Comparing Hand Evaluation Methods

Table of Contents

Introduction

This study attempts to compare several of the more widely known hand evaluation methods, with an emphasis on determining which of them, if any, are the most accurate when it comes to bidding making games, and staying out of games that do not make.

Background and Terms

The purpose of hand evaluation in Contract Bridge, is to estimate, using some metric(s), the number of tricks a pair of hands working together can take. Ever since Goren popularized the Work Point Count method this estimate has been arrived at by giving each hand a number of 'points'. Some of the features of a hand that earn positive 'points' are: high cards, length, shortness, fit with partner and so forth. A hand with negative features can also earn negative points for characteristics such as: a misfit with partner, minor honors in short suits, square shape and so on.
For the purposes of this study we shall call the set of factors used by the various methods a 'metric'. Many Bridge experts, other than Goren, have devised their own 'metric'; hence we have hand evaluations using Marty Bergen's metric, the Four Cs metric, and Richard Pavlicek's metric to name but a few.
We also refer to a pair of hands in a partnership as a 'side'; thus we have the North-South side, and the East-West side. It should be apparent that what matters most is the strength of the two hands of a side working together, not the strength of each hand of a side in isolation.

While most of the metrics that have been devised share many attributes in common, there are also many differences among them. Even the values assigned to the Honor cards, Aces, Kings, Queens, Jacks and Tens, the so called High Card Points (HCP) can differ from metric to metric. And the HCP are the factor that varies the least among all the factors that go into making up the various metrics.

The Meaning of Success

The most profitable payoff in Bridge comes from bidding and making games. The bonus on the scorecard for bidding and making a game is much more than the penalty incurred from bidding a game and failing in your contract. Games are also much more common than slams. So the focus of hand evaluation should be to allow a side to bid all the games that are worth bidding, and stay out of the games that are not worth bidding. If a hand evaluation metric could do that, it would be enormously successful. Of course such perfection is impossible in practice.

Jan Eric Larsson, in his book,"Good Better Best", demonstrates that, given the large amount of empirical data available, it is highly unlikely that your choice of bidding system, whether it be a 1940's style 4 card Major system, or the latest Big Club system, will alter your results by much more than 10%. His simulations also show that your choice of hand evaluation method, can have a bigger impact on your results than your choice of bidding system!
So it seems worthwhile to try to study which metric, if any, can get us that extra 10%.

Earlier we stated that the hand evaluation method should allow us to bid all the games that are worth bidding. What does that mean?

It does not mean that we never want to go down; it also does not mean that we want to bid game on every hand regardless of our overall strength.

It means that we know we can never be certain that we will always be successful in our contracts; we might be defeated by a very unlikely lie of the cards; a 5-0 trump split, with a key finesse losing say (a 2% chance). But if we know we have a 98% chance of success we will always bid the game and we will not deem our evaluation method a failure because in this instance we were defeated.
In fact the scoring table is structured so that if we are Vulnerable, we will break even in the long run even if only 37.5% of our game bids make. If we are Not Vulnerable we need a 45% chance of success to break even in the long run. This assumes of course that we are allowed to play the game undoubled. Jeff Reubens in his classic, Secrets Of Winning Bridge, says that we would like our Vulnerable games to have a 40% chance or better, and our NonVulnerable games to have a 50% chance, in order to turn a profit.

So that is our measure of success; we will judge the metric by how often it fails to get us to a making game, as well has how often it gets us to a game that fails. How many IMPs in the long run will we lose by following its advice?

Methodology

Population of Hands Studied

In this study the focus will be on bidding games. If we have enough honor cards then every metric under consideration will get us to the game, and the game will make most of the time on sheer power. We will not gain much insight if we limit our study to contracts of 3NT with 28+ HCP between the two hands. No, what we would like is to find a metric that will identify those factors thtat will get us to making games, when we have only 22, 23, 24 or perhaps 25 HCP. And these factors will be much more likely to be relevant in suit contracts.
Accordingly in this study, we will be studying hands that might be able to make 10+ tricks, with spades as trump. These hands will have either 8, 9, or 10 spades between them, and will have between 22 and 25 HCP. If a given metric can consistently get us to making games with only 23 HCP then that is something worth having.
The reason for choosing this population of hands is that it represents the most likely scenarios in practice. If we do not have at least 8 trump between us, then we will most likely not be playing our games in a suit contract. (We are not worried about part scores, since all partscores, are roughly equal in value when playing IMPs.)
Also hands with more than 10 trump between them are relatively rare; and if we have a fit of as much as 11 cards, it means the opponents are shapely also, and it could be anybody's hand.

