

Journal of Quantitative Analysis in Sports

Volume 5, Issue 2

2009

Article 1

2008 NORTHERN CALIFORNIA SYMPOSIUM ON STATISTICS
AND OPERATIONS RESEARCH IN SPORTS

A Statistical Analysis of NFL Quarterback Rating Variables

Derek Stimmel*

*Menlo College, dstimmel@menlo.edu

Copyright ©2009 The Berkeley Electronic Press. All rights reserved.

A Statistical Analysis of NFL Quarterback Rating Variables

Derek Stimel

Abstract

Using data from NFL seasons 1960-2007, we examine the quarterback rating and the four variables of which it consists: average yards per attempt, completion percentage, interception percentage, and touchdown percentage. We test for structural breaks in the means and standard deviations of each variable. The analysis finds evidence that there are structural breaks in the series likely associated with rule changes designed to promote the passing game and the implementation of the salary cap. The break test results as a whole suggest that comparisons of quarterbacks from different regimes are inappropriate unless the regime differences are taken into account. There appears to have been a simultaneous improvement in quarterback performance and reduction in volatility suggestive of the idea that the relative difference between above average and average quarterbacks has been reduced. Using graph theory and the information gleaned from structural break tests, we examine the causal relationships among the four quarterback rating variables over the most recent stable period, which is 2000-2007. The causal analysis shows that completion percentage is commonly caused by interception percentage and average yards per attempt over the course of a season. Also, touchdown percentage causes average yards per attempt. We suggest possible explanations of the findings and suggest avenues for future research.

KEYWORDS: quarterback rating, structural break test, graph theory, causation

Section 1, Introduction

Perhaps no statistic is more commonly mentioned, least well understood, and more criticized than the quarterback (QB) rating. Watch an episode of ESPN's Sportscenter during the NFL season or the broadcast of an NFL game and you will see the statistic commonly listed, often followed by a discussion by the commentators that they don't understand the rating or that it makes little sense. In this paper we aim to understand the QB rating, not by a treatise on the formula itself, but rather by trying to understand how the variables that makeup the rating (average yards per attempt, completion percentage, touchdown percentage, and interception percentage) relate to each other and how they have changed over time. To do this, we will use two sets of statistical techniques: structural break tests and graph theory.

In particular, we use tests for unknown structural breaks developed in Bai and Perron (1998), Bai and Perron (2003a), and Bai and Perron (2003b). We test for breaks in the means and standard deviations of season average QB rating and the variables it consists of using NFL season data from 1960-2007. We then try to link any found breaks to rule changes or other events that have affected quarterback play.

Graph theory uses arrows to represent causal relationships identified in data from correlations among variables. The idea is that there is an isomorphism between the graph and the probability distributions of variables in the graph (Hoover 2005). We will implement a particular algorithm for this called the PC Algorithm discussed and developed in Spirtes et al (1996), Pearl (2000), and Spirtes, Glymour, and Scheines (2001). We apply this method to the most recent set of season data that we are reasonably sure contains no structural breaks based on the break test results. We then use the structural break results to add information to the graph theory analysis.

Overall we find that there is evidence of structural breaks in the means and the standard deviations of most of the individual series. This suggests that comparisons of QB ratings between quarterbacks in different regimes (over different break periods) are inappropriate. One plausible explanation for the breaks is that they are mainly due to rules changes designed to promote passing. For at least one set of breaks, the implementation of the salary cap may also be a factor. Further, the evidence from these breaks suggests that while overall quarterback performance has been improving, the spread in performance has been shrinking. The results from the graph theory analysis suggest that over a season, average yards per attempt and interception percentage affect completion percentage. Also, touchdown percentage affects average yards per attempt. We offer some possible reasons for these findings in the conclusion.

Section 2 provides a basic explanation of the QB rating and shows the basic relationship between the underlying variables. *Section 3* describes the break testing methodology. *Section 4* provides the results from the break tests. *Section 5* describes the PC Algorithm. *Section 6* shows results from applying the algorithm and incorporates the break test results. *Section 7* concludes and discusses future research.

Section 2, QB Rating Basics

To start, it is useful to have a basic understanding of the quarterback rating formula. The formula is a linear combination of four variables that are meant to capture quarterback play. The four variables are average yards per attempt (*AYPA*), completion percentage (*CP*), touchdown percentage (*TP*), and interception percentage (*IP*). Each variable is defined relative to the pass attempt. For example, over a season, *AYPA* is total passing yards divided by the total number of pass attempts, *CP* is total completions divided by total number of pass attempts and so on. The NFL’s website provides a reasonably concise explanation of the precise QB rating calculation (NFL 2008). The calculation for the quarterback rating (*QBR*) is shown in equation (1).

$$QBR = \frac{0.25(AYPA - 3)}{6} + \frac{0.05(CP - 30)}{6} + \frac{0.2(TP)}{6} + \frac{2.375 - 0.25(IP)}{6} \quad (1)$$

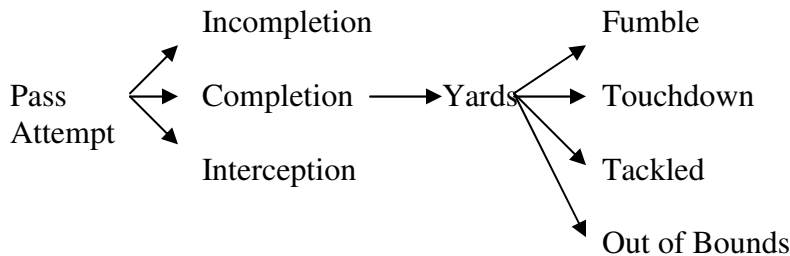
<i>Average Yards</i>	<i>Completion %</i>	<i>Touchdown %</i>	<i>Interception %</i>
<i>Per Attempt</i>	<i>Component</i>	<i>Component</i>	<i>Component</i>
<i>Component</i>			

Further complicating the formula in equation (1), no component is allowed to be negative. For example, a completion percentage less than 30% enters the formula as equivalent to a 30% completion rate and the numerator of that component is entered as zero rather than negative. Also, no numerator of any component is allowed to be larger than 2.375. So for example, any completion percentage greater than 77.5% enters the formula as equivalent to a 77.5% and the numerator of that component is 2.375 rather than a larger number. This produces a quarterback rating that can vary between 0 and 158.3.

