Hedging and Skill Scores for Probability Forecasts

ALLAN H. MURPHY

National Center for Atmospheric Research[1], *Boulder, Colo. 80302*

(Manuscript received 29 June 1972, in revised form 20 October 1972)

ABSTRACT

An individual skill score (*SS*) and a collective skill score (*CSS*) are examined to determine whether these scoring rules are strictly proper or improper. The *SS* and the *CSS* are both standardized versions of the Brier, or probability, score (*PS*) and have been used to measure the "skill" of probability forecasts. The *SS* is defined in terms of individual forecasts, while the *CSS* is defined in terms of collections of forecasts. The *SS* and the *CSS* are shown to be *improper* scoring rules, and, as a result, both the *SS* and the *CSS* encourage hedging on the part of forecasters.

The results of a preliminary investigation of the nature of the hedging produced by the *SS* and the *CSS* indicate that, while the *SS* may encourage a considerable amount of hedging, the *CSS*, in general, encourages only a modest amount of hedging, and even this hedging decreases as the sample size K of the collection forecasts increases. In fact, *the CSS is approximately strictly proper for large collections of forecasts* ($K \geq 100$).

Finally, we briefly consider two questions related to the standardization of scoring rules: 1) the use of different scoring rules in the assessment and evaluation tasks, and 2) the transformation of strictly proper scoring rules. With regard to the latter, we identify standardized versions of the *PS* which are strictly proper scoring rules and which, as a result, appear to be appropriate scoring rules to use to measure the "skill" of probability forecasts.

1. Introduction

Meteorologists have often used the Brier, or probability, score (*PS*) (Brier, 1950), a measure of the "accuracy" of individual forecasts (see Murphy and Winkler, 1970, 1971), to evaluate probability forecasts. The *PS* is also a *strictly proper* scoring rule (e.g., Murphy and Epstein, 1967; Winkler and Murphy, 1968), and, as a result, the *PS* discourages hedging on the part of forecasters.[2]

The results of empirical studies indicate that the "accuracy" of precipitation probability forecasts, as measured by the *PS*, is related to the climatological probability of precipitation (see, for example, Sanders, 1963; Hughes, 1965; Glahn and Jorgenson, 1970). Specifically, for climatological probabilities less than one-half, the "accuracy" of the forecasts increases, in general, as the climatological probability increases, while for climatological probabilities greater than one-half, their "accuracy" decreases, in general, as the climatological probability increases. In order to measure the "accuracy" of these forecasts relative to climatology, "skill scores" have been formulated in terms of the *PS* for the forecasts and the *PS* for the climatological probabilities (see Sanders, 1963; Hughes, 1965; Glahn and Jorgenson, 1970). Specifically, an "individual" skill score (*SS*) has been defined in terms of individual forecasts and a "collective" skill score (*CSS*) has been defined in terms of collections of forecasts. The *CSS* is presently used by the National Weather Service (NWS) to measure the "skill" of their precipitation probability forecasts (Hughes, 1965; Glahn and Jorgenson, 1970).[3] The purposes of this paper are to show that the *SS* and the *CSS* are *not* strictly proper scoring rules, to present the results of a preliminary investigation of the nature of the hedging produced by the *SS* and the *CSS*, and to briefly consider two questions related to the standardization of scoring rules.

In Section 2 we define the *PS*, the *SS*, and the *CSS*. The concepts of hedging and strictly proper and improper scoring rules are briefly discussed in Section 3. In Sections 4 and 5 we show that the *SS* and the *CSS*, respectively, are not, in general, strictly proper scoring rules. We present the results of a preliminary investigation of the nature of the hedging produced by the *SS* and the *CSS* in Section 6. The results indicate that while the *SS* may encourage a considerable amount of hedging, the *CSS*, in general, encourages only a modest amount of hedging, and even this hedging decreases as the sample size of the collection of forecasts increases. Two questions related to the standardization of scoring rules are briefly considered in Section 7: 1) the use of

[1] The National Center for Atmospheric Research is sponsored by the National Science Foundation.

[2] For a recent detailed description of the nature and properties of the *PS*, refer to Murphy (1970).

[3] The need for such standardized scoring rules, in order to be able to compare forecasters, forecast offices, etc., has recently been emphasized by Hughes (1971).

different scoring rules in the assessment and evaluation tasks, and 2) the transformation of strictly proper scoring rules. With regard to the latter, we identify standardized versions of the *PS* which are strictly proper scoring rules. Section 8 consists of a brief summary and conclusion.

2. The individual and collective skill scores

a. The PS

Consider a meteorological variable whose range has been divided into N mutually exclusive and collectively exhaustive states $\{s_1,\cdots,s_N\}$. Let the vector

$$\mathbf{r}=(r_1,\cdots,r_N)\ (r_n\geqslant 0,\sum_n r_n=1\ ;n=1,\cdots,N)$$

denote the forecaster's forecast, where r_n is the forecast probability of state s_n, and let the vector $\mathbf{d}=(d_1,\cdots,d_N)$ denote the relevant observation, where $d_n=1$ if state s_n obtains and $d_n=0$ otherwise.

Then, the *PS* for the forecast $\mathbf{r}$ is defined as follows:

$$PS(\mathbf{r},\mathbf{d})=\sum_{n=1}^{N}(r_n-d_n)^2. \tag{1}$$

The range of the *PS* is the closed interval $[0,2]$ and the *PS* has a negative orientation (i.e., the smaller the score the better the score). When state s_j obtains $PS(\mathbf{r},\mathbf{d})$ becomes $PS_j(\mathbf{r})$, where

$$PS_j(\mathbf{r})=1-2r_j+\sum_{n=1}^{N}r_n^2. \tag{2}$$

Now, consider a collection of K forecasts $\mathbf{r}_k=(r_{1k},\cdots,r_{Nk})$ and the K relevant observations $\mathbf{d}_k=(d_{1k},\cdots,d_{Nk})$ $(k=1,\cdots,K)$. Then, the total *PS* for the collection of K forecasts is $TPS(\mathbf{r},\mathbf{d})$, where, from (1),

$$TPS(\mathbf{r},\mathbf{d})=\sum_{k=1}^{K}\sum_{n=1}^{N}(r_{nk}-d_{nk})^2. \tag{3}$$

b. The SS

Let the vector

$$\boldsymbol{\pi}=(\pi_1,\cdots,\pi_N)\ (\pi_n\geqslant 0,\sum_n \pi_n=1\ ;n=1,\cdots,N)$$

denote the set of relevant climatological probabilities, where π_n is the climatological probability of state s_n.

