RR-73-15

On the Generalizability of Experimental Results*

Robert L. Winkler** and Allan H. Murphy***

Abstract

The age-old question of the generalizability of the results of experiments that are conducted in artificial laboratory settings to more realistic inferential and decision making situations is considered in this paper. Conservatism in probability revision provides an example of a result that 1) has received wide attention, including attention in terms of implications for real-world decision making, on the basis of experiments conducted in artificial settings and 2) is now apparently thought by many to be highly situational and not at all a ubiquitous phenomenon, in which case its implications for real-world decision making are not as extensive as originally claimed. In this paper we consider the questions of generalizations from the laboratory to the real world in some detail, both within the context of the experiments regarding conservatism and within a more general context. In addition, we discuss some of the difficulties inherent in experimentation in realistic settings, suggest possible procedures for avoiding or at least alleviating such difficulties, and make a plea for more realistic experiments.

1. Introduction

Considerable interest exists, among psychologists and others, in human behavior in inferential and decision making situations. This interest is evidenced by the amount of research conducted in this area in the past decade. The research has included a considerable amount of experimental work, much of which has involved purposely simple, artificial

*This paper will be published in a forthcoming issue of Theory and Decision.

Research was supported in part by the U.S. National Science Foundation Grant GA-31735.

Much of the material in this paper appeared in Winkler and Murphy [46]. We are grateful to Academic Press, Inc., for their permission to include that material here.

**Indiana University, U.S.A.; research scholar at the International Institute for Applied Systems Analysis.

***National Center for Atmospheric Research, U.S.A.

situations. Such simple situations are easy to deal with and to explain to typical subjects, and they also possess the advantage of being relatively simple to analyze. However, their very simplicity and artificiality makes the justification for generalizing the results of these experiments to more realistic inferential and decision making situations questionable.[1]

In Section 2 we consider one major result that has emerged from the extensive experimental work regarding human behavior in inferential and decision making situations conservatism in probability revision. The situational nature of conservatism and potential differences between realistic situations and artificial laboratory situations are discussed, and it is suggested that the conservatism observed in the laboratory experiments may be artifactual.[2] The question of generalizations from the laboratory to the real world is considered in a more general vein in Section 3, and we indicate that such generalizations should be made with a considerable degree of caution. In Section 4, some difficulties involved in realistic experiments concerning human behavior in inferential and decision making situations are examined, and we suggest that these difficulties are not as serious as some might believe. Section 5 contains a brief summary and discussion.

2. Conservatism: An Artifact?

Much of the research in the area of "man as a processor of information" has involved the phenomenon of conservatism in the revision of probabilities. Such conservatism should not be confused with conservatism in decision making situations, which is related to the notions of risk and risk aversion. Conservatism in probability revision refers to the failure of individuals, in reporting their probability revisions, to extract from data nearly as much information as the data contain. In Bayesian terms, this implies that posterior probabilities assessed judgmentally are closer to the corresponding prior probabilities than are posterior probabilities calculated via Bayes' theorem. In the decade following the initial research regarding conservatism, which was conducted in 1960-62 and reported in Phillips, Hays, and Edwards [29], numerous experiments have been conducted and a considerable amount of debate has ensued regarding the existence of and (assuming existence) the possible explanations for conservatism in human information processing. Extensive reviews of the literature concerning conservatism can be found in Du Charme [7] and Slovic and Lichtenstein [37]; we will touch on some of this literature, but we refer the interested reader to these reviews for more details and further references.

The degree of conservatism observed in probability revision experiments has varied considerably, and in some cases (e.g. Schum, [32]; additional references will be given later in the paper) the subjects were not conservative. Du Charme [7, p. 13] notes that

> The upshot of the research discussed thus far is that conservatism is not a stable phenomenon. The amount of conservatism evident in Ss' probability estimates has been shown to differentially depend on the diagnostic value of the data, prior odds, sample size, sample order and the diffuseness of the hypothesis under consideration.

This suggests that the existence of conservatism and (if existent) the degree of conservatism in any given situation may be highly dependent upon the exact nature of the situation.

Various explanations have been suggested for the conservatism that has been observed in experimental studies. The explanations include misperception (the individual simply misperceives the inferential impact of each datum), misaggregation (the individual perceives the impact of each datum correctly but errs in aggregating the data), and response bias (the individual exhibits a tendency to avoid extreme values on the probability scale). Although there has been much debate and many experiments have been designed in an attempt to ascertain the existence of conservatism and to isolate the "cause" of conservatism (where it exists), the matter remains unsettled. Perhaps this is due at least in part to the situational nature of conservatism, as discussed in the preceding paragraph, in which case the situations used in conservatism experiments should be examined carefully.

A commonly used device in probability revision experiments is the bookbag-and-poker-chip paradigm, in which the experimenter has a number of (usually two) bookbags filled with poker chips. The composition of the bookbags differs with respect to the number of poker chips of a particular color; the first bookbag may contain 70 red chips and 30 blue chips, for example, whereas the second bookbag contains 30 red chips and 70 blue chips. The subject is told that one of the bookbags has been chosen by a random device, and his task is to assess probabilities for the possible bookbags. Since the subject is told that the bookbag has been selected randomly, the initial, or prior, probabilities should be equal. The subject is then given information in the form of a series of draws, at random and with replacement, from the chosen bookbag. Upon observing the results of these draws, the subject assesses his posterior probabilities for the possible bookbags.

There are endless variations on the basic bookbag-and-poker-chip theme. The experimenter can vary the number of bookbags, the composition of the bookbags, the prior odds, the response mode (e.g. probabilities vs. odds vs. log odds), the sampling plan, and so on. To make the situation somewhat more complex, elements such as nonstationarity (e.g. random shocks that change the composition of the bookbags), conditional dependence of sample information, and unreliability in the reporting of the sample results (e.g. the possibility that the individual reporting the sample results is color blind or simply careless) can be introduced. Numerous variations on the bookbag-and-poker-chip theme have been investigated experimentally (see Du Charme [7], and Slovic and Lichtenstein [37]).

