T · H · E
OHIO
STATE
UNIVERSITY

Enroute Flight Planning: The Design of Cooperative Planning Systems

Philip J. Smith and Chuck Layton
Cognitive Systems Engineering Laboratory
The Ohio State University

Elaine McCoy
Department of Aviation
San Jose State University

NASA Ames Research Center
Moffett Field, California 94035

Contract No. NCC 2-615
Final Report
RF Project 767591/722399

July 1990

Enroute Flight Planning: The Design of Cooperative Planning Systems

Philip J. Smith
Chuck Layton
Cognitive Systems Engineering Laboratory
The Ohio State University

Elaine McCoy
Dept. of Aviation
San Jose State University

This work has been supported by NASA Ames Research Center

Abstract

We are developing and evaluating design concepts and principles to guide in the building of cooperative problem-solving systems. In particular, we are studying the design of cooperative systems for enroute flight planning. Our investigation involves a three-stage process, modeling human performance in existing environments, building cognitive artifacts, and studying the performance of people working in collaboration with these artifacts. This report focuses on the most significant design concepts and principles we have identified thus far.

Introduction

Broadly speaking, we are studying design concepts and principles to guide in the development of cooperative problem-solving systems. The assumption behind this approach is that there are many important problems that computers alone cannot solve acceptably well. Hence, there is a need to study methods for enhancing cooperative performance by humans and computers.

More specifically, we are studying the design of cooperative planning systems to aid in the replanning of flights while enroute. We have selected enroute flight planning because it is an important practical problem for commercial aviation, and because it serves as an excellent testbed for studying the design of cooperative problem-solving systems.

Research Approach

There are many ways to gain insights into the design of aids for enroute flight planning. One approach is to study performance in existing environments. Surveys can be conducted to identify problems and critical incidents. The ASRS database can be explored. Accident investigation reports can be analyzed. Simulation studies can be conducted.

A second approach is to study the general literature on human planning and problem solving. Behaviors observed in other contexts, and models of the cognitive processes underlying these behaviors, may provide insights useful for understanding flight planning performance.

A third approach is to study the literatures on the design and use of computer aids. This includes reviewing the fields of artificial intelligence, knowledge-based systems, applied optimization and human-computer interaction.

The first three approaches are all very valuable. They help us to identify problems in enroute flight planning that need to be solved. They help us to understand the nature of the task and the task environment. They also help us to gain insights into the strengths and weaknesses of people performing planning tasks, and the availability of computer aids to assist with such tasks.

We have therefore made use of all three approaches. We have conducted surveys; we have run simulator studies in existing cockpit environments; we have reviewed relevant literatures. Appendix A contains reports discussing such efforts.

These approaches merely provide the foundation for our work, however. It is useful to know that people sometimes fixate on hypotheses they have generated; it is useful to know that the efficiency of computerized planning systems can be markedly increased by the use of abstraction hierarchies; it is useful to know that pilots currently play a major role in detecting the need to replan, and in generating alternative plans. In order to make significant progress, though, we need to synthesize such background information into an understanding of the problem area and to specify a set of design principles to guide in system development. We then need to apply these principles to develop specific implementations.

Finally, we need to empirically test these implementations and, indirectly, test the underlying design principles.

In order to progress toward such empirical evaluations, we have built a Flight Planning Testbed. Below, we describe characteristics of this testbed. Then we discuss the specific design concepts we are exploring, and the relationships of these design concepts to various aspects of the flight planning problem.

Development of a Prototyping Tool

Our research plan calls for a two-stage approach to testing design concepts. The first stage involves the use of a part-task simulation in order to develop design concepts and complete an initial evaluation. Those concepts that prove most promising based on this initial evaluation will then be used in the second stage of testing. This second stage will involve evaluation in the NASA Ames Advanced Concepts Simulator.

In order to run experiments using a part-task simulation, we had to design a suitable development environment. We consequently built a prototyping tool that can support the development and testing of a variety of design concepts.

This prototyping shell, designed to run on a Mac II, provides a general environment for developing application software, but does not prohibit programmers from modifying the environment if necessary. Written in Lightspeed C, the system can control displays on up to six color monitors.

This prototyping tool supports the creation and use of multiple window displays on each screen and the use of both mouse and keyboard inputs. The tool also provides both real-time and simulation-time clocks to control the timing of events and to record response times. The system records the time and nature of all actions made by a subject, and can replay the entirety of a subject's actions at a later time.

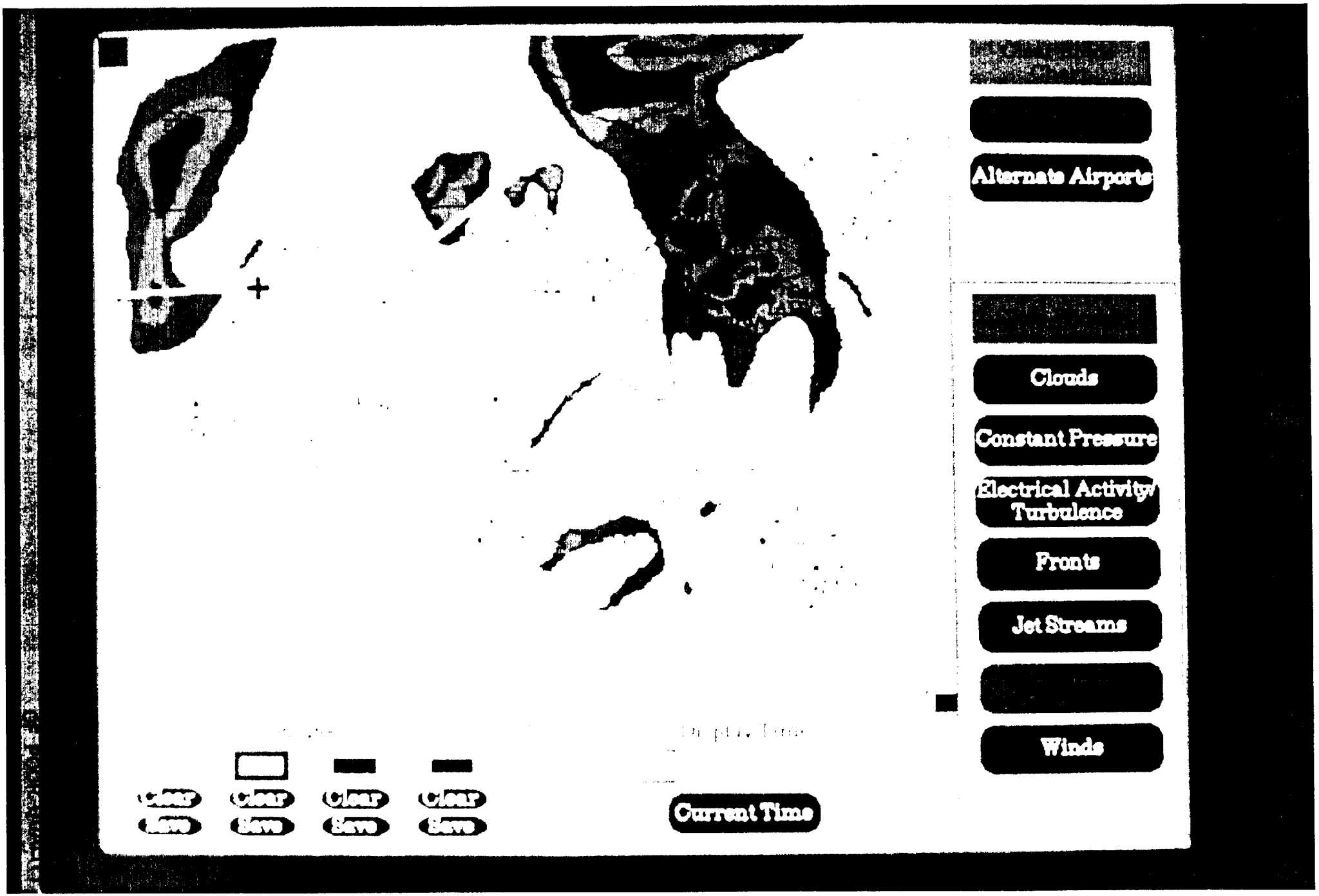
Development of a Flight Planning Testbed

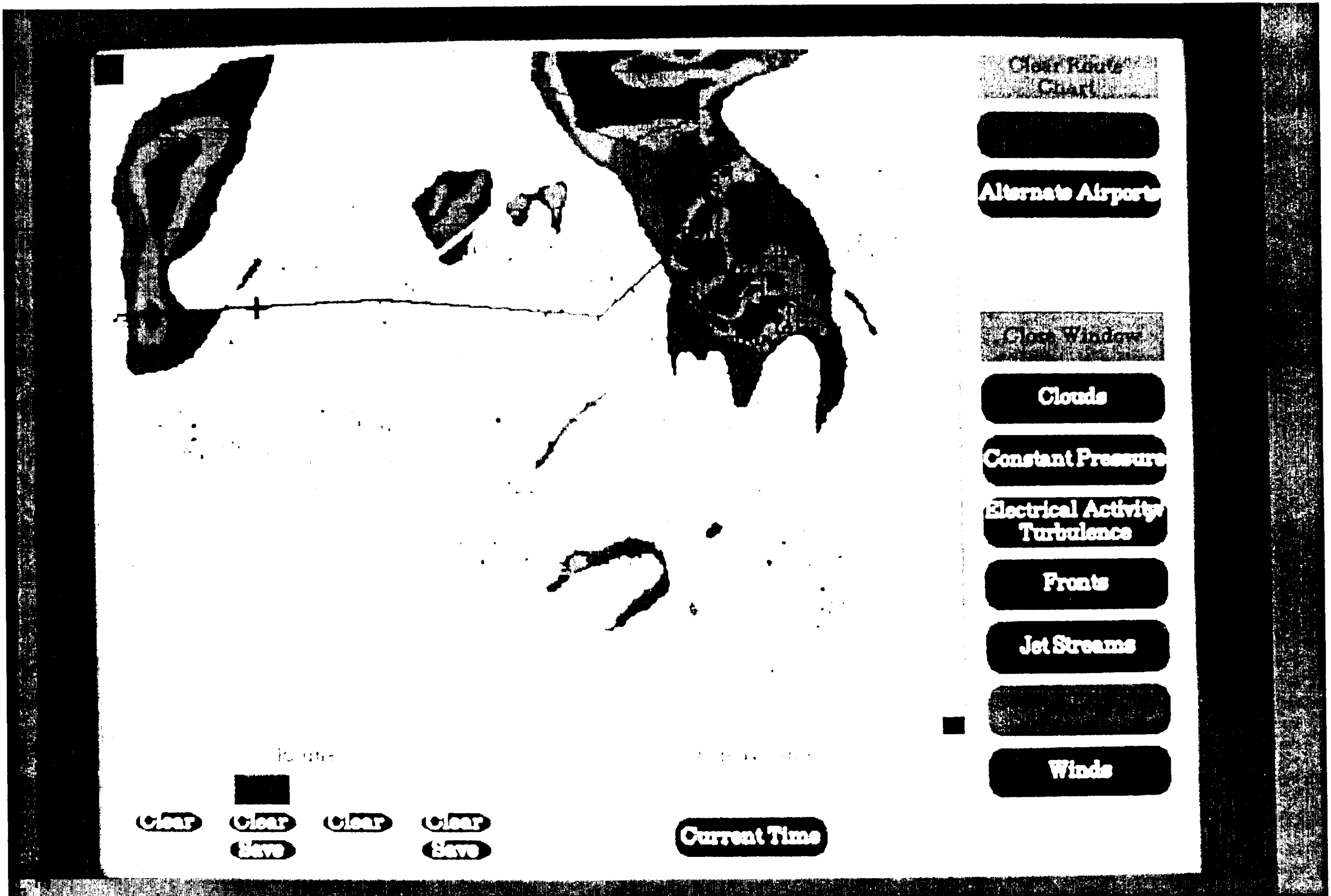
Using our prototyping tool, we have developed a Flight Planning Testbed that will provide the foundation capabilities for our studies. This testbed has several important features which are described below.

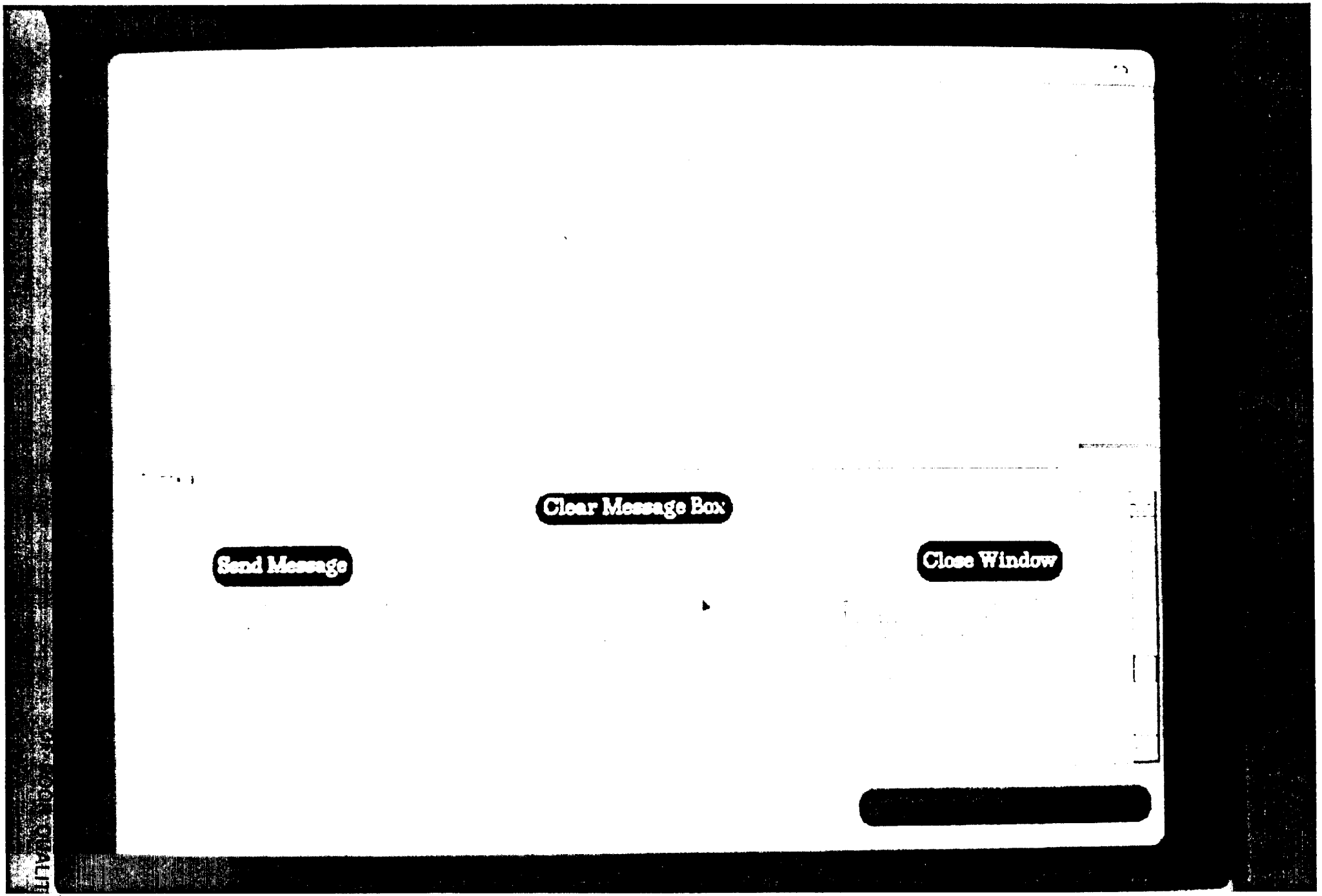
Map Display

Our testbed is capable of generating an accurate map display for any portion of the world. To accomplish this, we have ported to the Mac II a program (and associated database) that was developed using data from the U.S. Geological Survey. This program can produce accurate displays of any portion of the world, using any one of several available map projections.

Our testbed also allows for easy, rapid display of weather information on this map display. By simply pressing buttons with a mouse, the pilot can select a variety of weather overlays (radar weather, jet streams, fronts, etc.) to display on the map (see the second photo). In this manner, the

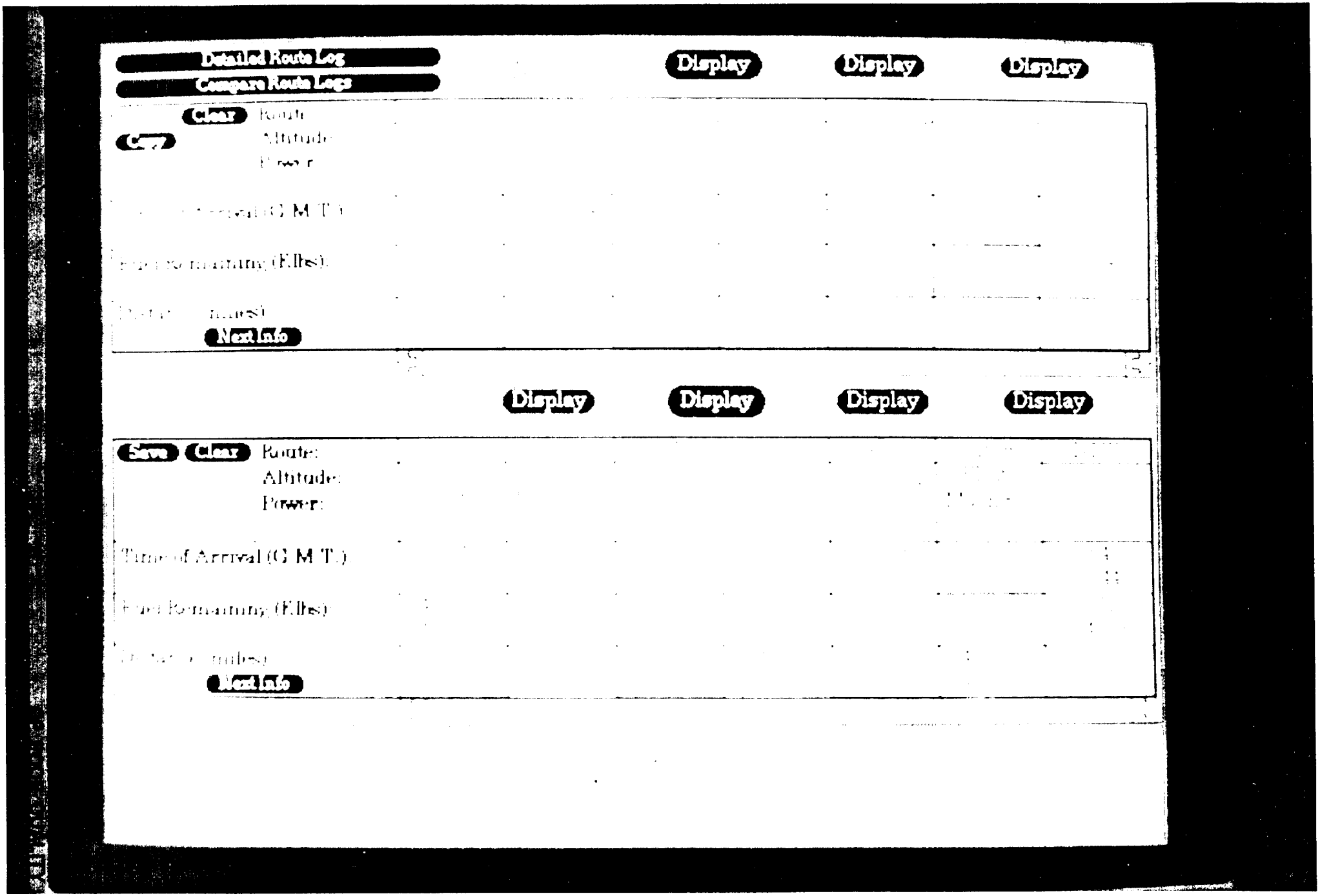

Alternate Airports
Clouds
Constant Pressure
Electrical Activity/ Turbulence
Fronts
Jet Streams
Winds
Current Time

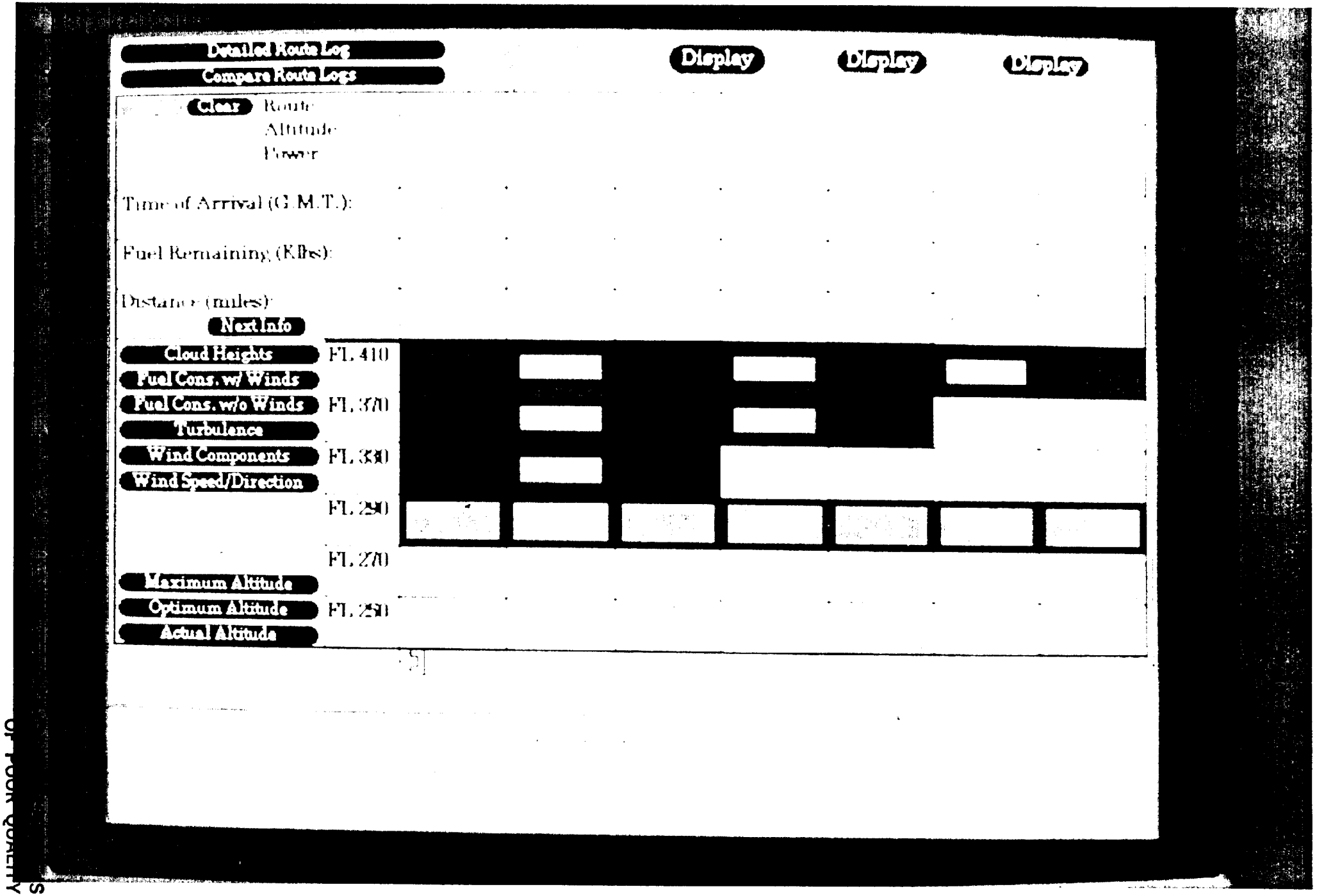

Clear Route Chart
Alternate Airports
Close Window
Clouds
Constant Pressure
Electrical Activity Turbulence
Fronts
Jet Streams
Winds
Clear
Clear
Save
Clear
Clear
Save
Current Time


Clear Message Box
Send Message
Close Window

Prepare Message for Dispatch

Airports


Detailed Route Log
Compare Route Logs
Display
Display
Display
Clear
Route:
Altitude:
Power:
Next Info
Display
Display
Display
Display
Save
Clear
Route:
Altitude:
Power:
Time of Arrival (G M T.)
Next Info



pilot can personalize the weather display to meet his/her current needs. Furthermore, by double-clicking with the mouse on any portion of the map display, the pilot can zoom in on the region, seeing a close-up display.

In order to facilitate viewing trend information, the pilot can also view weather sequences over time on the map display. This is accomplished by moving the plane along its route on the map. The plane is moved using a scroll bar controlled by the mouse.

The map display can also show radar weather information at different altitudes. (The National Weather Center has indicated that such data will be available nationally within the next five years.)

In addition to presenting weather information, the map display can show up to four alternative routes for the plane. It also displays the location of the plane on the active route. Both the plane's location and the weather displays are updated over time during the simulation.

Routes can be created or changed on the map display in two ways. One way is by direct manipulation of routes on the map itself using the mouse. With the mouse, the pilot can bend routes to deviate around some area. The pilot can also create new legs off an existing path. Finally, the pilot can create a totally new route (see the third picture).

A second way to create or change routes is described in the section on the Route Information Display. In that window, changes to routes can be made using the keyboard.

Information Alert Window

This Flight Planning Testbed also includes a window that can display important alerts to the pilot at appropriate times during the simulation.

Communications Window

The testbed has another window that provides a text editing environment for preparing and sending written messages to Dispatch (see the fourth photo). Routes drawn by the pilot on the Map Display could be transmitted to Dispatch along with this text.

Airport Information Window

This window displays both static information (number of runways, etc.) and changing information (weather, NOTAMS, etc.) about specific airports. The pilot can request such information by typing in the airport's identifier or by scrolling through an alphabetical list and selecting the airport with the mouse (see the fifth photo).

Route Information Display

The Map Display provides a graphic presentation of weather data. There are other types of information, however, that are probably better

displayed in a text format. We have developed a spreadsheet concept to present such information.

The sixth photo shows one of the spreadsheet displays available in our Testbed. Several important features are illustrated. First, the layout of data in the form of a spreadsheet seems well suited to this application. The horizontal sequence of information on the spreadsheet corresponds to the horizontal sequence of waypoints and jet routes along the flight path. Information specific to particular waypoints and jet routes is displayed under the column with the corresponding waypoint or jet route label.

Second, the spreadsheet allows the pilot to immediately view the implications of a change in the flight plan. The pilot can make changes in the plane's route on the spreadsheet by simply adding or deleting the appropriate waypoints. These changes in the route are immediately drawn on the Map Display. (Alternatively, the pilot can change the route by direct manipulation of the path shown on the Map Display. These changes are propagated to the spreadsheet.) The pilot can also make changes in the planned altitudes and power settings on the spreadsheet.

When a change is made in the flight plan, the system will appropriately change the other information displayed (such as arrival time and fuel consumption). The spreadsheet allows the pilot to view a variety of such information, such as wind components and distances between waypoints, as well as fuel consumption and arrival time information.

A third important feature of the display shown in the sixth photo is the ability to display information about two routes at the same time. This

facilitates comparisons of two alternative routes under consideration.

The seventh photo shows a second spreadsheet display that pilots can access. Using this display, the pilot can only look at a single route at a time, but can see information about that route in greater detail. In particular, this spreadsheet allows the pilot to easily compare different information about altitudes along his/her route. The pilot can display information such as turbulence, fuel consumption and wind components at these different altitudes. To facilitate such comparisons, the pilot can display the current altitude profile, optimal altitude profile and maximum altitudes. These kinds of information are displayed graphically within the spreadsheet itself.

Flight Planning Testbed - Summary

Using our prototyping tool, we have implemented a number of design concepts to support flight planning. The concepts were developed based on our cognitive task analysis of the performances of pilots in the simulation study (Galdes and Smith, 1990).

Design Concepts

In studying the design of aids for enroute flight planning, we have encountered a number of relevant design concepts that apply. These are discussed below. The value of such a list of concepts (and examples of their applications) is their ability to stimulate the thinking of system

designers. The designer must still consider his or her particular context in order to assess the validity of a particular design concept, and to generate ideas on how to apply it to his/her specific problem area. By considering such a list of concepts, however, the designer of some new system may come up with solutions that might otherwise be overlooked.

Concept 1. Use data abstractions to help planners deal effectively with large quantities of data.

In the near future, the amount of information that could be provided to the people responsible for enroute flight planning could be greatly increased. Data about passenger connections, flight crew schedules and air traffic congestion is already available for use. In addition, the technology exists to provide detailed, frequently updated weather information. Every plane in the sky is a potential weather sensor transmitting data about turbulence, winds, etc. to ground stations. (United and Northwest Airlines are already experimenting with this.) In addition, wind profiles, NexRad, ACARS and automated weather stations will be available to provide further detailed weather data.

Three questions arise:

1. What data should actually be provided to planners?
2. How should this data be displayed and utilized?
3. Who should have access to what data (ATC or Dispatch or the flight crew)?

In this section we deal with one answer to the second question.

Consider a system where an international turbulence map is available and updated regularly. The quantity of data to consider is huge.

Clearly, the planner for a particular flight can begin by focusing attention on the airspace along that flight's route. With up to 20 flight segments for longer flights, however, the number of relevant pieces of data is still very large.

We need some way to help the planner focus attention on potential problem areas, and on likely solutions. Our current design illustrates one solution, using a data abstraction.

Consider our detailed spreadsheet display. The spreadsheet displays turbulence reports for each of several altitudes along the route. It also displays (as a colored line) the planned altitude profile.

It would be impossible to display detailed turbulence data within such a compact display. Indeed, the pilots we have tested with our system indicate that, for just one individual flight segment, there could be considerable variation in turbulence levels at different points.

We could simply create a listing of all the turbulence information for all of the points along the route for all of the nearby altitudes. Instead, we are using our spreadsheet display to present an abstraction of this turbulence information. The label (light, moderate, etc.) in the spreadsheet cell indicates the maximum turbulence level along that segment at that altitude.

Imagine a planner who wants to ask:

> Am I likely to encounter significant turbulence in the next segment of my flight?

He/she can simply scan along the altitude profile as displayed in the spreadsheet and see whether any of the flight segments show significant turbulence. If, for instance, one segment indicates moderate turbulence, he/she can click on that cell, opening a window which describes in detail the nature and extent of the turbulence along that segment. (Graphics could also be included in this text window.)

Imagine this same planner asking:

> Can I avoid this moderate turbulence by changing altitude?

He/she can simply scan the spreadsheet cells, looking for an altitude corresponding to that flight segment that has less turbulence indicated.

Thus, our flight planning testbed will allow us to study the effectiveness of such data abstractions in helping the planner to detect potential problems in a timely manner, and to generate potential solutions. The above described form of abstraction can be applied to both turbulence data and wind components. An analogous form of data abstraction applies to the map display, where the planner can zoom in on a region and get more detailed information about weather and airport locations.

Concept 2. Allow direct manipulation of graphic displays to enhance exploration.