Then, the individual skill score *SS* for the forecast $\mathbf{r}$ is defined as

$$SS(\mathbf{r},\boldsymbol{\pi},\mathbf{d})=[PS(\boldsymbol{\pi},\mathbf{d})-PS(\mathbf{r},\mathbf{d})]/PS(\boldsymbol{\pi},\mathbf{d}),$$

or

$$SS(\mathbf{r},\boldsymbol{\pi},\mathbf{d})=1-[PS(\mathbf{r},\mathbf{d})/PS(\boldsymbol{\pi},\mathbf{d})]. \tag{4}$$

The range of the *SS* is the half-open interval $(-\infty,1]$ and the *SS* has a positive orientation (i.e., the larger the score the better the score).[4] Note that the *SS* represents the fractional amount by which the *PS* for the forecast $(r_1,\cdots,r_N)$ improves upon the *PS* for the climatological probabilities $(\pi_1,\cdots,\pi_N)$. When state s_j obtains, $SS(\mathbf{r},\boldsymbol{\pi},\mathbf{d})$ becomes $SS_j(\mathbf{r},\boldsymbol{\pi})$, where

$$SS_j(\mathbf{r},\boldsymbol{\pi})=1-[PS_j(\mathbf{r})/PS_j(\boldsymbol{\pi})]. \tag{5}$$

The total *SS* for the collection of K forecasts is $TSS(\mathbf{r},\boldsymbol{\pi},\mathbf{d})$, where, from (4),

$$TSS(\mathbf{r},\boldsymbol{\pi},\mathbf{d})=K-\sum_{k=1}^{K}[PS(\mathbf{r}_k,\mathbf{d}_k)/PS(\boldsymbol{\pi}_k,\mathbf{d}_k)]. \tag{6}$$

c. The CSS

The collective skill score *CSS* for the collection of K forecasts is defined as

$$CSS(\mathbf{r},\boldsymbol{\pi},\mathbf{d})=[\sum_{k=1}^{K}PS(\boldsymbol{\pi}_k,\mathbf{d}_k)-\sum_{k=1}^{K}PS(\mathbf{r}_k,\mathbf{d}_k)]/\sum_{k=1}^{K}PS(\boldsymbol{\pi}_k,\mathbf{d}_k),$$

or

$$CSS(\mathbf{r},\boldsymbol{\pi},\mathbf{d})=1-[\sum_{k=1}^{K}PS(\mathbf{r}_k,\mathbf{d}_k)/\sum_{k=1}^{K}PS(\boldsymbol{\pi}_k,\mathbf{d}_k)]. \tag{7}$$

The range of the *CSS* is also the half-open interval $(-\infty,1]$ and the *CSS* also has a positive orientation.[5] Note that the *CSS* represents the fractional amount by which the *PS* for the K forecasts $\mathbf{r}_k$ improves upon the *PS* for the K climatological probabilities $\boldsymbol{\pi}_k$ $(k=1,\cdots,K)$. In addition, note that $CSS(\mathbf{r},\boldsymbol{\pi},\mathbf{d})$ is *not*, in general, equivalent (i.e., linearly related) to $TSS(\mathbf{r},\boldsymbol{\pi},\mathbf{d})$.[6]

3. Hedging and strictly proper and improper scoring rules

Let the vector

$$\mathbf{p}=(p_1,\cdots,p_N)\ (p_n\geqslant 0,\sum_n p_n=1\ ;n=1,\cdots,N)$$

denote the forecaster's "degree of belief," or judgment, where p_n is the judgmental probability that state s_n will obtain. *Hedging* is said to occur whenever a forecaster's forecast $\mathbf{r}$ does not correspond to his judgment $\mathbf{p}$, i.e., whenever $\mathbf{r}\neq\mathbf{p}$.

A scoring rule, $S(\mathbf{r},\mathbf{d})$, with a positive orientation is said to be *strictly proper* if and only if the forecaster

[4] The half-open interval $(-\infty,1]$ is the *possible* range of the *SS*. The *actual* range of the *SS* depends upon the climatological probabilities $(\pi_1,\cdots,\pi_N)$ and is a closed interval unless $\pi_n=1$ for some n $(n=1,\cdots,N)$.

[5] The half-open interval $(-\infty,1]$ is the *possible* range of the *CSS*. The *actual* range of the *CSS* depends upon the climatological probabilities $(\pi_{1k},\cdots,\pi_{Nk})$ $(k=1,\cdots,K)$ and is a closed interval unless $\pi_{nk}=1$ for some n for all k $(n=1,\cdots,N;\ k=1,\cdots,K)$.

[6] In fact, $CSS(\mathbf{r},\boldsymbol{\pi},\mathbf{d})$ is equivalent to $TSS(\mathbf{r},\boldsymbol{\pi},\mathbf{d})$ only if $PS(\boldsymbol{\pi}_j,\mathbf{d}_j)=PS(\boldsymbol{\pi}_k,\mathbf{d}_k)$ for all j and k $(j,k=1,\cdots,K;\ j\neq k)$, in which case $CSS(\mathbf{r},\boldsymbol{\pi},\mathbf{d})=(1/K)TSS(\mathbf{r},\boldsymbol{\pi},\mathbf{d})$.

maximizes his (subjective) expected score $ES(\mathbf{r},\mathbf{p})$, where

$$ES(\mathbf{r},\mathbf{p})=\sum_{j=1}^{N} p_j S_j(\mathbf{r}), \tag{8}$$

by making his forecast $\mathbf{r}$ correspond to his judgment $\mathbf{p}$ (e.g., Murphy and Epstein, 1967; Winkler and Murphy, 1968).[7] A strictly proper scoring rule, then, discourages hedging on the part of forecasters. On the other hand, the scoring rule $S(\mathbf{r},\mathbf{d})$ is said to be *improper* if the forecaster's expected score $S(\mathbf{r},\mathbf{p})$ attains a unique maximum for some forecast $\mathbf{r}$ which does not correspond to his judgment $\mathbf{p}$. An improper scoring rule, then, encourages hedging on the part of forecasters.

