Application of the Hungarian Algorithm in Baseball Team Selection and Assignment

S.S. Britz and M.J. von Maltitz

Keywords: Hungarian algorithm, Kuhn-Munkres algorithm, team selection, position assignment, baseball team

University of the Free State
Department Mathematical Statistics and Actuarial Science

Abstract

Measuring and comparing the performance of athletes in all forms of team sports require more than a modest amount of data reflecting the athletes' abilities. When investigating the skill of baseball players in South Africa, however, one has to make do with a combination of personal opinion, gut feeling and a few descriptive statistics.

This research is aimed at developing a system with which to assess the abilities of baseball players in all practical aspects of the sport and compose a team in which all the players are assigned to positions such that the collective team skill is maximised for a specific goal. The system itself is of such a nature that any group of baseball players can apply it, with some help from a statistician.

The optimisation process that selects the team is carried out using the Hungarian algorithm, a mathematical method of optimally assigning a set of persons to a set of jobs.

Table of Contents

1 Introduction **5**

2 Literature Review **9**

2.1 Team Selection Strategies 7

2.2 The Development of the Hungarian Algorithm 9

3 Methodology **13**

3.1 The Baseball Aptitude Tests 11

3.1.1 Hitting Test 12

3.1.2 Speed Test 12

3.1.3 Pitching Test 12

3.1.4 Throwing Tests 13

3.1.5 Fielding Tests 13

3.1.6 Reach Test 14

3.1.7 Reaction Test 14

3.2 Data Transformation Procedures 15

3.2.1 Success Percentages 15

3.2.2 Less is More 16

3.2.3 The Reach Test and Hitting Test 17

3.2.4 Relative Scores 18

3.3 Dimension of the Data 19

3.4 Introducing Weights 20

3.5 How the Hungarian Algorithm Works 22

4 Application **27**

4.1 Testing for Multidimensional Data 27

4.2 The Optimal Baseball Team 29

4.2.1 Optimal Defence 30

4.2.2 Medium Offence 34

4.2.3 Strong Offence 38

4.2.4 Maximum Aggression 39

4.3 Introducing Relief Pitchers 42

4.4 An Overview of the Team Selections 45

5 Conclusion **49**

5.1 Further Research 47

Chapter 1 - Introduction

The sport of baseball, also referred to as 'the great American pastime' (Bergan, 1982), has been steadily growing in popularity worldwide for the past few decades. The Olympic Games first recognized baseball as an official sport at the 1992 Summer Olympics. The sport was played at seven previous Olympic Games stretching as far back as 1912, but only as a demonstration sport.

Rising interest and newly participating nations have led to the creation of the World Baseball Classic, equivalent to the World Cup in other sports. The establishment of this tournament in 2006 has indeed helped to put baseball on the world map and change the misconception that the sport is only played in the USA. After the second of these championships in 2009, Schlegel (2009) stated that, "…just as [the World Baseball Classic] did in 2006, it showed how large the world of baseball is becoming." The 16 participating teams represented 6 different continents including Africa, with South Africa taking part in both competitions but failing to make it past the first round on both occasions. South Africa has had no significant success on the international baseball circuit other than qualifying for the 2000 Sydney Olympics, where they failed to win a single match in the round robin stage and subsequently finished last.

Though there are many factors influencing the quality of performance in baseball, team selection arguably plays the pivotal role in obtaining positive match results. Tidhar, *et al*. (1996: 369) states that "[t]eam selection, the process of selecting a group of agents with complementary skills to achieve a goal, is an important collaborative task in multi-agent systems". However, simply choosing the most skilled group of players for the task at hand does not automatically ensure that the team is as optimally designed as can be. In a sport like baseball, where each position requires a player with very specific skills, not only the selection of the team, but also the positioning of each individual is of the utmost importance (Baeva *et al*., 2008).

The aim of this research is to create a system of assigning the most appropriate player to each position on the field in order to design the optimal team, *i.e.* the combination of players that possesses maximal common efficiency. Note that the term “efficiency” in context of this paper does not refer to efficiency in the classical statistical sense, but rather as a measure of how much of the team’s ability is utilised through the chosen selection.

South Africa is still finding its feet in terms of establishing a system through which talented baseball players can be identified and then be exposed to competitive play. In my opinion the fastest way of developing this process and giving the players this essential exposure is to create opportunities for these athletes to test their skills against the toughest competition the region can provide within their league. The emphasis should, however, be on creating these scenarios for teams on any level, not just teams competing in official leagues. Not only will the traditional statistics gathered on the players more closely resemble their true potential, but this competitive play will also give the players valuable match experience, which makes the transition to higher levels of competition much less disconcerting.

Therefore, our aim should be to create a system of selecting the optimal team that can be easily applied to even recreational teams, but is, at the same time, powerful enough to distinguish between two or more closely matched players vying for the same position at any competitive level.

An effective selection system also directly improves training productivity. Once every player’s strengths and weaknesses in terms of their positional responsibilities are known, these specific skills can be developed. Due to the practicality of the proposed system each player can be constantly monitored and their abilities continuously assessed. The final result of this analysis is that training sessions can be optimised in order to refine the athlete’s expertise in a shorter space of time. For an in-depth look at the impact that sport-specific training has on expert decision-making and an athlete’s level of expertise, see Baker *et al.* (2003).

As mentioned above, the goal of the selection system is not necessarily to assign to each position the best available player for that role, but rather to choose a team optimized in its entirety for a specific task. A constant compromise between positions has to be considered to balance the team as much as possible with the available group of players. The computational implication of this compromise is that optimal efficiency is not obtained through simply calculating a single score statistic for each individual and selecting a team comprised of the nine highest scores.

One must rather create a set of required abilities covering all aspects of baseball athleticism and assign scores to every player with regard to each of those specific skills. Once this information is available the appropriate player for each position can be identified by deciding which combination of the tested abilities is most significant with respect to said position and assessing each player's combined score for this selection of abilities.

The intuitive solution to this assignment problem is to select at each position the player with the highest aforementioned score. However, this paper shows that this approach does not necessarily (and, in fact, rarely does) guarantee a team selection yielding the optimum overall efficiency.

Methods of obtaining the players' test scores and identifying the optimal combination on the basis of this information is explored in detail in this paper. Although the aim of the application is on novice baseball players, the system is by design of such a nature that, should it need to be altered to accommodate for factors that only arise at more specialised levels of competition, this adjustment can be done by simply changing the appropriate tests, combination functions or weights.

An important consideration to bear in mind is the balance required within a baseball team in terms of offense and defence. With this in mind, this paper considers different weighted combinations of the roles that a baseball team is required to fulfil on the field.

After the players' test scores are processed and the desired weights are applied to these scores, the research objective can be met by solving the resulting set of values for optimum efficiency. The method used for solving the assignment is the Hungarian algorithm, also known as the Kuhn-Munkres algorithm or Munkres assignment algorithm. This simple, yet powerful algorithm ensures optimal assignment of the players with regards to the chosen roles in the different positions. Sensitivity analysis is done in the application section by adjusting the weights applied to the scores and using alternate methods of processing the scores obtained from the tests.

An outline of the structure and content of this paper is as follows:

2. Literature Review

   The review mainly focuses on the aspects of baseball important to this research and team selection strategies that have been suggested and used. The development and usage of the Hungarian algorithm is also discussed.

3. Methodology

   The methods used to gather the data, *i.e.* the skills tests run on the players, are described in detail in this section. The processing of this data is also considered. After explaining how to compile efficiency matrices with the help of weights we then explore the steps involved in applying the Hungarian algorithm to square and rectangular matrices.

4. Application

   In this section the results of the skills tests are given and the processing of the data begins. After different methods of computing the final score matrix are considered in terms of their applicability, one is chosen and first analysed through Principal Components Analysis to test for multidimensionality in the data. Application of the system is carried out using a wide range of sets of weights and team combinations.

5. Conclusion

   A summary of the results and their implication are given in this section. We also look at the shortcomings of the system and discuss in which way the system can be developed and improved in future.

Chapter 2 - Literature Review

The main goal of this research is to create a team selection system that is both effective and easily applied to any standard of baseball competition. In order to attain this goal one needs a measure of each player's ability with regards to certain aspects of baseball.

The first section in this chapter explores the details of the sport in terms of its positional requirements, which lays the foundation of understanding upon which the assessment of players is built. Modern team selection strategy in baseball is also discussed.

The literature review then gives background information regarding the Hungarian algorithm, the iterative solution used to finally optimize the team.

2.1 Team Selection Strategies

Berger & Deely (1988) notes that statistical methodology has been used (by now) for more than half a century to develop selection and ranking procedures. The procedures developed have been applied in a wide variety of fields and disciplines in which a subset of elements needs to be selected form a larger group. Berger & Deely (1988) also apply their methodology, a Bayesian approach, to the matter of ranking and comparing baseball players.

Selecting a team of agents to attain a predetermined goal involves a process of calculations which is guided by the specification of roles and allocations (Tidhar *et al*., 1996). These roles can either be directly determined by the system itself (as is the case in baseball) or in more complicated systems these roles might be determined by a field expert to introduce a clearer definition of requirements. The allocations require problem-specific solutions.

When considering the selection of a baseball team, the roles to be fulfilled are simply the different positions in the team. To better understand the relative importance of the different roles and place them in context, let us first discuss the ultimate collective goal of the team. The following is a layman's explanation of the basic concept of baseball:

"Two teams of nine players each compete [against] each other in this game. The goal of the sport is to score runs by striking a thrown ball with a bat and hitting a series of four bases sorted at the [corners] of a 90 feet square, or diamond. In this sport, each player of the [batting] team has to take turns striking against the pitcher of the other team who tries to stop them from making runs by getting hitters out in any of several ways." (Washingtonprepsports.com, 2010)

The first aspect of the sport to note is that there is a compromise between batting (referred to as offence) and fielding (defence) capability. Should one choose a team in which each player excels in batting but has no defensive skills whatsoever, dismissing the batting team would be nigh on a miracle. The converse also causes a problem, as a team of players all trained purely as fielders but incapable of laying bat on the ball would not score any runs.

