Assessment of Soccer Referee Proficiency in Time-Sensitive Decision-Making

Nathan Jones, Andrew Cann, Saud Almashhadi, Hina Popal

Abstract - **Soccer is the world's most popular sport. The success of the sport is dependent on referees making correct calls and maintaining game integrity.**

Referee performance is determined by attributes: game flow understanding, fitness, and call decision making. Evaluating referee quality using these attributes is vital to improving on-field performance.

MDCVSRP oversees all referees within Virginia. Because MDCVSRP has no program to assess junior referee fitness or game flow understanding, it is unable to identify high quality referees within its system. A program is needed to predict the call accuracy of all junior referees within MDCVSRP based on their Fitness and/or game flow understanding attributes.

Four alternative programs were identified to assess the quality of junior referees within MDCVSRP: 1) Baseline Physical Fitness Tests assessing referee fitness attributes; 2) Game Flow Evaluations assessing game flow understanding attributes; 3) Combined Evaluations assessing both fitness and game flow understanding attributes; and 4) No Assessments conducted for fitness or game flow understanding (status quo).

A two part analysis was conducted to determine the utility of each program. Part I used a discrete event simulator to quantify the effects of fitness and game flow understanding on call accuracy. Using this analysis, part II allocated utilities to each program based on attributes assessed. Analysis of part I concluded that call accuracy varies nonlinearly with both fitness and game flow understanding. Part II concluded that the Fitness Test (0.749) had the highest utility followed by Combined Evaluation (0.742), Game Flow Evaluation (0.727), and No Assessment (0.721).

Based on a cost benefit analysis, it was determined that the benefit of implementing any program to assess the fitness and/or game flow understanding of junior referees is outweighed by cost. Therefore, it is recommended that No Assessments be conducted for fitness and/or game flow understanding on junior referees within MDCVSRP.

I. Introduction to Referees

SOCCER is the world's most popular sport. Between 2009 and 2010, European soccer alone generated 16.3 billion euro in revenue [1]. Much of soccer's recent success can be attributed to improvements in viewer experience. Fans can now watch games from multiple angles and view replays of key events. However, this improvement in broadcasting has had indirect repercussions, particularly with regards to fan perception of game integrity.

In soccer, the administration and integrity of the game is overseen by one main referee and two assistant referees. The purpose of the referee is to make accurate calls on the field, to administer penalties when needed, and to ensure that calls do not interrupt the flow of the game. Most importantly, referees are responsible for instilling in fans a belief that the game is fair and impartial.

The governing bodies of soccer have been mostly unwilling to implement referee support technology, such as replays, for fear that it will interfere with game flow [2]. Thus, as the quality of soccer broadcasting has improved, the tools available to the referee have remained the same. This imbalance of technology between referees and fans has lead to an asymmetry in information where fans often have better information for judging the accuracy of a call than the referees on the field. This allows fans to easily identify injustices in the administration of the game, and has caused backlashes against the sport when incorrect calls alter the outcome of a match [2]. Therefore, poor referee performance can be considered one of the greatest threats currently facing the sport of soccer.

II. Organization of American Referees

Within the United States, soccer referees undergo a structured training and evaluation process. This process is broken into eight levels of seniority (grades) in which grades 8-7 represent entry level referees, 6 -5 contain state referees, 4-3 comprise national referees, and 2-1 are reserved for FIFA international referees [3,4]. Grade 8 referees are typically referred to as "junior" referees where referees in grades 7 – 1 are referred to as "senior" referees. Progression of referees beyond grade 8 is voluntary and requires classes, written examinations, fitness tests, and game performance evaluations. A referee's grade determines the level of game he is recommended to officiate [4].

The United States Soccer Federation (USSF) oversees all referees in grades 4-1 where those in grades 8-5 are overseen by state level referee organizations [4]. The state

Manuscript received April 2, 2012. This project was sponsored by the Metro DC Virginia State Referee Program (MDCVSRP). All authors are students at the Volgenau School of Engineering, Dept. of Systems Engineering and Operations Research, George Mason University, Fairfax, VA 22030 USA.