4. Hedging and the individual skill score

The SS, when state s_j obtains, becomes, from (2) and (5),

$$SS_j(\mathbf{r},\boldsymbol{\pi})=1-[(1-2r_j+\sum_{n=1}^{N} r_n{}^2)/PS_j(\boldsymbol{\pi})]. \tag{9}$$

Then, the expected score is $ESS(\mathbf{r},\boldsymbol{\pi},\mathbf{p})$, where

$$ESS(\mathbf{r},\boldsymbol{\pi},\mathbf{p})=\sum_{j=1}^{N} p_j SS_j(\mathbf{r},\boldsymbol{\pi}),$$

or, from (9),

$$ESS(\mathbf{r},\boldsymbol{\pi},\mathbf{p})=1-\sum_{j=1}^{N} p_j[(1-2r_j+\sum_{n=1}^{N} r_n{}^2)/PS_j(\boldsymbol{\pi})]. \tag{10}$$

Now, taking the derivative of $ESS(\mathbf{r},\boldsymbol{\pi},\mathbf{p})$ with respect to r_m, we obtain

$$\partial ESS(\mathbf{r},\boldsymbol{\pi},\mathbf{p})/\partial r_m=[2p_m/PS_m(\boldsymbol{\pi})]-(2r_m)\sum_{n=1}^{N}[p_n/PS_n(\boldsymbol{\pi})],$$

and setting this derivative equal to zero yields

$$p_m-r_m{}^*PS_m(\boldsymbol{\pi})\sum_{n=1}^{N}[p_n/PS_n(\boldsymbol{\pi})]=0,$$

or

$$r_m{}^*=[p_m/PS_m(\boldsymbol{\pi})]/\sum_{n=1}^{N}[p_n/PS_n(\boldsymbol{\pi})], \quad (m=1,\cdots,N), \tag{11}$$

where $r_m{}^*$ is the "optimal" value of r_m, i.e., the value of r_m which maximizes[8] the forecaster's expected SS for particular values of the climatological probabilities $(\pi_1,\cdots,\pi_N)$ and the forecaster's judgments $(p_1,\cdots,p_N)$. Note that $r_m{}^*\neq p_m$ unless $\pi_n=1/N$ for all $n(n=1,\cdots,N)$ or unless $p_m=0$ or 1. Thus, the SS is *not*, in general, a strictly proper scoring rule.

5. Hedging and the collective skill score

The concepts of hedging and strictly proper (and improper) scoring rules relate to individual probability forecasts and to a forecaster's behavior when making such forecasts. Specifically, in order to determine whether a scoring rule is strictly proper or improper, we must assume that the forecaster has a definite strategy in mind when he makes a forecast. In this paper we shall assume that the forecaster has already made K forecasts $\mathbf{r}_k$ $(k=1,\cdots,K)$, that he wants to maximize his expected score on the $(K+1)$st forecast $\mathbf{r}_{K+1}$, and that the $(K+1)$st forecast is *independent* of the previous K forecasts.[9]

The CSS, when state s_j obtains on the $(K+1)st$ occasion, becomes, from (3) and (7),

$$CSS_j(\mathbf{r}_{K+1},\boldsymbol{\pi}_{K+1})=1-\frac{[TPS(\mathbf{r},\mathbf{d})+PS_j(\mathbf{r}_{K+1})]}{[TPS(\boldsymbol{\pi},\mathbf{d})+PS_j(\boldsymbol{\pi}_{K+1})]}. \tag{12}$$

Then, the expected CSS is $ECSS(\mathbf{r},\boldsymbol{\pi},\mathbf{p})$, where

$$ECSS(\mathbf{r},\boldsymbol{\pi},\mathbf{p})=\sum_{j=1}^{N} p_{j,K+1}CSS_j(\mathbf{r}_{K+1},\boldsymbol{\pi}_{K+1}),$$

or, from (2) and (12),

$$ECSS(\mathbf{r},\boldsymbol{\pi},\mathbf{p})=1-\sum_{j=1}^{N} p_{j,K+1}\times\frac{[TPS(\mathbf{r},\mathbf{d})+(1-2r_{j,K+1}+\sum_{n=1}^{N} r^2{}_{n,K+1})]}{[TPS(\boldsymbol{\pi},\mathbf{d})+PS_j(\boldsymbol{\pi}_{K+1})]}. \tag{13}$$

Now, taking the derivative of $ECSS(\mathbf{r},\boldsymbol{\pi},\mathbf{p})$ with respect to $r_{m,K+1}$, we obtain

$$\partial ECSS(\mathbf{r},\boldsymbol{\pi},\mathbf{p})/\partial r_{m,K+1}=\{2p_{m,K+1}/[TPS(\boldsymbol{\pi},\mathbf{d})+PS_m(\boldsymbol{\pi}_{K+1})]\}-(2r_{m,K+1})\sum_{n=1}^{N}\{p_{n,K+1}/[TPS(\boldsymbol{\pi},\mathbf{d})+PS_n(\boldsymbol{\pi}_{K+1})]\}.$$

Setting this derivative equal to zero yields

$$p_{m,K+1}-r^*{}_{m,K+1}[TPS(\boldsymbol{\pi},\mathbf{d})+PS_m(\boldsymbol{\pi}_{K+1})]\sum_{n=1}^{N}\{p_{n,K+1}/[TPS(\boldsymbol{\pi},\mathbf{d})+PS_n(\boldsymbol{\pi}_{K+1})]\}=0,$$

[7] We assume that the forecaster's utility function is linearly related to the scoring rule $S(\mathbf{r},\mathbf{d})$. That is, we assume that the forecaster attaches a "value" to the score he receives for his forecast which is linearly related to the score itself.

[8] Note that $\partial^2 ESS(\mathbf{r},\boldsymbol{\pi},\mathbf{p})/\partial r_m{}^2<0$.

[9] Other strategies are, of course, available to the forecaster. For example, we could have assumed that the forecaster, prior to making any forecasts, wants to maximize his expected score on each of K occasions and that the forecasts on the K occasions are *dependent*. Dynamic programming could be used to solve this problem for the set of K optimal forecasts $\mathbf{r}_k{}^*(k=1,\cdots,K)$.

TABLE 1. The forecast probability r_1^* which maximizes the forecaster's expected *SS* in the two-state situation for selected values of the climatological probability π_1 and the forecaster's judgment p_1.

		Forecaster's judgment p_1										
		0.0	0.1	0.2	0.3	0.4	0.5	0.6	9.7	0.8	0.9	1.0
Climatological probability π_1	0.0	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	†
	0.1	0.0000	0.0014	0.0031	0.0053	0.0082	0.0122	0.0182	0.0280	0.0471	0.1000	1.0000
	0.2	0.0000	0.0069	0.0154	0.0286	0.0400	0.0588	0.0857	0.1273	0.2000	0.3600	1.0000
	0.3	0.0000	0.0200	0.0439	0.0730	0.1091	0.1552	0.2160	0.3000	0.4235	0.6231	1.0000
	0.4	0.0000	0.0471	0.1000	0.1600	0.2286	0.3077	0.4000	0.5091	0.6400	0.8000	1.0000
	0.5	0.0000	0.1000	0.2000	0.3000	0.4000	0.5000	0.6000	0.7000	0.8000	0.9000	1.0000
	0.6	0.0000	0.2000	0.3600	0.4909	0.6000	0.6923	0.7714	0.8400	0.9000	0.9529	1.0000
	0.7	0.0000	0.3769	0.5765	0.7000	0.7840	0.8448	0.8909	0.9270	0.9561	0.9800	1.0000
	0.8	0.0000	0.6400	0.8000	0.8727	0.9143	0.9412	0.9600	0.9714	0.9846	0.9931	1.0000
	0.9	0.0000	0.9000	0.9529	0.9720	0.9818	0.9878	0.9918	0.9947	0.9969	0.9986	1.0000
	1.0	†	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