Our preliminary tests indicate that pilots are very enthusiastic about the ability to graphically create and manipulate routes. The ability to make

the changes directly on the map display appears to make it much easier to explore alternate routes to avoid bad weather.

Using our map display, the planner can also move the plan along the route and watch the forecast weather change. This helps the planner to assess trends in the weather and their potential impact on the flight. It also helps the planner to answer questions such as:

- Am I likely to encounter bad weather at my destination?

If the answer is affirmative, the pilot may want to add extra fuel (if this potential problem is noted before takeoff) or identify suitable alternate airports.

In our detailed spreadsheet display, the planner will also be able to manipulate the altitude profile graphically. The planner can edit the text display to change the altitude for a flight segment. He/she can also, however, simply drag the the altitude profile (displayed as a colored line) up or down in order to explore alternative altitudes.

Concept 3. Support planning and plan evaluation at many levels of detail.

Sacerdoti (1974) discusses the use of abstraction hierarchies to improve the efficiency of planning systems. Based on an analogy to this idea, we have developed a system where the human planner can develop plans at several levels of detail.

Flight planning is well characterized in terms of such an abstraction hierarchy. Imagine, for instance, a pilot flying from San Francisco to

Detroit who learns of a line of thunderstorms crossing his flight path over the Plains States. His primary decision is whether to deviate north or south of this storm. In order to evaluate this choice, however, it is necessary to specify additional details. Waypoints, altitudes and power settings must also be specified.

In order to support this goal, we propose a system where:

1. The pilot first sketches out a general solution (such as a northern deviation around the storm). This sketch is drawn on the Map Display;
2. By default, the computer fills in the lower level details, finding waypoints that approximate the pilot's sketch, finding an "optimal" altitude profile for this path and finding suitable power settings;
3. The pilot then evaluates the details of this solution by looking at the spreadsheet displaying route information. If he chooses to, he can alter the computer's recommendations for the lower level details and compare his choices with the computer's. (He may, for instance, note that the computer's "optimal" altitude profile flies the plane through areas with unacceptable turbulence.)

Consider another situation where the pilot encounters turbulence. He/she wants to decide whether to go higher or lower. Using the spreadsheet display, he/she can directly generate and evaluate alternative altitudes.

Thus, we have designed a system where:

1. Displays exist corresponding to different levels of detail in the planning hierarchy;
2. The pilot can view and make changes on any of these displays. He/she can therefore make changes at any level of detail desired. He/she can also, therefore, look at the data needed to evaluate

decisions at that level of detail;

3. The computer, by default, handles lower levels of details. The pilot can, however, compare the computer's recommendations with his/her own ideas and make changes as desired at any level of detail.

Thus, using this architecture, the pilot can easily explore "what if" questions at any level of detail desired.

Concept 4. Create a microworld in which the planner can actively explore "what-if" questions and get useful feedback to help in evaluating alternative plans.

In our simulation studies and interviews pilots repeatedly express the desire to ask "what-if" questions of the computer. They generally feel that it's fine for the computer to suggest a plan and indicate the predicted fuel consumption and arrival time, but that it is also extremely important to give them the ability to explore alternatives. Our testbed environment provides such an environment.

The planner can change any of the flight parameters (route segments, altitudes or power settings). The system fills in the lower level details (see Concept 3) and estimates fuel consumption and arrival times for this new flight path.

Concept 5. Support a variety of planning "models" to accommodate different situations and people.

In our simulator studies, we observed several different planning "models" in use. An effective cooperative system should probably accommodate all

four of these "models."

Planning Model 1. The most common cause of flight amendments is some localized disturbance that makes the plane's original flight plan undesirable or impossible. Typical causes include:

1. the development of areas of turbulence;
2. the unexpected formation of localized storms;
3. changes in winds at different altitudes;
4. the appearance of other air traffic that prevents planned altitude changes.

Example 1. ATC informs the crew that they will have to stay at flight level 250 longer than planned (instead of going to FL270) because of other air traffic. The crew considers the effects of this change on fuel consumption and arrival times. The effects are significant but acceptable if no alternatives are available. The crew generates one minor modification to try to make up some of the fuel consumption. They ask ATC whether, once cleared to FL270, they can then proceed up to FL330 earlier than planned. This will now be possible because of the extra fuel consumption, which will lighten the plane ahead of the planned schedule.

Example 2. The flight crew notes that they are behind schedule and burning up more fuel than expected under the original plan. They conclude that the problem is a headwind that is stronger than expected under their original plan. The crew asks ATC whether there are any reports on winds at other altitudes. They learn that the headwinds are now favorable at lower altitudes. They compare the tradeoff between the benefits of the lower headwinds and the cost of flying at the lower altitude, and decide it is preferable to fly at a lower altitude. They request clearance from ATC

to do so.

Example 3. The flight crew encounters light to moderate turbulence. They consider changing altitudes to avoid it, or slowing the plane to reduce its effects. They check for pilot reports on the likely duration and magnitude of the turbulence at that altitude, and on turbulence levels at other altitudes. The turbulence is reported to be very localized, so they decide to ride it out.

Example 4. The onboard radar indicates the presence of a thunderstorm along the current path. The crew requests information about tops, and concludes the storm is too high to climb over. Since the storm is not too extensive, they ask ATC to simply vector them south of the storm.

Planning Behavior. Our data indicate that, currently, flight crews generally respond to such localized disturbances by generating solutions that are minor modifications of the original plan. In most cases, the crew doesn't replan the entire remainder of the flight, they simply select an immediate response to the local problem and act on it. They assume that they will be able to find additional minor modifications for the remainder of the flight when the need arises (Suchman, 1987).

Model 1 - Discussion. Three points merit discussion. First, under these circumstances, plans are generated by attempting to make minor modifications to the original flight plan. It is assumed that, because the modifications are small, the potential implications for later in the flight do not have to be considered in detail. It is assumed that any later

modifications made necessary by the current change will again be minor, and that acceptable modifications will be possible.

A second point is that such planning is very decentralized. ATC looks at the local implications in terms of air traffic, but other than that, no one evaluates the effects of the requested amendments on the overall system. No one says "there's been a disturbance, let's now replan everyone's flight" to ensure "optimal" or good overall system performance.

This decentralized approach to planning makes strong assumptions about the "world." It assumes that the flight plans of different planes are not tightly coupled. Small changes in one plane's plan do not result in significant disruptions of other plane's plans, or of overall system performance. It also assumes that the "world" generally allows a variety of small changes to be made. Consequently, it is unnecessary to anticipate the availability of future modifications that will be made necessary by the current minor modification. It is assumed that some acceptable modification will always be available to meet future needs.

The third point is that, at present, such localized planning is accomplished in one of two ways. The first method can be characterized as a simple forward search with a short planning horizon. The pilot looks at the immediately available alternatives (changes in altitude, vectoring around the storm or turbulence, slowing down to reduce the effects of turbulence, etc.) and picks the one that seems to best solve his/her immediate problem. The second method is somewhat analogous to case-based reasoning (Riesbeck and Schank, 1987), except that the pilots access a broader "institutional" memory. They ask ATC whether any other pilots

have already found a solution to the immediate problem and then make use of that solution.

Our present design currently supports such decentralized, localized planning. The planner can use the map display to find a set of waypoints that take the plane around a storm. The planner can also view the detailed spreadsheet and look at fuel consumption, winds and turbulence for the next flight segment in order to decide whether to change altitude. It would also be possible to support the case-based reasoning solution by providing the planner with access to already tried localized solutions that have been successful.

Planning Model 2. Under Planning Model 1, the planner doesn't worry too much about a complete path to his/her destination. He/she simply finds an amendment that solves the immediate problem and assumes that the remainder of the solution can be worked out when the time comes. We also saw cases where the pilots in our simulator study worked out the entire flight plan after proposing an amendment.

In such cases, planning was again very decentralized. No one asked: What's best for the whole system? ATC did, to some extent, look at the interactions among planes and put constraints on the solutions. The flight crew simply searched for a solution for their own plane alone that met these constraints.

There are several ways in which a flight planning aid could support such planning. The first would be to provide the raw data and calculations (winds, turbulence, fuel consumption, etc.) necessary for the (human)

planner to work out a complete solution using forward search methods. The second would again mimic case-based reasoning approaches, borrowing from already generated solutions used by other planes.

The third approach mimics current human-to-human interactions. In our simulation studies, we sometimes saw pilots develop fairly abstract plans and then let ATC or Dispatch work out the details. They would say things like:

> "Can you find us a route north of this storm?" or
> "We need a new destination airport."

By supporting planning at different levels of abstraction, our testbed mimics some aspects of this human-to-human interaction. Additional features worth considering based on this model, however, include allowing the (human) planner to specify a goal or constraint (such as "find a route that gets me to my destination within 10 minutes of my scheduled arrival time" or "find me an alternate destination" or "find a good airport that I can reach within 30 minutes" or "find an airport that I can reach and still have adequate holding fuel."

Planning Model 2 - Discussion. Planning Model 2 has two important characteristics. First, like Planning Model 1, the planner doesn't worry (too much) about finding global solutions that lead to good overall solutions for all of the air traffic. Second, unlike Planning Model 1, the planner works out the entire remainder of the flight. He/she uses a much longer planning horizon.

Finally, as discussed above, our simulation data suggests that pilots currently use a variety of solutions to generate such plans. They use

forward search methods; they use case-based reasoning; they plan at higher levels of abstraction and then offload planning to another agent by merely specifying a goal or constraint. All of these methods have potentially important implications for building computer aids.

Planning Model 3. Planning Models 1 and 2 involved looking for solutions from a decentralized perspective. The planner (the flight crew in this case) looked for a plan that was good for him/her without directly considering whether that plan was good from a global perspective. (The global perspective was still partially considered by ATC when deciding whether to approve a requested change in altitude, etc.)

A third planning model that we have seen in use involves explicitly considering the bigger picture. · Such planning is currently done by ATC and Dispatch. This model is typically invoked when there is some large, systemic disturbance (a line of thunderstorms, airport closings, etc.). In such a case, ATC and Dispatch look for broader solutions that consider the overall implications for all of the air traffic (or at least that airline's). At present, this global planning involves both elements of cooperation and competition. Dispatch would like to get the best solutions for his/her airline. ATC would like to get good overall solutions.

From the flight crew's perspective, such planning often takes the form of case-based reasoning. The crew is informed that ATC has developed a preferred alternate plan for planes along that path, or that Dispatch has a recommendation. The crew then evaluates this plan to ensure that it is acceptable to them.

Concept 5 - Discussion. Above, we describe a variety of planning "models" and methods that we have observed in use under current circumstances. These observations are of considerable importance, as it is likely that an effective cooperative system should support such alternative "models" and planning methods.

Concept 6. Use graphics to enhance perceptual processes, helping the planner to "see" the important patterns instead of making him/her laboriously "reason" about the data in order to infer their presence.

The attention literature makes a distinction between automatic recognition processes and controlled processes. Larkin and Simon (1987) suggest this concept can be fruitfully applied to designing aids for problem-solving.

The most interesting application of this concept to flight planning is with the map display. By allowing the planner to view the plane moving along its route, viewing concomitant changes in the weather, the planner may find it much easier to judge trends and note important patterns.

The detailed spreadsheet illustrates another simple application of this concept. By embedding graphics identifying the current flight plan, "optimal" plan and maximum altitudes into the spreadsheet, it should be much easier for the planner to identify pertinent data and make comparisons at different altitudes. We may also embed cloud TOPS into the spreadsheet at some point.

Concept 7. When using graphics, provide a "natural" mapping between the features of the display and the corresponding concepts or real-world objects.

The map display is an obvious application of this concept. The detailed spreadsheet is also consistent with it, however. The spreadsheet depicts the horizontal movement along jetways as a horizontal sequence of cells on the spreadsheet. Each successive column represents the next waypoint or jet route in sequence. (An interesting conflict arises, though, when the plane is flying east to west. Should the sequence on the spreadsheet now go from right to left to be consistent with the orientation of the map display?)

The altitude information at the bottom of the spreadsheet is also consistent with this principle. Higher altitudes for a flight segment are represented as higher cells in the spreadsheet.

There is also another inconsistency with this principle. The length of flight segments is not reflected at all in the graphics on the detailed spreadsheet. All spreadsheet columns are equally wide, even though the flight segments they represent differ in length. We have experimented with displays where segment lengths were drawn to scale. Segment lengths differ greatly, however, and our preliminary judgment was that it would be better to tradeoff in favor of compactness of the display (allowing the planner to see more flight segments at a time) rather than having pictorial realism.

Concept 8. Consider distributing the problem-solving to simplify the tasks for individual participants.

At present, there are several parties involved in flight planning. The flight crew plays a major role in detecting problems that require replanning. The flight crew also does much of the replanning. ATC sometimes generates some of the details of a plan, but often ATC plays a reactive role, telling the flight crew whether an amendment they have proposed is feasible given other air traffic. Similarly, Dispatch often plays a reactive role, relying on the flight crew to detect a problem and to suggest a solution.

These roles depend very much on the time-criticality of the problem and its nature. Dispatch is more likely to play a major role in selecting an alternative destination, for instance, than in proposing a change in altitude to avoid turbulence.

It is clear, though, that there is a decomposition of flight planning activities that allows different parties to deal with different aspects of the flight planning problem. Such task decompositions need to be considered when deciding who should have access to what information and computer aids.

Concept 9. Consider including redundancy in a distributed problem-solving environment to increase the likelihood that good solutions will not be overlooked and that bad solutions will not be accepted.

In addition to reducing the cognitive load by distributing tasks among

different parties, such shared problem-solving may benefit from intentional or chance occurrences of redundancy. Dispatch, for example, may notice that a flight amendment proposed by the flight crew leaves very little holding fuel and recommend finding an alternative plan.

In designing the planning environment, we may want to use computers and advanced communication capabilities to enhance such intended and incidental redundancy. There may be data and information that we want to deliberately present to multiple parties. This may include presenting the computer's conclusions, explorations and warnings to the flight crew and Dispatch (and in some cases, to ATC as well).

The literature on human error discusses such things as the generation of false assumptions (Smith, Giffin, Rockwell and Thomas, 1986), and fixations on incorrect hypotheses or unwise solutions. In our simulation study we saw one example of such behavior. One crew appeared to fixate on Toledo as an alternate destination after Detroit was closed. Initially, it appeared to be a reasonable alternative, but given the questionable weather in the area and the progressively lower fuel levels, it was a very dubious choice to commit to while over Gopher. The crew never asked: Do we have enough fuel to go elsewhere if the weather at Toledo turns bad (or if air traffic congestion develops)?

Similarly, we saw several cases where flight crews failed to consider the implications of certain events (being held at a lower than planned altitude) or actions (flying faster than normal cruise speeds). Appropriate aids to enhance distributed problem-solving might help reduce such "errors."

Concept 10. Design assuming that novel situations will arise that will make invalid certain inferences and conclusions made by the computer system.

It is clear that knowledge-based systems and optimization programs have limited scope. It is quite probable that situations will arise that were not anticipated by the system designers.

One solution is to provide the computer system with explicit error detectors (Smith, Smith, Svirbely, Miller, Glades, Fraser, Blazina and Kennedy, in press) and with metaknowledge. To the extent that the computer knows what it does and doesn't know, it will be better able to detect situations where it is "over its head."

This solution simply reduces the likelihood that the computer will unknowingly generate a questionable plan. There is still the likelihood that the system designers will leave out important metaknowledge to detect some novel situations.

A second solution, therefore, is to keep people actively engaged in the planning activities, and to attempt to ensure that they consider important data as well as recommendations by the computer (or another person). This requires careful consideration of the roles of various agents (human and computer) as well as the design and distribution of data displays.

Concept 11. Let the computer actively monitor the data and provide alerts when important changes are detected.

Planners sometimes make assumptions that are invalid and consequently fail to check for important information. One pilot we interviewed provided an example of such behavior:

> A flight into Cleveland had Columbus as the scheduled alternate airport. As they approached Cleveland, the weather progressively worsened. Cleveland was shut down before they arrived there. The crew therefore changed course to Columbus. As they approached Columbus, they talked to ATC and announced their intention to land on a particular runway. ATC responded that this runway was under construction and was only half its normal length. ATC gave them an alternative which required circling and approaching from a different direction. The crew responded that, due to the bad weather they had encountered and the diversion from Cleveland, they did not have enough fuel to circle and land on the alternative runway. They had to make an emergency landing on the shortened runway (and did so successfully).

If they had checked the available NOTAMS for Columbus, the crew would have learned of the runway construction. They failed to do so, however, assuming that since this was their scheduled alternate, there were no problems. This failure, compounded by the excessive fuel consumption, led to the emergency.

A partial solution to such problems is to develop more intelligent alerting capabilities for the computer. Such alerts need to be as context sensitive as possible. Their goal should be to focus the planner's attention on important data or problems that might otherwise be unnoticed.

Concept 12. Be sure there is a clear, easy to understand conceptual model for controlling and understanding the computer's processing.

Lehner (1987) suggests that computers need not think exactly like their human partners. The human simply needs to understand how the computer is reasoning so that he/she can assess the appropriateness of its recommendations.

Current flight planning systems often use optimization techniques to generate recommendations (Sorensen, 1982). These have a major problem:

> What does it mean to the human planner when the computer states that it has found the "best" plan to optimize a particular objective function that weights fuel consumption, arrival time and turbulence?

It may be possible to reduce some of these problems by placing a knowledge-based system (a "translator") between the person and the optimization program. The knowledge-based system could accept inputs such as "find a plan that gets me to my destination by 0400," and then translate this into an appropriate constraint on the optimization. This solution reduces, but does not necessarily eliminate, problems of cognitive compatibility. The person still has to understand how the computer arrives at its solution. (Does it, for instance, consider turbulence and, if so, how?). This understanding must include a recognition by the user of the limitations or brittleness of the computer's algorithms.

Concept 13. Try to predict the errors that components of the system, individually or jointly, could make. Try to design the overall system to prevent errors. Equally important, try to design the system so that errors (including those that haven't been predicted) are likely to be caught, or failing that, so that their impacts are not serious.

In our interviews and in our simulator studies, the most serious situations seem to result from a combination of three factors:

1. Using a short planning-horizon to solve some immediate problem (thus failing to consider long-run implications);
2. Failing to discard the current plan early enough, while there are still many alternative options available;
3. Experiencing the occurrence of a series of events that, taken together, seriously threaten the plane's safety, even though each one alone would normally be a minor problem.

Under Concept 5, we describe a model of planning in which planning is very localized, in which the pilot finds a solution to the immediate problem without considering in detail the implications for later in the flight. This form of planning assumes a "friendly" world, where there are numerous alternatives to select from to solve the next step in developing a plan. Under such an assumption, there is no great need to look beyond solving the immediate problem.

In flight planning, the assumption of a "friendly" world is normally quite viable. The plane has reserve fuel, keeping many options open. The plane

can land somewhere else if fuel, weather, etc. make this necessary. Finally, the pilot can request priority clearances if the situation is becoming sufficiently difficult, thus gaining additional options.

Occasionally, however, the flight crew finds itself in a less "friendly" world. Based on our interviews, this seems to arise for one of two reasons:

1. The plane encounters a series of problems that require flight amendments and use up extra fuel. The solution to each problem taken alone is quite reasonable, but, taken together, fuel levels get unacceptably low. Thus, by failing to consider a longer planning horizon, and by failing to anticipate potential "worst case" possibilities, the crew ends up in a situation where they have few good options left;
2. The crew "fixates" on their current plan too long, failing to notice that their other options are disappearing (due to low fuel). If the "worst case" arises and they can't complete their current plan, they are in a difficult situation.

Plan Generation. Generally, a less serious error that occurs is the failure of a crew to consider all of the alternatives available. In our simulator studies, we frequently saw cases where crews found one solution that was satisfactory and failed to even consider others that might have proved superior.

Other Sources of Error. Other problems may arise due to "slips," failures to detect problems in a timely fashion and overreliance or incorrect reliance on the computer.

Solutions. One solution would be to make the world "friendlier." The obvious (but expensive) way to accomplish this would be to require

greater fuel reserves. A second would be to develop computer aids that help the planner to use a longer planning horizon and to anticipate possible "worst case" situations. A third would be to develop aids that monitor the situation and warn the planner when the number of options is becoming dangerously low. A fourth would be to facilitate distributed planning on the assumption that Dispatch, for example, might be less likely to share a fixation that the crew has developed (or vice versa).

To assist in problem detection and plan generation, the computer could use both passive and active means. Displays can be designed that provide help in detecting problems in a timely fashion are suggestive of alternative plans that the crew might otherwise overlook. (Our detailed spreadsheet is an example of this.) As an active assistant, the computer could also monitor for problems and search for additional solutions, calling them to the attention of the planners. Enhancing distributed planning would again be an additional approach to improving performance.

"Slips" may be reduced by better design of the system, and their effects perhaps reduced by asking the computer to critique performance.

Finally, the issue of overreliance or incorrect reliance on the computer is an interesting one. One concern is the ability of the human planner to understand what the computer is and is not considering, and to decide appropriately whether to trust the computer's analysis. A second concern is whether, even if the planner understands the functioning, he/she will detect situations where the computer is operating outside of its intended scope.

The studies we are planning with our testbed and with the Advanced Cab should provide data to better understand the causes of such errors and the effectiveness of alternative solutions.

Concept 14. Facilitate communication and cooperation by designing a system that can infer the planner's current goals.

In selecting a flight amendment to deal with some problem, the solution space to be searched is often quite large. If the computer can determine what the planner is trying to accomplish, it can begin this search on its own.

One example of such aiding involves avoiding bad weather. Assume the planner sketches a solution on the map display that creates a route south of some storm activity. Given this input, the computer can infer the planner's goal and automatically begin searching for alternative solutions (e.g., going north of the storm, or flying above the storm). If a promising alternative solution is found, this can be displayed to the planner for consideration. (The planner might also want to view some (organized) listing of the plans the computer has considered.)

Concept 15. Consider appropriate functional groupings to facilitate important comparisons.

As part of our cognitive task analysis, we identified a variety of data and information that pilots wanted to view, and comparisons they wanted to make. In the spreadsheet display we have attempted to group information to facilitate such comparisons, supporting questions such as:

1. Which plan will use up more fuel?
2. What is my ETA at Denver?
3. What is the jet route we will take between St. Louis and Cincinatti?

Concept 16. Allow the planner to tailor the information and data displays to fit his/her needs and preferences.

On the one hand, it is plausible that different planners may prefer to display different information at different times. This may increase their acceptance of the computer aids and improve their ability to deal with different circumstances.

On the other hand, the literature indicates that people don't always know what the best design is, even for themselves. Thus, we must be cautious in deciding what flexibility to provide to the system user.

Our testbed currently has one simple application of this concept. The planner can select the combination of weather information that he/she prefers to display at any given time.

Summary

Our Flight Planning Testbed allows us to explore a wide variety of design concepts and principles that are important to the design of computer aids in general, and to the design of aids for enroute flight planning in particular. Above, we have described the prototyping environment we have

developed, and some specific implementations we have designed.

As discussed above, our studies of flight planning activities (through simulations and interviews), and our work in implementing specific concepts, have been very fruitful in identifying some very interesting design concepts. Our preliminary tests of the system, for instance, indicate that pilots view both the types of information provided and the form of interaction to be very useful:

> "Real-time [weather] is absolutely worthwhile....The left screen [would be] invaluable. [We] now lack big picture [in existing systems]"
>
> "[The left screen would be] invaluable in the cockpit. Absolutely."
>
> "This is much easier to learn. [Pilots] could learn [this system] faster with less training."

In particular, *having a functional system has proved to be a very valuable aid to stimulate ideas and discussion.* This implementation should be equally valuable in our next stage, in which we will empirically test these design concepts and develop more detailed cognitive models.

References

Ericsson, K. and Simon, H. (1984). Verbal Protocol Analysis. MIT Press: Cambridge, MA.

Galdes, D. and Smith, P. J. (1990). A Cognitive Analysis of Enroute Flight Planning Activities by Flight Crews. NASA Technical Report (in Press).

Larkin, J. and Simon, H. (1987). Why a diagram is (sometimes) worth ten thousand words. Cognitive Science, 11, 65-99.

Lehner, P. and Zirk, D. (1987). Cognitive factors in user/expert-system interaction. Human Factors, 20, 97-109.

Reisbeck, C. and Schank, R. (1989). Inside Case-Based Reasoning. Hillsdale, NJ: Erlbaum.

Sacerdoti, E. (1974). Planning in a hierarchy of abstraction spaces. Artificial Intelligence, 5, 115-135.

Smith, P. J., Smith, J., Svirbely, J., Galdes, D., Fraser, J., Rudmann, S., Thomas, D., Miller, T., Blazina, J. and Kennedy, M. (in press). Coping with the complexities of multiple-solution problems: A case study. International Journal of Man-Machine Studies.

Smith, P. J., Giffin, W., Rockwell, T. and Thomas, M. (1988). Modeling fault diagnosis as the activation and use of a frame system. Human Factors, 28(6), 703-716.

Sorensen, J., Waters, M. and Patmore, L. (1983). Computer Programs for Generation and Evaluation of Vertical Flight Patterns. NASA Report #3688.

Suchman, L. (1987). Plans and Situated Actions: The Problem of Human-Machine Communication. Cambridge, England: Cambridge University Press.

Wilensky, R. (1983). Planning and Understanding. Addison-Wesley: London.

Appendix A

Related Publications

A CATALOG OF ERRORS

Jane M. Fraser *
Philip J. Smith *
Jack W. Smith, Jr. **

* Department of Industrial and Systems Engineering, The Ohio State University, 1971 Neil Avenue, Columbus, Ohio, 43210.
** Department of Pathology, The Ohio State University.

Acknowledgements: This publication was supported in part by grant NCC2-165 from NASA Ames Research Center and grant HL-38776 from the National Heart, Lung and Blood Institute.

Abstract

This paper reviews various errors that have been described by comparing human behavior to the norms of probability, causal connection, and logical deduction. For each error we review evidence on whether the error has been demonstrated to occur. For many errors, the occurrence of a bias has not been demonstrated; for others errors a bias does occur, but arguments can be made that the bias is not an error. Based on the conclusions of this review, researchers and practitioners are cautioned in referring to well known biases and errors.

Introduction

Whenever human performance is compared to a standard, the notion of error arises. A simple list of the errors that humans make is not useful, but a coherent categorization of errors occurring in different domains, under different conditions, and in different forms is helpful.

A well-known categorization of errors is that of Tversky and Kahneman (1974); their categorization involves two levels. First, they create categories of errors on the basis of obvious groupings; for example, they classify together errors that involve ignoring base rates. Second, they create categories by grouping together errors that they argue are caused by the use of the same heuristic. For example, Tversky and Kahneman attribute the errors involving ignoring base rate, as well as other errors, to the operation of the representativeness heuristic; they attribute the occurrence of yet other errors to other heuristics.

This paper is a review of the literature on specific categories of cognitive errors. The categories are at the first level of Kahneman and Tversky. Thus, this paper is a catalog of errors, not a catalog of explanations of errors.

Scope. The errors included in this catalog were selected because they have received widespread attention by researchers and have been widely labeled as errors. The articles reviewed under each error were selected to portray the history of research and the range of thought regarding the error. Where possible the labels used in the original publication were used to determine to which category of error a behavior was assigned.

This is not an exhaustive literature review. References are intended to include the earliest source that labeled the error, studies that indicate the development of thought about the error, and studies that indicate the agreements and disagreements regarding the error. Since this is a catalog of errors, not a catalog of explanations of errors, discussions of the cognitive processes causing the error are omitted.

Errors due to lack of domain knowledge by a subject and errors that are specific to a domain of discourse are omitted. Studies are included that use adult subjects, whether novices or experts in the domain in which the error is being tested, but not studies using only children as subjects.

Motivation. The study of human error has long been of interest to researchers and practitioners. Our own interest arises from the desire the build computer systems that critique practitioners or that teach students. Such systems must be able to detect and diagnose errors and to provide effective remediation. In particular, the catalog is being used to predict and classify errors made by immunohematologists (laboratory technicians who analyze a patient's blood in order to provide matching blood for a transfusion) and

flight crews of commercial aircraft.

This paper is also useful as a history of ideas in one area of cognitive research. The errors in this catalog share an underlying structure, especially in the norms against which error is defined and in the development of ideas concerning the various errors.

Historical trends. Typically, a particular type of behavior is identified by comparison to a norm. Many of the errors in this paper are labeled by comparison with three norms based on treating the subject as a naive scientist: norms based on probabilistic reasoning, norms arising from the findings of science regarding causality, and norms developed from deductive logic. An example of an error in probabilistic reasoning is conservatism in revision of beliefs in the face of evidence. The fundamental attribution error is an example of an error of causal reasoning since, contrary to scientific findings, subjects overemphasize the role of personality as compared to the role of environment in causing a person's behavior. An example of an error in deductive logic is the failure of subjects to name the correct cases necessary for verification of a particular "if P, then Q" rule.

Many of these errors share a similar history. In the course of studying some task, a researcher discovers that some subjects use a simplifying strategy. The strategy is only a heuristic, and therefore it will occasionally lead to errors. Noting this, the researcher devises a task on which the heuristic is likely to lead people astray and performs experiments which demonstrate that a significant proportion of subjects do exhibit the error on the task. Other researchers attempt to replicate the experiments and to perform similar but slightly different experiments. Different results are observed by different researchers; some successfully replicate the original results, while others find contradictory results. A dispute follows concerning whether the error actually occurs; the debate usually changes into efforts to describe the conditions under which the error occurs and the conditions under which it does not occur. Sometimes researchers show that, while the error is common under laboratory conditions, under realistic conditions the error occurs infrequently. Some researchers argue that the heuristic is robust, that is, it is likely to give good results in natural situations. Some researchers seek to discover what kind of training is necessary to eliminate the error. Other researchers mount larger objections by objecting to labeling the subjects' behavior as an error. Some argue that subjects misinterpret the task because of linguistic ambiguities in instructions or because of expectations arising from the laboratory situation. Others argue that subjects use proper probability reasoning, use proper causal arguments, or are properly logical, but frame the problem differently from the way intended by the researcher. In a more fundamental attack, others argue that subjects use a logic or use concepts of probabilistic reasoning or causality different from the norms of a scientist but still worthy of being called rational.

Overview. Rarely is there a final consensus on any of these errors. Many of the disputes mentioned in this generic history are still in process concerning some of the errors. Indeed, even the fact of occurrence of some of the errors is still in dispute.

We begin with the errors that are labeled by comparison with the norms of probability. Since information received in real-world situations is often insufficient to reach a conclusive statement, people must be able to judge the proper weight to give to their knowledge and to new information in arriving at degrees of belief. The first error discussed is that people give *judgmental probabilities that do not equal correct combinatorial probabilities or correct empirical frequencies.* This error can be viewed as evidence that people's internal degrees of belief as expressed on a probability scale do not represent correct beliefs. More such evidence is provided by the *miscalibration* of many people's probability estimates.

