasons given when he used A2 on other episodes) so she's all finished, I think that's a summary
5	A1	It is an evaluation statement, that's an indication that she has finished something
FE		
6	NC	So she's decided that that's the end of that task, but she hasn't really reviewed it, did she assess, she has decided that is it, I don't have a choice I have to leave it hanging
7	A5	She's evaluating that she's ended it up with that
9	E3	She's saying that she's done, she doesn't even say that it's good, I don't see any code for that, I have to find the closest one, so she's accomplished, ok so that's my menu all right

(table continues)

Table 6 (continued)

Group & judge no.	Selected Codes for "that's my menu"	Reasons (protocol extracts)
10	A2	Entire task is that the three course meal or is it one course, I like A2 better, to actually evaluate the entire menu at the end of the task it would have to be an explicit evaluation
MNE		
11	A2	That's just summarization of what has been done
12	A2,A3	He's assessing a portion of the task, summarizes what has been done maybe an A2
13	A2	So that's the summary of what has been done
14	A5	It's not much of an evaluation but that's what it is
15	P4	I guess that's the last part of your strategy
FNE		
16	A1	She's doing some assessing here, locating herself, she's reminding herself that she's done but it's not a big summary or anything, she says now I've done one, what do I do next
17	A2	Doesn't assess the task, a summary of the three courses, that's summarizes what has been done
18	A2	She's sort of rounding off everything, summarizing what has been done not really, a partial summary, we don't know for sure if it's an assessment, it's a summary statement
20	S3,M1	That's a decision statement in the same way, M1 is operating in the background all the way through because she remembers the task and she's done with it

an evaluation, thus leaving "locates position with respect to the entire task" possibly a better decision. Complete definitions for these codes are in Appendix C.

To further analyze the differences between written and verbal indices a graph comparison of the two indices of agreement is presented in Figures 1 and 2. The solid bars represent the final decisions of the judges - the written indices, and the striped bars represent the justifications for choosing a particular code - the verbal indices. The contrast provides a rough gauge of the constant increase in agreement when verbalized justifications (verbal indices) are used for judging the agreement rather than the final decision (written indices). It is useful to note both agreement and disagreement when comparing the two indices. The particular episodes on which the two indices resemble or differ from each other are listed in Table 7.

Agreement. If high agreement (over 80%) results when both indices are used, the link between verbalizations and codes is direct in all these cases. The episodes on which both indices are similar are highlighted in Table 7. When high agreement resulted with the verbal index only, many of these episodes required a higher level of inference – assessing and monitoring activities, or dependence upon larger context. These are presented as the second group of episodes with high agreement in Table 7. The agreement was high on the interpretation of statements (verbal index) although not necessarily on the codes to match the interpretation (written index). From

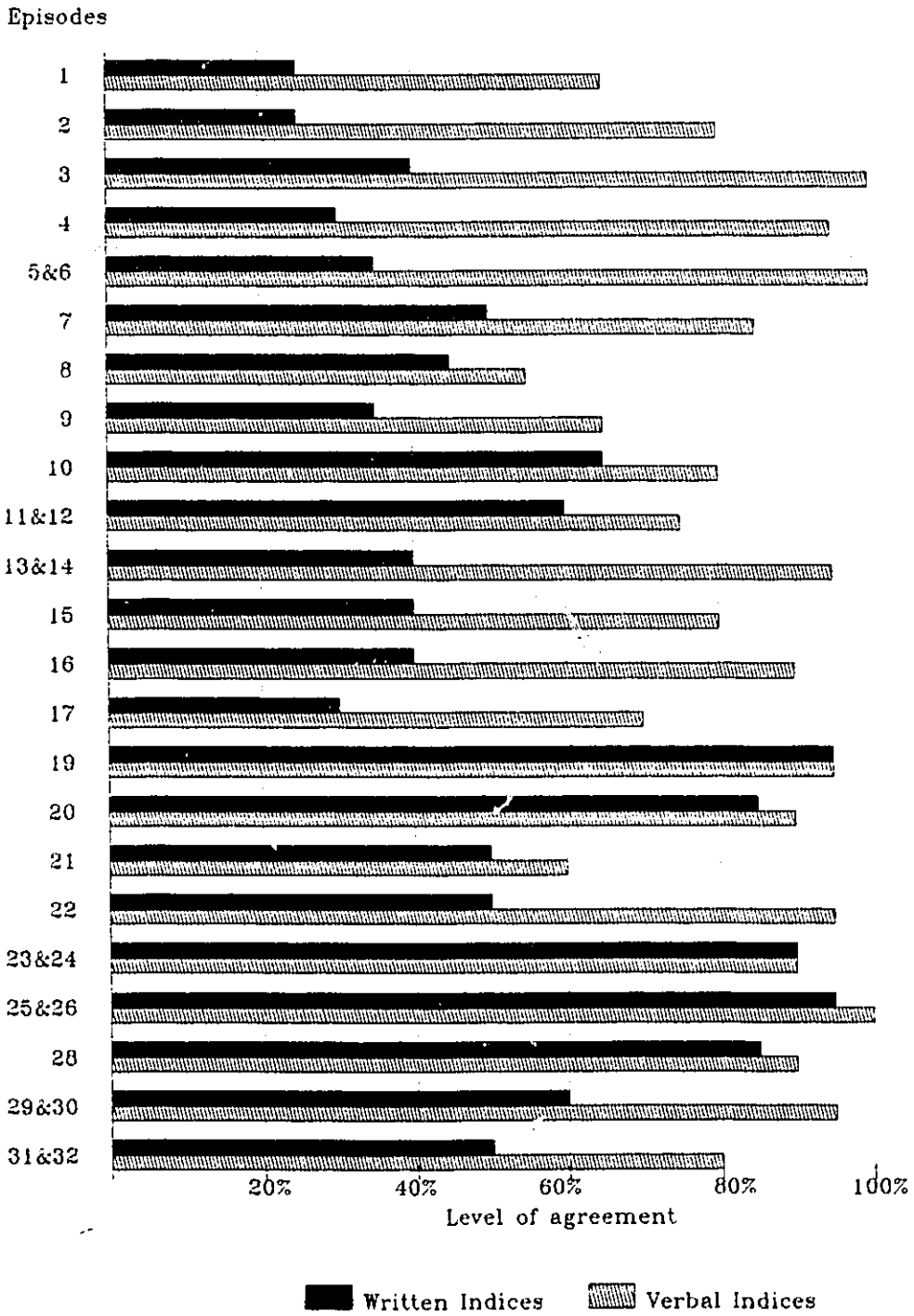

Figure 1. Bar graph representation of verbal and written indices of judge agreement on episodes of protocol 2.

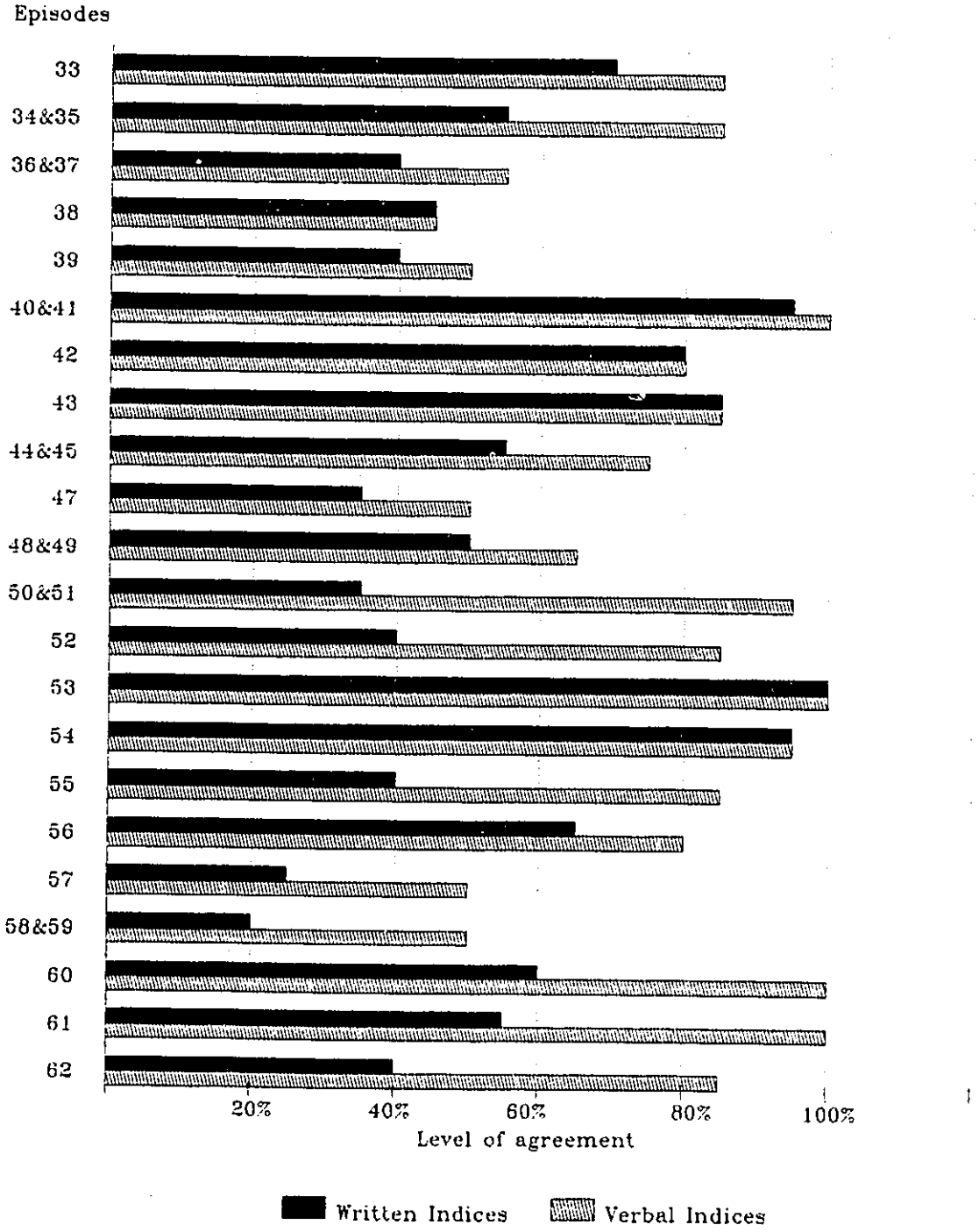

Figure 1 (continued)