† The value of r_1^* is not defined for this combination of values of π_1 and p_1.

or

$$r^*_{m,K+1}=\{p_{m,K+1}/[TPS(\boldsymbol{\pi},\mathbf{d})+PS_m(\boldsymbol{\pi}_{K+1})]\}/\sum_{n=1}^{N}\{p_{n,K+1}/[TPS(\boldsymbol{\pi},\mathbf{d})+PS_n(\boldsymbol{\pi}_{K+1})]\},\quad (m=1,\cdots,N),\quad (14)$$

where $r^*_{m,K+1}$ is the optimal value of $r_{m,K+1}$.[10] Note that $r^*_{m,K+1}\neq p_{m,K+1}$ unless $\pi_{n,K+1}=1/N$ for all n $(n=1,\cdots,N)$ or unless $p_{m,K+1}=0$ or 1. Thus, the *CSS* is *not*, in general, a strictly proper scoring rule.

6. The nature of the hedging produced by the individual and collective skill scores

a. Hedging and the SS

With regard to the nature of the hedging produced by the *SS*, we can identify several specific questions of interest:

1) How much hedging is produced by the *SS*, i.e., how large are the differences in probability between the optimal forecast $\mathbf{r}^*$ and the judgment $\mathbf{p}$?

2) How "flat" is the expected *SS* function, $ESS(\mathbf{r},\boldsymbol{\pi},\mathbf{p})$, e.g., how large are the differences in expected score between the optimal forecast $\mathbf{r}^*$ and the judgment $\mathbf{p}$?

3) How does the flatness of $ESS(\mathbf{r},\boldsymbol{\pi},\mathbf{p})$ compare with that of the expected *PS* function, $EPS(\mathbf{r},\mathbf{p})$, e.g., is the *SS* more or less likely to encourage hedging than the *PS* is to discourage hedging?

The amount of hedging produced by the *SS* can be measured in terms of the difference in probability between the optimal forecast $\mathbf{r}^*$ and the judgment $\mathbf{p}$. In the two-state $(N=2)$, i.e., "precipitation–no precipitation," situation the optimal forecast is $\mathbf{r}^*=(r_1^*,r_2^*)$, where, from (11),

$$r_1^*=p_1/[p_1+(\pi_2/\pi_1)^2p_2] \quad (15)$$

$(r_2^*=1-r_1^*,\ \pi_2=1-\pi_1$, and $p_2=1-p_1)$. We present the

[10] Note that $\partial^2 ECSS(\mathbf{r},\boldsymbol{\pi},\mathbf{p})/\partial r^2_{m,K+1}<0$.

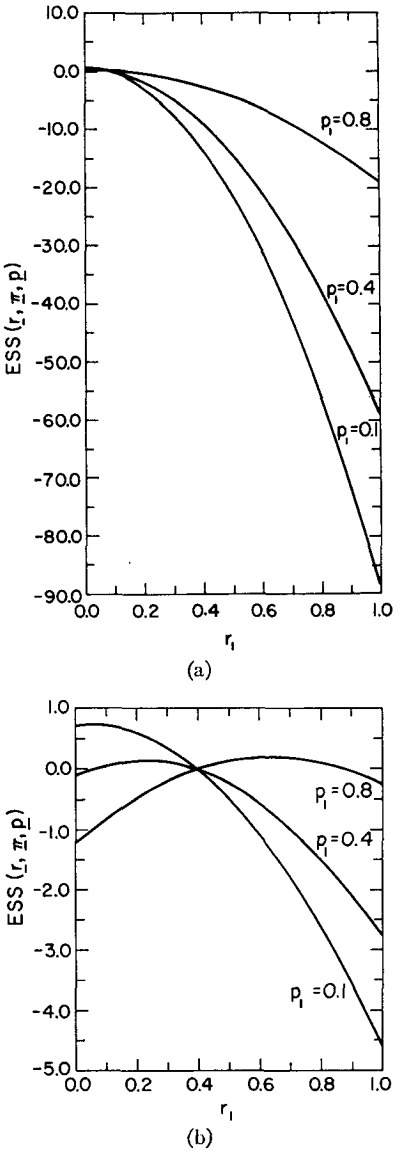

FIG. 1. The expected *SS*, $ESS(\mathbf{r},\boldsymbol{\pi},\mathbf{p})$, in the two-state situation for selected values of the forecaster's judgment p_1 when the climatological probability π_1 equals 0.1, (a), and 0.4, (b).

TABLE 2. The loss in expected score in the two-state situation if the forecaster makes his forecast r_1 correspond to his judgment p_1 instead of maximizing the function $ESS(r,\pi,p)$ by setting r_1 equal to the optimal forecast r_1^*, i.e., $ESS(\mathbf{r}^*,\boldsymbol{\pi},\mathbf{p})-ESS(\mathbf{p},\boldsymbol{\pi},\mathbf{p})$.

		Forecaster's judgment p_1										
		0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Climatological probability π_1	0.0	†	†	†	†	†	†	†	†	†	†	†
	0.1	0.000	0.877	3.112	6.113	9.288	12.044	13.792	13.938	11.899	7.111	0.000
	0.2	0.000	0.197	0.692	1.348	2.025	2.585	2.893	2.818	2.250	1.139	0.000
	0.3	0.000	0.065	0.227	0.432	0.634	0.782	0.836	0.762	0.547	0.226	0.000
	0.4	0.000	0.016	0.056	0.102	0.143	0.166	0.167	0.139	0.089	0.031	0.000
	0.5	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	0.6	0.000	0.031	0.089	0.139	0.167	0.166	0.143	0.102	0.056	0.016	0.000
	0.7	0.000	0.226	0.547	0.762	0.836	0.782	0.634	0.432	0.227	0.065	0.000
	0.8	0.000	1.139	2.250	2.818	2.893	2.585	2.025	1.348	0.692	0.197	0.000
	0.9	0.000	7.111	11.899	13.938	13.792	12.044	9.288	6.113	3.112	0.877	0.000
	1.0	†	†	†	†	†	†	†	†	†	†	†

† The loss in expected score, $ESS(\mathbf{r}^*,\boldsymbol{\pi},\mathbf{p})-ESS(\mathbf{p},\boldsymbol{\pi},\mathbf{p})$, is not defined for $\pi_1=0.0$ or 1.0.

