Aaron Brown

Stumbling on Economics

Armchair coaching manual proves a veritable exemplar of all that ails economics lit

My last two columns have been generally critical reviews of what might be considered the top of the economics profession. *This Time is Different* is a highly influential work by top practitioners on the big issues of economic policy. *The Squab Lake Report* represented the consensus of the world's top financial economists on how to fix the financial system. In this issue, I want to go to the other extreme. Not the bottom of the profession, but to a work far from the center of theoretical and policy debates, on the borderline between economics and hobby writing. *Stumbling on Wins* (FT Press, 2010) claims to apply economic reasoning to issues in sports. It is written by two less famous economics professors, David Berri and Martin Schmidt.

You will quickly discover that I didn't like this book. It has an unpleasant tone of smugness, as the authors criticize coaches and executives mercilessly for deviating from the behavior their models suggest. There is no hint that either author knows or cares anything about the games or the people involved, has ever played or even watched the games in question, or is knowledgeable about them beyond running some regressions on box score data. At no point in the book do the authors mention debating their conclusions with the people they criticize, nor making other efforts to falsify their results. The impression given is that the regression is run, the authors are incredibly smart to have run it, and that anyone who disagrees is much less smart. Let me emphasize, I am not saying that any of this is true; the authors may be experienced coaches with a deep love of the games and they may have listened carefully to all contrary opinions. But they failed to indicate any of those things in the book.

The flaws in the book are things that I find distressingly common in economic writing, and in this simple context they are easier to describe than with more complicated analyses. The first one is pretty basic. Economics often studies data that have major flaws in their definitions. Gross domestic product, for example, leaves out unpaid work and the underground economy, and accepts valuations that may be far from reality. The consumer price index applies only to a specific basket of goods, and one that changes over time in ways that cannot be fully adjusted, and deals badly with expenses such as interest and taxes. These flaws are well known inside the profession. They do not make conclusions worthless, but they do create significant gaps between model statements, say a relation between the unemployment rate and the inflation rate, and reality. Good writers make this clear, and temper conclusions and recommendations with solid doses of humility as a result. Moreover, they attempt to relate model predictions to more robust observations – for example, checking an assertion about GDP against electricity consumption or inflation data against the price of a Big Mac.

Inside the box

Stumbling on Wins almost never raises its eyes from the box scores and a few other pieces of easily available data, like salaries and draft positions. Even if you came up with a model that successfully related all the numbers, it does not follow that you have understood the sport. For one example, almost all box score data in basketball refer to offensive production. So, when rating basketball players, the authors assume that all players on a team are equally good at defense. This is justified (in an appendix) by, "The validity of this assumption is bolstered by the fact that teams typically play defense together." It's not clear what that means. That players are on the floor at the same time? That players on a team cooperate? Both of those are equally true of offense. It does suggest that the authors have never seen a basketball game. Anyway, elimination of defensive ability from player evaluation, on the grounds that it is missing from your data set, should at least merit some qualification of conclusions (especially because conventional wisdom is that defense is more important than offense).

Another objectionable flaw is less-than-forthright references and data. This is not a problem limited to economics. I often find myself concentrating on a small aspect of a paper, and chasing down references or data to support it. I think that anyone who has done this will agree that it is sometimes disappointing; the references or data are either unavailable (or may not exist) or don't say what the paper says they say. My friend Stan Young of the National Institute of Statistical Sciences documented this for a range of health science journals.

In this case, virtually the entire analysis of basketball is based on WP48, the authors' evaluation of basketball player ability. The text says that the details are in an appendix. The appendix says that the details are online. Online, there are two different explanations (one of which appears to predate the three-point shot in the NBA in 1979), with no clue as to which one applies. Neither matches the one in the book precisely and neither gives key details, such as how to deal with players who switch teams, players who play multiple positions, overtime, injuries, and so forth. These are not huge quantitative factors, but they are important for testing the model (e.g., players who switch teams help you decide if the player rating reflects player ability or situation on the team). Most importantly, the data are missing, and no specific source is given. This is not an issue for the more recent years, for which there are a number of consistent databases that mostly agree. But 20th century data sources have considerable diversity and numbers are usually adjusted for known biases. The Website does claim that the details are in five other books, none available online, with a combined list price of over $300.

Ice age

As Carl Sagan was fond of pointing out, absence of evidence is not evidence of absence. *One of the most newsworthy claims of Stumbling on Wins is that goalies in hockey don't make much difference.* This is supported by two arguments, one of which is that the squared correlation between successive season save percentages is only 6 percent. In Sagan's term, the authors have failed to find evidence of consistency in performance, and therefore concluded that performance is inconsistent.