Hands with less than 22 HCP between them will almost never make game with only an 8 fit, and even with a 10 fit the odds of success are too low to form a basis for a hand evaluation method. In contrast hands with 'only' an 8 fit, but with 26 or more HCP will make game almost always regardless of which hand evaluation method one uses. Just using HCP with no other factors considered would tell us that the game is worth bidding. Therefore the population of interest is the one just described.

Evaluating the Hands

The process of evaluating the hands is as follows:

1. Choose a the number of spades "F" for this batch where F is one of 8, 9 , or 10.
2. Choose a value of HCP "P" for this batch., where "P" is one of 22,23,24, or 25.
3. Deal 500 hands where NS have a total of exactly "F" spades and "P" HCP between them.
4. Evaluate each of these NS hands using each of the metrics of interest, and append each value thus obtained to the record for the side.
5. Choose the first of these 500 hands. Keeping the NS hands fixed, deal a hand at random for EW.
6. Solve the resulting deal using a Double Dummy Solver to determine the number of tricks NS can take with spades as trump. Append this value to the record for the NS hands. At this point we have one example of a solution for NS hand #1, along with how many 'points' it has according to each metric.
7. Repeat steps 5 and 6 ninety-nine more times so that the single NS hand has been 'played' a total of 100 times against random EW hands.
8. Repeat steps 5, 6 and 7 for each of the 500 NS hands in turn.

At the end of step 8 we will have 50,000 datapoints for hands with "F" spades and "P" HCP. We can then count the number of times that game makes, or does not. And we know how each metric values these NS hands.
It only remains to decide if the metric would have bid game or not on the hand in question.
In case of an 8 fit, the criteria was that the hands needed 26 'points' (typically a total of HCP, support points, length points, and points assigned for other factors) to bid game with an 8 fit. This corresponds to Pavlicek's criteria, and also to Jeff Reuben's advice.
In the case of a 9 fit the 'floor' was set at 25 points, and in the case of a 10 fit, the floor was set at 24 points. This represents what I believe would be the most common approach taken by most bridge players.
For example the NS hands according to the Bergen metric might evaluate to 26 points, and he would therefore bid game. Pavlicek on the other hand might evaluate the hands as being worth only 25 points and he would not bid game. If game would make (double dummy) then Bergen was right; if game would fail then Pavlicek was right. The question then becomes, for the 100 times that that particular hand was played, who was right most often? And how many IMPs would be lost, either by leaving IMPs on the table for not bidding game, or by going minus when a plus (part) score was available?

A total of 24 sets of hands were analyzed in this way; each set generating 50,000 datapoints.
12 of the sets fixed the 'HCP' strength of the hand using a simple 4-3-2-1 Work Point count. The other 12 sets used the Kaplan and Reubens four C's method to fix the strength of the hands.

The average number of IMPs per board that would be lost by each method was then totalled up for each metric for each set. The best metric is the one that lost the fewest number of IMPs, not necessarily the metric that had the most success bidding and making games, since it also valuable to stay out of games that don't make.

