Assessing probabilistic forecasts about particular situations

Kesten C. Green

Business and Economic Forecasting Unit,
Monash University, Melbourne, Australia.
Contact: PO Box 10800, Wellington 6143, New Zealand.
Email kesten@kestencgreen.com
T +64 4 976 3245
F +64 4 976 3250

April 25, 2004

Abstract

How useful are probabilistic forecasts of the outcomes of particular situations? Potentially, they contain more information than unequivocal forecasts and, as they allow a more realistic representation of the relative likelihood of different outcomes, they might be more accurate and therefore more useful to decision makers. To test this proposition, I first compared a Squared-Error Skill Score (*SESS*) based on the Brier score with an Absolute-Error Skill Score (*AESS*), and found the latter more closely coincided with decision-makers' interests. I then analysed data obtained in researching the problem of forecasting the decisions people make in conflict situations. In that research, participants were given lists of decisions that might be made and were asked to make a prediction either by choosing one of the decisions or by allocating percentages or relative frequencies to more than one of them. For this study I transformed the percentage and relative frequencies data into probabilistic forecasts. In most cases the participants chose a single decision. To obtain more data, I used a rule to derive probabilistic forecasts from structured analogies data, and transformed multiple singular forecasts for each combination of forecasting method and conflict into probabilistic forecasts. When compared using the *AESS*, probabilistic forecasts were not more skilful than unequivocal forecasts.

Key words: accuracy, error measures, evaluation, forecasting methods, prediction.

When forecasters are presented with a small set of possible outcomes, some of them might prefer to assign probabilities to the options rather than choose one of them. Moreover, it is possible that probabilistic forecasts contain more information than unequivocal forecasts and, because they allow a more realistic representation of the relative likelihood of different outcomes, they might be more useful to decision makers. In order to test this proposition, it is necessary to be able to compare the accuracy of the forecasts. Comparisons of accuracy are straightforward when each forecaster chooses a single option, but not when they provide probabilistic forecasts.

Brier Score and the Squared-Error Skill Score

Brier Score

The Brier Score was developed by Brier (1950) as a means of assessing the relative accuracy of probabilistic forecasts. It is recommended for this purpose by later authors including Lichtenstein, Fischhoff, and Phillips (1982), and Doggett (1998). The Brier Score is an average of the sums of the squared errors of probabilistic forecasts. For one set of probabilities for the possible outcomes of a single event, the formula for the Brier Score (*BS*) is more succinct (because the average calculation is avoided) and, arguably, more easily comprehended. On that basis the measure is the sum of the squares of the errors, or the dot-product of the error vector and itself. The formula for the Brier Score is shown in Exhibit 1.

Exhibit 1
Brier Score (*BS*)

$$BS = \left(F - A\right) \bullet \left(F - A\right)$$

F is the vector of probabilistic-forecast (F_1,…, F_k) for the k possible outcomes of the forecasting objective or target situation. A is the vector of actual outcomes (A_1,…, A_k) for the k outcome options of the forecasting objective or target situation, where the element representing the actual outcome is coded as one and all other elements are coded as zero.

The Brier scores are in the range between zero and two. A completely accurate forecast would have a Brier score of zero, and one that was completely wrong would have a Brier score of two[1]. That is, a forecast that allocates one to the outcome option that actually occurs will have a Brier score of zero, irrespective of the number of options. In a case with four possible outcomes, the Brier score for such a forecast could be represented as $BS(1*,0,0,0) = 0$, where the asterisk marks the actual outcome. A forecast that allocates 1.00 to any other option will have a *BS* of 2: for example, $BS(0*,1,0,0) = 2$. The actual-outcome vector would be (1,0,0,0) in both examples.

[1] A potential source of confusion is that the Brier score has also been formulated as the squared error of the probability assigned to a single event that may or may not occur (e.g. Fuller 2000). Using this alternative formulation, a forecast of 0.9 for the chance of rain tomorrow would have a *BS* of 0.01 if it did rain ($(0.9-1)^2$) and 0.81 if it did not ($(0.9-0)^2$). With this formulation, the Brier score takes values between zero and one.