Other errors in probability are exhibited by noting that subjects' responses on several questions do not have the numerical relationship to each other that they should have. These errors mean that a subject's judgments are internally inconsistent. These errors include *insensitivity to sample size, committing conjunctive and disjunctive probability errors, incorrectly revising probabilities, ignoring base rates,* and *exhibiting hindsight bias.*

Next we discuss errors that occur when people must judge the causal connection or the co-occurrence of two events. Several attribution errors are described, including the *fundamental attribution error* and various *egocentric biases.* Two errors regarding probabilistic dependencies are included: *illusory correlation* and *reliance only on positive hits.* The final errors in this group are *overprediction* and the *illusion of control.*

Four errors involving logical deduction are covered. *Incorrect testing of a conditional rule* and the *confirmation bias* both involve the incorrect selection by subjects of items with which to test a theory. *Differential treatment of positive and negative information* involves the inability to use negative information as efficiently as positive information. *Hypothesis fixation* involves continuing to hold a hypothesis which has been disproved.

This review focuses on whether each error has been shown to occur. The conclusion is that, for many of the errors, there is considerable doubt about whether the behavior occurs frequently. Furthermore, even if the behavior has been shown to occur, there is considerable disagreement over whether it should be labeled an error. Therefore, researchers and practitioners should be cautious in referring to certain biases as if their existence is well established or in referring to those biases as errors.

Judgmental probabilities don't equal corresponding empirical frequencies

Definition. People's probability judgments differ from the corresponding combinatorial probabilities or from the corresponding empirical frequencies.

In a variety of experiments, subjects have been asked to express their beliefs regarding

the likelihood of events in the form of probabilities. The experimenter provides no information or evidence to the subject. The quality of the subject's answer thus depends on his knowledge and on his ability to express his belief in the form of a probability. Generally, it is easy to show that the probabilities given by subjects often do not match beliefs that should be arrived at through combinatorial calculations or through observation of evidence.

The birthday problem is often used to show that intuition regarding combinatorial probabilities is poor. Among a group of only 23 people, the odds *favor* the occurrence of identical birthdays, a result counter to most people's intuition. Similarly, people generally believe that more 2-member committees than 8-member committees can be formed from 10 people, while the correct answer is that the same number of such committees can be formed (Kahneman and Tversky, 1974).

People also give probabilities that disagree with data on past occurrences. Extensive comparisons have been made between people's judgments of the number of deaths per year from various sources and public health statistics. Slovic, Fischhoff, and Lichtenstein (1980) summarized these studies: "[W]hile more likely hazards generally evoked higher estimates, ... [i]n general, rare causes of death were overestimated and common causes of death were underestimated." People also seem to see patterns in series of events that are actually random sequences, as summarized in the phrase "gambler's fallacy." Gilovich, Vallone, and Tversky (1985), for example, found that people believe in streaks in basketball shooting, when, in fact, the data show that the sequences are essentially random.

Shanteau (1978) argued that results showing the operation of such biases illustrate only how people make probability estimates about events for which they lack direct knowledge. There is some evidence that substantive knowledge improves probability judgments. Shanteau (1978) pointed out that experiments on subjects' perceptions of word frequencies and occupation frequencies showed less bias; presumably subjects have more direct experience with these categories than with public health data. J. Christensen-Szalanski, Beck, C. Christensen-Szalanski, and Koepsell (1983) found that while physicians overestimated the mortality rates of various diseases, with a median overestimation of 150%, the physicians were substantially more accurate than the judgments of college undergraduates, who had a median percentage error of 2,583%.

One way to control for subjects' substantive knowledge is to provide them with the experience on which they are to assess a probability. Sheridan and Ferrell (1974) summarized experiments by various researchers and concluded that: "A variety of experiments point strongly to the conclusion that people are, on the average at least, good transducers of relative frequency and proportion from observed events. Furthermore, they can even

describe the stimuli accurately by assigning numerical probability values." They also comment that little learning seems to occur over repeated trials, implying that this is a task subjects already know how to do.

Einhorn and Hogarth (1981) (as well as others) have pointed out the problem with labeling as errors any discrepancies between probabilities and frequencies: "[I]n the Bayesian framework subjective probabilities represent statements of personal belief and therefore have no objective referent."

Summary. It is difficult to argue that people should, on their own, know results concerning combinatorial probabilities that are usually the subject of specialized instruction at the college level. Furthermore, no evidence has been presented that such results are particularly useful for real-world decisions and thus should have been learned through experience. It is similarly unclear that subjects should be expected to know public health statistics.

A consensus seems to be emerging that these discrepancies should not be called errors, but that the discrepancies do have patterns and thus can be described as a bias, that is, a tendency to provide estimates that do not average to the true value, but which have a systematic deviation. There is some evidence that the bias is reduced by substantive knowledge.

In an effort to test probability judgments without using an objective reference and without simultaneously testing the subject's substantive knowledge, experimenters have used the idea of calibration.

Miscalibration

Definition. "If a person assesses the probability of a proposition being true as .7 and later finds that the proposition is false, that in itself does not invalidate the assessment. However, if a judge assigns .7 to 10,000 independent propositions, only 25 of which subsequently are found to be true, there is something wrong with these assessments. The attribute that they lack is called calibration ... " (Lichtenstein, Fischhoff, and Phillips, 1982). Adams (1957) seems to be the original source for this concept.

Early history. Many calibration studies have been performed. Lichtenstein *et al.* (1982) summarized the results: "The most pervasive finding in recent research is that people are overconfident with general-knowledge items of moderate or extreme difficulty." These are questions such as "Which magazine had the largest circulation in 1970, *Playboy* or *Time*?" in which the subject selects an answer and then gives his probability that his selected answer is correct. Overconfidence is highest for very difficult tasks and decreases as tasks get easier. While weather forecasters have been found to be well calibrated (Murphy and Winkler, 1984), the calibration of physicians is poor (Christensen-Szalanski and Bushy-

head, 1981). Some studies have found good calibration on future events, while others have found overconfidence in this situation also (Lichtenstein *et al.*, 1982). Fischhoff and MacGregor (1982) concluded "calibration for confidence assessments regarding forecasts is largely indistinguishable from that pertaining to general knowledge questions."

In a series of experiments, Howell (1971, 1972) found overconfidence for what he calls internal probabilities, that is, probabilities concerning events arising from the person's own behavior (for example, hitting a target with a dart). Vertinsky, Kanetkar, Vertinsky, and Wilson (1986), however, found that the members of a field hockey team were well calibrated in giving probabilities for outcomes of games.

Efforts to improve subjects' calibration by increasing their motivation or by cautioning them did not seem to have an effect (Lichtenstein *et al.*, 1982), but asking subjects to provide reasons why their answer might be *wrong* did seem to help (Koriat *et al.*, 1980). Arkes, Christensen, Lai, and Blumer (1987) found that less direct methods such as simply providing feedback or having subjects anticipate a group discussion of their answers also reduced overconfidence.

Hypothesis fixation (discussed elsewhere in this paper) can be viewed as overconfidence in the probability that one's hypothesis is correct. Fischhoff, Slovic, and Lichtenstein (1978), Mehle, Gettys, Manning, Baca, and Fisher (1981), and Mehle (1982) found that subjects exhibited overconfidence in the completeness of the set of hypotheses they considered.

Summary. While calibration is meant to measure a subject's ability to give probabilities apart from his substantive knowledge, there is some evidence that calibration improves with increased knowledge. Lichtenstein *et al.* (1982) found that calibration was better on easier questions; weather forecasters are well calibrated; the field hockey players were well calibrated on their own abilities. Counter evidence is provided by the poor calibration of physicians. Again, it is unclear whether subjects provide poor probabilities because of cognitive error or because of lack of substantive knowledge. As with incorrect probabilities, it may be that overconfidence occurs mostly when the subject lacks substantive knowledge.

Subjects' understanding of reasoning under uncertainty can also be tested by their ability to provide probabilities that are in the correct logical relationship to each other. In the next two errors, people's probability judgments are not sensitive to sample size and do not follow rules of set intersection and union.

Insensitivity to sample size

Definition. The law of large numbers implies that the probability of seeing an extreme result decreases as the size of the sample increases, but subjects often fail to use this fact.

Early history. Tversky and Kahneman (1974) found that "when subjects assessed

the distributions of average height for samples of various sizes, they produced identical distributions. For example, the probability of obtaining an average height greater than 6 feet was assigned the same value for samples of 1000, 100, and 10 men."

However, Peterson and Beach (1967) earlier summarized experiments that imply that subjects *are* sensitive to sample size in some questions. The subjects were not asked for probabilities but instead for their confidence in their conclusions. "Data that vary along a dimension are sampled from one of two populations, and subjects decide from which of the two populations the data have been drawn. ... In both experiments [of Irwin, Smith, and Mayfield, 1956], confidence increased with the size of the sample. ... . Little and Lintz (1965) performed a similar experiment and found that on a trial-by trial basis, confidence increased with sample size."

Later history. Evans and Dusoir (1977) found considerable individual variability in the use of sample size. They also found that simplifying the problem increased the fraction of subjects giving correct answers. Kahneman and Tversky's (1972) version of one problem was:

"A certain town is served by two hospitals. In the larger hospital about 45 babies are born each day and in the smaller about 15 babies are born each day. As you know, about 50% of all babies are boys. The exact percentage of baby boys, however, varies from day to day. Sometimes it may be higher than 50%, sometimes lower.

"For a period of 1 year, each hospital recorded the days on which more than 60% of the babies born were boys. Which hospital do you think recorded more such days?"

With Kahneman and Tversky's version, Evans and Dusoir found that 11 out of 20 subjects gave the correct answer. With the following simplified second paragraph:

"Which hospital do you think is more likely to find on one day that more than 60% of the babies born were boys?"

14 out of 20 subjects gave correct answer. With this simplified second paragraph:

"Which hospital do you think is more likely to find on most days recorded in a year that all the babies born were boys?"

17 out of 20 subjects gave the correct answer. With both simplifications, that is:

"Which hospital do you think is more likely to find on one day that all the babies born were boys?"

again, 17 out of 20 gave the correct answer. Evans and Dusoir pointed out that the original version requires subjects to assess a compound event with repeated trials, which introduces more complexity.

Bar-Hillel (1979) made similar modifications to some of Kahneman and Tversky's (1972) questions and found that subjects did better. In other experiments, she also found,

as Peterson and Beach reported, that confidence increased with sample size.

While Bar-Hillel noted that people are sensitive to sample size, she reported results based on subjects' rank ordering of their confidence in various surveys to support her hypothesis that they do so on the basis of the ratio of the size of the sample to the size of the population, which may not always be an appropriate method. Bar-Hillel concluded "[w]hile there still remains the problem of convincing people that it is absolute rather than relative sample size which typically determines a sample's worth, people's intuitions regarding the role of sample size in sample evaluation may not be as discouraging as previously thought."

Fischhoff *et al.* (1979) performed similar experiments in a within-subjects, rather than between-subjects framework, to see if this would draw the subjects' attention to the importance of sample size. Unlike the results with base rates and with overprediction (reported later), this manipulation did *not* cause subjects' responses to vary with sample size.

Summary. In simple questions many subjects do seem to be appropriately sensitive to sample size. They seem to be aware of the fact that confidence should increase more with a larger sample and thus do not seem to generally make any cognitive error. If subjects do make such an error in more complicated questions it may be due to other cognitive errors than a lack of understanding of the effect of sample size.

Conjunctions and disjunctions

Definition. Two errors in the estimation of probabilities are discussed here. The first is the overestimation of the probability of a conjunctive event, that is, an event such as drawing a red marble in *all* draws of seven successive draws from a bag containing red and white marbles. The second error is the underestimation of the probability of a disjunctive event, that is, an event such as drawing a red marble in *at least one* draw of seven successive draws from a bag containing red and white marbles. Both examples are from Bar-Hillel (1973).

Early history. Evidence for these errors is provided by subjects' choices between bets. Bar-Hillel (1973) found that in pairwise choices between bets, many subjects chose to bet on a simple event (that had probability .500 of winning) rather than to bet on a disjunctive event that had higher probability (.522); many subjects also chose to bet on a conjunctive event (.478) rather than to bet on a simple event with higher probability (.500). Slovic (1969) reported similar results and reviewed earlier work by Cohen and Hansel (1958).

Other experimenters asked subjects only to rank statements by likelihood. A conjunctive fallacy is said to have occurred when a subject ranks as more likely a compound event than a component of the event (Yates and Carleson, 1986). For example, subjects were

given the following description:
"Linda is 31 years old, single, outspoken and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations."
Many subjects then ranked the statement "Linda is a bank teller and is active in the feminist movement" as more likely than the statement "Linda is a bank teller" (Tversky and Kahneman, 1982). Yates and Carlson found conjunction fallacy rates of 29.5% to 70.5%, with the lowest rate occurring in abstract forms of the problem.

Later history. Markus and Zajonc (1985) suggested that people may be misinterpreting sentences such as "Linda is a bank teller" to mean "Linda is a bank teller and not active in the feminist movement" because of the presence of the alternative "Linda is a bank teller and is active in the feminist movement". Morier and Borgida (1984) found that explicitly including the statement "Linda is a bank teller who is not a feminist" in the list of statements to be ranked reduced the percentage of subjects making the conjunctive error. Evidence against this explanation is provided by Tversky and Kahneman (1983) in a study showing that physicians committed the conjunction fallacy. In a separate question the physicians said they were interpreting symptoms as being *among* the symptoms (not the only symptoms) the patient had, implying that their incorrect judgment was not due to misinterpretation of the statements. However, one must always be careful when using a subject's explanation of his thinking as evidence of his actual cognitive processes.

Tversky and Kahneman (1983) reported that "the conjunction of an unlikely symptom with a likely one was [incorrectly] judged more probable than the less likely constituent" alone. Similarly, Wells (1985) found that the conjunction fallacy occurred with high frequency only if a representative event (for example, "Linda is active in the feminist movement," which is perceived to match the given description of Linda) is added to an unrepresentative event (for example, "Linda is a bank teller," which is perceived not to match the description).

Summary. As with earlier errors, it seems that subjects cannot intuitively compute correct combinatorial probabilities, but this is again hard to label a cognitive error. However, the conjunction fallacy, that is, the incorrect ranking of two statements, does seem to be a cognitive error by subjects. Tversky and Kahneman (1983) attribute the conjunction fallacy to the operation of the representativeness heuristic, that is, subjects assess the likelihood of a statement by its degree of correspondence with a model, in the example, a model of Linda. More specifically, subjects may be assessing the probability that Linda would have the given description under the various possible answers. They are responding correctly that the description of Linda is more likely if the subject knows Linda is

a bank teller and active in the feminist movement than if the subject knows only that she is a bank teller. However, subjects are responding to the wrong question, since they have reversed the order of conditioning. This explanation still leaves open why subjects sometimes reverse the order of conditioning and sometimes don't.

The Linda question differs from previously cited experiments in that the experimenter provides information to the subject, in this case a description of Linda, and asks the subject to give probabilities after using the information. Other experiments have looked even more explicitly at subjects' use of information provided to them.

Incorrect revision of probabilities

Definition. Evaluated against Bayesian revision of probabilities, people tend to make two errors: failing to revise probabilities enough when data are processed simultaneously, and, conversely, revising probabilities too much when data are processed sequentially.

The following example (from Edwards, 1968) illustrates conservatism. A coin is flipped to select one of two bookbags, either predominantly red (700 red poker chips and 300 blue) or predominantly blue (700 blue and 300 red). After observing that in a set of 12 chips, selected randomly with replacement, 8 are reds and 4 are blues, one can use Bayes's theorem to calculate that the probability that the predominantly red bag was selected has increased to .97 due to the data. However, subjects typically estimate probabilities in the range .7 to .8.

Early history. In 1968, Edwards summarized the results of several studies on conservatism. "An abundance of research has shown ... that opinion change is very orderly, and usually proportional to numbers calculated from Bayes's theorem - but it is insufficient in amount." Alker and Hermann (1971) replicated the results with other versions of the problem and found that "subjects were more conservative as the decisions to be made became more lifelike in nature."

Evidence is mixed, but seems to indicate that training helps to reduce conservatism. "Peterson, DuCharme, and Edwards (1968) ... found that conservatism in a book-bag-and-poker-chip task was only slightly reduced by training [subjects] in the sampling distribution. Wheeler and Beach (1968), on the other hand, found a sharp reduction in conservatism after training [subjects] by having them place bets on the source of the sample and then giving feedback on the correct urn" (Messick and Campos, 1972). Messick and Campos found that training in both the sampling distribution and the posterior distribution helped reduce, but not eliminate, conservatism.

In contrast to conservative *single*-stage inference, most studies of *multi*-stage inference have shown that subjects are radical in their revision of opinions.

Later history. Navon (1979, 1981) argued that in many real-world situations data are

not independent. Furthermore, he argued that conservatism is called for in situations where data are correlated in a way that is likely in real-world situations. For example, among income, education, and social class, each pair is positively correlated. If one is trying to infer the social class of a person from data about income and education, the second piece of data should be discounted somewhat, that is, the observer should be conservative in revision of probabilities relative to the assumption of independence of data.

Summary. The correct probabilistic inference to draw from data depends on one's model of the process that generated the data. It is unclear whether subjects use a different model from the experimenter or make an error in inference. In either case, they have made an error under the experimental conditions, but their incorrect model, which assumes dependence of data, may often be correct in real world situations.

Bayes's theorem is also the standard for defining the next error.

Ignoring base rates

Definition. In many situations involving probabilistic inferences, two types of information are relevant: base rate information about the characteristics of the population at large, and specific information about an individual from that population. In statistical inference, both types of information should be considered, since the specific information leads to only probabilistic, not definitive statements about an individual. However, many subjects ignore base rate information in making probabilistic statements about an individual drawn from a population; equivalently, they place too much emphasis on the particular information about the individual.

Early history. Kahneman and Tversky (1973) reported on several experiments which illustrate this error. In one experiment, each subject was told that psychologists had interviewed 30 engineers and 70 lawyers and created 100 thumbnail descriptions of them. The subject was told that one description had been picked at random and he was to "indicate your probability that each person described is an engineer, on a scale from 0 to 100." Another group was given the same instructions, but was told that the description had been randomly selected from a group of 70 engineers and 30 lawyers. Both groups were also asked to give probabilities assuming they had no description of the individual. Kahneman and Tversky found "that explicit manipulation of the prior distribution had a minimal effect on subjective probability," although subjects did use the prior probabilities correctly when they were given no description. On the other hand, when given a description which was basically uninformative, the median response was .5, regardless of priors.

Another widely cited example in which subjects ignore base rates is the cab problem. "A cab was involved in a hit and run accident at night. Two cab companies, the Green and the Blue, operate in the city. You are given the following data:

a. 85% of the cabs in the city are Green and 15% are Blue.
b. A witness identified the cab as Blue. The court tested the reliability of the witness under the same circumstances that existed on the night of the accident and concluded that the witness correctly identified each one of the two colors 80% of the time and failed 20% of the time.
What is the probability that the cab involved in the accident was Blue rather than Green?" (Tversky and Kahneman, 1982).
Applying Bayes's theorem leads to a probability of .41. However, Tversky and Kahneman (1982) found that "[t]he median and modal answer [of subjects] is typically .8, a value which coincides with the credibility of the witness and is apparently unaffected by the relative frequency of Blue and Green cabs."

Later history. Lyon and Slovic (1976) used a light bulb problem which "takes the situation out of a court of law, employs a 'mechanical witness', emphasizes random selection from the population, and presents the individuating information in the same form as the base rate is given, i.e., as a percentage." Lyon and Slovic report that "the change in cover story provided by the Light-Bulb problem ... had little effect on the use of base-rate information." The median estimate "continued to match the stated accuracy levels regardless of whether the accuracy was 80%, 50%, or 20%." Several authors have pointed out that subjects' answers showed more variation in the light bulb problem than in the cab problem. The correct answer in the cab problem is 41%; the median answer was 80% with interquartile range 60-80%. The correct answer in the light bulb problem was also 41%; again the median estimate was 80% but with interquartile range 25-80%. This may indicate that while the median answer was still non-Bayesian, more subjects were using the base rate information.

Bar-Hillel (1983) pointed out that the error is quite common even in professional judgments. "[A]ccording to Dershkowitz (1971), Eddy (1982), Foltz and Kelsey (1978) and Lykken (1975), to cite but a few, this fallacy is still prevalent in such important areas of judgment and decision making as preventive detention, mammography, Pap smears, and lie detection, respectively." Using some of the standard questions (for example, the cab problem) as well as medical examples, Borak and Veilleux (1982) found the error in physicians' judgments, but the rate decreased with statistical training and with experience.

Christensen-Szalanski and Bushyhead (1981) found opposite evidence concerning physicians.
"Since base-rate information is necessary to calculate predictive value, the significant positive correlation between a symptom's actual predictive value for pneumonia and that obtained from the physician's estimate of a patient's probability of pneumonia suggests

that physicians do use base-rate information. These results may not be surprising to anyone who has heard his or her physician say, 'You probably have ___. There's a lot of that going around now.' Such a statement implies the use of base-rate information."
Wallsten (1983) summarized research on when base rates *are* used: "thus far at least three factors have been identified that affect their use. These are (i) expertise of the judge, (ii) relevance or specificity of the information, and (iii) salience of the information." Bar-Hillel (1983) listed several factors that enhance the impact of base rates: a causal relation between the population data and the individual; specificity of the base rate; and vividness of the base rate. Group discussion and estimation of the probability seems to accentuate the effect, not reduce it (Argote, Seabright, and Dyer, 1986).

Some authors have argued that subjects ignore base rates because they interpret language in a way different from that intended by the experimenters. Adler (1983) cited Grice's (1975) theory that conversation relies on cooperativeness between listener and speaker. "In one base-rate study, the experimenter's most salient contribution is a biographical sketch; this will be of greater selective relevance to the judgment requested if subjects judge on the basis of its diagnostic qualities rather than (less cooperatively) integrating it with the base-rate data" (Adler, 1983). Misinterpretation can possibly explain the results from the lightbulb problem; Spielman (1983) argued that subjects' error may lie in interpreting the word "random" to mean that defective and nondefective bulbs are equally likely to be chosen. Given this incorrect assumption, the correct answer is the base rate.

Birnbaum (1983) argued that the usual analysis of the problem requires that the observer's accuracy (.8) not vary with the base rate, but that signal detection theory implies that it should. The accuracy rate determined by the court, .8, presumably used equal numbers of each type of cab, while the city conditions have only 15% blue cabs. Birnbaum concluded that "the modal response of .8 by untrained subjects may be closer to the best normative solution than the value of .41 used in previous research." Birnbaum applied similar analysis to the light bulb and lawyer/engineer problems. He commented, for example, that a report of mathematical interest "may be highly diagnostic of a rare engineer in a population consisting mostly of lawyers," but have little diagnostic meaning in a population consisting mostly of engineers.

Cohen (1979) mounted a more fundamental objection, arguing that subjects may be performing a perfectly valid reasoning process even though it does not conform to standard probability. A rationality closer to forensic proof argues for ignoring base-rate information. "Imagine a rodeo into which 400 people are known to have been admitted through an automatic turnstile after paying the proper sum. Then 1,000 people are counted on the

seats, and a hole is discovered in the fence. A man is picked at random from the seats, who turns out to be John Smith. That is all the evidence before the court when John Smith is sued by the management of the rodeo for non-payment of his entry-money."
Cohen contrasted traditional Baconian probability with Pascalian probability which is concerned with the degree to which an assertion has been proven by the facts. In Cohen (1981) he again argued that subjects may be justified in ignoring the 85/15 base rate and using the the 50/50 rate instead even in a disease example which does not have a judicial setting.

The commentators on Cohen's 1981 article saw considerable problems with his analysis, but many agreed that the choice of priors is crucial. For example, Levi (1981) argued that, in the absence of information regarding the percentage of blue and green cabs in accidents, subjects could reasonably assume equal numbers are in accidents, thus leading to the correct conclusion of a .8 probability. Furthermore, he noted: "Kahneman and Tversky have reported on variants of the taxicab experiment. According to one variant, the subjects are told that 85% of the taxicabs involved in accidents are blue and the remainder green. In that case, the subjects do not neglect base rates" (Levi, 1981).

Einhorn and Hogarth (1981) also emphasized the question of selecting an appropriate prior. They noted that "a base rate can only be defined conditional on some population (or sample space)" and they argued that "there is no generally accepted normative way of defining the appropriate population." For example, "consider an inference concerning whether someone has a particular propensity to heart disease. What is the relevant population to which this person should be compared? The population of people in the same age group? The population of the United States? Of Mexico?"

Fischhoff, Slovic, and Lichtenstein (1979) converted some of these experiments from a between-subjects design to a within-subjects design. Using the cab problem and the light-bulb problem, an average of 16% of the respondents always answered .8 (that is they completely ignored the base rate), an average of 10% always answered with the base rate, an average of 63% gave answers that were ordered according to the base rate, and the remainder had other patterns of responses. They commented "[t]he robust finding of these studies is that most people know (or guess) the direction in which base rates should influence their judgments," but they do not adjust enough for the base rate.

Some authors have argued that the error has not been demonstrated. Birnbaum and Mellers (1983) gave "three reasons for a skeptic to doubt the claim that humans neglect base rate." First, the empirical evidence is inconsistent; the effect does not appear with certain versions of the problems and, as shown by Fischhoff *et al.* (1979) and by Birnbaum and Mellers themselves, the effect does not appear in within-subject designs. Second, the

use of many subjects "assumes that the mapping from subjective values to overt responses is the same for all subjects ... ." Third, vagueness in the problems allows "several rational solutions, depending on how the subject interprets the problem."

Summary. While there is evidence that some subjects do not use base rate information as efficiently as they could, there is also strong evidence that base rate information is used under some circumstances. Furthermore, there are reasonable arguments that even subjects who completely ignore base rates may not be committing an error. Again is it unclear whether subjects use the same model of the process as that used by the experimenter.

The next error also relates to the issue of whether people draw correct inferences from information.

Hindsight bias

Definition. Telling subjects the outcome of an event increases their *post hoc* estimate of the probability of that outcome, without the subject being aware that this has happened.

Early history. Fischhoff (1975) originated research in this area. In one experiment, he asked subjects to read a passage describing historical events in Nepal and then to give probabilities for four possible outcomes. Half the subjects were also told the actual outcome before they were asked to provide the probabilities. "[R]eporting an outcome's occurrence approximately doubles its perceived likelihood of occurrence" averaged across subjects and questions. Even when told to report probabilities as if they didn't know the true outcome, subjects' probabilities were affected by reported outcome. Experiments also showed that, 2 weeks to 6 months later, subjects "remembered having given higher probabilities than they actually had to events believed to have occurred and lower probabilities to events that hadn't occurred."

Later history. The bias seems to be robust. "Attempts to undo this *knew-it-all-along* effect by exhorting subjects to work harder or telling them about the bias failed' (Fischhoff, 1977). Wood (1978) found that other manipulations also failed to eliminate the bias. However, Arkes, Faust, Guilmette, and Hart (1988) successfully reduced the bias using a method that reduces overconfidence: asking subjects "to provide a reason why each of the other diagnoses might have been correct."

Some authors argue that the bias is only of concern for historians. As von Winterfeldt and Edwards (1986) stated, "[o]f all the cognitive illusions, the hindsight illusions would seem to be the easiest to correct: one simply writes down the probability estimate before the event occurs, or before the estimator learns the answer." However, Fischhoff (1975) argued that this bias has deleterious effects on judgments. As a result of their changed perceptions, subjects "overestimated what they would have known without outcome knowledge ..., as well as what others ... actually did know without outcome knowledge. It is argued that

this lack of awareness can seriously restrict one's ability to judge or learn from the past." He also found that "[r]eporting an outcome's occurrence alters the judged relevance of data describing the situation preceding the event."

Arkes, Wortmann, Saville, and Harkness (1981) found that physicians exhibit hindsight bias. "Thirty-eight out of 60 hindsight subjects gave higher probability estimates to the known-to-have-occurred diagnosis than the corresponding probability estimate obtained for the foresight group ...." They also noted: "However, the bias was restricted to the two diagnoses assigned the lowest probability estimates by the foresight group." "This result replicates the findings of Fischhoff (1977) and Wood (1978) who found that the hindsight bias is strongest for those events initially judged to be least plausible." Dawson, Arkes, Sicilian, Blinkhorn, Lakshamanan, and Petrelli (1988) also found a hindsight bias among physicians who were attending clinicopathologic conferences, except among more experienced physicians who were estimating probabilities for more difficult cases.

Summary. The effect does seem to occur in a number of situations, but its practical significance has been questioned. At first glance, the task of retrospectively estimating one's probabilities would seem to have little relevance to real-world situations. However, Fischhoff has raised concern about the implications of these results for the ability of people to learn from case histories. The lesson would be that teachers should be careful to present evidence first, asking students to draw conclusions and state their confidence, before revealing the correct solution. In real-world situations, of course, information may not be presented in such an order.

The next set of errors involves errors in assessing the degree of relationship (causal or statistical) between events.

Attribution errors

Definition. Attribution theory involves studying the causes people give for their own behavior and for other people's behavior. Two errors are specific to the context of explaining human action, the fundamental attribution error and egocentric bias. Ross and Anderson (1982) describe the fundamental attribution error as "the tendency for attributers to underestimate the impact of situational factors and to overestimate the role of dispositional factors in controlling behavior." The label egocentric bias has been used for several errors.

Early history. Ross and his colleagues have performed many experiments on errors in attribution, demonstrating the occurrence of the fundamental attribution error. For example, in experiments by Ross, Amabile, and Steinmetz (1977) participants were arbitrarily assigned the roles of questioner and answerer; even observers who knew that the assignments were arbitrary attributed more knowledge to the questioners than to the answerers. Other researchers have shown similar effects.

The rate of occurrence of the fundamental attribution error depends on whether the attributer is explaining his own or another's actions. Jones and Nisbett (1971) proposed that "there is a pervasive tendency for actors to attribute their actions to situational requirements, whereas observers tend to attribute the same actions to stable personal dispositions." Kelley and Michela (1980) summarized research on this topic: "[t]he preponderance of studies confirms Jones & Nisbett's ... hypothesis: actors tend to make more situational attributions and observers, more dispositional ones."

A second category of attribution errors, the egocentric bias, has been applied to several behaviors. Thompson and Kelley (1981) described one egocentric bias:
"When partners in a relationship judge the extent of their contribution to an activity, each one tends to claim more responsibility for the event than the other is willing to attribute to them. This phenomenon ... was first identified by Ross and Sicoly (1979) who demonstrated that egocentric biases occur in a variety of relationships, including marital couples, basketball teams, and temporary dyads in the laboratory."
Ross, Greene, and House (1977) also discussed the egocentric attribution or false consensus bias: "people's tendency ... to see their own behavioral choices and judgments as relatively common and appropriate to existing circumstances while viewing alternative responses as uncommon, deviant, and inappropriate." Elstein and Van Pelt (1969) found this effect in the perception of psychiatric patients by hospital staff: "almost without exception, each member of the staff says in effect that other staff see the patient the same way that he does when, in fact, the real similarity correlations show otherwise."