Tools

Since Bridge is a game of probabilities we must use some form of statistical analysis to arrive at any conclusions. And since there is not nearly enough time to play several thousand hands manually we must resort to computers. In the 1950s M. RenéVernes, of France, derived his Law Of Total Tricks

by studying no more than 2000 hands, all played at a high level by experts. But today we have cheap powerful computers available to take much of the drudgery out of the process. However to use computers, requires that the success or failure of the contract be determined by using a Double Dummy Solver (DDS). The validity of this approach, in a statistical sense, has been confirmed, by many authors of many other studies. For a discussion of some of the issues behind the use of such a tool, see Richard Pavlicek's excellent website at www.rpbridge.net

The main tool used in this study was the program "dealerv2". This program is based on the original "dealer" program written by Hans van Staveren in the 1990's, but heavily modified and expanded by the author (JGM) to make it much easier to use in simulation scenarios. Dealerv2 can be automated with a script to generate the deals as required with the number of spades and HCP desired, and export them for later use in the evaluation and DD solving phase. This second phase can also be automated with a script to access each of the 500 deals generated in phase 1, calculate all the metrics of interest for the hand in question, generate 100 EW hands and solve each of the resulting complete deals Double Dummy, and then write the combined results to a CSV file for later analysis by a spreadsheet or database program.

The second tool is a spreadsheet program (Libre Office Calc, but Excel would also work) to total the number of successes and failures, calculate the average IMPs lost per board, and total the results.

All of this software and raw data is available on Github.

Factors Considered by the Various Metrics

The following factors are the ones most commonly used to evaluate the trick taking potential of a hand. However even though many of the factors are the same from metric to metric, there are several variations of each of them that have all been used at one time or another. Outside of the Optimal Point Count (OPC) metric no one metric uses all of these factors.

1. Value of Honor Cards. The most common variation is Ace=4, King=3, Queen=2, Jack=1. The variations used in this study include the 'BumWrap' scale which T. Andrew's research suggests is better for suit contracts where : Ace=4.5, King=3, Queen=1.5, Jack=0.75, T=0.25 Danny Kleinman has described 'Little Jack Points' in his book, 'The NoTrump Zone' where Ace=13, King=9 Queen=5, Jack=2. and Tens get some points depending on whether they are isolated or have 'companions'. The resulting total is divided by 3, to bring the scale back to the recognized 0-40 points. There are also the values assigned by the four C's algorithm.
2. Value of Singleton Honors and Honors in two card suits. Each authority has its own recommendation as to how to evaluate singleton honors. The usual scale is to subtract one point from its normal value for a singleton K, Q, or J. There is no deduction for a singleton Ace. There is even more variation when adjusting for honors in two card suits. Combinations with the Ace (AK, AQ, AJ) are seldom discounted. Kx and KJ usually keep their full value. KQ is often discounted by 1 point, as are QJ, Qx and Jx.
3. Value of Shortness. Some authorities recommend counting for shortness when deciding whether or not to open the bidding. Goren's method for example counts for shortness. Other authorities who follow that approach are Sheinwold, and Pavlicek. The usual scale is 3 points for a Void, 2 points for a Singleton, and 1 point for a Doubleton. Variations are whether to discount the first doubleton, and whether to count both shortness points and high card points for short honors. Sheinwold would count a stiff King as being worth 4 points; 3

minus 1 for the honor plus 2 points for the shortness. Pavlicek's advice is to count the higher of the honor value or shortness value, but not both. A stiff King is thus worth 3 (for the honor) and a stiff Jack is worth 2 (for the shortness). Qx and Jx are counted as 2 and 1 respectively.