Thus, although many different situations have been studied, in most cases the experimental vehicle has been the bookbag-and-poker-chip paradigm or some similar paradigm. Paradigms like this are useful in determining the psychological effects of various probability assessment and aggregation procedures. It is not clear, however, whether the results of experiments using such paradigms can be generalized to the more ill-structured, vague situations encountered in the real world. The information sources used in the formulation of probability forecasts in most substantive areas cannot be modeled as clearly and unambiguously as the procedure of randomly drawing poker chips from a bookbag, which obviously follows a Bernoulli process (or more generally, a multinomial process). Yet, despite the simplicity of the bookbag-and-poker-chip paradigm, it may not be as familiar to subjects as many real-world processes that are considerably more complex.

Ironically, then, the subject in a bookbag-and-poker-chip experiment may be faced with a situation with which he is quite unfamiliar and for which he has little intuitive "feel" even though it is much simpler than most real-world situations he encounters. One possible mode of behavior for the subject is to treat the experimental situation at least in part as if it were similar to realistic inferential situations. Such behavior may yield inferences that are conservative in the bookbag-and-poker-chip situation.

One basic assumption that is made in most of the psychological experiments concerning probability revision is that of conditional independence of the trials, or information sources.[3] If poker chips are selected at random and with replacement from a bookbag, the color of the kth poker chip drawn is independent (conditional on a given bookbag) of the colors of the first k-1 poker chips that are drawn. In many (probably most) real-world inferential and decision making situations, the information sources are conditionally dependent. Often there are cases in which successive items of information are somewhat redundant, so that the total impact of several items of information is less than the combination of their marginal impacts (i.e. the impact of each item assuming that the other items had not been observed) (e.g. see Winkler and Murphy, [45]). If an individual treats conditionally <u>independent</u> items of information as if they were redundant,he will revise his probabilities less than he should and will appear conservative. Therefore, one possible explanation for conservatism in simple bookbag-and-poker-chip experiments is that the subject is behaving as he does in more familiar situations involving redundant information sources.

Another assumption in most of the experiments regarding information processing is that of stationarity. It is assumed that the composition of the bookbag remains constant over time. Parameters of real-world processes may be expected to change over time, and the additional uncertainty caused by the presence of nonstationarity may lead an individual to revise his probabilities less (for a given sample) than he would if the process were stationary. This implies that treating stationary bookbag-and-poker-chip situations as if they were nonstationary could lead to conservatism.

Of course, real-world processes may differ from processes used in experimental situations in many respects other than conditional dependence and nonstationarity. For example, unreliable information sources may be encountered frequently in the real world, individuals may tend to make inferences at various levels of hierarchical systems ("cascaded inference"), or data encountered in actual situations may usually be less diagnostic than typical data generated for experimental use. With regard to diagnosticity, Schum

[33, pp. 237-238] makes the following remarks:

> One repeated result is that men are most conservative when the diagnostic impact of the evidence is very large. They become less conservative and, indeed, excessive in their revisions when the diagnostic value of evidence is reduced.

Therefore, conservative subjects may simply be generalizing from real-world situations in which data tend to be relatively undiagnostic.

Some obvious potential differences between real-world situations and simple, artificial experimental situations, then, are conditional dependence, nonstationarity, unreliability, cascaded inference, and lack of diagnosticity. Factors like these may cause individuals to adjust their likelihoods when revising probabilities. Some experiments have been conducted to examine these potential differences, and modifications to make the experimental situation more realistic have frequently resulted in less conservatism or no conservatism on the part of the subjects. Modifications investigated experimentally have included conditional dependence (e.g. Gustafson [17], Schum [32], Schum, Southard, and Wombolt [35], Domas and Peterson [6]), nonstationarity (e.g. Chinnis and Peterson [4,5]), and unreliability (e.g. Schum, Du Charme, and De Pitts [34], Snapper and Fryback [38]). However, many of these modifications have been attained experimentally by changes in the bookbag-and-poker-chip paradigm or equally artificial situations. As a result, most of these experiments still do not address directly the question of human behavior in realistic situations.

Moreover, in addition to obvious differences mentioned above, there may be more subtle, but nevertheless important, differences between human behavior in operational inferential and decision making situations and human behavior in laboratory experiments. For example, unlike the well-structured bookbag-and-poker-chip paradigm, in which the information sources (draws from the bookbag) are straightforward, many real-world situations may be extremely vague and information may suddenly arrive from unexpected sources. The point is that experimental results may be artificial in the sense that the subjects are inappropriately generalizing from real-world situations to artificial experimental situations. Individuals who make inferences and decisions in a reasonably "optimal" fashion in the real world may behave suboptimally in experimental situations.

We claim no originality for the suggestion that conservatism may be an artifact caused by dissimilarities between the laboratory and the real world, although others have only made such a suggestion "in passing" and have not investigated it in any detail. In an early paper, Edwards and Phillips [13] note that the subjects' poor performance (i.e. conservatism) might simply reflect poor understanding of the experimental situation. Edwards [10, p. 78] comes even closer to the point:

> Although the evaluation of uncertain evidence is an ingredient of much everyday behavior, . . . it would appear that most persons do it rather badly. There are at least two explanations of this apparent contrast between the experiment and everyday behavior. One is that, although they are very inefficient, persons are not usually troubled by their inadequacies because everyday situations are undemanding. The other is that the experimental results are artifactual and persons are not necessarily as inefficient outside the laboratory.

Criticism is aimed directly at the bookbag-and-pokerchip paradigm in some references. In Edwards [12], the paper is written as a hypothetical debate among Ward Edwards, Lee Beach, and Cam Peterson; in the dialogue Cam Peterson argues against "the very artificial bookbag and poker chip situation," noting that he has found that Bayes' theorem is a very good descriptive model of subjects' behavior in other kinds of situations, and contending (pp. 18-19) "that the conservative behavior obtained in those experiments that obtain it, is essentially an artifact." Phillips [28, p. 259] states, "Conservatism is always found in bookbag-and-pokerchip experiments; for these tasks most subjects have had very little prior experience with the binomial data-generators." Schum, Southard, and Wombolt [35, p. 19] make the following observation:

> It is somewhat curious but true that men's estimates of $P(H|D)$ are generally closer to optimal values in more complex inference situations than they are in the binomial two-hypothesis case which is the

> simplest inference situation imaginable. Apparently the binomial task is too abstract, artificial, and perhaps not analogous to any real-life inferences that all of us make from time to time.