Squared-error skill score

The Brier score does not take account of the difficulty of the forecasting problem, or class of forecasting problem, that is being considered. To remedy this, one can use a measure that provides an assessment of forecast accuracy relative to the accuracy of forecast from a default method: the method that is usually used for similar problems, or the most basic method that might be used. Such a measure is sometimes referred to as a skill score. The Brier skill score is a Squared-Error Skill Score (*SESS*) and can be defined as shown in Exhibit 2a.

Exhibit 2a
Squared-Error Skill Score (*SESS*)

$$SESS = 1 - \frac{BS}{BS_D}$$

The ratio of the Brier Score of the forecast from the method under consideration (*BS*) to the Brier score of the forecast from the default method (BS_D) is subtracted from one. This results in a skill score with a value of one when the forecast is completely accurate and a lower value otherwise.

In the case of forecasting the decisions people make in conflicts, the evidence shows that the forecasts of experts using their unaided judgment are no more accurate than chance. A reasonable default method for these forecasting problems is therefore the equal-likelihood forecast, whereby equal probabilities are allocated to each outcome option. The denominator in the Exhibit 2b formula is a simplification of the Brier score calculation for the equal-likelihood forecast, and *k* is the number of outcome options.

Exhibit 2b
Squared-Error Skill Score (*SESS*)
where the default method is the equal-likelihood forecast

$$SESS = 1 - \frac{BS}{(k-1)/k}$$

Problems with squaring

Armstrong (2001) examined the evidence on measures for evaluating forecasting methods and found that measures based on squaring forecast errors were unreliable and difficult to interpret. On the other hand, he found that the relative absolute error (*RAE)* measure performed well against other error measures.

Relative Absolute Errors and the Absolute-Error Skill Score

For probabilistic forecasting, the skill score analogue of the *RAE* is the Absolute Error Skill Score or *AESS*. It is the sum of the absolute errors of a set of probabilities assigned to the outcome options, divided by the sum of the absolute errors from a default method, all subtracted from one in order to obtain a skill score that takes a value of one when the forecast is accurate and a lower value otherwise. The formula for the *AESS* is shown in Exhibit 3a.

Exhibit 3a
Absolute error skill score (*AESS*)

$$AESS = 1 - \frac{|F - A| \bullet U}{|F_D - A| \bullet U}$$

U is the unity vector of the same dimension as the probabilistic-forecast vector and the actual-outcome vector, and F_D is the probabilistic-forecast vector from the default method.

As discussed above, in the case of forecasting the decisions people make in conflicts a reasonable default method is the equal-likelihood forecast. Analogously with the *SESS* (Exhibit 2b), the denominator in the Exhibit 3b formula is a simplification of the sum of the absolute errors from the equal-likelihood forecast ($|F_D - A| \bullet U$), and k is the number of outcome options.

Exhibit 3b
Absolute error skill score (*AESS*)
where the default method is the equal-likelihood forecast

$$AESS = 1 - \frac{|F - A| \bullet U}{2(k-1)/k}$$

For the purposes of this paper I examine the use of skill scores in which the default method is the equal-likelihood forecast.

SESS and *AESS* compared

A forecast is perfectly accurate if a probability of one is allocated to the outcome that actually occurs. Both the *SESS* and *AESS* take on a value of one in the case of perfect accuracy. Both scores take on a value of zero (no skill) when probabilities are evenly distributed across outcome options.

The greatest inaccuracy occurs when there are only two possible outcomes and a probability of one is allocated to the outcome that does *not* occur. In this case, the *SESS* takes on a value of -

3.0 and the *AESS* takes on a value of -1. The forecasting task is harder the more outcome options there are, and both skill scores reflect this by giving higher (less negative) scores when the number of outcome options is higher. When a probability of one is allocated to an outcome that does not occur, the *SESS* tends asymptotically to -1, and the *AESS* tends asymptotically to zero, as the number of outcome options increases.