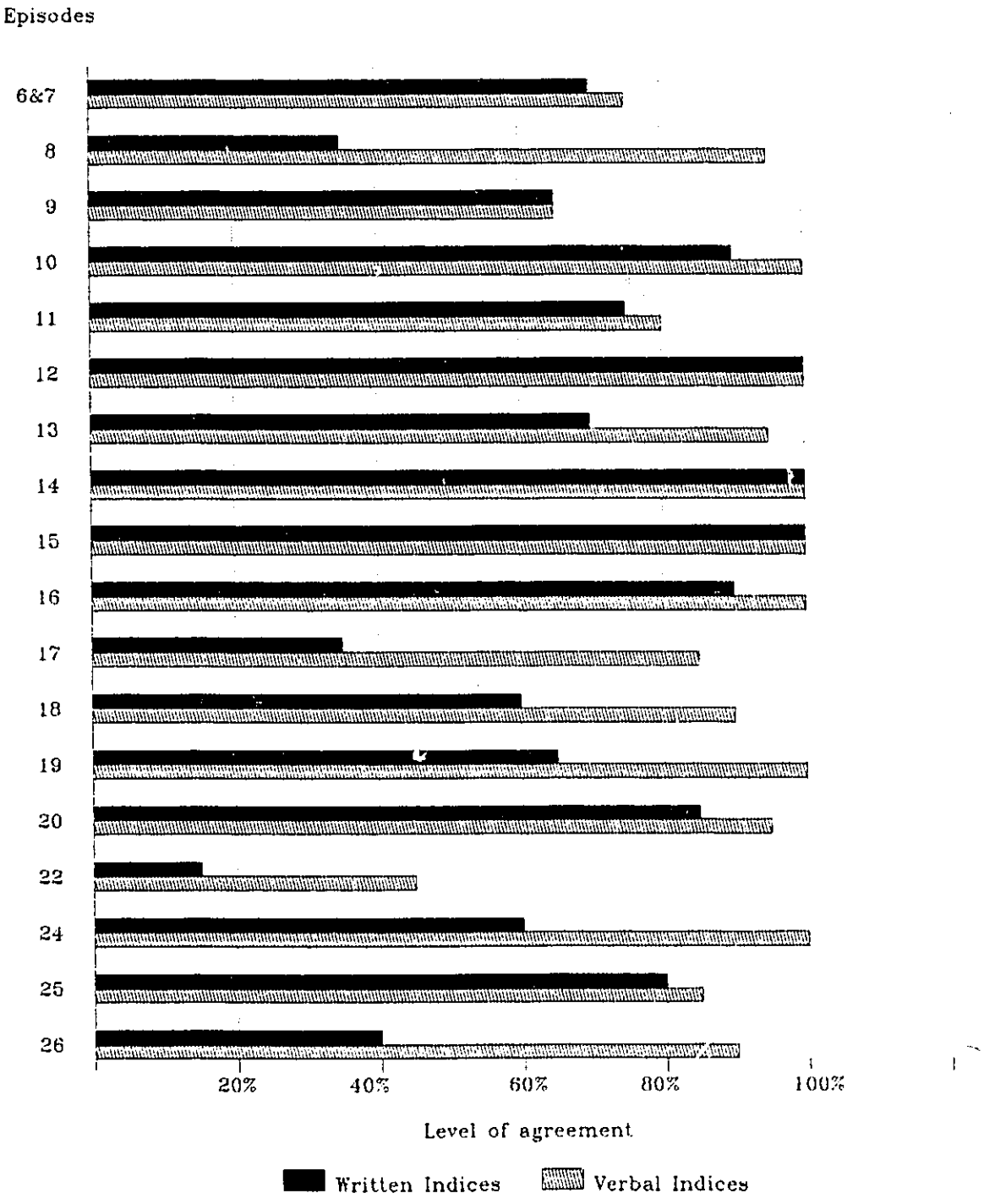

Figure 2. Bar graph representation of verbal and written indices of judge agreement on episodes of protocol 3.

Table 7

Written and Verbal Indices Compared in Relation to Agreement and Disagreement on Episodes

Indices characteristic on both protocols	Episodes	
	Agreement (over 80%)	Disagreement (under 80%)
V & W Indices Are Similar		
protocol 2	19, 20, 23-24, 25-26, 28, 40-41, 42, 43, 53, 54	1, 9, 11-12, 17, 21, 36-37, 38, 39, 44-45, 47, 48-49, 57, 58-59
protocol 3	10, 12, 14, 15, 16, 20, 25	6-7, 9, 11, 22
V Indices Are Higher Than W		
protocol 2	2, 3-8, 10, 13-16, 22, 29-35, 50-52, 55, 56, 60-62	
protocol 3	8, 11, 13, 17-19, 24, 26	

Note. V = verbal; W = written.

these results, one could say that reliability can be achieved when encoding verbal protocols collected while performing an ill-defined task when problems with stimulus information and the descriptions of categories (codes) are resolved.

Disagreement. When a high level of disagreement occurs, the indices are low (under 80%) no matter which index is used. The reasons in these cases (disagreement column, in Table 7) were of three different types: (1) not enough information on the perceptual environment, (2) not enough time to become familiar with the categories, and (3) missing categories or they were not exclusive of one another leading to different levels of analysis. Choosing a different level of analysis may have been triggered by the grid; the generality or specificity of a definition may have led the judge to analyze the text in a global perspective or from a more microscopic viewpoint.

The findings from the comparison analysis can be summarized as being important indicators for increasing reliability of encoding verbal protocols: preparation of the judges prior to encoding, and elaboration of complete and discriminating categories. With this information it may be argued that the level of agreement using the written codes, would increase substantially; it would approach or exceed verbal index results. To complete the analysis, patterns of agreement/disagreement among groups of judges are studied.

Results from Judge Characteristics Analysis

Numerous observations seemed to indicate that judges approached the problem in different ways. Intra-judge frequencies of codes used were compiled and results indicated noteworthy individual differences: every judge had a favourite code, and the number of codes used ranged between 10 and 22. A global search through the recorded practice session, protocols, and written notes followed. Table 8 summarizes the observations made on each judge, including how they partitioned the protocol to be analyzed, whether they read or commented on the presented material. Individual approaches to encoding verbal protocols were extracted from this information.

According to Ericsson and Simon (1984) three variables have an influence in studies on reliability of encoding: (1) the extent to which encoders must make inferences (the closer the match between categories and verbalizations the more reliable the coding), (2) the extent to which each decision is independent of the other, and (3) the extent to which segments are independent of each other (how many segments must be taken into account before applying a code). The first variable was observed when agreement/disagreement patterns were examined in the inter-protocol effect, agreement was higher on low-level inference coding than on high-level inference coding. In the index comparisons, high levels of agreement were also found when the match between category and verbalization was explicit (low-level inferences). The last two variables are of interest when related

Table 8

Judge Characteristics Analysis

Group & judge no.	Partition style	Comments		Checks			No.of words	No.of codes	Fav. code
		Grid	Pro	GR	WS	IN			
ME									
1	by section	X		X	X	X	5981	13	P5
2	by section	X		X			6543	13	S5
3	by section	X	X		X		6241	12	P5
4	by section	X	X	X	X	X	—[a]	19	A3
5	by section	X		X	X	X	8076	16	P5
FE									
6	by section	X	X	X	X	X	10851	13	S5
7	line by line			X		X	5751	18	S5
8	by section				X		5775	14	S1,S2,S3
9	by section						5213	13	S5
10	line by line			X		X	5213	15	S5
MNE									
11	by section	X	X	X	X		8599	18	S5
12	by section	X	X	X	X		9430	18	S1,S2,S3
13	line by line		X	X			8186	17	none
14	line by line	X	X	X			6476	22	S1,S2,S3
15	by section		X	X			4256	14	P3,P4
FNE									
16	by section	X	X	X	X	X	8885	13	S1,S2,S3
17	by section	X	X	X	X	X	12675	19	S1,S2,S3
18	by section	X	X	X	X	X	9287	18	P2
19	by section	X	X		X	X	10552	22	S1,S2,S3
20	by section	X	X	X	X	X	6754	10	S5,M1,S1

Note. X = judges have performed the mentioned task; GR = grid; WS = students' work sheets; IN = students' instruction sheet; Fav.code = judges' favourite codes.

[a] Part of transcript was loss due to mechanical problem.

to the findings of the present research. Four judges adopted a line by line analysis approach (Table 8, column 2); these judges were trying to keep each segment independent of the other. The line by line analyzers gave very little personal interpretations of segments or codes (Table 8, column 3), except for judge #14. Judges who chose to analyze the verbal protocols by larger sections tried to interpret in context rather than trying to encode in terms of the content of each segment. Some cognitive processes, requiring a high level of inference, can only be assessed by using a much larger context than what has been considered in direct semantic equivalent encoding studies (Ericsson & Simon, 1984). This can be seen when judge #20 decides that a substituting strategy has been used, in Protocol 2 when encoding episode 15 she has to read down to episode 19 where she finds out that the appetizer will be onion soup instead of an avocado salad:

196 substitutes one food item for another she does that/
197 um ... (READS) let's see no avocados how about onions/
198 so really there's an S8 right there (referring to episode 15) ok/
199 but she's also thinking I think about something there's something
200 particular in her mind of course we got it right there (referring to episode 19) it's french
201 onion soup/ (judge #20)

Comparison of agreement was made between judges who tried to analyze each segment for its own content and judges who chose to analyze by larger sections. Results can be found in Tables 9 and 10 where written indices and verbal indices were compared for 3 different groups: line by line

Table 9

Level of Agreement Among "Independent Segment Analyzers" and "Context Analyzers" in Protocol 2

	Line by line analyzer			Context analyzer with global approach			Context analyzer with specific approach		
Episodes	Codes	W	V	Codes	W	V	Codes	W	V
1	P1	66	100	—	33	33	NC	66	0
2	NC	66	100	NC	66	100	NC	66	100
3	P3	66	100	—	33	100	—	33	100
4	—	33	66	—	33	100	—	33	100
5-6	P5	100	100	—	33	100	—	33	100
7	P2	66	100	—	33	100	—	33	66
8	—	33	66	P2	66	66	—	33	100
9	—	33	0	P4	100	33	P2	66	100
10	—	33	66	P4	100	100	—	33	66
11-12	S1	66	100	S1	66	100	S1	100	100
13-14	S5	66	100	S1,S5	66	66	S5	66	100
15	S8	100	33	S1	66	66	S5	66	100
16	S5	100	66	S1	100	100	—	33	100
17	—	33	100	—	33	66	—	33	66
19	S1	100	100	S1	100	100	S1	100	100
20	S5	100	100	S1,S5	66	100	S5	100	100
21	—	33	33	S1	100	66	S5	66	100
22	S5	66	100	S1	100	66	—	33	100
23-24	—	33	66	R2	100	100	R2	100	100
25-26	S5	100	100	S1,S5	100	100	S5	100	100
28	S5	100	100	S1	100	33	S5	66	66
29-30	S5	100	100	S1	100	66	S5	100	100
31-32	—	33	66	S1	100	66	S8	66	100
33	S5	100	100	S1	66	33	S5	66	100
34-35	S5	100	100	S1	66	33	—	33	100

(table continues)

Table 9 (continued)

Episodes	Line by line analyzer			Context analyzer with global approach			Context analyzer with specific approach		
	Codes	W	V	Codes	W	V	Codes	W	V
36-37	—	33	33	S1	100	66	R5	66	66
38	P4	66	33	S1	100	33	R1	66	66
39	A3	66	33	S1,A2	66	66	A2	66	66
40-41	S2	100	100	S2	100	100	S2	66	100
42	S2	100	100	S2	66	66	S2	100	100
43	S2	66	66	S2	66	66	S2	100	100
44-45	S5	100	33	S2	66	33	S5,S2	66	66
47	—	33	66	—	33	33	—	33	33
48-49	A2	66	66	A2	66	66	S2,A2	66	100
50-51	A3	66	100	P5	66	100	—	33	100
52	—	33	66	S3	66	66	—	33	100
53	S3	100	100	S3	100	100	S3	100	100
54	S5	100	100	S3	100	66	S5	100	100
55	—	33	66	S5,A3	66	100	S5	66	100
56	S3	100	100	S3	66	66	A2	66	66
57	—	33	33	S4	66	33	A4	66	66
58-59	—	33	33	S4	66	33	—	33	33
60	S3	66	100	S3	100	100	A4	66	100
61	—	33	100	S3	66	100	A4	66	100
62	A2	66	100	A2	66	66	A2	66	100
Means		67	77.6		73.7	72.3		63.3	87.2

Note. W = written indices; V = verbal indices; Codes = code chosen by the majority.