Not only should one create a certain balance within the team with regards to offence and defence, but (more importantly) one must also choose players who in turn possess both the basic necessary skills. In other words, each position should be filled by a player with both adequate batting ability and fielding aptitude for that specific position.

The most accurate judgement of a player's suitability for a position is through the assessment of his performance in match situations. However, this requires a fairly large amount of data from recorded matches. In both the major leagues in the USA, namely the American League (AL) and National League (NL), each team plays 162 matches in the regular season. The task of comparing different players is simplified considerably by the immense amount of statistics gathered on each player every year. The result of this abundance of information is that team selection in the major leagues is mostly an economically-driven process in the modern era. For information on bargaining and arbitration in Major League Baseball and the impact that free agents have on the labour market in baseball, see Faurot (2001) and Raimondo (1983) respectively.

This paper attempts to create a system of selecting a team even with the absence of any match data whatsoever. If such information does exist it can be incorporated to make further adjustments to the team if necessary. The method used in this paper is largely based on the work done by Baeva *et al*. (2008). The methodology is explained in detail in the next chapter.

2.2 The Development of the Hungarian Algorithm

The fact that statistical and mathematical methodology have contributed enormously to selection and ranking procedures is undisputed. In this modern day and age optimisation in practically every facet of society is all-important. Luckily, many engineering problems can be simplified to what is referred to as assignment problems, with application in subjects ranging from time scheduling to satellite communication (Martello, 2010).

Harold W. Kuhn was the first mathematician to formally define a solution to the problem of assignment. In Kuhn's much celebrated 1955 publication he defines the personnel-assignment task as a problem "ask[ing] for the best assignment of a set of persons to a set of jobs, where the possible assignments are ranked by the total scores or ratings of the workers in the jobs to which they are assigned." (Kuhn, 1955: 83).

Kuhn's paper is based to a large extent on the work done by two Hungarian mathematicians, namely D. Kőnig (1916) and E. Egerváry (1931), hence the name, the "Hungarian algorithm". However, it was recently discovered that Carl G. Jacobi (1804 – 1851) came across a solution method very similar to that of Kuhn (see Martello, 2010). Jacobi's findings were posthumously published in 1890.

Many historical notes on the subject of assignment problem development as well as a summary of Kuhn's personal account are contained in Schrijver (2003). Frank (2004) contains a very neat summary of the work done by both Kőnig and Egerváry and then shows how the Hungarian algorithm was developed from this work.

Shortly after Kuhn published his ground-breaking article an American by the name of James Munkres reviewed Kuhn's work and made several important contributions to the theoretical aspects of the algorithm. Munkres (1957) found that the algorithm is (strongly) polynomial and elaborated on the usage of the algorithm by slightly altering the input design in order to apply it to a transportation problem.

Munkres' contribution to the development of the Hungarian algorithm has led to the algorithm also being referred to as the Kuhn-Munkres algorithm or the Munkres assignment algorithm.

Since the Hungarian algorithm by design optimizes the assignment of any two sets of values that need to be matched it has since been used with supreme confidence by in a vast number of fields since its inception (Tichavský & Koldovský, 2004).

This paper's research problem is exactly the same as the personnel-assignment task for which the Hungarian algorithm was initially created in 1955. In context of this paper the baseball players are the personnel and the different fielding positions are the jobs. The steps in the application of the algorithm is detailed in the following chapter and in chapter four it is applied to the set of players in question.

Chapter 3 - Methodology

We now embark upon the endeavour of answering the research question: "What composition of players from my group will yield the most effective baseball team?" One would be inclined to respond with "Well, let us see how good they are". This approach is in fact the first of three stages of our research process. A series of baseball trials and tests of athleticism provides an unbiased assessment of each individual's aptitude, or lack thereof. After the group has been evaluated on their baseball prowess we then need to present a definite criterion of global team efficiency (Baeva *et al.*, 2008). Rounding off the research process is the matter of optimising the group of players for the desired goal by using their test results.

3.1 The Baseball Aptitude Tests

The aim of the skills tests is to cover the entire spectrum of athletic prerequisites in baseball whilst considering the specific positions in which a player would be required to display these skills. Some of the tests are aimed at testing ability for a specific fielding (or the batting) position, others cover a set of similar positions. Table 1 presents a list of the ten tests run in this research, as well an initial indication of their applicability.

Test	Defence/Offence	Positional Application
1. Hitting	Offence	Designated Hitter; All
2. Speed	Both	N/A
3. Pitching	Defence	Pitcher
4. Base-distance Throwing	Defence	N/A
5. Long-distance Throwing	Defence	Outfielders
6. Base-distance Catching	Defence	N/A
7. Fly Ball Catching	Defence	N/A
8. Ground Fielding	Defence	N/A
9. Reach	Both	N/A
10. Reaction Time	Both	N/A

Table 1: Skills tests

Although most of the tests are self-explanatory, we shall now discuss the practical aspects of each one and consider the meaning of each of the recorded values. All the factors influencing performance in the tests were kept constant (or as close to as possible), including the equipment used, distances and the scorer, as some tests required subjective interpretation.

3.1.1 Hitting Test

This test was conducted by launching balls at the batter using a bowling machine positioned at the pitcher's mound (18.39 m from the batter). Although there was some variation in the machine-pitched deliveries, this slight inconsistency was present for all the participants. Each player received a total of ten opportunities, where an opportunity is defined as a ball that was either deemed to be hittable or that was swung at in any instance by the test subject.

The statistics that were measured and later used in calculating a player's score for the test are the hitting percentage and the average distance of all the hits that reached the outfield. If a ball was hit in play a hit was registered but was only measured if it came to rest outside of the bases. The hitting percentage is simply the proportion of the ten opportunities that the participant managed to hit fair.

3.1.2 Speed Test

The aim of this test is to measure the time it takes a player to run between the bases. Running times over two distances were recorded:

- From home plate to first base (27.4 m) from a standing start and holding a bat.
- Around the bases from first base to home plate (82.2 m).

3.1.3 Pitching Test

One of only a few tests conducted to test for a capable player in a specific position, this test almost completely replicates the setting a pitcher is faced with in a match situation, barring the presence of a batter.

The participants stood at pitching distance from the home plate and attempted to throw the ball through the strike zone (see Figure 1). Each player was allowed ten attempts. The pitches were scored as follows: two points were awarded for a strike, one point was awarded for a ball (outside the strike zone) and a player received zero points for a wild pitch. A wild pitch is one that is so skew that the catcher can't reach or stop the ball with an ordinary effort. The points are tallied to produce a test score.

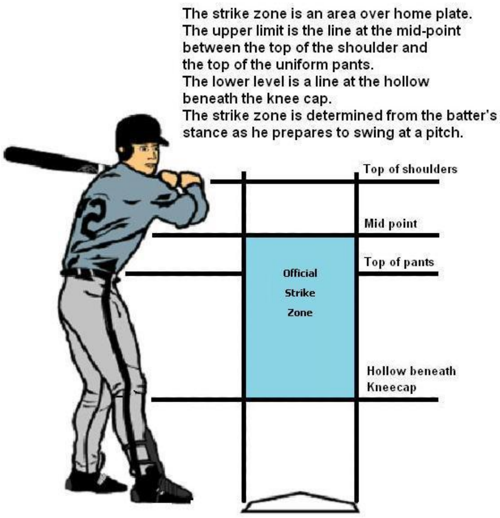


Figure 1: The official strike zone

3.1.4 Throwing Tests

Testing of the throwing ability of the participating group was done by conducting two separate tests. Both of these tests involved aiming at a target 1 meter wide and 2 meters tall and in both tests each participant was given ten attempts to hit the target.

In the test carried out at base-distance (27.4 m) the thrower was awarded one point for hitting the target either directly or after the first bounce.

In the long-distance test the target was placed at a distance of approximately 40 meters from the thrower. Since the target is quite a small one at such a distance, one point was awarded for throws only narrowly missing. Furthermore, two points were awarded for an accurate throw that bounced twice, four points for one bouncing once and eight points if a player managed to hit the target on the full.

3.1.5 Fielding Tests

These three tests assess ball handling skills and catching ability essential in all the fielding positions. As in the aforementioned tests, each player's score was calculated as the aggregate points scored from ten attempts.

A bowling machine was again used to launch balls at the players; as straight as possible at base-distance for the first test and then at approximately twice this distance at a higher angle to simulate fly balls. The result was recorded either as a one if the player caught the ball or a zero if they failed to do so.

The last fielding test saw the players having to run and field balls thrown along the ground to the left of them. Two points were awarded for an attempt where the ball was stopped completely and control was shown by immediately returning the throw. If the ball was blocked but not fielded cleanly or fumbled, a player received one point. Players did not receive any points if they allowed the ball to get away from them.

3.1.6 Reach Test

A significant determining factor in many a play in baseball is the flexibility or reach of the fielder responsible for executing the play. This applies especially to a baseman receiving a wide throw and to catchers in general.

In order to measure how far each participant is able to reach, the right foot is fixed against a block and the distance measured (in cm) is from this block to the furthest point the player is able to touch the ground with the left hand.