To see why these four variables would capture passing ability irrespective of the sensibility of the formula itself, it is helpful to use graph theory, a visual representation of causal relations among variables. At this point it is worth noting that when we use the phrase “cause” such as “A causes B”, we do not mean that A is sufficient to cause B to occur. What we mean is in a statistical sense, the event B depends on the event A given the conditional correlations of variables in the system. For a single pass attempt, the causation is clear because it is defined by

time. *Figure 1* shows that relationship. We do not claim that *Figure 1* is causally sufficient. By that we mean there are other factors, other variables and arrows connecting them to the variables shown, that affect the relationships. For example,

Figure 1
Causal Relations For a Single Pass Attempt



the abilities of the defensive secondary will have an affect on the probability of the events in each step of the pass play. However, we see that the components in the QB rating capture something about each step in the possible outcomes of a pass attempt. Recall that the QB rating variables are all defined relative to the pass attempt (divided by attempts) and that affects the causal relationships among them. For example, it is clear that on a single play a completion causes passing touchdowns because a completion must come before a passing touchdown, but over the course of a game or a season does the completion percentage cause touchdown percentage? That question is explored in *Section 4* and *Section 5*.

We believe the basic criticisms against the QB rating can be grouped into the following three categories:

- 1) The formula itself is difficult to understand and the results difficult to interpret.
- 2) The rating is an inaccurate measure of the performance and value of a quarterback by not accounting for a variety of factors like quarterback rushing yards and yards after catch attributable to receivers.
- 3) The inability to compare quarterbacks of different eras to one another because QB ratings have been rising over time and rules changes to promote passing have occurred.

The first criticism is true superficially. It is difficult to tell from looking at the formula in equation (1) the reasoning behind how the four variables are combined. For example, because yards per attempt and completion percentage are not measured on the same scale each must be adjusted so they can then be combined together in the overall formula. The normalization is based on the

average performance of quarterbacks in the 1970s, which is a difficult frame of reference the further away the current period is from it. Also, the fact that each component is potentially truncated and that the scale is 0 to 158.3 rather than a more natural 0 to 100, are both added complications. A lot of discussion of the QB rating, how it was developed, and these types of complaints against it can be found in popular media articles such as Steinberg (2001), Sandomir (2004), Mushnick (2007), and Bialik (2008).

The second criticism is especially relevant for studies that are interested in determining the relative value of a quarterback to other position players. For example, if a study is trying to determine optimal relative salaries of players, the quarterback rating may not be an accurate measure of the value-added by a quarterback. This criticism has led to alternative suggestions such as the “QB Score” of Berri, Schmidt, and Brook (2006). However, the quarterback rating may be a valid overall measure of the passing game. As *Figure 1* shows, it does include variables that capture something about each step in the passing game.

The third criticism appears to stem from the fact that the formula is based on the average quarterback performance of a 1970’s era quarterback. That alone, though, does not invalidate comparisons between eras. For example, in economics, it is quite common to have a reference year or base year for valid comparisons (the consumer price index or CPI is the most visible example of this). Rules changes or other structural breaks that separate eras would be a valid concern. Their presence would not make comparisons impossible, but any comparison would have to adjust for the different regimes. The structural break tests described and implemented in *Section 3* and *Section 4*.

Here is a brief description of the dataset used. Data is seasonal data, publicly available from the NFL website (NFL 2008). We limit the data to quarterbacks that averaged a minimum of 14 attempted passes per game each season. This is the standard used by the NFL (NFL 2008) and it eliminates spot starters, relief appearance quarterbacks, and other position players that attempted a pass (a running back throwing a halfback option pass for example). Thus the sample is starting quarterbacks with significant playing time over the course of a season. Data on QB ratings is available as far back as 1937, however we limit the sample to starting in 1960. There are some noticeable jumps in the number of quarterbacks that qualify under the minimum 14 attempted passes criteria due to league expansion and an increase in passing more generally. Prior to 1960, there were no more than 12 quarterbacks in a season that met the minimum number of attempts. After 1960, other than 1973 (17 quarterbacks) there have been at least 20 quarterbacks meeting the minimum number of attempts and since around 1980, the number of quarterbacks has hovered around 30. 1960 is the year that the American Football League or AFL (precursor to the American Football Conference or AFC) began. This added a lot of teams and quarterbacks to the

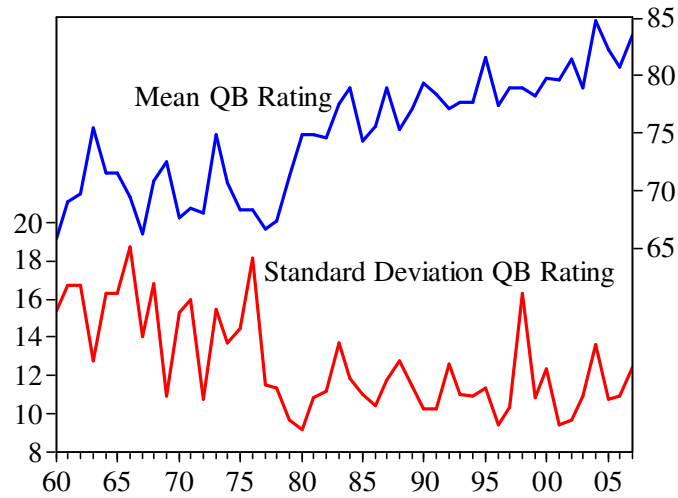
sample and makes a natural starting point. The AFL was also known to be a more passing offense oriented league. 1960 is the beginning of what would eventually become the merged NFL (combining the old AFL and NFL into AFC and NFC under the NFL banner).

We start by calculating the annual mean and standard deviations for the QB rating and the four variables that comprise it.¹ We graph these over time in *Figure 2*, *Figure 3a*, and *Figure 3b*. From *Figure 2* it is fairly clear that the mean QB rating has risen over time. Further, there appears to be a noticeable fall in the standard deviation of QB ratings over time as well. In particular each series in *Figure 2* has a noticeable jump in the late 1970s. This may correspond to rules changes put in place to promote passing and we will examine this in *Section 3*. In looking at the graphs for the individual variables (*Figure 3a* and *Figure 3b*), the standard deviations reveal a similar drop that may correspond to the late 1970s. However, in looking at the means in *Figure 3a*, it is hard to see a particular break in completion percentage. Completion percentage appears to have been rising fairly consistently over time since 1960. Average yards per attempt have changed little over time. Touchdown percentage has actually fallen, which should lower QB ratings. Interception percentage though, has fallen even more.