N. Jones (email: njonesh@gmu.edu), A. Cann (email: acann@gmu.edu), S. Almashhadi (email: s3ood@gmail.com), and H. Popal (email: hpopal@gmu.edu).

level organization within the Commonwealth of Virginia, the Metro DC Virginia State Referee Program (MDCVSRP), serves as the sponsor for this project. Within MDCVSRP, 96.8% of referees reside within grade 8 while the remaining 3.2% of referees are distributed over grades 7-1 [5].

The success of efforts to improve on-field performance hinges on an ability to evaluate referee quality. Evaluating referee quality is key to progressing referees to more senior grades and properly assigning referees to games [4, 6].

III. Referee Call Making Process

A referee's quality is defined as the percent of correct calls during games. A referee's call accuracy is dependent on how effectively he is able to carry out a standard decision making process whenever a call event is triggered (Fig. 1).

When a call event occurs, a referee must visually recognize that a decision needs to be made through a *Sensory Function.* Once an event is detected, a referee must make an accurate decision regarding the nature of the call (infraction, no infraction) using a mental model of what occurred in the event and knowledge of the laws of soccer. This process is combined into a function known as *Cognition for Making Calls* and determines a referee's call accuracy.

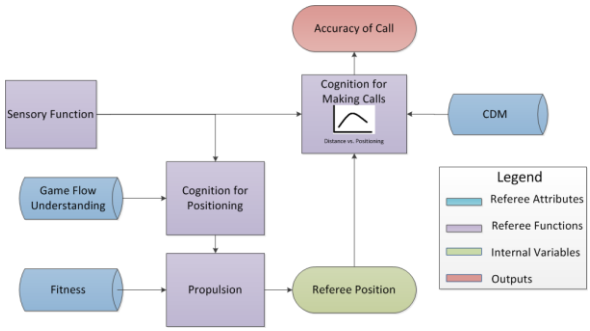

Fig.1. Referee call making process

The ability of a referee to make correct calls through the *Cognition for Making Calls* function is dependent on a referee's distance from the call, which is determined through the interaction of two functions. The first function, *Cognition for Positioning*, defines a referee's ability to choose an optimal position to make calls. This function is dependent on a referee's *Sensory Function* and mental model of game flow. The second function, *Propulsion*, is a physical function determining a referee's ability to move to the position identified during the *Cognition for Positioning* function in a time effective manner.

The ability of a referee to carry out the *Cognition for Making Calls*, *Cognition for Positioning*, and *Propulsion* functions is assumed to depend on three attributes. Game flow understanding (GFU) is the ability of a referee to perform *Cognition for Positioning*. Fitness is the ability of a referee to carry out *Propulsion*. Call decision making (CDM) is the ability of a referee to carry out *Cognition for Making Calls.*

IV. Evaluation of Referee Quality

Evaluating referee quality focuses on assessing referees in terms of their GFU, fitness, and CDM attributes.

GFU is currently evaluated indirectly through annual on field assessments conducted by official assessors for all referees grades 7 - 1 [4]. Fitness is evaluated through various fitness tests including a series of sprints and long distance runs. They are administered annually to referees grades 7 - 1 [4]. CDM is evaluated through written examinations administered to all referees and annual on field assessments for referee's grades 7-1 [4].

This current assessment methodology has significant gaps in assessing referees based on attributes. In particular, referees in grade 8, which account for the vast majority of referees within the Commonwealth of Virginia, do not receive any evaluations for GFU or fitness [5].

V. Need Statement

An evaluation system is needed to predict the quality (call accuracy) of grade 8 referees overseen by MDCVSRP based on their fitness and/or GFU attributes.

VI. Design Alternatives

Four evaluation system concepts have been identified to assess the quality of grade 8 referees. The specifics of design and implementation of these concepts are considered outside the scope of this project. The cost of each alternative is defined as the investment necessary to purchase required physical resources and carry out a one-time quality evaluation for all grade 8 referees. Three alternatives would involve a Baseline Fitness Test and a Game Flow Evaluation, either singly or in combination. A fourth alternative would involve no testing (status quo).

A. Baseline Fitness Test

A baseline fitness test would be administered to all grade 8 referees within MDCVSRP at an estimated cost of $26,990. The results of the baseline fitness test would be used to assign each referee a fitness attribute rating as a means of assessing overall quality. This would be the same fitness test currently administered to referee grades 7 –1.

