THE SEARCH FOR THE PERFECT HANDICAP

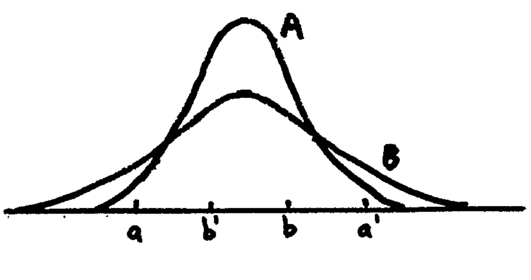
Francis Scheid

1. Introduction.

Suppose that two competitors A and B have symmetric score distributions. If B's mean score is greater than A's subtract the difference in means to obtain B's net scores. The result is suggested by Figure 1, in which B's net scores show greater variability. In a particular competition the actual net scores shot might be a and b. Assuming the objective is to score low, A wins this match. But for any such pair of scores there is a symmetric pair a',b' of equal probability for which B is the winner, so clearly each player ought to win just as often as the other. The competition is fair. Put in another way, the mean score is a perfectly accurate measure of ability in this situation. The same can, of course, be said of the median. With B's game having wider spread there is the somewhat unfortunate fact that A can play his very best and still lose, but in compensation he can also play his very worst and still win. Things are about as fair as a cruel world allows.

FIGURE 1

Symmetric distributions of net scores with equal means.



In most types of competition scores are not symmetrically distributed. In golf, for example, it is well known that poor scores range farther from the mean than good scores do, and more so for weak golfers than for strong ones. The question arises, what is then the most accurate measure of ability and do the mean and median still perform well? To find the answers to this and related questions in a golf setting the following experiments were carried out.

2. The types of measure tested.

Hole by hole scores of rounds shot by more than 1500 golfers at eighteen courses in various parts of the country form the data bank on which the experiments are based. For each player twenty rounds were used to compute a variety of ability measures. The present official procedure of the USGA was included. It begins by reducing hole scores if necessary to a maximum size dependent upon the ability level of the player. This process is called stroke control. It then subtracts from each score the rating of the course at which the score was shot, averages the ten lowest of these "differentials" and finally multiplies by 96 percent. The result is called the USGA handicap. The official British procedure was also included, in a way. This method depends somewhat upon the discretion of local handicap committees, but guidelines are provided which effectively make most players' handicaps their second best differential. These methods of measuring ability find broad use throughout the golfing world.

Over the years many alternative handicapping methods have been discussed by golfers. These are the basis for the variety of ability measures tested. Most prominent are the averages of certain score differentials such as the best five, the best ten, best fifteen, all twenty (the mean), the middle ten, and other sets obtained by trimming one or more of the best and worst, symmetrically or not. Also included were single differentials such as the very best, the second best (as an approximation to the British handicap), fourth best, sixth best and so on, and pairs such as the best and worst averaged, or the sixth and fifteenth averaged, or the tenth and eleventh (the median). Of somewhat different character are the point style handicaps which have been much discussed in golfing circles. In the 1248 system, for example, one point is given for a hole score of par plus one (bogey), two for par, four for par minus one (birdie), and eight for better (eagle). The average number of points per round is subtracted from 36 to obtain the handicap. A 1234 system is similar and probably clear without further description. In the 12 system one point is given for par and two for better, the average number of points per round being subtracted from 18 and the result limited to the interval 0 to 18. Different again are the selected hole types of handicap. For example, only the nine best holes of each round, relative to par, are selected and averaged, the total score for this nine then being doubled. Subtraction of course rating then produces a new differential for each round. The average of these differentials or of selected groups of them may be taken as ability measures. Similarly one may take the best twelve holes per round, or the best fifteen. A few measures of dispersion were included, such as the spread (worst score minus best) and the standard deviation. Even one or two measures of skewness were tried. For most of these types, all for which it was appropriate, a recomputation was made after stroke control, and counting these the list runs to more than a hundred. Some are clearly inferior but were left in to avoid prejudice, the experiment itself serving as impartial judge of merit.

