Coder's Materials and Reliabilities for Analysis of Thinking Aloud Protocols from Study on the Use of Conditional Probabilities in Probabilistic Inference.

Robert M. Hamm, Cynthia M. Lusk, Michelle A. Miller, Deborah L. Smith, and Ingrid E. Young

Institute of Cognitive Science
University of Colorado
Box 345
Boulder, CO 80309-0345
303/492-2936

Institute of Cognitive Science
Publication Number 88-13

September 1988

This work was sponsored by the U.S. Army Research Institute for the Behavioral and Social Sciences, Contract MDA-903-86-K-O265.

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S)	5. MONITORING ORGANIZATION REPORT NUMBER(S)
6a. NAME OF PERFORMING ORGANIZATION: Institute of Cognitive Science; 6b. OFFICE SYMBOL (If applicable)	7a. NAME OF MONITORING ORGANIZATION
6c. ADDRESS (City, State and ZIP Code): University of Colorado, Box 345 Boulder CO 80309-0345	7b. ADDRESS (City, State and ZIP Code)
8a. NAME OF FUNDING/SPONSORING ORGANIZATION: Army Research Institute; 8b. OFFICE SYMBOL (If applicable)	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER: MDA-903-86-K-0265
8c. ADDRESS (City, State and ZIP Code): 5001 Eisenhower Avenue Alexandria VA 22333-5600	10. SOURCE OF FUNDING NOS.

PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	WORK UNIT NO.

11. TITLE (Include Security Classification): Coder's Materials and Reliabilities...Inference.

12. PERSONAL AUTHOR(S): Robert M. Hamm, Cynthia M. Lusk, Michelle A. Miller, Deborah L. Smith and Ingrid E. Young

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Yr., Mo., Day)	15. PAGE COUNT
Technical	FROM _____ TO _____	1988 September	

16. SUPPLEMENTARY NOTATION: Contracting Officer's Representative was Mike Drillings.

17. COSATI CODES		
FIELD	GROUP	SUB. GR.

18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number): Protocol Analysis, Conditional Probability, Probabilistic Inference, Bayes' Theorem, Expert/Novice

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

This document contains the coder's materials for the protocol analysis used in a study which examined how individuals with varying experience in mathematics interpreted and used conditional probability information ("Confusion of Conditional Probabilities", Hamm and Miller, Institute of Cognitive Science, University of Colorado, Boulder, 1988). It explains the coding procedures used in the analysis of the strategies and answers given by subjects during the experimental sessions. The document also presents information on the reliability of the various coding schemes used in the study.

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS ☐	21. ABSTRACT SECURITY CLASSIFICATION
22a. NAME OF RESPONSIBLE INDIVIDUAL	22b. TELEPHONE NUMBER (Include Area Code) / 22c. OFFICE SYMBOL

DD FORM 1473, 83 APR EDITION OF 1 JAN 73 IS OBSOLETE.

SECURITY CLASSIFICATION OF THIS PAGE

Table of Contents

Hamm, Lusk, Miller, Smith, and Young
September 1988
Coder's Materials and Reliabilities.

1. Introduction.

This paper contains the coder's materials for the second think aloud study in the Probabilistic Inference project supported by ARI Contract MDA-903-86-K-0265 (Confusion of Conditional Probabilities, Hamm and Miller, 1988). Coders needed to understand the information in the introduction, as well as the specific instructions for their own coding tasks, in order to perform accurately.

1.1. Overview.

Subjects thought aloud while answering three probabilistic inference word problems, and while subsequently recalling the first of these problems in a prompted recall procedure. In the first and third problems four paragraphs were presented, and the subject was required to produce an estimate of p(h) after each.

1. The first paragraph simply laid out the situation and identified the two mutually exclusive possibilities.
2. The second paragraph presented base rate information concerning the likelihood that each of the possibilities is the case.
3. The third paragraph presented "evidence", i.e., an observation, indicating that the less likely possibility is the case.
4. The fourth and final paragraph presented conditional probability information pertinent to the reliability of the evidence. [Only the "reliability" coders were told the following: In half of the problems, this was technically the "reliability of the evidence", p(e/h), and in the other half it was the reversed conditional probability, p(h/e).]

In the second problem, after receiving the first paragraph, the subject was presented with descriptions of paragraphs 2, 3, 4a, and 4b and asked to choose the order in which to receive them. That is, the paragraphs were shown with the key information blanked out. The subject read the paragraphs, determined the order in which to receive their information, and then received the information one paragraph at a time and answered after each. [Only the "reliability" coders were told the following: The order in which paragraphs 4a [p(e/h)] and 4b [p(h/e)] were presented was counterbalanced.] In what follows, these are called paragraphs 1, 2, 3, and 4, rather than 2, 3, 4a, and 4b.

Three different problems were used, Doctor, Insurance (Termination) and Twins, and they were presented in all possible positions, across subjects.

1.2. Description of the transcripts.

Tape recordings of the subject's sessions were made. The researcher in these sessions was instructed to have the tape recorder off when the subject read the paragraph, then turn it on and have the subject identify the paragraph by number, read the question (for orientation), and answer the question. The purpose of this editing was to remove the exact wording of the information presented in the paragraph [specifically, whether the conditional probabilities presented had the form p(e/h) or p(h/e)] from the transcript. This tape editing was not always done accurately by the researchers/interviewers, and so the typists were instructed to leave out the reading of the paragraph, if it happened to have been recorded. And finally, Hamm read through the transcripts (as word processing files) and edited out any reading of the paragraphs that had made it through.

New identity codes were assigned to each transcript, so that the numbers did not indicate the subject's group identity (undergraduate, graduate student, or professional).

Each separate paragraph or section (e.g., memory test, or ordering section on problem 2) was printed starting on a separate page, to minimize coder confusion.

2. Coding of the "Answering Problems" (First and Third Problems).

The first and third problems are coded with respect to the information people attended to, and the strategies they applied to that information in arriving at their answers. Two separate coding schemes are applied to the transcripts. The first focusses on the use of various kinds of information and strategies. The second focusses on the subject's interpretation of the conditional probability (reliability) information, and is described in Section 5 below.

2.1. Codes for problems 1 and 3.

A set of codes was developed to be applied to each paragraph (see Appendix 1). The codes are as follows. Instructions are found in sections 2.1.1, 2.2, and 2.3.

I. Coding of subjects' responses to the second paragraph (baserate).

1. Code group 1: Mentioning of base rate.
   a. mentions correct baserate number
   b. mentions incorrect baserate number
   c. mentions nonspecific baserate
   d. no mention of baserate
2. Code group 2: Use of baserate.
   a. uses baserate as the answer
   b. uses baserate in an adjustment process (coder notes the adjustment process)
   c. no use of baserate (coder notes the reason)
3. Code group 3: Strategies of interpolation or anchoring and adjustment.
   a. interpolation - no justification
   b. interpolation - evidence justification
   c. interpolation - reliability justification
   d. interpolation - other justification (coder notes justification)
   e. anchor and adjust - no justification
   f. anchor and adjust - evidence justification
   g. anchor and adjust - reliability justification
   h. anchor and adjust - other justification (coder notes justification)
   i. some other adjustment made (coder notes the adjustment process)
   j. no adjustment is made

   For interpolation, the coder notes the first number, second number, and the subject's new answer. For the first and second numbers, the coder notes whether each is
   a. the previous answer
   b. the baserate
   c. the evidence
   d. the subject's estimate of reliability or the problem's conditional probability
      i. the correct one from the problem
      ii. the complement of the correct one from the problem - this represents an error

   For anchoring and adjustment, the coder notes the anchor and the new answer, and notes whether the anchor is one of the same five categories as used with interpolation.

II. Coding of subject's responses to the third paragraph (evidence).

1. Mention of baserate (as in paragraph 2).
2. Original "use of baserate" codes have been replaced by code set 4.
3. Strategies (as in paragraph 2).
4. Subject's use of baserate information.
   a. uses baserate only - no justification
   b. uses baserate only - reliability justification
   c. uses baserate only - other justification (coder notes the justification)
   d. no use of baserate (coder lists reason)
   e. uses baserate in conjunction with other information (coder is required to use a code from set 3, strategies)
5. Subject's mentioning of evidence information.
   a. mentions correct evidence
   b. mentions incorrect evidence
   c. no mention of evidence
6. Subject's use of evidence information.
   a. uses evidence only
   b. uses evidence to adjust (coder is required to use a code from set 3, strategies)
   c. no use of evidence -- reliability justification
   d. no use of evidence -- other justification, or no justification (coder notes the justification)
7. Subject's mention of the reliability issue (specific information is not provided until the 4th paragraph).
   a. mentions reliability issue - goes beyond provided information
   b. mentions reliability issue - repeats information provided
   c. no mention of reliability

III. Coding of subject's responses to the fourth paragraph (conditional probability pertinent to the reliability of evidence).

1. Mention of baserate (as in paragraphs 2 and 3).
2. Original "use of baserate" codes have been replaced by code set 4.
3. Strategies (as in paragraphs 2 and 3).
4. Subject's use of baserate information (as in paragraph 3).
5. Subject's mentioning of evidence information (as in paragraph 3).
6. Subject's use of evidence information (as in paragraph 3).
7. Subject's mention of the reliability issue (code set 7) has been replaced with code set 8).
8. Subject's mention of the reliability issue.
   a. mentions correct reliability (conditional probability) number [this is the correct number that is in the problem; the coder makes no judgment of whether the logic is correct]. Coder was given a sheet that specified which numbers were "correct" and which "incorrect", for each problem. This information did not specify which kind of conditional probability was given. This is the sheet:

```
-----------------------------------------------------------------------

Doctor problem:

    baserate: 85% hepatitis, 15% toxic uremia
    evidence: toxic uremia
    reliability: 20% error rate, 80% right

Insurance problem:

    baserate: 65% will have accident, 35% will not
    evidence: no accident
    reliability:  25% lie, 75% tell truth

Twins problem:

    baserate: 80% Paul, 20% Stephen
    evidence: Stephen
    reliability: 60% correct, 40% wrong

The reliability information states the extent to which the evidence
will be correct or will be in error. The probabilities associated
with each are as follows:

                Correct   Error

  Doctor          80%      20%
  Insurance       75%      25%
  Twins           60%      40%

-----------------------------------------------------------------------
```

b. mentions incorrect reliability number

c. mentions nonspecific reliability

d. no mention of reliability

9. Use of the conditional probability information.
   a. uses "correct" reliability only (that is, does not mistakenly substitute one hypothesis for the other)

   b. uses "error" reliability only

   c. uses reliability to adjust (coder is required to use a code from code set 3, strategies)

   d. no use of reliability

10. Subject's use of Bayes' Theorem.
   a. uses Bayes' Theorem (Coder does not judge whether subject used the Theorem correctly)

   b. mentions Bayes' Theorem but does not use it

   c. no use of Bayes' Theorem

2.1.1. Instructions for coding Problems 1 and 3.

To help the coder use the above coding scheme, the following instructions were given:

Before beginning coding, write in the subject number and circle the problem number on the coding sheet. Read through each paragraph before beginning the coding for the paragraph.

In general, each number should receive a circle for one and only one letter (i.e., 1a or 1b or 1c receives a circle). However, if none of the letters within a number seem appropriate, put a question mark and note the problem. If more than 1 letter is circled, please write the reason why.

1. Code group 1. Code whether the subject mentions the correct or incorrect baserate number, mentions a nonspecific baserate (e.g. "its usually Paul"), or makes no mention of the baserate. The correct baserate can be about either of the entities (e.g. hepatitis, toxic uremia) or both.

2. Code group 2. Code the extent to which the subject makes use of baserate information. Circle "a" if the subject uses the baserate as the answer. Circle "b" if the baserate is used to adjust the answer following the first paragraph. If "b" is circled, please list how the adjustment was made. If no use is made of the baserate, circle "c". If a reason for not using the baserate was given, please list it.

3. Code group 3. "a", "b", "c", or "d" is circled when an interpolation process is used. "a" is circled when no justification is given. "b" is circled when evidence justification is used. "c" is circled when reliability is used as an explanation. If "c" is circled, there should be a circle for "7a" or "7b" if you are coding the second or third paragraph. "d" is circled when some other justification for interpolating is given. If "d" is circled, list what the justification was. When interpolation occurs, the two numerical values involved in the interpolation must be listed. In addition, the role that number plays in the problem must be noted by checking the appropriate box. Two boxes can be checked ONLY if one of them is the "previous" answer box. 3 "e", "f", "g" or, "h" is circled if an anchor and adjustment process occurs. "e" is circled when no justification is given. "f" is circled when an evidence justification is given. "g" is circled when reliability is used as an explanation. If "g" is circled, list what the justification was. If anchoring and adjustment occurs, the numerical value of the anchor and of the new answer must be listed. In addition, the role that the anchor plays in the problem must be noted by checking the appropriate box (as when interpolation occurs above). "i" is circled when some other adjustment is made. If "i" occurs, please list the nature of the other adjustment. "j" is circled when the answer is the same as after the previous paragraph. (Although they were considering some sort of adjustment, they did not actually change their answer.)

4. Code group 4. Code the extent to which the subject makes use of the baserate information. "a", "b", "c" are circled if the subject ignores the evidence from the third paragraph, and uses only the baserate information. "a" is circled when the subject uses the baserate as an answer without providing any reason for using it. "b" is circled when the subject uses the baserate as an answer and argues that the evidence may be unreliable. If "b" is circled, then "6c" and either "7a" or "7b" should also be circled. "c" is circled when the subject uses the baserate as an answer and gives some explanation other than reliability. If "c" is circled, list what the explanation was. "d" is circled when the baserate is not used in arriving at the answer. When "d" is circled, list the reason the baserate was ignored, if possible. "e" is circled when the baserate is used in conjunction with some other information. When "e" is circled, there should be something circled in part 3.