Table 10

Level of Agreement Among "Independent Segment Analyzers" and "Context Analyzers" in Protocol 3

	Line by line analyzer			Context analyzer with global approach			Context analyzer with specific approach		
Episodes	Codes	W	V	Codes	W	V	Codes	W	V
6-7	P4	66	66	P4	100	100	—	33	33
8	—	33	100	—	33	100	—	33	100
9	P5	66	66	P6	66	66	P6	66	66
10	S10	100	100	S10	100	100	S10	100	100
11	S2	100	100	S2	100	100	S2	66	100
12	S10	100	100	S10	100	100	S10	100	100
13	S2	66	100	S2	100	100	S2	66	100
14	S10	100	100	S10	100	100	S10	100	100
15	S10	100	100	S10	100	100	S10	100	100
16	S10	100	100	S10	100	100	S10	100	100
17	—	33	33	—	33	100	—	33	100
18	S1,S10	66	66	S10	100	100	S1,S10	100	100
19	S3	100	100	S3	100	100	A4	66	100
20	S10	100	100	S10	100	100	S10	100	100
22	—	33	33	—	33	33	P7	66	66
24	S2	100	100	S2	66	100	—	33	100
25	S10	100	100	S10	100	100	S10	100	100
26	P3	66	100	S12,P4	66	100	A1	66	100
Means		79.4	86.9		83.2	94.4		73.8	92.5

Note. W = written indices; V = verbal indices; Codes = code chosen by the majority.

analyzers, section analyzers who used a general approach (as described above), and section analyzers who used a more microscopic approach. All 3 groups were equal in size, 3 judges per group.

The most noticeable result in the comparison study is with the "context analyzers" with a global approach to the task. The overall level of agreement on their verbal indices did not increase over the written indices in Protocol 2 (see bottom of Table 9, Means on verbal indices, for "context analyzers" with a global approach). A potential factor contributing to disagreement on verbal indices is that the verbal indices came from the justifications for final decisions and judges with a global approach did not provide many justifications. Protocol 2 contained many higher level processes and this particular group remained satisfied with a general code such as S1, S2, and S3, not going into a deeper level analysis. Another distinguishing result from this particular group is the agreement on the written indices, it was higher in both protocols for this group than for any other groups up to this point in the analysis. See bottom of Tables 9 and 10 for all three groups and Table 1 for originally selected groups. The written indices were calculated using the majority rule and as previously mentioned this does not mean validity; reverting to a general approach formed the majority in many instances. Taking into account the small number of judges per group one can cautiously conclude that the line by line analyzers and the context analyzers with a specific approach to the task provided evidence consistent with previously found patterns of agreement/disagreement. Results in Tables 9 and 10 indicate that judges who used context in

identifying mental activities reached a higher level of agreement provided they remained concerned with higher level inferences.

In examining protocols from the "context analyzers" with a specific approach, (they had the highest level of agreement on verbal indices), two problems were found with coding non-independent segments: The first problem was not restricting the influence of previous coded information on subsequent coding decisions. In essence, when a judge decided that activities like "confusing" or "misjudging" were going on, his decision could guide or bias his subsequent decisions. For example judge #20 says: "there's no design in this I'll tell you that, I'll never give her no (sic) S1, S2, S3"; in this case the judge never allowed herself to assign one of the mentioned codes for that protocol. Numerous such situations are present in the collected data, an extract from judge #18 is cited below:

228 (READING LINE 13) I wonder if there's any avocados on this/
229 that's checking the list after coming up with the idea/
243 substitutes one food item for another to adhere to restrictions of
244 food list I would say I would have to say S8/
245 but it's not a direct substitution/
246 like in his mind he thought um avocados would be nice but they're
247 not on the list/ (this information from line 14)
248 so I'll put S8 (on line 13) for now but um.../

When judge #18 says they are not on the list she is talking about the next episode which says: "no ... no avocados", and with this information she goes back and codes line 13 with an S8.

The second problem with encoding non-independent segments was adding words to what the subject said rather than adhering to the literal transcript. The following extract taken from the protocol of judge #6 illustrates this process.

189 Because of that something classic subject is not satisfied with
190 just French onion soup/
191 although that's what I had for lunch/
192 that's they're looking for something trying to make something more
193 out of what they've got/
194 and they're going for French onion soup because it is easier to
195 make it's much easier to make/.

Nothing in the protocol indicated that the student was not satisfied with French onion soup or that he chose to make it because it was easy.

Returning to the observations collected from the judges, two main characteristics emerged from the analysis: (1) task control judges, the ones who were concerned with choosing a description which would match the statement – finding the correct answer on the grid; (2) and the self control judges who were concerned with accurate interpretations of statements providing as many personal comments as possible on the categories as well as on the protocols. They finally chose a code which was the closest to their own interpretation of the statements.

Six judges (#7, 8, 9, 10, 13, and 15) used the "task control" approach, they chose a description from the grid which would match the statement. They provided no comments on the grid (see Table 8, column 3); three of the line by line analyzers were part of this group. These judges expressed

concern with providing a large selection of codes, not wanting to use the same ones all the time.

Eleven judges (#3, 4, 6, 11, 12, 14, 16, 17, 18, 19, 20) followed the "self control" path, they all commented profusely on the categories and the protocols (Table 8, column 3). The personal interpretations are valuable to the researcher whose purpose is to assess the reliability of categories. It is of value since he could use the judges' comments on the grid. They often give suggestions as to how a definition can be rewritten, or why certain categories may not quite describe a behavior. For example judge # 3 says:

255 um ... now he's interrupting work on one dish to go back to
256 elaborate on an already design element/
257 oh dear these levels are so incongruent/
258 or they seems to be so many different levels/
259 so he's interrupting to go back to elaborate on a design component/
260 well I guess that's strictly speaking design/
261 but I don't again I don't like that/
262 I'll use the S6 code by default/
263 stop the designing task to look for the utensil/
264 which to me is an interrupt/.

This type of comment can help regrouping behaviors under broader categories and finding precise definitions of behaviors. It is important to keep categories within a single group on the same level of analysis; for instance, describing overt behavior or what the subject is doing should not include categories describing hidden behavior or what the subject is thinking.

Eight judges from the "self control" group had no experience with protocol analysis. This might not be related to their relative experience with

the technique, it may be due to individual differences. However the judges who read through the provided material did give more accurate judgments. In Table 8, column 4, 6 judges who read all presented material and 5 judges who read parts of the presented material were identified as the "self control" group.

Agreement/disagreement indices were compared between "task control" judges and "self control" judges". The means for each group are presented at the bottom of Tables 11 and 12. The results revealed that agreement is generally higher for "self control" judges. The indices are noticeably higher for this group in Protocol 2, in which the higher levels of inference are better represented. According to existing models of the decision process (Montgomery, 1983) the process may be characterized by a search for good arguments supporting a chosen alternative. "Self control" judges seem to adopt this route. The dynamics of the decision process would have to be studied microscopically to understand about whether "task control" judges neglect some important information or dispense with making a final decision.

Table 11

Level of Agreement Among "Task Control" Judges and "Self Control" Judges in Protocol 2

	Task Control Judges			Self Control Judges		
Episodes	Codes	W	V	Codes	W	V
1	P1	67	50	P2	27	73
2	NC	33	83	NC	36	91
3	P3	50	100	P2	36	100
4	P2	33	67	P2	36	100
5-6	P5	50	67	P2	36	100
7	P2	50	67	P2	55	100
8	P3	50	17	P2	55	82
9	—[a]	17	17	P2	45	82
10	P4	50	50	P4	64	91
11-12	S1	67	83	S1	73	91
13-14	S8	50	100	S1	45	91
15	S8	83	67	S1	45	82
16	S5	50	83	S1	55	91
17	P4	50	83	S1	27	64
19	S1	100	100	S1	100	100
20	S5	83	83	S5	91	100
21	S5	50	50	S1	64	82
22	S5	83	100	S1	64	91
23-24	R2	83	83	R2	91	91
25-26	S5	100	100	S5	91	100
28	S5	83	100	S5	82	82
29-30	S5	50	100	S5	55	91
31-32	S8	67	83	S8	64	82
33	S5	67	83	S5	73	82
34-35	S5	100	83	S1	55	82

(table continues)

Table 11 (continued)

	Task Control Judges			Self Control Judges		
Episodes	Codes	W	V	Codes	W	V
36-37	S1/S5,R5	33	67	S1	55	45
38	S1,P4	33	17	R1	64	64
39	A3	33	33	A2	55	64
40-41	S2	83	100	S2	91	100
42	S2	100	100	S2	82	82
43	S2	83	83	S2	91	91
44-45	S5	67	67	S2	73	64
47	S2	33	67	S2	45	55
48-49	S5	33	33	A2	73	82
50-51	A3	50	100	A3	27	91
52	S3	33	67	S3	45	100
53	S3	100	100	S3	100	100
54	S5	83	83	S5	100	100
55	S5	50	83	A3	45	82
56	S3	100	100	S3	64	73
57	—	17	50	A4	27	55
58-59	—	17	17	S2	27	64
60	S3	50	100	S3	73	100
61	S3	50	100	S3	64	100
62	A2	33	83	A2	45	91
Means		58.8	74.4		60.2	85

Note. W = written indices; V = verbal indices; Codes = the code chosen by the majority.

[a] No agreement found, each judge chose a different code.

