

MEASURING THE POWER OF A PLAYER

**Francis Mechner
The Mechner Foundation**

INTRODUCTION

Ways of Describing Strength

The strength of chess or go players is usually described by numerical rating scales. Although the numbers used in the existing and widely used rating systems are arbitrary, they make strength comparisons possible.

In chess Elo rating system, for example, most beginners have a rating slightly below 1,000, and the world champion usually has a rating of somewhat over 2,800. In go, the rating scale used by the American Go Association for amateurs ranges from around -350 (35kyu) for rank beginners to about +800 (8 dan) for the strongest amateurs, and in Japan there is another rating scale from 1 to 9 for the professionals.

Current rating systems periodically readjust an active player's rating by the use of mathematical formulas that take into account the player's performance against other rated players. The formula used in chess was devised by the Hungarian mathematician Arpad Elo. In go, a similar formula is used. Both rating systems are now maintained by computers.

The present paper proposes that the strength of a player be described not only by his rated playing strength, but also by a power measure that is not based on performance against other players. Instead, it is based on results of computer-administered tests, and is related to power as defined in physics.

Measuring Power

Power is a measure of a player's skill and knowledge, without considering playing strength. Here is how power can be measured:

The player is presented with a game position on a computer screen, and is asked to indicate what he believes to be one of the best moves. After each try, the computer responds with either "Correct" or "Try again." The player keeps trying until the computer responds "Correct."

This process is repeated for many positions, taken from the beginning to the end of a game. Any previously played game can be used. In the case of chess, the set of possible alternative "best moves" for each position can now be determined by a computer of grandmaster strength. In the case of go, the "best

moves" would have to be established by a previous analysis of the game made by players of professional strength. In most go positions, there are only one or two best moves, but in the very early stages of a game there can be five to ten.

The computer registers both the number of tries and the time the player uses to find a best move in each position. It uses this data to calculate the player's power score for each individual position, for types of positions, and for the game as a whole. In other words, power is a function of the number of tries and the length of time needed to arrive at one of the best moves in each position¹.

Comparing Power and Practical Playing Strength

Power is a measure of the player's skill and knowledge, which is not necessarily highly correlated with the player's practical playing strength. Practical playing strength depends not only on skill and knowledge, but also on stamina, fighting spirit, performance under stress, persistence in the face of setbacks, energy level, steadiness, and ability to sustain vigilance. It is strongly affected by the frequency and severity of mistakes and lapses of concentration. In practical play, one mistake can, and often does, undo the result of many superb moves, and can lose the entire game.

Power, on the other hand, is a measure only of the ability to find best moves. Occasional mistakes have little effect on the power measure.

In common parlance, the terms "power" and "strength" are often used interchangeably. Here, the term "strength" will be reserved for practical playing strength, as measured by current rating systems, and the term "power" will refer to the measure of skill and knowledge provided by the computerized test described above and the formula described below.

Information Deficit or Uncertainty

The difficulty of any question or problem, for a given individual, can be expressed as an "information deficit," "knowledge gap", or "uncertainty." We will call it U for short. The uncertainty U can be thought of as the number of yes-no questions you would need to ask to get the answer. For example, if asked "In what city was Mr. X born?" you would need to ask fewer yes-no questions if you already had some knowledge about Mr. X's nationality or race than if you knew nothing at all about Mr. X. Furthermore, if you recognized that he speaks with a Texas accent, you would have a still smaller U and would need to ask still fewer yes-no questions.

¹ A refinement of this procedure would be to give "partial credit" for trying second-best and third-best moves. Partial credit could take the form of counting such tries as fractional, rather than whole guesses.

The field of mathematics called "information theory", developed by Claude Shannon around 1950, permits us to quantify U even when the knowledge gap is measured by means other than number of yes-no questions that need to be asked. For example, we can count the number of guesses or tries needed to arrive at the correct answer.

Suppose I tell you that I tossed a coin three times, and asked you how it landed on each of the three tosses. Your uncertainty U , when confronted with that question, is 3 bits, because you would need to ask the 3 yes-no questions: (1) "Did it land heads on the first toss?" (2) "Did it land heads on the second toss?" and (3) "Did it land heads on the third toss?" In situations where yes-no questions or other types of questions are not used, and the subject just keeps trying or guessing, one can take the log to the base 2 of the number of tries or guesses. The resulting number U is an approximation to the number of yes-no questions the subject would have needed to ask if that had been the procedure. For example, if 8 tries are needed, the U measure would be 3, because the log of 8 is 3.

The U Measure When Other Probing Strategies Are Used

This U measurement approach is applicable even when the "tries" are based on partial information or knowledge, and even when there are various kinds and more elaborate types of questioning strategies that could be used. In our measurement of U , we consider only the number of tries, and don't permit the subject to use any questioning strategies at all. According to Shannon, the log of the number of tries provides the desired U measure, which is logically and mathematically related to the number of yes-no questions that would have been needed if some type of yes-no questioning strategy had been used².