Abbreviations for the various types will be convenient. For example, (2) is used for the second best differential, 6-15 for the average of the sixth to fifteenth best, B12 when the best twelve holes per round are selected, PT18 for the point system in which from one to eight points are given, STEF for the average of the sixth, tenth, eleventh, and fifteenth best differentials and 1,20 for the average of the best and worst. Other abbreviations may be understood from these examples.

3. Simulation of head-to-head play.

The most popular form of individual head-to-head competition is match play. Here each hole is a separate contest and the player winning the most holes wins the match. For even play the weaker player is usually allowed to reduce his scores for a number of holes (usually the more difficult ones) equal to the difference in handicaps. Because of its greater interest match play was the focus of the head-to-head part of this study. Stroke play, in which the total scores for rounds of eighteen holes are compared, is not so common when only two players are concerned but was included anyway partly because it was so easy to do and partly for comparison with the match play results. For both kinds of competition extensive play simulations were made.

At each of the courses for which data was in hand fifty pairs of golfers were chosen, using current USGA handicaps to assure a variety of ability differences within pairs. For each pair the stronger player A gave the weaker B an initial number of strokes also determined by the USGA handicaps. Matching each of A's rounds against each of B's in the 400 possible combinations, say at individual match play, the number of wins by A was noted. If not 200, a change in stroke allowance was made and the simulation repeated. This was done at least four times and until the 200 level was straddled, at which point a least-squares line was fitted to the four (wins by A, strokes given) pairs and an interpolation made to estimate the number of strokes actually needed to equalize the pair. For each club then the output of the simulation was of this sort, figures in column three corresponding to match play,

Club number	Players		Strokes needed
	A	B	
01	01	03	2.1
01	01	02	0.9
01	01	12	11.9
01	03	10	7.0

and the full list running to fifty pairs. At least 1600×50 or 80,000 matches were played at each club to produce these results. For stroke play output was in the same form, the strokes needed being 2.1, 0.9, 11.8 and 6.4, differing only slightly from the match play figures. Another 80,000 or more matches were played at each club to produce these results. A little arithmetic will show that for the two types of play combined almost three million golf matches were simulated.

4. Orthogonal polynomials and smoothing.

At this point the natural question asks how well each type of handicap might be used to predict the number of strokes needed to equalize pairs. To find the answer we consider strokes needed as a function of handicap difference, at both individual match and stroke play, and treat each type of handicap in its turn for each course in the collection. For smoothing purposes this function is represented as a combination of orthonormal polynomials on the discrete set of handicap differences, say

$$z(k) = \sum_{j=0}^{M} c(j)Q(j,k) .$$

The Q(j,k) are found as follows. Let x(k) be the handicap difference of a golfing pair and y(k) the corresponding strokes needed to equalize as found by the simulation. This discrete function is first shrunk by removing duplicate arguments x(k), letting w(k) stand for the number of duplicates and y(k) the average of the associated function values. Orthogonal polynomials P(j,k) are then determined by the recursion

$$P(j,k) = (x(k)-A(j))P(j-1,k) - B(j-1)P(j-2,k)$$

with P(0,k) = 1 and P(1,k) = x(k) - xav, where xav is the mean x(k), and

$$A(j) = \Sigma\, w(k)x(k)P(j-1,k)^2/\Sigma\, w(k)P(j-1,k)^2$$

$$B(j-1) = \Sigma\, w(k)P(j-1,k)^2/\Sigma\, w(k)P(j-2,k)^2 \ .$$

All sums are over the k which remain after the removal of duplicates. The orthonormal Q(j,k) are then

$$Q(j,k) = P(j,k)/\mathrm{sqrt}(\Sigma\, w(k)P(j,k)^2) \ .$$

Standard least-squares procedure then determines the coefficients

$$c(j) = \Sigma\, w(k)y(k)Q(j,k)$$

and the minimum error

$$\mathrm{rms} = \mathrm{sqrt}(\Sigma\, w(k)(z(k) - y(k))^2/50) \ .$$

One advantage of this procedure is that scanning the c(j) suggests where the representation might be truncated, that is, it allows one to choose the degree M of the smoothing polynomial. At course 1, for match play, using the USGA handicap, these coefficients prove to be