Table 12

Level of Agreement Among "Task Control" Judges and "Self Control" Judges in Protocol 3

	Task Control Judges			Self Control Judges		
Episodes	Codes	W	V	Codes	W	V
6-7	P4	83	83	P4	64	64
8	S1	50	100	S3	27	100
9	P5	33	50	P6	91	91
10	S10	100	100	S10	100	100
11	S2	83	67	S2	82	91
12	S10	100	100	S10	100	100
13	S2	67	100	S2	82	100
14	S10	100	100	S10	100	100
15	S10	100	100	S10	100	100
16	S10	100	100	S10	82	100
17	—[a]	17	67	S3	36	91
18	S10	67	83	S1	82	91
19	S3	83	100	S3	64	100
20	S10	83	83	S10	82	100
22	A1	33	50	P7	27	45
24	S2	67	100	S2	55	100
25	S10	83	83	S10	82	91
26	S12	50	100	A1	45	91
Means		72.1	87		72.3	91.9

Note. W = written indices; V = verbal indices; Codes = the code chosen by the majority.

[a] No agreement found, each judge chose a different code.

Summary of Findings

The results summarized in Table 13, highlight the key results presented in tables 1 to 12, on all parameters studied. A wide range of agreement among judges was reported when considering their final decision (from 54.1% to 83.2%):

- agreement is higher in smaller groups but the risk for group bias is also higher;
- experience with the task presented seems to be more important for agreement than experience with the technique of protocol analysis;
- protocols from novices at encoding protocols yield higher encoding reliability than protocols from experts because of the more direct links between categories and verbalizations; and,
- more complete record of perceptually observable behaviors would yield higher agreement.

Level of agreement was higher when verbalized justifications for the final decisions were taken into consideration (ranging from 72.3% to 94.4%). The constant increase in agreement when calculating verbal indices instead of written indices was observed. Problems with semantic, time needed to familiarize oneself with the grid, and incongruity of categories does not seem to have an effect on verbal indices.

Table 13

Summary of Results in Percentages on Judges' Agreement

Group of Judges	n	Protocol 2		Protocol 3	
		W	V	W	V
Total	20	54.1	81.0	68.6	89.0
Experienced	10	59.6	79.2	71.1	86.2
Inexperienced	10	65.6	79.2	75.0	86.2
Male	10	58.2	78.7	73.4	86.7
Female	10	66.9	83.4	72.8	87.8
Line by line analyzers	3	67.0	77.6	79.4	86.9
Context analyzers					
global Approach	3	73.7	72.3	83.2	94.4
specific Approach	3	63.3	87.2	73.8	92.5
Task Control	6	58.8	74.4	72.1	87.0
Self Control	11	60.2	85.0	72.3	91.9

Note. W = written indices; V = verbal indices.

The main findings from the judge characteristic analyses indicated that total independence of segments is not always possible when encoding protocols collected while performing an ill-defined task; problems may arise when using context, decisions are not always independent of each other; and, higher agreement is found among "self control" judges who are more concerned with accuracy than accomplishing the task.

The outstanding feature of protocol analysis is its exploratory nature. Many characteristics associated with the judges were uncovered, although a deeper analysis of the decision process used by each judge is beyond the scope of this thesis, it would complement the information already obtained.

Chapter 5

Discussion and Recommendations

When an experimenter wishes to check the reliability of encoding verbal protocols with a developed grid, judges have to be selected and then a task must be designed. Some criteria for developing a reliable grid, choosing judges, and designing an encoding task can be derived from the present study.

Criteria for Choosing Judges

A higher level of agreement among females than among males was observed. Task experience, preparing a three course meal, was associated with this gender difference. If task dependent knowledge increases reliability among judges when applying a grid to verbal protocols, it could be concluded that the knowledge base is a relevant factor when choosing a judge.

Some decisions were guided by the grid and other decisions were the personal justifications of the judges. The "self control" judges contributed valuable personal justifications and comments for the experimenter. According to Ranyard and Crozier's study (1983), decision making can be seen as involving two interdependent activities – making the choice and justifying it. They suggest that justifications are generated by decision

processes. In these preliminary findings judges who provided reasons for their choices seem to characterize the "critical" judges. A deeper analysis of the verbal protocols would allow insights into the decision process.

Empirical evidence from this study suggest that a person with less experience in protocol analysis may provide a richer bank of information than a person with extensive experience with the technique. A non-experienced person tends to display every step of the process under study.

Some judges continuously wondered about what the student was thinking while talking. Others preferred limiting themselves to what was said. For a more general analysis (low level inference), where the explicit is being analyzed, an experimenter could choose the more literal type of judge (the line-by-line analyzer), while a deeper level processing analysis would benefit from a judge who analyzes using context information and with a "specific" approach to the task.

Newell and Simon's (1972) position is that the task in which people engage structures to a great extent the information selected from the situation. The task also influences how that information is processed. In other words, the subject's interpretation of the task is determined both by their available concepts and by their purposes in the particular situation. Judges were faced with choosing one or several alternatives (codes) of a possible 39 relating to preparation of menu. It was not feasible to consider all the options within a reasonable time frame. If only a few alternatives had been presented instead, the judges might have chosen to explore them all. If a different subject matter was presented, judges would be using a

different knowledge base and the processing could be different. In the "judges characteristic analysis", the nature of the task is revealed but not necessarily the nature of the judge. Further research is needed to investigate whether their strategies would vary with a different task.

Criteria for Designing an Encoding Task

Three main criteria for designing a pertinent encoding task have emerged: extent of context to be used; a case for ill-defined problems; and, extent of the practice.

Extent of context to be used

When context is used in interpretation, the evidence provided by each statement is no longer wholly independent of the evidence drawn from other statements. The amount of context that must be considered when encoding a segment is a function of the type of encoding - low-level encoding or high-level encoding. In this study, an ill-defined task was being encoded and a knowledge of previous or subsequent statements was often needed for exact interpretations. For instance, when encoding Protocol 3, evidence provided in the first two lines of the protocol became important for an exact interpretation of the subsequent statements. The depth of the encoding varied, narrow use of context was possible, when encoding at a shallow level. For instance, total agreement was found, when the link between statement and mental activity was direct.

A case for the ill-defined tasks

An ill-defined problem can be encoded with a high level of agreement as the results from the verbal indices indicated. Since segmentation had an influence on the justifications for a choice, if each segment contained only one idea, it virtually guaranteed that the verbalized justifications of judges would be related to that idea. To illustrate: lines 36 and 37 of Protocol 2 were analyzed as one segment in the transcript and contained three ideas,

> "well can't grate Brie very well so I'll use the Mozzarella and I guess I don't have any Gruyère"
> (Gillespie, 1988).

Disagreement developed because judges used selective encoding, they decided to base their decision on either: (1) Brie can't be grated; (2) using Mozzarella; or, (3) checking to see if there is any Gruyère. Moreover, judges who used context information before making a decision solved the dilemma by saying that the student was (4) "substituting one food item for another to adhere to restrictions on the food list", which is an all-inclusive interpretation.

Semantic confusion was observed throughout the analysis; conceivably such problems could be particularly numerous in ill-defined tasks. Open discussion between judges and experimenter following the practice session is thus primary for reducing the various semantic problems.

Extent of the practice

The importance of discussing and agreeing on the meanings of words has been emphasized. Three protocol extracts have been chosen to further illustrate this point. Task confusion was created by the use of the words: course, ingredient, component, and dish (underlined in the text).

966 (READS) um oh um I'll have carrots/
967 ok so she's also ... she's listing the ingredients really but at
968 the same time so that's/
969 no not a particular dish well can carrots be considered part of a
970 dish part of the main course yes/
971 ... so it's ... I don't know how to code this/
(judge #18)

1076 now I'm thinking of a plate a main menu as the main dish or
1077 whatever main course/
1078 the model (stress) of the main course is a three part plate so I'll
1079 use that/ (judge #17)

139 basically ... she's not exactly going back to an already designed
140 component/
141 but she's trying to re.. you know to reestablish whether she can
142 have the .../
143 already what she had planned out earlier/
144 which was I think a salad/ (judge #7)

If training focuses on narrowing semantic problems, decisions will be congruent with agreed upon definitions, thus increasing agreement among judges.

Information about the physical environment is important because it is sometimes difficult to capture adequately. Newell and Simon (1972) in their theory of problem-solving emphasized the importance for the experimenter to help define the problem space, so that the accumulating knowledge can be correct about the problem situation. Applying this principle to ill-defined problems, opening statements in Protocol 2 were ambiguous because the perceptual context was not well understood or described. In this instance, an annotation could have been included: "subject simultaneously writing on two different task sheets". If a more complete record of the information that is perceptually available is preserved with the transcripts, this information can be discussed with the judges prior to encoding.

The practice can also include discussion on the type of information to be encoded. There is a difference between what the subject is doing and what he is thinking. Encoding processes that generated the "heeded" information or encoding the information itself are two different classes of activities, and these should be distinguished.

Some problems encountered with the present design of the task may be useful indicators for future studies:

(1) Pauses were not clearly indicated in the protocols, and although timing each pause is tremendously tedious, a general indicator of the time elapsed may capture relevant information for the coder without becoming ponderous, such as one or many dashes depending on the length of the pause. A three second interval as the basic unit based on empirical evidence (Rowe, 1985) suggests that such an interval may provide an optimal indication of behavior

over time. These naturally occurring units are supposed to provide a basis for reliable protocol analysis (Goor & Sommerfeld, 1974).

(2) The written instructions were not given to the judges, and since the problem was complex, it would have been useful to many of them; avoiding interruptions which may change the course of thinking in some cases.

(3) The segmentation of the protocol influenced the judges. The lines that were grouped (no double space between them) were analyzed as one segment by all judges. This would indicate that separating a verbal protocol into segments can influence the judges in their decision process.

Criteria for Developing a Reliable Grid

To maintain consistency in judgments, based on protocol information indicating that some categories were missing, a grid must contain a category for every segment to be encoded.

The generality or specificity of definitions should be maintained throughout the grid. When specificity led to naming the different categories of food which could be included in a course, all possible categories should be included (bread and butter, etc). When comprehensive definitions are given, the more specific categories must be components within the larger category. Otherwise, codes such as S1, S2, and S3, may become a disposal for ambiguous segments. These codes were comprehensive and general (designing the appetizer, designing the main course, or designing the dessert), and could have been used to code the whole protocol. Further

argument in support of unilevel categories was seen in the use of the P categories (problem identification and representation). These were not all on the same level: P7 and P8 are judgments while P1 to P6 are actions being performed by the student (see grid in Appendix C). This may partially explain why some codes are used as fillers.

A definition must discriminate well as was seen with the use of codes S13 and S14. These were not exclusive of code S10 and some judges decided to use the more global code and never the more specific one.

When these recommendations guide the instructions and the practice of an encoding task, agreement among judges could increase.

Implications for Theory, Research and Education

This exploratory study provides a basis for determining criteria which will improve the reliability of the encoding technique when verbal protocols are used as data. The findings support the notion that information provided in verbal protocols can reliably be extracted. In the present study, the lower consistency in encoding both protocols was caused largely by failure to agree on a specific category, not in interpreting verbal information. An analysis of the divergent interpretations would certainly increase information on the lower consistencies found.