5. Code group 5. Code whether the subject mentions the correct or incorrect evidence or makes no mention of evidence.

6. Code group 6. Code the extent to which the subject makes use of the evidence. If the evidence is taken to establish that the hypothesis it supports is 100% true (i.e., it is given as the answer), circle "a". If the evidence is used to make an adjustment, circle "b". If "b" is circled, there should be something circled in part 3. "c" is circled when no use is made of the evidence because of a reliability explanation. If "c" is circled, "7a" or "7b" should also be circled. If the evidence has no impact on the answer and some other justification or no justification is given, circle "d". If some justification is provided, list it.

7. Code group 7. Code whether the subject mentions reliability (i.e., whether there is any doubt about the evidence) or not. If the subject does not mention reliability, circle "c". If the subject repeats information provided in the paragraph, circle "b". If the subject goes beyond the information provided in the paragraph, circle "a".

8. Code group 8. Code whether the subject mentions the correct or incorrect reliability number, mentions a nonspecific reliability (e.g. people usually tell the truth), or makes no mention of reliability.

9. Code group 9. Code the extent to which the subject makes use of the reliability information. If the reliability is given as the answer, circle "a" or "b" depending on whether the "correct" or "error" reliability information is used. If the reliability is used to make an adjustment, circle "c". If "c" is circled, there should be something circled in part 3. If the reliability is not used circle "d" and list the reason, if possible.

10. Code group 10. Code whether or not the subject mentions and/or uses Bayes' Theorem.

2.2. Guidelines for coding problems 1 and 3.

In addition to the above instructions, the following materials were used by coders in coding the first and third problems:

1. The most important rule is that if the coder is uncertain (1) that a process -- anchoring and adjusting or interpolation -- is occurring, (2) that the subject mentions the evidence or reliability, (3) that the subject is using the baserate, evidence, or reliability to adjust, and so forth, the coder should not indicate that these processes are occurring or are being mentioned.

2. The most difficulty in coding will occur when trying to answer number 3 on the coding sheet. Below are some guidelines to assist the coder when coding this section:
   a. Subject must mention something about "moving"; however, it is not necessary that the subject state the anchor number in the particular paragraph. i.e.: "I'm going to raise the likelihood that Stephen broke the lamp to 70% because the babysitter thought she saw Stephen by the lamp." In the previous paragragh (paragraph 2) the subject gave the baserate of 20% as the likelihood that Stephen broke the lamp. Thus, 20% would be the anchor number.

   b. Anchoring and adjusting/interpolation--no justification: This category is used when the subject mentions "moving", yet he/she does not give a reason for moving the answer. (This will probably be a rare event.)

   c. Anchoring and adjusting/interpolation--some other justification: This category describes adjustments that are made but are not based on evidence or reliability issues. For example, the subject may anchor and adjust as a result of baserate information.

   d. Some other adjustment is made: This category was designed primarily to include any processes that cannot be placed in any of the other categories--that is it is to be used when interpolation, anchoring and adjusting, and "no adjustment" do not occur. An example may be that the subject just picks a number, states that the number is a little low/high and adjusts the number accordingly; or the subject may adjust based upon information that was not given in the problem, such as events that occurred in his/her own life that may be related to the problem.

   e. No adjustment is made: This category is reserved for those instances when the subject comes up with the answer, but the coder is unable to determine what processes the subject used in arriving at his or her answer; the subject appears to pick a number without any specific reason and uses this number as the answer; or, the process did not involve adjustment or interpolation. The subject may, in fact, be increasing or decreasing his/her answer from the previous paragraph, but the coder is unable to determine the process that the subject used.

   f. If the subject "revises" an answer and the revised answer is the baserate or the reliability number, anchoring and adjusting/interpolation are not occurring.

3. When using Bayes' Theorem, the subject will usually state the baserate, the evidence and the reliability numbers. Circle that the subject mentions these items.

4. If the subject uses or attempts to use Bayes' Theorem, but does not give the correct answer, the coder should still note that the subject used Bayes' Theorem.
5. If the subject uses Bayes' Theorem, 3i (some other adjustment is made) should be circled.
6. In order to circle that the subject uses the evidence only as his/her answer, the subject must do just that--use only the evidence to arrive at his/her estimate. If the subjects estimate is 1.0 after receiving the the evidence and the subject does not mention an adjustment process, then the subject has used only the evidence. If after receiving the evidence, the subject gives a "high estimate", but it is less than 1.0, and again the subject did not mention an adjustment process, the subject has used only the evidence.

2.3. Training materials for coding Problems 1 and 3.

Coders had the following (Sections 2.3 to 2.3.3) material:

Subjects in Think Aloud Study #2 may try to combine two (or more) numbers together to get a number as an answer. There are a large number of possible strategies they may adopt. We are interested in these particular strategies.

2.3.1. Anchor and adjust.

In this strategy, the subject would start with one number, the "anchor", and adjust it upward or downward in response to another number or piece of information. This strategy can be applied to any two pieces of information, as long as the first one is a number. There is a sense in which one of them is "prior" to the other. In this study, it may frequently be that the anchor is the previous answer, and the other piece of information might be the base rate (a number) or the evidence (not a number). When you see that this strategy has been used, note specifically

1. what is the anchor?
    a. what is the number?
    b. what role does that number play in the word problem?
        i. Is it the previous answer?
        ii. Is it the baserate, 1.0, .50, the reliability (% right or % wrong)?
2. How is the adjustment justified? If a piece of information is referred to as the basis for the adjustment, what is it?

Sometimes it will not be possible to identify a strategy, even though the subject picks a number different from their previous answer. Clearly they have "adjusted" their answer, but we do not have evidence concerning an "anchor and adjust" strategy.

Examples of anchoring and adjusting: "Well, I said before that the probability was 40% that he would not have any accidents, but this indicates that he was a safe driver before, so it is probably 55% that he won't have one." Anchor on previous answer, adjust up because of evidence.

"The reliability of the test is 90%, but I already thought he had it, so I'll say 95%." Anchor on reliability, adjust up because of prior belief (evidence).

"Well, the test says he has toxic uremia, but I don't quite believe the test so I'll say 98%". Anchor on 1.0 (note it was not explicitly stated), adjust down because of concerns about reliability.

2.3.2. Interpolate.

In this strategy, the subject would start with two numbers, and use the third piece of information to guide the selection of an answer that is between the other two numbers. Where "anchoring and adjusting" has one anchor, this has two anchors. In this study, we might find people interpolating between the baserate and the reliability. When you see the subject using this strategy, note specifically:

1. What are the two endpoints? For each,
   a. what is the number?
   b. what role does that number play in the word problem?
      i. Is it the previous answer?
      ii. Is it the baserate, 1.0, .50, the reliability (% right or % wrong)?
2. How is the interpolation justified? If a piece of information is taken into account, to determine the answer's exact position between the endpoints, what is it?

Example of interpolation: "Although Stephen is at fault only 20% of the time, the sitter saw him, which indicates a probability of 1.0 that Stephen did it. So I'm going to give an answer in between, 60%."

Difficult example of interpolation: "The mother said that 80% of the time Paul was the trouble maker, but the babysitter thought it was Stephen, so I'm going to say that there is a 40% probability it was Stephen." This is difficult because of the switch from talking of Paul (80%) to talking of Stephen: he would have 20% chance when Paul has 80% chance.

2.3.3. Bayes' Theorem.

One way of integrating between the baserate number and 1.0 is to apply Bayes' Theorem. Some of our subjects may try to do this. The answer the subjects are asked to produce is p(H/E), the probability that hypothesis H is true, given that evidence E has been observed. The base rate is p(H), the probability that hypothesis H is true, when one does not have any evidence for or against it. The reliability is p(E/H). The symbol ~ means "not", so p(~H) means the probability that hypothesis H is not true, in the absence of evidence for or against it. Bayes' theorem says that:

$$p(H/E) = \frac{p(E/H) \times p(H)}{p(E/H) \times p(H) + p(E/\sim H) \times p(\sim H)}$$

If the evidence is completely reliable (if p(E/H) = 1 and p(E/~H) = 0) then p(H/E) = 1.0. If the evidence is completely unreliable (if p(E/H) = .5, in this case with only two hypotheses) then p(H/E) = p(H), the base rate. Thus, Bayes' Theorem is a way of interpolating between the baserate and 1.0. If it occurs, it ought to be coded as both "Bayes' Theorem" and as "interpolation". If the subjects make a mistake in applying Bayes' Theorem, but it is still obvious that this is what they are trying to do, call it "Bayes' Theorem" anyway. If the subjects mention that they could apply Bayes' Theorem, but do not actually apply it [e.g., "I think I should apply Bayes' Theorem here, but I don't remember it"], code that as "mentions Bayes' Theorem".

3. Coding of the "Ordering Problem" (Second Problem).

The answering of the second problem was not coded for use of baserate, evidence, and strategies. The reason is that since the subject determined the order in which information was received, it would be very complicated to keep track of what information was available.

3.1. Coding scheme guidelines.

The main interest here is the conditional probabilities, paragraphs 3 and 4. A coding scheme was made using the following guidelines:

1. Expression of confusion between the two. Does the subject say "but isn't that the same?" or "what is the difference?" or something like that, when they read paragraph 4?
2. How is the confusion resolved? Do they end up stating that they understand the difference, or do they end up deciding that they are, in effect, the same thing?
3. Then, for each of 4 locations, a "reliability interpretation count", as described below (Section 6).
   a. Ordering discussion, paragraph 3.

b. Ordering discussion, paragraph 4.

c. Answering, paragraph 3.

d. Answering, paragraph 4.

Note that it was difficult to decide whether part of the ordering discussions is a "discussion of paragraph 3" in contrast with "discussion of paragraph 4." To indicate this on the coding sheet, coders wrote "CDP" for "couldn't differentiate paragraphs", and did not otherwise code the passage..

4. What did they say they were going to do with the information? Why did they want the information for paragraph 3? for paragraph 4?

5. And, why did they say they wanted paragraph 3 before/after paragraph 4, if they addressed this explicitly?

3.2. The coding scheme for the second problem.

The second problem was coded by Coder #1 on just the discussion of the apparently redundant information in the two conditional probability paragraphs. This was coded separately for when the subject was ordering the paragraphs, and when he or she was answering the question after the blanks in the paragraphs had been filled in with information.

1. Level 0: The subject does not spontaneously mention any comparisons between the information in the two paragraphs.

2. Level 1: The subject expresses some confusion or differentiation regarding the information in the two paragraphs, but does not resolve the differences or the confusion.

3. Level 2: The subject expresses some confusion or differentiation regarding the information in the two paragraphs, and concludes that the information in the two paragraphs is the same.

4. Level 3: The subject expresses some confusion or differentiation regarding the information in the two paragraphs, and concludes that the information in the two paragraphs is different.

The instructions given to the coder were:

The goal of this coding scheme is to capture the subject's differentiation or lack therof regarding the reliability information in paragraphs 3 and 4 of Problem 2. The coding will take place at two points: 1) when the subject is ordering the paragraphs regarding which paragraphs s/he wants first, second, etc., and 2) when the subject is reading the information in paragraphs 3 and 4 and making probability judgments.

At each of these points, one of four levels of differentiation between the information in paragraphs 3 and four will be coded. (Lists the definitions of the levels, above.) No special coding sheet is required for this coding. The coding can be done on regular lined paper, using one line for each subject. The column headings will be as follows:

```
Subject    Ordering   At judgment
```

In addition, the coder noted: "If subject says information in paragraphs 3 and 4 is irrelevant, or it doesn't matter which paragraph is taken before the other, I coded it as Level 1."

4. Coding of memory test.

The subjects were asked to reconstruct the whole first problem, i.e., to recall all the information that they were given, and to say what answer they gave, and how they got it.

4.1. Guidelines for construction of memory test coding scheme.

A coding scheme was made, using these guidelines:

1. what information they remembered:
   a. base rate (did they mention the idea or not? did they remember the numerical values correctly, incorrectly, or not at all?)
   b. evidence (did they remember what was observed correctly, incorrectly, or not at all?)
   c. reliability (did they mention it or not? did they remember the numerical values correctly, incorrectly, or not at all?)
2. How they interpreted the reliability information, if they remembered it (using the reliability coding scheme).
3. Whether they remembered their final answer correctly, or incorrectly (reporting an earlier answer, that appeared on the answer sheet after paragraph 1, 2, or 3), or incorrectly (reporting some other number), or not at all.

4.2. The memory test coding scheme.

The memory coding instructions were:

After completion of 3 problems, the subjects were asked to remember the information in their first problem. Specifically, they were given 4 probes:

1. to remember everything they could about the problem
2. to remember the information pertinent to answering the question in the problem
3. to remember their final answer
4. to remember how they arrived at their answer.

In response to each of the four probes above, the subject could remember any of the four pieces of information of interest:

1. baserate
2. evidence
3. reliability
4. their final answer.

The goal of the memory coding is to determine, after each of the probes, whether or not the subject mentioned any of the 4 pieces of information and whether they remembered the correct information.

Specifically, the coding procedure is as follows (refer to coding sheet). After each probe, read the subject's responses and determine whether:

1. they mention any of the four concepts listed on the coding sheet.
   a. If so, put a 1 in the appropriate column.
   b. If not, put a 0.
2. they mention a numeric value corresponding to each concept.
   a. If the subject mentions a correct numeric value, check the appropriate column.
   b. If the subject mentions an incorrect numeric value, list that value in the appropriate column.

The coding sheet had three columns,

Mention Concept	Give Correct Value	List Incorrect Value

And the row headings were (A) questions 1 to 4, (B) baserate, evidence, reliability, and final answer for each question.