It is worth recalling that these variables have pass attempts in the denominator. What is occurring is that the number of interceptions thrown over a season by a QB has been generally falling since the 1960s while the number of pass attempts has been rising. Touchdowns thrown actually dropped in the early 1970s compared to the 1960s, but have risen generally since then. It appears that rising mean completion percentages and falling mean interception percentages are driving the rising mean QB rating. However, it is not clear from *Figure 3a* that either the mean completion percentage or mean interception percentage contain the apparent structural break in the late 1970s that the QB rating appears to have in *Figure 2*. The next section describes and implements formal structural break testing. It is also worth noting that often the West Coast Offense or the advent of the short passing game is usually cited as causes of things like the rising completion percentage. While not denying that as a possible contributing factor, *Figure 3a* appears to show this has been happening since the 1960s. Addressing that issue further is beyond the scope here, but would be an interesting question for future research.

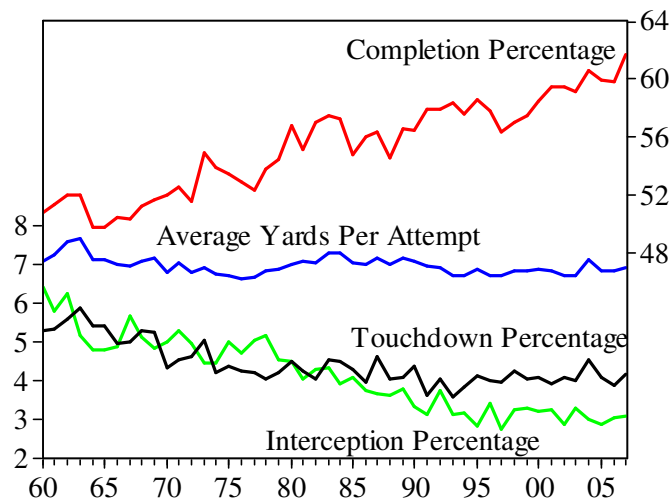
¹ We also looked at the median as well. Results with the median are similar to the results using the mean, so we dropped them for conciseness.

Figure 2
Mean and Standard Deviation of QB Rating 1960-2007



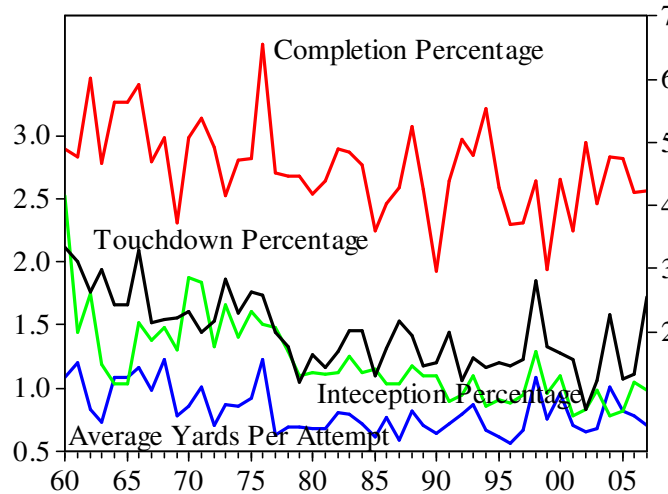
Notes: The bottom axis is years. The right axis is for the mean and the left axis is for the standard deviation.

Figure 3a
Means of QB Rating Variables, 1960-2007



Notes: The bottom axis is years. The right axis is for the mean completion percentage from each season. The left axis is for the means of the other variables.

Figure 3b
Standard Deviations of QB Rating Variables, 1960-2007



Notes: The bottom axis is years. The right axis is for the standard deviation of completion percentage from each season. The left axis is for the standard deviations of the other variables.

Section 3, Structural Break Testing

The objective is to test each mean series and standard deviation series for structural breaks. Bai and Perron (1998, 2003a, and 2003b) developed a testing procedure for unknown structural breaks. Andrews (1993), Garcia and Perron (1996) and Liu, Wu, and Zidek (1997) consider similar issues. Here we are interested in the following model.

$$y_t = \alpha_j + \varepsilon_t \tag{2}$$

$$t = T_{j-1} + 1, \dots, T_j, \text{ with } j = 1, \dots, m + 1, T_0 = 0 \text{ and } T_{m+1} = T$$

In equation (2), y is the series we are testing, α is a constant or intercept, and ε is the error term. As the model only has an intercept, we are testing for shifts in the means of the series (shift in the mean of the mean or shift in the mean of the standard deviation). There are m unknown breaks ($m+1$ regimes), and T observations. This particular model is a pure structural change model as all independent variables (the only independent variable in this case) are allowed to change. The Bai and Perron (1998) procedure is broader and can accommodate partial structural changes where some parameters are fixed.

The problem then is to estimate the parameters as well as the break dates (T_1, \dots, T_m) . Minimizing the sum of squared residuals for each m -partition can do this. Bai and Perron (1998 and 2003a) develop an efficient algorithm based on dynamic programming to find the estimated break dates. The least squares estimators for equation (2) are found over all possible break dates (T_1, \dots, T_m) . The estimated break dates $(\hat{T}_1, \dots, \hat{T}_m)$ are the ones that satisfy

$$(\hat{T}_1, \dots, \hat{T}_m) = \arg \min_{T_1, \dots, T_m} S_T(T_1, \dots, T_m) \quad (3)$$

where $S_T(T_1, \dots, T_m)$ is the sum of squared residuals. Thus the break dates are global minimizers of the sum of squared residuals. To choose the number of breaks, Bai and Perron (1998) explore a number of possibilities. Bai and Perron (2004) look at the various selection methods with simulation analysis. We use the sequential method developed in Bai and Perron (1998). The method uses a sup F test of l breaks versus $l + 1$ breaks. The test statistic $\sup F(l+1|l)$ is the least upper bound of F -statistics from the hypothesis of one additional break versus no additional breaks, varying the additional break over all possible breaks (Prodan 2008). Bai and Perron (1998 and 2003b) provide critical values for various trimming percentages and maximum possible number of breaks. The trimming percentage multiplied by the total number of observations gives the minimum length of a regime. Thus the process first examines whether there is one break versus no break, then conditional on one break, tests whether there are two breaks versus one break, and so on stopping when the smaller number of breaks is selected.