5.12, 34.02, -.16, -.06, .51, -.57, -.94,
-.18, 1.32, -.79, .72

and so on. After the first two these resemble the coefficient behavior of a random error function so a linear approximation seems appropriate. The rms error of this linear function was .68 and descended very slowly as higher degree terms were included, the value for a sixth degree approximation being .66. The standard analysis of variance test using the F distribution accepts the reduced (linear) model at the 99 percent level. This proved to be true for all but a few handicap types, for which quadratic approximation was required. For uniformity the rms error of the quadratic was adopted generally as the measure of accuracy of a handicap type.

5. The results.

The average errors for more than a hundred handicap types at each of eighteen golf courses form a fairly large matrix. Only a selected fragment is presented here, as the MATCH PLAY ERRORS table, in which each column represents a handicap type. There is, of course, a similar table for stroke play. Even a casual inspection shows differences between columns, and a Friedman test indicates that the differences are significant at a level far below one percent. The same was true at stroke play.

To compare the accuracy of the various types the median error of each column of the full matrix was found. Some of these are presented in the MEDIAN ERRORS, MATCH PLAY table. The same procedure at stroke play produced the MEDIAN ERRORS, STROKE PLAY table. All entries are for methods without stroke control, since it was soon apparent that whatever the merits of this process may be it almost always reduces accuracy by a small amount. Several things should be noticed. First of all the same types dominate both lists; ability measures that are accurate for match play also seem to be accurate for stroke play. Application of the Wilcoxon signed-rank test showed differences of .1 to be significant at the .05 percent level more often than not, while differences of .2 or more were almost always significant. Second, conspicuous among the top performers in each table are averages of differentials from which some of the best and worst are omitted, and it is interesting to see that the overall average is outdone by several of these. In other words, a certain amount of trimming at the ends of the score distribution improves the accuracy. Too much trimming, however, reduces it again. The median score, for example, does rather badly. Apparently one can afford to discard only so much of the information available. Numerous other trimmed averages of the same sort are not listed, often because they were not included in the experiment. Their median scores may be estimated to some extent by reading between the lines.

Finally note the types of ability measure that performed relatively poorly. The USGA handicap (with stroke control in this one case) had error .83 at match play and .80 at stroke play even when used optimally, which meant inflation by 108/96 at the former and slightly more at the latter. When used in the officially recommended way the errors were 1.1 and 1.2. The British handicap, as approximated, had errors of size .90 and .95 when used optimally. The most accurate single differential proved to be the eighth best. All the point and selected hole types were inferior. Measures of dispersion and skewness were entirely undependable for predicting ability, as anticipated, having errors of three to five strokes.

6. Events with many competitors.

Arranging fair play when there are many competitors is a somewhat different problem. Some insight into it can be gained from a simple example. Suppose that just three players compete and the best net score wins. A is perfectly consistent, always shoots 75, while B and C manage rectangular distributions between 70 and 80. For A to win it is necessary that both B and C play in the poorer halves of their games, and this happens with probability 1/4. B and C share the remaining wins equally, giving probability 3/8 to each. So A is at a slight disadvantage because of his consistent game, unless he is given strokes. For fair play A ought to win 1/3 of the time. To arrange this we could reduce his scores until the probability p that he outperforms B, and consequently the probability p^2 that he outperforms both B and C, is 1/3. In the same way if there are N competitors in all, A's fair share is 1/N and what is needed is

$$p^{N-1} = 1/N.$$

As N increases the limit of p is 1, which means that A's net score must equal the best of the other competitors.