B. Game Flow Evaluation

A video recording would be made of each referee's in-game performance. These videos would be transmitted to official assessors who would review the footage and assign each referee a GFU rating using expert opinion. An assigned GFU rating would be taken as a means of assessing overall referee quality. This test would be a

video based version of the same evaluation currently administered to referee grades 7 – 1. An evaluation of all grade 8 referees in this fashion would cost an estimated $337,995.

C. Combined Evaluation

Both a baseline fitness test and game flow evaluation would be used to assign each grade 8 referee fitness and GFU ratings as a means of assessing overall quality. Evaluating all grade 8 referees in this fashion would require an estimated cost of $341,870.

D. No Assessment

Under this alternative, no assessment is conducted to assess the GFU or fitness attributes of referees. This alternative exists as a point of reference against which to compare the cost and benefit of the three preceding alternatives and represents the status quo for assessments at the grade 8 level requiring no implementation cost.

VII. Evaluation of Alternatives

A two part analysis has been conducted to select the most beneficial evaluation system for grade 8 referees.

Part I utilizes a stochastic discrete event simulator modeling a referee's ability to position and make calls based on fitness and GFU attribute levels (Fig. 2). Through performance evaluation of 25 referee profiles defined as combinations of fitness and GFU attributes (scaled 0 – 100), the simulator is used to generate a regression equation describing the impact of fitness and GFU on a referee's call accuracy.

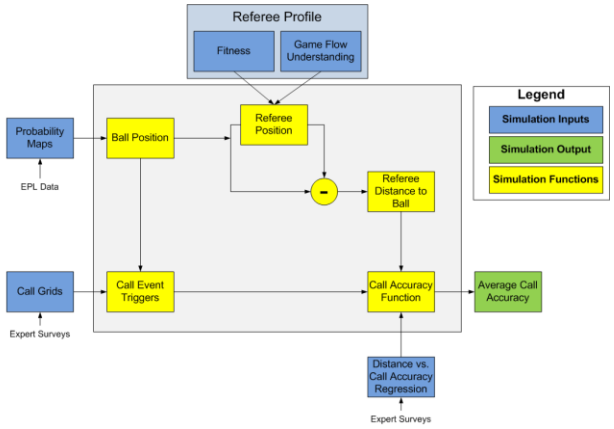

Fig.2. Discrete soccer game simulator design

Part II consists of a Monte Carlo analysis in which 5000 referees are randomly generated with independent fitness and GFU attribute levels. Utilizing the regression from part I, call accuracy is calculated for each of these referees. The utility of the No Assessment alternative is defined as the mean average call accuracy of referees within this pool over 30 scenarios. Each remaining evaluation program is used to identify the top 100 referees for each of 30 scenarios. The mean average call accuracy of these 100 referees is used to represent the utility for each alternative.

A. Part I: Discrete Soccer Game Simulator

The stochastic soccer game simulator divides a soccer field into a fine set of 8510 square cells where each cell represents a 1x1 yard area. Each of these cells is allocated to 1 of 60 movement polygons (Fig. 3) and 1 of 24 call grids (Fig. 4). Throughout a 90 minute simulated game, the ball moves from cell to cell adhering strictly to a play cycle of four events. This cycle begins with a pass reception (0.5s) and local dribbling (4.5 s) in which the ball moves within its current polygon. This is followed by either a shot on goal (0.5s) or a pass (0.5s). If a pass, the ball will move to its reception location over a period of time depending on the distance traveled. This play cycle repeats until the simulation terminates.

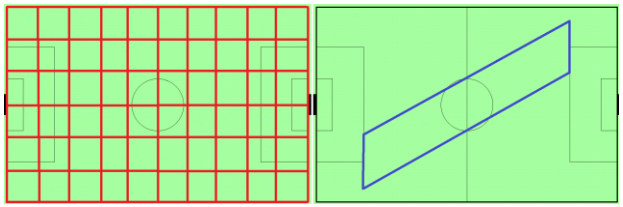

Fig.3. Red cells (left) are ball movement polygons; Blue area (right) is referee movement region.