The coder noted that in response to the first question, to remember everything about the problem, most subjects restated the question (what is the probability of the hypothesis being true). Because the subjects did not specifically mention the baserate, evidence, reliability, or final answer, she coded that they did not mention these concepts.

5. Coding of use of conditional probability information.

A set of codes were used in order to identify the exact form of any reliability idea that the subject uses, i.e., how the subject interprets the reliability information. If s/he states the idea (restates it; not just "uses it") more than once, then an additional "count" should be put in the idea. If s/he states a different idea, than that one should be counted. This is done at a different time than the main analysis of problems 1 and 3. The following locations in the transcript were coded:

1. The evidence paragraph (third paragraph) of problems 1 and 3.
2. The reliability paragraph (fourth paragraph) of problems 1 and 3.
3. The ordering of paragraph 3, in problem 2. (Whenever they talk about paragraph 3, in the ordering section.)
4. The ordering of paragraph 4, in problem 2.
5. The answering of problem 2, after they have received the specific information for paragraph #3 (no matter in which ordinal position they requested this information to be given).
6. The answering of problem 2, after they have received the specific information for paragraph #4 (no matter in which ordinal position they requested this information to be given).
7. The memory test - all mentions of reliability. (Only one sheet will be used for the memory test, no matter how often reliability is mentioned.)

Thus there are 9 locations where a "reliability coding" will be done.

The following table gives a quick guide to the identification of the concepts of evidence "e" and hypothesis "h", and their opposites, in each problem.

	Twins	Doctor	Insurance
h	was Stephen	is toxic uremia	have good past driving record
~h	was Paul	is hepatitis	have bad past driving record
e	say Stephen	test says t.u.	says he has a good past driving record
~e	say Paul	test says hep.	admit to a bad past driving record

At each location for which reliability interpretation coding and counting is required, identify each statement of the reliability concepts as one of the following interpretations. If more than one interpretation is stated, mark each. If an interpretation is stated more than once, put a mark for each statement, so that we have a count of how many times they used the idea. Note that the subjects will probably be loose in specifying who they are talking about. They do not have to use the same words specified below; just the same ideas. E.g., for the Twins problem, it might be "someone" or "the baby sitter" who has trouble identifying the twins.

The coders used the information on the following pages as guides in identifying each reliability concept.

Reliability interpretation coding scheme.
Doctor Problem.

At each location for which reliability interpretation coding and counting is required, identify each statement of the reliability concepts as one of the following interpretations. If more than one interpretation is stated, mark each. If an interpretation is stated more than once, put a mark for each time, so that we have a count of how many times they used the idea.

1. p(h/e). Doctor: probability of having toxic uremia given Spock test says it was toxic uremia.
2. p(h/~e). Doctor: probability that the disease is toxic uremia given that the test says it is hepatitis.
3. p(~h/e). Doctor: probability that the disease is hepatitis given that the Spock test says it is toxic uremia.
4. p(~h/~e). Doctor: probability of having hepatitis given that the Spock test says he has hepatitis.
5. p(e/h). Doctor: probability that the test will say the patient has toxic uremia given that the patient has toxic uremia.
6. p(e/~h). Doctor: probability of the test saying the patient has toxic uremia given that the patient has hepatitis.
7. p(~e/h). Doctor: probability of the test saying the patient has hepatitis given that the patient has toxic uremia.
8. p(~e/~h). Doctor: probability of the test saying someone has hepatitis given that they do have hepatitis.
9. p(correct evidence), i.e., p(e/h) or p(~e/~h). Doctor: probability of the test correctly identifying the disease.
10. p(wrong evidence), i.e., p(e/~h) or p(~e/h). Doctor: probability of the test being wrong, saying the person has the wrong disease.
11. p(correct inference), i.e., p(h/e) or p(~h/~e). Doctor: probability of the patient having the disease that the Spock test says he has.
12. p(wrong inference), i.e., p(h/~e) or p(~h/e). Doctor: probability of the patient having the other disease from what the test said they had.
13. p(h and e) [or p(e and h)]. Probability of having toxic uremia and having a "toxic uremia" result on the Spock test.
14. p(h and ~e) [or p(~e and h)]. Probability of having toxic uremia but having a "hepatitis" result on the Spock test.
15. p(~h and e) [or p(e and ~h)]. Probability of having hepatitis but having a "toxic uremia" result on the Spock test.
16. p(~h and ~e) or [p(~e and ~h)]. Probability of having hepatitis and having a "hepatitis" result on the Spock test.
17. They discuss the concept of reliability but it is difficult to understand which of the above concepts they were using.
18. They did not mention any reliability or conditional probability concept.

Reliability interpretation coding scheme. Insurance Problem.

At each location for which reliability interpretation coding and counting is required, identify each statement of the reliability concepts as one of the following interpretations. If more than one interpretation is stated, mark each. If an interpretation is stated more than once, put a mark for each time, so that we have a count of how many times they used the idea.

1. p(h/e). Insurance/termination: probability the client had a good past driving record given he says he had a good past driving record.
2. p(h/~e). Insurance/termination: probability that the client had a good past driving record given that he admits to a bad past driving record.
3. p(~h/e). Insurance/termination: probability that the client had a bad past driving record given that he said on the phone that he had a good past driving record.
4. p(~h/~e). Insurance/termination: probability that the client had a bad past driving record given that he said on the phone that he had a bad driving record.
5. p(e/h). Insurance/termination: probability a client will say he has a good past driving record given that he or she has a good past driving record.
6. p(e/~h). Insurance/termination: probability of a client saying he or she has a good past driving record given that he or she has a bad past driving record.
7. p(~e/h). Insurance/termination: probability of someone admitting to having a bad past driving record given that they had a good past driving record.
8. p(~e/~h). Insurance/termination: probability of someone admitting to a bad past driving record given that they have a bad past driving record.
9. p(correct evidence), i.e., p(e/h) or p(~e/~h). Insurance/termination: probability of someone truthfully reporting on their driving record.
10. p(wrong evidence), i.e., p(e/~h) or p(~e/h). Insurance/termination: probability of someone mistakenly reporting, or lying, about their driving record (Note that the motivations for these two would be different).
11. p(correct inference), i.e., p(h/e) or p(~h/~e). Insurance/termination: probability that the person had the kind of driving record that they said they had.
12. p(wrong inference), i.e., p(h/~e) or p(~h/e). Insurance/termination: probability of someone having a different driving record from what they report they had.
13. p(h and e) [or p(e and h)]. Probability that a person would both have a good driving record and say that they had a good driving record.
14. p(h and ~e) [or p(~e and h)]. Probability that a person would have a good driving record but say that they had a bad driving record.
15. p(~h and e) [or p(e and ~h)]. Probability that a person would have a bad driving record but say that they had a good driving record.
16. p(~h and ~e) or [p(~e and ~h)]. Probability that a person would both have a bad driving record and say that they had a bad driving record.
17. They discuss the concept of reliability but it is difficult to understand which of the above concepts they were using.
18. They did not mention any reliability or conditional probability concept.

Reliability interpretation coding scheme.
Twins Problem.

At each location for which reliability interpretation coding and counting is required, identify each statement of the reliability concepts as one of the following interpretations. If more than one interpretation is stated, mark each. If an interpretation is stated more than once, put a mark for each time, so that we have a count of how many times they used the idea.

1. p(h/e). Twins: probability that the twin who broke the lamp was Stephen given that the sitter thought she saw Stephen.
2. p(h/~e). Twins: probability that Stephen was the twin who broke the lamp given that the sitter thought she saw Paul.
3. p(~h/e). Twins: probability that Paul was the twin who broke the lamp given that the babysitter thought she saw Stephen.
4. p(~h/~e). Twins: probability that Paul was the twin who broke the lamp given that the babysitter thought she saw Paul.
5. p(e/h). Twins: probability of someone saying they saw Stephen given that it was Stephen.
6. p(e/~h). Twins: probability of someone thinking they see Stephen given that it was really Paul.
7. p(~e/h). Twins: probability of someone thinking that they saw Paul given that it was really Stephen.
8. p(~e/~h). Twins: probability of someone thinking they saw Paul given that it really was Paul.
9. p(correct evidence), i.e., p(e/h) or p(~e/~h). Twins: probability of someone accurately identifying a twin.
10. p(wrong evidence), i.e., p(e/~h) or p(~e/h). Twins: probability of someone misidentifying one of the twins.
11. p(correct inference), i.e., p(h/e) or p(~h/~e). Twins: probability that the twin really is the one that the babysitter thought she saw.
12. p(wrong inference), i.e., p(h/~e) or p(~h/e). Twins: probability of it being the other twin from the one the babysitter thought she saw.
13. p(h and e) [or p(e and h)]. Probability that the babysitter would have thought Stephen broke the lamp and Stephen would have actually been the one that broke the lamp.
14. p(h and ~e) [or p(~e and h)]. Probability that the babysitter would have thought Paul broke the lamp but Stephen would have actually been the one that broke the lamp.
15. p(~h and e) [or p(e and ~h)]. Probability that the babysitter would have thought Stephen broke the lamp but Paul would have actually been the one that broke the lamp.
16. p(~h and ~e) or [p(~e and ~h)]. Probability that the babysitter would have thought Paul broke the lamp and Paul would have actually been the one that broke the lamp.
17. They discuss the concept of reliability but it is difficult to understand which of the above concepts they were using.
18. They did not mention any reliability or conditional probability concept.

Reliability interpretation coding scheme answer sheet.

Subject Number_____ Coder _____ Date _____

At each location for which reliability interpretation coding and counting is required, identify each statement of the reliability concepts as one of the following interpretations. If more than one interpretation is stated, mark each. If an interpretation is stated more than once, put a mark for each time, so that we have a count of how many times they used the idea.

	Problem 1		Problem 2				Problem 3		Memory
	Par 3	Par 4	Ordering Par 3	Ordering Par 4	Answering Par 3	Answering Par 4	Par 3	Par 4	Test
p(h/e).									
p(h/~e).									
p(~h/e).									
p(~h/~e).									
p(e/h).									
p(e/~h).									
p(~e/h).									
p(~e/~h).									
p(correct evidence)									
p(wrong evidence)									
p(correct inference)									
p(wrong inference)									
p(h & e)									
p(h & ~e)									
p(~h & e)									
p(~h & ~e)									
Ambiguous.									
No mention.									

5.1. Training materials for reliability coding.

The following materials (Sections 6.1.1 to 6.1.5) were produced to give coders guidance in applying these codes.

5.1.1. Types of ambiguity.

There are several kinds of ambiguity, and the code "ambiguous" is used for just one particular kind.

1. In problems 1 and 3, the only ambiguity is that it is hard to tell which category of conditional probability a sentence is. For example, the subjects may clearly be talking about a conditional probability concept, but their words may match more than one of our categories -- or none at all.
2. In problem 2, in addition to difficulty of identifying the category, there is also difficulty in deciding whether the sentence is about paragraph 3 or paragraph 4. If it is difficult to tell which paragraph is being referred to, one should simply not count it. Only if a thought is clearly about the information that came in paragraph 3, or about the information that came in paragraph 4, should it be coded. Since we are contrasting the use made of these two paragraphs, any confusion between them only blurs our contrast and should be avoided.

When should we call it "ambiguous", and when should we guess one of the two concepts that it is ambiguous between? Sometimes a discussion of a conditional probability is simply hard to understand or interpret. If so, call it "ambiguous". But sometimes the ambiguity is that there are two possible interpretations, and it is hard to tell which one. Guidelines for calling it "ambiguous", or guessing, differ according to what the possible interpretations are.

The competing interpretations may both be of the same type. For example: A subject, on problem 2 (twins), is looking at paragraph #4 before having paragraph #3.

> Ok, so two times out of five, when you think it's Stephen, it's really Paul. So if she thought it was Stephen, and so, two times out of five, when she thinks it is Stephen, it's really Paul, oh boy. So... *she's a little bit better than half right, so she has a better chance, better than 50% chance of being right, so.. but it's still usually Paul.* Um, I'm going to say... um, I'll say uh, 50% I guess. Back to 50.

The italicized text speaks of a general chance of being right. One might think it is p(correct inference), i.e., referring equally to p(h/e) and to p(~h/~e). On the other hand, the subject knows full well that the evidence is that it is Stephen (the babysitter thought she saw Stephen), so the coder might view this as simply p(h/e). Hamm prefers the more general p(correct inference) here. But it is an ambiguous choice between two levels of generality **of inference**. Hamm would prefer that the coder choose one of the two, rather than call it "ambiguous", because that "ambiguous" category is a discard category.

The competing interpretations may be of different types, [evidence and inference]. On the other hand, if the ambiguity is between an "inference" interpretation and an "evidence" interpretation, then it is better to call it "ambiguous" than to make a mistake and add noise to the data.

5.1.2. Lying.

A superficial analysis of the word "lying" (which occurs with the Insurance problem) indicates that it is a case of purposefully wrong evidence, p(e/~h) or p(~e/h). The person has a true driving record, and chooses to misrepresent it. But it is not so simple. "Lying" can occur in both the p(e/h) [conditional evidence] and p(h/e) [conditional inference] contexts. [These are called "conditional" evidence, because in this study we also have "evidence", which is the report of the observation (Twins: babysitter saw Stephen; Doctor: Spock tests says "toxic uremia"; Insurance: Kenneth Peterson says he had no accidents for the previous 15 years).] Here is the distinction:

1. **"Lying" as p(e/~h).** If the idea that the subject is referring to is the chance that someone lied, that is conditioned on who they are: a person with a particular accident record, for example. Given who they are, do they report it accurately?

2. **"Lying" as p(~h/e).** Sometimes, the subject is starting with a particular piece of evidence, e.g., Kenneth Peterson's report that he had a good past driving record, and then asking about the possibility that the reality was otherwise. This is p(~h/e). But the subject may phrase it as "what is the chance he lied". Given that he said he had no accidents, what is the chance that he did have accidents (and must therefore have lied about them)?