Finally, Moskowitz [21, p. 10] states,

> Psychological experiments involving human versus Bayesian revision of probabilities almost always employ random data generating paradigms such as dice, urns, book bags-and-poker chips, etc. Although some may argue that such data producing vehicles provide more experimental control, they lack realism....

Specific attention has been paid to some of the potential differences between the real world and the laboratory that were discussed earlier in this section. In Phillips, Hays, and Edwards [29], for example, the authors speculate about the possibility that the observed conservatism could be caused by subjects behaving as if the data-generating process were characterized by nonstationarity or conditional dependence.

Du Charme and Peterson [9] work with continuous distribution and suggest that their experimental situation is more realistic than the simple bookbag-and-poker-chip paradigm (p. 541):

> Why do <u>S</u>s perform more optimally when revising estimates about the entire continuum of proportions than they do when the revision concerns only two proportions? One possibility is that tasks using a continuum are more representative of nonlaboratory inference

> tasks. The idea underlying this hypothesis is that people are basically Bayesian information processors and that as laboratory tasks become more representative of real life, Ss will perform more optimally.

It should be noted that their experimental situation is still quite artificial, involving urns and poker chips.

Chinnis and Peterson [5, p. 248] note that "it has been suggested that one reason people make conservative inferences is that they inappropriately generalize from a nonstationary environment to the stationary laboratory situations." Their experimental results appear to refute this hypothesis, but the experimental setting is artificial and the type of nonstationarity considered is very simple and may not be representative of realistic situations.

Youssef and Peterson [47] investigate cascaded inferences (inferences involving a hierarchy of several levels). They claim that "much of the real world is more complex than (a noncascaded situation)," and they make the following observations (p. 13):

> Why are people more excessive when they cascade inferences than when they do not? One possibility is that they perform better in the more realistic, hierarchical situation where it is necessary to cascade inferences. Such an explanation is in accord with previous research that has shown that quality of performance increases with task complexity, if the complexity is in the direction of being more representative of the environment in which people ordinarily behave (Peterson and Beach, [25]).

The notion of individuals being less conservative in complex situations if the situations are realistic is also noted by Du Charme and Peterson [8], who consider normal distributions of heights of men and women and find that conservatism is only about half as great as in experiments using binomial populations. They state that "the increase in optimality occurred in spite of the complexity of normal data generating processes," and they speculate as follows (p. 174):

> One reason for the increase in optimality in the present experiment may be that Ss were more familiar with the data generating process: both more familiar with the abstract nature of normal distributions and also more familiar with the particular distributions used, the heights of men and women.

In this section we have only touched upon some of the more obvious ways in which real-world situations may differ from the situations that have been used in most of the psychological experiments involving conservatism; there are undoubtedly other more subtle differences. The above quotes indicate that even at the time of the early experiments regarding conservatism, it was realized that such differences may account for some or all of the observed conservatism. Therefore, it may be useful to investigate these factors in more detail, particularly in a realistic setting.

3. Generalizations from the Laboratory to the Real World

In the abstract of Du Charme's review of "conservatism in human inference" (Du Charme, [7]), the author states,

> Much research has been directed at the variables affecting and causing errors in human probability estimation. The practical importance of the research lies in "real world" information processing systems where humans must be used to estimate probabilities.

In a similar vein, Slovic [36 p. 2] makes the following comment:

> As an experimental psychologist, my own interest is in laboratory experiments that are designed to test various hypotheses about bounded rationality. If these experiments are anything more than mere academic exercises, we should eventually be able to link their results to specific examples of suboptimal decision making in the real world.

Ultimately, then, the situations of most interest exist in the real world, not in the laboratory. As we pointed out in Section 2, however, the artificial nature of most of the experimental situations encountered in the literature regarding conservatism renders their potential generalizability to realistic situations very tenuous indeed. Thus, the implications of these experimental results for actual real-world situations are questionable.[4]

As the discussion in the previous section suggested, many researchers are aware of and concerned about the problem of generalizing results from the laboratory to the real world. Moreover, this concern is evident not only in conservatism experiments, but also in experiments concerning other aspects of inferential and decision making behavior. In an information-purchsing experiment, Fried and Peterson [15, p. 528] comment that their results (a bias toward purchasing greater amounts of less diagnostic information) "could be an overgeneralization from nonlaboratory situations where data may be relatively inexpensive when compared with payoffs." The following quote from Wallsten [40, pp. 30-31] states the case clearly:

> Virtually all previous studies of individual risky decision making have employed simple gambles in the belief that the important dimensions were value and probability, and that this would allow their study in the most pure form. Once the skeleton of the decision process has been understood, it could be fleshed out as progressively richer situations are studied. One conclusion from the present analysis is that the psychological changes from simple to progressively more complicated situations is qualitative--not quantitative--and that what is learned about choices among simple gambles is insufficient to understand the sorts of real life situations in which we are ultimately interested.

In order to investigate the possibility of generalizing from the laboratory to the real world, it is necessary to consider potential differences between laboratory situations and real-world situations. In the previous section such differences were considered within the context of research in the area of conservatism. In this section, we look at potential differences in a more general vein.

One distinguishing feature of most realistic situations is that the choice of a statistical model to represent the data-generating process is not an obvious choice. In simple situations such as the bookbag-and-poker-chip paradigm, the appropriate model is clearly a Bernoulli process, and the use of this process to determine likelihoods for probability revision tasks provides a normative model with which the subjects' probability assessments can be compared. In realistic situations, there may be some question about the appropriateness of potential statistical models, so that there is no single model that everyone would agree represents the actual situation (i.e. the model is not "public"). Hence, there is no single normative model that can serve as a basis for comparison with subjects' probability assessments. With regard to conservatism, for instance, it might be possible to claim that certain assessment procedures tend to result in smaller probability revision than other assessment procedures, but it is not possible to tell which procedure is closest to optimality in a normative Bayesian sense.