Finally, the self-serving bias, is the tendency of people "to attribute their own success to internal factors and to blame external factors for their failures" (Kruglanski and Ajzen, 1983). Many studies have shown this effect (see, for example, Weary Bradley, 1978, and Miller, 1976).

Later history. Some researchers have questioned whether attribution errors should be called errors. Harvey and Weary (1984) summarized these objections:
"The issues raised in this controversy include: (a) what questions research subjects think they are answering in studies purporting to show the working of the fundamental attribution error (Hamilton 1980); (b) what criteria to use in establishing the accuracy of attribution (Harvey, Town, and Yarkin 1981); (c) whether under certain conditions overattribution to situational factors also may represent incorrect judgment (Harvey *et al.* 1981); and (d) the idea that an important difference between situational and dispositional attributions is level of analysis; that is, situational attributions describe environmental circumstances associated with behavior, while dispositional attributions are intended to describe how a given action fits into the larger pattern of the actor's behavior over time

(Funder, 1982)."
Miller and Ross (1975) and Markus and Zajonc (1985) raised similar objections.

Summary. In attribution experiments, subjects are asked to function as amateur psychologists; it is not surprising that their intuitive judgments do not always agree with the findings established by psychologists. For example, subjects and observers in the experiment of Ross *et al.*, tended to be affected by the advantage conferred by the role of questioner. It would take a sophisticated subject or observer to recognize that such an effect might be operating, to reason that such an effect might be inappropriate in the situation of arbitrarily assigned roles, and finally to correct for the effect. In fact, in most real-world situations, evidence suggesting the superior ability of another person *should* lead to a revision of opinion.

The next error involves assessing statistical dependencies from data.

Learning probabilistic dependencies

In a real world situation a person may need to judge how often the occurrence and nonoccurrence of two events have been associated in order to decide whether one event can be used as an indication of the other event. Several errors appear in subjects' learning of such probabilistic dependencies. We discuss two here, illusory correlation and reliance only on positive hits. The first is the perception of positive correlation where none actually exists. The second is the estimation of association based only on the number of cases where both events occur.

Definition. Chapman (1967) proposed the term illusory correlation "for the report by observers of a correlation between two classes of events which, in reality, (a) are not correlated, or (b) are correlated to a lesser extent than reported, or (c) are correlated in the opposite direction from that which is reported."

Early history. Chapman presented subjects with pairs of words and then asked for an estimate of how often particular words were paired. Subjects tended to overestimate the co-occurrence of word pairs with "high associative connection," for example, hat and head. The amount of illusory correlation declined with repeated testing.

In a more realistic setting, Chapman and Chapman (1967 and 1969) found the effect in the association of clinical symptoms and diagnoses with naive subjects and with practicing psychodiagnosticians. Both groups tended to see a relationship between male homosexuality and the person's answers on DAP (Draw-a-Person) tests or on Wheeler-Rorschach cards. "The reported relationships corresponded to rated associative strength between symptom and drawing characteristic" (Chapman and Chapman, 1967). "[The] popular invalid signs were found to have much stronger rated, verbal associative connections to male homosexuality than the unpopular valid signs" (Chapman and Chapman, 1969).

Later history. Alloy and Tabachnik (1984) reviewed the literature on illusory correlation and concluded: "Our examination of the work on humans' covariation detection abilities demonstrates that under some conditions, people detect event contingencies accurately." Similarly, Kunda and Nisbett (1986) examined the conditions under which illusory correlation occurs. They "found substantial accuracy for correlation estimates if two conditions were met: (1) subjects were highly familiar with the data in question and (2) the data were highly 'codable', that is, capable of being unitized and interpreted clearly. ... Subjects were particularly inaccurate about correlations involving social behavior: They severely overestimated the stability of behavior across occasions."

In all of these experiments subjects were asked to estimate a relationship only from the data presented to them, that is, to ignore any prior beliefs they had concerning the relationship. This may be difficult for subjects to do and indeed, a Bayesian analysis says that they should *not* ignore prior beliefs but should revise them in light of the evidence. Golding and Rorer (1972) replicated Chapman and Chapman's 1969 study, but asked subjects for estimates of the relationships between cues and hypotheses both before and after exposure to training materials. They found that subjects *did* substantially revise their probability estimates of relationships in light of the evidence, although "the posttraining levels were still considerably biased, in an absolute sense, indicating that, while the bias was modifiable, it was still relatively resistant to change."

Summary. Naive subjects are likely to rely on unconscious or conscious prior associations in order to assess co-occurrence of events. In a real-world situation, it may be appropriate to use such prior beliefs and therefore subjects find it difficult to ignore these prior associations, even when instructed to do so.

Definition. The error of looking only at positive hits can be explained by the following table:

	B	not-B
A	*a*	*b*
not-A	*c*	*d*

where A and not-A represent the occurrence or nonoccurrence of event A, B and not-B represent the occurrence or nonoccurrence of event B, and *a*, *b*, *c*, and *d* represent the number of cases in each cell. The ability to predict B from A depends on there being a difference between $a/(a+b)$ and $c/(c+d)$. The error described here involves using only the magnitude of *a* to judge the relationship between events A and B.

Early history. Smedslund (1963) asked student nurses to estimate the relationship between a symptom and diagnosis by viewing a pack of 100 cards containing symptom and diagnosis pairs. He concluded: "Their strategies and inferences typically reveal a particularistic, non-statistical approach, or an exclusive dependence on the frequency of [positive symptom and positive diagnosis] instances." The subjects tended to estimate a large relationship between A and B when the value of *a* in the above table was high, regardless of the values of *b*, *c*, and *d*. Subjects were not presented with the summary table.

Jenkins and Ward (1965) asked subjects to judge the strength of a relationship by saying how much control they thought they had over which of two outcomes would occur. "The main finding of these experiments is that the amount of judged control was a function of the frequency of successful outcomes rather than of the actual dependency of outcomes upon response." The authors were aware of the difficulty of using control as a surrogate for the concept of contingency. Ward and Jenkins (1965) asked subjects to use data to judge "the degree of control over rainfall exerted by seeding." Subjects who received trial-by-trial data (with or without a later summary table being presented) made poorer judgments than those who received only the summary table. Of 22 subjects who saw only the summary table, 18 seemed to be basing their judgments appropriately on the difference in the ratios $a/(a+b)$ and $c/(c+d)$. However, "[t]hose who receive information on a trial by trial basis, as it usually occurs in the real world, generally fail to assess adequately the degree of relationship present."

Later history. Shaklee and others (Shaklee and Tucker, 1980; Shaklee and Mims, 1981; Shaklee and Mims, 1982) evaluated which of several possible rules subjects might be using in forming judgments of covariation. Shaklee and Mims (1982) summarized the results: "Conditional-probability rule patterns [the correct rule] were produced by sizeable minorities of subjects in the 10th-grade (17%–27%) and college (33–38%). Sum-of-diagonals judgment patterns [comparing $a+d$ to $b+c$] were common as early as seventh grade (50% of the sample) and persisted through the college years (35%–38%). Systematic *a*-versus-*b* rule patterns appeared as early as the fourth grade (29% of the sample) and remained common through college age (18%–38%). Cell-*a* patterns [the one suggested by Smedslund (1963)] were rare at all ages in the fourth grade to college age span (0%–8%)."

In all these experiments subjects used data presented in tables. Shaklee and Mims (1982) showed subjects the data sequentially. Some subjects were asked to estimate the frequencies in the four cells; all subjects were asked to assess covariation. The authors found that the covariation judgments of subjects asked to estimate frequencies were less accurate than those of subjects given summary tables, partly due to the use of less accurate judgment

rules. "In particular, conditional-probability and sum-of-diagonals patterns declined in frequency, whereas *a*-versus-*b* and cell-*a* patterns became more common. This preference for simpler rules compromises judgment accuracy at the same time that it simplifies the demanding decision environment."

Arkes and Harkness (1983) found that several factors affected which rules subjects applied, including the words used in labeling, memory load, salience, and method of estimation (a running estimate or only a final estimate). Beyth-Marom (1982) gave three factors accounting for the variability of results. Serial presentation of data encouraged the use of the *a*-only rule, while tabular presentation encouraged other rules. Some task instructions pointed to cell *a*, others to more cells. Finally, some studies used symmetric variables (for example, gender) in which both categories (male and female) have similar status as diagnostic clues, while other studies used asymmetric variables (for example, symptom) in which the two categories (symptom present and symptom absent) differ in perceived usefulness for a target category such as pneumonia present or absent.

Summary. There is considerable variation across and within experiments in the rule used by subjects in assessing covariation, with no simple pattern of people relying only on positive hits. When the task involves remembering data presented sequentially, subjects do seem to use simpler rules to judge relationships than when data are presented in a table.

Overprediction

Definition. In order to predict one variable from another, a person must be aware of the regression effect.

"A fundamental rule of the normative theory of prediction is that the variability of predictions, over a set of cases, should reflect predictive accuracy. When predictive accuracy is perfect [a correlation of 1 or -1], one predicts the criterion value that will actually occur. When uncertainty is maximal [a correlation of 0], a fixed value is predicted in all cases. ... With intermediate predictive accuracy, the variability of predictions takes an intermediate value, that is, predictions are regressive with respect to the criterion" (Kahneman and Tversky, 1973).

The error made by some subjects is to give similar predictions for situations with different degrees of uncertainty.

Early history. In Kahneman and Tversky (1973) two groups of subjects were given a verbal description of a college freshman and asked to give either (1) the percentage of freshmen likely to have a more impressive description, or (2) the percentage of freshmen likely to have a higher grade point average than this student. The subject should have greater uncertainty about the second question since the evaluation of the student could have been wrong or the evaluation might not have been an accurate prediction of grade

point average; the answers to the second question should have been more regressive, that is, should be closer to the mean. The results of this study were that the two groups produced estimates that did not differ in variability.

Kahneman and Tversky discussed the difficulty of teaching the regression effect even to graduate students. They illustrated this with an example in which instructors in a flight school found that "high praise for good execution of complex maneuvers typically results in a decrement of performance on the next try." Not one of the graduate students to whom this situation was described suggested regression as an explanation.

Later history. Nisbett and Ross (1980) reported unpublished experiments by Amabile and by Ross, Amabile, and Jennings which showed that "very few individuals systematically applied anything like a simple linear 'prediction equation.'"
"When subjects believed the relationship between variables to be strong, their individual predictions were often well matched by a simple linear function, but then they believed the relationship to be weak, their predictions varied widely around any potential regression line. Specifically, when subjects believed the relationship between X and Y to be relatively weak, they did not respond to extreme values of X with predictions of Y that were consistently moderately close to the mean. Instead, they responded by *varying* the extremity of their predictions (for examples, by predicting one very extreme value of Y and one value of Y close to the mean when given two identical values of the predictor variable X)."

As in the experiments concerning the use of base-rate and sample size information, Fischhoff *et al.* (1979) replicated the experiments in a within-subjects, rather than between-subjects framework, in order to draw the subjects' attention to the regression effect. Each subject predicted the grade point average associated with the 5th, 15th, 25th, ..., and 95th percentiles of three distributions of different types of scores. "One set of percentiles was described as coming from the distribution of GPA's one came from scores on a test of mental concentration described as having a moderate correlation with GPA; and one came from a measure of sense of humor described as having a low but positive correlation with GPA." Fischhoff *et al.* found that "[t]he sense of humor judgments were markedly regressed relative to the GPA and mental concentration judgments, with mental concentration judgments somewhere in between GPA and sense of humor. These differences contrast with the virtual identity of Kahneman and Tversky's (1973) mental concentration and GPA groups and slight regression with the sense of humor group."

Einhorn and Hogarth (1981) questioned the underlying normative model used to label this error: "extreme predictions are not suboptimal in nonstationary processes. In fact, given a changing process, regressive predictions are suboptimal. ... [T]he optimal prediction is conditional on which hypothesis you hold."

Summary. In these experiments, the experimenter creates a situation in which he believes that unexplainable random variation contributes to a phenomenon. The subject may disagree with the experimenter in that he thinks that he can explain the variation. In the experiment, the experimenter has guaranteed that he is correct, but in real-world situation, the subject may often be correct.

Illusion of control

Definition. The illusion of control is "the perception of control over objectively chance-determined events" (Langer and Roth, 1975).

Early history. Langer (1975) demonstrated illusion of control in subjects by asking the size of bet the subjects were willing to make. For example, subjects, on the average, placed smaller bets against a confident opponent in the game of drawing a high card than against a less confident opponent. The subject was told the experiment concerned the measurement of skin resistance. Langer also found that the more similar the chance situation was to a skill situation, the greater was the illusion of control. "This illusion may be induced by introducing competition, choice, stimulus or response familiarity, or passive or active involvement into a chance situation."

Langer and Roth (1975) found that subjects given early successes were likely to experience an illusion of control even in a coin tossing experiment. "Most of the subjects in the random and ascending groups gave responses that characterize a mechanical reading of objective probabilities, but subjects in the descending group deviated from these two groups."

Langer (1975) comments that in real situations it is not always easy to distinguish skill situations from chance situations: "there is an element of chance in every skill situation and an element of skill in almost every chance situation. The former is obvious and needs no further explication here. Examples of the latter are knowing what a good bet is in a game of dice (i.e., knowing the odds) or knowing which slot machines are rigged to given the highest payoffs."

Later history. Various researchers have found that depressed persons are not as susceptible to the illusion of control as nondepressed persons. Alloy and Abramson (1982) noted that this idea agrees with "theories of depression that portray the depressive as a person who believes that he or she is ineffective and powerless to obtain desired outcomes."

Summary. The phrase "illusion of control" is a misnomer. Subjects are not asked directly whether they believe they can control an event; rather they are asked the probability that they will win on future trials or they are asked to select the size of a bet. "Illusion of control" means they overestimate the probability or bet too large an amount, compared to probabilities determined by the experiment. Subjects could overestimate the probability

of winning without having any illusion that they can control an event.

The essential difference between experimenter and subject is their beliefs concerning the probabilities that determine certain events. Alloy and Tabachnik (1984) commented: "Although Langer's ... findings are compatible with those of the human learning studies, their interpretation is unclear. Expectancies of success may reflect factors other than people's beliefs about the degree of contingency between their responses and outcomes ...; most notably, they may reflect beliefs about the stability of causes that produced past success."

The next four errors deal with logical deduction. The first two deal with the types of tests subjects make in testing or discovering a rule.

Incorrect testing of a conditional rule

Definition. In this error, a subject fails to identify the correct cases that must be checked in order to test the validity of a conditional rule, that is, a rule of the form "If ..., then ...."

Early history. In Wason (1968) "[s]ubjects were presented with the following sentence, 'if there is a vowel on one side of the card, then there is an even number on the other side,' together with four cards each of which had a letter on one side and a number on the other side. The task was to select all those cards, but only those cards, which would have to be turned over in order to discover whether the experimenter was lying in making the conditional sentence." If the sentence is represented symbolically as "if P, then Q," then, with four cards representing P (a vowel), not-P (a consonant), Q (an even number), and not-Q (an odd number), the correct answer is that the two cards representing P (a vowel) and not-Q (an odd number) must be turned over. Wason found that "[n]early all subjects select P, from 60 to 75 percent select Q, only a minority select not-Q and hardly any select not-P. Thus two errors are committed, the consequent is fallaciously affirmed [by selecting Q] and the contrapositive is withheld [by not selecting not-Q]."

This version has been repeated many times with similar results. The experiment has also been replicated with thematic versions, that is, versions with a more realistic setting. Some researchers reported that the error rate is reduced in thematic versions and other researchers found little or no effect of realism. They found poor performance on the thematic versions compared to the previously cited studies.

The following authors found improvement due to realism. Wason and Shapiro (1971) used the rule "Every time I go to Manchester I travel by car." Johnson-Laird, Legrenzi, and Legrenzi (1972) used the rule "If a letter is sealed then it has a 50 lire stamp on it." Rumelhart and Norman (1981) reported on experiments by D'Andrade using an abstract version involving checking the quality of labels produced and a thematic version involving

checking that a sale at Sears over a certain amount has the manager's signature. Griggs and Cox (1982) used the rule "If a student is drinking beer, then the person must be over 19 years of age." Griggs and Cox (1983) used the drinking age version and the Sears version. The following table shows that these researchers found that the percent of subjects who gave the correct answer (P and not-Q) increased dramatically with thematic versions:

Study	abstract	thematic	thematic rule
Wason and Shapiro (1971)	2/16=13%	10/16=63%	travel
Johnson-Laird *et al.* (1972)	2/24=8%	21/24=88%	postal
Rumelhart and Norman (1981)	13%	70%	Sears
Griggs and Cox (1982)	0/40=0%	29/40=73%	drinking
Griggs and Cox (1983)	1/20=5%	14/20=70%	drinking
		17/20=85%	Sears

Other authors have found that thematic material did not improve performance. Manktelow and Evans (1979) used rules involving food such as "If I eat macaroni, then I do not drink champagne," as well as the travel sentence of Wason and Shapiro. Griggs and Cox (1982) also used a travel sentence and the postal version of Johnson-Laird *et al.* (1972). Brown, Keats, Keats, and Seggie (1980) used a travel sentence. Yachanin and Tweney (1982) used various thematic versions. Reich and Ruth (1982) used a travel sentence. As shown in the following table these researchers found little or no effect of thematic versions:

Study	abstract	thematic	thematic rule
Manktelow and Evan (1979)	8/24=33%	7/24=29%	food (% saying not-Q)
	1/16=6%	2/16=13%	travel (% saying not-Q)
Griggs and Cox (1982)	0/32=0%	3/32=9%	travel
	1/24=4%	2/24=8%	postal
Brown *et al.* (1980)	1/24=4%	2/24=8%	travel
Yachanin and Tweney (1982)	-	5/160=3%	various
Reich and Ruth (1982)	-	3/24=13%	travel

Many of these differences between these two tables are still unexplained. The difference between these two tables has been explained for the postal version of Johnson-Laird *et al.* (1972); successful replication depends on having a population familiar with such postal regulations (Griggs and Cox, 1982). Manktelow and Evans (1979) argued that this

indicates the postal version is a memory task, not a reasoning task. Griggs and Cox (1982) agreed, arguing that subjects call up past experience rather than performing conscious reasoning. Rumelhart and Norman (1981) argued that reasoning is involved, but not abstract reasoning; the effect they observed "is exactly the kind of effect expected if our knowledge is embedded in a relatively inaccessible procedural format rather than as general rules of inference."

Later history. Several authors have argued that the meaning of an "if-then" sentence varies considerably; a thematic version allows an unambiguous interpretation (Klayman and Ha, 1987). Indeed, other authors (starting with Wason, 1968) have noted that even philosophers disagree on the interpretation of "if P, then Q" with some arguing that situations where not-P holds are simply irrelevant to the truth of the conditional. Others have noted that a sentence like "If you mow the lawn, I'll give you five dollars" "carries the invited inference (pragmatic implication) that if the hearer does not mow the lawn, he will not get five dollars" (Harris and Monaco, 1978). Furthermore, "if-then" sometimes means disjunction: "If it isn't John, then it's John's brother" (Wason and Johnson-Laird, 1972).

Several authors have addressed the larger issue of whether humans are generally logical in their reasoning. Henle (1962) argued that even when subjects appear to violate syllogistic reasoning they have, in fact, formed a valid syllogism, but by altering the information in some way, for example by changing one of the premises. Smedslund (1970) pointed out that one cannot test a subject's logical thought without assuming that the subject understands the information; vice versa, one cannot test his understanding without assuming he is logical in his thinking.

Cheng, Holyoak, Nisbett, and Oliver (1986) contrasted two extreme views of human reasoning: (1) people use "syntactic, domain-independent rules of logic;" and (2) people "develop much narrower rules tied to particular content domains in which people have actual experience." They discussed a middle approach: (3) people use "*pragmatic reasoning schemas*: clusters of rules that are highly generalized and abstracted but nonetheless defined with respect to classes of goals and types of relationships." For example, certain versions of the "if P, then Q" task led subjects to use a schema for "permission." Cheng and Holyoak (1985) found that simply rephrasing the abstract rule to emphasize permission increased the percent of subjects giving correct solutions from 17% to 67%.

Summary. A consensus seems to be growing that the ambiguity in the meaning of "if-then" leads many subjects astray. Since in a real-world situation the meaning is likely to be clearer, this error may occur infrequently in natural settings.

In the next error, subjects are not given a rule to test, but must generate it themselves.

Confirmation bias

Definition. The phrase confirmation bias has been used to label various phenomena. We will use Wason's original definition: a bias toward testing instances expected to be positive instances under one's current hypothesis, in preference to testing instances expected to be negative instances.

Early history. Bruner, Goodnow, and Austin (1956) found that some subjects used a strategy which they called successive scanning to solve a concept attainment task. Successive scanning involves focusing on a particular hypothesis and testing it by its ability to predict exemplars. Bruner *et al.* found that "in dealing with the task of sorting out relevant from irrelevant cues in the environment, subjects persist until they are able to make direct tests with positive instances."

Wason (1960) noted that this strategy could lead to errors. "For example, if the correct concept is 'red figures' ..., it is possible to attain the incorrect concept, 'red circles,' by consistent use of confirming evidence ... ." Wason therefore designed an experiment in which this strategy was likely to lead subjects astray: "[s]ubjects were told that the three numbers 2, 4, 6 conformed to a simple relational rule and that their task was to discover it by making up successive sets of three numbers, using information given after each set to the effect that the numbers conformed, or did not conform, to the rule." The correct rule was any three numbers in increasing order; many subjects initially formed the hypothesis numbers increasing by 2.

Wason looked at the compatibility between the hypothesis the subject said he was testing and the instance the subject generated to test it. Those who solved the problem correctly produced a higher average number of incompatible instances, thus demonstrating that success in the task was aided by generating instances that disconfirm the subject's current hypothesis.

It is important to examine carefully the error made by Wason's subjects. Subjects generated compatible instances by asking about instances for which they expected the answer "yes" (for example, 10-12-14 under the hypothesis of numbers increasing by 2). We cannot say that they sought only confirming evidence, since confirming evidence can be sought by asking about instances to which one expects the answer "no" (for example, 10-12-13). Also, in the subjects' minds, 10-12-14 could possibly have led to the answer "no" and thus could have generated disconfirming evidence. The subjects' bias is thus best characterized as saying that they asked about instances for which they expected the answer "yes" under their current hypothesis.

This bias is not always an error. Consider the subject's beliefs about what will happen if he asks about an instance to which he thinks he will receive the answer "yes." If the

subject believes that his current hypothesis is likely to be correct, then he will also think it likely that he will receive the answer "yes." If he also believes that should his hypothesis be wrong, the answer "yes" is very unlikely, then once he receives the answer "yes" he will have a very strong belief that his current hypothesis is correct. Finally, if he receives the answer "no," he can definitively rule out his current hypothesis; he, of course, believes this is unlikely to occur. With these beliefs, the strategy of testing instances to which he expects to receive the answer "yes" is not a bad one. Rather than attribute his error to his strategy, one might better attribute it to his incorrect beliefs that his current hypothesis is very likely to be correct and that a "yes" answer is unlikely if his hypothesis is wrong.

Later history. Other researchers have attempted to study this phenomenon using less artificial tasks. Snyder and Swann (1978), for example, performed a series of experiments asking subjects to test whether someone is extraverted (other subjects were asked to test introverted) by selecting 12 questions from a list of 26. Of the 26 questions, 11 were extraverted questions, that is, questions that "would typically be asked of people *already known* to be extraverts, for example, 'What would you do if you wanted to liven up things at a party?'"; 10 were introverted questions, "for example, 'In what situations do you wish you could be more outgoing?'"; and 5 were questions for which there was no consensus or which were classified as irrelevant, "for example, 'What kinds of charities do you like to contribute to?'" Subjects asked to test for extraversion used more extravert questions than subjects asked to test for introversion, even when told that extraversion was unlikely to be true, and even when offered monetary incentives for accurate testing. Snyder and Campbell (1980) found similar results.

The questions labeled as confirming questions in this study are those that corresponded to the category for which the subject was asked to test. However, these differ crucially from the similar questions in Wason's experiment. In Snyder and Swann's study, even an introverted person would probably answer with suggestions on how to liven up a party. An introvert could refute the subject's hypothesis only by uncooperatively refusing to accept the assumption "if you wanted to liven up things at a party." Under any of the possible hypotheses, this question is likely to generate similar answers; this is not the case for Wason's experiment. In Wason's experiment, if a subject tested an instance for which he expected the answer "yes," then if his current hypothesis were wrong, he could receive a disconfirming answer. In Wason's experiment the correct hypothesis (increasing numbers) does give the same answers as the subjects' incorrect hypothesis (numbers increasing by 2) to instances likely to be generated by the subjects, but this fact is not known by the subjects. The error made by subjects in Snyder and Swann's study is thus to use questions which they should know are very unlikely to generate disconfirming evidence under any

hypothesis. Wason's subjects did not make this error.

The following tables show the possible situations that might occur in Wason's experiment and in Snyder and Swann's experiment.

Wason's experiment:	Subject "yes"	receives "no"
Subject expects "yes"	A	B
"no"	C	D

Snyder and Swann's experiment:	Subject extraverted	receives introverted
Subject expects extraverted answer	A	B
introverted answer	C	D

(Subject is testing for extraversion.)

The confirmation bias is that subjects tend to ask questions in the first row in each experiment. For subjects who formulated the hypothesis numbers increasing by 2, situation B is impossible since this hypothesis is a special case of the true rule, increasing numbers; however, the subject does not know this. While in Snyder and Swann's experiment, situation B is very unlikely, the subject can realize this.

Some authors questioned the methodology of the experiments by Snyder and Swann and by Snyder and Campbell. "As there were only 5 neutral questions in the set of 26 possibilities, subjects still had to choose many nonneutral ones in their set of 12 test questions" (Fischhoff and Beyth-Marom, 1983). Trope, Bassok, and Alon (1984) suggested that several characteristics of Snyder and Swann's task, including the presence of biased questions, may have led to subjects to think "that his/her task is to focus on the hypothesized trait and make discrimination within it, that (s)he is to discriminate between those who match the description of, say, the fully extroverted person and those possessing more moderate levels of extroversion." For the subject who realized that what we call situation B was very unlikely, Trope *et al.*'s suggestion might be a reasonable inference about why the experimenter included biased questions.

Trope *et al.* (1984) allowed subjects to spontaneously formulate questions and concluded: "The strategy emerging from the kinds of questions our subjects formulated is clearly at variance with the confirmatory strategy. ... Out of the 586 questions formulated by the Israeli and Canadian samples, our judges were able to find only two questions that

assumed that the respondent possessed the hypothesized trait." Dallas and Baron (1985) found a confirmation bias among psychotherapists when replicating Snyder and Swann's constrained choice, but did not find this when the subjects designed their own questions.

Strohmer and Newman (1983) were concerned about the implications of the confirmatory strategy for the questioning strategies counselors might use. However, after a series of experiments, they concluded: "Our results do not support the suggestion that counselors preferentially seek information to confirm hypotheses about clients. ... When asked to write questions to test a hypothesis about a target client, all participants constructed a strategy favoring unbiased questions. When asked to test a hypothesis about a client by asking them questions, all but 3 of the 40 participants used an unbiased strategy." Furthermore, those 3 used a disconfirmatory strategy. Sackett (1982) replicated several of Snyder and Swann's (1978) studies as well as others, but used campus recruiters rather than college students as subjects. Sackett did not find consistent use of the confirmatory strategy.

Subjects may use a diagnostic strategy rather than a confirmatory strategy. "Trope and Bassok [1982] suggested that lay interviewers test their hypothesis by selecting questions according to their diagnostic value, i.e., the extent to which the probabilities of the answers depend on whether or not the respondent possesses the hypothesized trait" (Trope and Bassok, 1983). Support for this theory is provided by experiments by Trope and Bassok (1983), Trope, Bassok and Alon (1984), and Skov and Sherman (1986). "Subjects preferred information that was most useful for distinguishing between the hypothesis and the alternative. ... In fact, diagnosticity was the main determinant of question selection. Given a choice between a high diagnostic and hypothesis disconfirming question vs a low diagnostic and hypothesis confirming question, subjects almost always chose the former" (Skov and Sherman, 1986).

Swann and Giuliano (1987) replied to criticisms of the Snyder and Swann 1978 study. They replicated the experiment of Snyder and Swann but, like Trope and Bassok, asked subjects to write questions, rather than to choose from a list. They found that subjects are more likely to probe for evidence of dominance (rather than submissiveness) when they were testing for dominance and more likely to probe for evidence of extraversion (rather than introversion) when they were testing for extraversion. Strikingly, whether testing extraversion or introversion, subjects were more likely to probe for extraversion; similarly, whether testing dominance or submissiveness, subjects were more likely to probe for dominance.

Mynatt, Doherty, and Tweney (1977) asked subjects to discover the rules governing the dynamics of moving particles in a computer generated system. The correct answer was that

dim figures (but not bright ones) stopped particles. Subjects were led to form an initial hypothesis that triangles stop particles (other shapes were squares and circles). Subjects were then given 10 pairs of choices in which they could pick 1 of 2 experiments. We can form a table similar to our previous ones to discuss the subjects' choice of experiment:

		Subject	receives: particle
Mynatt *et al.'s* experiment:		stops	continues
	Subject expects: particle stops	A	B
	continues	C	D

Mynatt *et al.* labeled as confirmatory selections of experiments in the first row in which situation B was impossible; they found that subjects with the initial triangle hypothesis made 71% such choices, significantly different from the 50% expected by chance. However, as in Wason's experiment, subjects did not know situation B was impossible; choices counted as confirmatory were constructed so that the hypothesis "triangles stop particles" was a special case of the correct rule "dim figures stop particles."

In an effort to create a more realistic task, Mynatt, Doherty, and Tweney (1978) used a more complex environment and let subjects explore in a less constrained manner. Objects could be of three different shapes, three different sizes, and three different brightnesses. At certain points determined by the objects in its vicinity a moving particle was deflected at an angle. The universe contained all 27 possible objects, but subjects could view only a portion of the universe at once although the behavior of a particle in that portion could be influenced by objects off the screen. Again, subjects were told to discover the rules of particle motion. The correct rules were complicated; the angle of deflection depended on size and brightness of nearby objects, but not on shape.

The space of hypotheses in this experiment is huge; for example, one subject formed hypotheses based on lines between objects, a feature which was, in fact, irrelevant. No subject solved the system; judges ranked subjects' performances on how close they came to the correct rules.

Mynatt *et al.* remark:
"Outcomes of hypothesis tests were categorized after the fact by the experimenters. It was often difficult to determine unambiguously what type of evidence subjects were *seeking* (as opposed to what they actually got). There was, however, almost no indication whatsoever that they intentionally sought disconfirmation. In those few cases where there was enough information in the protocols to allow a judgment to be made, subjects almost always

seemed to be seeking confirmation."