4. Value of Length. Most often a 5 card suit counts as an extra point for length. Some metrics (OPC among them) want at least a Queen or better in a 5 card suit in order to count for length. There is some variation on how to value 6 card or longer suits. Many award 2 points for each card over 5, so that a 7 card suit gets 5 points for length alone. Others only award one point for every card over 4 so that a 7 card suit only gets 3 points for length.
5. Support Points. The most common measure is with 4 card trump support to count 5 points for a Void, 3 for a Stiff, and 1 for a Doubleton. With 3 card trump support the most common scale is 3 points for a Void, 2 points for a Stiff and 1 point for a Doubleton. Some metrics insist on a '9 card' fit before awarding support points; it is not clear whether they really mean 4 trump or whether a 6-3 fit would qualify. Other metrics count 3/2/1 for an 8 fit and 5/3/1 for a 9 fit; again it is not clear whether they really mean 4 card support (so that a 4-4 fit would use the 5/3/1 scale, and a 6-3 fit would use the 3/2/1 scale) or whether they really mean 9 trump and 8 trump no matter how distributed. Other questions are whether each short suit gets counted (so that two doubletons are worth 2 points even with only 3 trump), and a stiff and a doubleton would be worth 3 points (again even with only 3 trump), or whether only the shortest short suit should receive support points. There are many other variations also, such as how to count when one has 5 card trump support.
6. Synergy Points. Bergen refers to these points as points for a 'Quality suit'. The usual case is to give 1 synergy point for a 5 card suit with 3 of the top 5 honors, and 2 synergy points for a 6 card or longer suits, with 3 of the top 5. Variations include awarding an extra point if the suit contains 4 of the top 5. Kleinman uses a different form of synergy points, giving an extra point to a King if it is with the Ace, and an extra point to any Queen or Jack in a suit that includes either the Ace or the King. Larsson also has a different definition of synergy points. The Four C's metric has a very detailed and precise way of evaluting honors in long suits.
7. Points for Fit. Some metrics specifically award extra points for a fit with partner. This is usually expressed as an upgrade to Declarer's hand once a fit is found. The most common scale is to award one point for an 8 fit, two for a 9 fit, and 3 for a 10 card or longer fit.
8. Points for Minor Honors in Partner's suit. These points are usually given only to Dummy but some metrics count these 'Hf' points for either hand. The usual scale is to award an extra point for Qxx(x), Jxx(x) , or QJx(x) in partner's 5+ suit. Some metrics limit this to only the trump suit. Jeff Reubens refers to this as 'In/Out' valuation, where we upgrade minor honors In our fit suits, and downgrade them when they are Outside our fit suits.
9. Misfits. Most metrics give lip service to the concept of misfits but few quantize this with numbers. Misfits are of two kinds, Wasted Values (Honor cards) facing shortness (so that KQx facing a singleton is not worth 5 points, particularly in a suit contract) and not counting for shortness in partner's long suits, so that a small stiff, x, is not worth 2 points if partner bids the suit naturally (presumably showing 4 cards or more).
10. Miscellaneous Factors. These include such things as deducting for 4333 shape, adding for 'body' when the hand is rich in Ten's and Nine's, and deducting for Mirror shape.

Metric Names Used in the Code

BERG is Bergen's method from information collected from his pamphlet,"Points Schmoints" and other internet sources.Bergen's writings and books on hand evaluation have received widespread attention and are quite commonly used. This study has tried to stick as closely as possible what is described in the book, "Points Schmoints" and other internet sources. Bergen counts for length when opening the bidding, has an 'Adjust-3' method to emphasize Aces and Tens and de-emphasize Queens and Jacks, adds points for 'quality suits' which have 3 or 4 of the top 5 honors, and uses the most common method of counting support points, and length points. He also deducts for misfits, but does not give much detail.

KAR is the Karpin method as described on R. Pavlicek's website. Points for a long suit. Very conservative support points: 3-2-1 with 4 Trumps.
JGM is the same as KAR but using the BumWrap HCP scale, where Ace=4.5, K=3, Q=1.5, J=0.75 and Ten=0.25.

LAR is the method as described by J.E. Larsson in his book, "Good Better Best". Points for long suits; unique synergy points for two suits with 2+ of the top 3 honors. Also conservative support points: The book says none with an 8 Fit and to use 5-3-1 with a 9 fit. JGM modified this to only do support points with a 9 fit but 5-3-1 with 4 trumps (5-4 say) and 3-2-1 with 3 trumps (6-3 say).
MOR is the above LAR but using Modified BumWrap points: Ace=4.5, K=3, Q=1.5, J=0.75, T=0.25 But a Queen with an A or K in a 3+ suit gets an extra 0.5 making AQ worth 6.5 and KQ worth 5. A Jack with a Q or T in a 3+ suit, gets an extra 0.25 making it worth 1.0 and a Ten with a Jack or 9 in a 3+ suit gets an extra 0.25 making it worth 0.5.

PAV is as described on RP's website; Goren (points for short suits) with some minor tweaks by RP. It is the only entry in the above list that uses short suit points before a fit is found. For short honors, it counts the greater of the honor value or the shortness value, but not both.