Of even greater and more fundamental importance than the absence of "public" models is the vague, ill-structured nature of many realistic situations. In the laboratory, the situation facing the subject is usually quite well-structured and orderly. For example, parcels of information frequently arrive in convenient units at regular time intervals at a fixed cost in laboratory settings, and the relationship of the information to the situation at hand is generally at least reasonably straightforward. In the real world, information is not always available in convenient units, the timing may be irregular, the costs involved may not always be obvious, and some of the ramifications of the information with respect to the situation of interest may be quite subtle. The discussion in Section 2 regarding factors such as conditional dependence, nonstationarity, unreliability, and diagnosticity is also relevant to this point regarding the structure of real-world situations.

Another distinguishing feature of realistic situations is that in order to maintain the realistic nature of the situation, the individual assessing the probabilities should possess some expertise in the substantive area of interest. It would do little good to carefully design an experiment involving the prediction of future economic variables such as Gross National Product, for instance, if the subjects are novices with respect to economics. Lichtenstein and Feeney [20] suggest that in simple experimental situations, subjects sometimes behave as though a different model is applicable. They note (p. 67) that "Caution must be taken not to confuse the subject's accuracy in probability estimation problems with his conceptual structuring of the problem." Perhaps lack of experience or expertise with a task contributes significantly to such behavior. With regard to the bookbag-and-poker-chip paradigm, Pitz [30, p. 203] states,

> Of course, it will not tell us much about the way experienced, intelligent people assess probabilities, but it will help us find out what happens to naive subjects when information is not related to outcomes in a deterministic way.

However, if research concerning human behavior in inferential and decision making situations is to have any practical signifiance, it should at least attempt to relate to situations involving experienced people.

Brief training is certainly no replacement for considerable expertise and experience in the area of interest, but even training within artificial laboratory experiments may be of some value. For example, Wheeler and Beach [41] find that training reduces conservatism, and Schum, Southard, and Wombolt [35, pp. 28-29] make the following comments:

> One is continually impressed by the relatively high degree of inference accuracy which trained subjects achieve even when the task is quite complex. In real-life diagnostic contexts there are, of course, some superb diagnosticians.

Goodman [16, p. 139] goes so far as to say, "Perhaps conservatism as a general phenomenon can be extinguished with a very few minutes of appropriate training." But in most cases even reasonably extensive training may not overcome a lack of expertise. The important question here, as in Section 2, is the generalizability of the results. "The individuals in real-life diagnostic situations, to whom we wish the results to generalize, will bring considerable experience to their tasks" (Schum, Southard, and Wombolt [35, p. 43]).

A further distinction that is useful within the context of this discussion is that between operational expertise and non-operational expertise. Of course, this difference is not a simple dichotomy, but a continuum relating to expertise and experience as well as to the ultimate use of the judgments provided by an expert. For example, consider two weather forecasters who are similar in terms of training and experience with the exception that the first prepares only deterministic (i.e. categorical) forecasts of precipitation, whereas the second prepares probabilistic forecasts of precipitation. Only the second forecaster has operational expertise with regard to the assessment of precipitation probabilities. Moreover, if the first forecaster begins to make probabilistic forecasts purely for the purposes of an experiment, while the precipitation probabilities of the second forecaster are used for experimental purposes and for dissemination to the public, then the second forecaster is preparing forecasts in more of an operational setting than is the first forecaster.

This discussion indicates that the question of laboratory situations versus real-world situations involves not a simple dichotomy, but an entire spectrum of possibilities. Realistic situations range from relatively simple and straightforward to extremely complex and ill-structured. Laboratory situations also vary considerably in complexity. The degree of generalizability in any specific instance therefore depends upon the particular nature of the real-world situation to which one wishes to generalize. The closer the laboratory experiment can be made to approximate the real world, the more confidence the experimenter should have in attempting to generalize the results. When the experimenter has a particular realistic situation in mind, every effort should be made to structure the experiment to mimic the realistic situation. In a sense, then, a direct analogy may be drawn between the building of a mathematical model of a realistic situation and the design of an experiment which is to be conducted in the laboratory but which is to be similar to a realistic situation. A tradeoff between realism and cost must be considered.

Most experiments, however, are designed to investigate general propositions rather than specific real-world situations. Therefore, if the experimenter would like to generalize the results to a large variety of real-world situations, it is necessary to conduct the experiment under a large variety of conditions in the laboratory. In this manner it can be seen whether a proposition is satisfied only under particular sets of conditions, under a broad spectrum of conditions, or under virtually no conditions. In attempting to consider the implications of the proposition for a particular realistic situation, then, one can compare the realistic situation with the various situations considered in the laboratory. The closer the various laboratory settings approximate realistic settings, the easier it is to make such comparisons and the more confident one can feel in the resulting generalizations.

To illustrate the importance of the generalizability question, consider once again the area of conservatism. Despite the questions concerning the generalizability of laboratory results regarding conservatism, a considerable amount of work regarding "information processing systems" has been based, at least in part, upon such results. Several extensive psychological experiments have been conducted in an attempt to compare various aggregation procedures (e.g. Edwards [10,11], Edwards, Phillips, Hays, and Goodman [14]). The general result of these experimental efforts has been that procedures involving judgmental aggregation (e.g. the direct assessment of posterior probabilities or odds ratios) tend to produce smaller probability revisions than procedures involving formal aggregation (e.g. the assessment of likelihoods or likelihood ratios and the use of Bayes' theorem to aggregate the information). Some argue that the latter procedures are better than the former procedures because they are less conservative. In other words, the argument essentially is that conservatism is to be combated, implying that procedures leading to more extreme probabilities are always "better." This sort of argument is implicit in the actual development of many real-world "information processing systems," and the question of whether conservatism occurs in realistic situations is seldom considered. Interestingly, in an experiment conducted by Murphy, Snapper, and Peterson [22] in a realistic weather forecasting situation, the judgmental aggregation procedure performs better than the formal aggregation procedure; instead of the former being conservative, the latter appears to be excessive. Similar results have been obtained by Domas and Peterson [6] in an artificial laboratory situation involving conditional dependence. Hence, universally equating "extremeness" with "goodness" in probabilistic information processing systems is likely to be an unwarranted generalization.