3.1.7 Reaction Test

Baseball is a sport involving a projectile flying around at great velocities, with two players positioned within twenty meters of the point at which this projectile abruptly changes direction, namely the pitcher and the catcher. Therefore, the ability to react quickly is not only paramount to the successful execution of plays, but also to these fielders' safety.

There are various ways of assessing a person's reaction time, for example the light board test used in testing boxers. This paper suggests a method that can easily be conducted without any specialised equipment. The procedure is as follows: First, fix the test subject's arm on a surface such that it can't move down. Then suspend the bottom edge of a ruler or similarly shaped object between the subject's fingers, which should be the same distance apart as all the other participants', before dropping it after an indeterminate time. Reaction time is then initially measured as the distance the ruler dropped before the subject catches it between their fingers.

To relate this distance to a reaction time (in seconds), the equation for falling bodies is used:

$$t = \sqrt{2d/g}$$

Where d is the distance the object dropped (in meters) and $g \approx 9.8\, m/s^2$ is the free-fall acceleration on earth.

3.2 Data Transformation Procedures

Just as important as the collection of the raw data itself, the arrangement of the data in a format suitable for further analysis should be done with the utmost care and consideration. It is crucial that all the tests' scores give an accurate reflection of the talent displayed by the participating group and, moreover, it is essential that the different test scores are fully comparable. Comparability is understood as having the same scale, meaning that a higher score in one test than another implies a proportionally superior performance in said test.

However, we shall later discuss the rationale behind an approach that sees test scores arranged in a format such that they are more easily compared with the performance of the same player in another test, relative to the rest of the group. First of all, let us convert each test score to a value in the range [0, 1].

The following sections focus on the rationale behind the conversion of the test scores. The mathematical application is illustrated with the use of a recorded data set in the next chapter.

3.2.1 Success Percentages

The most convenient way of processing the data to a scaled score is to simply relate the number recorded to a success percentage. This applies to tests in which the recorded value is the sum of the points scored for each attempt. A player's success percentage is then calculated as the score obtained divided by the maximum possible score. The tests to which this applies are the following:

- Pitching
- Base-distance throwing
- Base-distance catching

- Fly ball catching
- Ground fielding

Note that the long-distance throwing test is omitted from this list. The reason is that a large range of points are chosen to reward certain achievements, with one being the least a player could achieve and eight being the highest. Therefore, eighty is the maximum possible score, which can only be achieved by hitting the target on the full with all ten attempts. Since this is highly unlikely to be achieved at such a distance, a new "score limit" should rather be chosen on the basis of the range of scores produced by the group of players.

The maximum possible score in the pitching and ground fielding tests is twenty, whilst ten is the highest score a player can achieve in the base-distance throwing, base-distance catching and fly ball catching tests.

3.2.2 Less is More

Whereas a high score is desirable in most of the skills tests, there are two tests in which the best performance is the one yielding the lowest score. The two tests referred to are the speed and reaction tests. Both of these tests' measurements are time, with as fast a time as possible being the ideal.

The speed test possesses the added complication of consisting of two separate times. To combine these two times to a single value, each is multiplied by a weight, with the weights adding up to one. The first of these times, recorded over a shorter distance, naturally receives a much larger weight since acceleration and sprinting are of greater importance in baseball than stamina.

Now, to order these tests' times such that the highest value indicates the fastest time and the lowest value the slowest, we first subtract each recorded time from either the largest value in the set or a suitable larger number, depending on the range of values produced.

Finally, to produce a set of scores scaled between zero and one, we divide each score by a constant. This constant obviously has to be equal to or larger than the largest of the transformed values.

However, this approach exposes a critical difference in some of the tests. Although all the test scores are converted to a percentage, one cannot claim that they are on exactly the same scale. The reason for this being that the time trials just discussed as well as the hitting and reach tests do not have a benchmark which can be considered as "full marks" and, therefore, a player cannot score a one in these tests on the basis of a predetermined calculation.

A simple solution to this problem can be devised, but let us first consider the two remaining problematic skills tests.

3.2.3 The Reach Test and Hitting Test

The format of the data gathered from the reach test is in essence similar to that of the time trials, after the times have been converted such that the score obtained is directly proportional to performance. Once again we are faced with the predicament of having to choose a maximum score to divide the scores by in order to produce a percentage.

The hitting test also requires unique consideration. The main measure of a batter's performance in baseball is the batting average (BA), which is defined as the ratio of hits to at bats. Other commonly used batting statistics are the slugging percentage (SLG), which also indicates a batter's power, and runs created (RC).

Our aim is to identify the strongest batting candidates using only the data gathered from the batting test. The two statistics chosen to calculate a batting score are the player's hitting percentage and the average distance of the player's successful hits (provided these hits reached the outfield). The reason for including distance as a measurement is to avoid identifying a weak, yet accurate batter as a strong offensive player. A player with the ability to hit a home run every fourth at bat will score a lot more runs than a player who taps the ball straight to the infield every single at bat.

To combine the hitting percentage and average distance into a single score, we first convert the distance to a percentage as well. This is achieved by dividing each player's average by a suitable value. As will be shown in the following chapter, it turns out 100 meters is a fair benchmark for the chosen group of players.

One is now fronted with the task of choosing how large an emphasis to place on the hitting and distance percentages respectively. This choice is one of many throughout the application of the system which illustrates its adaptability and the ease at which it can be customised. For the purposes of this research it was decided to assign weights of 0.75 to the hitting percentage and 0.25 to the distance percentage.

Once again, however, it can be argued that the batting test's scores are not fully comparable with, say, the fly ball catching test. It is fair to say that it is a lot easier to catch ten looped balls than to hit ten balls in a row an average distance of 100 meters. To overcome this discrepancy we judge each player's performance in a test relative to the other players'.

3.2.4 Relative Scores

There are two factors motivating the use of relative scores:

1. Some of the tests are more difficult than others and will, therefore, produce lower overall scores.
2. Some of the tests do not have a score limit which causes the scale of these tests' scores to be subjectively determined by the user.

The reason why it is important to arrange all the tests' scores on exactly the same scale is that weights are to be applied to certain tests in order to relate a player to a specific position. A problem would occur if one of the tests applicable to a certain position has much lower overall scores than the other applicable tests; the result being that the weight assigned to this test is in effect diminished by the difficulty of the test (or a poor choice of scale).

Converting the actual recorded scores to relative scores is quite simple. The conversion entails identifying the highest and lowest scoring participants in each test and then assigning scores to the rest of the players relative to these two scores.

Let l_i and u_i be the lowest and highest scores respectively for test i, $i = 1, \dots, 10$.

The relative score (RS) for each observation j in test i is then a transformation of the absolute score (AS):

$$RS_{ij} = \frac{(AS_{ij} - l_i)}{(u_i - l_i)},\ j = 1, \dots, 21$$

The highest score in each test now has a score of one, whilst the lowest score in the test is indicated by zero. A player whose score is halfway between the two extremes will have this score converted to a half. Not only does this method increase the score of a player who did comparatively well in a difficult test, but it also solves the problem of having to decide what the maximum score is for, say, the reach test.

Once all the tests' scores have been converted, this information can be summarized in a matrix. This matrix compiled will be referred to as

$$P = \begin{pmatrix} RS_{11} & \cdots & RS_{1t} \\ \vdots & \ddots & \vdots \\ RS_{n1} & \cdots & RS_{nt} \end{pmatrix}$$

where $t = 10$ is the number of tests; $n = 21$ is the number of players in this study.

Although the use of relative scores is inspired by the lack of a definite measure of performance and difficulty inherent in some of the skills tests, this method of processing the scores actually ideally prepares the data for the purpose of this research. Our goal is to optimally assign each player in the group to a position. A player is related to a specific position through certain tests.

The core idea of this research is to compare different players within the same group with each other and, therefore, we need not evaluate the performance of players in a test judged on what they could have achieved. We should rather compare their test scores with the rest of the group's scores and rank them accordingly.

It should be noted that the actual scores initially calculated must not be disregarded completely. These scores are a pure assessment of a player's performance in the tests and are to be used when comparing players in different groups or to monitor a player's development over time.

3.3 Dimensionality of the Data

Before we can confidently use the relative scores to optimize the team, there is first the matter of determining the number of dimensions captured in the data. In fact, we are mainly interested in checking whether or not the data consists of more than one distinct dimension. A dimension can be interpreted as "a cause of variation".

The reason for checking this fact is to ensure that the test results do not simply indicate the overall sporting ability of each candidate but that the variation present in the data can be attributed to the different facets of baseball athleticism that the skills tests aim to quantify.

To determine the number of dimensions we run a principal component analysis (PCA) on the data gathered. Whereas PCA is usually used to reduce the number of dimensions (Rencher, 2002), we are merely concerned with determining how many components are present. Indeed, though the composition of the components may prove useful in creating a reduced-dimension data set or in advising on the choice of the weights, it must be reiterated that this is not the primary purpose of the PCA.

Once we have established multidimensionality we can safely regard certain tests as indicative of a player's capability in a specific task. However, most of the positions on a baseball field require a few of these tasks to be performed and most of these tasks are crucial in more than one position. The introduction of weights solves the problem of incorporating this interdependency into the system.

3.4 Introducing Weights

The purpose of the skills tests on which we focused in this chapter is to relate each player to a position on the field. The bridge we use to make this connection is a matrix of weights. To explain the usage of weights more clearly, let us first consider a single position. For any one position we introduce a vector of weights, each containing as many elements as there are tests.

The weight vector for position i is then represented as $\boldsymbol{w^i} = (w_1^i, w_2^i, \dots, w_t^i)'$, $w_j^i \geq 0$ and $\sum_{j=1}^{t} w_j^i = 1$, $j = 1, \dots, t$. Each w_j^i represents the importance of test j corresponding to the position i.