For a top player, who would always get one of the best moves on the first try, U would be zero bits, since the log of 1 is zero. That means he would need to ask zero yes-no questions. A weaker player's U , who may need, say, sixteen tries to arrive at the best move for that same position, would be 4 bits (i.e., 4 yes-no questions). So, the weaker the player, the greater the U .

² Claude Shannon, in his 1950 book "The Entropy of Printed English," showed that from the point of view of a data processor, such as a person, the entropy, or information content, of any complex system, such as English text, can be measured by taking the log of the number of guesses needed to arrive at the answer to any question regarding the system (e.g., to guess each letter of every word in a passage), and that this entropy measure, expressed in bits of information, corresponds to the number of yes-no questions that the person would need to ask if a questioning strategy were used. This approach has come to be called "the Shannon guessing game."

The Power Measure

The formula used to measure power takes into account not only U but also the time taken. The formula for U is closely related to formulas for the physical entities of entropy, heat, and work. When time is taken into account, the "rate of doing work" or "heat flow" correspond to the physical concept of power. Therefore, U with time taken into account, can justifiably be called "power."

In the proposed formula, power is defined in such a way that the power of the player who always arrives at a best move (or one of the best moves) in zero seconds is 1.00, the maximum possible power. A player who needs the maximum number of tries and the maximum time would get a power rating of close to zero. Therefore, in the proposed formula, a function of time must appear in the denominator.

Benefits and Advantages of the Power Measure

Problems of the type "What's the best move?" are widely used in game magazines and standard instructional books. The measurement of skill and knowledge by means of tests rather than competitive play is commonplace. Such tests usually provide either a "percent correct" score or a total number of points that reflects the quality of the answers chosen. The power measure is more valid, more meaningful, and more useful than any such measure, for the reasons discussed below.

Many conventional tests use scores based on "percent correct." One disadvantage of such scores is that a large number of positions must be used for the measurement to be sufficiently reliable. That's because only the first try counts, as in a game, and therefore each individual position yields only a single right/wrong score, usually of zero or 1. With the power measure, on the other hand, each individual position yields a score that can fall anywhere between zero and 1.00.

Another disadvantage of "percent correct" measures is that they severely penalize stylistic quirks, such as, for example, a tendency to try original or eccentric second-best moves on the first try. The power measure, on the other hand, would not penalize such tendencies significantly.

In some tests where the subject is asked to choose between alternative moves, different choices are sometimes given arbitrary point scores. In such tests, the point score value assigned to an incorrect choice does not reflect the magnitude of the knowledge deficiency that generated the incorrect choice. With the power measure, on the other hand, the log of the number of tries provides a sensitive measure of the knowledge deficiency.

Finally, the power measure differs from other non-play measures in that it takes time into account. This is the feature that makes it a measure of power.

Practical Uses of the Power Measure

A practical advantage of the power measure is that it is quick, convenient, and does not require any opponents. Each determination of power may take anywhere from a few minutes to a few hours, depending on the desired degree of accuracy and reliability. A serious go or chess student would be able to obtain daily reliable readings regarding his progress.

A second advantage is that it allows us to isolate the skill/knowledge component of playing strength, and measure it separately from the other components, like stamina and fighting spirit. Currently used rating systems cannot do that — they can measure only practical playing results. For example, rating systems do not permit a player to measure his skill if he has a psychological characteristic that interferes with his practical play.

The power measure would make it possible to compare players from different countries, cities, or clubs whose rating systems may be different or whose anchorage points may have drifted apart over time. In the distant future, the power measure would be usable for comparing players of different epochs, without the problem of long-term drift, and opponents who are not in their prime, or even alive, at the same time.

Players usually differ in their playing strength in different phases of the game, such as the opening, middle, and end game. These types of differences cannot be measured independently in competitive play. But the power measure can be applied separately to each phase of the game.

The chess and go literatures are replete with references to the "difficulty" of moves. Until now, difficulty of a move has been a subjective concept, one that cannot be quantified. The power measure provides, for the first time, a meaningful way to quantify the difficulty of a move for players of a particular category. Difficulty would be defined simply as one minus the average power score on that move for various players of that category, divided by one minus their power score on the average move.

Using the Power Measure to Compare Chess and Go

The power measure makes it possible, also for the first time, to compare chess and go in a meaningful way. Here are some of the questions that can be asked and answered by using the power measure:

Is a given player more powerful in chess or in go?

Do the top chess grandmasters approach the maximum power of 1.00 more closely than the top go professionals?

How much study and experience, and how long, does it take to reach the same level of power in chess and in go?

