KATHLEEN M. WHITCOMB and P. GEORGE BENSON

EVALUATING SECOND-ORDER PROBABILITY JUDGMENTS WITH STRICTLY PROPER SCORING RULES

ABSTRACT. Empirical studies have demonstrated that uncertainty about event probabilities, also known as ambiguity or second-order uncertainty, can affect decision makers' choice preferences. Despite the importance of second-order uncertainty in decision making, almost no effort has been directed towards the development of methods that evaluate the accuracy of second-order probabilities. In this paper, we describe conditions under which strictly proper scoring rules can be used to assess the accuracy of second-order probability judgments. We investigate the effectiveness of using a particular strictly proper scoring rule – the ranked probability score – to discourage biased assessments of second-order uncertainty.

KEY WORDS: Second-order probability, ambiguity, probability assessment, proper scoring rules, ranked probability score.

1. INTRODUCTION

Subjective expected utility theory (SEUT) maintains that all the uncertainty relevant for decision making is adequately represented by precise (first-order) probabilities (Goldsmith and Sahlin, 1982). Within this framework, uncertainty about event probabilities or ambiguity is at best regarded as a useful fiction for deriving precise probability judgments.[1]

This view has been disputed in a number of empirical studies that have investigated the effects of ambiguity on decision makers' preferences for decision alternatives. Operationalizing ambiguity as a second-order probability density on probability, $f(p)$, these studies demonstrated that subjects' choice preferences were systematically impacted not only by the mean of $f(p)$, as implied by SEUT, but by the variance of $f(p)$ as well (Ellsberg, 1961; Becker and Brownson, 1964; Yates and Zukowski, 1976; Goldsmith and Sahlin, 1982; Curley and Yates, 1985; Einhorn and Hogarth, 1985; Kahn and Sarin, 1987)[2]. More recently, Boiney (1993) reported experimental results indicating that the skewness of $f(p)$ also affects choice.

Theory and Decision **41:** 165–178, 1996.

If second-order probability judgments are to be accepted and routinely used by decision makers, methods for evaluating and encouraging their accuracy must be developed. Strictly proper scoring rules are used extensively for these purposes in the assessment of precise probabilities. No comparable measures have been constructed for evaluating second-order probability judgments. In fact, we know of only one study that has even attempted to evaluate second-order probability judgments (Benson and Whitcomb, 1993). In the current study, we examine whether strictly proper scoring rules – measures that are based on the expected utility criterion – can be used to effectively evaluate the external correspondence of second-order probabilities. We investigate the effectiveness of using a particular strictly proper scoring rule – the ranked probability score – to discourage biased assessments of second-order uncertainty.

The paper is organized as follows: Section 2 provides a brief discussion of strictly proper scoring rules, with emphasis on the ranked probability score. Section 3 describes a method for applying proper scoring rules to second-order probability functions, and explores the sensitivity of the ranked probability score to biased assessments of second-order uncertainty. The results are discussed in Section 4.

2. PROPER SCORING RULES

Scoring rules are measures of the correspondence between an individual's reported probabilities and the relevant observations. Under a strictly proper scoring rule (PSR), an assessor optimizes her expected score by reporting her true probability judgments, i.e., those that reflect her beliefs. Any other assessment strategy results in a suboptimal expected score. The logarithmic, spherical, and quadratic scoring rules are the most frequently used forms of strictly proper scoring rules (cf. Winkler, 1967).

While clearly preferred to other scoring rules, PSRs do have a drawback: all are relatively insensitive to departures from the optimal strategy of reporting one's true (first-order) probability judgments (Edwards and van Winterfeldt, 1986, pp. 421–426). While some PSRs are more sensitive than others, none yield a sharp rewarding scheme (Murphy and Winkler, 1970). The issue of sensitivity is of particular concern when evaluating second-order probabilities. We will return to this point in Section 3.

In conducting the sensitivity analysis reported later in the paper, we restricted our attention to a variant of the quadratic scoring rule known as the ranked probability score. This scoring rule takes rank order relationships into account, and is the most appropriate for the evaluation of probability judgments of ordinal predictands (Murphy, 1970), such as those considered in this study.