This same conclusion can be reached by an intuitive argument. When there are many competitors someone is almost certain to be on his best game, or near it. In a field of one hundred players, for

Handicap types

	USGA*	(2)	MEAN	MED	6-15	2-19	2-17	(8)	B12	B15	PT18	PT14	PT12
Courses	.68	.67	.59	.45	.49	.50	.35	.27	.76	.68	.48	.41	1.02
	1.11	.96	.96	.98	.60	.89	.58	.62	1.44	1.02	1.40	1.28	2.21
	.50	.43	.24	.47	.39	.19	.20	.52	.40	.39	.35	.27	.52
	.25	.29	.15	.29	.25	.13	.16	.31	.50	.24	.35	.17	.20
	.74	.94	.47	.81	.36	.36	.33	.68	.98	.71	1.22	1.12	2.83
	.90	.84	.37	.59	.43	.32	.32	.36	.72	.51	.44	.32	1.02
	1.13	1.27	.71	.77	.51	.58	.57	.50	.78	.58	.83	.76	1.56
	.64	.66	.19	.49	.30	.21	.30	.26	.32	.24	.42	.36	.84
	.70	.77	.40	.62	.53	.33	.33	.38	.77	.47	1.44	1.08	2.72
	.89	1.23	.42	.50	.63	.49	.28	.55	.37	.44	.34	.35	.46
	1.04	1.40	.46	.75	.47	.56	.37	.54	.85	.53	.88	.70	1.70
	1.26	1.36	.49	.31	.30	.37	.56	.53	.64	.66	.53	.51	1.39
	.81	.72	.40	.50	.48	.42	.50	.49	.62	.44	.75	.62	2.09
	.86	1.62	.46	.64	.47	.36	.50	.58	1.21	.71	1.05	.66	1.52
	.72	2.00	.52	.49	.49	.41	.40	.97	.74	.55	1.44	1.48	3.56
	.55	.78	.42	.52	.41	.39	.52	.46	.72	.56	1.50	1.54	2.47
	.86	.99	.38	.36	.30	.27	.60	.82	1.15	.84	1.67	1.29	2.62

MATCH PLAY ERRORS

Selected handicap types

MEDIAN ERRORS, MATCH PLAY

A. The most accurate types tested.

Error	Types
.4	2-17 2-19 4-17 3-18 2-15 5-16 MEAN 4-15 8,13
.5	3-15 1-15 6-15 STEF 3-14 4-14 2-14 (12) MED 4-13 (8) 6,15 (10) B15 2-13

B. Certain other types, selected.

Error	Types	Error	Types
.7	PT14 B12	1.0	(2) 1-5
.8	USGA	1.6	PT12
.9	PT18 B9 2,19		

MEDIAN ERRORS, STROKE PLAY

A. The most accurate types tested.

Error	Types
.2	3-18 4-17 2-19 2-17 5-16 4-15
.3	3-15 6-15 2-15 4-14 3-14 MEAN 1-15 STEF
.4	8,13 2-14 4-13 6,15 3-13 2-13 (8) 4-12

B. Certain other types, selected.

Error	Types	Error	Types
.5	MED	1.0	B9
.6	B15	1.2	1-5
.7	PT14	1.7	PT12
.8	PT18 B12 USGA	4.4	SPRD
.9	(2) 2,19		

instance, it can be argued roughly that only one will probably play in the top hundredth of his score distribution, and doing so ought to be the winner. This suggests that fair play calls for equalizing the percentile 1 scores of the competitors, which is not the same as equalizing means, the weaker player B's scores being shifted only by the amount shown in Figure 2.