Section 4, Structural Break Test Results

We implement the break testing assuming a maximum of 5 breaks (6 possible regimes) and a trimming percentage of 0.15. With 48 observations the minimum regime length (trimming percentage*number of observations) is 7.2 or 7 years when rounded to whole years. In estimating equation (2) we use robust standard errors, which correct for the possibility of heterogeneity and serial correlation in the error terms using the method in Andrews (1991). Further, we use 1% significance level for the sup F tests as Prodan (2008) provides some evidence on size distortions of the Bai and Perron (1998) procedure. The estimated break dates with 90% confidence intervals as well as the estimated mean for each regime are in *Table 1a* for the mean series and *Table 1b* for the standard deviation series.

Table 1a
Structural Break Tests, Means, Equation (2), $y_t = \alpha_j + \varepsilon_t$, 1960-2007

Quarterback Rating						
Break Dates	1979 (1978,1981)	1994 (1989,1996)	Regimes	1960-1979	1980-1994	1995-2007
			α	69.64 (0.51)	76.79 (0.59)	80.45 (0.63)
Average Yards Per Attempt						
Break Dates	No Breaks Found		Regimes	1960-2007		
			α	6.98 (0.05)		
Completion Percentage						
Break Dates	1979 (1978,1984)	1999 (1996,2002)	Regimes	1960-1979	1980-1999	2000-2007
			α	52.06 (0.28)	56.89 (0.28)	59.84 (0.45)
Interception Percentage						
Break Dates	1983 (1982,1989)	1992 (1991,1997)	Regimes	1960-1983	1984-1992	1993-2007
			α	4.99 (0.09)	3.68 (0.15)	3.10 (0.12)
Touchdown Percentage						
Break Dates	1969 (1968,1971)		Regimes	1960-1969 1970-2007		
			α	5.35 (0.09) 4.19 (0.05)		

Notes: For the break dates, 90% confidence intervals are in parentheses. The intercept “ α ” is the average of the variable over that period. The standard error of the estimate is in parentheses. Break dates are statistically significant at 1% level using sequential sup F test.

Table 1b

Structural Break Tests, Standard Deviations, Equation (2), $y_t = \alpha_j + \varepsilon_t$, 1960-2007

Quarterback Rating				
Break Dates	1976 (1974,1978)	Regimes	1960-1976	1977-2007
		α	15.20 (0.43)	11.29 (0.32)
Average Yards Per Attempt				
Break Dates	1976 (1974,1982)	Regimes	1960-1976	1977-2007
		α	0.97 (0.03)	0.74 (0.03)
Completion Percentage				
Break Dates	1976 (1970,1983)	Regimes	1960-1976	1977-2007
		α	5.09 (0.15)	4.34 (0.11)
Interception Percentage				
Break Dates	1978 (1978,1983)	1990 (1984,1991)	Regimes	1960-1978 1979-1990 1991-2007
			α	1.51 (0.05) 1.12 (0.07) 0.95 (0.06)
Touchdown Percentage				
Break Dates	1976 (1974,1978)	Regimes	1960-1976	1977-2007
		α	1.72 (0.05)	1.28 (0.04)

Notes: For the break dates, 90% confidence intervals are in parentheses. The intercept “ α ” is the average of the variable over that period. The standard error of the estimate is in parentheses. Break dates are statistically significant at 1% level using sequential sup F test.

First, from *Table 1a*, we see the test finds 2 structural breaks in the QB rating; one in 1979 and one in 1994. Both breaks seem to match rules changes that made completing passes easier. In 1977, a rule was implemented that allowed defenders to make contact with receivers only once. In 1978, a rule was implemented that allowed defenders to make contact with receivers within 5 yards of the line of scrimmage, but limited contact after that (NFL 2008). Both changes allowed receivers to maneuver more freely and should have increased the productivity of quarterbacks. In the mid 1990s those same rules were re-enforced. For example, in 1996, the league agreed to enforce the 5-yard contact rule more strongly (Steelers Fever 2008). Alternatively, it is possible that the break in 1994 is related to the implementation of the NFL salary cap that went into effect that year. That may have prevented teams from stockpiling talent and so led to a structural break in a wide range of performance measures including QB rating. In looking at the 4 QB rating variables, the breaks in completion percentage and interception percentage are largely consistent with the breaks in QB rating. For the 1979 break, the completion percentage has a break at the same year, and the interception percentage a few years later. For the 1994 break, the situation is reversed with the interception percentage break coming just before the QB rating break and the break in completion percentage comes a few years later. The tests find no break in the average yards per attempt. The break found in the touchdown percentage is more difficult to attribute to a rule change. Of course, it's close to when the AFL and NFL officially merged into one entity, but there's no obvious reason to think that would affect touchdown percentage. One possibility is that there is a noticeable drop in mean touchdown passes thrown per season in the early 1970s. The reason for that drop-off may lead to the structural break date in 1969.

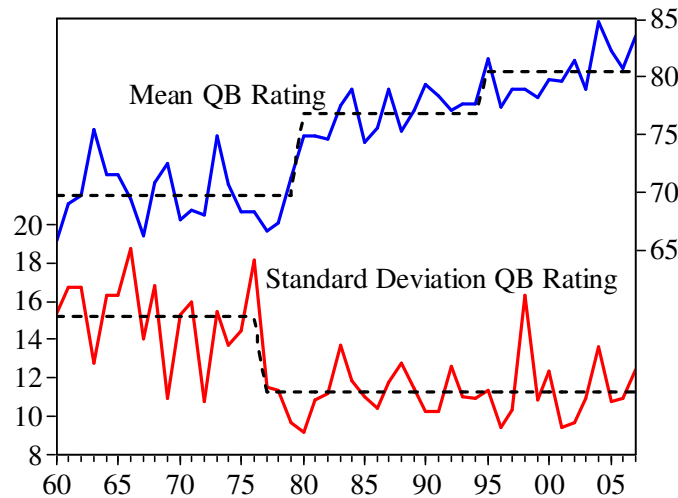
In looking at the break test results for the standard deviations shown in *Table 1b*, one result jumps out. Four of the series, QB rating, average yards per attempt, completion percentage, and touchdown percentage, all show evidence of one break in the standard deviation in 1976. Interception percentage has two breaks: 1978 and 1990. Thus all five series show evidence of a break in the mid to late 1970s likely associated with rules changes designed to promote the passing game. Further, in looking at the estimates, we can see that all series show evidence of a reduction in the standard deviation of the series after the 1970s.