values of r_1^* in Table 1 for selected values of π_1 and p_1. Note that $r_1^* \neq p_1$ unless $\pi_1=0.5$ or unless $p_1=0.0$ or 1.0. In addition, note that the amount of hedging is quite large when the climatological probability π_1 is either greater than 0.6 or less than 0.4. For example, $|r_1^*-p_1| \geqslant 0.1$ for $0.2 \leqslant p_1 \leqslant 0.8$ when $\pi_1 \geqslant 0.6$ or $\pi_1 \leqslant 0.4$, while $r_1^* \geqslant (\leqslant) 0.9(0.1)$ for all values of p_1 when $\pi_1 \geqslant (\leqslant)$ 0.9(0.1). Specifically, if $\pi_1=0.2$ and $p_1=0.4$, then $r_1^*=0.04$, while if $\pi_1=0.6$ and $p_1=0.8$, then $r_1^*=0.90$. Thus, if a forecaster is aware that his forecasts are being evaluated by the *SS* and if his utility function is linearly related to the *SS* (see Footnote 7), then the *SS* could produce a considerable amount of hedging.

The amount of hedging which is produced by the *SS* also depends, in part, upon the flatness of the expected *SS* function, $ESS(\mathbf{r},\boldsymbol{\pi},\mathbf{p})$. For example, if this function is quite flat over a range of values which includes the optimal forecast $\mathbf{r}^*$ as well as the judgment $\mathbf{p}$, then the forecaster has little incentive to hedge and he may well make his forecast $\mathbf{r}$ correspond closely to his judgment $\mathbf{p}$. The function $ESS(\mathbf{r},\boldsymbol{\pi},\mathbf{p})$ in the two-state (i.e., precipitation–no precipitation) situation is depicted in Figs. 1a and 1b for selected values of p_1 when π_1 equals 0.1 and 0.4, respectively. Note that in a "dry" climate, i.e., when $\pi_1=0.1$, the function $ESS(\mathbf{r},\boldsymbol{\pi},\mathbf{p})$ is certainly not flat. Moreover, while the range of this function decreases as p_1 increases, the loss in expected score if the forecaster makes his forecast r_1 correspond to his judgment p_1 instead of maximizing the function by setting r_1 equal to the optimal forecast r_1^*, i.e., $ESS(\mathbf{r}^*,\boldsymbol{\pi},\mathbf{p})-ESS(\mathbf{p},\boldsymbol{\pi},\mathbf{p})$, increases from 0.877 when $p_1=0.1$ to 11.899 when $p_1=0.8$. On the other hand, in a relatively "wet" climate, i.e., when $\pi_1=0.4$, the function $ESS(\mathbf{r},\boldsymbol{\pi},\mathbf{p})$ is quite flat (note the change in the scale on the ordinate). In addition, the range of $ESS(\mathbf{r},\boldsymbol{\pi},\mathbf{p})$ decreases as p_1 increases and the loss in expected score, $ESS(\mathbf{r}^*,\boldsymbol{\pi},\mathbf{p})-ESS(\mathbf{p},\boldsymbol{\pi},\mathbf{p})$, increases from 0.016 when $p_1=0.1$ to only 0.089 when $p_1=0.8$. Thus, the shape of the function $ESS(\mathbf{r},\boldsymbol{\pi},\mathbf{p})$ is such that the *SS* encourages a considerable amount of hedging on the part of forecasters, particularly when the climatological probability π_1 is small.[11] The loss in expected score, $ESS(\mathbf{r}^*,\boldsymbol{\pi},\mathbf{p})-ESS(\mathbf{p},\boldsymbol{\pi},\mathbf{p})$, in the two-state situation is presented in Table 2 for selected values of π_1 and p_1.

In order to compare the *SS* and the *PS* with regard to the degree to which the former encourages and the latter discourages hedging, the expected scores must be transformed in such a way that the range and orientation of $ESS(\mathbf{r},\boldsymbol{\pi},\mathbf{p})$ and $EPS(\mathbf{r},\mathbf{p})$ are identical. We present these transformed expected scores, $ESS^*(\mathbf{r},\boldsymbol{\pi},\mathbf{p})$ and $EPS^*(\mathbf{r},\mathbf{p})$, in the two-state situation in Figs. 2a and 2b, respectively, for selected values of p_1 when π_1 equals 0.1 and 0.4.[12] A comparison of $ESS^*(\mathbf{r},\boldsymbol{\pi},\mathbf{p})$ and $EPS^*(\mathbf{r},\mathbf{p})$ is difficult because these functions attain their maxima, and occasionally even their minima, for different values of r_1.[13] However, with reference to their respective maxima, the function $ESS^*(\mathbf{r},\boldsymbol{\pi},\mathbf{p})$ appears to be slightly flatter, i.e., less "sensitive," than the function $EPS^*(\mathbf{r},\mathbf{p})$. Thus, in terms of relative expected scores, the *SS* is slightly less likely to encourage hedging than the *PS* is to discourage hedging. The reader should be aware that the discussion in this paragraph has been concerned with relative expected scores, while the discussion in the previous paragraph was concerned with absolute expected scores. In this regard, the absolute sensitivity of a scoring rule can be changed by performing a linear transformation on the scoring rule; on the other hand, such a transformation will not change its relative sensitivity. However, the use of such a transformation may produce a scoring rule whose range is such that the forecaster's

[11] This statement is equally valid when the climatological probability π_1 is large. For example, $ESS(\mathbf{r},\boldsymbol{\pi},\mathbf{p})$ when $\pi_1=0.9$ is a mirror image of $ESS(\mathbf{r},\boldsymbol{\pi},\mathbf{p})$ when $\pi_1=0.1$, and, in addition, the loss in expected score, $ESS(\mathbf{r}^*,\boldsymbol{\pi},\mathbf{p})-ESS(\mathbf{p},\boldsymbol{\pi},\mathbf{p})$, when $\pi_1=0.9$ and $p_1=0.2$ is the same as that when $\pi_1=0.1$ and $p_1=0.8$.

[12] A different linear transformation is required for each combination of values of π_1 and p_1 for the *SS*. For example, when $\pi_1=0.1$ and $p_1=0.8$, $ESS^*(\mathbf{r},\boldsymbol{\pi},\mathbf{p})=9.952ESS(\mathbf{r},\boldsymbol{\pi},\mathbf{p})+0.997$. On the other hand, a different linear transformation is required for each value of p_1 for the *PS*. For example, when $p_1=0.1$, $EPS^*(\mathbf{r},\mathbf{p})=-0.617EPS(\mathbf{r},\mathbf{p})+1.111$.