The *SESS* distinguishes among forecasts that are completely inaccurate—those that allocated a probability of 0.0 to the outcome that actually occurred—and that have the same number of outcome options (*k*). Forecasts in which probabilities were evenly allocated across outcomes that did not occur are rated as having higher (less negative) skill than those in which a high probability was allocated to a single outcome that did not occur. The *AESS* does not distinguish among completely inaccurate forecasts for situations with the same number of outcome options.

It seems a reasonable and useful property of a skill score that for a set of forecasts in which a probability of 1.0 is allocated to one or other of the outcome options that the average skill score should be 0.0 (no skill) if the proportion of accurate forecasts equals $^1/_k$. This is the case for the *AESS*, but not for the *SESS*.

I examined the average *SESS* and *AESS* scores for an illustrative set of probabilistic forecasts that would have led a decision maker to expect one outcome ahead of others. In other words, these are forecasts in which one outcome option is allocated a probability greater than that allocated to any of the other options. In Exhibit 4, the first column, headed "Probabilistic forecasts", shows the prbabilities allocated to each of the four possible outcomes. In each forecast, the first outcome option has been allocated the largest probability and is shown in bold. Columns two to seven of the Exhibit show the average skill score for each probabilistic forecast assuming the forecast were repeated many times and the first outcome option occurred with the frequency indicated by the percentage figure in the column header. For example, the average *SESS* of a forecast (0.75, 0.09, 0.08, 0.08) repeated many times, with the first outcome occurring with a frequency of 90% is 0.89.

Exhibit 4
Average skill scores for illustrative forecasts by outcome frequencies
(Forecasts for four outcome options)

Squared-Error Skill Scores (SESS)

Probabilistic forecasts	Frequency with which 1st outcome occurs * (percent)					
	100	**90**	**75**	**50**	**25**	**0**
(**1.00**, 0.00, 0.00, 0.00)	1.00	0.73	0.33	-0.33	-1.00	-1.67
(**0.95**, 0.02, 0.02, 0.01)	1.00	0.75	0.37	-0.25	-0.87	-1.49
(**0.95**, 0.05, 0.00, 0.00)	0.99	0.74	0.37	-0.25	-0.87	-1.50
(**0.75**, 0.09, 0.08, 0.08)	0.89	0.71	0.44	0.00	-0.44	-0.89
(**0.75**, 0.25, 0.00, 0.00)	0.83	0.66	0.39	-0.06	-0.50	-0.94
(**0.50**, 0.17, 0.17, 0.16)	0.56	0.47	0.33	0.11	-0.11	-0.33
(**0.30**, 0.24, 0.23, 0.23)	0.13	0.11	0.08	0.04	0.00	-0.05

Absolute-Error Skill Scores (AESS)

Probabilistic forecasts	Frequency with which 1st outcome occurs * (percent)					
	100	**90**	**75**	**50**	**25**	**0**
(**1.00**, 0.00, 0.00, 0.00)	1.00	0.87	0.67	0.33	0.00	-0.33
(**0.95**, 0.02, 0.02, 0.01) (**0.95**, 0.05, 0.00, 0.00)	0.93	0.81	0.62	0.31	0.00	-0.31
(**0.75**, 0.09, 0.08, 0.08) (**0.75**, 0.25, 0.00, 0.00)	0.67	0.58	0.44	0.22	0.00	-0.22
(**0.50**, 0.17, 0.17, 0.16)	0.33	0.29	0.22	0.11	0.00	-0.11
(**0.30**, 0.24, 0.23, 0.23)	0.07	0.06	0.04	0.02	0.00	-0.02

* Other outcomes are assumed to occur in equal proportion

In Exhibit 4 I have framed the average skill scores with values of zero (no skill) or less. The *SESS* finds no skill or worse in forecasts for which the outcome option allocated a probability of 0.75 or greater occurs 50% of the time. An evaluator using the *SESS* might therefore mistakenly reject a method or forecaster as having no skill. The *AESS*, on the other hand, would not mislead an evaluator in this way, and would lead to the sensible conclusion that any set of forecasts likely to lead a decision maker to correctly anticipate actual outcomes more frequently than chance, should be preferred to the equal-likelihood forecast.