As the ball moves throughout the play cycle, it refreshes its position every 0.5 seconds of simulated game time. At any instant, the ball is possessed by one of two teams, each executing its own unique strategy. For each team, a set of probability maps represents that team's strategy and style of play. For each of the 60 polygons, these maps specify the probability that the ball moves to any other polygon or is shot at the goal. A further dimension of the map indicates probabilities that a pass or shot is successful.

Based on data collected from 80 English Premier League games, probability map sets were formulated for 4 teams: Wigan, Manchester United, Arsenal, and Stoke. These teams were chosen to give a broad representation of different play styles and enable the simulator to replicate a vast number of game flow situations.

Using pass completion as a metric representing team strategy, an ANOVA analysis was conducted (Table I) to determine if teams changed their strategy based on score differential (ahead, behind, tie) or elapsed game time (divided into discrete 15 minute time periods) . The results of this analysis were used to determine how many probability maps were needed to encapsulate each team's strategy and when maps should be changed, based on situation, to reflect strategy alterations. It was concluded that Arsenal utilizes a single probability map for all game situations. Stoke, Manchester United, and Wigan utilize six probability maps representing situations where the team is ahead, behind, or tied in the first and second half respectively.

TABLE I
EFFECT OF GAME SITUATION ON PASS COMPLETION

Situation	Arsenal	United	Stoke	Wigan
Time	p = 0.777	p = 0.142	p = 0.001	p = 0.001
Score	p = 0.231	p = 0.001	p = 0.000	p = 0.000
Time*Score	p = 0.338	p = 0.000	p = 0.000	p = 0.116

Upon concluding the dribbling event in the play cycle, the probability that the ball is passed (versus shot at the goal) depends on game situation as determined using the active probability map of the team with possession. If a pass occurs, the active map indicates the destination polygon of the pass and chance of success. If a shot occurs, the active map indicates the probability that the shot will result in a goal. Executing passes and shots in this fashion allows the simulator to accurately represent the flow of a soccer game in which a referee must interact.

To ensure the timing of ball movement accurately represents that of a soccer game, whenever the ball is being dribbled or passed, a single destination cell is set. The ball moves to that destination in a straight line trajectory which it follows for a duration of simulated game time (1).

$$Travel\ Time\ (s) \cong \frac{\sqrt{(Start\ x - Finish\ x)^2 + (Start\ y - Finish\ y)^2}}{Speed\ (\frac{yds}{s})} \quad (1)$$

In the simulation, a single referee is modeled running within a standard diagonal system of control 2-dimensional area (Fig. 3). The speed of the referee is calibrated to represent the fitness level of the referee profile being tested.

Every 0.5 seconds, the referee sets his desired position using one of two movement scripts. In script I, the referee sets his destination to the closest cell within 11 – 13 yds of ball's current location. In script II, the referee sets his destination to closest cell within 11 – 13 yds of next most probable pass destination as determined using the active probability map. Upon setting his destination using script I or II, a referee will begin moving towards his destination using the same straight line movement algorithm described previously for ball movement (1).

At the beginning of each play cycle, the probability that the referee utilizes script II is determined by the referee's GFU level (higher GFU yields higher probability). Furthermore, this same GFU probability is used to determine the likelihood that if a call will occur in the current cycle, the referee will recognize the buildup to the call and switch to script I until the call transpires.

At the beginning of each play cycle, the ball location is used to reference a set of probabilities indicating probability that the referee will need to make a call in that cycle (Fig. 4). These probabilities were developed using an expert survey administered to 16 senior referees within MDCVSRP and tailored to ensure that roughly 65 call events occur per game. Data from the survey were also used to determine the probability of the call event occurring at the receiving (0.21), dribbling (0.44), passing (0.21), or pass en route (0.15) events of the play cycle.

0.121	0.061	0.023	0.042	0.041	0.073
0.037	0.288	0.084	0.102	0.218	0.155
0.052	0.193	0.095	0.125	0.291	0.023
0.274	0.086	0.041	0.024	0.045	0.062

Fig.4. Call event probabilities based on field location

When a call event occurs, the probability of a correct referee decision is determined based on the referee's distance to the ball (assuming calls occur at ball location). Using the MDCVSRP senior referee survey, a regression was performed relating the probability of making a correct call to a referee's distance from the call (Fig. 5). The average standard deviation for the 12 distances polled on the survey was 21.3%, indicating disagreement among participants.