FIGURE 2

Equalizing the percentile 1 scores

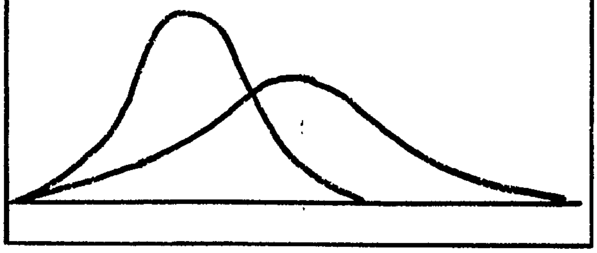
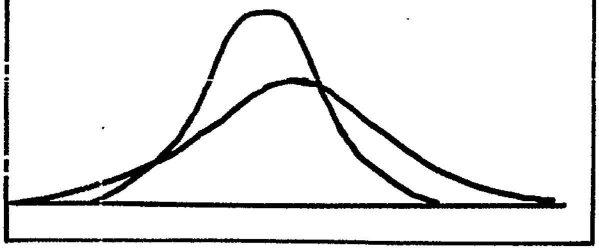


FIGURE 3

Equalizing the percentile 25 scores.



The above argument applies if equal chance to finish in first place is the criterion of fairness. In large field events, however, prizes of some sort will usually be given to about a quarter of the field and if Figure 2 prevails most of these would go to the stronger players. In fact the bottom of the order of finish would be heavily populated with high-handicappers. For equal chance to finish in the top quarter B's scores should be shifted until his percentile 25 score coincides with those of the other competitors, as shown in Figure 3. Both criteria of fairness will now be explored.

7. The simulation.

In order to determine the percentile scores just discussed the twenty rounds available for each player were taken as a base and eighty more simulated using random numbers to select hole-by-hole scores from this base. After sorting the hundred total scores so obtained the lowest serves as the percentile 1 score and other percentiles are easily identified. This procedure was followed for each player in the data bank. The output of the simulation thus consisted of a list of players identified by club and number together with percentile 1 and percentile 25 scores, relative to course rating. The list began as follows.

Club	Player	Pctile 1	Pctile 25
1	1	-1.0	5.0
1	2	2.0	6.0
1	3	2.0	7.0
1	4	4.0	7.0

8. Analysis and results.

And how accurate are our handicap types at predicting success in many-competitor events? To answer this question the idea of equalizing scores at some percentile level can be implemented. This might be done by bringing everyone to a specified score, say the appropriate percentile score of an average scratch player. Strokes needed would then be P(i) - S, where P(i) is player i's percentile score and S is scratch. The function (strokes needed, h'cap) is then smoothed by the same orthogonal polynomial routine used before, with the individual golfers forming the experimental unit at each course instead of the fifty selected pairs. The rms errors again serve as measures of accuracy. In carrying out this process the subtraction of S can be omitted, making it unnecessary to estimate S, since smoothing of the function (P(i),h'cap) will lead to exactly the same rms errors, the new smoothing function having values just S units greater. From another point of view what this means is that if the latter function requires very little smoothing for some handicap type then that type will be a good predictor of strokes needed.

Applying this procedure first to the percentile 1 data the usual large matrix of rms errors was produced, but will again be omitted. Certain medians are displayed as before in the MEDIAN ERRORS, PERCENTILE 1 table. The new abbreviation NORM appears, representing the measure MEAN - 2.326SIG which is the percentile 1 level of a normal model of the player's score distribution. The first thing to notice here is the larger size of the errors compared with those reported for head to head play. This is probably due to the fact that the tails of a distribution are harder to define with precision, and it suggests that it may be harder to arrange for fair play by this criterion. Even so it is reassuring to note that prominent among the most accurate types are those which rely more heavily upon the better part of a player's game. The normal model also does about as well as any, but it is the one best of the basic twenty differentials that leads the list. The Wilcoxon signed rank test found a difference of .1 to be undependable, but a difference of .2 was quite trustworthy at the .05 percent level of significance. Among other things this means that the official USGA and British handicaps take a close second place. Of the averages after trimming which dominated the lists for head-to-head play several can be found along with the MEAN only slightly lower down.