2.1. *The Ranked Probability Score*

The ranked probability score is the squared difference between an assessor's cumulative vector of reported probability judgments and the cumulative outcome vector. Consider an event E consisting of N mutually exclusive and exhaustive outcomes $E_1, \ldots, E_N$, where a rank ordering of the alternative outcomes is a given or inherent feature of the decision problem. Thus, $E_1, \ldots, E_N$ represents an *ordered* listing. For an individual making a probability judgment for such an event, let p_i and r_i denote the assessor's true and reported probability assessments for E_i, respectively, where $\sum_{i=1}^{N} p_i = \sum_{i=1}^{N} r_i = 1.0$. Then $R_i = \sum_{k=1}^{i} r_k$ is the cumulative probability assessment corresponding to E_i. Further, let D_i denote an element of the cumulative outcome vector and the subscript j refer to the outcome that obtains. Define $D_i = 0$ if $i < j$, and $D_i = 1.0$ if $i \geq j$. The ranked probability score can then be defined as follows:

$$RPS = \sum_{i=1}^{N} (R_i - D_i)^2.$$

RPS is negatively oriented (smaller scores are better) and ranges over the interval $[0, N-1]$. Letting $j = 1, \ldots, N$, the assessor's expected ranked probability score can be expressed as:

$$E(RPS) = \sum_{j=1}^{N} p_j RPS_j,$$

where p_j refers to the assessor's true probability for the outcome that obtains (outcome j), and RPS_j is the ranked probability score, conditioned on the jth outcome obtaining. It has been shown that the assessor can optimize $E(RPS)$ only by setting $(r_1, \ldots, r_N) =$

$(p_1, \ldots, p_N)$, and, hence, that *RPS* is a strictly proper scoring rule (Murphy, 1969).

The following example demonstrates how the conventions described above are used and how the scores are computed.

Imagine that an investor is interested in two financial stocks, Stock A and Stock B. In particular, he is interested in the outcomes that neither, exactly one, or both stocks will increase in value by 10% or more in the next year. Denote these outcomes as E_1, E_2, and E_3, respectively. Suppose that a stock broker reports to the investor that she believes that the probabilities corresponding to E_1, E_2, and E_3 are 0.1, 0.2, and 0.7. Suppose further that, not wanting to appear too pessimistic, the stock broker biased her reported probabilities and, in truth, believes that the probabilities corresponding to E_1, E_2, and E_3 are 0.3, 0.4, and 0.3. Accordingly, $(r_1, r_2, r_3) = (0.1, 0.2, 0.7)$; $(R_1, R_2, R_3) = (0.1, 0.3, 1.0)$; and $(p_1, p_2, p_3) = (0.3, 0.4, 0.3)$.

If it happens that E_1 obtains, the cumulative outcome vector becomes $(D_1, D_2, D_3) = (1,1,1)$ and $RPS = (0.1 - 1)^2 + (0.3 - 1)^2 + (1.0 - 1)^2 = 1.30$. Similarly, if E_2 obtains, then $(D_1, D_2, D_3) = (0,1,1)$ and $RPS = (0.1 - 0)^2 + (0.3 - 1)^2 + (1.0 - 1)^2 = 0.50$. If E_3 obtains, $(D_1, D_2, D_3) = (0,0,1)$, and $RPS = (0.1 - 0)^2 + (0.3 - 0)^2 + (1.0 - 1)^2$ 0.10. The stock broker's expected ranked probability score is then computed as: $E(RPS) = 0.3\ (1.30) + 0.4\ (0.50) + 0.3\ (0.10) = 0.62$. The reader may verify that had the broker reported her true probabilities her $E(RPS)$ would be 0.42, which is lower and better.