[13] The function $ESS^*(\mathbf{r},\boldsymbol{\pi},\mathbf{p})$ attains its maximum at $r_1=r_1^*$ and its minimum at $r_1=0$ if $p_1\pi_1^2 \geqslant p_2\pi_2^2$ and at $r_1=1$ if $p_1\pi_1^2 \leqslant p_2\pi_2^2$, while the function $EPS^*(\mathbf{r},\mathbf{p})$ attains its maximum at $r_1=p_1$ and its minimum at $r_1=0$ if $p_1 \geqslant 0.5$ and at $r_1=1$ if $p_1 \leqslant 0.5$.

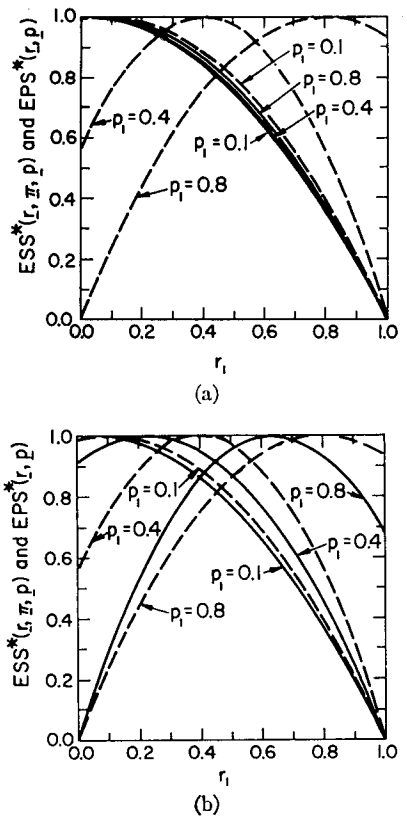

FIG. 2. The transformed expected *SS* and expected *PS*, $ESS^*(\mathbf{r},\boldsymbol{\pi},\mathbf{p})$ and $EPS^*(\mathbf{r},\mathbf{p})$, respectively, in the two-state situation for selected values of the forecaster's judgment p_1 when the climatological probability π_1 equals 0.1 (a) and 0.4 (b). The solid (dashed) curves denote the expected *SS* (*PS*) function.

utility function is no longer linear in the score (see Footnote 7). For example, a forecaster's utility function may be linear in $SS(\mathbf{r},\boldsymbol{\pi},\mathbf{d})$ but not in $SS^*(\mathbf{r},\boldsymbol{\pi},\mathbf{d})$ (or vice versa). Thus, while linear transformations of scoring rules provide a means of changing the absolute flatness of expected scoring rule functions, the use of such transformations may introduce nonlinear utility considerations.

b. *Hedging and the CSS*

Since the *CSS* is defined in terms of a collection of forecasts, in order to investigate the nature of the hedging produced by the *CSS* we must consider collections of forecasts rather than individual forecasts. In this paper we shall be primarily concerned with the amount of hedging produced by the *CSS* for hypothetical collections of forecasts in the two-state situation.

The optimal forecast in the two-state situation is $\mathbf{r}^*_{K+1}=(r^*_{1,K+1},r^*_{2,K+1})$, where, from (14),

$$r^*_{1,K+1}=\{p_{1,K+1}/[TPS(\boldsymbol{\pi},\mathbf{d})+\pi^2_{2,K+1}]\}/\{p_{1,K+1}/[TPS(\boldsymbol{\pi},\mathbf{d})+\pi^2_{2,K+1}]+p_{2,K+1}/[TPS(\boldsymbol{\pi},\mathbf{d})+\pi^2_{1,K+1}]\} \quad (16)$$

($r^*_{2,K+1}=1-r^*_{1,K+1}$, $\pi_{2,K+1}=1-\pi_{1,K+1}$, and $p_{2,K+1}=1-p_{1,K+1}$). Note that $r^*_{1,K+1}$ depends upon the *PS* for the K climatological probabilities π_k $(k=1,\cdots,K)$, $TPS(\boldsymbol{\pi},\mathbf{d})$, but not upon the *PS* for the K forecasts $\mathbf{r}_k$ $(k=1,\cdots,K)$, $TPS(\mathbf{r},\mathbf{d})$. Thus, we are actually concerned not with the forecasts in the hypothetical collections but with the climatological probabilities which correspond to the forecasts.

Consider a collection of ten forecasts ($K=10$) in the two-state situation and suppose that the climatological probabilities of states s_1 and s_2 are equal to 0.2 and 0.8, respectively, on all of the ten occasions [i.e., $\pi_{1k}=0.2$ and $\pi_{2k}=0.8$ $(k=1,\cdots,K)$]. Further, suppose that states s_1 and s_2 obtain on two and eight of the ten occasions, respectively [i.e., $\bar{d}_1=0.2$ and $\bar{d}_2=0.8$; $\bar{d}_n=(1/K)\sum_n d_{nk}$ $(n=1,2;\ k=1,\cdots,K)$]. Then, $TPS(\boldsymbol{\pi},\mathbf{d})$ for this collection of forecasts is, from (3), 3.20. Now, suppose that the forecaster wants to maximize his expected *CSS* on the eleventh forecast ($K+1=11$) when $\pi_{1,11}=0.2$ and $p_{1,11}=0.4$. Then, from (16), $r^*_{1,11}=0.36$. Note that $r^*_{1,11}\neq p_{1,11}$. However, the amount of hedging produced by the *CSS*, 0.04, is considerably smaller than the amount of hedging that would have been produced by the *SS*, 0.36 ($r_1^*=0.04$ when $\pi_1=0.2$ and $p_2=0.4$; see Table 1). Moreover, even this modest amount of hedging decreases as the sample size K of the collection of forecasts increases. For example, if the forecaster had made forecasts on twenty occasions ($K=20$) with the same climatological probabilities, i.e., $\pi_{1k}=0.2$ and $\pi_{2k}=0.8$ $(k=1,\cdots,K)$, and if states s_1 and s_2 had obtained on four and sixteen occasions, respectively, i.e., $\bar{d}_1=0.2$ and $\bar{d}_2=0.8$, then $TPS(\boldsymbol{\pi},\mathbf{d})$ would equal 6.40. If, in addition, $\pi_{1,21}$ and $p_{1,21}$ are equal to 0.2 and 0.4, respectively, as in the previous situation, then $r^*_{1,21}=0.3788$. Thus, increasing the sample size from 10 to 20 forecasts reduces the amount of hedging from 0.04 to 0.0212. We present the optimal forecast $r^*_{1,K+1}$ in the two-state situation in Table 3 for collections of forecasts with selected sample sizes K, when the sample relative frequency $\bar{d}_1$ and the

TABLE 3. The optimal forecast $r^*_{1,K+1}$ in the two-state situation in the presence of the *CSS* for collections of forecasts with selected sample sizes K, when the sample relative frequency $\bar{d}_1$ and the climatological probability, π_{1k} $(k=1,\cdots,K)$, are (a) equal and (b) unequal. The climatological probability, π_{1k} $(k=1,\cdots,K+1)$, equals 0.2 for all of the collections of forecasts and the forecaster's judgment, $p_{1,K+1}$, equals 0.4.