All the columns (positions) are then combined, producing a single weight matrix W:

$$\begin{pmatrix} w_{11} & \cdots & w_{1k} \\ \vdots & \ddots & \vdots \\ w_{t1} & \cdots & w_{tk} \end{pmatrix}$$

where k is the number of positions and again t is the number of tests. The number of positions is usually nine, namely all the fielding positions. However, should the addition of an extra position, for instance a designated hitter (DH) be required, this can easily be done by adding the appropriate column of weights to W.

Once again the choice of weight vectors is the responsibility of the field expert and should be chosen with careful consideration. The subjective opinion of the relative importance of certain tests depends on the required composition of the team, especially in terms of offense versus defence. Therefore, the ultimate objective of the team directly determines the assignment of weights.

We now approach the point of defining a player's efficiency in terms of the team's requirements. Note that the term efficiency is used only in this context; the role to be fulfilled by the group. Efficiency can, therefore, be regarded as an empirical assessment of a player's suitability for a position in the light of what the team is trying to achieve.

Now the matrix P relates the players to the tests, whilst W relates the tests to the positions accordingly. To relate each player to a position, we multiply the two matrices harbouring the applicable information:

$$Q = P \cdot W = \begin{pmatrix} RS_{11} & \cdots & RS_{1t} \\ \vdots & \ddots & \vdots \\ RS_{n1} & \cdots & RS_{nt} \end{pmatrix} \begin{pmatrix} w_{11} & \cdots & w_{1k} \\ \vdots & \ddots & \vdots \\ w_{t1} & \cdots & w_{tk} \end{pmatrix} = \begin{pmatrix} Q_{11} & \cdots & Q_{1k} \\ \vdots & \ddots & \vdots \\ Q_{n1} & \cdots & Q_{nk} \end{pmatrix}$$

The solution to our research problem is found by identifying the combination of values in Q resulting in the greatest summed efficiency, subject to certain constraints. These conditions are:

- Exactly one value must be chosen in each column (ensures a player is selected for every position)
- At most one value can be selected in each row (ensures no players are assigned to more than one position)

Stated mathematically, the optimal team efficiency is defined as

$$E = \max_{Y} Z(Y) = \sum_{i=1}^{n} \sum_{j=1}^{k} y_{ij}\, Q_{ij}$$

where $y_{ij} = \{0, 1\}$, subject to $\sum_{i=1}^{n} y_{ij} = 1, j = 1, \dots, k$ and $\sum_{j=1}^{k} y_{ij} \leq 1, i = 1, \dots, n$.

The objective function Z represents the overall efficiency of the team. The combination of players yielding E is the optimal selection and placement of agents out of all possible combinations to achieve the defined goal. A player (i) is assigned to a position (j) if the corresponding y_{ij} equals 1.

In summary, the weight matrix links each player to a position through the scores obtained from the skills tests in such a way that the most appropriate player for each position is instantly identified, whilst still adhering to the collective team agenda. However, it is prudent to constantly bear in mind that simply choosing the player with the highest efficiency for each position does not ensure optimal team efficiency. The optimal solution is rather more complex and involves optimally assigning a set of persons (players) to a set of jobs (positions), as described by Kuhn (1955) himself. Kuhn's solution is still the essential ingredient in trying to solve a problem of this nature. In order to apply the Hungarian algorithm, however, we must first explore the algorithm's methodology.

3.5 How the Hungarian Algorithm Works

The assignment problem we are faced with in this research differs from the original setting the Hungarian algorithm was designed for in two distinct ways. Firstly, the algorithm was developed to assign n "persons" to n "jobs". Another discrepancy is the fact that the algorithm gives a solution in terms of minimum cost, whereas our problem is one of maximisation. Ways of overcoming these obstacles are explained in the following steps, as the algorithm is an iterative and possibly repetitive procedure.

Given a matrix containing the "value" a set of agents (rows) contributes to certain tasks (columns), this procedure ensures optimal assignment:

Step 1: If necessary, convert the values in the matrix from maximum profit to minimum cost. This is achieved by subtracting each element from the maximum value in the matrix. Therefore, replace Q_{ij} with $C - Q_{ij}$, where $C = max(Q_{ij})$ for $i = 1, \dots, n$ and $j = 1, \dots, k$.

Step 2: If the matrix is rectangular, *i.e.* of the dimension $n \times m$ where $n > m$, transform it to a square matrix by adding $n - m$ columns, all containing a constant larger than C. Conversely, add $m - n$ rows if $m > n$.

Step 3: In each row, subtract the minimum value in the row from every element.

Step 4: In each column, subtract the minimum value in the column from every element.

These last two steps ensure that there is at least one zero in each row and column.

Step 5: There are a few slightly different ways of continuing from this point, some of which immediately produce a solution in simple cases. Since the matrix Q in context of this research is rather large, the following approach proves fastest.

Draw lines through the rows and columns in such a way that all the zeros are covered using the minimum number of lines required. Let k be the number of lines used.

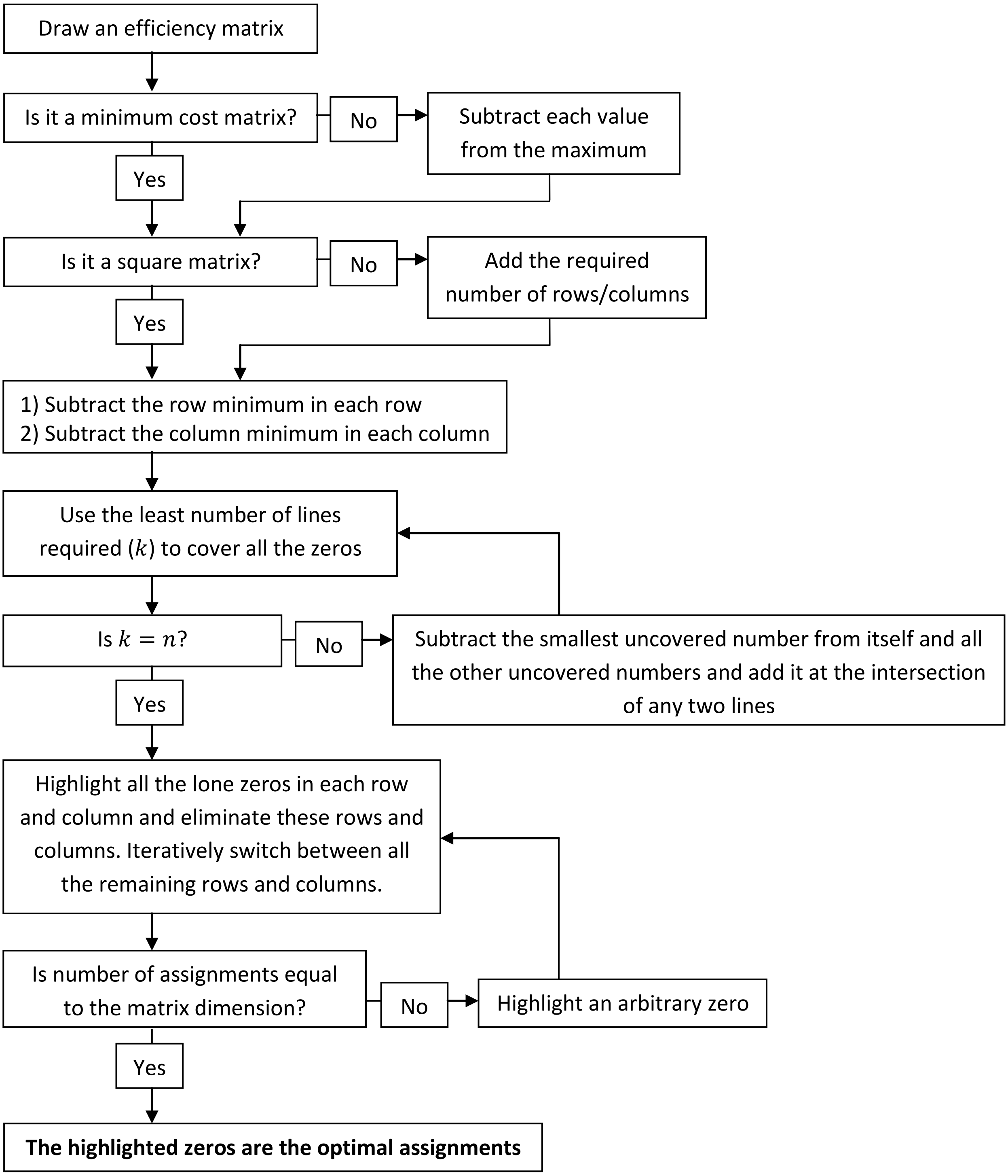
- If $k = n$, move on to Step 6.
- If $k < n$, let m be the smallest number that is not covered by any lines. Now look at every element in the matrix. If a number is not covered by any lines, subtract m from this number (including from m itself). Add m to all numbers in a position where two lines intersect. Keep the rest of the values as they are. Repeat Step 5 with the resulting matrix until $k = n$.

Step 6: Evaluate each row, starting at the top and highlight the zeros which are the solitary zeros in the row. The positions of these zeros are unique assignments and, therefore, the corresponding row and column can be deleted from further consideration. If all n assignments have not been made by applying this step, repeat this procedure for the columns, starting from the left.

Continue iterating between rows and columns until all n assignments have been made. If a final complete solution cannot be reached this means that there is no unique solution yielding the minimum overall cost. An arbitrary zero can then be selected and Step 6 can be repeated if necessary for the remaining rows and columns, producing a final assignment solution.

In terms of the application done in this research, the assignment will simply be interpreted as follows: For all the highlighted zeros at the position Q^*_{ij}, assign player i to position j.