Exhibit 4 also shows that when outcomes occur with equal frequency, the average *SESS* is negative (worse than no skill) if the forecast probabilities are not equal for all outcome options. The average *AESS* does not distinguish between forecasts in such situations and takes on a value of zero or "no skill" in all cases.

The average *SESS* is higher the more closely forecast probabilities match the frequency of actual outcomes. In contrast, the average *AESS* is higher the greater the forecast probability allocated

to the outcome that occurs most frequently. Look down the 75% column of Exhibit 4 for an illustration of this phenomenon.

In summary, the *AESS* is a better measure than the *SESS* for choosing between forecasting methods and forecasters when forecasts are in the form of probabilities allocated to outcome options. While the *SESS* makes more distinctions than the *AESS*, these distinctions are unlikely to help an evaluator choose the best method. Most importantly, the *SESS* can lead an evaluator to reject a useful method as having no skill and should therefore be avoided.

AESS and percent correct forecasts compared

Research on forecasting decisions in conflict situations has to date compared forecasting methods using the the relative percentage of correct forecasts from the methods. Chance accuracy—the accuracy one could expect to achieve from choosing a decision at random from a finite set of possible decisions—has been used to make the broad distinction between useful and invalid methods. The *AESS* as defined in this paper measures the skill of the forecaster or method relative to chance and so the *AESS* should lead to similar assessments of the relative and absolute merits of methods. Indeed it is desirable that it should do so.

To determine whether use of the *AESS* would lead to similar assessments of conflict forecasting methods as percent correct forecasts, I calculated probabilistic forecasts from the singular forecasts obtained in research on forecasting decisions in conflicts. For each combination of method tested and conflict situation used in the research, I calculated probabilities for each decision option from the relative frequency of singular forecasts for that option. The result of these calculations is shown in Exhibit 5.

Exhibit 5

Accuracy, and skill of multiple singular forecasts transformed into probabilistic forecasts for each combination of method and conflict situation *

Percent correct forecasts; *AESS*; (number of forecasts)

	Chance	Unaided judgment						Game theorist			Structured analogies			Simulated interaction		
		novices			experts			experts			experts			novices		
		%	*AESS*	n	%	*AESS*	n	%	*AESS*	n	%	*AESS*	n	%	*AESS*	n
Artists Protest	**17**	5	-0.14	(39)	10	-0.08	(20)	6	-0.13	(17)	27	0.13	(11)	29	0.14	(14)
Distribution Channel	**33**	5	-0.27	(42)	38	-0.18	(17)	23	-0.13	(13)	50	0.22	(12)	75	0.67	(12)
55% Pay Plan	**25**	27	0.02	(15)	18	-0.09	(11)	29	0.06	(17)	57	0.43	(14)	60	0.47	(10)
Nurses Dispute	**33**	68	0.52	(22)	73	0.60	(15)	50	0.25	(14)	57	0.36	(14)	82	0.73	(22)
Personal Grievance	**25**	44	0.26	(9)	31	0.08	(13)	43	0.24	(7)	36	0.14	(14)	60	0.47	(10)
Telco Takeover	**25**	10	-0.20	(10)	0	-0.33	(8)	0	-0.33	(7)	8	-0.22	(12)	40	0.20	(10)
Water Dispute	**33**	45	0.18	(11)	50	0.25	(8)	75	0.63	(8)	92	0.88	(12)	90	0.85	(10)
Zenith Investment	**33**	29	-0.07	(21)	36	0.04	(14)	22	-0.17	(18)	38	0.06	(8)	59	0.38	(17)
Average, unweighted	28	29	**0.04**	(169)	32	**0.04**	(106)	31	**0.05**	(101)	46	**0.25**	(97)	62	**0.49**	(105)

* Data from Green (2005), Green and Armstrong (2007a), and Green and Armstrong (2007b).