Over the course of the simulated game, the call accuracy of a referee is defined as the number of correct calls divided by the total number of calls made.

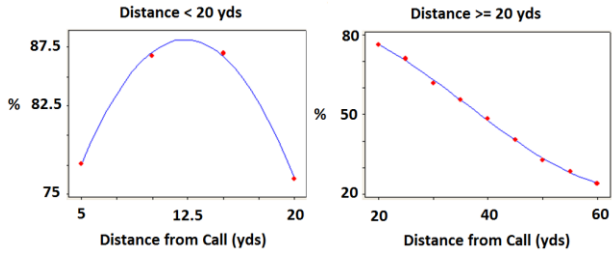

Fig.5. Call accuracy vs. distance from call

To determine the impact of fitness and GFU on call accuracy, each of 25 distinct referee profiles representing different combinations of fitness and GFU (scaled from 0 – 100) was simulated through 2000 games representing 200 games for each combination of the Arsenal, Manchester United, Stoke, and Wigan play styles. Referee speeds corresponding to profile fitness ranged linearly from 2.023 yds/s at fitness = 0 to 3.911 yds/s at fitness = 100. Probabilities corresponding to profile GFU ranged linearly from 0.25 at GFU = 0 to 0.90 at GFU = 100. The average call accuracy for each profile over the simulated games was used to formulate a multivariate regression for call accuracy as a function of fitness and GFU level.

B. Monte Carlo Analysis

For each Monte Carlo scenario, 5000 referees are randomly generated under the assumption that each referee's fitness and GFU levels are uncorrelated and represent independent draws from normal distributions (mean 50, standard deviation 15). Using the regression from the part I analysis, average call accuracy for each

referee was determined. Using normal CDFs to ensure selection of roughly 100 top referees, fitness and/or GFU cutoffs were defined for the first three alternatives based on attributes evaluated (Table II). If a generated referee met all of the cutoff criteria for an alternative, he would be selected by that alternative as one of the top 100 referees. The mean average call accuracy of selected referees over 30 scenarios was used to define the utility of these alternatives. The utility of the No Assessment alternative was defined simply as the mean average call accuracy of referees within each pool. This analytical method assumes that each alternative has an idealized ability to evaluate the attributes assessed.

TABLE II
DESIGN ALTERNATIVE ATTRIBUTE CUTOFFS

Alternative	Attributes Evaluated	Cutoff	Average # Referees Selected
Fitness Test	Fitness	Fitness > 81	97
Game Flow Evaluation	GFU	GFU > 81	97
Combined Evaluation	Fitness, GFU	Fitness > 66 GFU > 66	102
No Assessment	N/A	N/A	N/A

VIII. RESULTS

A. *Discrete Soccer Game Simulator Results*

Analysis of each of the 25 referee profiles over 2,000 simulated games yielded results for average call accuracy as a function of fitness and GFU (Fig. 6). Fitness and GFU levels are scaled where a rating of 0 is the worst possible and 100 the best possible. Across the referee profiles, call accuracy ranged from 71.22% to 75.67%. Over the 25 profiles, the average 95% CI half-width for mean call accuracy was $2.866e^{-3}$. This indicates an acceptable level of confidence in the data points.

A multivariate regression for call accuracy was computed with an R-squared value of 99.51% representing a strong fit (2).

$$\textbf{Accuracy (Fitness, GFU)} = 0.713491 + 0.000923486\ \text{Fitness} + 1.28791e^{-5}\ \text{GFU} - 6.4846e^{5}\ \text{Fitness}^2 + 1.12504e^{-6}\ \text{GFU}^2 + 1.26193e^{-6}\ \text{Fitness}^3 - 6.75305e^{-9}\ \text{Fitness}^4 \quad (2)$$

The regression analysis indicates that accuracy varies nonlinearly with fitness and GFU. Adding polynomial terms for fitness and GFU until p-values for leading terms jumped above acceptable levels ($p > 0.05$) resulted in a fitness degree of 4 and GFU of degree 2 (Table III). The generated regression does not include an interaction term between fitness and GFU, since adding an interaction term resulted in a p-value of 0.813.