The point of this section is not to suggest that generalizations from the laboratory to the real world should never be made. Rather, we wish to emphasize that considerable care should be exercised in making such generalizations and that giving serious consideration to potential generalizations before the experiment is conducted (i.e. in the design stage) may result in laboratory situations that allow such generalizations to be made with greater confidence. Moreover, we by no means claim that studies of human inferential and decision making behavior in even the most simple, artificial laboratory situations are of no interest. On the contrary, they are of considerable interest provided that the results are kept in context. They provide valuable information about behavior in simple laboratory situations, and even more importantly, they suggest hypotheses for investigation in more realistic settings. This next step, the conducting of realistic experiments, is the subject of the next section.

4. Are Experiments in a Realistic Setting Feasible?

As indicated in Section 3, it is necessary to design and conduct some realistic experiments in order to generalize to realistic situations with a reasonable degree of confidence. Few realistic experiments have actually been conducted, however, and the purposes of this section are to explain this somewhat surprising state of affairs and to investigate the feasibility of realistic experiments regarding human inferential and decision making behavior.[5]

There is some evidence that the absence of "public" models for most real-world data processes has been an important factor in dissuading experimenters from looking at realistic situations. Edwards [11, p. 38] comments that "in most studies of systems it is not possible to assess the amount of conservatism shown by the subjects because an adequate criterion for validation is absent." In Edwards, Phillips, Hays, and Goodman [14], the authors argue against the use of empirical evaluation by comparing predictions with outcomes. In situations where a suitable normative model is available, of course, the performance of subjects can be evaluated directly by comparison with the normative model. Normative evaluation may be more informative than empirical evaluation involving comparisons of subjects' predictions with actual outcomes, since uncontrollable "noise" in the system may lead to a difference between the prediction and the outcome even if the prediction is in perfect accord with the normative model. When no normative model is available, however, prediction-outcome comparisons provide valuable information concerning the performance of the subjects, particularly if the sample size is large enough to minimize the effect of the uncontrollable noise.

Mathematical functions that provide measures of the relationship between assessed probabilities and actual outcomes are called scoring rules. The role of scoring rules in probability assessment and evaluation has received considerable attention; see, for example, Winkler [42], Murphy and Winkler [23], Staël von Holstein [39], and Savage [31]. For studies in which scoring rules are used to evaluate probability assessments in real-world situations for which no normative model is available, see Winkler and Murphy [44], Staël von Holstein [39], and Winkler [43]. The experiment by Murphy, Snapper, and Peterson [22] is of particular interest because unlike most other studies referenced here, it involves a probability revision task. Two procedures are considered, a subjective assessment of posterior probabilities after receiving new information and a subjective assessment of likelihoods, which are then aggregated formally via Bayes' theorem. The former procedure leads to smaller probability revisions than the latter procedure, so in relation to the latter procedure, it is conservative. An evaluation using scoring rules indicates, however, that the former procedure leads to better scores on the average. Therefore, instead of the former procedure being conservative, the latter procedure may be excessive in the sense that it leads to probability revisions larger than the data justify.

Although scoring rules do not involve comparisons with normative models, they can be used to compare the performance of alternative models, including probabilities assessed by different subjects, probabilities determined by statistical procedures such as regression analysis, and so on. Certain types of comparisons may be especially useful in probability revision experiments. For example, if conservatism is present, a procedure that calibrates subjects' probability revisions by making them more extreme might lead to higher average scores. It has been suggested that conservatism in dichotomous situations can be represented by implied likelihood ratios (i.e. likelihood ratios calculated by dividing subjectively assessed posterior odds ratios by prior odds ratios) of the form LR^c, where LR is a normative likelihood ratio and c is a constant between zero and one (c has been called an accuracy ratio; see Peterson, Schneider, and Miller [26]). The smaller c is, the greater the conservatism. But if this is the case, then it might be possible to improve upon subjectively assessed posterior odds ratios by using $LR^{*1/c}$ as the likelihood ratio, where LR* is the implied likelihood ratio. Scoring rules could be used to evaluate this procedure for various values of c. If the value of c between zero and one leads to the highest scores, conservatism is indicated; on the other hand, if a value of c greater than one leads to the highest scores, the subjectively assessed posterior odds ratios would appear to be excessive. If individuals are "good" at revising probabilities in realistic situations, high scores would correspond to values of c near one. Of course, alternative representations of conservatism could also be considered and evaluated via scoring rules.

The desire to use experts to achieve realism in inferential and decision making experiments surely acts to discourage realistic experiments. It is much more difficult to persuade experts to serve as experimental subjects than it is, e.g. to recruit student subjects. The administration of the experiment is often more time-consuming, for the experimenter may have to "go to the subjects" instead of having the subjects come to him. Once the experts are recruited and the administrative details are set up, however, the experts may become interested and cooperative subjects, particularly if the experimental task relates to their usual occupational tasks. An example of this phenomenon is a study by Bartos [1], in which security analysts assessed probability distributions for future security prices, where the securities considered were included in or were being considered for inclusion in the portfolio of the investment firm employing the analysts. Other examples of realistic inferential and decision making experiments involving experts are experiments conducted by Gustafson [17] in the area of medical decision making, by Murphy, Snapper, and Peterson [22], Peterson, Snapper, and Murphy [27], and Murphy and Winkler [23] in the area of weather forecasting, and by Kelly and Peterson [19] in the area of intelligence analysis.