5.1.3. "Accurate", "right", "error".

The "conditional probability" paragraphs **all** have general phrases concerning accuracy. These were intended to be ambiguous between p(e/h) and p(h/e). When you see these phrases, you should realize that:

1. they were intended as ambiguous
2. every subject saw them in the 4th paragraph of Problems 1 and 3, and in both the 3rd and 4th paragraphs of Problem 2.

These ambiguous, general phrases are:

1. **Doctor problem.** "The Spock test is not perfect. It has an error rate of 20%, and is right 80% of the time." The first sentence occurs in all problem, but the second sentence occurs only in Problems 1 and 3. This can be interpreted as, that the error is an error of evidence (given the disease that the subject really has), or an error of inference (given the test result, the inference to the disease may be in error).
2. **Insurance problem.** "Miss Shelton realizes that Mr. Peterson might have deceived her; however, she is also aware that when people are asked a surprising question, they usually answer truthfully." The "ambiguity" is weaker here. Rather than it being possible to read each of the sentences in both ways, the first sentence is strongly p(~h/e), and the second is strongly p(e/h). Every subject got this introductory section, which mentions both conditional probabilities, and then got a specification (with numbers) that was explicitly one or the other. Coders can then interpret "deception" as p(~h/e) and "answer truthfully" as p(e/h) or p(correct evidence), rather than being forced to call them "ambiguous". [But see the discussion of "lying", above.]
3. **Twins problem.** "New people have trouble telling the boys apart. They only identify them correctly 60% of the time." People often rephrase this as "distinguish" the boys. This can be interpreted as an error of evidence (given that it is a particular boy, people sometimes think it is the other boy) or an error of inference (given that people think the boy is a particular boy, sometimes it is really the other boy).

This puts us in a hard spot:

1. On the one hand, these phrases were intended to be ambiguous.
2. On the other hand, common usage of them is more like "p(e/h)" than "p(h/e)" [except for "deception" in the insurance problem]. If you look at the words, e.g., "tell them apart", it seems like "them" is more fundamental, is given, and "tell" follows, is conditional on the "them" which is given.

For our coding, we need to stick with the intention of the study, and acknowledge that these phrases were present in every case, and were intended to be ambiguous. So if the phrase is simply used without elaboration, mark it as ambiguous. However, if from context you can tell that one of the interpretations was being intended, code it as that. [Additionally, for the insurance problem, see the considerations above.]

5.1.4. The distinction between "answering the question" and "conditional p(h/e)".

The subjects are always asked to estimate the probability that some proposition h is true, p(h). If they have already read the "evidence" paragraph, then they know e, and they could view the probability they are asked to estimate as p(h/e). However, we do not state it for them in that way: we do not ask "what is the probability of h, given e?"; rather, we ask "what is the probability of h, given what you know now?"

When we give the "conditional probability" information, in the context of a paragraph discussing the reliability of the evidence, we sometimes phrase it as p(e/h) [which is technically the reliability], and sometimes as p(h/e) [which is technically the answer to their question, rather than just some information

which is pertinent to discovering that answer].

Coders who are judging the form of the subject's understanding of the "reliability" or conditional probability information need to be aware that sometimes, very rarely, the subject is stating his/her goal [the question he/she is trying to answer] as p(h/e). This is his/her conception of his/her goal, and **not** his/her conception of the conditional probability information that we have given him. One place this occurs is with a subject who applied Bayes' Theorem, and explicitly stated that p(h/e) = p(e/h) times p(h), all divided by (a complicated term). In this context, the p(h/e) should not be coded as the subject's interpretation of the conditional probability information we have given him/her. Rather, it should not be coded, for it is something else. We do not know how often this occurs, but keep an eye out for it. Sometimes, during the course of a process that develops into an application of Bayes' Theorem, the p(h/e) concept will be articulated without clearly being "the answer". In these cases, its occurrence should be counted.

5.1.5. Meta-negatives and *Vice-Versas.*

Sometimes people reject a possible interpretation. For example, "It doesn't say" p(~e/h). If we coded that as "p(~e/h)", we would fail to give the subject credit. Worse, we would probably count it as an instance of a confusion of the meanings of the conditional probabilities when in fact the subject was distinguishing them.

Since there is no simple way to determine what conditional probability one would mean if one negated p(~e/h), these statements should simply not be counted as anything. [But if it is the **only** mention of conditional probability, do not check "no mention", even though you will not be checking any categories.]

A similar dilemma occurs when subjects describe one concept, e.g., their understanding of paragraph 3 in problem 2, and then say that the meaning of the other paragraph is the opposite (or *vice versa*). These must be counted as ambiguous, since we do not know exactly what transformation the subject had in mind.

6. Reliability of Coding.

In this and the next section we present information on the reliability of the coding. Section 6 presents the reliability of Coder #1, who coded strategies and the use of information. Section 7 presents the reliability of Coder #3, who coded conditional probabilities.

6.1. Reliability of the coding of strategies and of the use of information.

6.1.1. Reliability of coding of problems 1 and 3.

Four randomly chosen subjects, #s 2, 3, 8, and 17 (10% of the total N), were coded by Coder #2 as well as by Coder #1 (who coded all subjects). The coders' agreements and disagreements are counted in the following table, for each paragraph coded (paragraphs 2, 3, and 4) for each problem. The table counts the number of Code Groups (see Section 2.1 and Appendix 1) on which the two coders agreed (A) and disagreed (DA) about what is circled, and what is not circled. That is, if one coder felt that none of the categories in Class 3 applied, while the other checked 3g, it would be counted as a disagreement.

Subject	2		3		8		17		Overall	
	A	DA	A	DA	A	DA	A	DA	A	DA
Problem 1										
Paragraph 2	3	0	3	0	3	0	3	0	12	0
Paragraph 3	6	0	6	0	6	0	4	2	22	2
Paragraph 4	8	0	5	3	8	0	7	1	28	4
Problem 3										
Paragraph 2	3	0	3	0	3	0	3	0	12	0
Paragraph 3	5	1	6	0	6	0	4	2	21	3
Paragraph 4	8	0	7	1	8	0	7	1	30	2
Total	33	1	30	4	34	0	28	6	125	11
% Agreement	97		88		100		82		92	

6.1.2. Reliability of coding of Problem 2.

Coder #2 coded 4 subjects and agreed 100% with Coder #1, on both categorization schemes they used in Problem 2.

6.1.3. Reliability of coding of memory test.

The coding of the memory test involved checking each of four locations (after probe questions 1 to 4) for the mention of the baserate, evidence, reliability, and final answer. The coder had to decide whether the concept had been mentioned (1) or not (0). Coder #2 did four subjects, so there are 16 possible items of agreement between Coder #2 and Coder #1. Their results are in the following table:

Baserate

	Coder #1: 0	Coder #1: 1
Coder #2: 0	9	1
Coder #2: 1	0	6

$X^2 = 12.34$, df = 1, $p < .001$
Marginal $X^2 = 0.13$, df = 1, NS
$\pi = .870$
$\kappa = .871$

Evidence

	Coder #1: 0	Coder #1: 1
Coder #2: 0	8	0
Coder #2: 1	3	5

$X^2 = 7.27$, df = 1, $p < .01$
Marginal $X^2 = 1.17$, df = 1, NS
$\pi = .611$
$\kappa = .625$

Reliability

	Coder #1: 0	Coder #1: 1
Coder #2: 0	11	0
Coder #2: 1	0	5

$X^2 = 16$, df = 1, $p < .001$
Marginal $X^2 = 0.0$, df = 1, NS
$\pi = 1.0$
$\kappa = 1.0$

Final Answer

	Coder #1: 0	Coder #1: 1
Coder #2: 0	11	0
Coder #2: 1	0	5

$X^2 = 16$, df = 1, $p < .001$
Marginal $X^2 = 0.0$, df = 1, NS
$\pi = 1.0$
$\kappa = 1.0$

6.2. Reliability of coding of conditional probabilities.

Conditional probability statements in the transcripts for 10 subjects were coded by Coder #3, with duplication by Coder #4 so that the reliability of this coding could be assessed. This coding differed from the other coding schemes carried out by Coder #1 and Coder #2, because here individual statements were categorized, while in the other codings, all verbalizations following the presentation of a given paragraph of information were coded.

Thus, the definition of the objects to be coded was not at all a problem for Coder #1 and Coder #2, but was an important issue for the conditional probability codes.

The conditional probability coders were instructed to identify the sentences, categorize them, and count the number of occurrences of each category, for each locus. Thus, the data that will be used in the later analyses are counts of category occurrences in each location (*locus*, juncture, paragraph).

But the most appropriate way of measuring the reliability of this categorization is to focus on the sentences or phrases that were the units of judgment.

We will do both. In the next subsection, we present the sentence by sentence analysis of the reliability of coding of conditional probabilities. In the following subsection, we present the reliability analysis on the basis of their locations.

6.3. Reliability of conditional probability coding, assessed on a passage by passage basis.

Coder #4 and Coder #3 both coded 10 subjects. For every phrase or passage that any one of them coded, over all 10 subjects, we identified the codings from each coder. If only one of them coded the phrase, the other was categorized as "no mention"; or, if the other coder was confused with respect to whether the sentence referred to paragraph 3 or 4 on problem 2, as "cdp" for "couldn't determine paragraph".

We want to look at the κ and π indices for these (see Appendix II). There are two ways that we have reduced the complexity of the categorizations:

1. Reduction of the number of categories:
   a. Collapsing specifics [e.g., p(e/~h), p(~e/h) and p(wrong evidence)] into a general category [e.g., wrong evidence].
   b. Collapsing even further, e.g., into "evidence", whether it is wrong or right.
2. Stripping off all the sentences where there was no possibility of real disagreement or agreement because the coders did not both code it.
   a. ignoring combinations of ambiguous and not mentioned
   b. ignring combinations of non mentioned and a "real" category
   c. ignoring combinations of ambiguous and a real category

   This leaves just instances that both coders considered to be "real" categories.

In addition, the data are aggregated across

1. all 10 subjects
2. all 3 problems plus memory test
3. all paragraphs/*loci* within problem.

We can look at problems separately, or at paragraphs within problems. It would take some additional effort to look at paragraphs over problems (1 and 3) separately. This would, however, be appropriate.

We want to report analyses on each of these levels. First, let us look at the tables collapsed across all paragraphs and all problems.

The fullest results are in the table of the categories that were actually used in the coding There were 19 categories (coders used 18, but we break the "no mentions" into two categories), but neither coder used two of the categories, so we are down to 17. Of these, one coder used 14 and the other 15, so there were some all-0 rows and columns.

The overall X^2 for this table is 1072, df = 182, $p < .0001$. But this does not guarantee us that the non-predictability of the counts (from the marginals) is due to the agreement between the coders; it could be due to them consistently having different ideas about the same sort of sentence.

Testing the marginals, as recommended by Zwick (1988), shows that the two coders distributed their answers differently ($X^2 = 112.2$, df = 16, $p < .0001$. [To be accurate: Zwick (1988) recommends tesing the marginals. We do so using a simple X^2; she advocates a procedure that we did not bother to program. She did not specifically discuss our procedure, and why it was not sufficient.] Inspection shows that Coder #3 used "ambiguous" much more often (109) than Coder #4 (24; only 13 in common). Other differences are: Coder #4 did not mention a lot more sentences than Coder #3 ignored; Coder #4 used "p(correct evidence)" more than twice as often (23 to 10); Coder #4 used p(e/~h) slightly more often (32 to 24); Coder #4 did not use any of the "and" codes, while Coder #3 used 3 of them for a total of 6 sentences; Coder #4 used p(h/e) 31 times to Coder #3's 10; Coder #4 used p(wrong evidence) much more often (15) than Coder #3 (3); Coder #3 used p(~h/e) more often (24) than Coder #4 (14), though they agreed frequently (10 of Coder #4's 14).

Because the marginals have significantly different patterns, Zwick would not trust the π or κ, particularly the π. Nonetheless, we report them: $\pi = .190$ and $\kappa = .230$. That is, only about 20% of the possible agreement, above what would be expected by chance, was attained between the two coders when their overall coding is looked at.

Many of the disagreements between the coders involve sentences that one person counted, the other one not. In order to get a better conception of the agreement when both coders agreed that something was really happening, we can eliminate some sentences from the table. There were, out of 241 sentences:

1. 82 sentences where both coders said they were not easy to categorize, of which:
   a. 13 sentences were considered by both to be ambigous
   b. 9 **paragraphs**/*loci* were considered by both to be "no mentions". This is not so much a judgment of a sentence, but of a whole paragraph. If our data say that both coders used "no mention", it is because both checked it on their coding sheets. If our data say that just one coder coded "no mention", it is because the other coder said something about the sentence, but this coder did not, so we added "no mention" at the time of analysis.
   c. 7 sentences were considered by both to be "can't differentiate the paragraphs" on problem 2.
   d. and 53 were given <u>different</u> categories among these three.

   All these can be excluded from our discussion of how reliable the coders were with respect to the important categories, i.e., evidence versus inference.
2. 88 sentences on which one coder thought it possible to code it, while the other did not:
   a. 54 where the other thought the sentence "ambiguous"
   b. 26 where the other did not code the sentence ('no mention')
   c. 8 where the other could not tell whether the sentence came from paragraph 3 or 4 of Problem 2.
3. We will exclude these from the discussion of how reliable the coders are. This is perhaps a little questionnable.
4. The remaining 71 sentences were considered by both coders, to be codable.
   a. They agreed, giving the sentences exactly the same category, on 47 sentences
   b. They disagreed on 24 sentences.