Overall, there are a number of conclusions and implications that can be drawn from these break tests. One is that it is indeed not valid to compare quarterbacks with the quarterback rating over certain eras without accounting for the regime differences. So for example, comparing a quarterback from 2007 with one from 1977 would not be appropriate with the quarterback rating as they are in separate regimes and played under distinctly different rules that affected their performance. This supports the third criticism against the QB rating discussed in

Section 2. Another result is that changes in the average completion percentage and the average interception percentage are the likely drivers of the rising QB rating as their breaks align closely with the breaks in QB rating. A possible implication of *Table 1a* and *Table 1b* is that the average performance of quarterbacks or the passing game in general has improved while at the same time the spread between great performance and average performance shrunk as indicated by the standard deviation results. This could imply that the value of a great quarterback relative to an average quarterback purely on passing ability has been reduced. Confirming or even addressing that possibility in a meaningful way is beyond the scope of this paper. Also, note that there is no evidence of a break in any series from the 2000 season onward. That period will be the period we focus on in the next sections to learn about the causal relationships between the 4 QB rating variables.

Finally, we graph the series again with their estimated breaks. These are *Figure 4*, *Figure 5a*, and *Figure 5b*. What these figures show is that there may not be trends as much as there are mean shifts in variables. That suggests that it is exogenous “shocks” that drive these changes such as rules changes rather than somehow an endogenous response of the interrelationships between these variables (such as the spread of the short passing game).

Figure 4
Mean and Standard Deviation of QB Rating 1960-2007 With Breaks



Notes: The bottom axis is years. The right axis is for the mean and the left axis is for the standard deviation. The dashed lines are the estimated means or intercepts shown in *Table 1a* and *Table 1b*.

Figure 5a
Means of QB Rating Variables, 1960-2007 With Breaks

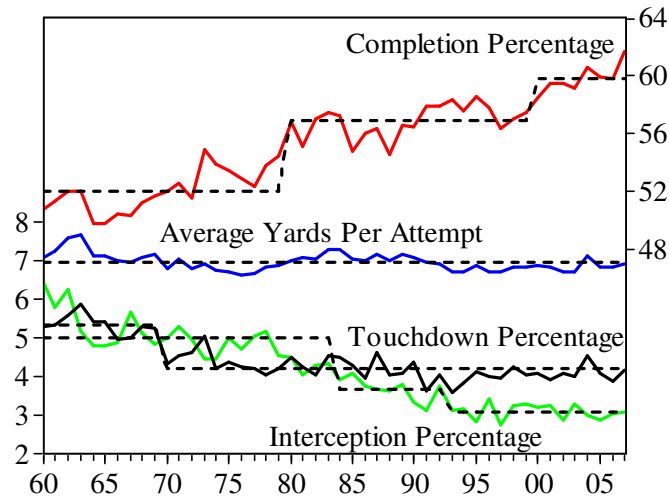
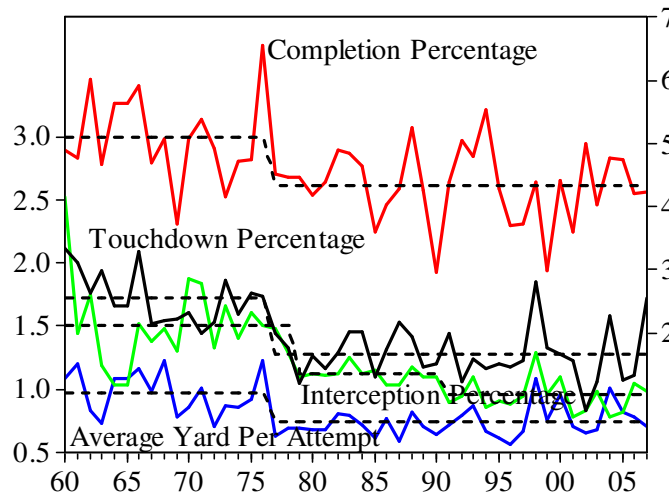


Figure 5b
Standard Deviations of QB Rating Variables, 1960-2007 With Breaks



Notes: The bottom axis is years. The right axis is for completion percentage from each season. The left axis is for the other variables. The dashed lines are the estimated means or intercepts shown in *Table 1a* and *Table 1b*.

Section 5, Graph Theory and the PC Algorithm

The concept of graph theory is that from information in the correlations among variables in a system, the dependency, non-dependence, and direction of the dependence might be extracted. These relationships can be mapped to a graphical representation of the causal relationships (Hoover 2003). The main underlying theory is the “causal Markov condition”. It states that two correlated variables, A and B , that have a common cause or set of causes C such that $A \leftarrow C \rightarrow B$, or an intervening cause or set of causes C such that $A \rightarrow C \rightarrow B$, will be uncorrelated conditional on that set C . Intervening causes are often referred to as “screens”. In addition to the “causal Markov condition”, if A and B are conditionally independent given a set of variables excluding C and its descendants (variables directly or indirectly caused by C) and $A \rightarrow C \leftarrow B$, then A and B will be correlated conditional on C . Variable C is called an “unshielded collider”. Directly linking A and B would make C a “shielded collider” and A and B would not be conditionally independent (Hoover 2003 and Demiralp and Hoover 2003).

Algorithms to find these relationships in data have been developed. The most commonly used algorithm is the PC Algorithm (Pearl 2000 and Spirtes et al 2001). Spirtes et al (1996) have developed a program, *Tetrad III*, to implement this algorithm. The following is a description of the steps of the PC Algorithm, also discussed in Spirtes et al (1996), Cooper (1999), Pearl (2000), Spirtes et al (2001), Demiralp and Hoover (2003), and Hoover (2005).

- 1) Assume each variable in the system is connected to every other variable in the system. This is by an undirected link such as $A - B$.
- 2) Test the unconditional correlation of each pair of variables. The test is a Fisher z-test where the null is that the two variables are independent. Remove the link between variables where the test fails to reject the null hypothesis.
- 3) For remaining linked variables, test the conditional correlation of each pair conditional on a third variable using Fisher’s z-test. Remove the links where the test fails to reject the null. Repeat the test for remaining linked pairs conditional on 2 variables, 3 variables, and so on.
- 4) Test remaining linked variables to identify unshielded colliders. Where found, orient the links as $A \rightarrow C \leftarrow B$ where C is an example of an unshielded collider.
- 5) If there are sets of variables such that $A \rightarrow B - C$, orient the undirected link as $A \rightarrow B \rightarrow C$.
- 6) If there are two variables with $A - B$ and there are directed links from A to B through other variables, orient the link as $A \rightarrow B$.