	(a) $\bar{d}_1=\pi_{1k}$ $(k=1,\cdots,K)$		(b) $\bar{d}_1\neq\pi_{1k}$ $(k=1,\cdots,K)$	
Sample size K	Sample relative frequency $\bar{d}_1$	Optimal forecast $r^*_{1,K+1}$	Sample relative frequency $\bar{d}_1$	Optimal forecast $r^*_{1,K+1}$
10	0.20	0.3600	0.40	0.3760
20	0.20	0.3788	0.35	0.3862
50	0.20	0.3912	0.30	0.3936
100	0.20	0.3956	0.25	0.3962
200	0.20	0.3978	0.25	0.3981

climatological probability, π_{1k} $(k=1,\cdots,K)$, are equal, as in the two previous situations, and unequal. Note that for sample sizes $K>50$, the amount of hedging is less than 0.01 in both cases and that, for a given sample size, the amount of hedging is greater when the sample relative frequency and climatological probability are equal than when they are unequal. This relationship holds in N-state $(N>2)$ as well as in two-state $(N=2)$ situations. In addition, note, from (16), that, since the PS for the climatological probabilities, $PS(\boldsymbol{\pi},\mathbf{d})$, is a maximum when the climatological probabilities, π_n $(n=1,\cdots,N)$, are all equal to $1/N$, the amount of hedging encouraged by the CSS decreases as these probabilities approach equality (i.e., as π_1 approaches one-half in the two-state situation). While we have considered the hedging produced by the CSS for only a few hypothetical collections of forecasts, we believe that the hedging which is produced for these collections is representative of the hedging which would be produced for other collections with similar sample sizes.[14]

In this regard, consider the expression for the optimal forecast $r^*_{m,K+1}$ in (14). Note that as the sample size K increases, $TPS(\boldsymbol{\pi},\mathbf{d})$ increases, and, for K large (i.e., $K\geq 100$), $TPS(\boldsymbol{\pi},\mathbf{d})\gg PS_n(\boldsymbol{\pi}_{K+1})$ for all n $(n=1,\cdots,N)$. Then, $r^*_{m,K+1}$ in (14) becomes

$$r^*_{m,K+1}\approx[p_{m,K+1}/TPS(\boldsymbol{\pi},\mathbf{d})]/\sum_{n=1}^{N}[p_{n,K+1}/TPS(\boldsymbol{\pi},\mathbf{d})],$$

or, since

$$\sum_n p_{n,K+1}=1 \quad (n=1,\cdots,N),$$

$$r^*_{m,K+1}\approx p_{m,K+1}. \tag{17}$$

Thus, *the CSS is approximately strictly proper for large collections of forecasts.*

Clearly, then, the length of the period over which a forecaster attempts to maximize his expected score in part determines the amount of hedging produced by the CSS on a particular occasion. For example, if the period is a month, during which the forecaster might make twenty forecasts $(K\approx 20)$, then the amount of hedging would probably be greater than 0.02 during most of the period, while if the period is a season, during which the forecaster might make sixty forecasts $(K\approx 60)$, then the amount of hedging would probably be less than 0.02 during most of the period (see Table 3). These results suggest that, if the CSS is used to evaluate precipitation probability forecasts, then the period over which the scores are accumulated should be three or more months in length.

[14] The hedging produced by the CSS in the two-state situation is such that $r^*_{1,K+1}<p_{1,K+1}$ for $p_1<0.5$ and $r^*_{1,K+1}>p_{1,K+1}$ for $p_1>0.5$.

7. The standardization of scoring rules: Some questions

A number of questions arise regarding the standardization of scoring rules and we briefly consider two such questions in this section.

a. The assessment and evaluation tasks

Since assessment and evaluation can be considered to be distinct tasks and since the standardization process may often produce an improper scoring rule, could different scoring rules be used for assessing, or formulating, and for evaluating probability forecasts? For example, could the PS, a strictly proper scoring rule, be used in the assessment task to discourage hedging on the part of the forecaster, while the SS, an improper scoring rule, is used in the evaluation task to measure the "skill" of his forecasts? In this regard, Winkler (1969) has indicated that, while the same scoring rule need not be used in the assessment and evaluation tasks, game theoretic considerations could arise if different scoring rules were used in the two tasks. Specifically, if the forecaster is aware that he is being rewarded or penalized according to one scoring rule (e.g., the PS) to discourage hedging and evaluated according to another scoring rule (e.g., the SS) to measure "skill," then hedging may still be to his advantage rather than to his disadvantage. Thus, while different scoring rules could be used in connection with the assessment and evaluation tasks, their use might present certain problems.

b. The standardization of strictly proper scoring rules

If we assume that the scoring rule to be used in the evaluation task is a transformation of the strictly proper scoring rule used in the assessment task, then what transformations of this scoring rule are permissible (in the sense that the transformed or standardized scoring rule is also strictly proper)? In this regard, a *linear transformation* of a strictly proper rule yields a strictly proper scoring rule (e.g., Winkler and Murphy, 1968). That is, if a scoring rule $S(\mathbf{r},\mathbf{d})$ with a positive orientation is strictly proper, then the scoring rule $S^*(\mathbf{r},\mathbf{d})$, where

$$S^*(\mathbf{r},\mathbf{d})=aS(\mathbf{r},\mathbf{d})+b, \tag{18}$$

in which a and b are constants, is strictly proper.[15] However, since $S^*(\mathbf{r},\mathbf{d})$ and $S(\mathbf{r},\mathbf{d})$ are equivalent, such transformations are of little interest relative to measuring the "skill" of probability forecasts. On the other hand, a nonlinear transformation of a strictly proper scoring rule does not, in general, yield a strictly proper scoring rule. With regard to linear and nonlinear transformations of strictly proper scoring rules, the SS

[15] Since $S(\mathbf{r},\mathbf{d})$ has a positive orientation, $S^*(\mathbf{r},\mathbf{d})$ has a positive orientation if $a>0$ and a negative orientation if $a<0$.

appears to be a linear transformation of the PS, in which $a=-1/PS(\boldsymbol{\pi},\mathbf{d})$ and $b=1$ [see (4)]. However, $PS(\boldsymbol{\pi},\mathbf{d})$ is not a constant; the value of $PS(\boldsymbol{\pi},\mathbf{d})$ depends upon the state that obtains. That is, in general, $PS_j(\boldsymbol{\pi})\neq PS_k(\boldsymbol{\pi})$ $(j,k=1,\cdots,N;\ j\neq k)$. Now, while a, in (18), must be a constant (in the sense that the value of a cannot depend upon the state that obtains), no such restriction need be placed upon the constant b. In this regard, a *modified* skill score MSS for the forecast r can be defined as

$$MSS(\mathbf{r},\boldsymbol{\pi},\mathbf{d})=PS(\boldsymbol{\pi},\mathbf{d})-PS(\mathbf{r},\mathbf{d}).^{16} \tag{19}$$