A more microscopic analysis of the collected protocols and identification of decision process patterns may help differentiate further

between the two selected groups, the judges with experience in protocol analysis and the judges with no experience in protocol analysis.

Identifying decision process patterns from the verbal information is difficult because the characteristics of a given decision may vary and be task dependent. To improve the search for patterns, the task specific simulations should be done in different situations with continuous variations of various parameters associated with information processing. Focusing on key attributes such as: a well-defined task, and/or judges with/without task dependent experience, may prove valuable in future research.

The results of the present study have helped established criteria to improve the reliability of a process tracing technique. Cognitive models induced from more reliable protocol analysis will help create more testable educational theories. Educators are placing more emphasis upon the understanding of processes underlying performance in learning situations, and reliable identification of processes within a problem-solving context can benefit instruction and learning in the educational system.

There has been an increasing interest in teaching critical thinking, this is motivated by evidence that learning large amounts of information does not provide information to analyze, synthesize, and evaluate that information for personal use. Diagnostic methods based on information processing can provide insights about what it is students are doing while they seek to understand and learn, and guide them in their thinking.

In preserving the naturalistic context and allowing for diversity, the quality of intellectual abilities are studied. In knowing on which basis differentiation is really made among students would certainly be of value to educational qualitative research.

Concurrent verbal reports provide researchers and educational practitioners with information on cognitive processes that have been unknown in the past. Integration of reliable verbal data and product oriented data would give a more developed and comprehensive description of cognitive structures and processes, as contended by Messick and Sternberg on measurement of mental abilities.

References

Anderson, J. R. (1982). Acquisition of cognitive skill. Psychological Review, 89,(4), 369-406.

Anderson, M. A. (1986). Protocol analysis: A methodology for exploring the information processing of gifted students. Gifted Child Quarterly, 30(1), 28-32.

Anzai, Y., & Simon, H. A. (1989). The theory of learning by doing. In H. A. Simon (Ed.), Model of thought (pp. 116-133). New Haven: Yale University Press.

Arbib, M. A. (1982). From artificial intelligence to neurolinguistics. In M. A. Arbib, D. Caplan, & J. C. Marshall (Ed.), Neural models of language processes (pp. 77-94). New York: Academic Press.

Ashcraft, M. H. (1982). The development of mental arithmetic: A chronometric approach. Developmental Review, 2, 213-236.

Ashcraft, M. H. (1983). Procedural knowledge versus fact retrieval in mental arithmetic: A reply to Baroody. Developmental Review, 3, 231-235.

Ashcraft, M. H., Fierman, B. A., & Bartolotta, R. (1984). The production and verification tasks in mental addition: An empirical comparison. Developmental Review, 4, 157-170.

Baroody, A. J. (1983). The development of procedural knowledge: An alternative explanation for chronometric trends of mental arithmetic. Developmental Review, 3, 225-230

Bereiter, C., & Scardamalia, M. (1983). Levels of inquiry in writing research. In P. Mosenthal, L. Tamor, & S. A. Walmsley (Ed.), Research on writing (pp. 3-25). New York: Longman.

Bashkar, R., & Simon, H. A. (1977). Problem solving in semantically rich domains: An example from engeneering thermodynamics. Cognitive Science, 1, 193-215.

Bower, A. C., & King, W. L. (1967). The effect of number of irrelevant stimulus dimensions, verbalization, and sex on learning bi-conditional classification rules. Psychonomic Science, 8, 453-454.

Brehmer, B. (1974). Hypotheses about relations between scaled variables in the learning of probalistic inference task. Organizational Behavior and Human Performance, 11, 1-27.

Bruner, J. (1984). Notes on the cognitive revolution. Interchange, 15(3), 1-8.

Carey, S. (1986). Cognitive science and science education. American Psychologist, 41(10), 1123-1130.

Caverni, J.-P., (1988). La verbalisation comme source d'observables pour l'étude du fonctionnement cognitif. Dans J.-P. Caverni, C. Bastien, P. Mendelsohn, G. Tiberghien, (Ed.), Psychologie cognitive: Modèles et méthodes (pp. 253-273). Grenoble: Presse Universitaire de Grenoble.

Chaiklin, S. (1984). On the nature of verbal rules and their role in problem-solving. Cognitive Science, 8, 131-155.

Churchland, P. M. (1988). Matter and consciousness: A contemporary introduction to the philosophy of mind (rev. ed.). Cambridge, Massachusetts: The MIT Press.

Cooper, M., & Holzman, M. (1983). Talking about protocols. College Composition and Communication, 34, 284-296.

Cronbach, L. (1984). Essentials of Psychological Testing. New York: Harper & Row, Publishers.

Cronbach, L. (1986). Signs of optimism for intelligence testing. Educational Measurement Issues and Practice, 5(5), 23-24.

Dobrin, D. N. (1986). Protocols once more. College English, 48(7), 713-725.

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, Massachusetts: The MIT Press.

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 (Ed.), Analysing and aiding decision processes. Amsterdam: North Holland.

Flaherty, E. (1973). Cognitive processes Used in Solving Mathematical Problems, unpublished doctoral dissertation, Boston University.

Flaherty, E. G. (1974). The thinking aloud technique and problem solving ability. Journal of Educational Research, 68, 223-225.

Frick, T., & Semmel, M. I. (1978). Observer agreement and reliabilities of classroom observational measures. Review of Educational Research, 48(1), 157-184.

Gagné, R. H., & Smith, E. C. (1962). A study of the effects of verbalization on problem solving. Journal of Experimental Psychology, 63, 12-18.

Gardner, H. (1985). The mind's new science: A history of cognitive revolution. New York: Basic.

Gillespie, S. (1988). [Verbal protocols collected at a College of Agriculture and Technology]. Unpublished data.

Gilmartin, K. J., & Newell, A. (1976). A program modeling short-term memory under strategy control. In H. A. Simon (Ed.), Model of thought (pp. 84-93). New Haven: Yale University Press.

Glaser, R., & Pellegrino, J. W. (1978). Uniting cognitive process theory and differential psychology: Back home from the wars. Intelligence, 2, 305-319.

Goor A., & Sommerfeld, R. E. (1974). A comparison of problem-solving processes of creative students and noncreative students. Journal of educational Psychology, 67, 495-505.

Greeno, J. G., & Simon, H. A. (1974). Processes for sequence production. In H. A. Simon (Ed.), Model of thought (pp. 344-356). New Haven: Yale University Press.

Haines, G. H. (1974). Process models of consumer decision making. In G. D. Hughes & M. L. Ray (Ed.), Buyer/consumer information processing.

Chapel Hill: University of North Carolina Press.

Hayes, J. R. (1981). The complete problem solver (pp. 51-92). Philadelphia, Penn.: The Franklin Institute Press.

Hayes, J. R., & Flower, L. S. (1981). Identifying the organization of writing processes. In L. W. Gregg & E. R. Steinberg (Ed.), Cognitive processes in writing (pp. 3-30). Hilldale, NJ: Erlbaum.

Hayes, J. R., & Flower, L. S. (1983). Uncovering cognitive processes in writing. An introduction to protocol analysis. In P. Mosenthal, L. Tamor, & S. Walmsley (Ed.), Research in writing: Principles and methods (pp. 207-220). New York: Longman.

Johnson, D. M. (1972). A systematic introduction to the psychology of thinking. New York: Harper Row.

Karpf, D. A. (1973). Thinking aloud in human discrimination learning. Dissertation Abstracts International, 33, 6111B (University Microfilms International no. 73-13625).

Kaye, D. B., Post, T. A., Hall, V. C., & Dineen, J. T. (1986). Emergence of information-retrieval strategies in numerical cognition: A developmental study. Cognition and instruction, 3(2), 127-150.

Klatzky, R. L. (1980). Human Memory: Structures and Processes (pp. 6-26). New York: W. H. Freeman and Company.

Klinger, E. (1974). Utterances to evaluate steps and control attention distinguish operant from respondent thought while thinking out loud. Bulletin of the Psychonomic Society. 4, 44-45.

Lesgold, A. M. (1984). Human skill in a computerized society: Complex skills and their acquisition. Behavior Research Method, Instruments, & Computers, 16(2), 79-87.

Lesgold, A. M. (1988). Problem solving. In R. J. Sternberg, & E. E. Smith (Ed.), The psychology of human thought (pp. 188-213). Cambridge: Cambridge University Press.

Martindale, C. (1981). Cognition and consciousness. Homewood Illinois: The Dorsey Press.

Messick, S. (1973). Multivariate models of cognition and personality: The need for both process and structure in psychological theory and measurement. In J. R. Royce (Ed.), Multivariate analysis and psychological theory. New York: Academic Press.

Messick, S. (1984). The psychology of educational measurement. Journal of Educational Measurement, 21, 215-237.

Miller, G. A. (1956). The magical number seven, plus or minu two: some limits on our capacity for processing information. The Psychological Review, 63(2), 81-97.

Minsky, M. (1975). A framework for representing knowledge. In P. H. Winston (Ed.), The Psychology of Computer Vision (pp. 211-277). New York: Mcgraw-Hill.

Montgomery, H. (1977). A study in intransitive preferences using a think aloud procedure. In H. Jungermann & G. de Zeeuw (Ed.), Decision making and change in human affairs. Dordrecht: D. Reidel Publishing.

Montgomery, H. (1983). Decision rules and the search for dominance structure: towards a process model of decision making. In P. Humphreys, O. Svenson, A. Vari (Ed.), Analysing and aiding decision process. Amsterdam: North Holland.

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(3), 231-259.

Norman, D. A. (1970). Models of human memory. New York: Academic Press.

Randall, A., Fairbanks, M. M., & Kennedy, M. L. (1986). Using think-aloud protocols diagnostically with college readers. Reading Research and Instruction, 25(4), 240-253.

Ranyard, R., & Crozier, R. (1983). Reasons given for risky judgment and choice: A comparison of three tasks. In P. Humphreys, O. Svenson, A. Vari (Ed.), Analysing and aiding decision process. Amsterdam: North Holland.

Resnick, L. B., & Ford, . W. (1981). The psychology of mathematics for instruction. Hillsdale, New Jersey: Lawrence Erlbaum Associates, Publishers.

Rowe, H. A. (1985). Problem solving and intelligence. New Jersey: Lawrence Erlbaum Associates, Publishers.

Rumelhart, D. E., & Norman, D. A. (1985). Representation of knowledge. In A. M. Aitkenhead, & J. M. Slack, Issues in cognition modeling. New

Jersey: Lawrence Erlbaum Associates, Publishers.

Rumelhart, D. E., & Ortony, A. (1977). The representation of knowledge in memory. In R. C. Anderson, R. J. Spiro, & W. E. Montague (Ed.), Schooling and the acquisition of knowledge (pp. 99-135). Hillsdale, NJ: Erlbaum.