3. APPLYING STRICTLY PROPER SCORING RULES TO SECOND-ORDER PROBABILITY JUDGMENTS

Now consider an expert asked to judge the probability that a trial of a new drug therapy will have a successful outcome. Since the therapy is still experimental, the assessor may not believe that he has enough information to justify the assignment of a precise probability, but instead may prefer to express his probability vaguely. Suppose he characterizes p as a uniform random variable over the interval [0.6,0.8], so that his uncertainty is expressed in terms of a second-order probability density on p, $f(p) = 1/(0.8-0.6)$ for $0.6 \le p \le 0.8$. Can SEUT-based measures such as the ranked probability score be used to evaluate and encourage unbiased assessment of second-order probability expressions such as this? That is the question that motivated our research. The following well-known theorem due to de Finetti (1937) facilitates an answer to the question:

THEOREM. *Let x_j represent a binary random variable, where the events 'failure' and 'success' are denoted by 0 and 1, respectively. Assume further that a sequence of such events is exchangeable and infinitely extendable. Given these assumptions, the predictive probability for a given sequence $\{x_1, \ldots, x_n\}$ that yields r successes in n trials may be obtained by acting as if the x_js are independent, identically distributed, and as if their density were averaged over p. That is,*

$$(3.1)\quad p\{x_1, \ldots, x_n \mid r, n\} = \int_0^1 p^r(1-p)^{n-r} dF(p),$$

where F(.) is a unique distribution function.

An equivalent expression of de Finetti's theorem (Press, 1989, p.60) which is intuitively appealing and more useful for our purposes is the following:

THEOREM. *Let S_n denote the number of successes in n exchangeable Bernoulli trials in which the probability of a success on a single trial, p, is expressed in terms of a (second-order) probability density on p, f(p). Then:*

$$(3.2)\quad \text{a. } P\{S_n = r\} = \int_0^1 \binom{n}{r} p^r(1-p)^{n-r} dF(p)$$

where dF(p) = f(p) dp, and

$$\text{b. } lim_{n\to\infty} S_n/n = p$$

with probability one for any distribution function, F(p).

If $F(p)$ represents the distribution function for the probability of a success on a single trial, then for each prescribed n, Equation (2) yields a precise predictive probability distribution that can be interpreted as a mixture of binomial probability distributions, where $F(p)$ is the mixing distribution. Since this predictive distribution is, in effect, specified by $F(p) = \int f(p)dp$, any evaluation of the predictive distribution is an evaluation of $f(p)$. Accordingly, an assessor's second-order probability density on p, $f(p)$, can be evaluated by applying a strictly proper scoring rule to the predictive probability distribution derived from $f(p)$ using Equation (2).

To demonstrate, imagine that three clinical trials of the new drug are to be conducted and we would like to obtain the predictive probabilities for the number of successful outcomes (S_3). Substituting the expert's assessment, $f(p)dp = 1/(0.8 - 0.6)dp$, for $dF(p)$ in Equation (2), the following predictive probability distribution is obtained, assuming the trials to be exchangeable and infinitely extendable[3]: $P\{S_3 = 0\} = 0.03$, $P\{S_3 = 1\} = 0.19$, $P\{S_3 = 2\} = 0.43$, $P\{S_3 = 3\} = 0.35$. These predictive probabilities can be evaluated by means of strictly proper scoring rules such as the ranked probability score. For instance, if exactly two of the three drug therapy trials were successful, the cumulative outcome and cumulative probability vectors would be (0,0,1,1) and (0.03,0.22,0.65,1.0), respectively. Recalling that the ranked probability score is simply the squared distance between these vectors, *RPS* is computed as 0.1718.

But to what extent would this evaluation strategy motivate the assessor to report her true second-order probability density, denoted as $g(p)$, rather than some other probability density function? Under an ideal scoring system, the assessor should expect to:

(1) achieve her best score only by reporting $g(p)$; and
(2) incur substantial penalties for reporting any second-order density other than $g(p)$.

We refer to the former requirement as the optimization criterion and the latter as the sensitivity criterion. We consider each in turn.

3.1. *The Optimization Criterion*

To determine whether the proposed evaluation strategy satisfies the first criterion, we need the following result: Equation (1), which applies to every n, is equivalent to the condition that the probability that every n of each prescribed set of n of the x_js takes the value 1 is

$$(3.3) \quad \int_0^1 p^n dF(p),$$

(cf. Savage, 1954, pp. 51–53; Good, 1965, p. 12).