The percentage of correct forecasts from unaided judgment by novices and experts, and from game theory experts, was close to the average value for chance of 28%. Exhibit 5 shows that for those same methods, the average *AESS* figures for the probabilistic versions of the forecasts were close to zero. Recall that an *AESS* of zero indicates that the forecasting method has not

shown skill, which is consistent with the analysis by percentage of correct forecasts. Forecasts from structured analogies and simulated interaction were, on average, substantially more accurate than chance and the *AESS* figures of 0.25 and 0.48 are positive and well above zero, thereby providing a consistent assessment.

Moreover, Exhibit 5 illustrates that the *AESS* is a better measure of method or forecaster skill than percent correct. The percent correct figures convey no information about skill relative to the default method, in this case chance or random selection. The *AESS* figures, on the other hand, can be interpreted as the error reduction as a proportion of potential error reduction, or the extent to which the skill of the forecaster or method has reduced residual error. In mathematical terms,

$$AESS = \frac{\Pr(f^*) - 1/k}{1 - 1/k}$$

where Pr(f*) is the probability assigned to the actual outcome or the proportion of forecasts that were accurate, and k is the number of outcome options and $1/k$ is chance accuracy.

Careful readers will notice that this formula does not apply to the Distribution Channel conflict. That is because participants in the original research were offered four choices, one of which was "either A or B". In the papers reporting the result for this conflict, the findings were adjusted to account for the fact that there were in effect only three real choices and where respondents opted for the either A or B option, their response was coded as half right (0.5). I did not make equivalent adjustments to the AESS calculations here.

Assessment of probabilistic forecasts obtained in conflict forecasting research

Green and Armstrong (2007) described an experiment in which novices (university students) were asked either to pick the most likely decision, or to assign relative frequencies to each possible decision, in a set of from 3 to 6 possible decisions. The authors converted the relative frequency forecasts into singular forecasts—the most probable decision was used as the singular forecast—in order to compare accuracy with the singular forecasts. On that basis, there was on average no difference in accuracy between the two sets of forecasts, and little difference from chance accuracy (Exhibit 6).

I converted the singular and frequencies forecasts into probabilistic forecasts in order to assess whether the frequencies forecasts were the product of more skill than they appeared to be on the basis of percent correct forecasts. They were not. As Exhibit 6 shows, when converted to probabilistic forecasts, neither the singular nor frequencies forecasts exhibited any useful level of skill (mean *AESS* of 0.04 and 0.05 respectively) and the conclusion was the same as when they were both assessed as singular forecasts.

Exhibit 6

Accuracy and skill of novices' singular and relative frequency forecasts treated as probabilistic forecasts *

Percent correct forecasts; mean *AESS*; (number of forecasts)

	Chance	Singular			Frequencies		
	%	%	*AESS*	n	%	*AESS*	n
55% Pay Plan	25	9	-0.19	(11)	0	-0.22	(12)
Artists Protest	17	0	-0.20	(11)	10	-0.06	(10)
Personal Grievance	25	46	0.25	(13)	11	-0.08	(9)
Distribution Channel	33	38	-0.17	(13)	23	0.07	(13)
Zenith Investment	33	42	0.17	(12)	40	0.02	(10)
Telco Takeover	25	25	0.00	(12)	50	0.03	(12)
Nurses Dispute	33	58	0.38	(12)	64	0.28	(11)
Water Dispute	33	42	0.08	(12)	67	0.36	(12)
Averages (unweighted)	28	33	0.04	(96)	33	0.05	(89)

* Data from Green and Armstrong (2007a)

Aside from the study just described that directly compared singular and relative frequency forecasts, other conflict forecasting studies gave all participants the opportunity either to choose what they considered the most likely decision from a list, or to assign percentage likelihood figures. In practice, the overwhelming majority of respondents picked single decisions. For this paper, I converted the small number of percentage likelihood forecasts into probabilistic forecasts and calculated skill scores (Exhibit 8).

I also calculated skill scores from the conflict forecasting data in two other ways. The first of these was to set the option (decision) with the highest probability to one, and other options to zero. The second was to use the information that structured-analogies participants provided about their analogies to derive probabilities using the following rule. The probability of each decision is equal to the highest rating given to an analogy that suggests the decision, plus one-third of the sum of ratings given to any other analogies that suggest that decision, all divided by the sum of these aggregates across all decisions. Decisions not suggested by any of a participant's analogies are given a probability of zero.