Russel, R. L. & Ginsburg, H. P. (1984). Cognitive analysis of children's mathematics difficulties. Cognition and Instruction, 1(2), 217-244.

Schank, R., & Abelson, R. (1977). Scripts, plans, goals and understanding. Hillsdale, N. J.: Lawrence Erlbaum Associates.

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. Psychological Review, 84(2), 127-175.

Simon, H. A. (1976). The information storage system called "Human memory". In H. A. Simon (Ed.), Model of thought (pp. 62-83). New Haven: Yale University Press.

Simon, H. A., Langley, P. W., & Bradshaw, G. L. (1981). Scientific discovery as problem solving. Synthese, 47, 1-27.

Simons, H. D. (1971). Reading comprehension: the need for a new perspective. Reading Research Quarterly, 4, 103-107.

Smith, E. R., & Miller, F. D. (1978). Limits on perception of cognitive processes: A reply to Nisbett and Wilson. Psychological Review, 85(4), 355-362.

Steinberg, E. R. (1986). Protocols, retrospective reports, and the stream of consciousness. College English, 48(7), 697-725.

Sternberg, R. J. (1984). Toward a triarchic theory of human intelligence. The Behavioral and Brain Sciences, 7, 269-315.

Svenson, O. (1983). Scaling evaluative statements in verbal protocols from decision processes. In P. Humphreys, O. Svenson, & A. Vari (Ed.), Analysing and aiding decision process. Amsterdam: North Holland.

Thorndike, R. (1986). Comments on Angoff's "Some contributions of the college board SAT...". Educational Measurement Issues and Practice, 5(5), 25.

Vocate, D. R. (1987). The theory of A. R. Luria: Functions of spoken language in the development of higher mental processes. New Jersey: Lawrence Erlbaum Associates, Publishers.

Waterman, D. A., & Newell, A. (1971). Protocol analysis as a task for artificial intelligence. Artificial Intelligence, 2, 285-318.

Waterman, D. A., & Newell, A. (1973). PAS-II: An interactive task free version of an automatic protocol analysis system. In Proceedings of the Third IJCAI. Menlo Park, CA: Stanford Research Institute, 431-445.

White, P. (1980). Limitations on verbal reports of internal events: A refutation of Nisbett and Wilson and Bem. Psychological Review, 87(1), 105-112.

Appendix A

Instructions to the judges

A method has been devised to allow some of the things going through a person's mind to be revealed. This technique is very simply to ask people to "think-aloud" while they work on something, and explain why they do the things they do. The verbal monologue is then transcribed.

Today you will be working with transcripts which have been collected. Your job will be to use a set of predetermined codes and assign them to segments of the transcript.

The transcripts are from College students who were presented with these instructions (show the documents). Their task was to design the menu for a dinner party for six people.

Designing a 3 course meal was what they had to focus on and two subtasks were included:

- determining amounts & costs of food,
- estimating a time schedule for the preparation of the meal.

The students' verbalizations were recorded and later transcribed into these protocols.

The researcher than developed a list of categories (show the grid) that would best describe the segments of the protocols, and each was assigned a code.

What I would like you to do is use these codes to categorize these two transcripts or protocols from two different students chosen from the group of 22 who performed the task.

The sections you will encode do not include the time task; and since you have only two subjects, some of the categories might not apply to the particular protocol. While you are doing this I would like you to explain aloud <u>why</u> you do what you do.

When you are reading any material, you can read out loud, and also tell me what is going through your mind as you read.

During that time I will be recording your verbalizations, making a few notes, and if necessary reminding you to continue thinking aloud. The tapes and transcripts will be kept confidential.

- Do you have any questions?

We can now try a sample, and feel free to ask any questions since this is only a practice.

After we start I don't want to interfere with your thinking process so I will only listen and remind you to keep talking if necessary.

Appendix B

Protocol 1

1 - S: ok what I'm going to do is just look over the different things

2 - that we have on our list and decide what I want for my meal

3 - we have an appetizer and then the main meal and then the desert

4 - so for my appetizer I think I'd like to have a punch of some sort

5 - so I'm going to take some strawberries and puree them later

6 - some lemon juice if I can find it

7 - lemon juice over here

8 - and I put that in the wrong place

9 - so I'll take ok I'll have lemon juice and I'll take

10 - um three tbsp of lemon juice

11 - E: could you reserve this for the cost

12 - S: ok the cost do you want cost of this one over here no just the

13 - persishable items are supposed to be costed

14 - strawberries I'll have uh a pint no a quart of those

15 - oh and we'll get some sugar

16 - ok um probably about 3 tbsp of sugar and some gingerale

17 - I'll get um for 6 guests I'll get uh 1 bottle

18 - um for my food for the appetizer

Protocol 1 (continued)

19 - um I'd like to have I guess I'll have some sausage rolls if

20 - there's any sausage

21 - no sausage so pause is there any crackers or anything either

22 - so I'll make something from scratch there's none of those

23 - um could we put maccaroni down for those shells those maccaroni

24 - shells you know

25 - ok I'll put pasta shells down

Protocol 2

my food list

well I'll have a cup of coffee first

now my food list the rest of my food list

spices and alcoholic beverages

my time sheet for preparing my menu

my sheet for writing down my menu

ok now food list and a menu sheet and a time sheet

design a three course menu

check off those food items you plan to use

so I pretty well have to know the recipes for what I'm using

so I could either start with the soup

or a salad

I wonder if there's any avocados on this

let's see no no avocadoes

how about onions no only cooking onions

there's mushrooms

well I guess we should go for something classic

E: tell me what you're writing as you go

S: ok French onion soup

and I check off my cooking onions

and I'm going to need something to sauté those in

so let me see do you have any shortening lard or vegetable oil

Protocol 2 (continued)

oh I better be a good dietitian here and use vegetable oil it

doesn't have as much saturated fat

and a little bit of flour and a little bit of sugar

and a little bit of salt where's the salt?

E: it's on the second sheet

S: ok there's my salt and then I want some beef some consommé

ok now I need some bread for some croutons hope I don't have to make the bread

here we are here's some fresh bread so that maybe we'll just put a slice of French bread instead of the croutons

and some cheeses somewhere

what do we have

we have Mozzarella, and Cheddar, and Brie

well can't grate Brie very well so I'll use the Mozzarella and I guess I don't have any Gruyère or Swiss cheese

well we'll mix it with Cheddar give it a little bit more flavour

so that takes care of the onion soup

and ok lamb chops are nice we could have broiled lamb chops I'll write that broiled lamb chops

and our vegetable here broiled lamb chops ok with broccoli

and we could put some rice and mushrooms with that so

ok mushrooms we need here's our rice need a bit of butter with the broccoli I don't see any butter

E: it's in the dairy section

S: the dairy products ok

Protocol 2 (continued)

ok that takes care of our broiled lamb chops, rice and mushrooms,

and some broccoli

and now I suppose if I serve a salad with that I've lost my three courses

ok so we need a dessert

how about a baked alaska that sounds nice

ice cream and eggs whites and sugar

I have sugar checked already so

baked alaska, now that's all and I suppose tea or coffee

what about bread and butter is that considered a add that as well

so maybe some nice rolls and I'll buy those so I don't have to make them instant dinner rolls and I already have my butter

and then a beverage

tea coffee or milk and I suppose we want a red wine with this

ok so that's my menu

Protocol 3

I'm taking the food list and I'm going to pick out the foods I'd like for this dinner for six people or six adult

um we'll have coffee

do I have to write the amount
(yes)

um 15 no 2000 ml. of coffee in the pot and um

we're going to have some um some chicken breasts

have 12 is this where I write them down (yes)

um oh um I'll have carrots

um 500 ml. um

dinner rolls 12 dinner rolls and

some potatoes 6 potatoes and

for dessert um oh umwhat goes with a three course meal

um cream of chicken soup 6 cans I guess

and um long pause ice cream for dessert

1500 ml.

E: could you speak out a little bit louder

S: oh ok um ok oh I'll have take the pantry stock or page two

E: yes that goes with your food list

S: oh um we're going to have dry red wine no chicken yes

2 bottles

then I'll um make up the time sheet

Appendix C

LOCAL PLANNING OPERATIONS

Subject uses a combination of these operations

P: Problem Identification and Representation

- P1 - Reads instructions
- P2 - Rereads or summarizes all or part of instructions
- P3 - Identifies or defines problem
- P4 - Chooses strategy for attacking the problem
- P5 - Refers to an element in the problem statement
- P6 - Questions E about problem (for clarification only)
- P7 - Fails to consider all important elements in problem
- P8 - Does not have specific vocabulary/domain specific knowledge to understand problem. Is confused, therefore, about the task

S: Solution Related Activity

- S1 - Designs appetizer - that course which comes before the main part of the meal; may consist of soup(s), salad(sl), hors d'oeuvres(h), beverage(b)
- S2 - Designs the main course of the meal; may consist of fish/meat(m), potatoes/pasta/rice(c), vegetable(v), salad(sl), beverage(b)
- S3 - Designs dessert; may consist of food(f), beverage(b)
- S4 - Garnishes a particular dish for aesthetic appeal
- S5 - Lists ingredients in recipe for a particular dish
- S6 - Interrupts work on one course/dish to go back to elaborate on an already designed component
- S7 - Interrupts work on time-planning task to go back to menu-planning task and make additions/changes to it
- S8 - Substitutes one food item for another to adhere to restrictions of the food lists
- S9 - Modifies or changes a dish for reasons other than #S8
- S10 - Calculates amounts/quantities of food items needed for six people
- S11 - Calculates cost of a particular item
- S12 - Works on time-planning task, deciding when and how to prepare foods for serving
- S13 - Misjudges amounts appropriate for six people
- S14 - Confuses amount of food/beverage served with amount of raw ingredient needed to prepare the dish

(continues)

M: Memory Related Activity
- M1 - Attempts to recall previous experience to use as a model
- M2 - Attempts to associate, compare, relate to previous experience outside task

C: Changing the Conditions of the Problem
- C1 - Requests that new items be added to the food lists
- C2 - Consciously uses items which are not on the food lists

R: Reasons for Particular Food Choices
- R1 - Explains that a food choice has been made for aesthetic reasons, relating to sensory appeal
- R2 - Explains that a food choice has been made for nutritional reasons
- R3 - Indicates awareness of cost factor when choosing a particular food
- R4 - Explains that a food choice has been made on basis of cost
- R5 - Explains that a food choice has been made to illustrate subject's expertise in the field

A: Assessing and Monitoring Activity
- A1 - Locates position with respect to entire task (monitoring)
- A2 - Summarizes what has been done
- A3 - Assesses a portion of the task
- A4 - Makes changes/additions to the menu as a result of the assessment process
- A5 - Evaluates the entire menu at the end of the task

E: Emotional Reactions
- E1 - Comments explicitly on the difficulty (-) or ease (+) of problem
- E2 - Expresses feeling of confusion or frustration with the task
- E3 - Expresses feeling of accomplishment, relief, or pleasure from the task