Mynatt *et al.* do not show that this behavior led to poor performance. They did find that even though subjects sought confirmation, they received disconfirmation much more often. Thus, although subjects tended to choose questions in the first row of our tables, they most often ended up in situation B. Performance did depend on how well the subject handled the information received in such a situation; this question is discussed in a later section of this paper on hypothesis fixation.

In Wason's task, subjects were forced to generate a hypothesis about the correct rule and then test the hypothesis by generating instances. Subjects failed to reach the correct conclusion because they tended to (1) generate a hypothesis that was a special case of the correct solution (for example, numbers increasing by two) and (2) test only instances for which they expected the answer "yes." The second behavior is the confirmation bias, but the first behavior can be viewed as a property of the problem rather than of the subjects. Klayman and Ha 1987 accepted that the confirmation bias exists (they call it the *positive test strategy*) but argued that Wason's task is unnatural because the nature of the correct solution leads subjects to form a hypothesis of the wrong "size." Usually one is seeking to define a minority phenomenon, rather than such a general one as any three increasing numbers. Thus, Klayman and Ha concluded: "Under commonly occurring conditions, this strategy can be well suited to the basic goal of determining whether or not a hypothesis is correct."

Other authors have noted that the norm against which this error is defined is that of the Popperian view of science, that is, that theories should be tested by attempts to disconfirm them. This view is not universally accepted as the appropriate procedure in science and thus is not unassailable as a norm against which naive subjects should be compared.

Summary. For the confirmation bias to occur on a task, people have to use a hypothesis generation and testing strategy. There is abundant evidence that, on many artificial and realistic tasks, people do use such a strategy. In addition, for the confirmation bias to occur, people must use what Klayman and Ha called a positive test strategy. They must test a hypothesis by looking for data that are expected to be present if that hypothesis is correct. There is mixed evidence regarding the occurrence of this behavior. On some tasks, it appears to occur, while on others it does not.

Even when both of these conditions are present, the efficiency of the positive test strategy depends on the relationships among the potential hypotheses and the data with which they are consistent. In some tasks, a positive test strategy can lead to efficient confirmation of a correct hypothesis or efficient disconfirmation of an incorrect hypothesis (Smith, Giffin, Rockwell, and Thomas, 1986); in other tasks, certain hypotheses, when

combined with the use of a positive test strategy, will lead a subject to collect only data that cannot reject the subject's hypothesis.

Whatever information a subject seeks, he may receive disconfirming information, depending on the relationship between his hypothesis and the correct solution. The next two errors deal with subjects' behavior in such situations.

Differential treatment of positive and negative information

Definition. Bruner *et al.* (1956) noted: "[A] general tendency is the inability or unwillingness of subjects to use efficiently information which is based on negative instances or derives from indirect test of an hypothesis." People process information more efficiently if it is presented in a positive form (for example, an event occurred) than if it is presented in a negative form (an event did not occur). This can be labeled an error in that a subject does not use all the information he could from a negative statement.

Early history. There is a long history of experiments on the learning of concepts from positive and negative instances. Smoke (1933) found that a concept can be learned equally quickly from both positive and negative instances as from positive instances only but that learning from positive and negative instances discouraged learning of incorrect concepts compared to learning from positive instances only. Smoke did not test subjects using only negative instances but Hovland and Weiss (1953) did. They found that "a significantly greater percentage of subjects working with positive instances than of subjects working with negative instances were able to identify the concepts correctly. This was true even when the total number of instances presented was equal for both groups and when the memory factor was eliminated" (Freibergs and Tulving, 1961).

Freibergs and Tulving (1961) found that practice had a very large effect on the time required to learn concepts with positive instances and with negative instances. They trained two groups of subjects. The P Group was trained to solve problems using positive instances only, while the N Group used negative instances only. The P Group began with a large advantage over the N Group, but after 20 trials, "the difference between the two medians is quite small, especially when compared with the initial difference." It also appeared that the P Group was leveling off in median performance while the N Group was still improving.

Later history. Newman, Wolff, and Hearst (1980) report on a series of experiments which show the feature-positive effect. For example, one group of subjects had to learn the concept that Good patterns included one diamond among 4 symbols; another group had to learn the concept that Not Good patterns included one diamond among 4 symbols. The average number of trials required to learn the correct solution was significantly higher for the second group. However, if all subjects were told "something about each of the

Not Good cards makes them Not Good" then the average performance of the two groups reversed.

A similar effect was found by Tweney *et al.* (1980). They tried various strategies to teach subjects to use a disconfirmatory strategy to solve Wason's 2-4-6 task. While they found effects on strategy, only one change in instructions led to improved performance. Rather than referring to one rule and labeling instances offered by the subject as "conforming" or "not conforming," the experimenter said there were two rules and labeled instances as "DAX" or "MED." The results were dramatic; in Wason's original experiments and various replications, 16 out of 98 subjects solved the problem with the first rule announced while with the variation in wording 21 out of 37 did so.

Most of the research on this effect has required the subject to identify which conjunctive hypothesis is correct from a set of possible conjunctive hypotheses. Many researchers have noted that it is usually the case that a positive instance allows the elimination of many more conjunctive hypotheses than is allowed by a negative instance. It is not at all surprising that people attend to and find it easier to use positive information in this context (see, for example, Nahinsky and Slaymaker, 1970). Bourne and Guy (1968) presented evidence that subjects are better able to process information that comes from a smaller and more homogeneous set. Whether the set of positive or negative information has that property depends on the nature of the problem.

Markus and Zajonc (1985) reviewed evidence for a *negativity* bias in which negative information is given greater weight in comparing alternatives. They concluded that such an effect exists but, like Bourne and Guy, question whether it is always an error. "Negativity is a bias to the extent that in the drawing of inferences negative information is given more weight on the a priori false grounds that it leads to valid inferences or better decisions. However, if in general negative items are in fact more informative than positive items, then negativity may well be regarded not as a bias but as a useful and defensible heuristic that serves the individual well on appropriate occasions."

In another task setting, Evans (1982) reviewed a series of experiments in which subjects were required to either verify or construct sentences that were either affirmative or negative, and either true or false. While results concerning true and false sentences were ambiguous, Evans found that "[t]he one undisputed result that all these experiments (and similar subsequent ones) have in common is that negatives are reliably more difficult to process." Similarly, Taplin (1975) used response time data to test the idea that positive information is easier to process than negative information. He asked subjects to classify an instance as a positive or negative instance according to the concept under test. He looked at four types of concepts: conjunction (A and B), inclusive disjunction (A or B), conditional (if A,

then B), and biconditional (A if and only if B). Errors were very infrequent. The average response time across all types was less for positive instances than for negative instances, but this varied by rule. The average time was less for positive instances for conjunction and more for positive instances for conditional; no significant difference was found on inclusive disjunction or biconditional.

Christensen-Szalanski and Bushyhead (1981) asked physicians to estimate the probability that a patient has pneumonia based on the presence or absence of particular symptoms. They regressed the subjective probability on the actual probability for symptoms present and symptoms absent and obtained a positive slope, although the coefficient was smaller for absent symptoms than for present symptoms. However, they note that if one outlier (out of 45 points) is omitted, the two coefficients are nearly equal. They concluded that "physicians in this study appear to use the absence of a clinical finding as efficiently as the presence of clinical finding when estimating the predictive value of the finding." The authors note that this result may be due to the fact that the procedures of their experiment (in particular, the use of a checklist) may have called physicians' attention to the absence of a symptom.

Summary. There is evidence that negative information is more difficult than positive information for humans to use both in concept learning and in reaching conclusions, but it is unclear whether that is due to the fact that positive information often has more useful information. Also, even though the response time data indicated that negative information takes longer and thus may be more difficult to process, errors were infrequent.

Hypothesis fixation

Definition. Hypothesis fixation occurs when a subject maintains a hypothesis that has been demonstrated conclusively to be false.

Early history. Wason's 2-4-6 experiment, one of the original works in the confirmation bias, is also one of the original works in hypothesis fixation. "After the announcement of an incorrect rule it would be expected, on a rational basis, that a high proportion of the immediately succeeding series would be inconsistent with the rule just announced, ie subjects would relinquish their hypotheses. In fact, over the entire experiment, sixteen such series were consistent and fifteen inconsistent with the rule announced. This experiment strongly suggested that the subjects were either unwilling, or unable to eliminate their hypotheses in this task."

Later history. Mynatt, Doherty, and Tweney (1977) performed a series of experiments (discussed earler in relation to the confirmation bias) in which they asked subjects to formulate and test hypotheses about the effects of objects of various shapes and brightness on the behavior of particles in a computer simulation. They found that subjects failed to

consider alternative hypotheses, but were able to use falsifying data if it were present.

In their 1978 study using a more complex environment, subjects received many more disconfirmations of hypotheses than confirmations. "[M]ost subjects showed a strong tendency to discount or disregard falsification. Of the 88 hypothesis tests ... which resulted in disconfirmation, only 26 led to the permanent abandonment of the hypothesis. Following the remaining 62 tests, the subjects either abandoned the hypothesis but returned to it later, revised it to attempt to account for the anomalous data, or simply ignored the disconfirmation and went on testing the same hypothesis."

The test strategy used seemed to have less impact on the success rate than the goodness of the initial hypothesis generated. From the 1977 study: "In all of our studies, there have been subjects who very quickly picked the correct hypothesis, tested it once or twice, and announced the rule, completing the task in very short time. Some of the subjects used confirmation and some used disconfirmation. 'Luck' may well be, ironically, the best predictor of success, as it seemed to be in one of our earlier studies." In some cases, subjects did better if they ignored disconfirming evidence. In commenting on the 1977 study, Tweney, Doherty, Worner, Pliske, Mynatt, Gross, and Arkkelin (1980) were surprised to find "that subjects who focused on disconfirmatory evidence could be led seriously astray. Nor were we prepared to find that extensive searches for confirmatory evidence, and some discounting of disconfirmatory evidence, would characterize the most *successful* subjects. But that is what happened." Mynatt, Doherty, and Tweney (1978) commented: "Occasionally, a subject would hit on a promising, but partially incorrect, hypothesis. Since the hypothesis was only partially correct, it would sooner or later receive clear disconfirmation. Three different subjects quickly abandoned promising hypotheses and began testing others which were, in fact, much less close to the 'truth.'"

Some authors have questioned whether hypothesis fixation should be viewed as an error. Cohen (1981) said: "retention of a falsified hypothesis would even be desirable if it explained quite a lot of the evidence and no unfalsified hypothesis were available that had as good explanatory value." Nisbett and Ross (1980) discussed the difficulty of knowing when a subject should abandon a hypothesis. Such a statement "presupposes that we can determine when the person is 'inappropriately' persisting in an impression or belief whose basis has been undermined."

Tweney *et al.* (1980) suggested that a complicated inference task first requires formulating a set of good hypotheses "in the sense that they must possess at least some correspondence to reality to be worthwhile candidates for test. Thus, seeking confirmatory data and ignoring disconfirmatory data may be the most appropriate strategy early in the inference process, since it could then produce good hypotheses." They found that

successful subjects did seem to use this strategy.

Biased assimilation, the interpretation of neutral or negative evidence as being supportive of one's already held beliefs, can be viewed as a weaker form of hypothesis fixation. In hypothesis fixation, a person holds onto a belief in the face of logically contradictory evidence, while in biased assimilation the evidence may not obviously contradict the subject's belief, but it also does not support it. However, biased assimilation also includes the idea that the person misinterprets information presented to him.

Skov and Sherman (1986) discuss "the case where a hypothesis tester gathers information and biases the interpretation of that information so that the hypothesis appears to be true. For example, Lord, Ross, and Lepper (1979) demonstrated that subjects' initial beliefs in the use of capital punishment led them to bias their interpretation of evidence that was objectively neutral with respect to capital punishment. In other words, subjects on both sides of the issue ended up believing more than ever that their initial hypotheses were true. Likewise, Regan, Straus, and Fazio (1974) showed how subjects bias their attributions for success and failures by friends and enemies so that their hypotheses about the traits of these people are confirmed."

Lord *et al.* (1979) also found that subjects rated as more sound studies that supported their initial beliefs. The authors comment, however, that such a rating is rational; studies that agree with a known truth should be given more credence. "Our subjects' main inferential shortcoming, in other words, did not lie in their inclination to process evidence in a biased manner. ... Rather, their sin lay in their readiness to use evidence already processed in a biased manner to bolster the very theory or belief that initially 'justified' the processing bias."

Summary. It is difficult to demonstrate the extreme case of a subject's maintaining a clearly refuted conclusion since the degree of refutation can be questioned. Such a demonstration would be even more difficult in real situations, where conclusively refuting evidence is rarely present or can always be interpreted as not being as decisive as might seem.

Conclusions

For many of the errors discussed in this review, significant doubts have been raised about whether the behavior is common. For example, in the results regarding sensitivity to sample size, estimation of covariation by relying only on positive hits, the confirmation bias, and the testing of a conditional rule, there is considerable variation across subjects and considerable variation among experimental results. It is difficult to use the literature on these subjects to conclude that the existence of a bias has been established.

In other cases, the existence of a behavior has been established but significant doubts

have been raised about whether the subject has made an error. Often it can be argued that subjects have made a reasonable interpretation of the experimental conditions presented to them and have made reasonable inferences, even if the interpretation or the inference do not match those intended by the experimenter. Such an argument has been made for base rates and for attribution errors. In the illusion of control and in overprediction, the experimenter and subject may disagree about the underlying model of the process generating the data.

In other cases, it has been established that the behavior occurs and that subjects make an unreasonable interpretation of the problem, but their interpretation is reasonable for a more realistic version of the problem. For example, in conservatism in revising probabilities, subjects may implicitly assume dependence in data, even though the experimental conditions make clear this is not the case. In more realistic conditions this assumption may be appropriate. In illusory correlation, subjects seem to disregard instructions to base their judgments only on data presented to them; in real situations, they are correct in incorporating previous information. Indeed, subjects are labeled as ignoring a base rate if they refuse to use prior information in other contexts. In overprediction, the experimenter has created a situation in which the subject's natural tendency to seek explanation of seemingly random variation is incorrect; however, such a tendency is appropriate in many real situations. In hypothesis fixation, it seems difficult to determine when such fixation is appropriate and when it is inappropriate; the experimenter has the advantage over the subject in knowing the correct hypothesis, but often the subject's behavior is not unreasonable without such knowledge.

Some biases may operate primarily when subjects are ignorant of the subject. For example, people generally give reasonable probability judgments when they have appropriate substantive knowledge on which to base their judgments. There is some evidence that miscalibration falls in this category also.

Several of the biases reviewed here do seem to exist and do seem to be errors. Hindsight bias does seem to occur, although it is unclear whether this is a significant cause of error in real situations. Subjects do seem to have more difficulty in processing negatively presented information, but they seem to be able to overcome the difficulty in order to reach correct conclusions. In the conjunctive fallacy, subjects may inappropriately reverse the order of conditioning.

Thus, researchers should use caution in referring to well known biases. The existence of the bias may not have been established or it may occur very infrequently. Further caution should be used in referring to such biases as errors. In discussing conservatism in revising probabilities, Wickens (1984) calls the assumption of interdependence among

data a "rational and legitimate bias." This phrase could well be applied to many of the phenomena discussed in this paper.

Finally, this review suggests that, to make greater progress in the understanding of human error, we need to focus more attention on modeling the cognitive processes that cause particular errors, and on the identification of the contexts in which these processes are likely to be active. In this way, it may be possible to unravel the seeming contradictions and confusions present in the literature on different categories of errors.

References

1. Adams, Joe K. (1957). A confidence scale defined in terms of expected percentages. *American Journal of Psychology,* **70**, 432-436.
2. Adler, Jonathan E. (1983). Human rationality: essential conflicts, multiple ideals. *The Behavioral and Brain Sciences,* **6**, 245-246.
3. Alker, Henry A., and Hermann, Margaret G. (1971). Are Bayesian decision artificially intelligent? The effect of task and personality on conservatism in processing information. *Journal of Personality and Social Psychology,* **19**, 31-41.
4. Alloy, Lauren B., and Abramson, Lyn Y. (1982). Learned helplessness, depression, and the illusion of control. *Journal of Personality and Social Psychology,* **42**, 1114-1126.
5. Alloy, Lauren B., and Tabachnik, Naomi. (1984). Assessment of covariation by humans and animals: The joint influence of prior expectations and current situational information. *Psychological Review,* **91**, 112-149.
6. Argote, Linda, Seabright, Mark A., and Dyer, Linda. (1986). Individual versus group use of base-rate and individuating information. *Organizational Behavior and Human Decision Processes,* **38**, 65-75.
7. Arkes, Hal R., Christensen, Caryn, Lai, Cheryl, and Blumer, Catherine. (1987). Two methods of reducing overconfidence. *Organizational Behavior and Human Decision Processes,* **39**, 133-144.
8. Arkes, Hal R., Faust, David, Guilmette, Thomas J., and Hart, Kathleen. (1988). Eliminating the hindsight bias. *Journal of Applied Psychology,* **73**, 305-307.
9. Arkes, Hal R., and Harkness, Allan R. (1983). Estimates of contingency between two dichotomous variables. *Journal of Experimental Psychology: General,* **112**, 117-135.
10. Arkes, Hal R., Wortmann, Robert L., Saville, Paul D., and Harkness, Allan R. (1981). Hindsight bias among physicians weighing the likelihood of diagnoses. *Journal of Applied Psychology,* **66**, 252-254.
11. Bar-Hillel, Maya. (1983). The base rate fallacy controversy. In Scholz, R.W. (editor). *Decision Making Under Uncertainty,* New York: Elsevier Science Publishers.
12. Bar-Hillel, Maya. (1973). On the subjective probability of compound events. *Organizational Behavior and Human Performance,* **9**, 396.
13. Bar-Hillel, M. (1979). The role of sample size in sample evaluation. *Organizational Behavior and Human Performance,* **24**, 245-57.
14. Barclay, Scott, and Beach, Lee Roy. (1972). Combinatorial properties of personal probabilities. *Organizational Behavior and Human Performance,* **8**, 176-183.

15. Beyth-Marom, Ruth. (1982). Perception of correlation reexamined. *Memory and Cognition,* **10**, 511-519.
16. Birnbaum, Michael H. (1983). Base rates in Bayesian inference: Signal detection analysis of the cab problem. *American Journal of Psychology,* **96**, 85-94.
17. Birnbaum, M.H., and Mellers, B.A. (1983). Bayesian inference: combining base rates with opinions of sources who vary in credibility. *Journal of Personality and Social Psychology,* **45**, 792-804.
18. Borak, Jonathan, and Veilleux, Suzanne. (1982). Errors of intuitive logic among physicians. *Social Science and Medicine,* **16**, 1939-47.
19. Bourne, Lyle E., Jr., and Guy, Donald E. (1968). Learning conceptual rules: II. The role of positive and negative instances. *Journal of Experimental Psychology,* **77**, 488-494.
20. Bower, Gordon H., Black, John B., and Turner, Terrence J. (1980). Scripts in memory for text. In *Human memory. Contemporary readings.* New York: Oxford University Press.
21. Brown, Carole, Keats, John A., Keats, Daphne M., and Seggie, Ian. (1980). Reasoning about implications: A comparison of Malaysian and Australian subjects. *Journal of Cross-Cultural Psychology,* **11**, 395-410.
22. Bruner, Jerome S., Goodnow, Jacqueline J., and Austin, George A. (1956). *A Study of Thinking.* New York: John Wiley & Sons, Inc.
23. Chapman, Loren J. (1967). Illusory correlation in observational report. *Journal of Verbal Learning and Verbal Behavior,* **6**, 151-155.
24. Chapman, L.J., and Chapman, J.P. (1959). Atmosphere effect re-examined. *Journal of Experimental Psychology,* **58**, 220-226.
25. Chapman, L.J., and Chapman, J.P. (1967). Genesis of popular, but erroneous psychodiagnostic observation, *Journal of Abnormal Psychology,* **72**, 93-204.
26. Chapman, L.J., and Chapman, J.P. (1969). Illusory correlation as an obstacle to the use of valid psychodiagnostic Signs. *Journal of Abnormal Psychology,* **74**, 271-290.
27. Cheng, Patricia W., Holyoak, Keith J. (1985). Pragmatic reasoning schemas. *Cognitive Psychology,* **17**, 391-416.
28. Cheng, Patricia W., Holyoak. Keith J., Nisbett, Richard E., and Oliver, Lindsay M. (1986). Pragmatic versus syntactic approaches to training deductive reasoning. *Cognitive Psychology,* **18**, 293-328.
29. Christensen-Szalanski, J.J.J., Beck, Don E., Christensen-Szalanski, Carlyn M., and Koepsell, Thomas D. (1983). The effects of expertise and experience on risk judgments. *Journal of Applied Psychology,* **68**, 278-284.

30. Christensen-Szalanski, Jay J.J., and Bushyhead, James B. (1981). Physicians' use of probabilistic information in a real clinical setting. *Journal of Experimental Psychology,* **7**, 928-935.
31. Cohen, John, and C.E.M Hansel. (1958). The nature of decisions in gambling: equivalence of single and compound subjective probabilities. *Acta Psychologica,* **13**, 357-370.
32. Cohen, L. Jonathan. (1981) Can human irrationality be experimentally demonstrated? *The Behavioral and Brain Sciences,* **4**, 317-370.
33. Cohen, L.J. (1979). On the psychology of prediction: Whose is the fallacy? *Cognition,* **7**, 385-407.
34. Dallas, Mercedes E. Wilson, and Baron, Robert S. (1985). Do psychotherapists use a confirmatory strategy during interviewing? *Journal of Social and Clinical Psychology,* **3**, 106-122.
35. Dawson, Neal V., Arkes, Hal R., Siciliano, Carl, Blinkhorn, Richard, Lakshamanan, Mark, and Petrelli, Mary. (1988). Hindsight bias: An impediment to accurate probability estimation in clinicopathologic conferences. *Medical Decision Making,* **8**, 259-264.
36. Dershkowitz, A. (1971). Imprisonment by judicial hunch. *American Bar Association Journal,* **57**, 560-564.
37. Eddy, David M. (1982) Probabilistic reasoning in clinical medicine: problems and opportunities. In Kahneman, Daniel, Slovic, Paul, and Tversky, Amos (editors). *Judgment Under Uncertainty: Heuristics and biases.* Cambridge: Cambridge University Press.
38. Edwards, Ward. (1968). Conservatism in human information processing. In Kleinmutz, Benjamin (editor). *Formal Representation of Human Judgment.* New York: Wiley.
39. Einhorn, Hillel J., and Hogarth, Robin M. (1981). Behavioral decision theory: Processes of judgment and choice. *Annual Review of Psychology,* **32**, 53-88.
40. Elstein, Arthur S. (1976). Clinical judgment: psychological research and medical practice. *Science,* **194**, 696-700.
41. Elstein, Arthur S., and Van Pelt, John D. (1969). Assumed similarity in staff perception of psychiatric patients. *Journal of Clinical Psychology,* **25**, 6-97.
42. Evans, J.St.B.T. (1980). Current issues in the psychology of reasoning. *British Journal of Psychology,* **71**, 227-239.
43. Evans, Jonathan St.B.T. (1982). *The psychology of deductive reasoning.* London: Routledge & Kegan Paul.

44. Evans, B.T., and Dusoir, A.E. (1977). Proportionality and sample size as factors in intuitive statistical judgment. *Acta Psychologica,* **41**, 129-137.
45. Fischhoff, B. (1975). Hindsight does not equal foresight: The effect of outcome knowledge of judgment under uncertainty. *Journal of Experimental Psychology: Human Perception and Performance,* **1**, 288-299.
46. Fischhoff, B. (1977). Perceived informativeness of facts. *Journal of Experimental Psychology: Human Perception Performance,* **3**, 349-358.
47. Fischhoff, B., and Beyth-Marom, R. (1983). Hypothesis evaluation from a Bayesian perspective. *Psychological Review,* **90**, 239-260.
48. Fischhoff, Baruch, and MacGregor, Don. (1982). Subjective confidence in forecasts. *Journal of Forecasting,* **1**, 155-172.
49. Fischhoff, Baruch, Slovic, Paul, and Lichtenstein, Sarah. (1978). Fault trees: sensitivity of estimated failure probabilities to problem representation. *Journal of Experimental Psychology: Human Perception and Performance,* **4**, 330-344.
50. Fischhoff, B., Slovic, P., and Lichtenstein, S. (1979). Subjective sensitivity analysis. *Organizational Behavior and Human Performance* **23**, 339-359.
51. Foltz, A.M., and Kelsey, J.L. (1978). The annual pap test: a dubious policy success. *Milbank Memorial Quarterly/Health and Society,* **56**, 426-462.
52. Freibergs, Vaira, and Tulving, Endel. (1961). The effect of practice on utilization of information from positive and negative instances in concept identification. *Canadian Journal of Psychology,* **15**, 101-106.
53. Funder, David C. (1982). On the accuracy of dispositional versus situational attributions. *Social Cognition,* **1**, 205-222.
54. Gilovich, Thomas, Vallone, Robert, and Tversky, Amos. (1985). The hot hand in basketball: On the misperception of random sequences. *Cognitive Psychology,* **17**, 295-314.
55. Golding, Stephen L., and Rorer, Leonard G. (1972). Illusory correlation and subjective judgment. *Journal of Abnormal Psychology,* **80**, 249-260.
56. Grice, H.P. (1975). Logic and conversation. In Davidson, D., and Harman, G. (editors). *The logic of grammar.* Dickensen.
57. Griggs, Richard A., and Cox, James R. (1983). The effect of problem content and negation on Wason's selection task. *Quarterly Journal of Experimental Psychology,* **35A**, 519-533.
58. Griggs, Richard A., and Cox, James R. (1982). The elusive thematic-materials effect in Wason's selection task. *British Journal of Psychology,* **73**, 407-420.
59. Hamilton, V. Lee. (1980). Intuitive psychologist or intuitive lawyer: alternative

models of the attribution process. *Journal of Personality and Social Psychology,* **39**, 767-772.

60. Harris, Richard J., and Monaco, Gregory E. (1978). Psychology of pragmatic implication: Information processing between the lines. *Journal of Experimental Psychology: General,* **107**, 1-22.
61. Harvey, John H., Town, Jerri P., and Yarkin, Kerry L. (1981). How fundamental is 'the fundamental attribution error'? *Journal of Personality and Social Psychology,* **40**, 346-349.
62. Harvey, John H., and Weary, Gifford. (1984). Current issues in attribution theory and research. *Annual Review of Psychology,* **35**, 427-459.
63. Henle, Mary. (1962). On the relation between logic and thinking. *Psychological Review,* **69**, 366-378.
64. Hoch, Stephen J., and Tschirgi, Judith E. (1985). Logical knowledge and cue redundancy in deductive reasoning. *Memory and Cognition,* **13**, 453-462.
65. Hovland, C.I., and Weiss, W.. (1953). Transmission of information concerning concepts through positive and negative instances. *Journal of Experimental Psychology,* **43**, 175-182.
66. Howell, William C. (1972). Compounding uncertainty from internal sources. *Journal of Experimental Psychology,* **95**, 6-13.
67. Howell, William C. (1971). Uncertainty from internal and external sources: a clear case of overconfidence. *Journal of Experimental Psychology,* **89**, 240-243.
68. Irwin, Francis W., W.A.S. Smith, and Jane F. Mayfield. (1956). Tests of two theories of decision in an 'expanded judgment' Situation. *Journal of Experimental Psychology,* **51**, 261-268.
69. Jenkins, Herbert M, and Ward, William C. (1965). Judgment of contingency between responses and outcomes. *Psychological Monographs: General and Applied,* **79**, Whole No. 594.
70. Johnson, Paul E., Ahlgren, Andrew, Blount, Joseph P., and Petit, Noel J. (1981). Scientific reasoning: garden paths and blind alleys. In Robinson, J.T. (editor). *Research in science education: New questions, new directions.* Louisville, CO: ERIC.
71. Johnson, Paul E., Hassebrock, Frank, Duran, Alicia S., and Moller, James H. (1982). Multimethod study of clinical judgment. *Organizational Behavior and Human Performance,* **30**, 201-230.
72. Johnson-Laird, P.N., Legrenzi, Paolo, and Legrenzi, Maria Sonino. (1972). Reasoning and a sense of reality. *British Journal of Psychology,* **63**, 395-400.
73. Jones, Edward E., and Nisbett, Richard E. (1971). The actor and the observer:

divergent perceptions of the causes of behavior. In Jones, Edward E., Kanouse, David E., Kelley, Harold H., Nisbett, Richard E., Valins, Stuart, and Weiner, Bernard (editors). *Attribution: perceiving the causes of behavior.* Morristown, NJ: General Learning Press.

74. Kahneman, Daniel, and Tversky, Amos. (1973). On the psychology of prediction. *Psychological Review,* **80**, 237-251.
75. Kahneman, Daniel, and Tversky, Amos. (1972). Subjective probability: a judgment of representativeness. *Cognitive Psychology,* **3**, 430-454.
76. Kelley, H.H., and Michela, J.L. (1980). Attribution theory and research. *Annual Review of Psychology,* **31**, 457-501.
77. Klayman, Joshua, and Ha, Young-Won. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review,* **94**, 211-228.
78. Koriat, Asher, Lichtenstein, S., and Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory,* **6**, 107-118.
79. Kruglanski, Arie W., and Ajzen, Icek. (1983). Bias and error in human judgment. *European Journal of Social Psychology,* **13**, 1-44.
80. Kunda, Ziva, and Nisbett, Richard W. (1986). The psychometrics of everyday life. *Cognitive Psychology,* **18**, 195-224.
81. Langer, Ellen J. (1975). The illusion of control. *Journal of Personality and Social Psychology,* **32**, 311-328.
82. Langer, E.J., and Roth, J. (1975). Heads I win, tails it's chance: the illusion of control as a function of the sequence of outcomes in a purely chance task. *Journal of Personality and Social Psychology,* **32**, 951-955.
83. Levi, Isaac. (1981). Should Bayesians sometimes neglect base rates? *The Behavioral and Brain Sciences,* **4**, 342-343.
84. Lichtenstein, S., Fischhoff, B., and Phillips, L. (1982). Calibration of probabilities; the state of the art. In Jungermann, H., and de Zeeuw, G. (editors). *Decision making and change in human affairs.* Dordrecht, The Netherlands: Reidel, 1977. Reprinted (in revised version) in Kahneman, Daniel, Slovic, Paul, and Tversky, Amos (editors). *Judgment under uncertainty: Heuristics and biases.* Cambridge: Cambridge University Press.
85. Little, Kenneth B., and Lintz, Larry M. (1965). Information and certainty. *Journal of Experimental Psychology,* **70**, 428-432.
86. Lord, Charles G., Ross, Lee, and Lepper, Mark R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology,* **37**, 2098-2109.