The table is given here:

Coder #3's Coding

Coder #4's Coding	Evidence: The Probability of e/h	e/~h	~e/h	ce	we	Inference: The Probability of h/e	h/~e	~h/e	ci	Ambiguous: The Probability of h&e	h&~e	~h&e	
Evidence													
P(e/h)	1	0	0	0	0	1	0	0	0	2	0	0	4
P(e/~h)	0	16	1	1	1	0	0	5	0	0	0	0	24
P(~e/h)	0	0	8	0	0	0	1	0	0	0	0	0	9
P(~e/~h)	0	0	0	0	0	0	0	0	0	0	0	1	1
P(correct evidence)	0	0	0	4	0	1	0	0	0	0	0	0	5
P(wrong evidence)	0	1	0	0	2	1	0	0	0	0	0	0	4
Inference													
P(h/e)	1	0	0	1	0	6	0	2	0	1	0	0	11
P(h/~e)	0	0	0	0	0	0	0	0	0	0	1	0	1
P(~h/e)	0	1	0	0	0	0	0	10	0	0	0	0	11
P(correct inference)	0	0	0	1	0	0	0	0	0	0	0	0	1
Ambiguous													
P(h&e)	0	0	0	0	0	0	0	0	0	0	0	0	0
P(h&~e)	0	0	0	0	0	0	0	0	0	0	0	0	0
P(~h&e)	0	0	0	0	0	0	0	0	0	0	0	0	0
Total	2	18	9	7	3	9	1	17	0	3	1	1	71

The coders agreed 47 times out of 71. The most common category was p(e/~h), which Coder #3 used 18 times and Coder #4 24 (agreeing 16 times). Other common categories were p(~e/h) (agreed on 8 times), p(~h/e) (agreed on 10 times) and p(h/e) (agreed on 6 times). X^2 for this table is 360.4, df = 99, $p < .001$. Within these important categories, the marginals are in agreement: $X^2 = 10.5$, df = 12, p near .50. So we can regard the π and κ with confidence: $\pi = .591$, and $\kappa = .593$.

6.3.1. Collapsing categories preserving the distinction between correct and incorrect, within evidence, hypothesis, and ambiguous.

Some of these disagreements may be between similar categories, e.g., p(correct evidence) versus p(e/h). For example, in the table above, in the sub-table having to do with inference, there are four sentences off the diagonal.

1. One which Coder #3 called p(e/~h), and Coder #4 called p(wrong evidence).
2. One which Coder #3 called p(~e/h), and Coder #4 called p(e/~h).
3. One which Coder #3 called p(correct evidence) and Coder #4 called p(e/~h).
4. One which Coder #3 called p(wrong evidence) and Coder #4 called p(e/~h).

The first and third are the same idea but at different levels of abstraction. The second involved an evidence relation between the opposite categories. Perhaps the coder made a mistake in entering this on the sheet. Only the fourth is a major disagreement. It is evident that the full categorization scheme offers very fine distinctions. In our next analysis, we collapse across these potentially confusable categories, preserving

1. Whether the conditional probability was recognized to be an inference, a statement of evidence, or an "and".
2. Whether the hypothesis and evidence that were linked were the same ("correct") or different ("wrong").

The distinction between "ambiguous" and "and". In the full categorization scheme, there were four specific categories for where subjects spoke of p(hypothesis and evidence). In a sense, these are ambiguous forms of conditional probability. In another sense, they are exactly what they claim to be: conjunctions. In the "third level of collapse", below, we combine these with the "ambiguous" category, which is only fair under the former interpretation of the specific "and" forms.

For the full table (including sentences called "ambiguous", not mentioned, etc.; see p 6 of ta2cprel1.out), the X^2 was 297.9, df = 48. The X^2 test of whether the marginals were different was 105.0, df = 8, which is highly significant (π = .176 κ = .220). This is, of course, the same data as before, only looked at slightly differently; and the differences we have made are in the substantive categories. When we look at the table of only the substantive categories, the pattern is very similar to the table above.

"Correct" and "wrong" conditional probabilities. In the following table, and every other, the terms "correct" and "wrong", in this coding scheme, do not refer to an error on the coders' part or on the subject's report. Rather, they refer to a lack of correspondence between evidence and hypothesis in the word problem that the subject was talking about.

Coder #3's Coding

Coder #4's Coding	Evidence Correct	Evidence Wrong	Inference Correct	Inference Wrong	Ambiguous Correct	Ambiguous Wrong	Total
Evidence							
Correct	5	0	2	0	2	1	10
Wrong	1	29	1	6	0	0	37
Inference							
Correct	3	0	6	2	1	0	12
Wrong	0	1	0	10	0	1	12
Ambiguous							
Correct	0	0	0	0	0	0	0
Wrong	0	0	0	0	0	0	0
Total	9	30	9	18	3	2	71

The total X^2 for this table is 91.8, df = 15, $p < .001$. The marginal test had a X^2 of 7.4, df = 5, $p < .25$. $\pi = .573$ and $\kappa = .576$.

The change between this and the previous table is that some of the sentences on which coders previously disagreed, are now given the same higher-level category. For example, among those that were considered "evidence" by both coders, 3 of the sentences moved from being disagreements to being agreements, leaving only one sentence that Coder #3 called p(correct evidence) and Coder #4 called p(e/~h). [This distinction will be collapsed over in the next stage.]

Note that the indices of agreement have slipped. The number of categories has decreased, and hence the probability of agreement by chance has changed, compensating for the increased number of agreements.

Collapsing over the correct/wrong distinction. The final stage of collapsing of categories ignores whether the coders thought the subjects were talking about the same hypothesis and evidence, or the opposite. It can be argued that the particular evidence being referred to may be difficult to distinguish, while the coder may be able to discriminate whether the subject is talking about inference (p(h/e)) or evidence (p(e/h)) generically.

The coding here is different from the previous in that the "ambiguous" category, which was excluded from our smaller, "extracted", tables above (which looked only at p(h and e), etc., the specific statements of ambiguity) and the specific statements of ambiguity, are collapsed into one. We will consider them when we look at the whole table, but not when we look at the extracted "substantive" subset.

The whole table is here:

Coder #4's Coding	Coder #3's Coding: Evidence	Inference	Ambiguous	CDP	No Mention	Total
Evidence	35	9	37	6	2	89
Inference	4	18	15	0	11	48
Ambiguous	4	3	13	0	4	24
CDP	1	1	0	7	0	9
No Mention	7	6	49	0	9	71
Total	51	37	114	13	26	241

The overall X^2 was 166.7, df = 16, $p < .001$. The test of whether the marginals are significantly different had a X^2 of 92, df = 4, very significant. The π was .132, and the κ was .186.

When we exclude the ambiguous, no mention, and cdp categories, this table emerges:

	Evidence	Inference
Evidence	35	9
Inference	4	18

$X^2 = 22.8$, df = 1, $p < .001$. The test to determine whether the marginals differed had a X^2 of .811, df = 1, $p < .50$. The π was .578, and the κ was .581.

Summary of the total tables. The following table summarizes our results so far.

Degree of categorization detail.	Overall Test			Test of Marginals			Agreement	
Level of detail	X^2	df	p <	X^2	df	p <	π	κ
Evidence/Inference, Specific Correct/Incorrect								
Full Table	1072	182	.0001	112.2	16	.0001	.190	.230
Extracted Table	360	99	.001	10.5	12	.50	.591	.593
Evidence/Inference, General Correct/Incorrect								
Full Table	298	48	.0001	105	8	.0001	.176	.220
Extracted Table	92	15	.001	7.4	5	.25	.573	.576
Evidence/Inference only								
Full Table	167	16	.001	92	4	.0001	.132	.186
Extracted Table	23	1	.001	0.8	1	.50	.578	.581

Broken down by problem position. Problems 1 and 3 were combined into a sub-table which excludes categories relating to "no mention", "cdp", and the general "ambiguous". This table also collapses categories by "correct" and "wrong" evidence, "correct" and "wrong" inference, and "correct" and "wrong" ambiguity (specific).

Coder #3's Coding

Coder #4's Coding	Evidence Correct	Evidence Wrong	Inference Correct	Inference Wrong	Ambiguous Correct	Ambiguous Wrong	Total
Evidence							
Correct	3	0	0	0	1	1	5
Wrong	0	5	0	3	0	0	8
Inference							
Correct	2	0	4	2	1	0	9
Wrong	0	1	0	7	0	1	9
Ambiguous							
Correct	0	0	0	0	0	0	0
Wrong	0	0	0	0	0	0	0
Total	5	6	4	12	2	2	31

The coders' judgments of sentences as evidence, inference or specifically ambiguous agreed for 19 sentences. They disagreed on 12 sentences. Overall, $X^2 = 41.2$, df = 15, $p < .001$. The test to determine whether the marginals differed had a X^2 of 6.64, df = 5, $p < .25$. The π was .492, and the κ was .500.

The data this table ignores can be categorized as:

1. 7 sentences were considered "ambiguous" by both coders.
2. 8 paragraphs were coded as "no mention" by both individuals.
3. 18 sentences were considered by both coders to be either "ambiguous", "no mention" or "cdp", yet not the same category within these three.
4. 27 sentences were thought ambiguous by one coder but identifiable as evidence, inference, or specific ambiguous by the other.
5. 14 sentences were not considered by one coder ('no mention'), while the other coded them as evidence, inference, or specific ambiguous.

Problem 2.

Problem two, as one and three, was collapsed into categories of "correct"/"wrong" and evidence/inference/ambiguous. "No mention", the general "ambiguous", and "cdp" are, once again, omitted.

	Coder #3's Coding						
	Evidence		Inference		Ambiguous		
	Correct	Wrong	Correct	Wrong	Correct	Wrong	Total
Coder #4's Coding							
Evidence							
Correct	2	0	1	0	1	0	4
Wrong	1	24	0	3	0	0	28
Inference							
Correct	0	0	1	0	0	0	1
Wrong	0	0	0	3	0	0	3
Ambiguous							
Correct	0	0	0	0	0	0	0
Wrong	0	0	0	0	0	0	0
Total	3	24	2	6	1	0	36

The coders agreed on 30 of the sentences, and disagreed on 6. Overall, $X^2 = 58.7$, df = 12, $p < .001$. The test to determine whether the marginals differed had a X^2 of 2.8, df = 4, which is not significant. The π was .631, and the κ was .635. Thus there was slightly higher agreement between the coders on Problem 2 than on Problems 1 and 3.

The data this table ignores can be categorized as:

1. 4 sentences were considered "ambiguous" by both coders.
2. Both coders could not differentiate which paragraph the subject was referring to for 7 sentences
3. 11 sentences were not considered by one coder ('no mention'), while the other coded them.
4. One coder thought she could differentiate the paragraphs for 8 sentences, while the other one couldn't.
5. 28 sentences were considered by both coders to be either "ambiguous", "no mention" or "cdp", yet not the same category within these three.
6. 21 sentences were thought ambiguous by one coder but identifiable by the other as a specific category

Memory Test. This table depicts the collapsing of categories into evidence/inference and general correct/incorrect, for the memory test part of the transcript.

	Coder #3's Coding						
	Evidence		Inference		Ambiguous		
	Correct	Wrong	Correct	Wrong	Correct	Wrong	Total
Coder #4's Coding							
Evidence							
Correct	0	0	1	0	0	0	1
Wrong	0	0	1	0	0	0	1
Inference							
Correct	1	0	1	0	0	0	2
Wrong	0	0	0	0	0	0	0
Ambiguous							
Correct	0	0	0	0	0	0	0
Wrong	0	0	0	0	0	0	0
Total	1	0	3	0	0	0	4

Coders agreed 1 time, disagreed 3 (X^2 = 1.3, df = 2, NS). The test to determine whether the marginals differed had a X^2 of 1.2, df = 2, which is also not significant. The π was -.412, and the κ was -.333. This is very sparse! The data this table ignores can be categorized as:

1. 2 sentences were considered "ambiguous" by both coders.
2. 1 paragraph was coded as "no mention" by both individuals.
3. 6 sentences were thought ambiguous by one coder but identifiable by the other.
4. 1 sentence was not considered by one coder ('no mention'), while the other coded it.
5. 7 sentences were considered by both coders to be either "ambiguous", "no mention", yet not the same category within these.

Problems 1 and 3, collapsed into evidence and inference categories. The data from problems 1 and 3 combined, collapsed into general evidence/inference categories, and ignoring any sentences where at least one of the coders thought the sentence uncategorizable, yield:

	Coder #3's Coding	
	Evidence	Inference
Coder #4's Coding		
Evidence	8	3
Inference	3	13

The coders agree on 21 of these sentences, and disagree on 6. Overall, X^2 = 7.9 df = 1, $p < .01$. The test to determine whether the marginals differed had a X^2 of 0.0, df = 1. The π was .540, and the κ was .540.

Problem 2, collapsed into evidence and inference.

	Coder #3's Coding	
	Evidence	Inference
Coder #4's Coding		
Evidence	27	4
Inference	0	4

The coders agreed 31 times, disagreed 4 times (X^2 = 15.2, df = 1, $p < .001$). The test to determine whether the marginals differed had a X^2 of 1.6, df = 1, $p < .25$. The π was .598, and the κ was .607.

Memory test, collapsed into evidence and inference. Display of the memory test results, when the categories are collapsed so that only general evidence and inference distinctions are maintained and then only those sentences that both coders agreed fell into the substantive categories, shows the following pattern:

	Coder #3's Coding	
	Evidence	Inference
Coder #4's Coding		
Evidence	0	2
Inference	1	1

The coders agreed 1 time, disagreed 3 times (X^2 = 1.33, df = 1, $p < .25$). The test to determine whether the marginals differed had a X^2 of .53, df = 1, $p < .50$. The π was -.60, and the κ was -.50.