Let's ask another question. Suppose goalie performance was completely consistent; what squared correlation would we expect to observe? The answer is not 100 percent. Consider a goalie with the ability to stop 90 percent of the shots on goal. Over a season with 1,600 shots, he expects to allow 160 goals, but the standard deviation on that number is 12 goals, assuming independence. So, he could easily end the season with a save percentage between 0.885 and 0.915. But the standard deviation among goalies is only about 0.010. The squared correlation coefficient can be thought of as one minus the ratio of squared deviations for the same goalie over a series of years to the total squared deviations. If the random variation among years for completely consistent goalies is roughly the same size as the variation among goalies, you expect a squared correlation coefficient of about 50 percent.

The problem is much worse for goalies who face fewer shots on goal. The authors partially correct for this by weighting by minutes on ice (they don't specify exactly how they do the weighting). The result is that you expect a squared correlation near 10 percent if every goalie is completely consistent. The number is driven down significantly by a few outlier results from little-used goalies. A 6 percent result for any two seasons is well within standard confidence bands. The authors claim that result over eight seasons (I can't replicate it), which is significant at the 5 percent level. But even accepting the number, it means only that all goalies are probably not completely consistent. It would take only a few inconsistent goalies, or other complicating factors like injuries or team effects, to explain the results. It is entirely beyond the pale to claim that no goalies are consistent on the basis of this evidence. In fact, there is overwhelming evidence for consistency of goalies who face over 1,500 shots per season, which is what matters for the argument.

There is a second argument for goalie inconsistency. Only 30 percent of the goalies in the top ten for save percentages in one season are also in the top ten in the next season. No significance testing is given. I ran a simulation, taking the actual shots faced by each goalie and a random draw of save percentage, assuming independent random stops at the goalie's lifetime success rate (i.e., perfect consistency across entire careers). In none of the simulations did as many as 30

Elimination of defensive ability from player evaluation, on the grounds that it is missing from your data set, should at least merit some qualification of conclusions

percent repeat. The effect of goalies with few shots, plus injuries and retirements, kept repeat rates low. So, the mystery is why so many goalies repeat, not so few.

All of this is based on the authors' assertion that save percentage is the only important way that a goalie affects the game. They note that goals given up is the product of shots on goal multiplied by one minus save percentage. But "a goalie has little control over" shots on goal. No evidence or argument is given for this. Statistically, there is a strong and consistent relation between goalie and shots on goal, even for two goalies on the same team, or for goalies who switch teams.

Let's think from the perspective of a player who has a potential shot on goal. He has to weigh the probability of success versus the potential to improve the shot with a move or a pass. The potential for improvement, and the risk of loss in attempting it, is likely determined by the situation and by players other than the goalie. But whatever its value, the better the goalie, the less the value of the immediate shot; thus, the greater the appeal of trying for improvement. For example, if there's an 80 percent chance of improving the shot prospects by 2 percent but a 20 percent chance of losing the puck, you'll shoot now if you have an 8 percent or better chance of getting the goal, but hold off if you have less. The better the goalie, the fewer the shots.

Of course, I don't think every NHL forward is doing probability calculations in the heat of play. Nevertheless, it seems reasonable to work to develop better shots, and thereby get fewer shots, versus a good goalie. Moreover, you'll also likely make more difficult shots because the good goalie is more likely to be in a position to block the easy shots. Some of those difficult shots will miss by too much to be counted as shots. And then there's the fact that a good goalie can do more than stop the goal, he can control the puck or direct it to his team. This also results in fewer shots, and many goals in hockey are scored on rebounds.

Basketball Jones

The first extended discussion in the book claims that professional basketball coaches and executives err by preferring players who score a lot of points to players with high shooting percentages. This is a staggering claim. It means that the best players in the NBA sit on the bench every night, and even better players are cut before the season begins. The starters and stars are no better than average players. Those coaches must be incredible dummies if the authors are correct.

The claim also flies in the face of conventional analysis. Sports statistics are usually stated either as averages or totals. Averages are used when players face more or less identical situations repeatedly: batting average in baseball, extra point percentage for kickers in football, free throw percentage for basketball players. But when situations differ significantly (and in ways that do not average out) and your team decides whether or not to use you, totals make more sense. Consider a field goal kicker in football. The better he is, the more long field goals he will be asked to try, so he might end up with a lower average than an inferior kicker. His value is measured by the number of field goals he kicks, not the percentage. In baseball, the best power hitters are given the most RBI opportunities, so we measure total RBIs, not RBIs per at-bat.

It means that the best players in the NBA sit on the bench every night, and even better players are cut before the season begins

Averages suffer from two potential problems. First is that the situations may not be completely identical. Second is that you have to figure out a way to weigh various outcomes. Averages measure only a specific skill – hitting a baseball, kicking an extra point, making a free throw – not overall value. Totals are better in these respects because they have a built-in correction mechanism. Suppose one football running back plays on a team with a great quarterback, and another plays for a team with a weak quarterback. The first running back's total yards will suffer because the team will call more passes, but will be helped because the defense will be concentrating more on passes, so he will get more yards each time he does get the ball. If a basketball player plays good defense, he will get more minutes on the floor, and pick up points and rebounds. The big problem with totals is that they presume that the player is being used to maximum advantage.