Appendix D

Table D-1

Frequencies of Codes Used in Protocol 2

	Codes																														
Episodes	P1	P2	P3	P4	P5	P6	P7	P8	S1	S2	S3	S4	S5	S6	S8	S9	S10	H1	H2	R1	R2	R3	R5	A1	A2	A3	A4	A5	E1	E2	E3	N/C
1	5	5	5		2																			1								5
2		1	3	5					2															1								8
3	2	6	8		3																			4								
4	3	6	6		4								1											4								
5-6	3	6	5	1	7																			4								
7	1	10	4		4																			2	3							
8	2	8	9		2																											1
9	3	7	2	4	4								1													1						1
10		1	2	13	1				1									1	1					2		3						
11-12			1	6	1				12						1																	
13-14		3	1		2				8				7		4																	
15		3			2				8				4	1	8											1	1					
16		3			2				8				5	1	3				1							1	1					
17				6					4					1	1			3	1	3			2			1	1					
19				2					19					1	2			1								1	1					
20		1		1					9				17		1			1	1					1		1	1					
21			1	1					10				7	1	1			1					2			4						
22		2		2	2	1			10				10		1			1	1				1			2						
23-24				1	1				5				1			1					18		2				1					
25-26		1				7			9				19					1	1							1						
28		1							9				17					1								2						
29-30		1			1				7		1		12					2								2			4	2		
31-32		1			1				8			2	6		10	4	1	1	1				1			1						
33		1		1	3				8				14					1	1							2						
34-35		3			3				8				11					1						1		2						

(table continues)

Table D-1 (continued)

	Codes																													
Episodes	P1	P2	P3	P4	P5	P6	P7	P8	S1	S2	S3	S4	S5	S6	S8	S9	S10	H1	H2	R1	R2	R3	R5	A1	A2	A3	A4	A5	E1	E2	E3	N/C
36-37		1		1	1				8				6		4	1		2					7			3						
38				2	1				8			1	8		3	1		1		9			3			1						
39					1				5							1								2	8	6				2		
40-41		1	1	1						19			4			1		1	1					1	1	1						
42		1		1	1					16			6					2	1	1			1		2	1						
43		1		1						17			7					3													
44-45		2		1	2		1			11			11					3					1		1	2			1		
47		3		1	3					7			3					1							1	1					2
48-49					1					7			3					1						1	10	3			1	1		
50-51					4		1	1		2				1				1						2		7	3		1			
52		2	3	2	2						8							1						4								
53											20	1	1					2		4			2							2		
54											8		19					1	1				1	1								
55					2						4		8					2							3	7						
56				1						2	13		3					1	1					1	8	3		1		1		
57		1	1		1	3	1	1	1	1	1	2	2	2		1		2	1						1	2	5	2				
58-59		1		1	3				1	4	2	2	4	3		1		1		1		1			1	2	5			2		
60				1	1					4	12		2	1				1								2	2					
61				1	1					6	11		4	2				1								2	2					
62				1	1						1							1						2	8	1		4		2		

Note. N/C = no code was chosen. Codes not used have been omitted.

Table D-2

Frequencies of Codes Used in Protocol 3

	Codes																								
Episodes	P1	P2	P3	P4	P5	P6	P7	S1	S2	S3	S4	S5	S6	S9	S10	S12	S13	S14	H1	H2	R1	A1	A2	A3	A4	N/C
6-7		1	4	14	2			1	1	1					1									1		1
8				2	1			4	3	5		3												1	1	3
9		1	2		2	13									1							1				
10	1									1		4			18		5		1					1		
11		1							15			4			1									1	1	
12		1	1			7			2						20		2		1							
13			1	1	1				14			6												1	1	
14				1					2						20		4							1		
15								1	12			6			20									1	1	
16				1					13			6			18		1							1	1	
17			2	1	3	1	1	1		7		1			1				3	2		4		3	1	
18				1	1			12		1		2	4		14		7	1						1	1	
19			1	1	1					13		3	2		1					1				1	2	
20												1			17		4									
22	2	2		1	3	3	2		3	1	1											2		2		1
24				1					12			3	2	1							1		1	2	2	1
25									4			1			16		2								1	
26			3	3	2											5						8				

Note. N/C = no code was chosen. Codes not used have been omitted.

Appendix E

Prototype proposition extraction

Episodes 13 & 14

judge #16
289 I wonder if there's any avocados on this/
290 ok so she's checking the food menu/
291 wondering if there's any avocados um .../

306 she's just wondering if there's avocados/

317 well she's not really ... she's just looking for avocados/

323 ... she's just listing she's just thinking I wonder if there's
324 avocados/
325 and there aren't any/
326 so is she listing the ingredients in a recipe for a particular
327 dish/
328 she hasn't started saying well I'm going to make guacamole and I'll
329 see if there's avocados/
330 she's just um ... she's just ... I'll just put it under um .../
331 (mumble)
332 lists ingredients in recipe for a particular dish/
333 lists ingredients in recipe for a particular dish/
334 I can't see where else this would come under/

343 S: ok I wonder if there's any avocados on this/
344 let's see no no avocados ok/
345 but it doesn't have here, check the list to see if there's any
346 avocados/
347 (laugh) shouldn't have avocados/

judge #17
472 so (READS) or a salad, I wonder if there's any avocados avocados
473 on this/
474 on this list I imagine/
475 (READS) let's see no no avocados/
476 so there's no avocados/

477 so mmmm what is she doing/
478 she's looking at the food list to see if something would fit with
479 an appetizer/

489 ok (READS) I wonder if there's any avocados on this/
490 no no avocados/
491 so that's still the design of the appetizer/
492 both of those statements about avocados/

judge #1
119 (READS) I wonder if there's any avocados on this, let's see no
120 avocados, how about onions no only cooking onions, there's
121 mushrooms/
122 so the person must be as I'm doing now going to the food list and
123 asking him or herself questions/
124 I'm trying to check against the food list/
125 so um ... that whole little blurb starting with I wonder and ending
126 with there's mushrooms/
127 could be considered under one that's one kind of scheme but which
128 one is the problem/

135 because what this person is doing is asking questions and then
136 checking against the list/
137 how would that be categorized/
138 um ... I guess if you considered the food list as part of
139 instructions then it could come under P2/
140 because the person rereads summarizes/
141 so I'll bet you it's P2 because the person is asking questions and
142 then checking questions against the food list for an answer/
143 so that whole line or the four lines 1 2 3 4 starting with I wonder
144 and ending with there's mushrooms/
145 I'd put under P2 rereads or summarizes all or part of instructions/

judge #6
118 so line 11 to 14 they're looking for what's available to them/
(lines 11 to 14 are 13 to 16 in the standard protocol)

119 so is there anything in the sheet here that says lists/
120 I'm looking for something that says lists ingredients in a recipe
121 for a particular dish/
122 ok so that's an S5/
123 but it's more than that because they're not really listing the
124 ingredients they're looking for the ingredients/

145 so salad would have been her first choice/
146 and there's no avocados so that's lists ingredients/
147 so that's what she must have been going through is the list of
148 ingredients for the salad/
149 so these two lines 11 and 12 must have been listing/
150 I'm writing S5/
151 because that must have been listing the ingredients for the salad/
152 and when it wasn't positive they went to the soup/

judge #18
46 I continue to read (READS) I wonder if there's any avocados on
47 this, let's see no no avocados/
48 so he's checking the list ... checking the list after coming up
49 with an idea .../

243 substitutes one food item for another to adhere to restrictions of
244 food list I would say I would have to say S8/
245 but it's not a direct substitution/
246 like in his mind he thought um avocados would be nice but they're
247 not on the list/
248 so I'll put S8 for now but um.../

288 why doesn't it say somewhere on here that he is looking at
289 instructions/
290 rereads or summarizes all or part of the instructions well if the
291 food list is part of the instructions then I guess it would um
292 grrr/

302 well I would definitely qualify the food list as part of the
303 instructions/

304 because he's got to limit himself to that/
305 so consequently when he's checking the list after coming out with
306 his idea ok the four statements (READS 13, 14, 15, 16)/
307 all those four ok I would have to put as part of rereads reads
308 instructions or part of the instructions/
309 I would put P2 but temporarily/

judge #11
125 now under (READS) I wonder if there's any avocados in this/
126 he's or she is thinking again I would assume there's on this
127 perhaps on the list so .../
128 yeah next line (READS) let's see no no avocados/
129 ok so the person has referred back to the list/

133 lists ingredients well it's starting to get into ingredients/

137 I don't see anything S5 potentially ... ... /

145 obviously that individual is thinking avocados might be nice/
146 again going probably using passed experience of having had avocados
147 in either salad or soup/
148 and thinking that that might be a good ingredient/
149 then checks the list of of items on available/
150 and what I'm trying to do is find where those two fit into this
151 set of categories /

162 by instructions do you refer to the actual menu that's available
163 to the individual/
164 E: the instructions are this/
165 S: just this first page all right .../
166 ok so doesn't apply .../
167 I'm going to put an S5 for the line I wonder if there's any
168 avocados/
169 simply because they're starting to list but then quickly realize
170 that it's not on the list so/

187 I seem to recall that there might have been something there on this
188 idea of going to the list that's what's bothering me/

189 lists of food to be chosen/
190 I just can't seem to find anything on that/
191 ... so they're still designing the appetizer/
192 ... and I don't see any under the S's/
193 so I'm going to have to put S5 for that as well/
194 might have to come back to it later/

judge #7
89 now she is looking at the ... ingredients/
90 are there any avocadoes on this/
91 she's going over the list/
92 then ... so she sees that/
93 let's see no avocadoes/
94 so she doesn't find avocadoes/

97 this is right now she's going through the .../
98 list ingredients yes so it'll be S5/
99 then she's going through the list and sees that there are no
100 avocadoes/
101 then she thinks about other things/
102 so she has to try ... and substitute .../
103 could that be it S8 .../
104 for to adhere to restrictions of the food list yeah/
105 because this is not on the food list/
106 so she has to do that/

judge #3
13 (READS) if there's any avocados, let's see no no avocados,
14 avocados, how about onions no only cooking onions, there's
15 mushrooms/
16 he's just surveying the list/

173 (READS) I wonder if there's any avocados/
174 I'll have to go back to the problem statement the P5 category/
175 let me see there's no no no avocados/
176 so he's referring to an element of the problem statement as
177 restrictions in that list/
178 ok so P5 so/
179 one might assume he's he's um listing ingredients for particular
180 dish/
181 although I don't know that he is at this stage/
182 listing ingredients for particular (stress) dish/
183 I think he's still trying to search for a particular dish/

184 so therefore he's not listing them yet/
185 so he's still working/

judge #12
401 (READS) I wonder if there's any avocadoes, let's see/
402 in a way he yeah ok he's listing ingredients for a particular dish/
403 he's not giving a recipe/
404 but he is he's identifying ingredients that could come in handy/