Summary of the comparison of problems. We have looked at the reliability of the coding of conditional probability statements in Problems 1 and 3, in Problem 2, and in the Memory test. The following reports the results for the extracted tables only, not the full table.

Degree of categorization detail.	Overall Test			Test of Marginals			Agreement	
Problem	X^2	df	p <	X^2	df	p <	π	κ
Evidence/Inference, Specific Correct/Incorrect								
All Problems	360	99	.001	10.5	12	.50	.591	.593
Evidence/Inference, General Correct/Incorrect								
All Problems	92	15	.001	7.4	5	.25	.573	.576
Problems 1 & 3	41.2	15	.001	6.6	5	.25	.492	.500
Problem 2	58.7	12	.001	2.8	4	NS	.631	.635
Memory Test	1.3	2	NS	1.2	2	NS	-.412	-.333
Evidence/Inference only								
All Problems	23	1	.001	0.8	1	.50	.578	.581
Problems 1 & 3	7.9	1	.01	0.0	1	NS	.540	.540
Problem 2	15.2	1	.001	1.6	1	.25	.598	.607
Memory Test	1.3	1	.25	0.5	1	.50	-.600	-.500

6.3.2. Inspection of coding reliability by paragraph.

The data can also be broken down by paragraph for each problem. The tables which display these figures are collapsed into the evidence/inference categories. Often the counts are small or zero.

Problem 1, paragraph 3. None of the codings done for this paragraph involve evidence and/or inference for both coders on the same sentence. Hence, the table would consist of all 0's. We omit it.

The data that don't fall into these conjunctive categories are as follows:

1. 2 sentences were considered "ambiguous" for both coders.
2. 4 paragraphs were coded as "no mention" by both individuals.
3. 7 sentences were coded as "no mention" by one coder, when the other chose "ambiguous".
4. 4 sentences were coded as "ambiguous" by one coder, while the other found them identifiable as evidence or inference statements.

Problem 1, paragraph 4.

	Coder #3's Coding	
	Evidence	Inference
Coder #4's Coding		
Evidence	5	0
Inference	2	6

The coders agreed 11 times, disagreed 2 times. Overall, $X^2 = 6.96$, df = 1, $p < .01$. The test to determine whether the marginals differed had a X^2 of .62, df = 1, $p < .50$. The π was .690, and the κ was .698.

The data that don't fall into these (evidence/inference) conjunctive categories are as follows:

1. 5 sentences were considered "ambiguous" by both coders.
2. 0 sentences were coded as "no mention" by both individuals.
3. 4 sentences were coded as "no mention" by one coder, when the other chose "ambiguous".
4. 16 sentences were coded as "ambiguous" by one coder, while the other found them identifiable as evidence or inference statements.
5. 8 sentences which one coder didn't mention, were judged to involve inference or evidence by the other.

Note that there are more data, and better agreement, for Paragraph 4 (where it is pertinent to the comparisons we used in our data) than for Paragraph 3.

Problem 2, ordering paragraph 3.

	Coder #3's Coding	
	Evidence	Inference
Coder #4's Coding		
Evidence	3	0
Inference	0	0

The coders agreed 3 times, disagreed 0 times. $X^2 = 0.0$, df = 1.

The data that don't fall into these (evidence/inference) conjunctive categories are as follows:

1. 3 sentences were coded as "cdp" by both individuals.
2. 5 sentences were coded as "no mention" by one coder, when the other chose "ambiguous".
3. 1 sentence was coded as "ambiguous" by one coder, while the other found it identifiable as an evidence or inference statement.
4. 7 sentences could not be linked to a particular paragraph by one coder, while the other coded it as evidence or inference referring to Paragraph 3.

Problem 2, ordering paragraph 4. Once again both coders did not concur on the inference/evidence categories being applicable for any sentence. The table has been omitted.

The data that don't fall into these (evidence/inference) conjunctive categories are as follows:

1. 4 sentences were considered "cdp" for both coders.
2. 7 sentences were coded as "no mention" by one coder, when the other chose "ambiguous".
3. 1 sentence was coded as "ambiguous" by one coder, while the other found it identifiable as an evidence or inference statement.
4. 1 sentence which one coder didn't mention, was judged inference or evidence by the other.
5. 1 sentence could not be linked to a particular paragraph by one coder, while the other coded it as evidence or inference referring to Paragraph 4.

Problem 2, answering paragraph 3.

	Coder #3's Coding	
	Evidence	Inference
Coder #4's Coding		

Evidence	15	4
Inference	0	4

The coders agreed 19 times, disagreed 4 times. Overall, $X^2 = 9.1$, df = 1, $p < .01$. The test to determine whether the marginals differed had a X^2 of 1.8, df = 1, $p < .25$. The π was .549, and the κ was .566.

The data that don't fall into these (evidence/inference) conjunctive categories are as follows:

1. 1 sentence was considered "ambiguous" for both coders.
2. 9 sentences were coded as "no mention" by one coder, when the other chose "ambiguous".
3. 16 sentences were coded as "ambiguous" by one coder, while the other found them identifiable as evidence or inference statements.
4. 5 sentences which one coder didn't mention, were judged inference or evidence by the other.

Problem 2, answering paragraph 4.

	Coder #3's Coding	
	Evidence	Inference
Coder #4's Coding		
Evidence	9	0
Inference	0	0

The coders agreed 9 times, disagreed 0 times. Overall, $X^2 = 0.00$, df = 1.

The data that don't fall into these (evidence/inference) conjunctive categories are as follows:

1. 3 sentences were considered "ambiguous" for both coders.
2. 7 sentences were coded as "no mention" by one coder, when the other chose "ambiguous".
3. 4 sentences were coded as "ambiguous" by one coder, while the other found them identifiable as evidence or inference statements.
4. 5 sentences which one coder didn't mention, were judged inference or evidence by the other.

In conclusion for Problem 2, the agreement is fairly good on those that both saw as codable, for each paragraph.

Problem 3, paragraph 3.

	Coder #3's Coding	
	Evidence	Inference
Coder #4's Coding		
Evidence	0	1
Inference	0	1

The coders agreed 1 time, disagreed 1 time. Overall $X^2 = 0.00$, df = 1.

The data that don't fall into these (evidence/inference) conjunctive categories are as follows:

1. 3 sentences were coded as "no mention" by both individuals.
2. 2 sentences were coded as "no mention" by one coder, when the other chose "ambiguous".
3. 2 sentences were coded as "ambiguous" by one coder, while the other found them identifiable as evidence or inference statements.
4. 4 sentences which one coder didn't mention, were judged inference or evidence by the other.

Problem 3, paragraph 4.

	Coder #3's Coding	
	Evidence	Inference
Coder #4's Coding		
Evidence	3	2
Inference	1	6

The coders agreed 9 times, disagreed 3 times. Overall $X^2 = 2.74$, df = 1, $p < .10$. The test to determine whether the marginals differed had a X^2 of .18, df = 1, which is not significant. The π was .467, and the κ was .471.

The data that don't fall into these (evidence/inference) conjunctive categories are as follows:

1. 1 sentence was coded as "no mention" by both individuals.
2. 5 sentences were coded as "no mention" by one coder, when the other chose "ambiguous".
3. 9 sentences were coded as "ambiguous" by one coder, while the other found them identifiable as evidence or inference statements.
4. 2 sentences which one coder didn't mention, were judged inference or evidence by the other.

Again, there are more data and better agreement, on Paragraph 4 where it is needed.

[Note, while it is interesting to contrast Paragraph 4 (evidence only) with Paragraph 4 (conditional probability) on Problems 1 and 3, there is no reason to do so on Problem 2.]

6.3.3. Conclusions from the analysis of reliability on a passage by passage basis.

The sentence-by-sentence analysis of coder reliability has shown that if we consider all the categories, including those concerned with the coder's decision whether a sentence represented the use of a conditional probability or not, there was fairly low reliability. For example, about 18% of the opportunity for agreement, over and above chance agreement, was taken, and in the remaining 82% the coders disagreed.

However, if the questionnable cases were excluded, i.e., those where one or both of the coders thought the sentence not to involve a conditional probability, or to involve a conditional probability whose identification was ambiguous, then we get about 60% agreement, over and above chance.

This indicates that there is still a fair amount of unreliability in the categorization of the conditional probabilities. While this agrees with the main assertions of the research, i.e., that the conditional probabilities are hard for regular people to distinguish, and they are hard for coders to distinguish too, this gives scant comfort. In particular, we are in a situation where there is a danger of confirming the null hypothesis:

1. The hypothesis of the research is that is it difficult for people to distinguish two types of conditional probabilities.
2. The statistical significance of a finding is related to the product (multiplication) of the effect size

and the number of cases.

3. Because of the amount of effort involved in this work, the number of subjects used is very small.
4. The unreliability of the coding scheme makes the effect size estimate small.
5. Both the unreliable coding, and the small n, make it unlikely that we would find any statistically significant differences, which is what we wanted in the first place.

So the coding of the subject's use of conditional probabilities is problematic.

One possibility would be to use only those codes that both coders agreed on. This would allow there to be higher reliability (still, by chance, these could both be wrong). This approach is rejected because we have both coders' answers only on 10 of the 32 subjects. We would need to get the second coder to do it for all the others, in order to be able to apply this strategy.

Instead, we are relying on two other factors that allow us to argue that the reliability estimate is a low estimate of the actual reliability (validity) of the coding.

1. First, Coder #3 is the coder who coded all 32 subjects, and her coding is higher quality, i.e., more in agreement with Hamm's definitions. She used the "ambiguous" category more frequently. This includes some specific conditions where the Coder's Materials specified that it should be called ambiguous: the use of "she thought it was Stephen" in the Twins problem, because it was a phrase used in the text; the use of "probability of error" in the doctor problem, and analogous phrases in the other two, because these phrases were included in both versions of the conditional probability information. This factor means that the coding on all the subjects is probably better than the relatively low reliability score would indicate.
2. Second, for half of the subjects Hamm went over all judgments with the coders and a decision was made about what is the best category.
   a. Six of the subjects' transcripts were used as training, and Hamm, Coder #3, and Coder #4 went over them all, and arrived at an agreed upon answer.
   b. Ten of the transcripts were used for measuring reliability. Every sentence that Coder #3 and Coder #4 disagreed on was discussed by Hamm and Coder #3, and the best answer (which sometimes was an answer neither coder had used) was assigned to it.

Therefore, the coding on 16 of the subjects has the benefit of both coders' and Hamm's judgment; the coding on the final 16 subjects has only the benefit of Coder #3's judgment, but hers was more accurate than Coder #4's.

One might ask, what is the point of having two coders if one of them, who received less training, is less good? It gives us a lower limit measure of the reliability. The work of repeated coding and checking certainly improved the coding of Hamm and Coder #4.

6.4. Reliability of conditional probability coding, analyzed by location in the questionnaire.

The accuracy of coding of conditional probabilities has been described above in terms of the agreement between the two coders on a sentence by sentence basis. The way this coding will be used in the analysis of the data, however, will be on a locus by locus, juncture by juncture, i.e., paragraph by paragraph, basis. (All of Coder #1 and Coder #2's coding is also on a paragraph by paragraph basis.) This section describes the reliability of the coding with respect to the unit of analysis used in the whole study.

The number of times each type of conditional probability concept (category) was used at each locus. For the purposes of the study, the focus of the analysis is on the use of each category at each *locus* in the procedure, each paragraph. This can be measured by

1. counting how often the category was used at each locus
2. giving each category a yes/no score

3. determining whether each category was used more often, equal to, or less often than its complement.

Reliability on each can be measured. We will deal only with the first, the count of how often the category was used at each locus.

6.4.1. Reliability of the count of uses of categories.

How reliable is the count of the number of times the subject used each category at each locus? Several methods are available to analyze this:

1. Cross tabulations: how often did the two coders say the same number? This can be done for each of 18 categories at each of 9 loci.
2. Correlation of how often a category was used, over the 10 subjects. This can be done for each of 18 categories at each of 9 loci.
3. Within each locus, correlation between coders over the 18 categories. This is calculated for each subject, and then the mean over the 10 subjects can be looked at.
4. Within each of the 18 categories, correlation between coders over the 9 loci where the category was used. This can be calculated for each subject, and then the mean over the subjects can be looked at.
5. Correlation between coders over all 18 categories times 9 loci can be calculated for each subject, and then the mean over subjects taken.

In addition, we recognize that the concepts are perhaps too detailed. Parallel with the collapsing of categories done in the sentence- by- sentence reliability analysis (Section 7.1), we can collapse the categories that are being counted here, and look at some of the above types of reliability measure.

6.4.2. Cross tabulations.

The cross classification compares, for a given location and a given category, how often (out of 10 subjects) the two coders had exactly the same count. It does not tell us if they were counting the same sentences.

Since most categories were seldom used, most of these show 10 0's. An example is the revelation of the different uses of "ambiguous", between the two coders. In the memory test, here is the table:

		Coder #4		
		0	1	2
Coder #3	0	2	1	0
	1	2	0	0
	2	1	0	0
	3	1	0	1
	4	1	0	0

We see that only on 2 of the 9 [one subject did not do a memory test that was coded] subjects did the two coders agree, both saying there were no ambiguous conditional probabilities.

For the most part, this information is too detailed to be informative.

6.4.3. Correlation between coders, over the 10 subjects, for each category, each locus.

The counts (most often 0's) can be correlated. If there was no variation (all 10 subjects were given 0's), no correlation is calculated.