B. *Monte Carlo Analysis Results*

The Monte Carlo analysis implied that the most effective evaluation method was the Fitness Test followed by the Combined Evaluation, Game Flow Evaluation, and No Assessment (Table IV). 95% confidence interval half widths indicate a high level of confidence in the results (Table IV).

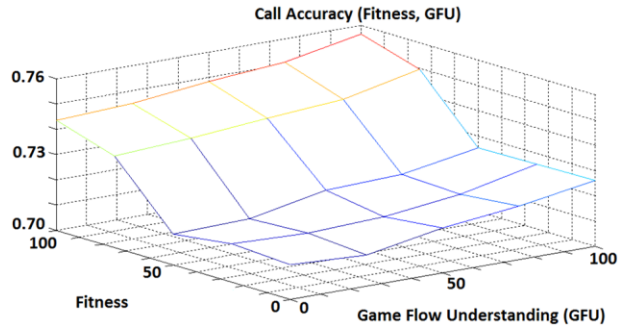

Fig.6. Analysis of 25 profiles: Call Accuracy (Fitness, GFU)

TABLE III
CALL ACCURACY REGRESSION ANALYSIS

Term	T	P - value
Constant	1042.55	0.000
Fitness	7.84	0.000
Fitness^2	-10.97	0.000
Fitness^3	13.35	0.000
Fitness^4	-14.36	0.000
GFU	0.55	0.590
GFU^2	4.99	0.000

TABLE IV
UTILITIES FOR GRADE 8 EVALUATION ALTERNATIVES

Alternative	Average Call Accuracy	95% Half-Width Call Accuracy
Fitness Test	0.74926	0.00012
Game Flow Evaluation	0.72693	0.00028
Combined Evaluation	0.74174	0.00021
No Assessment	0.72099	0.00004

IX. UTILITY/COST ANALYSIS AND RECOMMENDATION

Based on a cost vs. utility analysis conducted on alternatives (Fig. 7) it can be concluded that the Fitness Test dominates both the Combined Evaluation and Game Flow Evaluation due to its higher utility and lower cost. Therefore, the choice of alternatives lies between conducting a Fitness Test at grade 8 (74.9% Accuracy, $26,990) or conducting no assessments at this level (72.1% Accuracy, $0). As the average accuracy of the top 100 referees selected using the Fitness Test exceeds the overall referee accuracy by only 2.8 percentage points, the improvement in selection due to implementing the Fitness Test over the status quo can be considered statistically but not practically significant. Thus, the benefit of implementing Fitness Tests for all grade 8 referees is outweighed by its cost. It is therefore the recommendation of this project that the status quo be maintained and no referee evaluations be conducted for fitness and/or GFU at the grade 8 level.

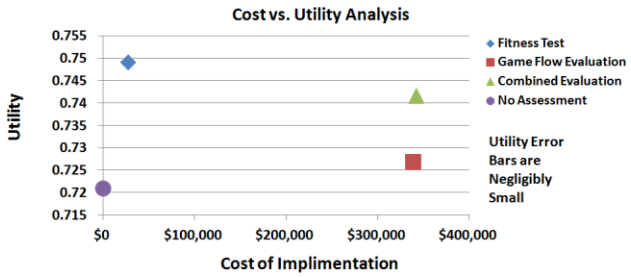

Fig.7. Cost vs. utility analysis for alternatives

X. Additional Findings

To determine the effect of game flow on a referee's call accuracy, an analysis was conducted on the extent to which referee call accuracy was affected by the playing styles of the teams competing in a game (Fig. 8). Based on the range of performance from best performing referee profile to worst performing profile (indicated by error bars), it can be concluded that team playing styles can have a significant impact on referee performance.