Yet another difficulty encountered in realistic experiments is the possible lack of sufficient controls to make reliable inferences from the experiment. In an operational setting with an expert serving as a subject, it may be difficult to control such factors as the prior odds, the information sources available to the subject, the particular information obtained from these sources, the feedback available to the subject, and perhaps even the time available for the experimental tasks. For example, consider a weather forecaster participating in an operational experiment concerning the probability of precipitation. The forecaster must perform his usual duties, which include forecasting temperature, wind, and other variables; such duties necessitate keeping an eye on various weather charts, reports, and other data as they are received and updated. It would be extremely difficult to attempt to control the order in which the forecaster considers his information sources and to separate the information he considers in assessing a probability of precipitation from the information he considers in forecasting other variables.[6] Furthermore, the experimenter would be unlikely to be able to replicate exactly any particular set of conditions. Hence, the experimental data may contain a considerable amount of noise, and it is necessary to take this into account in determining sample size and other aspects of the design and in attempting to generalize the results.

Using historical data may enable the experimenter to control the availability and order of examination of the information sources, thus enabling him to design a carefully controlled yet realistic experiment. In this regard, he can vary the type, number, form, and order of examination of the information sources available to the subject and investigate the effect of such variations upon the inferential behavior of the subject. In the experiment conducted by Murphy, Snapper, and Peterson [22], sets of nine information sources (weather charts) were obtained from past data. Experienced weather forecasters then assessed the probability of precipitation on the basis of each set of charts, either directly assessing the probability after looking at all nine charts or assessing a likelihood ratio for each chart as it was observed (in which case Bayes' theorem was used to formally aggregate the likelihoods). By preparing sets of weather charts from historical data, the experimenters were able to control the nature and type of information sources presented to the subjects.

Although experiments that involve the use of historical data may enable the experimenter to exert more control over some details of the experiment while still retaining realism and using experts, other types of experiments may also provide some insight into the nature of problems that may be encountered in an operational setting. In such a setting, an individual is exposed to many information sources, some of which are difficult to describe, and one would expect to find certain conditions (some of which might be quite subtle) that cannot be reproduced in a non-operational setting. Of course, problems of implementation (e.g. cooperation or time constraints) may arise in operational experiments. Nevertheless, since inferences and decisions in actual situations are, by definition, made in operational settings, is incumbent upon the experimenter, if he wishes to generalize his results to such situations, to make every attempt to come as close as possible to the operational setting in his experiment.

In this section, we have attempted to discuss some of the difficulties that have dissuaded experimenters interested in human behavior in inferential and decision making situation from conducting more realistic experiments. Some of these difficulties, such as the evaluation of assessed probabilities and the inclusion of experimental controls, can be circumvented. Problems of implementation, on the other hand, are unavoidable to a certain extent, but these problems may be regarded as detours on a journey. Such detours may occasionally be quite irritating, but they certainly need not cause cancellation of the journey. Therefore, the question posed in the title of this section can be answered in the affirmative; experiments in a realistic setting are feasible.

5. Summary and Discussion

In this paper we have discussed the difficulty of generalizing from simple, artificial laboratory situations to more realistic situations that are often encountered in practice. Although the behavior of individuals in unrealistic situations may be of some interest, it is of little practical interest with regard to making inferences and decisions in the real world unless the results can be generalized appropriately. Therefore, the need for more realistic experiments seems clear.

In Section 2 we considered a particular subset of the experimental work regarding human behavior in inferential and decision making situations: the experimental evidence regarding conservatism. Although the phenomenon of conservatism has been found in numerous experiments, the degree of conservatism has varied considerably, suggesting that its existence and strength in any given situation may be highly dependent upon the exact nature of the situation. But if moderate experimental manipulations within the context of the simple bookbag-and-poker-chip paradigm can cause variations in the degree of conservatism, surely the differences between such artificial situations and realistic situations might produce much larger variations. This is not to say that conservatism does not exist in realistic situations; it may well exist to a considerable degree. We feel, however, that the experimental evidence presented to date is not sufficient to draw any conclusions one way or the other concerning conservatism in realistic inferential and decision making situations.

The question of generalizations from the laboratory to the real world was considered in a more general framework in Section 3. Realistic situations differ from laboratory situations in terms of how "public" the data-generating model is, how well-structured the situation itself is, how expert and experienced the subjects are, and how operational the setting is. The closer a laboratory situation approximates a real-world situation in terms of these and other factors, the more confidence the experimenter can have in attempting to generalize the results. Experiments in very simple, artificial laboratory settings are of some interest in their own right and may suggest hypotheses for further examination in more realistic settings, but the generalizability of results from these simple experiments to the real world is very questionable indeed.

The numerous quotes presented in Section 2 indicate that many of the investigators in the area of conservatism realize the difficulty of generalizing from artificial experimental situations to realistic situations. Nevertheless, until recently few investigators have attempted to conduct more realistic experiments (however, see Section 4). Of course, it is always easier to work in a laboratory situation than to design and conduct more realistic types of experiments. Besides, there is undoubtedly some feeling (perhaps rightly so) that descriptions of behavior in simple, artificial situations are of some value and that the practical implications of experimental results for actual inferential and decision making procedures are not the primary province of the experimental psychologist. Moreover, difficulties such as the evaluation of probabilities in the absence of an agreed-upon normative model, the recruitment of experts in an area of application to serve as subjects, and the problem of including controls in an operational setting contribute to the disinclination of investigators to conduct experiments in realistic situations. But the evaluation of probabilities in the absence of a normative model can be handled by scoring rules; some previous experiments have used experts, who usually make good subjects one they agree to participate; and devices such as scenarios "lifted" from historical data enable the experimenter to include some controls while maintaining realism. Thus, the difficulties are not as serious as they may seem at first glance, and experiments in a realistic setting are indeed feasible.

The need for more realistic experiments is particularly pressing in view of the rapid increase in the practical application of modern inferential and decision making models.