	Problem 1		Problem 2				Problem 3		Memory
	Par 3	Par 4	Ordering Par 3	Ordering Par 4	Answering Par 3	Answering Par 4	Par 3	Par 4	Test
P(h/e)	.	-.41	.	.	.75	.	.29	.55	.50
P(h/~e)	.	.	.	.	.	.	.	.66	.
P(~h/e)	.	.93	.	.	.66	-.13	.	.58	.
P(~h/~e)	.	.	.	.	.	.	.	.	.
P(e/h)	.	-.11	.	.	.	1.00	.	.	.
P(e/~h)	.	.93	1.00	.	.98	.81	-.13	-.38	.
P(~e/h)	.	.	.	.67	1.00	1.00	.	.	.
P(~e/~h)	.	.	.	.	.	.	.	.	.
P(correct evidence)	.	-.02	.	.	.	.66	.	.50	-.89
P(wrong evidence)	.	.	1.00	.	.	1.00	.	.66	.
P(correct inference)	.	.	.	.	.	.	.	.	.
P(wrong inference)	.	.	.	.	.	.	.	.	.
P(h & e)	.	.	.	.	.	1.00	.	.	.
P(h & ~e)	.	.	.	.	.	.	.	.	.
P(~h & e)	.	.	.	.	.	.	.	.	.
P(~h &~e)	.	.	.	.	.	.	.	.	.
Ambiguous	.06	-.05	-.22	-.30	.51	-.61	-.19	-.22	.16
No Mention	.82	.	.	.	.	.	.79	.	.50

6.4.4. Correlations over categories/loci, within coder.

The next categories of correlations required a procedure be written, using numerical transformations, to produce the correlation. There are 36 input variables for this procedure, for Coder #3's 18 categories and Coder #4's 18 categories. There are 4 output variables, for the Coder #3 times Coder #4 crossproduct, the standard deviation of Coder #3 and Coder #4 category-use-counts, and the correlation of Coder #3 and Coder #4 category use counts. And there is also the mean of the Coder #3 counts, the mean of the Coder #4 counts, and the count of the number of variables that go into each of these means.

The code gets the mean of the 18 Coder #3 variables, and then transforms them into their deviation

scores. Similarly for Coder #4. It then takes the standard deviation of the Coder #3 scores, over the 18 categories, and of the Coder #4 scores, and the sum of their crossproducts. The correlation is then the quotient of the sum of the crossproducts divided by the product of the standard deviations.

If any loci were given exactly the same profile of counts, even if that was all 0's except for 1 no mention, they would be given a correlation score of 1.0. If any loci were given all 0's (as happened if the information was missing and not coded; or if, on problem 2, the idea of conditional probability was mentioned but it was difficult to link it to paragraph 3 or 4), then the correlation was called missing, due to the division by 0 (no variance).

The mean over-category correlations for these 9 loci, each across the 10 subjects, are given in the following table:

	Mn r	SD r	Min	Max	N
Prob 1, para 3 (evidence)	.576	.547	-.06	1.0	10
Prob 1, para 4 (cond prob)	.414	.419	-.17	1.0	10
Prob 2, ord, para 3	.314	.430	-.06	.7	4
Prob 2, ord, para 4	.122	.376	-.08	.7	4
Prob 2, ans, para 3	.472	.478	-.20	1.0	10
Prob 2, ans, para 4	.437	.512	-.09	1.0	9
Prob 3, para 3 (evidence)	.377	.526	-.06	1.0	9
Prob 13 para 4 (cond prob)	.296	.374	-.06	.8	9

With N = 18 categories, a correlation would have to be .40 to be significant at the .05 level, and .542 at the .01 level.

6.4.5. Correlations between coders, over *loci*, within category, within subject.

This analysis, as the last, required use of numerical transformations to compute. It was done for each of 18 categories.

The data are:

	Mn r	SD r	Min	Max	N
p(h/e)	.702	.321	.15	1.0	6
p(h/~e)	1.000	.	1.00	1.0	1
p(~h/e)	.637	.485	-.27	1.0	7
p(~h/~e)	.	.	.	.	0
p(e/h)	.884	.	.88	.9	1
p(e/~h)	.527	.472	-.19	1.0	8
p(~e/h)	.719	.563	-.13	1.0	4
p(~e/~h)	.	.	.	.	0
p(correct evidence)	.112	.436	-.29	.7	5
p(wrong evidence)	.591	.448	.11	1.0	3
p(correct inference)	.	.	.	.	0
p(wrong inference)	.	.	.	.	0
p(h&e)	1.0	.	1.00	1.0	1
p(h&~e)	.	.	.	.	0
p(~h&e)	.	.	.	.	0
p(~h&~e)	.	.	.	.	0
ambiguous	.128	.381	-.61	.7	10
no mention	.675	.214	.40	1.0	6

With N = 9 loci (it was in most cases), a correlation would have to be .582 to be significant at the .05 level, and .750 at the .01 level.

6.4.6. Correlations between coders over all categories and all loci, within subject.

The first of these analyses of reliability by calculating within-subject correlations looked across categories, within locus; the next one looked across loci, within category. This one will look across both categories and loci. That is, for all 162 category/locus entities, how much did the two coders agree? The results for each subject are:

Subject	r	Number of locus/category combinations	Statistical Significance of r
s1	.62	162	<.0005
s2	.06	144	NS
s4	.16	144	NS
s5	.32	162	<.0005
s7	.12	162	NS
s8	.40	162	<.0005
s11	.32	162	<.0005
s12	.60	162	<.0005
s14	.72	126	<.0005
s15	.39	162	<.0005

The average of these correlations, across the 10 subjects whose transcripts were categorized by both coders, is .369, SD = .222, $p < .005$.

6.4.7. Correlations between coders, after categories have been collapsed into more general categories.

There were two levels of collapsing. These are identical to those used in the sentence by sentence analysis (Section 7.1). In sum:

1. Collapsing preserving both the evidence/inference distinction and the correct/wrong distinction.
   a. evidence
      i. correct evidence: p(e/h), p(~e/~h), or p(correct evidence)
      ii. wrong evidence: p(e/~h), p(~e/h) or p(wrong evidence)
   b. inference
      i. correct inference: p(h/e), p(~h/~e), or p(correct inference)
      ii. wrong inference: p(h/~e), p(~h/e), or p(wrong inference)
   c. ambiguous
      i. correct ambiguous: p(h and e) or p(~h and ~e)
      ii. wrong ambigous: p(h and ~e) or p(~h and e)
   d. In addition:
      i. ambiguous
      ii. no mention
2. Collapsing preserving only the evidence/inference distinction.
   a. evidence: p(h/e), p(~h/~e), or p(correct inference); p(h/~e), p(~h/e), or p(wrong inference)
   b. inference: p(h/e), p(~h/~e), or p(correct inference); p(h/~e), p(~h/e), or p(wrong inference)
   c. ambiguous: p(h and e) or p(~h and ~e); p(h and ~e) or p(~h and e); or "ambiguous"
   d. or "no mention"

We did the correlations within row, for each of these new variables. These are:

	Mn r	SD r	Min	Max	N
e/i and c/w					
correct evid	.076	.397	-.26	.66	6
wrong evid	.534	.446	-.22	1.00	9
correct inf	.560	.475	-.29	1.00	7
wrong inf	.761	.466	-.27	1.00	7
correct amb	1.000	.	1.00	1.00	1
wrong amb	~	~	~	~	~
"ambig"	.128	.381	-.61	.70	10
no mention	.675	.214	.40	1.00	6
e/i					
evidence	.416	.457	-.24	.97	9
inference	.624	.454	-.29	1.00	10
ambiguous	.159	.394	-.55	.70	10
no mention	.675	.214	.40	1.00	6

Each of these correlations is between coders, across 9 loci (though a few of the subjects were lacking some of the later loci), so the appropriate df is 8. Hence the level of statistical significance is .582 for .05, and .750 for .01.

Conclusion. When we measure reliability as the correlation between the counts used by the coders, over the 9 loci, the average reliability is r = .42 for evidence, r = .62 for inference.

References.

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20, 37-46.

Hubert, L. (1977). Kappa revisited. Psychological Bulletin, 84, 289-297.

Scott, W. A. (1955). Reliability of content analysis: The case of nominal scale coding. Public Opinion Quarterly, 19, 321-325.

Stuart, A. A. (1955). A test of homogeneity of marginal distributions in a two-way classification. Biometrika, 42, 412-416.

Zwick, R. (1988). Another look at interrater agreement. Psychological Bulletin, 103, 374-378.

Zwick, R., Neuhoff, V., Marascuilo, L. A., and Levin, J. R. (1982). Statistical tests for correlated proportions: Some extensions. Psychological Bulletin, 92, 258-271.

Appendix I. Coding sheets for Problems 1 and 3.

TA2 THINK-ALOUD CODING SHEET: PROBLEMS 1 AND 3

Subject Number_____ Problem: 1 or 3 Content: doctor insurance twins

After Second Paragraph (Baserate)

1a. mentions correct baserate number
1b. mentions incorrect baserate number
1c. mentions nonspecific baserate
1d. no mention of baserate

2a. uses baserate as the answer
2b. uses baserate in an adjustment process (list adjustment process)
2c. no use of baserate (list reason)

3a. interpolation - no justification
3b. interpolation - evidence justification
3c. interpolation - reliability justification
3d. interpolation - other justification (list justification)

	LIST	PREVIOUS ANSWER	BASERATE	EVIDENCE	RELIABILITY CORRECT	RELIABILITY ERROR
FIRST NUMBER:						
SECOND NUMBER:						

NEW ANSWER:_____

3e. anchor and adjust - no justification
3f. anchor and adjust - evidence justification
3g. anchor and adjust - reliability justification
3h. anchor and adjust - other justification (list justification)

	LIST	PREVIOUS ANSWER	BASERATE	EVIDENCE	RELIABILITY CORRECT	RELIABILITY ERROR
ANCHOR NUMBER:						

NEW ANSWER:_____

3i. some other adjustment made (list adjustment process)
3j. no adjustment is made

After Third Paragraph (Evidence)

1a. mentions correct baserate number
1b. mentions incorrect baserate number
1c. mentions nonspecific baserate
1d. no mention of baserate

3a. interpolation - no justification
3b. interpolation - evidence justification
3c. interpolation - reliability justification
3d. interpolation - other justification (list justification)

	LIST	PREVIOUS ANSWER	BASERATE	EVIDENCE	RELIABILITY CORRECT	RELIABILITY ERROR
FIRST NUMBER:						
SECOND NUMBER:						

NEW ANSWER: _____

3e. anchor and adjust - no justification
3f. anchor and adjust - evidence justification
3g. anchor and adjust - reliability justification
3h. anchor and adjust - other justification (list justification)

	LIST	PREVIOUS ANSWER	BASERATE	EVIDENCE	RELIABILITY CORRECT	RELIABILITY ERROR
ANCHOR NUMBER:						

NEW ANSWER: _____

3i. some other adjustment made (list adjustment process)
3j. no adjustment is made

4a. uses baserate only - no justification
4b. uses baserate only - reliability justification
4c. uses baserate only - other justification (list justification)
4d. no use of baserate (list reason)
4e. uses baserate in conjunction with other info (circle something in part 3)

5a. mentions correct evidence
5b. mentions incorrect evidence
5c. no mention of evidence

6a. uses evidence only
6b. uses evidence to adjust (circle something in part 3)
6c. no use of evidence-reliability justification
6d. no use of evidence (list justification)

7a. mentions reliability issue - goes beyond provided info
7b. mentions reliability issue - repeats info provided
7c. no mention of reliability

After Fourth Paragraph (Reliability)

1a. mentions correct baserate number
1b. mentions incorrect baserate number
1c. mentions nonspecific baserate
1d. no mention of baserate

3a. interpolation - no justification
3b. interpolation - evidence justification
3c. interpolation - reliability justification (circle 7a or b)
3d. interpolation - other justification (list justification)

	LIST	PREVIOUS ANSWER	BASERATE	EVIDENCE	RELIABILITY CORRECT	RELIABILITY ERROR
FIRST NUMBER:						
SECOND NUMBER:						

NEW ANSWER: _____

3e. anchor and adjust - no justification
3f. anchor and adjust - evidence justification
3g. anchor and adjust - reliability justification (circle 7a or b)
3h. anchor and adjust - other justification (list justification)

	LIST	PREVIOUS ANSWER	BASERATE	EVIDENCE	RELIABILITY CORRECT	RELIABILITY ERROR
ANCHOR NUMBER:						

NEW ANSWER: _____

3i. some other adjustment made (list adjustment process)
3j. no adjustment is made

4a. uses baserate only - no justification
4b. uses baserate only - reliability justification
4c. uses baserate only - other justification (list justification)
4d. no use of baserate (list reason)
4e. uses baserate in conjunction with other info (circle something in part 3)

5a. mentions correct evidence
5b. mentions incorrect evidence
5c. no mention of evidence

6a. uses evidence only
6b. uses evidence to adjust (circle something in part 3)
6c. no use of evidence-reliability justification
6d. no use of evidence (list justification)

8a. mentions correct reliability number
8b. mentions incorrect reliability number
8c. mentions nonspecific reliabilty
8d. no mention of reliability

9a. uses "correct" reliability only
9b. uses "error" reliability only
9c. uses reliability to adjust (circle something in part 3)
9d. no use of reliability (list)

10a. uses Bayes' theorem
10b. mentions Bayes' theorem but doesn't use
10c. no use of Bayes' theorem

Appendix II. Reliability of conditional probability coding, analyzed by sentences.

In order to compute reliability for the kind of table that we have here, we refer to the procedure recommended by Rebecca Zwick (1988). She notes that two coders may use their categories with different frequencies. If we think of an n by n table of the first coder's categorizations versus the second coder's categorizations of the same set of objects, then these profiles (distributions of a coder's categorizations, over the possible categories) are the marginals.

Zwick argues that we must first determine whether the two coders have similar (or "homogenous") marginals. Only if the marginals are similar, does it make sense to use the statistical indices of agreement that are available for 2 by 2 or n by n tables.