The following flow diagram summarizes the procedure involved in applying the Hungarian algorithm.



Source: Castello, 2010

Chapter 4 - Application

After clearly defining the research objective and methodology to be used to reach this objective, we can now test the proposed system on a group of players. The participants in this study possess quite a wide range of sporting ability, the majority having had at least some experience playing baseball whilst a few players were complete newcomers to the sport. The group includes twenty males and one female, ranging in age from nineteen to twenty-nine.

As mentioned before, the skills tests were conducted under highly controlled circumstances. After capturing the data it is processed and arranged into a matrix in accordance with the methods explained in section 3.2. The resulting scores can be seen in Table 2. The test numbers correspond to the tests as given in Table 1.

	Test									
Player	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	0.512	0.764	0.5	0.3	0.72	0.8	0.4	0.75	0.813	0.822
2	0.759	0.748	0.7	0.7	0.24	0.9	1	0.8	0.733	0.751
3	0.384	0.899	0.5	0.1	0.16	0.8	0.8	0.9	0.804	0.655
4	0.424	0.618	0.7	0.5	0.48	0.8	0.8	0.7	0.896	0.751
5	0.556	0.636	0.75	0.5	0.48	0.9	1	0.75	0.833	0.596
6	0.368	0.773	0.55	0.4	0.08	1	1	0.85	0.863	0.718
7	0.467	0.681	0.65	0.4	0.4	0.4	1	0.7	0.908	0.465
8	0.558	0.861	0.7	0.6	0.6	0.5	0.2	0.65	0.854	0.465
9	0.266	0.545	0.6	0.1	0.16	0.3	0.8	0.6	0.792	0.686
10	0	0.790	0.65	0.2	0.12	0.2	0	0.9	0.792	0.569
11	0	0.616	0.35	0.2	0	0.4	0.6	0.8	0.888	0.515
12	0.389	0.121	0.5	0.2	0	0.4	0.8	0.65	0.854	0.393
13	0.233	0.890	0.7	0.6	0.76	0.7	0.9	0.85	0.833	0.655
14	0	0.856	0.65	0.1	0.12	0.9	1	1	0.917	0.655
15	0.190	0.732	0.2	0.1	0	0.7	0.9	0.75	0.921	0.348
16	0.233	0.370	0.35	0.1	0	0.4	0.2	0.35	0.842	0.655
17	0.310	0.666	0.5	0.4	0.84	0.8	0.9	0.9	0.829	0.625
18	0.586	0.748	0.65	0.2	0.72	0.9	0.9	0.9	0.904	0.515
19	0.210	0.521	0.7	0.3	0.2	0.7	0.3	0.6	0.796	0.751
20	0.460	0.898	0.6	0.4	0.2	0.9	0.8	0.55	0.879	0.686
21	0.841	0.810	0.75	0.4	0.36	0.9	0.9	0.8	0.792	0.596

Table 2: The original (actual) scores

The five tests of which the scores cannot be calculated as success percentages require special consideration. After observing the range of values produced in each of these tests, a player's actual score (AS) is calculated by transforming the recorded test value(s) using a unique formula. These formulas are:

- Hitting test
  $AS = 0.75 \times hitting\ percentage + 0.25 \times (average\ distance \div 100)$
- Speed test
  $AS = (9 - v) \div 3.5$, where
  $v = 0.8 \times time\ to\ 1^{st}\ base + 0.2 \times time\ to\ home\ base$
- Long-distance throwing test
  $AS = total\ score \div 25$
- Reach test
  $AS = distance\ reached \div 240$
- Reaction test
  $AS = (0.3 - t) \div 0.2$, where t is as calculated in section 3.1.7

The actual scores are useful when evaluating the chosen team's strength in certain aspects of play, for instance pitching. However, for the purpose of comparing the players and selecting a team, we need to convert these actual scores to relative scores using the method explained in section 3.2.4. Table 3 presents the relative scores.

	Test									
Player	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	0.609	0.813	0.545	0.333	0.857	0.75	0.4	0.615	0.422	1
2	0.902	0.827	0.909	1	0.286	0.875	1	0.692	0	0.849
3	0.456	0.994	0.545	0	0.190	0.75	0.8	0.846	0.378	0.647
4	0.504	0.636	0.909	0.667	0.571	0.75	0.8	0.538	0.867	0.849
5	0.661	0.666	1	0.667	0.571	0.875	1	0.615	0.533	0.524
6	0.437	0.840	0.636	0.5	0.095	1	1	0.769	0.689	0.779
7	0.555	0.713	0.818	0.5	0.476	0.25	1	0.538	0.933	0.246
8	0.664	0.943	0.909	0.833	0.714	0.375	0.2	0.462	0.644	0.246
9	0.317	0.538	0.727	0	0.190	0.125	0.8	0.385	0.311	0.712
10	0	0.844	0.818	0.167	0.143	0	0	0.846	0.311	0.465
11	0	0.642	0.273	0.167	0	0.25	0.6	0.692	0.822	0.353
12	0.463	0	0.545	0.167	0	0.25	0.8	0.462	0.644	0.095
13	0.276	0.984	0.909	0.833	0.905	0.625	0.9	0.769	0.533	0.647
14	0	0.933	0.818	0	0.143	0.875	1	1	0.978	0.647

Table 3: The relative scores

	Test									
Player	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
15	0.226	0.772	0	0	0	0.625	0.900	0.615	1	0
16	0.276	0.320	0.273	0	0	0.250	0.200	0	0.578	0.647
17	0.369	0.716	0.545	0.500	1	0.750	0.900	0.846	0.511	0.584
18	0.696	0.798	0.818	0.167	0.857	0.875	0.900	0.846	0.911	0.353
19	0.250	0.506	0.909	0.333	0.238	0.625	0.300	0.385	0.333	0.849
20	0.547	1	0.727	0.500	0.238	0.875	0.800	0.308	0.778	0.712
21	1	0.892	1	0.500	0.429	0.875	0.900	0.692	0.311	0.524

Table 3 (Continued): The relative scores

4.1 Testing for Multidimensional Data

As explained in section 3.3, we first have to confirm that there is more than one component accounting for the variation present in the data. The results given in Tables 4 and 5 are produced by running a PCA on the relative scores using the statistical package STATA (StataCorp, 2007).

Component	Eigenvalue	Difference	Proportion	Cumulative
Comp1	.291973	.145864	0.3671	0.3671
Comp2	.146109	.0565506	0.1837	0.5507
Comp3	.0895588	.00476817	0.1126	0.6633
Comp4	.0847906	.0330074	0.1066	0.7699
Comp5	.0517832	.0107042	0.0651	0.8350
Comp6	.041079	.00269203	0.0516	0.8867
Comp7	.0383869	.00965742	0.0483	0.9349
Comp8	.0287295	.0117599	0.0361	0.9710
Comp9	.0169695	.0109085	0.0213	0.9924
Comp10	.00606108	.	0.0076	1.0000

Table 4: Principal components (eigenvalues)

Variable	Comp1	Comp2	Comp3	Comp4	Unexplained
Batting	0.3520	-0.0884	0.0425	-0.4991	.01571
Speed	0.2356	0.1271	-0.0515	0.4877	.02039
Pitching	0.3378	-0.2299	0.0373	0.0082	.02934
SD_Throws	0.4498	-0.2305	0.2240	-0.2130	.02176
LD_Throws	0.4406	-0.1021	0.5421	0.3174	.0171
SD_Catch	0.4042	0.3436	-0.3828	0.0111	.01349
LD_Catch	0.2610	0.5965	-0.1237	-0.3616	.01257
Ground_F	0.1503	0.2851	-0.0231	0.4292	.01912
Reach	-0.1191	0.4908	0.3911	0.0502	.01945
Reaction	0.2040	-0.2522	-0.5785	0.2221	.01407

Table 5: Principal components (eigenvectors)

Interpreting the results in Table 4, we see that four components explain more than 75% of the variation in the relative scores. Identifying these components, however, is not as trivial.

The second and third components in Table 5 might be seen as indicative of throwing ability and glove dexterity, but this is not clear.

It does appear as though the reach and reaction tests are somewhat confounding variables. These “abilities” are characteristics that are not of as fundamental importance as the rest of the tests, although they might be a determining factor should two players have similar test results in all the other tests applicable to a certain position. Now since these two variables appear to be completely uncorrelated with the remaining eight tests and are not of great significance, we run another PCA excluding the reach and reaction tests as variables.

Component	Eigenvalue	Difference	Proportion	Cumulative
Comp1	.28099	.157766	0.4303	0.4303
Comp2	.123224	.0397734	0.1887	0.6189
Comp3	.0834507	.0326808	0.1278	0.7467
Comp4	.0507699	.00615133	0.0777	0.8244
Comp5	.0446186	.0131772	0.0683	0.8928
Comp6	.0314413	.00844122	0.0481	0.9409
Comp7	.0230001	.00741133	0.0352	0.9761
Comp8	.0155888	.	0.0239	1.0000

Table 6: Principal components (eigenvalues) excluding reach and reaction tests

Variable	Comp1	Comp2	Comp3	Comp4	Unexplained
Batting	0.3593	-0.1227	-0.4791	-0.0936	.01659
Speed	0.2430	0.1187	0.4919	0.3727	.01378
Pitching	0.3285	-0.2726	-0.0018	0.4818	.01927
SD_Throws	0.4522	-0.3551	-0.1974	0.2772	.0168
LD_Throws	0.4553	-0.3301	0.3842	-0.7082	.000732
SD_Catch	0.4144	0.4205	-0.0542	0.0442	.0212
LD_Catch	0.3032	0.6228	-0.3081	-0.1745	.01381
Ground_F	0.1688	0.3112	0.4940	0.0975	.01247

Table 7: Principal components (eigenvectors) excluding reach and reaction tests

Again we see from Table 6 that 75% of the variation is explained by four components and one of these components does appear to encompass the throwing tests collectively (Table 7). However, we are not concerned with identifying the components *per se.* Since it is clear that the data is not one-dimensional it can be safely concluded that the skills tests do not merely indicate general athleticism and the usage of weights applied to the tests is statistically justified.