For example, assume a fictitious participant used structured analogies to forecast a conflict with three decision options. Exhibit 7 shows the similarity ratings he provided (out of ten) for his three analogies. The ratings are shown in the columns corresponding to the decision options (A, B, or C) suggested by those analogies. The bottom line of the Exhibit shows the probabilities, derived using the rule just described, for the decision options.

Exhibit 7

Illustration of method for deriving probabilities from structured analogies data using a rule

	Decision options			
	A	B	C	Sum
Analogy 1 rating		7		
Analogy 2 rating			5	
Analogy 3 rating			3	
Probability of decision (derived using rule)	0 /13 = **0.00**	(7 + 0/3) /13 = **0.54**	(5 + 3/3) /13 = **0.46**	(7 + 0/3) + (5 + 3/3) = 13

Where participants considered that their analogies suggested more than one decision, the ratings were ascribed to each of the decisions that were suggested. In practice, there were never more than two decisions suggested by a single analogy and so, in such cases, the ratings were counted twice.

Exhibit 8

Skill scores for cases in which solo experts provided probabilistic forecasts, by derivation of probabilities [a] and forecasting method

	Absolute-Error Skill Scores (*AESS*)				
	One option set to 1.0 [b]		Participants' probabilities		Rule
	Unaided judgment	Structured analogies	Unaided judgment	Structured analogies	Structured analogies
Artists Protest	-0.20		-0.14		
	-0.20		-0.20		
	-0.20		-0.20		
Average	**-0.20**		**-0.18**		
Average 2					
Distribution Channel			*0.25*		
	-0.50	*1.00*	-0.05		*0.20*
	-0.50	*1.00*	-0.13		*0.40*
		0.25		0.14	-0.11
Average	**-0.50**	**0.75**	**-0.09**	**0.14**	**0.16**
Average 2		**0.25**		**0.14**	**-0.11**
55% Pay Plan		-0.33		-0.33	-0.33
	-0.33		-0.33		
Average	**-0.33**	**-0.33**	**-0.33**	**-0.33**	**-0.33**
Average 2		**-0.33**		**-0.33**	**-0.33**
Personal Grievance		-0.33		-0.33	-0.33
		1.00		0.73	0.20
	1.00		0.47		
Average	**1.00**	**0.34**	**0.47**	**0.20**	**-0.07**
Average 2		**0.34**		**0.20**	**-0.07**

Nurses Dispute	1.00		0.40		
		1.00	*0.25*		*1.00*
Average	**1.00**	**1.00**	**0.40**		**1.00**
Average 2					
Telco Takeover	-0.33	*-0.33*	-0.33		*-0.33*
	-0.33		-0.27		
		-0.33		-0.33	-0.33
	-0.33		0.20		
			0.20		*-0.33*
Average	**-0.33**	**-0.33**	**-0.13**		**-0.33**
Average 2		**-0.33**		**-0.33**	**-0.33**
Water Dispute	1.00		0.40		
	-0.50		-0.13		
			0.00		
Average	**0.25**		**0.14**		
Average 2					
Zenith Investment	-0.50	1.00	-0.05		0.55
		1.00		0.70	1.00
Average	**-0.50**	**1.00**	**-0.05**		**0.78**
Average 2		**1.00**		**0.70**	**1.00**
Total Average [c]	**0.05**	**0.40**	**0.06**		**0.20**
Conflicts, number	8	6	8		6
Observations, number	14	11	14		11
Total Average 2 [c]		**0.19**		**0.08**	**0.03**
Conflicts, number		5		5	5
Observations, number		6		6	6

Notes:

Figures in italics are excluded from the calculation of "Average 2" and "Total Average 2" in all cases, and from "Average" and "Total Average" in some. The criterion for exclusion was lack of a matching figure from the same method but an alternative derivation.