To test if the marginals are similar, she suggests using Stuart's 1955 test (if the sample is fairly large) (which involves matrix inversion), or a simpler thing, producing an index

$M = 1 - X^2/n.$

The source of the X^2 in Zwick (1988) was not clear. The solution we adopted was to use the following procedure:

1. Produce a 2 by N table, where the two rows are the marginals for the two coders, and the N columns represent the N categories the coders used.
2. The observed frequences Fo are the counts of how frequently each coder used each category -- the marginals from the N by N table.
3. The expected frequencies Fe are calculated in the usual way for a j by k table, that is,

$$Fe_{jk} = \frac{Fo_{j*} * Fo_{*k}}{Fo_{**}}$$

In the case where $Fo_{1*} = Fo_{2*}$, that is, where both coders coded the same number of objects, the two coders' expected frequencies for a category are equal, so

$$Fe_{1k} = Fe_{2k} = \frac{F0_{1k} + Fo_{2k}}{2}$$

4. The X^2 of this table should not reveal that the distribution of use of the coding categories of Coder 1 is different from the distribution of Coder 2.

Once we have decided that the marginals have sufficiently similar patterns, we can compute an index of agreement. The basic form of the index of agreement is:

$$A = \frac{P_O - P_C(A)}{1 - P_C(A)}$$

where

$$P_O = \sum_{i=1}^{k} p_{ii}$$

is the observed proportion of agreement, p_{ii} is the proportion of cases in the ith diagonal cell of the table, and $P_C(A)$ is the proportion of agreement expected by chance. This definition will vary for difference coefficients, A. (Zwick, 1988, p 374). "These coefficients represent an attempt to correct P_O by subtracting from it the proportion of cases that fall on the diagonal by 'chance'. the numerator is then divided by 1 - $P_C(A)$, the maximum nonchance agreement."

The Pi (π) index. In Scott's (1955) π coefficient, the chance agreement is defined by assuming that the two coders really have the same marginals, though they differ by chance. So, take the average of

their proportions, $(p_{i*} + p_{*i})/2$, as the estimate of the marginals, and then calculate the expected proportion to fall in the marginals by chance using the following formula:

$$P_C(\pi) = \sum_{i=1}^{k} \left(\frac{p_{ix} + p_{xi}}{2}\right)^2$$

The Kappa (κ) index. The κ index (Cohen, 1960) does not assume that the individuals have the same marginal distributions, and so it calculates the chance expectation of agreement using the same formula used to get expectations in a X^2 analysis.

The measure of chance expectation of agreement between the coders is:

$$P_C(\kappa) = \sum_{i=1}^{k} p_{ix} p_{xi}$$

These then get plugged into the other formula, to produce the π and κ indices, as follows:

$$\pi = \frac{P_O - \sum_{i=1}^{k} \left(\frac{p_{ix} + p_{xi}}{2}\right)^2}{1 - \sum_{i=1}^{k} \left(\frac{p_{ix} + p_{xi}}{2}\right)^2}$$

$$\kappa = \frac{P_O - \sum_{i=1}^{k} p_{ix} p_{xi}}{1 - \sum_{i=1}^{k} p_{ix} p_{xi}}$$

Hubert (1977, pp 293-294) gives an expected value and variance for π, allowing for tests of whether it is significant (given that we already think the marginals are similar).

Zwick, Neuhoff, Marascuilo, and Levin (1982) provide tests for each category.

Appendix III: Listing of the Fortran Program.

```
c    RELIND.FOR        Rob Hamm      6. 3. 88
c
c   This program will produce some reliability indices for N by N tables.
c   The issue is whether two coders agree on the assignment of cases to
c   categories. The analysis will produce both the Kappa (Cohen, 1960) and
c   Pi (Scott, 1955) indices, reviewed by Zwick (1988). Following Zwick,
c   the program will also test for whether the marginal distributions are
c   the same. Zwick advises that the other indices not be heeded, unless
c   the marginals are similar.
c
c   The program's input will be the dimensions, and then the cell values.
c   It will calculate the marginals, report an elementary Chi-squared
c   (indicating, grossly, whether the subjects have any agreement)
c   test for whether the marginals
c   are similar, report conclusions of the marginal
c   comparison, and then (no matter what)
c   calculate the Pi and Kappa indices. It will make a statistical test
c   of the Pi index.
c
c      chi-squared = Sum over cells of ((fo - fe)**2)/fe
c
c              The formula for one df is:
c
c      chi-squared = sum over cells of ((|fo - fe| - .5)**2)/fe
c
c
c Memory structure
c
      DOUBLE PRECISION CHISQ,FO(20,20),fe(20,20),fe2(2,20),
     2         frowmar(20),fcolmar(20),ftotal,pchance,pagree,pi,
     3         kappa,pchank
      INTEGER N,J, ERFLAG, nm1, nrows,ncolumns,nrowind,ncolind,
     2  degfre,degfr2,errflag
C
C INTERVIEW THE USER.
C
 100     WRITE (5,105)
 105     FORMAT (' How many categories in this Category Judgment ',
     2  'Reliability Analysis?'/' Maximum = 20')
      READ (5,110)nrows
 110     FORMAT (I4)
       ncolumns = nrows
 111    do 114, nrowind = 1,nrows
 112    write (5,113)nrowind,ncolumns
 113    format (' Enter Fo"s from row ',I4', 1/line, '/
     2  i3,' lines, use decimal pt.')
       read (5,116)(fo(nrowind,j),j=1,ncolumns)
1134    write(*,1135)
1135    format(' Is this row correct? 1 = correct, 2 = redo it.')
       read(*,332)icheck
       if (icheck.eq.1) go to 114
       if (icheck.eq.2) go to 112
       go to 1134
 114    continue
 116     FORMAT (F10.0)
```

```
C
C CALCULATE
C
1165   errflag = 0

c produce the marginals
      do 120, nrowind = 1,nrows
      frowmar(nrowind) = 0.0
      do 118, ncolind = 1,ncolumns
      frowmar(nrowind) = frowmar(nrowind) + fo(nrowind,ncolind)
 118   continue
 120   continue
c
      do 125, ncolind = 1,ncolumns
      fcolmar(ncolind) = 0.0
      do 122, nrowind = 1,nrows
      fcolmar(ncolind) = fcolmar(ncolind) + fo(nrowind,ncolind)
 122   continue
 125   continue


C REPORT TO USER
C
     write (*,201)
201   format(/' Confusion matrix between coders. '
    2 'Observed frequencies.'/
    3  ' (check this against your intended input)'/
    4  ' Coder 1 categories are the rows, Coder 2 cats across top'/)
c
      do 2065, nrowind = 1,nrows
     WRITE (6,206)(FO(nrowind,ncolind),ncolind=1,ncolumns)
 206   FORMAT (16F5.0/'  '4f5.0)
 2065  continue
c
c      write (*,4002)
c      read (*,4003)dummy

c  Give the user a choice of whether to reenter the whole table
c  Watch out, these numbers are out of sequence!
329    write(*,330)
330    format(/' Are the Observed Frequencies now correct?'/
    2 ' 1 = yes, 2 = no')
      read(*,332)icheck
332    format(i2)
      if (icheck.eq.1.and.errflag.eq.0) go to 340
      if (icheck.eq.1.and.errflag.eq.1) go to 1165
      if (icheck.eq.2) go to 335
      go to 329
c    Set the error flag to 1.
335    errflag = 1
      write(*,336)
336    format(' Indicate row and column of cell you need to change,'
    2 /' the row # on the 1st line, and the column # on the 2nd')
      read(*,337)nrowfx,ncolfx
337    format(i2)
      write(*,338)nrowfx,ncolfx
338    format(' What is the correct count for row 'i2', column'i2'?')
```

```
      read (*,339)fo(nrowfx,ncolfx)
339   format(f10.0)
      go to 329

c      Produce the grand total
340    ftotal = 0.0
 126   do 128, nrowind = 1,nrows
       ftotal = ftotal + frowmar(nrowind)
 128   continue
c
c     Calculate the expected values
       do 140, nrowind = 1,nrows
       do 130, ncolind = 1,ncolumns
       fe(nrowind,ncolind) = frowmar(nrowind)*fcolmar(ncolind)/ftotal
 130   continue
 140   continue
c for debugging
c       do 150,nrowind = 1,nrows
c       write(*,145)(fe(nrowind,ncolind),ncolind=1,ncolumns)
c145     format(10f8.3)
c150     continue

c
c Calculate Chi-squared.
 180   CHISQ = 0
       DO 200, nrowind=1,nrows
       do 190, ncolind = 1,ncolumns
       if (fe(nrowind,ncolind).eq.0) go to 190
       CHISQ = CHISQ + ((FO(nrowind,ncolind) - FE(nrowind,ncolind))**2)/
    2  FE(nrowind,ncolind)
 190   continue
 200     CONTINUE
C
c  Now show the expected frequencies
20659      WRITE (*,2066)
 2066  format (//' Expected frequencies for each cell.')
       do 2075, nrowind = 1,nrows
      WRITE (6,206)(FE(nrowind,ncolind),ncolind=1,ncolumns)
 2075  continue

       degfre = (nrows - 1)*(ncolumns - 1)
      WRITE (6,210)CHISQ,degfre
 210   format (/' Chi-squared = ',F10.4,' df = ',I3/
    2  ' [If any row or column has only 0"s, it is dropped in '
    3  /' calc of chi-sq. You must adjust the reported df.]')
c Now do it for the df = 1 case
       if (nrows.ne.2.or.ncolumns.ne.2) go to 350

 280      CHISQ = 0
      DO 300, nrowind=1,nrows
       do 290, ncolind = 1,ncolumns
      CHISQ = CHISQ + ((abs(FO(nrowind,ncolind) - FE(nrowind,ncolind))
    2   - .5)**2)/FE(nrowind,ncolind)
 290   continue
 300     CONTINUE
C
C REPORT TO USER
```

```
C
      WRITE (6,310)CHISQ
310   format(' Chi-squared, Yates" correction, df = 1, is ',f8.3)


c
c Where we will go from here:
c Test for the equality of the marginals
c  display them
c  do a simple 2 by n Chi-squared
c  do a more complicated procedure from Zwick
c Next, do the kappa test
c Next, do the pi test

c Test for the equality of the marginals
c We have a 1 by n vector of column marginals, and an n by 1 vector
c of row marginals. At issue is whether these have the same distribution.
c An easy way to test this is to do a Chi-squared on this, as a 2 by n
c table.
c
c Make the expected frequencies matrix. The expected frequenceis for the
c column marginals will be in
c the first row, and for the row marginals in the second row.
350   do 360, nrowind = 1,nrows
      fe2(1,nrowind) = .5*(frowmar(nrowind)+fcolmar(nrowind))
      fe2(2,nrowind) = .5*(frowmar(nrowind)+fcolmar(nrowind))
360   continue

c Calculate Chi-squared.

 380     chisq2 = 0
      DO 400, nrowind= 1,nrows
      chisq2 = chisq2 + (fcolmar(nrowind) - fe2(1,nrowind))**2/
     2                fe2(1,nrowind)
      chisq2 = chisq2 + (frowmar(nrowind) - fe2(2,nrowind))**2/
     2                fe2(2,nrowind)
 400     CONTINUE
C
C REPORT TO USER
C
 4001   write (*,4002)
 4002   format (/' Hit RETURN when ready ')
        read (*,4003)dummy
 4003   format (F8.3)
        write(*,401)
401     format(/' Coders" marginals. 1st line = row marginals.'
     2  /' 2nd line = column marginals. More than 10 categs '
     2  'are wrapped.')
        WRITE (6,406)(frowmar(nrowind),nrowind = 1,nrows)
        WRITE (6,406)(fcolmar(nrowind),nrowind = 1,nrows)
 406     FORMAT (10F7.1/' '10f7.1)

c
      write (*,410)nrows
410   format(/' Expected frequencies in 2 (coder) by ',i2,
     2  ' (category) matrix.')
      do 415, i = 1,2
```

```
      write (6,412)(fe2(i,nrowind),nrowind = 1,nrows)
 412    FORMAT (10F7.1/' '10f7.1)
 415   continue
c
      degfr2 = nrows - 1
      WRITE (6,430)chisq2,degfr2
 430   format (/' Chi-squared test of whether these marginals are '
     2 'different = ',F10.4/' df = ',i3/)

c [Here place the more sophisticated tests of marginal differences.]

c The Pi test. First we will produce the number that represents
c the expected total proportion of the cells in this table that
c would agree by chance, according to this index (which assumes that
c both coders have the same marginal distribution, distorted today by
c error, and takes their mean, for each category, to estimate the true
c marginal proportion).

      pchance = 0.0
      do 450, nrowind = 1,nrows
      pchance = pchance
     2 + ((.5*(frowmar(nrowind)+fcolmar(nrowind)))**2)/ftotal**2
 450   continue

c  Next, calculate the observed proportion of agreements
      pagree = 0.0
      do 460, nrowind = 1, nrows
      pagree = pagree + fo(nrowind,nrowind)/ftotal
 460   continue

c  Put these in the formula for Pi.

      pi = (pagree - pchance)/(1 - pchance)
      write (6,470)pi
 470   format(' Pi = ',f8.3)

c  Calculate the chance expectation, according to the Kappa index.
      pchank = 0.0
      do 550, nrowind = 1,nrows
      pchank = pchank + frowmar(nrowind)*fcolmar(nrowind)/ftotal**2
 550   continue

c  Calculate the Kappa index
      kappa = (pagree - pchank)/(1 - pchank)
      write (6,570)kappa
 570   format(' Kappa = ',f8.3)

 580   write(*,590)
 590   format(' Do you need to make further corrections?'
     2 /' 1 = yes, 2 = no')
      read(*,332)icheck
      if (icheck.eq.1) go to 335
      if (icheck.eq.2) go to 100
      go to 580
      END
```