To a first approximation, all players on a basketball team should have the same shooting percentage; it should be unrelated to skill. The team should give the shot to whoever has the best chance of making it. If one player has a higher percentage, he should get more shots, which will draw more defensive attention, which will equalize the percentages. The main adjustment to this is that some players, point guards for example, can get a shot almost every time, with no work from the rest of the team. Centers, on the other hand, generally need a well-executed play to get them the ball in position. Since that requires work and risk of turnover or shot clock violation, centers must have a higher shooting percentage to compensate. Therefore, shooting percentage is mainly a measure of how hard it is for you to get your shot. There is a negative correlation between free-throw shooting percentage, a measure of pure shooting ability, and field-goal shooting percentage.

So, how is this claim supported? By reference to the authors' measure of player skill, WP48. WP48 does not include field goal percentage, but it does have a subtraction for field goals attempted, which penalizes players with poor field goal percentages. Why is there a subtraction for field goals attempted? You have to go to the Web site for that answer. They claim that the key variable of basketball success is points scored divided by

possessions. Why? In their words, "We are going to simply take the link between wins and the efficiency measures as given." To get a deeper answer, you need to shell out $240 for another work by one of the authors.

I refused to pay that, but I used the Amazon "search inside this book" feature to see what I could find. It may not represent the full development of the idea. As best I can tell, the only support is that if you regress season wins on points scored divided by offensive possessions, and points given up divided by defensive possession, you get good t-statistics. This is the opposite mistake to the one with hockey goalies, assuming that a high t-statistic means that a model is good.

The problem is that the rules of basketball force a near equality between possessions of the two teams. So, dividing by possessions makes little difference to the regression. The usual way to estimate wins from points is Pythagorean winning percentage, points scored raised to some power, divided by the sum of points scored raised to that power, and points given up raised to that power. For professional basketball, the usual power is 14. If you stick Pythagorean prediction into the regression, the authors' variables drop out as insignificant. Points divided by possessions adds no information to points.

So, there turns out to be no basis at all for the assertion that shooting percentage is more important than points scored. But let's ignore that and accept the claim that the WP48 formula applied to team statistics is the best predictor of season wins. The authors apply the same formula to individual player statistics, to rate players and berate coaches, general managers, and fans for foolish player rankings.

Ecology

This has a name, the ecological fallacy, assuming that the properties of a group apply to individuals. The classic example is the observation that the US states with the highest percentage of African-Americans had the harshest Jim Crow laws in the 1950s. It is an ecological fallacy to claim that African-Americans supported these laws. It's more reasonable to argue that smaller white minorities felt the need for more rigid defenses of inequality.

In basketball, we know that the team that gets more points wins. But we don't know that the player credited with the points actually caused them. The same is true for other statistics in the box score. To justify WP48 as a player rating, we have to do a player-level analysis, and one that does not collapse to the team level (e.g., if we regress player point totals on team wins, the sum of the player betas has to equal the team beta, whether or not points scored is a good measure of player contribution).

Here's one example. If WP48 is a true measure of player ability, then a team's success should depend on the WP48s available to it. All the authors' tests weight WP48 by minutes played, which essentially converts it to a team-level total.

In basketball, we know that the team that gets more points wins. But we don't know that the player credited with the points actually caused them

But over the 1,225 team-seasons for which I can get complete data, the sum of a team's players' WP48s has a negative correlation with winning percentage (not statistically significant, however). If I follow the authors' preference and weight by minutes played, I do get a significant correlation of 0.17. But compare that to the 0.93 correlation I get with Pythagorean winning percentage, which is based on simple points scored and points given up, without any of the adjustments. If I put both in a regression, weighted total WP48 remains significant, but it explains only 0.9 percent of the residual variance after Pythagorean winning percentage.

Here's another example. If WP48 reflects true player ability, it should be the same for subsets of the data. For the 2008–2009 NBA season, I computed WP48 for all players in each of the four quarters. Dwyane Wade of Miami was the 72nd best player in the NBA in the first quarter, 8th best in the second, 15th best in the third, and 34th best in the fourth. Remember, these are per-minute statistics, so they don't reflect differential playing time. The game is not noticeably different in different quarters, so we have to assign most of the discrepancies to a sampling error of WP48. The quarter-by-quarter analysis suggests that the average NBA starter's WP48 has a season sampling error of 16 percent of its value. Only a handful of players are clearly better than the average starter, and it is impossible to reliably distinguish among them. Remember, we're talking only about sampling error here; this assumes that the measure itself is perfect.