If the results of artificial experiments do not carry over into realistic experiments, the implications with respect to the implementation of such models could be very significant. A potential example is the development of "information processing systems" discussed in Section 3. Another potential example involves the rapidly-growing field of "decision analysis." The spirit of these examples is to "divide and conquer," i.e. to simplify matters for the decision maker by breaking the decision making problem into its component parts and by considering each part, which presumably is simpler than the problem as a whole, separately. But it may be that without considerable training, decision makers are not very good at making inferences for the simple component parts (as opposed to inferences for the entire problem) (e.g. see Howard, Matheson, and North, [18]). In general, experiments concerning human behavior in realistic inferential and decision making situations could have important implications for the determination of inputs for formal models, the training and utilization of experts, the roles of humans and computers, the gathering and summarizing of information, and many other important questions. The ultimate <u>practical</u> question with regard to studies of human behavior in inferential and decision-making situations is this: How does a highly-motivated, experienced individual in an operational setting in his area of expertise, given appropriate feedback regarding past predictions and decisions, perform inferential and decision making tasks, and can his performance be improved upon in any manner?

References

[1] Bartos, J.A. "The Assessment of Probability Distributions for Future Security Prices." Bloomington. Indiana University, unpublished doctoral dissertation, 1969.

[2] Brown, R.V. "Do Managers Find Decision Theory Useful?," Harvard Business Review, 48 (1970), 78-89.

[3] Brown, R.V., Kahr, A., and Peterson, C.R. Decision Analysis for the Manager. New York. Holt, Rinehart and Winston, 1974.

[4] Chinnis, J.O., and Peterson, C.R. "Inference about a Nonstationary Process," Journal of Experimental Psychology, 77 (1968), 620-625.

[5] Chinnis, J.O., and Peterson, C.R. "Nonstationary Processes and Conservative Inference," Journal of Experimental Psychology, 84 (1970), 248-251.

[6] Domas, P.A., and Peterson, C.R. "Probabilistic Information Processing Systems: Evaluation with Conditionally Dependent Data," Organizational Behavior and Human Performance, 7 (1972), 77-85.

[7] Du Charme, W.M. "A Review and Analysis of the Phenomenon of Conservatism in Human Inference." Houston. Rice University, unpublished manuscript, 1969.

[8] Du Charme, W.M., and Peterson, C.R. "Intuitive Inference about Normally Distributed Populations," Journal of Experimental Psychology, 78 (1968), 269-275.

[9] Du Charme, W.M., and Peterson, C.R. "Proportion Estimation as a Function of Proportion and Sample Size," Journal of Experimental Psychology, 81 (1969) 536-541.

[10] Edwards, W. "Probabilistic Information Processing in Command and Control Systems." Ann Arbor. University of Michigan, Institute of Science and Technology, Report No. 3780-12-T, 1963.

[11] Edwards, W. "Nonconservative Probabilistic Information Processing Systems." Ann Arbor. University of Michigan, Institute of Science and Technology, Report No. 5893-22-F, 1966.

[12] Edwards, W. "Conservatism in Human Information Processing." In B. Kleinmuntz, ed., Formal Representation of Human Judgment. New York. Wiley, 1968.

[13] Edwards, W., and Phillips, L.D. "Man as Transducer for Probabilities in Bayesian Command and Control Systems." In M.W. Shelly and G.L. Bryan, eds., Human judgments and optimality. New York. Wiley, 1964.

[14] Edwards, W., Phillips, L.D., Hays, W.L., and Goodman, B.C. "Probabilistic Information Processing Systems: Design and Evaluation." IEEE Transactions on Systems Science and Cybernetics, SSC-4 (1968), 248-265.

[15] Fried, L.S., and Peterson, C.R. "Information Seeking: Optional versus Fixed Stopping." Journal of Experimental Psychology, 80 (1969), 525-529.

[16] Goodman, B.C. "Action Selection and Likelihood Ratio Estimation by Individuals and Groups." Organizational Behavior and Human Performance, 7 (1972), 121-141.

[17] Gustafson, D.H. "Evaluation of Probabilistic Information Processing in Medical Decision Making." Organizational Behavior and Human Performance, 4 (1969), 20-34.

[18] Howard, R.A., Matheson, J.E., and North, D.W. "The Decision to Seed Hurricanes." Science, 176 (1972), 1191-1202.

[19] Kelly, C.W., and Peterson, C.R. "Probability Estimates and Probabilistic Procedures in Current-Intelligence Analysis: Report on Phase I." Gaithersburg, Md., IBM, unpublished manuscript, 1971.

[20] Lichtenstein, S., and Fenney, G.J. "The Importance of the Data-Generating Model in Probability Estimation." Organizational Behavior and Human Performance, 3 (1968), 62-67.

[21] Moskowitz, H. "Conservatism in Group Information Processing Behavior under Varying Management Information Systems." Lafayette, Ind., Purdue University, unpublished manuscript, 1971.

[22] Murphy, A.H., Snapper, K.J., and Peterson, C.R. "On the Process of Subjective Probability Forecasting." Boulder, Colorado, National Center for Atmospheric Research, and Ann Arbor, University of Michigan, unpublished manuscript, 1972.

[23] Murphy, A.H., and Winkler, R.L. "Scoring Rules in Probability Assessment and Evaluation." Acta Psychologica, 34 (1970), 273-286.

[24] Murphy, A.H., and Winkler, R.L. "Subjective Probability Forecasting in the Real World: Some Experimental Results." Boulder, Colorado, National Center for Atmospheric Research, and Bloomington, Indiana University, unpublished manuscript, 1973.

[25] Peterson, C.R., and Beach, L.R. "Man as an Intuitive Statistician." Psychological Bulletin, 68 (1967), 29-46.

[26] Peterson, C.R., Schneider, R.J., and Miller, A.J. "Sample Size and the Revision of Subjective Probabilities." Journal of Experimental Psychology, 69 (1965), 522-527.

[27] Peterson, C.R., Snapper, K.J., and Murphy, A.H. "Credible Interval Temperature Forecasts." Bulletin of the American Meteorological Society, 53 (1972), 966-970.

[28] Phillips, L.D. "The 'True Probability' Problem". Acta Psychologica, 34 (1970), 254-264.

[29] Phillips, L.D., Hays, W.L., and Edwards, W. "Conservatism in Complex Probabilistic Inference." IEEE Transactions on Human Factors in Electronics, HFE-7 (1966), 7-18.