412 I'm going to say that's S5/
413 those those ones he's just identifying ingredients/
414 ... he's not really even looking at the list/
415 he's just saying on the top of his head he's got some notions I
416 think of of what he wants for either a soup or a salad/
417 and ... he should probably be looking at the list first and doing
418 it the other way/
419 but that's the way it is .../

449 ok so I'm going to throw an S1 up here where I've also got an S5/
450 where he talks about listing ingredients/
451 I think the S1 also applies to the first one where he's wondering
452 about avocadoes/

judge # 5
139 I wonder if there are any avocados on this/
140 now this one refers to a section of the problem actually/
141 and I think ... that's the some referring to the list of ... the
142 list of I don't know whether foods food list/
143 or I don't think whether it should fall under the perishables/
144 fails to consider, refers to an element in the problem, in the
145 problem statement/
146 so here fi.. I'll call this P5/
147 E: do you consider the food list an element of the problem?/

148 S: well since I think it is an element because part of what you
149 are given in the problem/

150 E: ok/
151 S: let's see no no avocado that's examine the food examining the
152 food list/
153 let see fails to consider, refers to an element/
154 so that's P5 again .../
155 this is after referring to the element she comes out with a
156 statement that there's nothing like that in her (mumble)/

judge #8
67 (READS) I wonder if there's any avocados on this/
68 so she's looking at her list/

75 so she's ah ... she's obviously um ... checking her strategies
76 there/
77 she has an idea and then she's going back to check at her food
78 list/
79 and she's saying to herself well it's not there/
80 so I'll call that/
81 again she's looking at um ... her solution related activities/
82 S5 lists ingredients in recipe for particular dish/
83 although she doesn't have one/
84 I might come back to that/

93 so she's looking with these several statements of the avocados the
94 cooking onions and the mushrooms/
95 I'm confirming that yes or no on the list/
96 her strategy her solutions her strategies for ah um designing her
97 first course or her appetizer/
98 I'm going to put those all that group as S1/
99 and maybe come back to that/

judge #9
107 (READS) I wonder if there's any avocadoes on this/
108 so she's going back/
109 and she has to assess/
110 or check make a check/
111 and I am trying to see where this is ... on the sheet .../
112 she's gone back then she's checked/

127 part of this she's checking on whether or not there are ingredients

144 she says substitutes one food item to adhere to restrictions on
145 the food list/
146 and then she modifies or changes/
147 no it's all S8/

judge #2
92 (READS LINES 13 TO 16)/
93 those four lines she's looking for ingredients for the/
94 she's still working on the appetizer here/

judge #13
311 (READS) I wonder if there's any avocados on this/
312 oh! she's looking at her list I guess/
313 I wonder if there's/
314 ok she's looking at the list/
315 she wants to see if she has to buy avocadoes/

366 I don't know what this is let's see no no avocadoes/

371 she's not listing ingredients/
372 because it's a list/
373 unless she needs that why not/
374 S5 no/
375 because she says there aren't any avocadoes/

381 I don't know I'll say it's ah ... identifies the problem I don't
382 know/
383 E: but you're not satisfied/
384 S: no I'm not I am not I can't find the appropriate category/

judge #14
(no recording due to technical problems)

judge #15
73 um I wonder if there's any avocados on this, let's see no avocados/
74 ok I would consider these 2 lines a P2 obviously having to go back
75 to the instructions/

judge #10
84 (READS) I wonder if there's any avocados on this/
85 she's referring to her food list/
86 so she must have looked at it/
87 um and that would be doing the ingredients/
88 so that's um S5/
89 ... and she checks the list/
90 (READS) let's see no no avocados/
91 so she's um ... ... that's problem solving/
92 ... .../
93 E: problem solving you mean problem identification/
94 S: no she says there's no avocados so that's she's designing she's
95 designing the appetizer/
96 um ... and has checked the list to see if there's avocados on it/
97 I'd say that's S1 still designing the appetizer/

417 so how about onions no only cooking onions she really that could
418 be I put it as an S1 but I'd also put it as S8/
419 because she has ... um ruled out the avocados because there's not
420 they're not on the list/

judge #19
289 the person is designing the appetizer/
290 I'm presuming this is an appetizer/
291 um so they're designing the appetizer/

325 let's see if there let's see no avocadoes/
326 the person went to the list/
327 checks list (laughs) pardon me .../
328 oh all right I want to ask you a question is there a place here
329 that says checks list/
330 E: no/
331 S: all right there's no place that says checks list/
332 for here now I think that there should be a thing that says checks
333 list/
334 there isn't one and I know I have to ... assign a number to each
335 one of these things statements/

judge #20
118 no no no avocados on it no no sir no ok see no avocados/
119 so what she did there is simply but she's designing/
120 but she has to refer back all the time/
121 because it's part of the problem/
122 but that's not obvious as so she's it's all part of the design/

125 um ok now there might be something there I wonder if there's any
126 avocados/
127 um she is just checking to see I thought I saw that um .../
141 that's not it either she's checking she's really just checking
142 she's checking all of the possibilities/
143 but with something particular in mind she's got something in her
144 mind/

147 but um you know I have the feeling that she's recalling/
148 inherent in this is a recall of previous experience/
149 because she's got she's got a um she's got something in her mind
150 with avocados/
151 so outside of all that there's likely to be some sort of memory
152 related activity going on/

164 but it's the real um it's almost like predesign/
165 because she hasn't made a decision about it as such/
166 but it is all part of her decision as to what kind of an appetizer
167 she's going to have/
168 so I'm going to call those ones um because that category seems to
169 include an awful lot of other things as one/
(talking about the S1)

judge #4
175 (READS) I wonder if there's any avocados on this/
176 so she's checking the food list/

190 so what she's doing is checking the food list to see if there's an

191 avocado on the food list/
192 and is there a coding instruction for checking the food list/

201 ok there's nothing in there I'm going to put S1 up in there/
202 so that's why I put S there/
203 (READS) I wonder if there's any avocados on this/
204 S1 I don't know what else to put there/

215 no ok I wonder if there's any avocados on this that's an S1/
216 she's checking the food list/
217 but I think she also did an S5/
218 she listed ingredients in recipes for a particular dish/
219 she had avocados in mind for the salad so I put S5 there/
220 and then she went back and checked it on the list/
221 so an S1/

222 (READS) let's see no no avocados/
223 she's made some sign of a .../
224 she's decided that she doesn't have the stuff/
225 and that's still under S1/

Appendix F

Table F-1

Indices of Agreement Among Judges in Protocol 2

	ME		FE		MNE		FNE		Total	
Episodes	W	V	W	V	W	V	W	V	W	V
1	40	60	20	60	60	60	40	80	25	65
2	40	80	20	80	40	80	60	80	25	80
3	40	100	40	100	60	100	60	100	40	100
4	40	100	20	80	40	100	60	100	30	95
5-6	60	100	40	100	40	100	60	100	35	100
7	40	60	60	100	40	80	80	100	50	85
8	60	100	40	80	40	80	60	100	45	55
9	60	100	40	40	40	40	40	100	35	65
10	40	80	60	60	80	80	60	100	65	80
11-12	40	40	80	100	60	60	80	100	60	75
13-14	40	100	60	100	60	80	60	100	40	95
15	40	100	80	80	40	60	60	80	40	80
16	40	100	40	100	60	80	60	80	40	90
17	40	40	40	100	40	60	60	80	30	70
19	80	80	100	100	100	100	100	100	95	95
20	80	80	100	100	60	80	60	100	85	90
21	40	40	80	40	60	60	80	100	50	60
22	40	60	100	100	60	80	80	100	50	95
23-24	100	100	80	80	100	100	80	80	90	90
25-26	100	100	100	100	80	80	100	100	95	100
28	100	100	80	80	80	80	80	80	85	90
29-30	100	100	80	100	80	80	100	100	60	95
31-32	40	80	80	100	40	60	80	80	50	80
33	60	100	80	80	60	80	80	80	70	85
34-35	60	100	100	80	60	80	80	80	55	85
36-37	40	40	60	80	40	40	60	60	40	55
38	60	60	40	20	40	40	60	60	45	45
39	40	40	40	20	60	80	40	60	40	50
40-41	80	100	100	100	100	100	100	100	95	100
42	40	40	100	100	80	80	100	100	80	80

(table continues)

Table F-1 (continued)

	ME		FE		MNE		FNE		Total	
Episodes	W	V	W	V	W	V	W	V	W	V
43	80	80	80	80	80	80	100	100	85	85
44-45	80	100	80	20	60	60	100	100	55	75
47	60	40	40	60	40	60	100	40	35	50
48-49	40	60	40	40	80	80	80	80	50	65
50-51	60	80	60	100	40	100	40	100	35	95
52	40	80	40	60	40	100	40	100	40	85
53	100	100	100	100	100	100	100	100	100	100
54	100	100	80	80	100	100	100	100	95	95
55	40	100	60	60	60	100	40	80	40	85
56	40	60	80	100	80	80	80	80	65	80
57	40	60	40	80	20	20	40	40	25	50
58-59	40	60	40	20	40	40	40	80	20	50
60	40	100	60	100	60	100	80	100	60	100
61	40	100	60	100	60	100	80	100	55	100
62	40	80	20	80	60	100	40	80	40	85
Means	56	80	63	79	60	78	71	88	54	81

Note. ME = males with experience in protocol analysis; FE = females with experience in protocol analysis; MNE = males with no experience; FNE = females with no experience; W = written indices; V = verbal indices; Total = entire group.

Table F-2

Indices of Agreement Among Judges in Protocol 3

	ME		FE		MNE		FNE		Total	
Episodes	W	V	W	V	W	V	W	V	W	V
6-7	60	80	80	80	80	80	60	60	70	75
8	40	60	60	60	60	60	20	100	35	95
9	80	80	20	60	60	60	100	100	65	65
10	80	80	80	80	100	100	100	100	90	100
11	60	80	80	80	80	80	80	80	75	80
12	100	100	100	100	100	100	100	100	100	100
13	60	80	60	100	80	100	80	100	70	95
14	100	100	100	100	100	100	100	100	100	100
15	100	100	100	100	100	100	100	100	100	100
16	100	100	100	100	100	100	80	100	90	100
17	80	100	40	80	40	80	40	80	35	85
18	80	80	80	100	80	80	80	100	60	90
19	20	100	80	100	80	100	80	100	65	100
20	100	100	80	80	80	100	80	100	85	95
22	40	40	40	60	40	40	20	20	15	45
24	60	100	60	100	60	100	60	100	60	100
25	80	80	80	80	80	100	80	80	80	85
26	40	80	40	100	40	100	80	80	40	90
Means	71	86	71	87	76	88	74	89	69	89

Note. ME = males with experience in protocol analysis; FE = females with experience in protocol analysis; MNE = males with no experience; FNE = females with no experience; W = written indices; V = verbal indices; Total = entire group.