It can be noted, however, that general athleticism does play a role in the explanation of the variation in the original data, since the first component in each PCA is definitely summarising a measure of sporting ability. Of course, this athleticism accounts for between 36% and 43% of the overall variation, not enough to justify calling the data one-dimensional.

4.2 The Optimal Baseball Team

The composition of any sports team is the result of first having a clear understanding of what is to be accomplished and then selecting the optimal combination of individuals to achieve this goal. The goal of any baseball team is to score runs and prevent the opposing team from doing so. Therefore, the main factor to take into consideration in composing the optimal baseball team is the balance between defensive and offensive capability.

Players displaying the specific strengths required by the team are given a certain amount of preference, the extent of which is determined by the weights placed on the skills test in question. Before we start composing all these weights in a matrix, let us first consider the defensive positions as well. Since we are not only choosing a group of players, but also positioning them optimally, it is important to understand the responsibilities of each fielder.

The different fielding positions and their numbers corresponding to the columns in which they will be added in the weight matrix are given in Table 8. Figure 2 shows the positioning of the fielders.

Number	Position	Abbreviation
1	Pitcher	P
2	Catcher	C
3	First Base	1B
4	Second Base	2B
5	Third Base	3B
6	Shortstop	SS
7	Left Field	LF
8	Centre Field	CF
9	Right Field	RF

Table 8: Fielding positions with numbering and abbreviations

Figure 2: Fielding positioning

Although the choice of weights is subjective, there are some logical conclusions that can be drawn by looking at the fielders' positioning. For instance, positions 7-9 demand throwing accuracy at a greater distance than any other position and therefore the weight placed on test number 5 will be the largest for these positions. Some of the other factors to bear in mind include the large number of balls thrown to first base, the amount of running outfielders are required to do, *etc.* This is perhaps best explained through an application.

4.2.1 Optimal Defence

A good starting point for establishing suitable weight proportions is to construct a team optimized for fielding, thus ignoring the batting test. The set of weights chosen in Table 9 can be altered as is thought necessary, as long as the columns add up to one.

	Position								
Test	**P**	**C**	**1B**	**2B**	**3B**	**SS**	**LF**	**CF**	**RF**
Hitting	0	0	0	0	0	0	0	0	0
Speed	0	0	0	0	0	0	0.1	0.1	0.1
Pitching	0.9	0	0	0	0	0	0	0	0
Base-Distance Throw	0	0.1	0.25	0.3	0.25	0.4	0.1	0.1	0.1
Long-Distance Throw	0	0	0	0	0.15	0	0.3	0.3	0.3
Base-Distance Catch	0	0.7	0.4	0.3	0.25	0.2	0	0	0
Fly Ball Catch	0	0	0	0	0	0	0.3	0.3	0.3
Ground Fielding	0	0	0.25	0.3	0.25	0.4	0.2	0.2	0.2
Reach	0	0.1	0.1	0.1	0.1	0	0	0	0
Reaction	0.1	0.1	0	0	0	0	0	0	0

Table 9: Weights for optimal defence

As explained in section 3.4, we generate the matrix to be optimized by multiplying this weight matrix with the matrix containing the relative scores given in Table 3. The result is shown in Table 10 and will be referred to, for lack of a better term, as the "efficiency matrix".

	Position								
Player	**P**	**C**	**1B**	**2B**	**3B**	**SS**	**LF**	**CF**	**RF**
1	0.5909	0.7006	0.5794	0.5518	0.5955	0.5295	0.6149	0.6149	0.6149
2	0.9031	0.7974	0.7731	0.7702	0.6847	0.8519	0.7069	0.7069	0.7069
3	0.5556	0.6274	0.5493	0.5166	0.4654	0.4885	0.5658	0.5658	0.5658
4	0.9031	0.7632	0.6879	0.6732	0.6612	0.6321	0.6494	0.6494	0.6494
5	0.9524	0.7849	0.7238	0.7004	0.6783	0.6878	0.7277	0.7277	0.7277
6	0.6506	0.8968	0.7862	0.7497	0.6505	0.7077	0.6164	0.6164	0.6164
7	0.7610	0.3429	0.4529	0.4799	0.4869	0.4654	0.6719	0.6719	0.6719
8	0.8428	0.4349	0.5382	0.5654	0.5891	0.5929	0.5443	0.5443	0.5443
9	0.7257	0.1898	0.1773	0.1840	0.1871	0.1788	0.4279	0.4279	0.4279
10	0.7829	0.0943	0.2843	0.3350	0.3057	0.4051	0.3132	0.3132	0.3132
11	0.2807	0.3091	0.3970	0.4149	0.3595	0.3936	0.3993	0.3993	0.3993
12	0.5004	0.2656	0.3215	0.3279	0.2840	0.3013	0.3490	0.3490	0.3490
13	0.8828	0.6388	0.7040	0.7216	0.7459	0.7660	0.8770	0.8770	0.8770

Table 10: Optimal defence efficiency matrix

	Position								
Player	**P**	**C**	**1B**	**2B**	**3B**	**SS**	**LF**	**CF**	**RF**
14	0.8010	0.7749	0.6978	0.6603	0.5880	0.5750	0.6362	0.6362	0.6362
15	0	0.5375	0.5038	0.4721	0.4101	0.3712	0.4703	0.4703	0.4703
16	0.3101	0.2974	0.1578	0.1328	0.1203	0.0500	0.0920	0.0920	0.0920
17	0.5493	0.6845	0.6876	0.6800	0.7251	0.6885	0.8608	0.8608	0.8608
18	0.7716	0.7555	0.6943	0.6575	0.6916	0.5801	0.7928	0.7928	0.7928
19	0.9031	0.5891	0.4628	0.4362	0.4048	0.4122	0.3223	0.3223	0.3223
20	0.7257	0.8114	0.6297	0.5826	0.5342	0.4981	0.5230	0.5230	0.5230
21	0.9524	0.7460	0.6792	0.6513	0.6122	0.6519	0.6762	0.6762	0.6762

Table 10 (Continued): Optimal defence efficiency matrix

To illustrate the concept that choosing the best available player in every position does not necessarily guarantee the optimal team, let us first select a team using this crude method and compare the result with the one produced by the Hungarian algorithm. Table 11 shows a team (the selection indicated by the red cells) selected by applying the steps explained afterwards.

	Position								
Player	**P**	**C**	**1B**	**2B**	**3B**	**SS**	**LF**	**CF**	**RF**
1	0.5909	0.7006	0.5794	0.5518	0.5955	0.5295	0.6149	0.6149	0.6149
2	0.9031	0.7974	0.7731	0.7702	0.6847	0.8519	0.7069	0.7069	0.7069
3	0.5556	0.6274	0.5493	0.5166	0.4654	0.4885	0.5658	0.5658	0.5658
4	0.9031	0.7632	0.6879	0.6732	0.6612	0.6321	0.6494	0.6494	0.6494
5	0.9524	0.7849	0.7238	0.7004	0.6783	0.6878	0.7277	0.7277	0.7277
6	0.6506	0.8968	0.7862	0.7497	0.6505	0.7077	0.6164	0.6164	0.6164
7	0.7610	0.3429	0.4529	0.4799	0.4869	0.4654	0.6719	0.6719	0.6719
8	0.8428	0.4349	0.5382	0.5654	0.5891	0.5929	0.5443	0.5443	0.5443
9	0.7257	0.1898	0.1773	0.1840	0.1871	0.1788	0.4279	0.4279	0.4279
10	0.7829	0.0943	0.2843	0.3350	0.3057	0.4051	0.3132	0.3132	0.3132
11	0.2807	0.3091	0.3970	0.4149	0.3595	0.3936	0.3993	0.3993	0.3993
12	0.5004	0.2656	0.3215	0.3279	0.2840	0.3013	0.3490	0.3490	0.3490
13	0.8828	0.6388	0.7040	0.7216	0.7459	0.7660	0.8770	0.8770	0.8770
14	0.8010	0.7749	0.6978	0.6603	0.5880	0.5750	0.6362	0.6362	0.6362
15	0	0.5375	0.5038	0.4721	0.4101	0.3712	0.4703	0.4703	0.4703
16	0.3101	0.2974	0.1578	0.1328	0.1203	0.0500	0.0920	0.0920	0.0920
17	0.5493	0.6845	0.6876	0.6800	0.7251	0.6885	0.8608	0.8608	0.8608
18	0.7716	0.7555	0.6943	0.6575	0.6916	0.5801	0.7928	0.7928	0.7928
19	0.9031	0.5891	0.4628	0.4362	0.4048	0.4122	0.3223	0.3223	0.3223
20	0.7257	0.8114	0.6297	0.5826	0.5342	0.4981	0.5230	0.5230	0.5230
21	0.9524	0.7460	0.6792	0.6513	0.6122	0.6519	0.6762	0.6762	0.6762

Table 11: Crude defensive team selection

Step 1: Identify the largest value in the matrix.

Step 2: Assign the player to the corresponding position and delete the row and column in which this selection occurred.

Step 3: Repeat steps 1 and 2 with the resulting matrix until 9 assignments have been made.