Equation (3) implies that, for a finite sequence of n exchangeable Bernoulli trials, an individual needs only to report a second-order probability density whose first n moments correspond to those of her true density in order to optimize her expected score. When $n = 1$, evaluation of second-order probability densities through corresponding predictive probability distributions generated using Equation (1) or Equation (2) provides no incentive for the assessor to report second-order uncertainty. This follows since the expected score can be optimized by reporting any second-order probability density whose mean is the same as that of the assessor's true density. At least two trials (n>2) are necessary for the derived predictive probabilities to reflect the higher-order moments of $f(p)$, and accordingly, the assessor's second-order uncertainty.

To clarify this point, we refer again to the drug therapy example. If only one trial is to be conducted, Equation (2) will yield the same predictive probability distribution, $P\{S_1 = 0\} = 0.3$ and $P\{S_1 = 1\} = 0.7$, whether the expert assesses $f(p) = 1/(0.8\text{-}0.6)$ for $0.6 < p \leq 0.8$, or reports a precise probability equal to the mean of $f(p)$, 0.7. (In the latter case, Equation (2) reduces to the binomial formula with $p = 0.7$.) However, when $n = 2$, the predictive probability distribution is $P\{S_2 = 0\} = 0.0933$, $P\{S_2 = 1\} = 0.4133$, and $P\{S_2 = 2\} = 0.4933$ for $f(p) = 1/(0.8-0.6)$, versus $P\{S_2 = 0\} = 0.0900$, $P\{S_2 = 1\} = 0.4200$, and $P\{S_2 = 2\} = 0.4900$ for $p = 0.7$.

The proposed evaluation strategy satisfies the first criterion for an ideal scoring system to the extent that the first n moments of $f(p)$ capture the relevant dimensions of the decision maker's uncertainty.

3.2. *The Sensitivity Criterion*

Since proper scoring rules are somewhat insensitive to departures from the optimal strategy of reporting one's true first-order uncertainty, it would seem to follow that they would be even less affected by biased reports for second-order uncertainty. In the remainder of this section, we demonstrate that their effectiveness is a function of n, the number of trials under consideration.

We analyzed the sensitivity of the ranked probability score to biased reports of second-order uncertainty. Since many biasing schemes are possible, for simplicity, we focused on the situation where the assessor has some degree of uncertainty about an event probability, p, but conceals it by reporting a precise probability equal to the mean of her true second-order probability density, i.e., $p = \int p dF(p)$. In this case, the difference between the ranked probability scores corresponding to her reported and true second-order probability densities measures the expected penalty for ignoring second-order uncertainty. This penalty, expressed relative to the optimal expected score, is:

$$\Delta E_*(RPS) = \frac{E_b(RPS) - E_*(RPS)}{E_*(RPS)},$$

where $E_*(RPS)$ represents the optimal expected score – the expected score when the assessor reports her true second-order probability density, $g(p)$; $E_b(RPS)$ represents the biased expected score. For our biasing scheme, $E_b(RPS)$ is the expected score when the assessor reports a precise probability equal to the mean of $g(p)$. In computing $E_*(RPS)$, the predictive probability distributions, $(p_1, \ldots, p_N)$ and $(r_1, \ldots, r_N)$, are identical and derived by substituting $g(p)$ d(p) for d$F(p)$ in Equation (2). For $E_b(RPS)$, $(r_1, \ldots, r_N)$ differs from $(p_1, \ldots, p_N)$, but is easily derived since, for our biasing scheme, Equation (2) reduces to the binomial formula for the probability of r successes in n trials, conditional on $p = E(p) = \int p dG(p)$.

We constructed and analyzed several examples using the biasing strategy described above. For each of the examples, we assumed that $g(p)$ could be represented by a beta density with parameters $n' > r' > 0$:

$$(3.4)\quad f(p) = \begin{cases} = \frac{(n'-1)!}{(r'-1)!(n'-r'-1)!} p^{r'-1}(1-p)^{n'-r'-1}, & \text{for } 0 \le p \le 1, \\ 0 & \text{elsewhere.} \end{cases}$$