In the algorithm, steps 1-3 identify what is called the “skeleton” of the causal graph, which is the causal links between variables but without the direction. Steps 4-6 direct the links. Identified causal graphs are “faithful” if they represent a unique mapping to the data’s probability distribution. There is no guarantee that the PC Algorithm will be able to resolve all of the links, though. Pearl (2000) refers to the “observational equivalence theorem”, which is that a probability distribution represented by a faithful graph could be characterized by another graph with the same skeleton and unshielded colliders. In cases where there is observational equivalence some directed links could go either direction. In that case, the *Tetrad III* program of Sprites et al (1996) leaves those links as undirected. Also, the *Tetrad III* program does not implement step 6, so that must be done by hand after running the program.

Section 6, Causal Relationships Among QB Rating Variables

We apply the PC Algorithm using *Tetrad III* to the QB rating variables using data from the 2000-2007 NFL seasons. From the structural break test results discussed in *Section 3*, we see no evidence of breaks in the various series during this time period. As an additional check that this is an appropriate sub-sample to focus on, we ran the same break test methodology from *Section 3* and *Section 4* on the simple correlations of the four QB rating variables. We found no evidence of any breaks over the entire sample except one break between completion percentage and interception percentage correlation. The estimated break date was 1999 (90% confidence interval between 1996 to 2003). Taking the point estimate as accurate adds to the evidence that there are no breaks over 2000-2007. There are two complications we have to address first before we can apply the algorithm.

One complication is that a graph identified by the algorithm is unlikely to be causally sufficient. That means that there are variables not included in our system that are important to the causal relationships. That is to be expected since we are only examining aspects of the passing game rather than aspects of the entire game. Another way to put this is that there may be a latent variables problem. That means that there is at least one additional variable that is a common cause of at least two variables in the system. As a result, we run the program not assuming causal sufficiency. Found links will be represented with an $A \circ \rightarrow B$. This indicates that either “A causes B” or that A and B have a set of common causes not included in the model or both.

The second complication is that the algorithm cannot be applied directly to trending data. This is often an issue with using graph theory in time series analysis and was first addressed in Swanson and Granger (1997). Here we have a panel dataset, so we may be concerned with the time series issue. However, from *Figure 4*, *Figure 5a*, and *Figure 5b*, we see that from 2000 onward, there is no

evidence of mean trends. As a result, we elect to assume that the series are stationary over this brief time.

We apply the PC Algorithm to four variables: average yards per pass attempt (*AYPA*), the completion percentage (*CP*), touchdown percentage (*TP*), and interception percentage (*IP*). There are 260 total observations between 2000-2007. *Table 2* shows the simple correlation matrix among these variables and *Table 3* shows the p-values from the Fisher z-test of the simple correlations (Step 2 of the PC Algorithm). From the results of those tests, no links are removed between any variables as each test rejects the null of independence.

Table 2
Correlations Among QB Rating Variables 2000-2007

Variables	<i>AYPA</i>	<i>CP</i>	<i>IP</i>	<i>TP</i>
<i>AYPA</i>	1.00			
<i>CP</i>	0.69	1.00		
<i>IP</i>	-0.20	-0.22	1.00	
<i>TP</i>	0.73	0.53	-0.21	1.00

Table 3
P-Values From Independence Tests

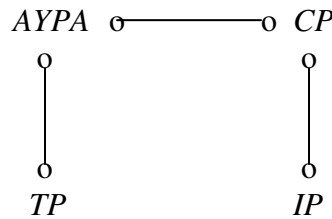
Variables	<i>AYPA</i>	<i>CP</i>	<i>IP</i>	<i>TP</i>
<i>AYPA</i>	0.00			
<i>CP</i>	0.00	0.00		
<i>IP</i>	0.00	0.00	0.00	
<i>TP</i>	0.00	0.00	0.00	0.00

Next, Step 3 of the algorithm is applied. There again, the conditional independence tests reject independence except in four cases. Those four test results are shown in *Table 4*. From *Table 4* we see that we have a decision to make about completion percentage and interception percentage. If we assume a 1% significance level, we fail to reject the null hypothesis on the conditional independence test. If we assume a 10% significance level, we will reject the null of independence. We assume a 10% level, in order to not unnecessarily eliminate links. With Step 3 complete, the skeleton of the model has been identified. We show that skeleton as *Figure 6*. There are three causal links identified and three links that have been removed.

Table 4
P-Values from Conditional Independence Tests

Link Examined	Tested Relationship	Conditional Correlation	P-Value
$CP \circ - \circ IP$	$Corr(CP, IP AYP A)$	-0.12	0.06
$IP \circ - \circ TP$	$Corr(IP, TP AYP A)$	-0.09	0.14
$AYPA \circ - \circ IP$	$Corr(AYPA, IP TP)$	-0.07	0.24
$CP \circ - \circ TP$	$Corr(CP, TP AYP A)$	0.05	0.38

Figure 6
Skeleton of Causal Graph



The next steps in the PC Algorithm are to resolve the directions of the links. Step 4 looks for unshielded colliders. Here we find that completion percentage is an unshielded collider (CP) between average yards per attempt ($AYPA$) and interception percentage (IP). That means that IP and $AYPA$ are unconditionally uncorrelated but conditional on CP they are correlated. There are no other unshielded colliders and there are no issues covered by Step 5 and Step 6 of the PC Algorithm. Thus we can resolve two of the links but not the link between $AYPA$ and TP .

Since we cannot resolve the direction of the link between $AYPA$ and TP with the PC Algorithm, we will need outside information in order to do so. As discussed in Section 2, the causation on a single play would be clear, but we have no particular intuition on how these variables once divided by attempts and aggregated over a season should relate. It turns out we may be able to use the structural break information to resolve the issue. This method is used in Hoover and Sheffrin (1992), Hoover and Siegler (2000) and Hoover (2001). From Table 1a, we only have one break in the means of these two variables, which is the break in touchdown percentage in 1969. Further, from Table 1b we see that the standard deviations of the two variables are stable from 1960 to 1976. Assuming that the basic causal relations are the same now as then, which is possibly a large assumption, we can exploit this information to resolve the final link. Assuming a linear relationship between these two variables, there are the following simple possibilities.