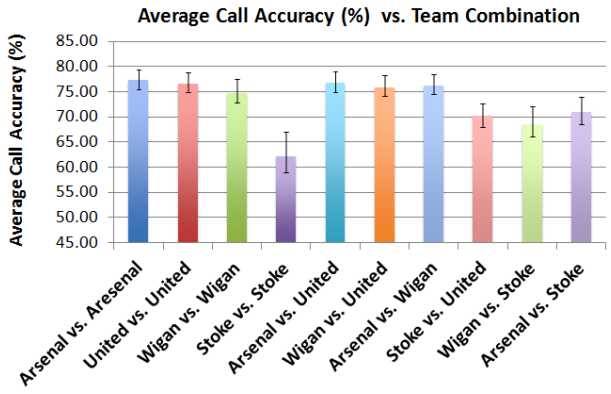

Fig.8. Impact of team combinations on referee call accuracy

Given this finding, further analysis was conducted to determine why certain team combinations result in decreased referee performance. 500 simulated games were run using a single referee profile (fitness = 50, GFU = 50) for United vs. United and Stoke vs. Stoke play styles. Over these games, the simulated referee made roughly 30,000 calls for each team combination. For all call events, the distance from the call was recorded and analyzed.

It was concluded that differences exist in the distributions of call distance as a result of team play styles. United vs. United games resulted in density concentrating heavily around 11-13 yards and decreasing consistently with further increases in distance (Fig. 9). However, Stoke vs. Stoke games resulted in a bimodal density concentrating around 11-13 yards and again at around 44-47 yards (Fig. 9). This second peak along with the increased density between peaks accounts for the decreased referee performance in Stoke vs. Stoke games due to the referee more frequently being out of position to make calls. This analysis shows that the same referee, when placed in two different games, can have a decreased performance and be out of position far more often in one game due exclusively to different team combinations and their effect on game flow.

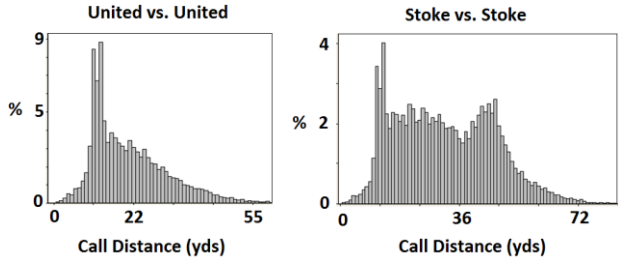

Fig.9. Distance from calls for United vs. United and Stoke vs. Stoke

The results of this finding are nontrivial and point to a key characteristic that is currently lacking in referee criticism and evaluation. When assessing the in-game call performance of a referee, the difficulty of the match being officiated (in terms of game flow) must be taken into account due to its large and unavoidable effect on call performance. Furthermore, when comparing the performance of different referees, the games in which referees are evaluated must be synchronized to ensure that team combination does not act as a confounding variable in the analysis.

Acknowledgment

The basic concept for the stochastic soccer game simulator was derived from a previous George Mason University student project [7]. This concept consisted of probability guided ball movement over a soccer field grid where a modeled referee would position and respond to randomly generated call events [7]. From this initial concept, the simulator used in this project was re-designed and coded independently of past work.

References

[1] C. Gordine-Wright, Z. Reilly, (2011, June 9) European football market grows to €16.3 billion. [online]. Available: http://www.deloitte.com/view/en_NL/nl/7fdea05260570310VgnVCM2000001b56f00aRCRD.htm

[2] J. Wilson. (2010, June) Soccer could use instant replay, but not at expense of sport's flow. [online]. Available: http://sportsillustrated.cnn.com/2010/soccer/world-cup-2010/writers/jonathan_wilson/06/28/soccer.technology/index.html

[3] (2003, April) United States Soccer Federation Referee Grades. [online]. Available: http://www.pawestsoccer.org/Assets/documents/Announcement forgradechanges.pdf

[4] Definitions of Referee Grades [online]. Available: http://www.vadcsoccerref.com/docs/DEFINITION%20OF%20REFEREE%20GRADES.pdf

[5] Pat Delaney (2011, January 3) Annual Assessor, Assignor, Instructor and Administrator Meeting. [Presentation].

[6] Pat Delaney (2011, November 10) MDCVSRP Sponsor Meeting [Verbal]

[7] A. Solomon, A. Paik, T. Phan, A. Alhauli, (2011) A Decision Support System for the Professional Soccer Referee in Time-Sensitive Operations. [online]. Available: http://catsr.ite.gmu.edu/SYST490/DSTSO_IEEE_SIEDS.pdf