Selecting $g(p)$ from among the class of beta densities offers three advantages. First, the beta density covers a wide range of unimodal shapes over the closed interval [0,1]. Second, it requires assessment of only two parameters, r' and n', which can be expressed in terms of the mean and variance of the beta density (cf. Raiffa and Schlaifer, 1961). Finally, substituting $f_\beta(p \mid r', n')d(p)$ for $dF(p)$ in Equation (2) the expression for the predictive distribution for the number of successes, r, in n trials reduces to the well-known beta-binomial distribution:

$$P(r \mid n, r'n') \tag{3.5}$$
$$= \frac{(r + r' - 1)!(n + n' - r - r' - 1)!n!(n' - 1)!}{r!(r' - 1)!(n - r)!(n' - r' - 1)!(n + n' - 1)!}.$$

Ten beta densities are used to represent the assessors true second-order probability densities, $g(p)$, and are depicted in Figure 1. The corresponding biased (precise) assessments for p are equal to the means of these beta densities, that is $p = E(p) = \int pdG(p)$.

Since the distance between the predictive distributions corresponding to the true and biased assessments of p increases as the number of trials increases, we used a wide range of n values in computing the sensitivities. The following values of n were used in conjunction with each beta density: $n = 2, 3, 5, 10, 20, 50, 100, 500, 1000$, and 5000. Increases in sensitivities were minimal beyond $n = 5000$. Figure 2 displays plots of the sensitivity values, $\Delta E_*(RPS)$, corresponding to each of the beta densities in Figure 1.

As would be expected, the penalty for concealing second-order uncertainty, $\Delta E_*(RPS)$, increases not only with n, but with the assessor's degree of second-order uncertainty. For example, when the assessor's second-order uncertainty is high, as is the case for $f_\beta(p \mid r' = 1, n' = 2)$ and $f_\beta(p \mid r' = 3, n' = 6)$, $\Delta E_*(RPS)$ may be large enough to encourage the assessor to report at least some of her second-order uncertainty for n as small as five. Conversely, when the assessor's degree of second-order uncertainty is fairly low, as for $f_\beta(p \mid r' = 27, n' = 30)$, a much larger value of n is required to generate the same penalty. Thus, the proposed evaluation strategy produces penalties that satisfy the sensitivity criterion when n is large and/or second-order uncertainty is high.

Of course, a different strictly proper scoring rule might be more or less sensitive to second-order uncertainty than the ranked probability score (Murphy and Winkler, 1970). However, of the commonly used PSRs, only the ranked probability score is sensitive to distance, and for this reason is the most relevant for evaluating discrete predictive probability distributions (Murphy, 1970; Matheson and Winkler, p.1092, 1976).[4]

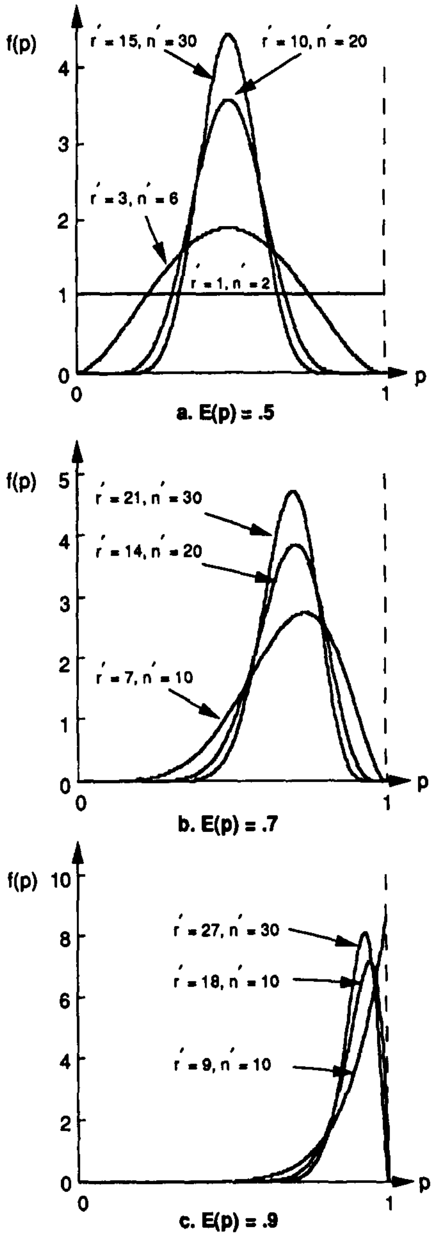

Figure 1.