Approximation of the percentile 25 data led to the comparable table MEDIAN ERRORS, PERCENTILE 25. The handicap type NORM which took first place here represents MEAN - .674SIG which is the percentile 25 level of the normal model. Behind it is a tight pack of trimmed averages and miscellania. The USGA handicap was expected to perform well here, since it is based upon the average better half of the

MEDIAN ERRORS, PERCENTILE 1

A. The most accurate types tested.

Error	Types
1.2	(1) 1-5
1.3	NORM
1.4	(2) 2,19 USGA 1-15

B. Certain other types, selected.

Error	Types
1.5	PT18 B12 MEAN 2-17 B15 PT14 6,15 1-15 2-15 2-19 3-18 1,20 4-17 B9 STEF 2-14 4-15
1.6	5-16 PT12
1.7	8,13 MED

MEDIAN ERRORS, PERCENTILE 25

A. The most accurate types tested.

Error	Types
.2	NORM
.4	2-19 MEAN 2-16 2-15 1-15 2-14 3-18
.5	2-10 to 2-13 3-10 to 3-15 4-10 to 4-17 B15 6,15 STEF 8,13 5-16 6-15 (6) (8) PT14

B. Certain other types, selected.

Error	Types
.6	MED USGA PY18 B12
.8	2,19
.9	(2) B9 1-5
1.2	(1) PT12

player's game, but finished slightly behind the pack. The British handicap and the one best differential were definite also-rans, while PT12 again brought up the rear. As before differences in medians of .2 are statistically significant at the .05 level.

9. Summary.

The experiments just described may be summarized by saying that for both head-to-head and large field competition a moderately heroic effort was first made to measure individual playing ability. This consisted for the former of simulating paired matches in great number, and for the latter simulating many rounds in addition to those available from scorecards. Hopefully those efforts brought their reward in the form of accurate ability measurement. But an operating handicap system must deal with millions of players, provide frequent updates and still be economically feasible, so it has to be based upon a more modest procedure. In the second part of the experiments the question was, which among numerous simpler procedures best approximates the heroic. No one method proved to be best for all cases, a result which was surely predictable. However, a wide variety of averages does establish a clear superiority, both at head-to-head play and when errors for the four types of competition are combined. This variety includes the overall mean, the others involving some trimming of extreme scores symmetrically or not, such as 1-15, 2-19, 3-18, 4-17, 5-16, 6-15, 2-14, 3-14 just to name a few. For predicting winners in large field competition these averages are outdone by certain measures which emphasize the best part of a player's game, such as the best of twenty rounds, or the second best or even the USGA handicap, but these prove to be inferior when measured by any other criteria. The various point systems which have been proposed, and the selected hole methods also prove to be disappointing.

Work has been begun on a follow-up experiment focusing upon events in which the best ball, hole-by-hole, of a team of two or more players is active. This will without doubt make things more complicated still. It is clear that the perfect handicap does not exist, but the search for the nearest thing goes on.

References

1. "The Search for the Perfect Swing"; Cochran and Stobbs; J. B. Lippincott Co., 1968.

2. "You're not getting enough strokes"; Scheid; Golf Digest, June 1971.

3. "A Least-squares Family of Cubic Curves with Application to Golf Handicapping"; Scheid; SIAM Journal on Applied Mathematics, 1972.

4. "A Basis for Golf Handicapping"; Scheid; presented at Austin meeting of SIAM, 1972.

5. "Does your handicap hold up on tougher courses?"; Scheid; Golf Digest, October, 1973.

6. "Computer-assisted Handicap Survey"; Bogevold; USGA report, 1974.

7. "A non-linear feature of golf course rating and handicapping"; Scheid; presented at TIMS 22nd international meeting, Kyoto, 1975.

8. "A Model of the USGA Handicap System"; Pollock; in Optimal Strategies in Sports, edited by Machol and Ladany; North Holland Publ. Co., 1977.

9. "An Evaluation of the Handicap System of the USGA"; Scheid; in Optimal Strategies in Sports.

10. "A Study of Selected Ball Play"; Scheid; report to USGA annual meeting, Atlanta, 1977.

11. "Fair Handicap Percentages for Individual Competition"; Scheid; report to USGA annual meeting, San Francisco, 1978.