[a] For a single participant for a single conflict, the probabilities from which the *AESS*s were calculated were derived in three different ways: 1/ see note b; 2/ the participants' own probabilities were derived from their percentage likelihood figures; 3/ probabilities were derived from participants' analogy decisions and ratings using the rule described in Exhibit 7. In all cases, forecasts of C for Distribution Channel were re-coded, with 0.5 allocated to option-A and 0.5 to option-B.

[b] When a participant allocated the highest percentage likelihood for a conflict to a single option, that option was re-coded as one and the rest were recoded as zero. When participants' percentage likelihoods were inconsistent with their own analogies, any forecasts from those percentages were coded to the method “unaided judgment” and, where this was reasonable, single forecasts were derived from the analogies.

[c] Means of conflict averages.

Exhibit 8 includes only forecasts for which participants provided probabilities for several outcomes. Few such forecasts were provided. Consequently, interpretation of the findings shown in Exhibit 8 can only be tentative. As measured by the *AESS*, the assessment of the skill of the unaided judgement and structured analogies methods is consistent with comparisons of percent correct. That is, the structured analogies method was on average more skilful than unaided judgment.

It is not clear that probabilistic forecasts were more skilful than forecasts in which the participants' most-likely decision options were set to one and other options to zero. The probabilistic unaided-judgement forecasts showed more (but still negative) skill than did the one-option-set-to-one forecasts, but this was not the case for structured analogies forecasts. The rule for deriving probabilistic forecasts from participants' analogies data offered no increase in skill relative to first-choice-set-to-one or relative to participants' probabilities.

Summary and conclusions

The Absolute-Error Skill Score (*AESS*) provides a useful measure for assessing the skill of a forecaster or forecasting method. It is intuitive in that a score of 1 indicates perfect accuracy, a score of zero indicates no skill compared to the default method, and a negative score indicates the forecasting method is worse than the default method. The *AESS* is easier to interpret and better behaved than the Brier skill score or Squared Error Skill Score (*SESS*).

Using the *AESS* and the probabilistic data from my conflict forecasting research, I could find no evidence in that probabilistic forecasts tend to be more accurate than unequivocal forecasts. Nevertheless, decision makers may find a forecast that provides information on the relative likelihood of different outcomes to be more useful than an unequivocal forecast. I did not investigate that possibility.

My comparison of probabilistic and unequivocal forecasts was limited in the amount of data available and the type of forecasting problem that the data were drawn from. Further research on the relative usefulness of probabilistic and unequivocal forecasts is warranted.

Acknowledgements

Scott Armstrong, Don Esslemont, Baruch Fischhoff, Dilek Önkal, and Thomas Stewart provided helpful suggestions on early drafts of this paper. Responsibility for any errors or omissions is mine.

References

Armstrong, J. S. (2001). Evaluating forecasting methods. In Armstrong, J. S. (Ed.), *Principles of forecasting: a handbook for researchers and practitioners*. Norwell, MA: Kluwer Academic Publishers, 443-472.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*(1), 1-3.

Doggett, K. (1998). Glossary of verification terms (revised June, 1998). National Oceanic and Atmospheric Administration. Retrieved November 13, 2002, from http://www.sel.noaa.gov/forecast_verification/verif_glossary2.html.

Fuller, S. (2000). Verification: probability forecasts. *NWP Gazette, December 2000*. Retrieved November 10, 2002, from http://www.met-office.gov.uk/research/nwp/publications/nwp_gazette/dec00/verification.html.

Green, K. C. & Armstrong, J. S. (2007a). Value of expertise for forecasting decisions in conflicts. *Interfaces*, *37*, 287-299.

Green, K. C. & Armstrong, J. S. (2007b). Structured analogies for forecasting. *International Journal of Forecasting*, *23*, 365-376.

Green, K. C. (2005). Game theory, simulated interaction, and unaided judgment for forecasting decisions in conflicts. *International Journal of Forecasting*, *21*, 463-472

Lichtenstein, S., Fischhoff, B., & Phillips, L. (1982). Calibration of probabilities: the state of the art to 1980. In Kahneman, D., Slovic, P., & Tversky, A. (Eds.), *Judgement under uncertainty: heuristics and biases*. New York: Cambridge University Press, 306-334.