4. CONCLUSION

We have described a procedure for evaluating second-order probability distributions in terms of precise predictive probabilities and

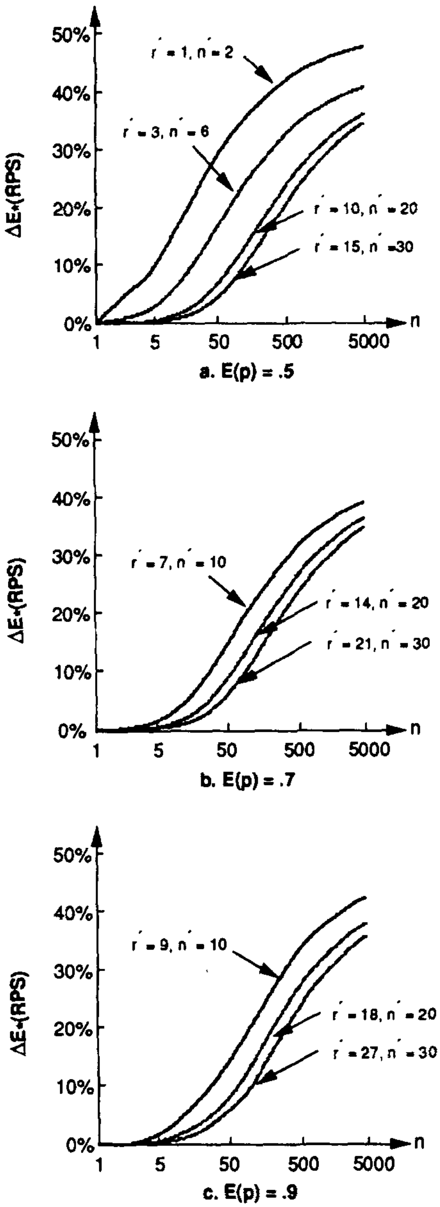

Figure 2.

observable outcomes. The procedure demonstrates that it is possible to apply SEUT-based scoring rules to second-order probability assessments.

The procedure can be utilized in forecasting tasks in which de Finetti's Theorem applies; that is, where the events of interest can be expressed in terms of finite sequences of binary trials that can (potentially) be extended to 'sufficiently large' and exchangeable sequences of trials. It can be used to discourage assessors from concealing second-order uncertainty, but only if the number of trials to be realized is relatively large.

It is not difficult to envision decision problems where events of interest conform to these criteria. For example, a public health agency may be interested in predicting how many of a large group of people exposed to a carcinogen will be adversely affected by their exposure. Or an insurance company may want to predict the number of new claims that will be submitted if the company is forced to amend its medical coverage to include a treatment previously classified as experimental. However, effective evaluation of second-order probability judgments in more general forecasting tasks will undoubtedly require the application of criteria other than SEUT-based scoring rules.

5. ACKNOWLEDGEMENT

We would like to thank Jim Lynch for his helpful comments during the preparation of this paper.

6. NOTES

[1] For example, Savage (1954, p.78) argues that if a first-order probability, p, is expressed as a distribution in terms of secondary probabilities, one should simply perform the necessary summation or integration to obtain the expectation of p, $E(p)$. $E(p)$ then serves the role of the precise probability.

[2] Throughout this paper, $f(p)$ is used to denote a probability density function of the continuous random variable, p. If an assessor believes that only a finite number of values for p are possible, $f(p)$ can be interpreted as a probability mass function of p.

[3] De Finetti's Theorem requires that there be an infinite or potentially infinite number of exchangeable trials. While de Finetti's Theorem does not hold exactly for a finite number of exchangeable trials, it does hold approximately for n finite but sufficiently large (Diaconis, 1977; Diaconis and Freedman, 1980).

[4] A scoring rule is considered to be 'sensitive' to distance if forecasts that concentrate their probability near the event that occurs receive better scores than those that do not (Murphy, 1970).