[30] Pitz, G.F. "On the Processing of Information: Probabilistic and Otherwise." Acta Psychologica, 34 (1970), 201-213.

[31] Savage, L.J. "Elicitation of Personal Probabilities and Expectations." Jounal of the American Statistical Association, 66 (1971), 783-801.

[32] Schum, D.A. "Inferences on the Basis of Conditionally Nonindependent Data." Journal of Experimental Psychology, 72 (1966), 401-409.

[33] Schum, D.A. "Behavioral Decision Theory and Man-Machine Systems." In K. De Greene, ed., Systems psychology. New York, McGraw-Hill, 1970.

[34] Schum, D.A., Du Charme, W.M., and De Pitts, K.E. "Research on Human Multistage Probabilistic Inference Processes." Organizational Behavior and Human Performance, 1973, in press.

[35] Schum, D.A., Southard, J.F., and Wombolt, L.F. "Aided Human Processing of Inconclusive Evidence in Diagnostic Systems: A Summary of Experimental Evaluations." Wright-Patterson Air Force Base, Ohio, Aerospace Medical Laboratory, AMRL-TR-69-11, 1969.

[36] Slovic, P. "From Shakespeare to Simon: Speculations - and Some Evidence - about Man's Ability to Process Information." Eugene, Oregon, Oregon Research Institute Research Monograph, 12 (1972).

[37] Slovic, P., and Lichtenstein, S. "Comparison of Bayesian and Regression Approaches to the Study of Information Processing in Judgment." Organizational Behavior and Human Performance, 6 (1971), 649-744.

[38] Snapper, K.J., and Fryback, D.G. "Inferences Based on Unreliable Reports." Journal of Experimental Psychology, 87 (1971), 401-404.

[39] Staël von Holstein, C.-A.S. Assessment and Evaluation of Subjective Probability Distributions. Stockholm, Stockholm School of Economics, Economic Research Institute, 1970.

[40] Wallstein, T.S. "Subjects' Probability Estimates and Subjectively Expected Utility Theory: Their Relationship and Some Limitations." Chapel Hill, University of North Carolina, unpublished manuscript, 1970.

[41] Wheeler, G., and Beach, L.R. "Subjective Sampling Distributions and Conservatism." Organizational Behavior and Human Performance, 3 (1968), 36-46.

[42] Winkler, R.L. "The Quantification of Judgment: Some Methodological Suggestions." Journal of the American Statistical Association, 62 (1967), 1109-1120.

[43] Winkler, R.L. "Probabilistic Prediction: Some Experimental Results." Journal of the American Statistical Association, 66 (1971), 675-685.

[44] Winkler, R.L., and Murphy, A.H. "Evaluation of Subjective Precipitation Probability Forecasts." Proceedings of the First National Conference on Statistical Meteorology. Boston, American Meteorological Society, 1968, 133-141.

[45] Winkler, R.L., and Murphy, A.H. "Information Aggregation in Probabilistic Prediction." IEEE Transactions on Systems, Man, and Cybernetics, SMC-3 (1973a), 154-160.

[46] Winkler, R.L., and Murphy, A.H. "Experiments in the Laboratory and the Real World." Organizational Behavior and Human Performance, 9 (1973b), in press.

[47] Youssef, Z.I., and Peterson, C.R. "Intuitive Cascaded Inferences." Organizational Behavior and Human Performance, 1973, in press.

Footnotes

[1]We recognize that this is an old problem with which many researchers are familiar. However, because some researchers may not always keep the problem in mind, because real-world experimentation frequently is avoided unnecessarily, and because some considerations relating to realistic experients are somewhat controversial, we feel that the issues involved should be examined in some detail.

[2]We claim no originality for this suggestion. Conservatism is considered in detail here because it provides an excellent example of a phenomenon that attracted a great deal of attention, including attention in terms of implications for real-world decision making, on the basis of some highly artificial experiments and because it has since been demonstrated to be less than a universal phenomenon.

[3]Although conditional independence or dependence refers to the output of the information sources (i.e. the events observed on the various trials) rather than to the sources themselves, for convenience we will simply refer to conditional independence or dependence of the sources.

[4]Problems of generalizability arise in settings other than laboratory experimentation, of course. For example, statisticians take great pains to emphasize the dangers involved in attempting to extend inferential statements based upon a particular sample beyond the immediate population from which the sample was drawn. Nevertheless, there are many cases of unwarranted generalizations in statistical applications. One famous example is the Literary Digest poll of 1936, which erred by 19 percent in predicting the percentage of votes that Franklin D. Roosevelt would receive in the presidential election of that year. The sample was taken from readily available lists such as lists of telephone subscribers and automobile owners, and individuals with high incomes were overrepresented in these lists in relation to their representation in the overall voting population.

[5]We want to differentiate between experiments conducted in a realistic setting and the application of formal inferential and decision making techniques to handle individual problems, although this dichotomy is admittedly an oversimplification (e.g. experiments may sometimes be conducted within the context of a particular application).

Applications of these techniques, particularly in the rapidly growing field of "decision analysis," represent a significant contribution to knowledge in this area and have multiplied in recent years, as exemplified by the work conducted by the Decision Analysis Group at Stanford Research Institute (e.g. see Howard, Matheson, and North [18]) and by others (e.g. see Brown [2], and Brown, Kahr, and Peterson [3]). Most of this work, however, has involved "one-shot" applications and can only be construed as experimental in a very marginal sense. Moreover, many of these applications are confidential. We believe that although some realistic experiments have been conducted (references will be given later in this section), the number of realistic experiments reported in the published literature is minimal.

[6]We recently took a first step in this direction by controlling the examination by weather forecasters of one particular information source, the guidance forecasts provided by the U.S. National Weather Service. The experiment was conducted in an operational setting at two National Weather Service Forecast Offices, and the forecasters assessed precipitation probabilities both before and after examining guidance forecasts (the last information source examined). Some preliminary results of this experiment, along with preliminary results of two other experiments in weather forecasting conducted in realistic settings, are presented in Murphy and Winkler [24].