Table 5
 OLS Estimates of Equations (4), (5), (6), (7)

Equation	$TP = \alpha_{TP} + \varepsilon_{TP}$	
Sample	Intercept	Slope
1960-1969	5.34 (0.12)	na
1970-1976	4.46 (0.13)	na
Test of Parameter Difference	22.50 [0.00]	na
Equation	$AYPA = \mu_{AYPA} + v_{AYPA}$	
Sample	Intercept	Slope
1960-1969	7.21 (0.07)	na
1970-1976	6.81 (0.07)	na
Test of Parameter Difference	14.30 [0.00]	na
Equation	$AYPA = \alpha_{AYPA} + \beta_{AYPA} TP + \varepsilon_{AYPA}$	
Sample	Intercept	Slope
1960-1969	5.10 (0.17)	0.40 (0.03)
1970-1976	5.02 (0.15)	0.40 (0.03)
Test of Parameter Difference	0.48 [0.49]	0.35 [0.55]
Equation	$TP = \mu_{TP} + \lambda_{TP} AYPA + v_{TP}$	
Sample	Intercept	Slope
1960-1969	-3.02 (0.63)	1.16 (0.09)
1970-1976	-4.01 (0.69)	1.24 (0.10)
Test of Parameter Difference	8.36 [0.00]	7.75 [0.01]

Notes: Standard errors of estimates are in parentheses. The test of parameter difference is an *F*-test or Chow test. The p-value is in brackets. There are 213 observations in the early sample and 156 in the later sample.

$$TP = \alpha_{TP} + \varepsilon_{TP} \tag{4}$$

$$AYPA = \alpha_{AYPA} + \beta_{AYPA} TP + \varepsilon_{AYPA} \tag{5}$$

$$AYPA = \mu_{AYPA} + v_{AYPA} \tag{6}$$

$$TP = \mu_{TP} + \lambda_{TP} AYPA + v_{TP} \tag{7}$$

In the above equations α and μ terms are intercepts, β and λ terms are the slope parameters, and ε and v terms are the errors. If equation (4) and equation (5) are accurate, touchdown percentage causes average yards per pass attempt but not the reverse. If equation (6) and equation (7) are accurate, then average yards per pass attempt causes touchdown percentage but not the reverse.

Assume that equation (4) and equation (5) represent the truth. Then the structural break in touchdown percentage should affect equation (4) but not equation (5). However, if we mistakenly believe equation (6) and equation (7) are accurate, neither equation (6) nor equation (7) should be invariant to the structural break. Well, the break identified in touchdown percentage in 1969 should produce exactly that. So we estimate these equations by OLS for the period 1960-1969 and the period 1970-1977. The results are in *Table 5*. From *Table 5*, we see that only equation (5) is stable between the two sub-samples. This is only consistent with $TP \rightarrow AYPA$ and not the reverse. Thus we can resolve the direction of all the causal links in *Figure 6*. We show the causal graph in *Figure 7*. Of course, there are multiple possible explanations for the links found, we suggest some plausible ones, but determining which explanation is accurate requires more analysis than we conduct here.

Figure 7
Causal Graph



From *Figure 7*, interception percentage causes completion percentage. The two variables are negatively related as seen in *Table 2*. Intuitively this makes sense. Over the course of a season, a quarterback that throws more interceptions is likely to be a less accurate quarterback overall and so this leads to a lower completion percentage as well. Essentially, this may be an example of a latent variables problem, where the latent variable, “QB accuracy” is the common cause

of *IP* and *CP*. If true, this may indicate interception percentage is a better proxy for accuracy than completion percentage.

The second found link suggests that average yards per attempt causes completion percentage, which are positively related. At first this seems counterintuitive. We might think lower average yards per attempt would cause higher completion percentage as shorter passes have a higher probability of being completed. One possibility is that there is a measurement problem. As discussed in *Section 2* the average yards per attempt include yards after catch, which may distort the result. Alternatively, though, it may be that over the course of a season, quarterbacks that complete longer passes open up the field more, and so create a higher probability of completing passes in general. In effect, they make defenses defend more of the field, spreading the defense out. We note, for example, the top 5 quarterback in terms of average yards per attempt in 2007 were Tom Brady, Tony Romo, Peyton Manning, Ben Roethlisberger, and Brett Favre, arguably 5 of the best quarterbacks in the league.

Finally, based on the structural break information, we concluded that touchdown percentage causes average yards per attempt, which are positively related. This makes intuitive sense. Quarterbacks that complete longer passes on average are probably more likely to actually throw touchdowns. As teams get closer to the goal line, the probability of rushing for a touchdown likely rises while the probability to throw a touchdown likely declines. This means that quarterbacks that complete longer passes on average are more likely to throw a touchdown than those that complete shorter passes on average due to team behavior in regards to calling running plays near the goal line. Again, this would be a latent variable issue where rushing ability of a team is common causing touchdown percentage and average yards per attempt. It may also mean that rushing percentage affects touchdown passing percentage more than the average yards per pass attempt.

Finally, an overall conclusion can be drawn from this causal analysis about valuations placed on these measures in the QB rating formula or any formula in general. It is important to account for the fact that these measures are not necessarily independent of one another. So for example, treating interception percentage and completion percentage as independent may unintentionally over estimate the contribution of interception percentage in the formula. This is because changes in *IP* cause changes in *CP* and so there is a partial double-counting effect. Further, in regards to running regressions to estimate the contributions of individual players, it is important that those regressions are of effects on causes not causes on effects.

Section 7, Conclusion and Future Research

In investigating the statistical relationships among QB rating variables we employed two tools: structural break tests and graph theory. Separately, both yield interesting conclusions. Combined, the structural break tests augmented the information from the graph theory analysis. We discuss those results and suggest some avenues for future research in this section.

The structural break tests results provide strong evidence that breaks exist in the QB rating and its underlying series. The results are suggestive of the fact that rules changes that affected the passing game in the 1970s and in the early 1990s, including the implementation of the salary cap, are likely sources of the breaks. This makes comparisons of quarterbacks in different break periods, in different regimes, inappropriate. In effect, it is an apples and oranges comparison if the rules changes are not accounted. If adjustments for the rules changes are made in a formula or in more formal statistical analysis, valid comparisons between quarterback play in the different regimes could be made and valid inferences drawn.

Another main result from the structural break tests is that the rules changes have led to rises in QB rating over time, especially driven by higher completion percentages and lower interception percentages. An investigation into the effects of the West Coast offense as either a cause or an effect (a response to the rules change) would be interesting. The 1970s break that leads to regimes with lower standard deviations generates an interesting implication. One possible interpretation is that while overall quarterback play, or if preferred, the overall passing game, as measured by these variables, rose over time; the spread between quarterbacks has been reduced. This implies that the relative difference in performance between a great quarterback and an average quarterback has shrunk. A further investigation into this possibility is warranted as teams try to find accurate measure of value of positions not just to each other but within position as well.