The individual MSS, in (19), *is* a strictly proper scoring rule with a positive orientation, the range of which is the closed interval $[-2,2]$. Note that the individual MSS represents the amount by which the PS for the forecast $(r_1,\cdots,r_N)$ improves upon the PS for the climatological probabilities $(\pi_1,\cdots,\pi_N)$. The collective MSS, $CMSS$, i.e., the MSS for a collection of K forecasts, is then defined as

$$CMSS(\mathbf{r},\boldsymbol{\pi},\mathbf{d})=TPS(\boldsymbol{\pi},\mathbf{d})-TPS(\mathbf{r},\mathbf{d}). \tag{20}$$

The $CMSS$ represents the amount by which the total PS for the K forecasts $\mathbf{r}_k$ improves upon the PS for the K climatological probabilities $\boldsymbol{\pi}_k$ $(k=1,\cdots,K)$. If we make the same assumptions in this situation as we made in connection with the CSS (see Section 5), then we can easily show that the $CMSS$ is a strictly proper scoring rule with a positive orientation, the range of which is the closed interval $[-2K,2K]$. The MSS and the $CMSS$, then, appear to be appropriate scoring rules to use to measure the "skill" of probability forecasts.

Note that the MSS for the forecast $\mathbf{r}$ is simply the numerator in the expression for the SS [cf. (4) and (9)]; that is,

$$SS(\mathbf{r},\boldsymbol{\pi},\mathbf{d})=MSS(\mathbf{r},\boldsymbol{\pi},\mathbf{d})/PS(\boldsymbol{\pi},\mathbf{d}). \tag{21}$$

Now, the SS appears to be a linear transformation of the MSS, in which $a=1/PS(\boldsymbol{\pi},\mathbf{d})$ and $b=0$. However, since $PS(\boldsymbol{\pi},\mathbf{d})$ is not a constant, the SS is an improper scoring rule. A similar relationship exists among the CSS, the $CMSS$, and the TPS [cf. (7) and (20)]. In this case, although the $CMSS$ is strictly proper, the CSS is improper because $TPS(\boldsymbol{\pi},\mathbf{d})$ is not a constant. Thus, only certain nonlinear transformations of strictly proper scoring rules are permissible; for example, those nonlinear transformations of the form $S^*(\mathbf{r},\mathbf{d})=aS(\mathbf{r},\mathbf{d})+b$, in which a is a constant and b depends upon the climatological probabilities.

8. Summary and conclusion

In this paper we have examined the individual and collective skill scores, SS and CSS, respectively, both of which are standardized versions of the Brier, or probability, score, PS, and have been used to measure the "skill" of probability forecasts. The SS is defined in terms of individual forecasts, while the CSS is defined in terms of collections of forecasts. We have shown that the SS and the CSS are, in general, *improper* scoring rules, and that, as a result, both the SS and the CSS encourage hedging on the part of forecasters.

We have also presented the results of a preliminary investigation of the nature of the hedging produced by the SS and the CSS. In this regard, in terms of both probability and expected score, the SS may encourage a considerable amount of hedging, particularly when the climatological probabilities are small (or large). However, the SS appears to be slightly less likely to encourage hedging than the PS is to discourage hedging. The CSS, on the other hand, encourages only a modest amount of hedging, and even this hedging decreases as the sample size of the collection of forecasts increases. In fact, we have shown that *the CSS is approximately strictly proper for large collections of forecasts.*

Finally, we have briefly considered two questions related to the standardization of scoring rules: 1) the use of different scoring rules in the assessment and evaluation tasks, and 2) the transformation of strictly proper scoring rules. First, while different scoring rules can be used in the assessment and evaluation tasks, e.g., the PS in the assessment task to discourage hedging and the SS in the evaluation task to measure "skill," the use of different scoring rules in these two tasks may introduce game theoretic considerations which could produce hedging on the part of forecasters even in the presence of the PS. Second, while a linear transformation of a strictly proper scoring rule is, of course, strictly proper, certain nonlinear transformations may also be strictly proper. Specifically, the additive constant in the linear transformation can be a function of the climatological probabilities. In this regard, the modified skill scores, MSS and $CMSS$, which represent the difference between the PS for the relevant climatological probabilities and the PS for the forecast or forecasts of concern and which, as a result, *are* strictly proper, appear to be appropriate scoring rules to use to measure the "skill" of probability forecasts.

Acknowledgments. The author would like to acknowledge the valuable comments of Lawrence A. Hughes and Robert L. Winkler on earlier versions of this manuscript.

[16] $MSS(\mathbf{r},\boldsymbol{\pi},\mathbf{d})$ has been described by Sanders (1958, 1963).

REFERENCES

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1–3.

Glahn, H. R., and D. L. Jorgenson, 1970: Climatological aspects of the Brier P-score. *Mon. Wea. Rev.*, **98**, 136–141.

Hughes, L. A., 1965: On the probability forecasting of the occurrence of precipitation. Wea. Bur. Tech. Note 20-CR-3, Washington, D. C., 36 pp.

——, 1971: Comments on "Nonlinear utility and the probability score." *J. Appl. Meteor.*, **10**, 335.

Murphy, A. H., 1970: The ranked probability score and the probability score: A comparison. *Mon. Wea. Rev.*, **98**, 917–924.

——, and E. S. Epstein, 1967: A note on probability forecasts and "hedging." *J. Appl. Meteor.*, **6**, 1002–1004.

——, and R. L. Winkler, 1970: Scoring rules in probability assessment and evaluation. *Acta Pschol.*, **34**, 273–286.

——, and ——, 1971: Forecasters and probability forecasts: Some current problems. *Bull. Amer. Meteor. Soc.*, **52**, 239–247.

Sanders, F., 1958: The evaluation of subjective probability forecasts. Sci. Rept. No. 5, Contract AF 19(604)-1305, Dept. of Meteorology, M.I.T., 62 pp.

——, 1963: On subjective probability forecasting. *J. Appl. Meteor.*, **2**, 191–201.

Winkler, R. L., 1969: Scoring rules and the evaluation of probability assessors. *J. Amer. Statis. Assoc.*, **64**, 1073–1078.

——, and A. H. Murphy, 1968: "Good" probability assessors. *J. Appl. Meteor.*, **7**, 751–758.