What is the correspondence between chess ratings and go ratings at a given power level?

Instructional Uses of the Power Measure

Large numbers of players of various power levels, working their way through numerous test games, would generate a gold mine of data. The data, separated according to players' power levels, would include the actual tries (not just the number of tries) that were entered in each position. Such data could be put to numerous instructional uses.

One type of instructional use would be to identify the positions that players of any given power level find most difficult. This information could be used to design and produce separate instructional materials specifically designed for each category of power. Such materials would be aimed at the empirically determined deficiencies of players in each category.

Such instructional materials could also be based on any typical recurring patterns of incorrect tries in given positions. Such patterns, once identified, would provide diagnostic clues to typical dysfunctional thought processes needing correction at each power level.

The data could also be used clinically, to provide an individual player with feedback after each position of the test game, if he requests such feedback. He could elect to see his power score for that position along with (a) the ratio of his power score to that of the average player of his level, and (b) the average power level of other players who averaged the same score as he did for that position.

This type of feedback would provide a powerful self-diagnostic tool for ascertaining the types of positions in which the player has weaknesses, taking his power level into account, and where remediation would be in order. Remedial instruction for that player can then be focussed on those specific categories of positions.

Finally, the power test could be used to measure the strength and power of computer programs, for both chess and go, not only by means of practical competitive play against humans or other computers, but also on a move by move basis. It is well known that all computer chess and go programs make occasional losing blunders interspersed among very strong moves. As in human play, one blunders can vitiate a lot of good play. The power measure could be used not only as a supplemental or alternative measure of the strength of computer programs, but also as a tool for pinpointing a program's strengths and weaknesses.

THE POWER FORMULA

As was explained earlier, U is the log of the number of tries used.

When U is zero, we want the power expression to have a value of 1.00, and when U is infinite we want it to have a value of zero. The simplest expression that satisfies this requirement is e where e is the base of natural logarithms. But this expression lacks the time factor, and is therefore not a yet a power expression.

For the time function too, we want the expression to have a value of 1.00 when the time consumed is zero, and a value of zero when the time consumed is infinite. So again, the simplest expression that satisfies this requirement is e , where t is the thinking time for each move. The definition of power requires the time function to be in the denominator.

So, if the player needs only one guess and makes it in zero seconds, his power score on that move is 1.00. As he expends more guesses and time, his power score approaches zero asymptotically. Putting the U function in the numerator and the time function in the denominator, the power equation would be

where k is simply the weighting of the time factor relative to the U factor.

When this measurement system is implemented in practice, it goes without saying that the computer software should automatically subtract out the time consumed by the physical act of keying in each try.

It is interesting to note, in passing, that this definition of power is similar to the physical concept of power: U has the same formula as the thermodynamic concept of entropy. In physics, entropy is closely related to heat and work, and

power is heat or work per unit time. In our formula, as in the physics definition of power, the time factor is in the denominator. In the power formula as written above, the minus sign brings the time factor into the numerator in accordance with algebra.

The Time Factor k

The power formula shows that within limits the player can trade off number of tries against thinking time in a way that leaves his power score unaffected. In other words, he can achieve the same power score by thinking longer so as to require fewer tries, or by trying more moves more quickly so as to cut down on thinking time.

This means that the player's power score would usually not be strongly affected by the amount of time he spends on each position, within limits. We will use the term "power constancy" for this type of compensation or complementarity between thinking time and number of tries.

The degree of power constancy will depend on the value of the time factor k selected. Obviously, if k is zero, time plays no role at all, and the way for the player to maximize his power score would be to think as long as possible on each try. In that case, there would be no power constancy at all. On the other hand, if k is set at an extremely high value approaching infinity, it is the number of tries that plays no role at all, and in this case the way to maximize the score would be to minimize thinking time by trying every reasonable move as quickly as possible. Again, no power constancy.

Therefore, power constancy is maximized at some intermediate value of the time factor k . It would seem that the value of k that maximizes constancy would be the one for which $U = k \log t$. Whatever value of k is used, the player should be assisted by appropriate computer-generated feedback to help him adjust his thinking time to that k setting, so that he may maximize his power scores. This assistance could take the form of beep signals at meaningful points, possibly combined with a visual display in a corner of the screen.

Other Applications of the Power Measure

Chess and go are certainly not the only games to which the power measure is applicable. With appropriate adaptation, it is applicable to any game in which the player can enter decisions or moves on the computer keyboard.

The power measure should also be useful for measuring mastery in knowledge domains not generally called games. It could become a generalized testing tool for complex skills involving timely judgment-based moves or decisions, as in

certain types of military, business, legal, counseling, and social situations. The power measure would also be useful as a computer-based testing tool for power tests in academic subjects like mathematics, the sciences, literature, history, etc.