DKP is based on Danny Kleinman's 'Little Jack Points' as described in his book, "The Notrump Zone". For the purposes of suit contracts JGM has added 1 point for a 5+ suit, and the common scale for support points with 4 trump (5-3-1) and 3 Trump (3-2-1), since DK did not specify what he uses when in a suit contract.

HCP uses the 4-3-2-1 count only. No points for length, shortness, or support. It is included as a Lowest Common Denominator for comparison purposes.

CCCC is the Kaplan and Reubens 4 C's method as written in the Bridge World article. Some bug fixes by JGM to the code published by Jeff Goldsmith.

KnR implements CCCC as above but then adds the points for Fit and Support as described in the original Bridge World article.

OPC is the Optimal Point Count as defined by P. Darricades in his books on the subject. OPC takes into account just about every factor one can think of that might affect hand evaluation. For example it counts for both length and shortness before a fit is found, it deducts for misfits, wastage, mirrored

hands, square hands, honors in short suits. Queens and Jacks have a variable value depending on whether they are un-accompanied or in a suit with other honors. It adds points for fits, for minor honors in any and all of partner's 4+ suits, for support, and for synergy when there are 3 of the top 4 honors in a single suit.

Results

The results are divided into two groups: those based on hands with a given number of HCP and those based on hands with a given number of Four C's points.
The HCP point count range of interest, as described earlier, is 22-25 since outside of that range there is seldom a difference of opinion among the various metrics as to whether to bid game or not.

Avg Vul IMPs Lost per board	HCP	CCCC	BERG	DKP	JGM	KAR	KNR	LAR	MOR	PAV	OPC
CCCC Base F10 Total	4.442	2.054	**1.739**	1.931	**1.740**	1.761	1.757	2.188	2.149	1.907	1.793
CCCC Base F9 Total	4.400	3.021	3.024	2.833	**2.728**	2.998	2.948	3.222	3.069	2.936	2.868
CCCC Base F8 Total	2.591	2.575	3.605	2.696	2.484	2.528	2.973	2.421	**2.389**	2.628	3.053
CCCC Base Grand Total	3.811	2.550	2.789	2.487	**2.317**	2.429	2.559	2.610	2.536	2.490	2.572
HCP Base F10 Total	5.274	1.158	1.202	1.177	1.185	1.202	**1.128**	1.206	1.178	1.168	1.218
HCP Base F9 Total	5.306	1.738	1.470	1.489	**1.393**	1.460	1.477	1.773	1.609	1.491	1.539
HCP Base F8 Total	5.248	2.349	2.299	2.303	2.245	2.305	2.268	3.020	2.706	**2.220**	2.443
HCP Base Grand Total	5.276	1.748	1.657	1.657	**1.608**	1.656	1.624	2.000	1.831	1.626	1.733

See the Appendix for a more detailed breakdown.

Conclusions

1. Relying on HCP alone is a losing proposition. One must take into account at least some other factors.
2. Simpler is better. The two metrics that score the best, the most consistently, are those of Pavlicek, and Karpin. One using short suit evaluation, and one using long suit evaluation. Larsson is also consistently good. We include CCCC and its more complicated sibling KnR for reference purposes since these are considered the most complete and accurate evaluation methods. They are also both too involved to be used at the table, so that they are not scored as 'best'. Even though KnR includes taking into account shortness, support points and fit, where CCCC does not, it does not gain much, and sometimes not at all, over the 'simpler' CCCC method.
3. T. Andrews' research that suggested that using the 'BumWrap' HCP scale would work better for suit contracts is validated by comparing JGM to KAR, and MOR to LAR. But the gains are again quite modest.

One recommendation that seems to be born out by the above results, is that there is a limit to how much you can count for a single suit. For example Larsson's approach to 'Synergy' requiring two suits with 2+ honors, seems to give better results than the Bergen 'Quality Suit' approach.
The reasons for the above results, and conclusions, are at this point a matter of conjecture. Perhaps worthy of further simulations trying to isolate factors that really matter from those that do not.
It might be interesting to add Larsson synergy to the Karpin and Pavlicek methods and see whether this gains anything, for example.

Appendix -- Results and Data Details