The problems with applying this method are immediately apparent. The first issue to deal with is deciding which player to choose should there be a tie with regards to the highest score. Another concern is that the selection with the lowest positional efficiency is the third base selection, namely 0.6122. Although this is not cause for concern in itself, it should be highlighted that third base is the position that the player in question is least suited for. Furthermore, there are seven players who obtained a higher score for that particular position. Therefore, one would expect that there exists a more balanced solution.

If we wish to compare selected combinations from the same matrix, we need an indication of overall team efficiency. This is calculated by summing all the highlighted values.

$\therefore E^* = \sum_{i=1}^{n} \sum_{j=1}^{k} y_{ij}\, Q_{ij} = 7.2149$ for the selection in Table 11. Since there are nine positions and therefore the theoretical maximum of E^* is 9, the calculated value can be expressed as a percentage.

$$E_p^* = {}^{7.2149}/_{0.09} = 80.17\%$$

We now finally evaluate the efficiency matrix using the Hungarian algorithm. The comparison between the initial selection and the optimal selection is shown in Table 12.

The red cells indicate the initial selections, the blue cells indicate selections made by the Hungarian algorithm and the green cells show the common selections. The Hungarian algorithm was run off-line using a commercial software package (MATLAB, 2009b) courtesy of a program written and made available by Cao (2008).

	Position								
Player	**P**	**C**	**1B**	**2B**	**3B**	**SS**	**LF**	**CF**	**RF**
1	0.5909	0.7006	0.5794	0.5518	0.5955	0.5295	0.6149	0.6149	0.6149
2	0.9031	0.7974	0.7731	0.7702	0.6847	0.8519	0.7069	0.7069	0.7069
3	0.5556	0.6274	0.5493	0.5166	0.4654	0.4885	0.5658	0.5658	0.5658
4	0.9031	0.7632	0.6879	0.6732	0.6612	0.6321	0.6494	0.6494	0.6494
5	0.9524	0.7849	0.7238	0.7004	0.6783	0.6878	0.7277	0.7277	0.7277
6	0.6506	0.8968	0.7862	0.7497	0.6505	0.7077	0.6164	0.6164	0.6164
7	0.7610	0.3429	0.4529	0.4799	0.4869	0.4654	0.6719	0.6719	0.6719
8	0.8428	0.4349	0.5382	0.5654	0.5891	0.5929	0.5443	0.5443	0.5443
9	0.7257	0.1898	0.1773	0.1840	0.1871	0.1788	0.4279	0.4279	0.4279
10	0.7829	0.0943	0.2843	0.3350	0.3057	0.4051	0.3132	0.3132	0.3132
11	0.2807	0.3091	0.3970	0.4149	0.3595	0.3936	0.3993	0.3993	0.3993
12	0.5004	0.2656	0.3215	0.3279	0.2840	0.3013	0.3490	0.3490	0.3490
13	0.8828	0.6388	0.7040	0.7216	0.7459	0.7660	0.8770	0.8770	0.8770
14	0.8010	0.7749	0.6978	0.6603	0.5880	0.5750	0.6362	0.6362	0.6362
15	0	0.5375	0.5038	0.4721	0.4101	0.3712	0.4703	0.4703	0.4703
16	0.3101	0.2974	0.1578	0.1328	0.1203	0.0500	0.0920	0.0920	0.0920
17	0.5493	0.6845	0.6876	0.6800	0.7251	0.6885	0.8608	0.8608	0.8608
18	0.7716	0.7555	0.6943	0.6575	0.6916	0.5801	0.7928	0.7928	0.7928
19	0.9031	0.5891	0.4628	0.4362	0.4048	0.4122	0.3223	0.3223	0.3223
20	0.7257	0.8114	0.6297	0.5826	0.5342	0.4981	0.5230	0.5230	0.5230
21	0.9524	0.7460	0.6792	0.6513	0.6122	0.6519	0.6762	0.6762	0.6762

Table 12: Optimal defensive team compared with initial selection

We can now calculate the team efficiency as defined in section 3.4.

$$E = \max_{Y} Z(Y) = \sum_{i=1}^{n} \sum_{j=1}^{k} y_{ij}\, Q_{ij} = 7.2942$$

Again, expressed as a percentage: $E_p = {}^{7.2942}/_{0.09} = 81.05\%.$

The improvement from the initial team efficiency of 80.16% is slight, but this increase nevertheless shows that the Hungarian algorithm produces a superior assignment.

Table 12 also shows the importance of placement, as opposed to mere selection. There is only one change of personnel between the two teams, with player 20 being preferred to player 14, yet only four of the other eight fielders retain their positions. The difference in player 20 and player 4's positional scores is 0.1136. However, the difference in the teams' total scores is only 0.0792. This discrepancy shows that, although some efficiency is lost through shuffling the fielders around, the defence is more balanced, with any weaknesses eliminated rather than compensated for. One might say all the team's bases are covered.

There are two more statistics regarding the team that would be most useful to calculate, namely a batting and a pitching score. The aim of gathering these statistics would be to compare them with those of another team, whether from the same group of players or not. Therefore, we need to calculate these scores using the actual scores as given in Table 2, since these raw scores are not affected by the team strategy or the relative strength of the players.

Calculating a batting and pitching score is quite simple, since two of the skills tests, *i.e.* the hitting and pitching tests, assess exactly these abilities. The batting score is calculated by summing the actual scores in test 1 over all the selected players. It is unnecessary to calculate a pitching score using the entire team's data since only a few pitchers (if at all more than one) are used in a match. The pitching score will be comprised of the top four scores in test 3. A summary of the team selected for optimal defence is given in Table 13.

Position	**P**	**C**	**1B**	**2B**	**3B**	**SS**	**LF**	**CF**	**RF**
Player	21	20	6	5	4	2	18	17	13
Pitching score		2.9	**Pitchers:**		21	5	2	4	
Batting score		4.535							

Table 13: Optimal defensive team selection and scores

4.2.2 Medium Offence

Selecting a team optimized for defence as in the previous section does not necessarily imply that said team does not possess any batting ability. We now attempt to discover how much, if it all, the team is altered to accommodate for stronger offensive players. Once again the composition of the team in terms of the strategy to be employed is determined by the weights placed on the different tests.

To adjust the overall balance of the team to that of a moderately offense-orientated team we need to add weights to the hitting test and also the speed tests, as a good batter should be able to reach first base very quickly. The question now arises whether the different positions should have different batting weights. In general pitchers have significantly lower batting averages than the rest of the team, but the question is by how much?

This research uses the Major League's batting records of the past season (2010) as an indication of whether or not professional teams sacrifice fielding ability in certain positions for batting performance or *vice versa.* The batting averages of the AL and NL in the 2010 regular season are combined and categorised per position. The data used is courtesy of MLB Advanced Media, L.P. (2010). Now to determine whether there is a significant difference between the batting averages for the different positions, we will be applying the method of bootstrapping.

This sampling method is chosen in favour of more traditional methods to negate the effect that averages calculated from little batting data might have. For instance, out of the 320 pitchers that stepped up to the plate, 164 failed to register a hit and therefore achieved a batting average of zero. Furthermore, 159 of these 164 pitchers faced less than 10 at bats.

A single bootstrap sample is taken by sampling, with replacement, n random averages from the population (position), where n is the size of the population. The average of this bootstrap sample is calculated and the process is repeated 1000 times. After sorting these averaged sampled batting averages in increasing order, the 25th and 975th can then be taken as the 2.5th and 97.5th percentiles. Values within this range then fall into an approximate 95% confidence interval for the batting average of the position in question. If the intervals of two positions do not overlap we can conclude that there is a statistically significant difference in the batting averages for these two positions, with $\alpha = 0.05$.

Table 14 shows the confidence interval for the each position and the graphical representation in Figure 3 illustrates the difference in these intervals. Since the weights used for all three outfield (OF) positions are the same, these statistics are combined.

Position	Number	Confidence Interval
P	1	[0.094; 0.135]
C	2	[0.215; 0.239]
1B	3	[0.214; 0.245]
2B	4	[0.238; 0.266]
3B	5	[0.229; 0.255]
SS	6	[0.224; 0.264]
OF	7	[0.235; 0.249]
DH	8	[0.227; 0.259]

Table 14: Confidence intervals for positional batting averages

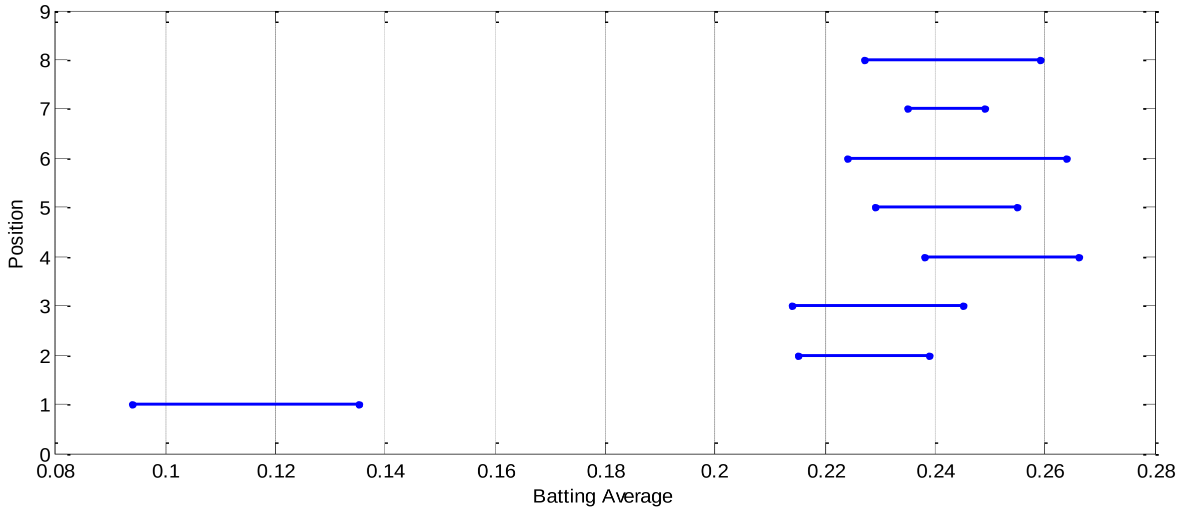


Figure 3: Confidence intervals for positional batting averages

As expected, the overall batting average for pitchers is significantly less than for the other positions. This position, however, is the only outlier. We will therefore decide to use the same batting weights for all the positions barring the pitcher, which will receive half the weight of the rest of the positions.