REFERENCES

Becker, S.W. and Brownson, F.O.: 1964, 'What Price Ambiguity? Or the Role of Ambiguity in Decision Making', *Journal of Political Economy* **72**, 62–73.

Benson, P.G. and Whitcomb, K.M.: 1993, 'The Effectiveness of Imprecise Probability Forecasts', *Journal of Forecasting* **12**, 139–159.

Boiney, L.G.: 1993, 'The Effects of Skewed Probability on Decision Making Under Ambiguity', *Organizational Behavior and Human Decision Processes* **56**, 134–148.

Curley, S.P., and Yates, J.F.: 1985, 'The Center and Range of Probability Intervals as Factors Affecting Ambiguity Preferences', *Organizational Behavior and Human Decision Processes* **36**, 273–287.

De Finetti, B.: 1937, 'La Prevision: Ses Lois Logiques, ses Sources Subjectives', *Annales de l'Institut Henri Poincaré*, **7** 1–68. Translated by H.E. Kyburg, Jr. and H.E. Smokler, (eds.): 1964, *Studies in Subjective probability* New York: John Wiley and Sons.

Diaconis, P.: 1977, 'Finite Forms of de Finetti's Theorem on Exchangeability', *Synthese* **36**, 271–281.

Diaconis, P. and Freedman, D.: 1980, 'De Finetti's Generalizations of Exchangeability', in: R. C. Jeffrey (ed.): *Studies in Inductive Logic and Probability* Berkeley, University of California Press, pp. 223–249.

Edwards, W. and von Winterfeldt, D.: 1986, *Decision Analysis and Behavioral Research*, Cambridge, U.K.: Cambridge University Press.

Einhorn, H.J., and Hogarth, R.M.: 1985, 'Ambiguity and Uncertainty in Probabilistic Inference', *Psychological Review* **92**, 433–461.

Ellsberg, D.: 1961, 'Risk, Uncertainty, and the Savage Axioms', *Quarterly Journal of Economics* **75**, 643–669.

Goldsmith, R.W., and Sahlin, N.E.: 1982, 'The Role of Second-order Probabilities in Decision Making', in: P.C. Humphreys, O. Svenson, and A. Vari (eds.); *Analysing and Aiding Decision Processes* Amsterdam, North Holland, pp. 319–329.

Good, I.J.: 1965, *The Estimation of Probabilities: An Essay on Modern Bayesian Methods*, Cambridge, Mass. The M.I.T. Press.

Kahn, B.E., and Sarin, R.K.: 1988, 'Modeling Ambiguity in Decision Under Uncertainty', *Journal of Consumer Research* **15**, 265–272.

Matheson, J.E. and Winkler, R.L.: 1976, 'Scoring Rules for Continuous Probability Distributions', *Management Science* **22**, 1087–1092.

Murphy, A.H.: 1969, 'On the "Ranked Probability Score",' *Journal of Applied Meteorology* **8**, 985–987.

Murphy, A.H.: 1970, 'The Ranked Probability Score and the Probability Score: A Comparison', *Monthly Weather Review* **98**, 917–924.

Murphy, A.H. and Winkler, R.L.: 1970, 'Scoring Rules in Probability Assessment and Evaluation', *Acta Psychologica* **34**, 273–286.

Press, S.J.: 1989, *Bayesian Statistics: Principles, Models, and Applications*, New York: John Wiley and Sons.

Raiffa, H. and Schlaifer, R.: 1961, *Applied Statistical Decision Theory*, Cambridge, Mass: Harvard University Press.

Savage, L.J.: 1954, *The Foundations of Statistics*, New York: John Wiley and Sons.

Winkler, R.L.: 1967, 'The Quantification of Judgment: Some Methodological Suggestions', *Journal of the American Statistical Association* **62**, 1105–1120.
Yates, J.F. and Zukowski, L.G.: 1976, 'Characterization of Ambiguity in Decision Making', *Behavioral Science* **21**, 19–25.

Kathleen M. Whitcomb
Department of Management Science
College of Business Administration
University of South Carolina,
Columbia, South Carolina 29208, U.S.A.

P. George Benson
Graduate School of Management
Rutgers, The State University of New Jersey,
Newark, New Jersey 07102, U.S.A.