87. Lykken, D.T. (1975). The right way to use a lie detector. *Psychology Today,* **8**, 56-60.
88. Lyon, D., and Slovic, Paul. (1976). Dominance of accuracy information and neglect of base rates in probability estimation. *Acta Psychologica,* **40**, 287-298.
89. Manktelow, K.I., and Evans, J.St B.T. (1979). Facilitation of reasoning by realism: Effect or non-effect. *British Journal of Psychology,* **70**, 477-488.
90. Markus, Hazel, and Zajonc, R. B. (1985). The cognitive perspective in social psychology. In Lindzey, Gardner, and Aronson, Elliot (editors). *Handbook of social psychology* (3rd edition), Reading, MA, Addison-Wesley.
91. Mehle, Thomas. (1982). Hypothesis generation in an automobile malfunction inference task. *Acta Psychologica,* **52**, 87-106.
92. Mehle, Thomas, Gettys, Charles F., Manning, Carol, Baca, Suzanne, and Fisher, Stanley. (1981). The availability explanation of excessive plausibility assessments. *Acta Psychologica,* **49**, 127-140.
93. Messick, David M., Campos, Francis T. (1972). Training and conservatism in subjective probability revision. *Journal of Experimental Psychology,* **94**, 335-337.
94. Miller, Dale T. (1976). Ego involvement and attributions for success and failure. *Journal of Personality and Social Psychology,* **34**, 901-906.
95. Miller, Dale T., and Ross, Michael. (1975). Self-serving biases in the attribution of causality: fact or fiction? *The Psychological Bulletin,* **82**, 212-225.
96. Morier, Dean M., and Borgida, Eugene. (1984). The conjunction fallacy: A task specific phenomenon? *Personality and Social Psychological Bulletin,* **10**, 243-252.
97. Murphy, Allan H., and Winkler, Robert L. (1984). Probability forecasting in meteorology. *Journal of the American Statistical Association,* **79**, 489-500.
98. Mynatt, C.R., Doherty, M.E., and Tweney, R.D. (1977). Confirmation bias in a simulated research environment: An experimental study of scientific inference. *Quarterly Journal of Experimental Psychology,* **29**, 85-95.
99. Mynatt, C.R., Doherty, M.E., and Tweney, R.D. (1978). Consequences of confirmation and disconfirmation in a simulated research environment. *Quarterly Journal of Experimental Psychology,* **30**, 395-406.
100. Nahinsky, Irwin D., and Slaymaker, Frank L. (1970). Use of negative instances in conjunctive concept identification. *Journal of Experimental Psychology,* **84**, 64-84.
101. Navon, D. (1979) The importance of being conservative. *British Journal of Mathematical and Statistical Psychology,* **29**, 33-48.
102. Newman, Joseph, Wolff, William T., and Hearst, Eliot. (1980). The feature-positive effect in adult human subjects. *Journal of Experimental Psychology: Human Learn-*

ing and Memory, **6**, 630-650.

103. Nisbett, Richard, Fong, Geoffrey T., Lehman, Darrin R., and Cheng, Patricia W. (1987). Teaching reasoning. *Science*, **238**, 625-631.
104. Nisbett, Richard, and Ross, Lee. (1980) *Human inference: Strategies and shortcomings of social judgment.* Englewood Cliffs, NJ: Prentice-Hall, Inc.
105. Peterson, Cameron R., and Beach, Lee Roy. (1967). Man as an intuitive statistician. *The Psychological Bulletin,* **68**, 29-46.
106. Peterson, Cameron R., DuCharme, Wesley M., and Edwards, Ward. (1968). Sampling distributions and probability revisions. *Journal of Experimental Psychology,* **76,** 236-243.
107. Regan, Dennis T., Straus, Ellen, and Fazio, Russell. (1974). Liking and the attribution process. *Journal of Experimental Social Psychology,* **10**, 385-397.
108. Reich, Shuli S., and Ruth, Pauline. (1982). Wason's selection task: verification, falsification and matching. *British Journal of Psychology,* **73**, 395-405.
109. Ross, Lee D., Amabile, Teresa M., and Steinmetz, Julia L. (1977). Social roles, social control, and biases in social-perception processes. *Journal of Personality and Social Psychology,* **35**, 485-494.
110. Ross, Lee, and Anderson, Craig A. (1982). Shortcomings in the attribution process: on the origins and Maintenance of erroneous social assessments. In Kahneman, Daniel, Slovic, Paul, and Tversky, Amos (editors). *Judgment under uncertainty: Heuristics and biases.* Cambridge: Cambridge University Press.
111. Ross, Lee, Greene, David, and House, Pamela. (1977). The 'false consensus effect': An egocentric bias in social perception and attribution processes. *Journal of Experimental and Social Psychology,* **13**, 279-301.
112. Ross, Michael, and Sicoly, Fiore. (1979). Egocentric biases in availability and attribution. *Journal of Personality and Social Psychology,* **37**, 322-336.
113. Rumelhart, David E., and Norman, Donald A. (1981). Analogical processes in learning. In Anderson, John R. *Cognitive psychology and its implications.* San Francisco: W.H. Freeman and Company.
114. Sackett, Paul R. (1982). The interviewer as hypothesis tester: The effects of impressions of an applicant in interviewer questioning strategy. *Personnel Psychology,* **35**, 789-804.
115. Schum, David A. (1973). Concluding comments about the special issue on hierarchical inference. *Organizational Behavior and Human Performance,* **10**, 427-431.
116. Shaklee, Harriet, and Mims, Michael. (1981). Development of rule use in judgments of covariation between events. *Child Development,* **52**, 317-325.

117. Shaklee, Harriet, and Mims, Michael. (1982). Sources of error in judging event covariations: effects of memory demands. *Journal of Experimental Psychology: Learning, Memory and Cognition,* **8**, 208-224.
118. Shaklee, Harriet, and Tucker, Diane. (1980). A rule analysis of judgments of covariation between events. *Memory and Cognition,* **8**, 459-467.
119. Shanteau, J. (1978). When does a response error become a judgmental bias? Commentary on 'Judged frequency of lethal events.' *Journal of Experimental Psychology: Human Learning and Memory,* **4**, 579-581.
120. Sheridan, Thomas B. (1981). Understanding human error and aiding human diagnostic behavior in nuclear power plants. In Rasmussen, J., and Rouse, W. (editors). *Human detection and diagnosis of systems failures.* New York: Plenum Press.
121. Sheridan, Thomas B., and William R. Ferrell. (1974). *Man-machine systems: Information, control and decision models of human performance.* Cambridge, MA: The MIT Press.
122. Skov, Richard B., and Sherman, Steven J. (1986). Information-gathering processes: diagnosticity, hypothesis-confirmatory strategies and perceived hypothesis confirmation. *Journal of Experimental Social Psychology,* **22**, 93-121.
123. Slovic, Paul. (1969). Manipulating the attractiveness of a gamble without changing its expected value. *Journal of Experimental Psychology,* **79**, 139-145.
124. Slovic, Paul, Fischhoff, Baruch, and Lichtenstein, Sarah. (1980) Facts and fears: understanding perceived risk. In Schwing, Richard C., and Albers, Jr., Walter A. (editors). *Societal risk assessment: how safe is safe enough?* New York: Plenum Press, 181-214.
125. Smedslund, Jan. (1970). Circular relation between understanding and logic. *Scandinavian Journal of Psychology,* **11**, 217-219.
126. Smedslund, Jan. (1963). The concept of correlation in adults. *Scandinavian Journal of Psychology,* **4**, 165-173.
127. Smedslund, Jan. (1966). Note on learning, contingency, and clinical experience. *Scandinavian Journal of Psychology,* **7**, 265-266.
128. Smith, P., Giffin, W., Rockwell, T., and Thomas, M. (1988). Modeling fault diagnosis as the activation and use of frame system. *Human Factors,* **28**, 703-716.
129. Smoke, Kenneth L. (1933). Negative instances in concept formation. *Journal of Experimental Psychology,* **16**, 583-588.
130. Snyder, Mark, and Campbell, Bruce. (1980). Testing hypotheses about other people: The role of the hypothesis. *Personality and Social Psychology Bulletin,* **6**, 421-426.
131. Snyder, M. and Swann, Jr., W.B. (1978). Hypothesis-testing processes in social

interaction. *Journal of Personality and Social Psychology,* **36**, 1202-1212.

132. Spielman, Stephen. (1983). Kyburg on ignoring base rates. *The Behavioral and Brain Sciences,* **6**, 261-262.
133. Strohmer, Douglas C., and Newman, Lisa J. (1983). Counselor hypothesis-testing strategies. *Journal of Counseling Psychology,* **30**, 557-565.
134. Swann, William B., Jr., and Giuliano, Toni. (1987). Confirmatory search strategies in social interaction: how, when, why, and with what consequences. *Journal of Social and Clinical Psychology,* **5**, 511-524.
135. Taplin, John E. (1975). Evaluation of hypotheses in concept identification. *Memory and Cognition,* **3**, 85-96.
136. Thompson, Suzanne C., and Kelley, Harold H. (1981). Judgments of responsibility for activities in close relationships. *Journal of Personality and Social Psychology,* **41**, 469-477.
137. Trope, Yaacov, and Bassok, Miriam. (1982). Confirmatory and diagnosing strategies in social information gathering. *Journal of Personality and Social Psychology,* **43**, 22-34.
138. Trope, Yaacov, and Bassok, Miriam. (1983). Information-gathering strategies in hypothesis-testing. *Journal of Experimental Social Psychology,* **19**, 560-576.
139. Trope, Yaacov, Bassok, Miriam, and Alon, Eve. (1984). The questions lay interviewers ask. *Journal of Personality,* **52**, 90-106.
140. Tversky, Amos, and Kahneman, Daniel. (1971). Belief in the law of small numbers. *The Psychological Bulletin,* **2**, 105-110.
141. Tversky, Amos, and Kahneman, Daniel. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review,* **90**, 293-315.
142. Tversky, Amos, and Kahneman, Daniel. (1974). Judgment under uncertainty: Heuristics and biases. *Science,* **185**, 1124-1131.
143. Tversky, Amos, and Kahneman, Daniel. (1982). Judgments of and by representativeness. In Kahneman, Daniel, Slovic, Paul, and Tversky, Amos (editors). *Judgment under uncertainty: Heuristics and biases.* Cambridge: Cambridge University Press.
144. Tweney, T.D., Doherty, M.E., Worner, W.J., Pliske, D.B., Mynatt, C.R., Gross, K.A., and Arkkelin, D.L. (1980). Strategies of rule discovery in an inference task. *Quarterly Journal of Experimental Psychology,* **32**, 109-123.
145. Vertinsky, Patricia, Kanetkar, Vinay, Vertinsky, Ilan, and Wilson, Gail. (1986). Prediction of wins and losses in a series of field hockey games: a study of probability assessment quality and cognitive information-processing models of players. *Organi-*

zational Behavior and Human Decision Processes, **38**, 392-404.

146. Wallsten, Thomas S. (1983). The theoretical status of judgmental heuristics. In Scholz, R.W. (editor). *Decision making under uncertainty,* New York: Elsevier Science Publishers.
147. Ward, William C., and Jenkins, Herbert M. (1965). The display of information and the judgment of contingency. *Canadian Journal of Psychology,* **19**, 231-241.
148. Wason, P.C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology,* **12**, 129-140.
149. Wason, P.C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology,* **20**, 273-281.
150. Wason, P.C., and Johnson-Laird, P.N. (1972). *Psychology of reasoning: structure and content.* Cambridge, MA: Harvard University Press.
151. Wason, P.C., and Shapiro, Diana. (1971). Natural and contrived experience in a reasoning problem. *Quarterly Journal of Experimental Psychology,* **23**, 63-71.
152. Weary Bradley, Gifford. (1978). Self-serving biases in the attribution process: a re-examination of the fact or fiction question. *Journal of Personality and Social Psychology,* **36**, 45-71.
153. Wheeler, Gloria, and Beach, Lee Roy. (1968). Subjective sampling distributions and conservatism. *Organizational Behavior and Human Performance,* **3**, 36-46.
154. Wickens, Christopher D. (1984). *Engineering Psychology and Human Performance.* Charles E. Merrill Publishing Company: Columbus, OH.
155. von Winterfeldt, Detlof, and Edwards, Ward. (1986). *Decision analysis and behavioral research.* Cambridge, England: Cambridge University Press.
156. Wood, Gordon. (1978). The knew-it-all-along effect. *Journal of Experimental Psychology: Human Perception and Performance,* **4**, 345-353.
157. Yachanin, Stephen A., and Tweney, Ryan D. (1982). The effect of thematic content on cognitive strategies in the four-card selection task. *Bulletin of the Psychonomic Society,* **19**, 87-90.
158. Yates, Frank J., and Carlson, Bruce W. (1986). Conjunction errors: Evidence for multiple judgment procedures, including 'signed summation.' *Organizational Behavior and Human Performance,* **37**, 230-253.

A COGNITIVE TASK ANALYSIS OF ENROUTE FLIGHT PLANNING ACTIVITIES BY FLIGHT CREWS

Deb Galdes
Philip J. Smith

Cognitive Systems Engineering Lab
Department of Industrial and Systems Engineering
The Ohio State University
Columbus, Ohio 43210

ABSTRACT

Currently there is a great deal of interest in putting knowledge-based systems into the cockpit. But along with this new technology comes an old problem - pilot acceptance. Although most pilots favor more automation (Chambers & Nagel, 1985), they do not want systems which move them further out of the decision-making loop.

One solution that keeps pilots in the loop is knowledge-based systems that act as cooperative decision aids rather than autonomous decision makers. The design of such systems, though, is not an easy task. For a cooperative system, the human-computer interface becomes the critical component, linking the pilot to the underlying system.

We are exploring a methodology to aid in the design of these interfaces and underlying systems. This methodology relies on performing a cognitive task analysis of the current cockpit environment, prototyping potential interfaces to a cooperative computer-based decision aid, and evaluating and iteratively redesigning this interface. The context is an intelligent flight path routing aid.

This report discusses the results of Phase II of our study. The purpose was

1) to perform a cognitive task analysis of the flight planning activities engaged in by flight crews while enroute, and
2) to evaluate these crews' reactions to one possible interface - the Flight Plan Advisor (FPA).

The results indicate that the cognitive task analysis is a critical step in designing the human-computer interfaces of these types of systems. It gives us insights into the role of the crew, what information assists the crew in this role, and how the crew actually uses the information. These insights, then, give us an objective basis for generating truly useful cooperative decision aids.

1. INTRODUCTION

Although most pilots favor more automation (Chambers & Nagel, 1985), they do not want systems that move them further out of the decision-making loop. That is, they do not want systems that

- are difficult to understand and, therefore, to judge for reliability, and
- take away the pilot's control.

Currently there is a great deal of interest in using knowledge-based systems in the cockpit to help alleviate problems while keeping the pilot "in the loop." The assumption is that since knowledge-based systems are designed to more closely match the structure and level of the users' thought processes (Norman, 1986), pilots will find these systems easier to understand.

Building a knowledge-based system to help the pilot with a task, however, *does not automatically guarantee* that he/she will find the system acceptable. For example, recent research (Rouse, Geddes & Curry, 1987; Rouse, Rouse & Hammer, 1982; Woods, 1986) has shown that experienced users do not want knowledge-based systems which act as autonomous decision makers because these systems do not give their users enough control. An alternative paradigm which keeps pilots in the loop is a knowledge-based system which acts as a cooperative decision *aid* rather than an autonomous decision maker.

We are currently developing an interface to a cooperative decision aid. Our long range goals are to

- develop a set of guidelines for interfaces to cooperative systems and
- specify the interface for our particular system, a cooperative decision aid for replanning a flight path while enroute.

The short term goals of this student project are to

- explore methods for knowledge extraction from experts and
- use these techniques to acquire the appropriate knowledge from experts in our domain.

The results from Phase I of this project (Galdes, Smith & Chappell, 1989) indicated the following:

1) Although we can collect valuable information from pilots during structured interviews, we still need a detailed understanding of what goes on in the cockpit environment.
2) It is difficult for pilots to make suggestions and imagine an interface to a system because typically they are not familiar with the technology available.

Based on these results, the goals of Phase II were to

1) perform a cognitive task analysis of the flight planning activities engaged in by flight crews while enroute, and
2) evaluate these crews' reactions to one possible interface - the Flight Plan Advisor (FPA).

Our first goal was to collect data on crews during a flight that required replanning while enroute, and to perform a cognitive task analysis based on this data. Specifically, we were concerned with the following questions:

1) What role does the flight crew play in identifying a need for replanning, in generating flight amendments, and in evaluating alternative flight plans?
2) What information assists the flight crew in playing its role?
3) How is this information used?

This type of analysis is critical if we are to build effective computer aids for the flight crew. At present, the literature does not provide an adequate description of how flight crews, in coordination with Air Traffic Control (ATC) and Dispatch, cope with situations requiring replanning.

The second goal of this project was to evaluate the interface of a particular computer aid built by NASA staff (Taylor & Dammann, 1989) - the Flight Plan Advisor (FPA). The purpose of this evaluation was to gain insights into the issues associated with putting a flight planning aid in the cockpit.

The remainder of this report is organized as follows. Section 2 describes our data collection and analyses methods including a description of our flight scenario and the FPA interface. Section 3 gives an overview and examples of the enroute flight planning task. Sections 4 and 5 discuss the results of the cognitive task analysis and the evaluation of FPA, respectively. Finally, section 6 gives some conclusions from this specific study and discusses the value of these kinds of studies in general.

2. METHODS

As already stated, the goals of this study were to

1) perform a cognitive task analysis of crew members' enroute flight planning activities, and
2) evaluate the design of a flight planning aid, the Flight Plan Advisor (FPA) system.

To achieve these goals, we first designed a scenario that would give us insight into the role of the flight crew in the flight planning process and the information which the crew needs to support their role. This scenario was based on a standard flight from San Francisco to Detroit.

The second step was carried out by NASA personnel and focused on designing and implementing the FPA system. This system provided the crew with current weather information and suggestions for better flight plans when one was available.

Third, both the scenario and FPA system were used in full-mission simulations using the 727 motion-based simulator at NASA Ames Research Center.

Fourth, we analyzed videotapes of these simulations to build a model of the flight planning behavior of the crew while enroute. We also used these

videotapes and post-flight interviews to evaluate the crews' reactions to the FPA system.

2.1 The Scenario

To achieve our goals, we first designed a scenario which consisted of several typical situations that might cause a crew to deviate from their current flight plan. For example, the winds might not be as forecasted or a thunderstorm might appear along the crew's current routing. These situations, or "events," were designed to give us insight into the role of the flight crew in the replanning process and what information the crew needs to support their role. In addition, these events were also designed to create a slowly worsening fuel situation. This worsening fuel situation gave us insight into how the crew's role might change as the seriousness of their problem increases, and how the information they need might also change over time. Our scenario was a modification of the scenario we developed and tested in Phase I of this study (Galdes, Smith & Chappell, 1989).

The scenario consisted of a standard flight from San Francisco to Detroit and is graphically depicted in Figure 1. It included the following six events:

1) During climb-out, ATC holds the aircraft at FL250 because of other air traffic. The crew's flight plan, however, specifies a cruise altitude of FL290. ATC informs the crew that they may be held at this lower altitude for some time. This event immediately forces the crew off of their scheduled flight plan and increases their fuel burn.

2) The crew is being held at a cruise altitude of FL250 instead of their flight planned altitude of FL290. FPA receives updated wind information that shows the actual winds at FL290 (the original cruise altitude) are well below forecasted winds, and the winds at FL250 will give a significant savings of time and fuel. FPA, therefore, recommends to the crew that they should maintain their current

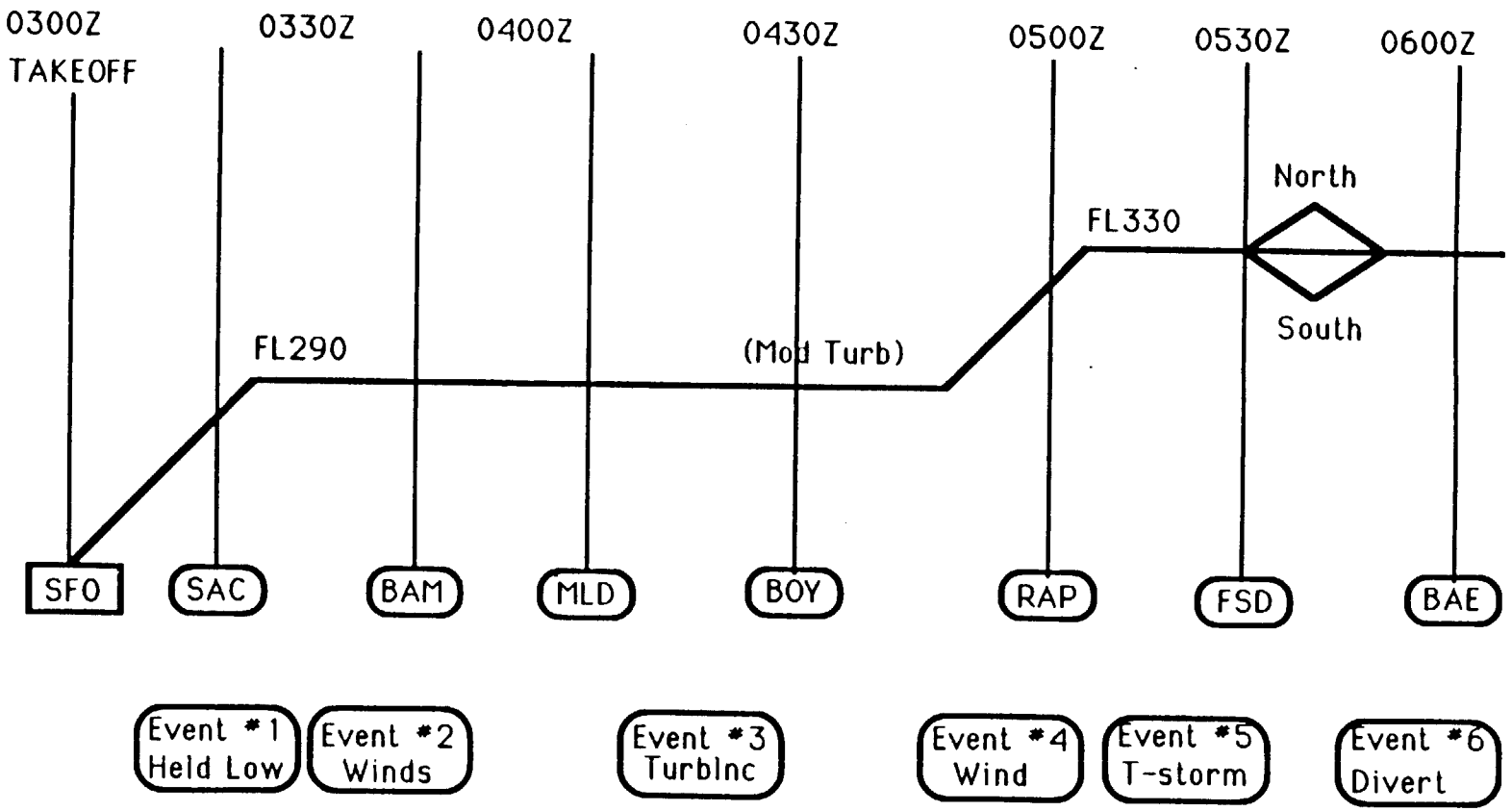


Figure 1. Our scenario - San Francisco to Detroit.

cruise altitude of FL250 and climb to FL290 at BOY. ATC clears the crew to a cruise altitude of FL290. The crew must now decide whether to immediately ascend to their originally flight planned cruise altitude or to follow FPA's recommendation and wait until BOY to climb.

3) The crew is approaching BOY. FPA receives a turbulence report that shows a short stretch of moderate turbulence at FL290 after BOY. It, therefore, recommends to the crew that they maintain their current cruise altitude of FL250 until RAP. Note that this recommendation is based on passenger comfort rather than a savings of time and/or fuel. The crew must now decide whether to immediately ascend to a cruising altitude of FL290 and possibly encounter moderate turbulence or to follow FPA's recommendation and wait until RAP to ascend.

4) The crew is approaching RAP where they are planning to ascend to FL330. FPA receives another wind update that shows the actual winds at FL330 are slightly less than the forecasted winds, and the winds at FL290 will give a marginal savings of fuel and time. It, therefore, recommends that the crew climb only to FL290 rather than FL330. The crew must decide, in this borderline case, whether to ascend to FL330 or to ascend to FPA's recommendation of FL290.

5) The crew has just finished their climb at RAP, and their fuel situation is becoming more serious due to winds less than forecasted and cruising at lower altitudes than anticipated. Also, the destination airport weather has been getting progressively worse over time.

FPA notices that the crew must soon make a decision about possibly deviating around the thunderstorm which they are approaching. For this event, FPA cannot make a single recommendation. It instead gives the crew three options - maintain their current path which goes through the middle of the thunderstorm, vector to the north of the thunderstorms, or vector to the south of the thunderstorms.

Maintaining the current path is very risky. The northern deviation is moderately risky, but does not use an extreme amount of extra fuel. The southern deviation is the least risky, but it uses a large amount of extra fuel which could become a problem given the worsening weather at the destination airport.

6) After the crew has deviated around the thunderstorm, their fuel situation is a definite problem. FPA receives a report that says the destination airport, Detroit, has gone below minimums (i.e., it is closed). It recommends a new destination of Toledo and the same alternate as before which is Cleveland. The crew must decide whether to take FPA's recommendation of going to Toledo, whether to stick with the original alternate and go to Cleveland, or whether to select an entirely different destination altogether.

2.2 The Flight Plan Advisor (FPA)

A secondary goal of Phase II was to evaluate the FPA system, a flight planning aid developed by NASA personnel. A detailed description of this system appears in Taylor and Dammann (1989).

Briefly, this system provided the crew with

1) current weather information,
2) suggestions for better flight plans when one was available, and
3) an explanation of why the new flight plan was better.

The crew had access to current weather information at all times. This information included

- wind charts at 400mb (24,000 ft.), 300mb (30,000 ft.), and 250mb (35,000 ft.),
- a radar summary chart,
- a significant weather chart which also depicted areas of turbulence, and
- hourly sequence reports, forecasts, and NOTAMs for all airports.

As an example of these displays, the radar summary chart is shown in Figure 2. Crews could request a hard copy printout of any of these displays.

Besides providing current weather information, FPA also recommended new flight plans to the crew if it found one that was significantly better than the current one. To make this recommendation, FPA would first alert the crew with both a message on the screen and an audio beeping. As soon as the crew acknowledged the beeping by pressing a button, FPA displayed a general explanation which included

- the reason for the new flight plan (e.g., better winds, turbulence, or thunderstorms),
- an abbreviated version of what to do next (e.g., continue at current altitude rather than climbing as planned), and
- the impact of this flight plan with respect to fuel and time (i.e., did the fuel burn and time increase, decrease, or remain unchanged).

Also at this time, FPA printed a copy of the new flight plan for the crew.

The FPA system was implemented using Apple Macintoshes and HyperCard software. During the experimental flights, the experimenter provided the "intelligence" of the FPA system by tracking the crew's progress along their route, calculating new flight plans for them, updating the weather information and sequence reports when appropriate, and triggering the events. Additional details of the implementation appear in Taylor and Dammann (1989).

2.3 Full-Mission Simulations

Five three-person crews participated in this study. Each crew consisted of a captain, first officer, and second officer from the same airline. All participants were currently qualified to operate a 727. Altogether, the five crews represented three different airlines. All subjects were paid and were informed that their responses and actions would remain completely confidential.

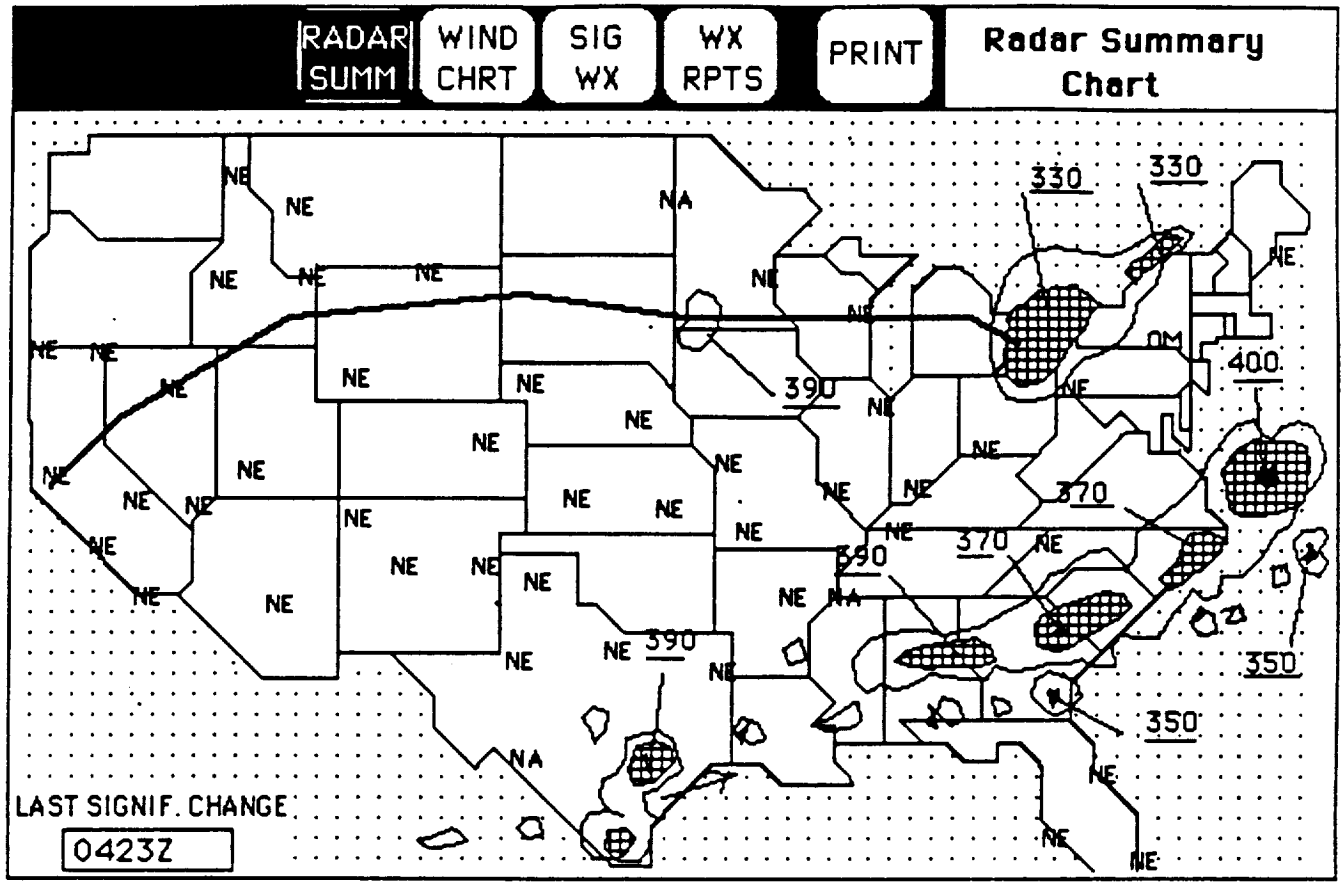


Figure 2. An example FPA display - the radar summary chart.