It also occurred to me that the best time to rate a basketball player is near the end of a close game. *A lot of statistics are gathered in "garbage time," when the game result is already determined.* Moreover, strategy might dictate holding back in various ways at early points in the game. But in the last five minutes of the game (or in overtime), with neither team ahead by more than five points, the game is pretty simple and performance counts. Lebron James was the top clutch WP48 player and Dwight Howard was the second. These are reasonable; both players are top overall WP48 players. But third was Carmelo Anthony, who is only an average WP48 starter on the season. There is only a 65 percent correlation between clutch performance and season performance.

That presents a severe problem. If the truth is that players have very different values in clutch time from other times, then no single player rating is possible. What seems more likely is that the different strategies at different times have a strong effect on WP48 measurement, while players' abilities remain fixed. Another indication of this is that there are large differences in WP48 when a player is measured in games versus good,

average, or bad teams. Al Jefferson is the 5th best player in the NBA versus bad teams, and 9th best against good teams, but only 24th best against average teams. That suggests he can play well versus poor defense, and also when the game is likely lost anyway (his clutch rating is 72nd in the league). On a different team, with different challenges and needs, he could look like a top player or merely a very good one.

The general fault here is failure to test the model. The authors develop it based on a theory: get some good-looking t-statistics, then insist that everyone who doesn't believe it is irrational. While lots of researchers in different fields neglect aggressive falsification efforts, it is peculiar to economists to call other people irrational as a result.

In black and white

Moving on to football, the book presents a classic example of selection bias. They want to prove that the NFL discriminated against black quarterbacks for many years. There's no doubt that they're correct, but that's no excuse for manufacturing weak evidence to bolster the charge. One of their main tests is to compare nine white Hall of Fame quarterbacks with four black quarterbacks with similar statistics according to the authors' measure, only one of whom made the Hall of Fame. They show that the black quarterbacks got to play in fewer games, and therefore had less total production, which presumably kept three of them out of the Hall.

The trouble is, we don't know how many white quarterbacks also had similar statistics and game counts. Comparing any group with a group selected for success will produce this kind of pattern.

Another example of selection bias is used to snipe at NFL quarterback drafting. The authors note that quarterbacks picked with the top ten draft slots each year have lower average productivity per play than quarterbacks picked in slots 11 to 90. They reason from this that the general managers make foolish choices. The trouble is that they are only comparing players to the extent that they made teams and continued to play. It's possible that every quarterback picked in slots one to ten was very good or better, and that only one quarterback picked in slots 11 to 90 ever made a team, but he was a star, better than the average top pick. That would be evidence of excellent discrimination on the part of draft decision makers, but the authors would draw the opposite conclusion.

It's hard to test this effect, since the details of the analysis will not be available, according to the book, until 2011. But the numbers presented make clear that teams got about 20 percent more total production out of the 11 to 90 slot players than the one to ten slot players. Assuming that the former represent eight times as many players, that means that the average 11 to 90 slot player produced about 15 percent as much as the average one to ten slot player.

That's misleading in itself, since the bulk of the value comes from the minority of players who become regular starters and enjoy long careers. If we assume that those players are equally good on average, regardless of draft position, then you have to pick about seven times as many 11 to 90 slot players than one to ten slot players in order to get one. The cost, in terms of salary, training and development, and draft slot usage of one top ten selection versus seven 11 to 90 selections, is hard to compute. Moreover, there are a number of assumptions in this computation that may not hold, and we also have to consider some partial value from backup or short-career players. Still, a quick look at the data suggests no evidence that draft decisions are foolish. The authors are misled by selection bias into an unsupported conclusion.

I know some people will tell me that I spent way too much time studying a dumb book on a minor topic. I should have put it aside after the third or fourth major blunder and read something enlightening instead

I have by no means exhausted the quantitative errors in this book, but unfortunately it gets worse on the rare occasions that the authors attempt qualitative analysis. They quote a number of basketball coaches to support their claim that points scored are not important for a player. What all the coaches actually say is that they like players to do what they're told, shoot when they're supposed to shoot, pass when they're supposed to pass.

I know some people will tell me that I spent way too much time studying a dumb book on a minor topic. I should have put it aside after the third or fourth major blunder and read something enlightening instead. I certainly shouldn't have used it as evidence against the entire economics profession.

The reason I disagree is that I see exactly these same errors in serious published work, but it's harder to demonstrate because the issues and data are more complex. Games are relatively simple to analyze, and therefore make better object lessons. I'm not claiming that all economists make these errors, but I don't think this book could have been written in the same tone by professors in another field. "Economist" is in the subtitle, and forms of the word are found nine other times on the cover, plus every third page in the text. Other fields have their own issues, but I think *Stumbling on Wins* can be laid at the door of economics.