Turning to the causal analysis, we find evidence that over the course of a season the relationship between the QB rating variables is fairly different than on a single play. This is probably due in part to the altered temporal nature of the variables as they are aggregated over the course of a season as well as being defined relative to the pass attempt. The interpretation of the variables and their implications change as a result. Completion percentage over the course of a season contains different information than completion percentage over the course of a game or a completion on one particular play.

Coupled with information from the structural break tests, we find evidence of three links among the four QB rating variables. We find that interception percentage and average yards per attempt cause completion percentage over the

course of a season. We also find that touchdown percentage causes average yards per attempt. We offer some plausible explanations for these. However, further study needs to be completed to discern between these explanations and other possibilities.

Overall, to understand the causal relations or the dependency relations among variables is important when conducting statistical analysis or attempting to create accurate measures of performance. Future research to incorporate other aspects of play like rushing and defense, to incorporate some of the broader criticisms against the QB rating like yards after catch and QB rushing, and to examine the relationships over seasons, over games, and on individual plays, will all be important. This will add to our understanding not just of the QB rating, but also of the overall game play and how different aspects causally relate to each other. That understanding will hopefully help decision makers be better informed.

References

- Andrews, D.W.K. (1991) "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation", *Econometrica*, 59, 817-858.
- Andrews, D.W.K. (1993) "Tests for Parameter Instability and Structural Change with Unknown Change Point", *Econometrica*, 61, 821-856.
- Bai, J. and P. Perron (1998) "Estimating and Testing for Multiple Structural Changes in a Linear Model," *Econometrica*, 66, 47-78.
- Bai, J. and P. Perron (2003a) "Computation and Analysis of Multiple Structural Change Models", *Journal of Applied Econometrics*, 18, 1-22.
- Bai, J. and P. Perron (2003b) "Critical Values for Multiple Structural Change Tests", *Econometrics Journal*, 1, 1-7.
- Bai, J. and P. Perron (2004) "Multiple Structural Change Models: A Simulation Analysis", in Corbea, D., Durlauf, S., and B.E. Hansen (Eds.), *Econometric Essays*. Cambridge University Press.
- Berri, D. J., Schmidt, M. B., and S. L. Brook (2006) *The Wages of Wins: Taking Measure of the Many Myths in Modern Sport*, Stanford University Press, Stanford, CA.

- Bialik, C. (2008) "The NFL's Most Mysterious Number", *The Wall Street Journal*, Retrieved July 30, 2008, from <http://blogs.wsj.com/numbersguy/the-nfls-most-mysterious-number-255/>.
- Cooper, G. (1999) "An Overview of the Representation and Discovery of Causal Relationships Using Bayesian Networks", in Glymour, C., and G. Cooper (Eds.), *Computation, Causation, and Discovery*, American Association for Artificial Intelligence, Menlo Park, CA and MIT Press, Cambridge, MA.
- Demiralp, S. and K.D. Hoover (2003) "Searching for the Causal Structure of a Vector Autoregression", *Oxford Bulletin of Economics and Statistics*, 65, Supplement, 745-767.
- Garcia, R. and P. Perron (1996) "Computation and Analysis of Multiple Structural Change Models", *Review of Economics and Statistics*, 78, 111-125.
- Hoover, K. D. (2003) "Some Causal Lessons From Macroeconomics," *Journal of Econometrics*, 112, 121-125.
- Hoover, K. D. (2005) "Automatic Inference of Contemporaneous Causal Order of a System of Equations", *Econometrics Journal*, 2(2), 167-191.
- Hoover, K. D. and S. M. Sheffrin (1992) "Causation, Spending, and Taxes: Sand in the Sandbox or Tax Collector for the Welfare State", *American Economic Review*, 82, 225-248.
- Hoover, K. D. and M. Siegler (2000) "Taxing and Spending in the Long View: The Causal Structure of U.S. Fiscal Policy After 1791", *Oxford Economic Papers*, 52, 745-773.
- Liu, J., S. Wu and J. V. Zidek (1997) "On Segmented Multivariate Regressions", *Statistica Sinica*, 7, 497-525.
- Mushnick, P. (2007) "QB Ratings? Throw 'Em Out", *New York Post*, Retrieved July 30, 2007, from http://www.nypost.com/seven/09162007/sports/qb_ratings__throw_em_out.htm?page=0.
- NFL (2008) Retrieved July 30, 2008, from NFL.com - Official Site of the National Football League Web site: <http://www.nfl.com/>.

- Pearl, J. (2000) *Causality: Models, Reasoning, and Inference*, Cambridge University Press, Cambridge.
- Prodan, R. (2008) "Potential Pitfalls in Determining Multiple Structural Change Models with an Application to Purchasing Power Parity", *Journal of Business and Economics Statistics*, January, 26, 50-65.
- Sandomir, R. (2004) "Pro Football; The NFL's Passer Rating, Arcane and Misunderstood", *New York Times*, Retrieved July 30, 2008, from <http://www.nytimes.com/2004/01/14/sports/football/14RATE.html?ei=5007&en=bbd66a76d6e4af0f&ex=1389416400&partner=USERLAND&pagewanted=print&position=>.
- Spirtes, P., R. Scheines, C. Meek, T. Richardson, C. Glymour, H. Hoijtink and A. Boomsma, (1996) *Tetrad 3: Tools for Causal Modeling*, program and user's manual on the world wide web at <http://www.phil.cmu.edu/tetrad/tet3/master.htm>.
- Spirtes, P., C. Glymour and R. Scheines (2001) *Causation, Prediction, and Search*, MIT Press, 2nd Edition, Cambridge, MA.
- Steinberg, D. (2001) "How I Learned to Stop Worrying and Love the Bomb: A Survival Guide to the NFL's Quarterback Rating System", *Gentleman's Quarterly*, October.
- Steelers Fever (2008) "History of NFL Rules", Retrieved July 30, 2008, from Steelersfever.com Web site: http://www.steelersfever.com/nfl_history_of_rules.html.
- Swanson, N.R. and C. W. J. Granger (1997) "Impulse Response Functions Based on a Causal Approach to Residual Orthogonalization in Vector Autoregressions," *Journal of the American Statistical Association*, 92, 357-367.