Initially, each crew member answered a questionnaire regarding his flying experience. This questionnaire is shown in Appendix A. Second, crews were given instruction in the differences between the 727s which their airlines used and the 727 simulator at NASA Ames.

Third, crews received training on the FPA system. This training consisted of a demonstration by the experimenter, followed by hands-on problem-solving by the crew members themselves. During the demonstration and hands-on problem-solving, the crews participated in two different events where FPA recommended a new flight plan. At the end of the experiment, all crews were asked if the amount of training which they received was sufficient. All agreed that it was.

Fourth, crews participated in the experimental flight from San Francisco to Detroit. In addition to the information available in a typical 727 cockpit, the crew also had access to the FPA system. All phases and aspects of the flight from take-off to top of descent were simulated including ATC and motion. The crew's original flight plan is shown in Figure 3. The data was collected using the 727 motion-based simulator at NASA Ames Research Center. The data included

- audio and video recordings among the crew,
- audio recordings between the crew and ATC,
- a video recording of the FPA screen during the flight and the crew's interactions with this system,
- the crew's decision for each "event," and
- the time and fuel situation for each event.

At the conclusion of the experimental flight, crew members participated in a post-flight interview session. This interview focused on what the pilots felt would be a reasonable design for a flight planning aid and a critique of our scenario. The interview questions are listed in Appendix B.

DTW	39300	3:59	YOU WILL LAND AT DTW WITH 12100 LBS	0230Z
ALT CLE	4100	0:18		ORIGINAL
FAR	6000	0:45		
HOLD	2000	0:13		
TOTAL	51400	5:15		

SAC J32 CZI J82 FSD J16 BAE J34 ALPHE

FIX	FL	OAT	TAS	WD	WV	COMP	ZD	ZT	ETA/ATA	CT	CF	RMG
SFO												
SAC	0	0	346	0	0	0	69	12		0:12	37	477
TOC	29	P01	346	0	0	0	44	8		0:20	62	452
FMG	29	P01	475	260	44	40	62	7		0:27	75	439
LLC	29	P01	475	270	44	36	61	7		0:34	87	427
BAM	29	P01	475	270	44	42	80	9		0:43	102	412
MLD	29	P01	475	290	45	31	224	27		1:10	148	366
BOY	29	P02	476	280	43	36	198	23		1:33	187	327
CZI	29	P02	476	290	43	31	87	10		1:43	204	310
RAP	29	P02	476	300	39	33	148	17		2:00	232	282
FSD	33	P01	467	300	34	30	271	33		2:33	284	230
MCW	33	P01	467	300	23	22	155	19		2:52	313	201
BAE	33	P01	467	300	20	17	222	28		3:20	353	161
ADALE	33	P00	466	310	10	8	101	13		3:33	372	142
GRR	33	P00	466	310	10	8	23	3		3:36	376	138
TOD	33	P01	467	310	10	10	8	1		3:37	378	136
ALPHE	0	0	250	0	0	0	41	10		3:47	385	129
DTW	0	0	250	0	0	0	52	12		3:59	393	121

Figure 3. The original flight plan.

2.4 Data Analysis

As part of the cognitive task analysis, the five videotaped flight sessions were analyzed for evidence of

- goals established by the flight crew (e.g., detecting when an alternative flight plan should be considered),
- information accessed by the flight crew to help accomplish various goals, and
- methods used to evaluate the information gathered.

The results of this analysis were used to build a model of the flight planning behavior of the crew while enroute.

The videotapes and post-flight interviews were studied for evidence of

- information the crew would have liked FPA to present,
- alternative functions that the crew wished FPA had performed, and
- problems in interacting with FPA.

3. THE ENROUTE FLIGHT PLANNING TASK

Not surprisingly, enroute flight planning is a group activity. The flight crew is involved, but so are Dispatch and Air Traffic Control (ATC). Perhaps more significant, however, is how strongly the data indicate that the flight crew plays a critical role in enroute flight planning. Some members of the aviation community have suggested that the flight crew needs little in the way of aids for replanning. They suggest this in the mistaken belief that replanning is primarily the responsibility of Dispatch. Our data makes it clear that this is a grossly simplistic view of enroute planning. This data clearly indicates that flight crews play a critical role in enroute planning.

This section provides evidence of flight crew involvement and also gives several examples that help us to begin to get "into the minds" of the flight crews. These examples help us to understand the role of the crew in enroute flight planning and also help us to begin to identify the types of data and aids

that could be of value to the crew. Briefly, the data suggests that flight planning aids could be very useful in assisting both the flight crew and Dispatch to

- detect the need to re-plan in a timely fashion,
- generate alternative flight plans,
- evaluate these alternative plans, and
- facilitate communications.

3.1 Evidence of Flight Crew Involvement

The simulations provide strong evidence that the flight crew plays a major role in detecting the need to consider alternative routes. Half way through the flight, for instance, one crew noted...

> "We could have some activity in Detroit, too. I think we're going to want to go north of that. North or south. It looks like north would be better... I wonder if we can figure a new route north of where that weather is out there and request it through the company?"

And at a later point...

> "What does it show on our charts as far as when it would be efficient to go to 33?"

In both cases, it was the flight crew that detected the need to consider an alternative route or altitude. The results reported here indicate that one important function of a computerized planning aid would be to help flight crews (and Dispatch) to detect the need to amend a flight in a timely fashion. Failure to detect such a need early enough could result in missing the opportunity to make the best adjustment.

These data, illustrated in the above quotes, also make it clear that the flight crew generates alternative routes. In some cases, their recommendation is forwarded to Dispatch for further consideration and approval. In other cases, the crew makes the change and simply informs Dispatch:

> "Normally what we do is we just request to Center for a slight deviation. But then we also tell Dispatch. We don't always, though."

Thus, the data indicate that both the flight crew and Dispatch are involved in generating and evaluating alternative routes. A computerized aid that helps to develop and evaluate alternative routes could be of value to both parties. Furthermore, such a system could prove to be a valuable communications tool between the flight crew and Dispatch:

> "That's something that would be real valuable with this type of system, would be to be able to have a direct request to Dispatch right there via the system."

3.2 Specific Instances of Flight Planning

Perhaps the clearest way to understand the nature of the flight crews' involvement in planning is to look at some specific examples. These are presented below.

Example 1. Fifteen minutes after takeoff, the pilot requested clearance to climb from FL250 to FL290. ATC denied this request because of other traffic. In response to this event, the flight crew did the following:

1) Asked ATC how long they would be at FL250.
2) Noted that they "ought to call Dispatch and tell them we're at a different altitude," but chose not to call Dispatch yet.
3) Asked themselves, "What do you think our difference in burn would be at 250?"
4) Determined the differences in fuel burn and time (actual vs. planned) at the next waypoint - "47.7--we're 200 pounds under."
5) Checked the wind speeds and directions - "Have the winds changed at all? We're coming up on Mustang. Mustang has winds at 290 of 44 knots."
6) Predicted the extra fuel burn resulting from staying at FL250 until Battle Mountain (the point at which ATC had indicated they could probably climb) - "I guess we know we're going to burn some more fuel staying down here, but probably as much as 500 pounds maybe."

7) Further evaluated the implications of staying at FL250 - "Twenty-five minutes down here. That'll let us get to 33 a little ahead of time because we'll have burned off fuel just a little ahead of time. Yeah. Possible. I don't know."
8) Planned their next change in path - "Battle Mountain. That's when I'm hoping to get 29,000."
9) Evaluated this plan by checking the winds at Battle Mountain.

As this example illustrates, the flight crew was extremely active in considering alternative flight paths. They collected a variety of data to determine the implications of the unplanned deviation from their route, and to decide what they should do next. Some of this data involved comparing actual performance (e.g., fuel burn) with that expected under the original plan. Other data required making predictions about future performance if the current altitude was maintained.

Example 2. In the first example, ATC instructions made it necessary for the pilots to consider the implications of a different route. In this second example which occurred 54 minutes into the flight, one crew detected data that caused them to consider a different route for other reasons...

Looking at the radar chart, the co-pilot noted:

"We could have some activity on the way to Detroit, too. I think we're going to want to go north of that. North or south. It looks like north would be better."

The crew then proceeded to develop such a plan:

"It seems like maybe we could reroute our flight up above there [North] rather than wait 'til we get up here... What kinds of VORs are we looking at then? Should we maybe go to Aberdeen flying up north and possibly Redwood Falls?"

The pilot then requested such a change:

"We have a routing request we'd like to have you pass on to our

dispatcher. We'd like to fly Jet 32 to Aberdeen, then Jet 70 to Badger. We'd like to remain at FL 250 for the time being."

This example illustrates the fact that the flight crew plays an active role in detecting the need to consider an alternative plan and in generating the alternative plan. One potentially valuable function of a computer aid would be to assist crews (and Dispatch) in detecting such situations as early as possible, so that all of the alternatives are still available to consider. A computer aid might also be useful to flight crews if it helped them to generate alternative plans and evaluate them.

Example 3. In a third situation, one hour and forty-four minutes into the flight, one crew began to consider alternative altitudes...

"What does it show on our charts as far as when it would be efficient to go to 33? [answer:] We should be at 33 now. We're at 8,000 below. Between the two altitudes, there's a difference of about 60 knots...We should be thinking about going to 33 here soon. Of course, turbulence always overrules."

Again we see the flight crew taking the initiative to consider an alternative. A computer aid could clearly help them by providing easy access to the data and calculations they wanted.

Example 4. Two hours and sixteen minutes into the flight, the same crew as in Example 2 began to consider the thunderstorm again...

"That looks kinda nasty. We tried to tell them a long time ago we wanted to go north of that. I'm not wild about going between those things. There's not 20 miles between them. I vote total deviation. Ask 'em for a vector around the north side of the weather. How far are we going to have to go? 100 miles? If we start down, we won't have to go as far out of our way. Just tell 'em we want to vector north of the weather and let them [ATC] do it. We don't have enough information to be that specific. There's no way we're going to fly into that... Holy shit! There's stuff behind it, too. Holy Mother!"

This example provides a nice illustration of the role of the crew in detecting a problem and considering alternatives. It also points out the importance of

coordination between the crew, ATC and Dispatch. In particular, the crew noted, "Taking our deviation a lot further back would have made a whole lot more sense."

Example 5. Two hours and forty-eight minutes into the flight, one crew began to worry about their destination...

> "I have a bad feeling about Detroit. Should have been starting to clear... The minimum there - we need a half mile... What did they show for the fuel [the reserve fuel for the new path around the thunderstorm] there? 18.6 - One thousand pounds less than original... I recommend, gentlemen, if Detroit doesn't look good we go direct to Cleveland and we go to the 100 Bomb Group for dinner, to the restaurant right next to the airport. [The pilot's joking response:] You wouldn't let that influence your decision, would you?... Chicago's pretty good. Milwaukee's not bad. Our landing fuel just gets lower and lower."

Here again we see the flight crew wanting to evaluate their plan. Detroit is still open, but the trends look unfavorable. In fact, just a few minutes after this example, the crew received a report that Detroit had closed.

Example 6. After receiving the bulletin that Detroit was closed, along with a recommendation to divert to Toledo, this same crew from Example 5 commented...

> "Do they expect an improvement in Detroit then? 1/4 mile and rain and fog... We would have a little time to hold, wouldn't we?... Call Dispatch and see what they want us to do. Let's see how much we've got in fuel, how much we can hold at someplace if we wanted to consider that."

In this case, the flight crew is generating an alternative and trying to evaluate it. An aid to help them evaluate this alternative in terms of fuel consumption would probably be useful.

4. COGNITIVE TASK ANALYSIS

The first step in our cognitive task analysis was to evaluate whether or not our scenario had truly caused the crew to consider replanning their current

flight path. Second, we analyzed the crews' performance to determine what goals were established by the flight crew, what information they accessed to help them accomplish the various goals, and what methods the crews used to help them evaluate the information.

4.1 Evaluating the Scenario

The overall planning performances of the flight crews can be summarized in terms of certain key descriptors including

- which direction they selected to deviate around the thunderstorm,
- which alternate airports they considered, and
- which airport they selected as their final destination.

The crews' decisions at each of these points are shown in Table 1.

This summary indicates that the scenario was in fact successful in inducing flight planning. The crews all diverted around the area of thunderstorms and selected alternate destinations. In addition, the more detailed analyses given in the next section demonstrate that each crew engaged in many planning activities in response to our scenario.

4.2 Task Analysis of Goals

The flight crew's planning activities can be described in terms of the following four major goals:

1) Detection of a need to consider an alternate flight plan.
2) Generation of alternative plans.
3) Evaluation of alternative plans that have been generated.
4) Determining what actions have to be taken to carry out a plan.

Each crew's transcript was analyzed to determine the circumstances where these goals appeared to be activated, and to identify the information used in achieving the goals. Because the data used to make such inferences consisted of discourse among the crew members, it is to be expected the crew did not

Table 1. The crews' decisions.

Crew	Direction of Deviation Around Storm	Alternate Airports Considered	Final Destination
1	north	Chicago Cleveland Milwaukee Toledo	Toledo
2	south	Chicago Cleveland Milwaukee Toledo	Milwaukee
3	north	Chicago Cleveland Grand Rapids Lansing Milwaukee Saginaw Souix Falls Toledo	Milwaukee
4	north	Cleveland Grand Rapids Lansing Milwaukee Minneapolis Toledo	Milwaukee
5	south	Chicago Cleveland Milwaukee Toledo	Milwaukee

mention all of the information which they actually used. Because of the teamwork required in the cockpit, though, the transcripts appear to provide a fairly good overall picture of flight planning activities.

4.3 Task Analysis of Information

Several coding categories were developed to indicate the information used by the flight crews. These consisted of the following:

1. Information on **fuel consumption**, including
   a. the current remaining fuel level,
   b. the original estimate of the fuel level at some waypoint,
   c. the difference between the actual fuel level and the original estimate,
   d. an updated estimate of the fuel level,
   e. the difference between the actual fuel level and the updated estimate, and
   f. the difference between the original fuel estimate and the updated estimate.

To generate updated or predicted estimates of fuel consumption, the crews also considered the weight of the aircraft, power settings, winds aloft and altitude. They furthermore considered the holding fuel likely to remain for particular paths.

2. Information on **arrival time**, including
   a. the actual time of arrival at some waypoint,
   b. the original estimate of arrival time at some waypoint,
   c. the difference between the actual arrival time and the original estimate,
   d. an updated estimate of the arrival time,
   e. the difference between the actual arrival time and the updated estimate, and

f. the difference between the original estimated arrival time and the updated estimate.

3. Information on **weather**, including
   a. turbulence,
   b. winds aloft,
   c. wind components,
   d. thunderstorm locations and tops,
   e. front locations, and
   f. weather at specific airports (ceiling height, precipitation, temperature-dewpoint spread, visibility).

These weather categories were also checked for trends over time and for accuracy by checking their own and other aircrafts' experiences at particular points along the route.

4. Information for **following their flight plan**, including
   a. jetways,
   b. waypoints,
   c. inbound and outbound paths,
   d. VOR frequencies, and
   e. groundspeeds and airspeeds.

4.4 Task Analysis of Methods

In addition to analyzing the transcripts for goals and information, we also analyzed the transcripts for the methods used to accomplish each of the previously mentioned goals and what information was used in each of these methods.

Methods for the Goal of Detection. Because they are on the scene, and because they have a primary role in deciding when a flight planning problem exists,

the crew is constantly monitoring for evidence of a need to consider an alternate route or destination.

First, the crew routinely monitors fuel consumption and arrival times at waypoints. This data is compared with original and updated estimates:

> "Time 3:46, fuel 446 on board. We're low 600 pounds, an additional 400 for that period, and one minute late."

They also predict the effects of unplanned route and wind changes on future arrival times and fuels:

> "I guess we know we're going to burn some more fuel staying down here. But probably as much as 500 pounds maybe."

> "We've dipped into our holds a little bit [already]. I think this detour up here is going to dip into the rest of it."

Second, the crew monitors for unpredicted changes in weather. In this scenario, the focus is on wind speeds and directions at different altitudes, on thunderstorm locations and tops and on airport ceilings, precipitation and temperature-dewpoint spreads. They also monitor for the presence of fronts and turbulence. In addition to monitoring the actual weather at times during the flight, the crew checks forecasts and looks at trends over time.

These types of data, then, help the crew to decide whether an alternative flight plan should be considered. When particular data deviate from expectations sufficiently, the crew begins to generate alternatives as illustrated below.

Methods for the Goals of Plan Generation and Evaluation. Having detected a need (or possible need) to replan, the flight crews began to generate and evaluate alternatives. Often they would develop complete alternative plans before requesting assistance or concurrence from Dispatch.

As one example of a specific planning episode, we can look at Crew #2 during their first planning episode. At 4:01 (the flight began at 3:00), this crew initiated their first planning episode. Prior to this time, they had been held at FL250, below their scheduled altitude. In this episode, they evaluate whether they should request a higher altitude at some point soon...

> "What's the maximum altitude we can get now? What are the winds up there? We have climb capability up to 33. What's the altitude where we'll get the best nautical miles per thousand pounds? The tail wind increases as we go up. Significantly? Better tail winds lower but not significantly. Ask the Dispatcher what if we go up at this point. Request a flight for 33 from Mallard City on this same routing. I'd say go up... Here's 33 and we're getting better tail winds up here all along and we're getting better fuel consumption up there."

This episode illustrates one type of planning script - attempting to find the best altitude along a given path. The factors considered in making this judgment include the aircraft's maximum altitude for its weight, the winds at different altitudes and fuel efficiency for the aircraft at different altitudes.

As a second example, we can look at this same crew during their third planning episode. At 4:37 this crew noted thunderstorms along their path...

> "See that one near the Minnesota-Iowa border. Tops 390."

At 5:07 they asked ATC,

> "Are you controlling that area over the Minnesota-South Dakota border? We had a forecast of some thunderstorms there. Any reports?"

At 5:15 they decided to seriously consider alternate routes...

> "What would be the shortest deviation? To the north, wouldn't it? Well, the wind would be pushing it to the south so we should go to the left. Yeah."
>
> "If we're already 200 miles away and it's that strong, you know we should start doing something about it."
>
> "When do you think we ought to start detouring?"
>
> "Do you have any reports on tops? Any traffic this way? Which way have they been going through this?"

"Mason City is just right smack in the middle of it. It's pretty extensive. It's pretty solid."

"What do you say we head to the left...We're going to have to go up north of Minneapolis. Why not just start here and miss the whole thing. Worst thing comes to worst, we'll land in Milwaukee."

"Actually, it would be smoother down there [a southern deviation]. There's almost nothing there. Sioux Falls-Des Moines is J45 and then after we get to Des Moines we can turn back in. J45 Des Moines-that's the south route and then how about... Des Moines direct Badger. What will this give us at landing?... So we've used all our holding [fuel] then...What's the new revised landing time?"

"I think we can save some money by not going up to Badger. Some time, too. What do you think? We'll go Des Moines, Northbrook, Pullman. It'd save us going all the way here and all the way back."

This episode, then, represents a second type of planning script, with the goal of finding an alternate horizontal path. In this case, the crew generates three alternative paths and evaluates them in terms of turbulence, storm activity (location, tops and direction of movement), fuel consumption, and landing time. They also consider the availability of an alternate landing site if needed.

All of the specific planning episodes for each crew are described in Appendix C.

Methods for the Goal of Determining if They Were Following a Flight Plan.

Having selected a flight plan, the crews had to collect certain data to ensure that they were following it. This included

- VOR frequencies and the inbound/outbound directions associated with particular waypoints,
- the jetways they were supposed to follow when going between waypoints, and the altitudes and power settings to be used along these jetways,
- the time and distance to a particular waypoint so that they could prepare to take appropriate actions, and

- actual and predicted groundspeeds and their differences, so that they could make adjustments in power settings if needed.

The primary insight provided by this study regarding this goal is that such information is used with great frequency. Given this high frequency of use, such information needs to be made readily accessible.

4.5 Cognitive Task Analysis Summary

Our analysis provides evidence of a number of goals that the flight crew must deal with. Included are

1) detecting a situation where replanning is desirable or necessary,
2) generating alternative paths for consideration,
3) evaluating the (actual or predicted) results of following a particular path,
4) communicating with Dispatch and ATC, and
5) collecting data and performing calculations to help accomplish goals 1-3 above.

The preceding subsections serve to indicate the kinds of data and inferences used to trigger and support these goals. The transcripts provide, for example, evidence of a number of different types of plans and planning situations. To avoid turbulence, for instance, we found evidence of plans involving changes in altitude, reductions in speed, or changes in the route.

It is clear that the 727 cockpit used in this simulation does not adequately support the planning activities of flight crews we observed. To our knowledge, while much improved, more recent cockpit designs still fail to meet many of the needs evidenced by this cognitive task analysis. There is a clear need to develop information displays that support the different types of problems that initiate planning episodes, and support the different types of plans that flight crews develop.

5. EVALUATION OF FPA

Why study pilots' flight planning performance in the simulator? Why analyze this performance data to find evidence of the goals and reasoning processes used by pilots to re-plan flights? Isn't what we need to do to help the flight crew obvious to anyone knowledgeable about aviation?

The data from this study make it clear that it is not "obvious" how to design a useful flight planning aid. We need better insights into the planning performances of pilots before we design such aids.

In particular, FPA was designed by NASA staff without the benefit of any performance data or cognitive task analyses. The results are well summarized by one pilot's evaluation three hours into the flight...

> "You know what I feel about this machine right now? I think they've got a great piece of equipment, but they need to figure out how to use it."

The results from our Phase I study demonstrated that there are at least four important issues to consider when designing interfaces to cooperative decision aids (Galdes, Smith & Chappell, 1989). These issues include

- the role of the system (e.g., a critic, a consultant, or a "what if" system),
- the amount of explanation necessary for pilots to accept the system,
- the specific information that pilots want/need from such a system, and
- the range of individual differences that the system must accommodate.

In Phase II, we again saw these same issues arise over and over again, primarily because the design of FPA was lacking with respect to how pilots felt these issues should be handled.

5.1 The Role of the System

As stated previously, recent research (Rouse, Geddes & Curry, 1987; Rouse, Rouse & Hammer, 1982; Woods, 1986) has shown that experienced users do not want knowledge-based systems which act as autonomous decision makers

because these systems do not give their users enough control. Yet, this is the classic AI paradigm for expert systems (Woods, 1986) and was also the basic paradigm for FPA. An alternative, and more acceptable, paradigm which keeps pilots in the loop is a knowledge-based system which acts as a *cooperative* decision aid rather than an autonomous decision maker.

In this experimental setup, FPA informed the crew when it found a better flight plan. It did not, however, allow the crew to *cooperate* in the decision-making process by possibly allowing them to ask FPA "What would happen if...?" or "Calculate the most fuel efficient flight plan that does not go above FL330 (because there is other traffic there)." This lack of cooperation was the most common complaint about the FPA system, and was mentioned by all five crews.

Some typical comments during the post-flight discussion were the following:

> "Here's the thing I have at the top of my list - I think you've got a great machine here, but the thing I found most frustrating was that you couldn't ask for a new flight plan." (Crew #1)
>
> "We should be able to ask 'what-if.' The system should also tell you if it wouldn't recommend that route and why. (Crew #2)
>
> "I would like to be able to change the cruise speed - the mach number - and see what would happen, see if slowing down would give me a little extra fuel when I landed." (Crew #3)
>
> "I found myself a couple of times turning around to the machine and going, 'Speak to me! We need some ideas here.'... You couldn't make it say anything." (Crew #4)
>
> "There wasn't any way to request a new flight plan. Sometimes we made a decision of what we wanted to do, but then we couldn't request a new flight plan and get information about it." (Crew #5)

5.2 Explanations

For each recommended flight plan, FPA gave two explanations - a very

general one and a more detailed specific one. Each explanation described

- the reason for the new flight plan,
- FPA's recommended changes to the current flight plan, and
- how this new flight plan compared to the current one in terms of fuel and time.

Examples of the general and specific explanations for Event #2, described in Section 2.1, are shown in Figures 4 and 5.

When FPA found a new flight plan for the crew, it alerted them with both audio and visual signals. When the crew acknowledged these signals, FPA would automatically display the brief, general explanation. In every instance after reading this general explanation, the crew then immediately requested the specific one.

All of the crews liked the explanations, but felt that one explanation would be enough. When asked if they preferred the general or the specific, detailed version, all crews responded that the specific one was better. One crew member came up with a way to keep both...

> "One [explanation] would be enough. You could have the specific explanation, then have the general explanation information bold or highlighted in some way in this more detailed information. That way, you could look at the fine details when you had time, but you could also just get the information in a glance if you were in a hurry." (Crew #3)

It is important, however, to note two points about the adequacy of these explanations. First, some pilots stated that the explanations were adequate because the pilots could check the explanations themselves using the raw data. This implies that we cannot provide an explanation capability without a mechanism which allows the pilots to check the system's reasoning. Second, although the explanation format was adequate for the role which FPA played, it may not provide the pilots with enough information if the system also supports a "what-if" role.

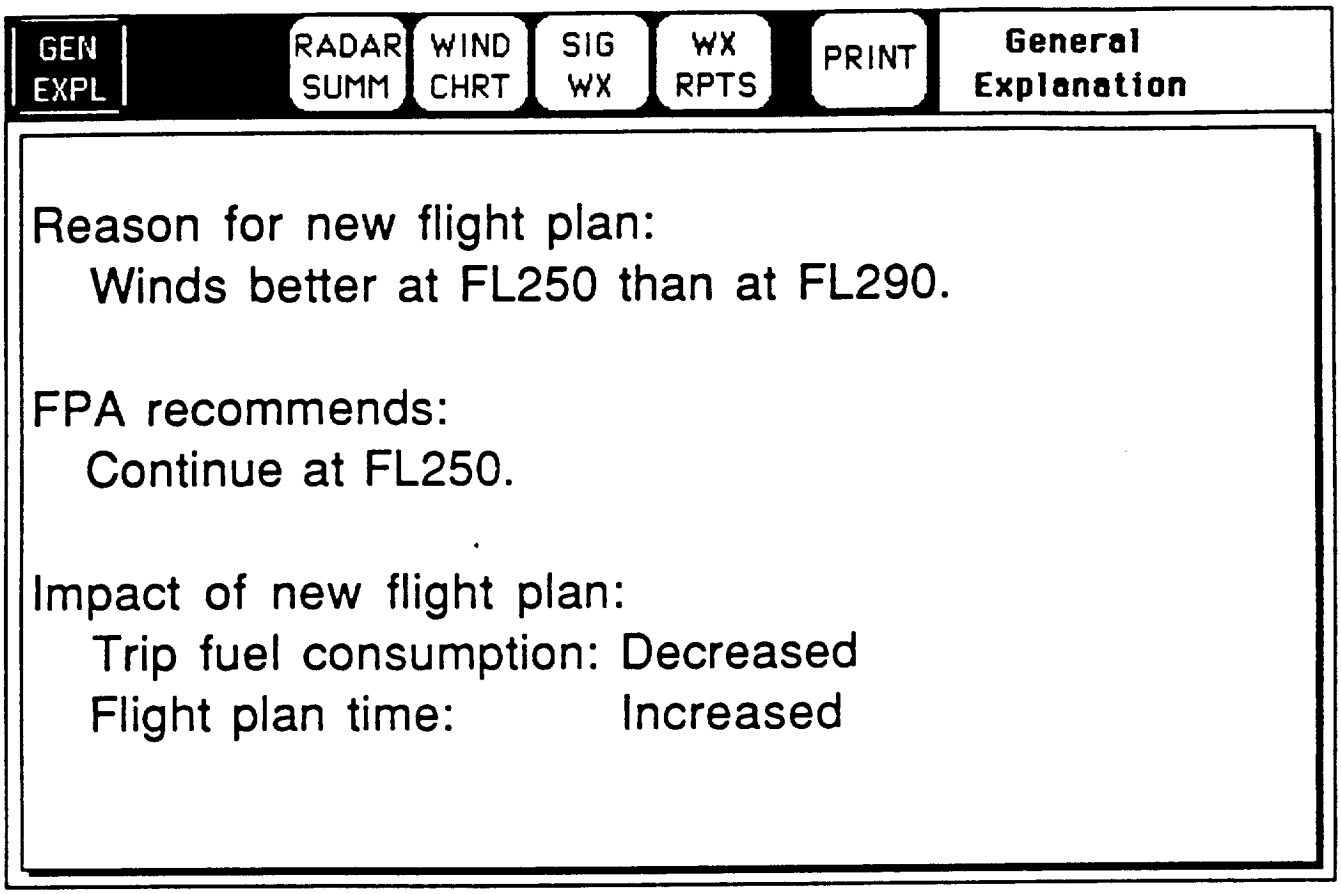


Figure 4. FPA's general explanation for Event #2.

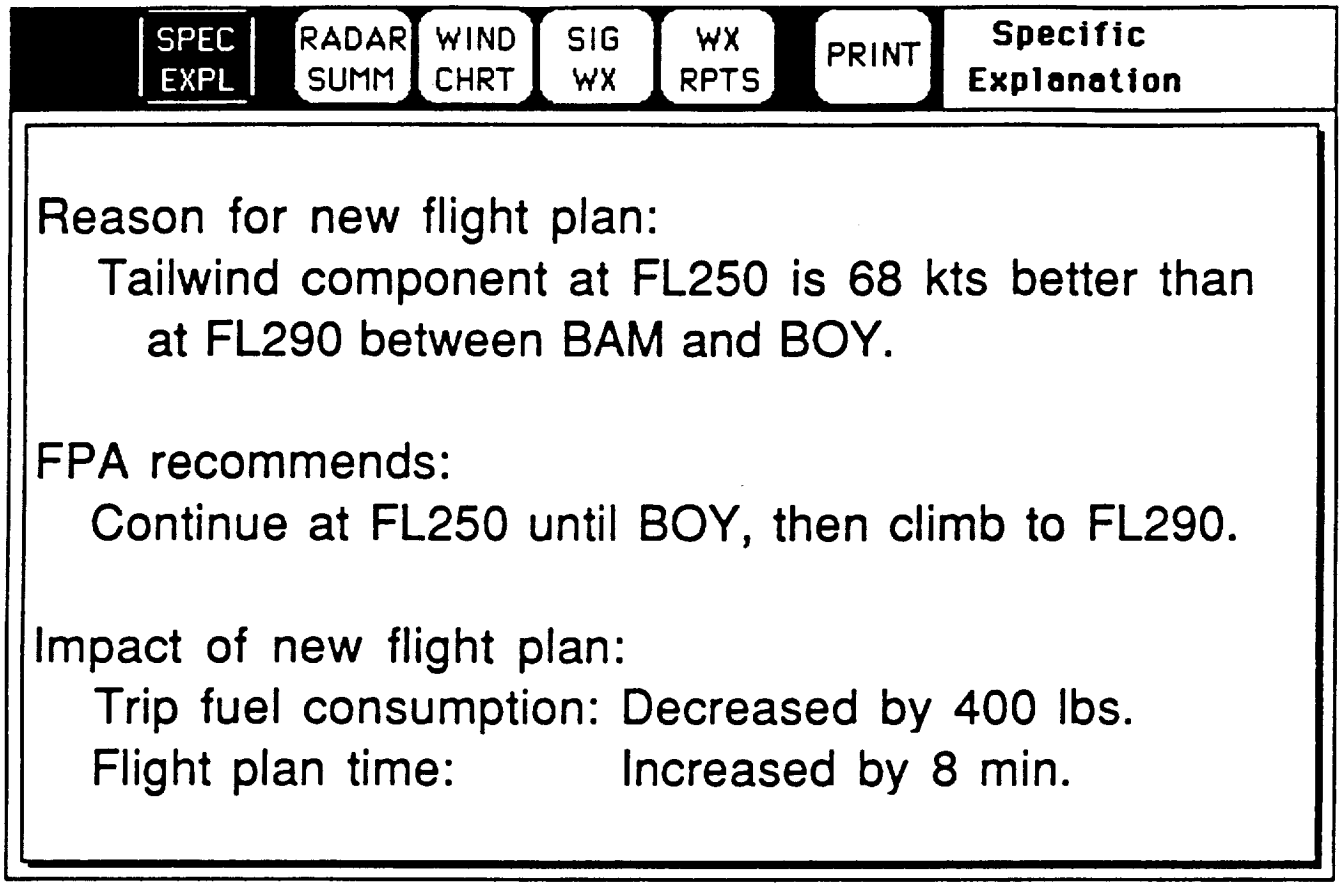


Figure 5. FPA's specific explanation for Event #2.

5.3 Information that Pilots Want

All of the pilots raved about the real-time weather information provided by FPA...

> "I like the real-time weather information. It helps us keep more in touch." (Crew #4)

They also commented that the weather data was especially important for confirming FPA's recommendations...

> "Pilots aren't going to accept these new flight plans without being able to verify it themselves... I think you're going to want to verify it yourself with some of the data." (Crew #1)

Some of the information that pilots felt was missing from FPA included

1) trends in the weather...

> "We would like to see not only the current weather [sequence report] but we also like to have access to what it's been over the last few hours. We want to see trends." (Crew #2)

2) being able to specify the time-of-arrival...

> "Time-of-arrival is becoming more and more important. For example, O'Hare is trying to get everyone to land as close as possible to their scheduled time-of-arrival. They want pilots to speed up or slow down during their flight so that their arrival time is accurate. I would think that for this FPA then you would want to be able to give it both a destination and a time of arrival." (Crew #1)

3) having the system highlight the arrival time in the forecasts (Crew #2),

4) information about delays over specific airports such as Chicago (Crew #2)

5) information about the distance to the closest airport...

"We couldn't find out how far Toledo was from anything. We wanted to be able to ask for what airport was closest. We're in a holding pattern and different airports are closer at different times." (Crew #4)

6) allowing the crew to put parameters into the system such as groundspeed and shortest time...

"We want to be able to put parameters into the system such as fastest groundspeed, different flight levels, shortest time, and least fuel burn for economy reasons." Experimenter's Question: When would you ask for shortest time as opposed to least fuel burn? Crew's Response: "Well, for example, you might see some weather moving along and you know that if you don't get there within the next hour then you probably won't be able to fly in there at all. Or you might have 150 people on board and you're trying to get them all on the last connecting flights out for the night. You know that if you don't make it, the company is going to have to put up 150 people and that's going to cost more than burning some extra fuel to get there fast." (Crew #4)

7) providing a zoom feature that would allow the crew to switch between a general overview of a large area and a detailed picture of a smaller area (all crews).

5.4 Allowing for Individual Differences

Whatever the final design of any decision aid, it should certainly allow for individual differences. Each crew in our study selected a slightly different route even though all crews began with the same original flight plan, began with the same amount of fuel, and encountered the same weather and traffic obstacles along the way. FPA did not provide this feature.

As one example of where individual differences are important, the experimenter asked one crew member if he would want to be responsible for picking his own cruise speed or would he rather have a system calculate it for him. His response was, "Personally, I'd want to pick it myself, but that's my way of operating." (Crew #3) Other crew members of different crews,

however, stated that the system could probably do a better job of picking the most efficient cruise speed so let the machine do the work.

In addition to individual differences/tolerances between pilots, pilots also change their own tolerances with respect to the current situation. As one pilot put it,

> "For example, if a pilot feels tight on fuel then he may want the system to inform him of better flight plans when "better" equates to as little as 1000 lbs. of fuel savings. If a pilot doesn't feel tight on fuel then he would want a flight plan that is much better before it [the system] interrupted him." (Crew #1)

5.5 FPA Evaluation Summary

FPA was capable of doing some useful computations, but the way in which it interacted with pilots greatly reduced its usefulness. The most important deficiency of FPA was that the crew needed and wanted the ability to explore paths they were interested in, and not just the paths FPA recommended. They wanted a system that played the role of a *cooperative* decision aid - one that would allow them to say, "Speak to me!" and it would speak back.

Furthermore, much of the data and calculations needed by pilots to replan flights simply wasn't available to them using FPA. This reaction is best summarized by one crew member who stated in the post-flight interview,

> "I started out thinking there was too much stuff that you didn't need. Then, as you got into the program, I was demanding more data and I was demanding more data than what the system was able to provide. I wanted more data on comparing the flight levels - should I go up, should I go down?" (Crew #4)

6. CONCLUSIONS

What is the value of a study such as this? First, the videotapes are an incredibly rich source of insights into the flight planning activities of pilots.

Anyone interested in building flight planning aids would be well advised to view these (or similar) tapes.

Second, the cognitive task analysis has identified the types of goals pilots establish and the information used to help accomplish these goals. These results are critical if we hope to build effective computer aids.

Third, the results are strongly suggestive of the role that the flight planning aid should play. It may be useful to build a system that suggests good alternatives. At least as important, if not more important, however, is the need for a system that

1) Helps pilots to detect the need for replanning in a timely fashion.
2) Aids pilots in generating alternative paths.
3) Provides computations to the pilots to aid in evaluating alternatives.
4) Allows the pilot to say "what if I took this path" to the computer, which then helps him to evaluate that alternative.
5) Facilitates communications between the flight crew and Dispatch.

In short, building effective flight planning aids will not be a trivial matter. We must identify the needs pilots actually have and how they think about flight planning. Such insights can then form the foundation for generating creative designs that are helpful. Data from studies such as this provide us with a more objective basis for building such a foundation.

7. ACKNOWLEDGEMENTS

This research was conducted as part of the NASA Graduate Student Researchers' Program. We would especially like to thank Sherry Chappell from NASA-Ames for being instrumental in getting this project off the ground, Ev Palmer for volunteering to take over while Sherry was pursuing her doctoral degree, Kim Coleman for being our aviation expert, the wonderful staff at the MVSRF Simulator Facility, Karen McNally for always being ready to lend a helping hand, and NASA personnel for building and maintaining the FPA system.

REFERENCES

Chambers, A.B. & Nagel, D.C. (1985) "Pilots of the future: Human or computer?" Computer, November, pp. 74-87.

Galdes, D.K., Smith, P.J. & Chappell, S.L. (1989) "Knowledge-based systems and pilot acceptance - new technology and old problems," Technical Report CSEL-1989-40, Department of Industrial and Systems Engineering, The Ohio State University.

Norman, D.A. (1986) "New views of information processing: Implications for intelligent decision support systems," Intelligent Decision Support in Process Environments, E. Hollnagel et al. (Eds.), Springer-Verlag.

Rouse, S.H., Rouse, W.B. & Hammer, J.M. (1982) "Design and evaluation of an onboard computer-based information system for aircraft," IEEE Transactions on Systems, Man, and Cybernetics, vol. SMC-12, no. 4, pp. 451-463.

Rouse, W.B., Geddes, N.D. & Curry, R.E. (1987-88) "An architecture for intelligent interfaces: Outline of an approach to supporting operators of complex systems," Human-Computer Interaction, vol. 3, pp. 87-122.

Taylor, L. & Dammann, K. (1989) "Inside ESI: An introduction to the ESI experimental environment and software reference - Volumes 1-3," NASA Technical Report, to be published.

Woods, D.D. (1986) "Paradigms for intelligent decision support," Intelligent Decision Support in Process Environments, E. Hollnagel et as. (Eds.), Springer-Verlag.

APPENDIX A - PRE-FLIGHT QUESTIONNAIRE

Subject number (experimenter will provide this) _____

Age _____ FAA Medical Certificate Class _____

Visual restrictions? _____

Cockpit position: Captain ____First Officer___Second Officer _____

Currently flying (aircraft, position) _____

Total flying time _____

Time in cockpit position _____ Time last 90 days _____

Please characterize your operational flying during the last 90 days:

Types of aircraft flown, approximate amount of flight time for each:

What % of time is military _____

airline _____

APPENDIX B - POST-FLIGHT INTERVIEW QUESTIONS

EXPLANATION OF INTELLIGENT FLIGHT PLANNING AID

What I want to do now is get some feedback and comments from all of you about your experiences interacting with this system. Again, I'd like to emphasize that this is only an experimental system at this stage, and it is not necessarily a system that we would put in the cockpit. The purpose is to give you an idea of what intelligent systems could do in the cockpit. It's meant to start you thinking about situations where you would find this kind of system useful and how you would want to interact with such a system.

In addition to giving you updated weather data, and giving you suggestions for new flight paths, a future version of FPA could also do things such as:

1) keep track of information from ATC
   - ride reports
2) or evaluate your own decisions
   - you choose an alternate airport, if there is another airport significantly better then it tells you that, otherwise it doesn't say anything

What we want to know is what would you really like the system to do.

GENERAL DISCUSSION QUESTIONS:

1) If you had such an intelligent system, how much of the underlying system do you think you'd need to know about in order to understand it or trust it? Were there specific events where you didn't trust FPA? In general, did you trust FPA? Why or why not?

2) What kind of roles do you see for such a system?
   - critic?
   - give you an answer?

- a giant number cruncher?

Does the role vary with the situation? with the amount of workload?

3) Did you like just getting one suggestion in most cases or would you like to see options in all cases? If the system did give you options most of the time, what kind of information would you want presented so that you could decide which option is best for you?

4) Did you like the way that the system alerted you to a new flight plan? Did it get your attention? Did you like the audio alarm?

5) If you had such a system in the cockpit, where would you like to see it placed and how would you like it to alert you?

6) Would you prefer a color system to one that is black and white?

7) Did you like getting explanations from FPA? Were those explanations too short or too long? Did you like having two, a general and a more specific one, or would a single explanation have been enough? Did the three distinct fields make sense? Would you have preferred different fields from "Reason for change", "FPA recommends", and "Impact" - in other words, is there other information that you wanted to see but was missing from these explanations?

8) Did you find the weather information useful? Did you like it in this "raw" form or would you have preferred a more intelligent form that would allow you to ask some questions - for example, for the wind charts, have options to highlight what has changed from the last wind chart or what changes have caused a new flight plan to be generated or click on portions of the US map and have it blow up that area and show more detail? Are there any other "intelligent" data displays that you would like to see?

9) Did you find the print option necessary? Would you have a preferred a system where you could go back to earlier displays or did you like having a hard copy?

10) If you could put all of the information that you wanted on a visual display, what would it look like? Would it be something similar to FPA or something completely different?

11) In what kinds of situations do you think this information about flight plans and their implications is most likely to be helpful to you? What kinds of situations can you imagine? In other words, if we wanted to create different scenarios to study how people responded to having to replan, what are the different circumstances to consider where it's really necessary/desirable to take another course?

12) Are there other kinds of information you'd want in these different circumstances to help you do some good planning? Is there any information that we missed?

13) Would you prefer to get this information verbally from someone on the ground like a Airink or ATC or would you prefer to have it visually displayed and always be allowed to read it off that way? Would one format feel better than the other?

14) Can you think of situations where this system or the extra information would get in your way or where you'd rather not have to mess with it? That is, can you think of times when this additional feature might actually interfere with your primary job of flying the aircraft?

15) When something like these situations occur, do you generally already have in mind what is probably a good or possible solution? If so, suppose that the computer displayed several possible solutions and the one you were thinking of wasn't in the computer's list. How would you react to

that? What would you like to be able to do in such a situation?

16) In terms of the general simulation here, did you feel that it was fairly realistic? Are there any aspects of the simulation that you can think of that would make it more realistic? How about in terms of the actual scenario - did it seem fairly realistic?

17) If you could only have three or four buttons of information, which ones would you keep, which would you throw away, and which might you add that weren't there today for you?

18) Suppose you had a choice of two options for how this system would be implemented. The first is a person on the ground who acts as a special ground service giving you all of this information that we've been talking about. They can put up displays as necessary so the communication isn't strictly over a phone line. The other option is to have a computer system that's providing the information. This system could be in the plane or on the ground. If you had a choice between those two, would you rather be dealing with a human or with a computer system? Why?

19) Are there any other comments you would like to make in terms of the study such as how it was run or comments about the scenarios chosen?

APPENDIX C - PLANNING EPISODES FOR ALL CREWS

CREW #1

Planning Episode A. At 3:24 (the flight began at 3:00), this crew began considering the implications of the request by ATC that they stay at FL250. They noted that this would burn up some extra fuel and began to consider when they should try to climb to FL330...

> "That'll let us get to 33 a little ahead of time because we'll have burned off fuel just a little ahead of time."

As part of this consideration they also looked at the winds aloft for different altitudes...

> "I'm going to look at the wind chart and see if I can pick that out."

Based on such data, they elected to stay at FL250.

Planning Episode B. At 3:54 the crew then noted the line of thunderstorms and began to consider possible deviations...

> "We're going to want to go north of that. North or south. It looks like north would probably be better...I guess when we get closer in we can turn the radar on and take a better look."

> "What we normally do, we wait'll we get up close and we say we want to go north of the weather and they give us a radar vector. It's showing our path going right through some thunderstorms. It seems like maybe we could re-route our flight up there [north] rather than wait 'til we get up here and get a radar vector. If we call the company, they can send us a new fight plan. What kind of VORs are we looking at then. Should we maybe Aberdeen flying up north and possibly Redwood Falls?... We have a routing request we'd like to have you pass on to our Dispatcher. We'd like to fly Jet 32 Aberdeen, Jet 70 to Badger. We'd like to remain at flight level 250 for the time being."

This episode illustrates the considerable amount of planning done by the flight crew. They are playing a significant role in deciding when an

alternative plan should be considered, and in generating details of an alternative. In evaluating this alternative plan, they considered a number of factors...

> "Landing fuel will be decreased by a thousand pounds...We added 10 minutes. It's going to cut on our hold fuel quite a bit. Ask him what the tops are. It shows turbulence in Wyoming, too. Check the wind chart."

They did not receive approval of their alternate plan soon enough. Consequently, at 5:16 the crew decided to request a deviation from ATC...

> "Ask them for a vector around the north side of the weather. How far are we going to have to go? A hundred miles. If we start down we won't have to go as far out of our way. Just tell 'em we want to vector north of the weather and let them [ATC] do it. We don't have enough information to be that specific."

This request was initiated by their conclusion that

> "It's a good possibility those holes could close up before we get there."

They chose the northern route because

> "...looks like Gopher'd be the better way to go as far as fuel burn and time. On the radar I don't see any significant difference."

They also noted the importance of making timely decisions...

> "...taking our deviation a lot further back would have made a whole lot more sense."

Planning Episode C. At 4:42 this crew again began to consider a higher altitude (as they were doing in Episode A). This time, the change was in part motivated by turbulence they had encountered...

> "Think it would smooth out if we went up higher?... It shows more turbulence up higher."

They continue this episode, considering other factors...

> "What does it show on our charts as far as when it would be efficient to go to 33?... Between the two altitudes, there's a difference of about 60 knots... at 33 we'd be 48 knots better... We should be thinking about going to 33 here soon. 'Course turbulence always overrules... since our fuel is holding now, I don't see anything to worry about now."

Planning Episode D. At 5:41 the crew checked the forecast for Detroit and Cleveland and noted,

"The weather keeps going down, doesn't it?"

At 5:48 one crew member commented,

"I have a bad feeling about Detroit. Should have been starting to clear... The minimum there, we need 1/2 mile. So far it would work. What did they show for fuel there? 18.6. One thousand pounds less than original."

At 5:59 another crew member stated,

"I recommend, gentlemen, if Detroit doesn't look good, we go to Cleveland."

At 6:03 it was noted,

"Chicago's pretty good. Milwaukee - not bad. Our landing fuel just gets lower and lower."

At 6:05 they were informed that Detroit was below minimums and that they should consider Toledo. They checked Toledo weather, and then commented,

"We would have a little time to hold [at Detroit], wouldn't we? If that weather's down, there's probably not a good chance it's gonna come up anyway, huh? Call Dispatch and see what they want us to do. Let's see how much we've got in fuel, how much we can hold at someplace if we wanted to consider that... We've got enough time [at Toledo] to make it on one turn... How much fuel does it take to go from Toledo to Cleveland?"

Again we see the flight crew playing an important role in the replanning process.

CREW #2

Planning Episode A. At 4:01 this crew initiated their first planning episode. Prior to this time, they had been held at FL250, below their scheduled altitude. In this episode, they evaluate whether they should request a higher altitude at some point soon...

"What's the maximum altitude we can get now? What are the winds up there? We have climb capability up to 33. What's the altitude where we'll

get the best nautical miles per thousand pounds? The tail wind increases as we go up. Significantly? Better tail winds lower but not significantly. Ask the Dispatcher what if we go up at this point. Request a flight for 33 from Mallard City on this same routing. I'd say go up... Here's 33 and we're getting better tail winds up here all along and we're getting better fuel consumption up there."

This episode illustrates one type of planning script - attempting to find the best altitude along a given path. The factors considered in making this judgment include the aircraft's maximum altitude for its weight, the winds at different altitudes and fuel efficiency for the aircraft at different altitudes.

Planning Episode B. At 5:05, the aircraft has gone to FL290, but the crew again considers an alternative altitude...

"To stay at this altitude, we need a tail wind component of about 13 knots to be favorable. At this weight they recommend 33. We've got better than that [a difference of 13 knots.] We're 44. We're better off at 29 and saving money and everything else than being at 33."

This episode, then, serves to make it clear that the wind components (headwinds and tailwinds) are of direct importance in considering alternate altitudes.

Planning Episode C. At 4:37 this crew noted thunderstorms along their path...

"See that one near the Minnesota-Iowa border. Tops 390."

At 5:07 they asked ATC,

"Are you controlling that area over the Minnesota-South Dakota border? We had a forecast of some thunderstorms there. Any reports?"

At 5:15 they decided to seriously consider alternate routes...

"What would be the shortest deviation? To the north, wouldn't it? Well, the wind would be pushing it to the south so we should go to the left. Yeah."

"If we're already 200 miles away and it's that strong, you know we should start doing something about it."

"When do you think we ought to start detouring?"

"Do you have any reports on tops? Any traffic this way? Which way have they been going through this?"

"Mason City is just right smack in the middle of it. It's pretty extensive. It's pretty solid."

"What do you say we head to the left... We're going to have to go up north of Minneapolis. Why not just start here and miss the whole thing. Worst thing comes to worst, we'll land in Milwaukee."

"Actually, it would be smoother down there [a southern deviation]. There's almost-nothing there. Sioux Falls-Des Moines is J45 and then after we get to Des Moines we can turn back in. J45 Des Moines - that's the south route and then how about... Des Moines direct Badger. What will this give us at landing?... So we've used all our holding [fuel] then... What's the new revised landing time?"

"I think we can save some money by not going up to Badger. Some time, too. What do you think? We'll go Des Moines, Northbrook, Pullman. It'd save us going all the way here and all the way back."

Episode C, then, represents a second type of planning script, with the goal of finding an alternate horizontal path. In this case, the crew generates three alternative paths, and evaluates them in terms of turbulence, storm activity (location, tops and direction of movement), fuel consumption and landing time. They also consider the availability of an alternate landing site if needed.

Planning Episode D. At 5:40, the crew notes that the weather at Detroit is deteriorating...

"400, overcast. It's going down."

They then began to evaluate alternative destinations...

"We've got 21,000 pounds. I'm going to have about 10,000 pounds at touchdown. I'm going to check Milwaukee. We go by Milwaukee, we're going to make a good check on the Detroit weather. If it doesn't look good, we'll go to Milwaukee. The hell with it. We don't have any holding fuel or anything now. 'Cause Cleveland isn't exactly good - 600 overcast... It was 800 overcast. That's going down."

At 6:04 they again evaluated their alternatives...

> "Detroit - 200 overcast, 1/4 mile, rain, fog. Okay, it's decision time. What are the minimums for two-one? I think we're going to Milwaukee. Look at that thing at Cleveland... Get the Dispatcher and tell him we're thinking about going to Milwaukee. We can load up with gas there and go on. O'Hare wouldn't be bad. O'Hare is 5000 overcast. Milwaukee - we won't get delayed as long... check that Toledo weather... The weather's not that swift at Toledo either. 52-50 is the temperature-dewpoint spread so I'd go to Milwaukee."

Thus, Episode D provides us with further evidence of how information is used in a third script - selecting a new destination (because of bad weather at the scheduled landing site).

CREW #3

Planning Episode A. At 3:27, this crew begins to consider alternative altitudes. They begin to explore the wind patterns...

> "Right about Reno it shows the wind at 320. We've actually got a tail wind."

And comment,

> "These flight plans ought to have the different fuel burns at different altitudes at different times."

No action was taken as a direct result, however.

Planning Episode B. At 3:49 this crew notes the thunderstorms along their path, stating:

> "We're gonna be in trouble when we get to South Dakota. Yeah. We're gonna go either north or south from the looks of it."

Again, no action was taken.

Planning Episode C. At 4:21, the crew begins to worry about the weather at their destination...

"We'll have 11,000 when we get to Detroit. I don't want to start getting much below 11,000. What do you think? No, not with the weather we've got there."

This leads them to start exploring alternatives...

"Alternate Cleveland. Cleveland is worse than Detroit. That's 0400 weather... any other ideas? Grand Rapids, Chicago, Saginaw. Actually, we're going to be going right over the top of Milwaukee. I was thinking that as we get down in the Detroit area and they start vectoring us around we might need someplace close like Toledo or Saginaw... Actually, Grand Rapids would be good 'cause its got a long runway and light traffic and it's not too far."

Planning Episode D. At 4:30 they run into some turbulence. They briefly consider going up but decide not to because

"Supposedly there's worse turbulence up higher. It's not bad yet [at their current altitude]."

Planning Episode E. At 4:57 the crew again considers changing altitudes...

"We may need to go up to 33 just to see around this weather as we go along here."

Planning Episode F. At 5:11 we see evidence of a new planning script, one in which the focus is on the effects of speed on fuel consumption...

"You're going a little fast. You'll burn up your reserve fuel."

Note that later at 5:23 after Episode G, they conclude,

"That brings us down to 10.3. Why don't you bring this thing down to save us some fuel."

Planning Episode G. At 5:17, the crew begins to worry about the line of thunderstorms...

"Looks like we need to go one way or the other. It's about a tie... Want to take the northern route? The only reason I suggest that is I think it's a little more direct flying to Detroit."

Planning Episode H. Worried about fuel consumption, at 5:31 this crew again worries about their altitude...

"I can't believe we can't get a better burn at a higher altitude as light as we are... It's real doubtful we're better off down here than at 33. But we might as well stay here. They say the winds are a lot stronger."

Planning Episode I. At 5:46 the crew considers minor deviations to their route...

"We're running a little short on fuel. If you [ATC] arrange any vectors toward Detroit, we'd appreciate it."

Planning Episode I. At 6:04 they learn that Detroit is below minimums. They conclude,

"I don't think Toledo's so hot... We're still real close to Grand Rapids. I'm beginning to like Milwaukee more and more all the time... Milwaukee's real good - 6000 overcast."

They make their decision and inform ATC,

"We have no holding fuel left. We'd like to land at Milwaukee."

CREW #4

Planning Episode A. This crew simply accepted the request from ATC to stay at FL250. It wasn't until 4:23 that they considered changing altitudes, and that was because they encountered turbulence...

"Did it say where the turbulence was? From 21 to 37."

They stayed at FL250 since there was no indication that the turbulence would be less at another altitude.

Planning Episode B. At 4:33, the crew does some very general advance planning, noting the possibility that they may have to land at an alternate airport...

"Let's call Dispatch and tell him that we're advising him that we're down to flight plan fuel and that we stayed down at 250, moderate turbulence at 29, and that if there's any change in our destination or if the altitude or if the weather at our alternate deteriorates any more we're gonna have to stop and get some gas someplace."

Planning Episode C. At 5:01 the crew does a little more advance planning...

"I'm trying to see how much closer Toledo is than Cleveland, just in case Dispatch falls asleep."

Planning Episode D. At 5:10 the crew begins to respond to the line of thunderstorms along their path...

"You want to deviate to the north. Well, I'm thinking, since we got to go south anyway, maybe south might be the way we want to go. It'd save us some gas 'cause we've got exactly 24 grand worth of gas right now."

"We're getting awful close to that. We're going to have to turn sometime. It's time to make a move."

"How high can we go? Can we make 37? Ask Center what kind of ride reports he's had. Ask him what kind of ride reports at 29 and 33."

"It looks to me like the best way to go is to go right down to Des Moines. Yeah, if these things don't move too far south."

"If you go south to Des Moines, decrease 1500 pounds. I suggest we go north... That's 800 pounds."

"Let's take a look at the winds."

"Let's see if we can go up to 33 and miss some of this. Any ride reports?"

Because they are low on fuel, one factor influencing the choice of a north or south deviation is the availability of an alternate destination...

"Got one pound, get to hold no time. Why don't you look up Minneapolis weather. Any specials on Detroit or Cleveland?"

They then continue to consider the best altitude for the deviation...

"You got time to compare the wind at this altitude and 29?... Ask him if he's got any top [of the thunderstorm] reports."

Finally, they decide to deviate north.

Planning Episode E. At 5:34, the crew is concerned about their fuel level. As a result they decide to slow down...

"We're going about 8.4. Want to pull her back a touch, save some gas?"

Planning Episode F. At 5:41 they again begin to consider alternate destinations...

"If we check our fuel at Badger, what's the next suitable alternate? Grand Rapids."

Planning Episode G. At 5:43, in response to turbulence, they consider revising their flight plan by slowing down...

"If it gets any rougher, I'll slow down to 7.8."

Planning Episode H. At 5:45, the crew begins to get seriously concerned about their fuel levels...

"We're into our FAR 6000 pounds. Okay. Let's shut the engine heat off. Anybody got any ideas how we can recover that 800 pounds? What was the wind at 29? Twenty knots difference. How about we talk them into changing the alternate to Toledo? It's closer... Cleveland is better [in terms of weather] than Toledo, but Toledo is closer."
"Call Dispatch and tell him we're 400 pounds below the required."

"How about something north of Detroit? How about Lansing?"

Planning Episode I. At 5:58 the crew looks for another way to conserve fuel...

"Try to cut the corner at O'Dale. Save a little bit there."

They request permission from ATC to do so.

Planning Episode J. At 6:03, they learn that Detroit is below minimums. They first consider Toledo...

"Let's go see the Mudhens. Anybody got a Toledo plate?"

At Badger, however, ATC tells them they must hold because of heavy traffic into Toledo. Their response is to...

"Tell him we've got a critical fuel state and we can't accept holding. We've got 14.4 now. We can hold a couple of minutes. What amount of

fuel do you want to leave Badger with?"

"What's the best holding altitude? The holding speed, we weigh now about 142, would be about 223."

Because ATC can't release them from the holding pattern soon enough, they decide to go to Milwaukee instead of Toledo...

"Take a look and see what the Milwaukee weather is. How far is Cleveland? Farther than Toledo. How far is Milwaukee? We're right over the top of Milwaukee. I know a good restaurant in Milwaukee."

They check with Dispatch to see what he recommends...

"Call Atlanta... Tell him we're considering going into Milwaukee... He says, yeah, let me know what you do. Typical Dispatch."

Finally they decide to land at Milwaukee.

CREW #5

Planning Episode A. At 3:36, Crew #5 noted that they were consuming more fuel than planned because they were being held at FL250. They therefore investigated going to a higher altitude assuming that it would be more fuel efficient since their original flight planned called for this higher altitude...

"Fuel is 42.2. We're running behind on gas, a little behind on time. Not too much. What kind of speed are we doing?"

"We're probably not at the optimum. We should be at 29. The flight plan's at 29. We'll call them [ATC] and see if we can get 29."

Planning Episode B. At 4:14 they note that they are behind schedule and below the scheduled ground speed. Consequently, they increase their speed...

"We're about three minutes late."

"Our winds are northwest."

"Why don't we take it up to about 320. That should bring our groundspeed closer in line."

Planning Episode C. At 4:31 they consider going higher but decide not to because of reported turbulence at the higher altitudes...

"Because of the turbulence, why don't we just hold with what we've got."

Planning Episode D. At 4:56 they again debate going up to FL290...

"We have to make a decision in 17 miles? Uh oh."

They decide to go up as scheduled and proceed to determine a good power setting...

"And we're gonna probably need an EPR for 29."

Planning Episode E. At 5:07, they begin to consider the potential implications of their fuel consumption...

"What kind of minimum fuel do you want to land with at Detroit? We could call Dispatch for anticipated delays over Detroit. It looks to me like we won't hold. We're down to five minutes of hold fuel. We'll make one pass over Detroit and if that doesn't work we'll go right over to Cleveland."

Planning Episode F. At 5:10 they note that there are

"...some cells right above Sioux Falls. Are we going to have to go around that?"

As an alternative, they explore the possibility of going above the storm...

"Wonder what the tops are on that. Might be better to go up. Turbulence might not be as bad. No steep gradients."

Planning Episode G. At 5:16 they again worry about weather at their destination...

"Why don't you call Dispatch and ask them what they recommend if Detroit is bad."

Planning Episode H. At 5:19 the captain decides to divert south around the storm...

"We got even worse stuff ahead. We gotta make a command decision..."

And requests from ATC,

"...direct Des Moines."

Planning Episode I. At 5:29 they consider changing altitude to save on fuel...

"What would be nice would be to see what our fuel burn would be if we went up to 35 - 37,000 feet."

Planning Episode I. At 5:41 they decide to go lower from FL330 to FL290 to try to reduce turbulence and take advantage of better winds...

"Think we ought to go back down? As soon as possible. The reason I say that is the tailwind component is considerably less than at 29."

Planning Episode K. At 5:58, the crew begins to consider alternate destinations...

"You guys want to go to someplace that's wide open instead of taking our chances at Cleveland?"

They reject Toledo, because it

"...looks shitty, too. With that much gas I don't like it."

They also consider Chicago...

"This is where it'd be nice to know what kind of delays there are in Chicago. Normally we get that from Dispatch."

They finally decide to land at Milwaukee.