Table 15 shows the weights used to compose a team with a medium offensive strategy. The weights used in the previous section serve as a benchmark and the ratios between the different defensive tests are kept more or less the same.

	Position								
Test	**P**	**C**	**1B**	**2B**	**3B**	**SS**	**LF**	**CF**	**RF**
Hitting	0.1	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2
Speed	0.1	0.1	0.1	0.1	0.1	0.1	0.2	0.2	0.2
Pitching	0.75	0	0	0	0	0	0	0	0
Base-Distance Throw	0	0.05	0.175	0.25	0.2	0.3	0.05	0.05	0.05
Long-Distance Throw	0	0	0	0	0.05	0	0.25	0.25	0.25
Base-Distance Catch	0	0.55	0.3	0.2	0.2	0.1	0	0	0
Fly Ball Catch	0	0	0	0	0	0	0.2	0.2	0.2
Ground Fielding	0	0	0.175	0.2	0.2	0.3	0.1	0.1	0.1
Reach	0	0.05	0.05	0.05	0.05	0	0	0	0
Reaction	0.05	0.05	0	0	0	0	0	0	0

Table 15: Weights for medium offence

After following the same procedure as in the previous section, *i.e.* calculating the efficiency matrix and optimizing using the Hungarian algorithm, the team selection and summary statistics are determined. These results are given in Table 16.

Position	P	C	1B	2B	3B	SS	LF	CF	RF
Player	21	20	6	8	5	2	18	17	13
Pitching score	2.9		Pitchers:		21	5	2	13	
Batting score	4.669								

Table 16: Moderately offensive team selection and scores

As expected, there is an increase in the batting score when compared to the purely defensive team. This increase is attributed to a single change in personnel. Even though player 4, who had been one of the top four pitchers is now omitted from the team, the replacement in player 8 is equally adept at pitching. Both these players scored 0.7 in the pitching test, resulting in the team's pitching score being unaltered. Table 17 gives a clear representation of the difference in team selections. The blue cells again indicate the new assignments and the green cells the common assignments, while the red cells represent the selections made in the optimal defensive team.

	Position								
Player	P	C	1B	2B	3B	SS	LF	CF	RF
1	0.6013	0.7033	0.6152	0.5806	0.6068	0.5627	0.6568	0.6568	0.6568
2	0.8972	0.8368	0.8218	0.8266	0.7909	0.8583	0.7365	0.7365	0.7365
3	0.5865	0.6544	0.5827	0.5288	0.5384	0.5196	0.5824	0.5824	0.5824
4	0.8384	0.6962	0.6438	0.6322	0.6275	0.6011	0.6182	0.6182	0.6182
5	0.9088	0.7661	0.7122	0.6901	0.6853	0.6708	0.7030	0.7030	0.7030
6	0.6439	0.8198	0.7280	0.6847	0.6645	0.6522	0.5812	0.5812	0.5812
7	0.7528	0.4038	0.4858	0.5117	0.5105	0.5189	0.6516	0.6516	0.6516
8	0.8548	0.5195	0.5984	0.6349	0.6290	0.6530	0.6278	0.6278	0.6278
9	0.6665	0.2370	0.2375	0.2346	0.2441	0.2450	0.4170	0.4170	0.4170
10	0.7213	0.1316	0.2772	0.3109	0.3097	0.3883	0.2975	0.2975	0.2975
11	0.2864	0.2688	0.3306	0.3354	0.3271	0.3469	0.3260	0.3260	0.3260
12	0.4601	0.2754	0.3097	0.3088	0.3004	0.3060	0.3070	0.3070	0.3070
13	0.8402	0.5981	0.6483	0.6675	0.6711	0.6970	0.7769	0.7769	0.7769
14	0.7393	0.6558	0.5797	0.5172	0.5243	0.4808	0.5223	0.5223	0.5223
15	0.0998	0.5161	0.4676	0.4204	0.4204	0.3695	0.4411	0.4411	0.4411
16	0.2965	0.2860	0.1911	0.1661	0.1661	0.1123	0.1592	0.1592	0.1592
17	0.5468	0.6376	0.6315	0.6151	0.6401	0.6242	0.7565	0.7565	0.7565
18	0.7807	0.7718	0.7043	0.6505	0.6850	0.6104	0.7860	0.7860	0.7860
19	0.7998	0.5201	0.4303	0.4025	0.3977	0.3784	0.3258	0.3258	0.3258
20	0.7357	0.7901	0.6521	0.6098	0.5967	0.5392	0.5847	0.5847	0.5847
21	0.9654	0.8372	0.7759	0.7432	0.7396	0.7344	0.7597	0.7597	0.7597

Table 17: Medium offensive team compared with optimal defensive team

4.2.3 Strong Offence

The addition of relatively small weights to the hitting and speed tests forced some positional changes from the team optimised for defence. We now set out to determine how many changes need to be rung to build a team that can be considered a specialised batting unit. Whilst still roughly adhering to the weight ratios used in previous team compositions, a major emphasis is now placed on the batting test as well as the speed test. The weights can be seen in Table 18.

	Position								
Test	**P**	**C**	**1B**	**2B**	**3B**	**SS**	**LF**	**CF**	**RF**
Hitting	0.2	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4
Speed	0.1	0.1	0.15	0.15	0.15	0.15	0.2	0.2	0.2
Pitching	0.7	0	0	0	0	0	0	0	0
Base-Distance Throw	0	0.05	0.1	0.15	0.15	0.2	0	0	0
Long-Distance Throw	0	0	0	0	0.05	0	0.15	0.15	0.15
Base-Distance Catch	0	0.35	0.25	0.15	0.1	0.1	0	0	0
Fly Ball Catch	0	0	0	0	0	0	0.15	0.15	0.15
Ground Fielding	0	0	0.1	0.15	0.15	0.15	0.1	0.1	0.1
Reach	0	0.05	0	0	0	0	0	0	0
Reaction	0	0.05	0	0	0	0	0	0	0

Table 18: Weights for strong offence

The resulting efficiency matrix and team selection, again compared with a purely defensive team is displayed in Table 19.

	Position								
Player	**P**	**C**	**1B**	**2B**	**3B**	**SS**	**LF**	**CF**	**RF**
1	0.5849	0.6751	0.6478	0.6202	0.6256	0.5994	0.6562	0.6562	0.6562
2	0.8995	0.8422	0.8729	0.8700	0.8405	0.8762	0.7883	0.7883	0.7883
3	0.5725	0.5957	0.6038	0.5711	0.5432	0.5336	0.6146	0.6146	0.6146
4	0.8009	0.6470	0.6052	0.5905	0.5816	0.5863	0.5886	0.5886	0.5886
5	0.8987	0.7232	0.7111	0.6877	0.6725	0.6772	0.6946	0.6946	0.6946
6	0.6169	0.7072	0.6778	0.6412	0.5960	0.6162	0.5841	0.5841	0.5841
7	0.7551	0.4649	0.4954	0.5223	0.5336	0.5348	0.6400	0.6400	0.6400
8	0.8634	0.5772	0.6302	0.6574	0.6744	0.6803	0.6374	0.6374	0.6374
9	0.6262	0.2753	0.2771	0.2838	0.2871	0.2775	0.4213	0.4213	0.4213
10	0.6572	0.1316	0.2279	0.2786	0.2857	0.2869	0.2749	0.2749	0.2749
11	0.2551	0.2188	0.2447	0.2626	0.2501	0.2585	0.2876	0.2876	0.2876
12	0.4744	0.3179	0.3104	0.3168	0.3043	0.3127	0.3513	0.3513	0.3513
13	0.7901	0.5284	0.5747	0.5923	0.6063	0.6027	0.6550	0.6550	0.6550
14	0.6660	0.4808	0.4587	0.4212	0.3846	0.3775	0.4580	0.4580	0.4580
15	0.1224	0.4363	0.4239	0.3922	0.3610	0.3610	0.4413	0.4413	0.4413

Table 19: Strong offensive team compared with optimal defensive team

	Position								
Player	**P**	**C**	**1B**	**2B**	**3B**	**SS**	**LF**	**CF**	**RF**
16	0.2782	0.2913	0.2210	0.1960	0.1835	0.1835	0.2045	0.2045	0.2045
17	0.5271	0.5613	0.5770	0.5693	0.5818	0.5568	0.6603	0.6603	0.6603
18	0.7917	0.7360	0.7182	0.6813	0.6804	0.6459	0.7862	0.7862	0.7862
19	0.7369	0.4450	0.4038	0.3772	0.3579	0.3626	0.3203	0.3203	0.3203
20	0.7185	0.7245	0.6683	0.6212	0.5894	0.6025	0.6053	0.6053	0.6053
21	0.9892	0.8622	0.8718	0.8439	0.8216	0.8251	0.8469	0.8469	0.8469

Table 19 (Continued): Strong offensive team compared with optimal defensive team

In this offensively oriented team the batting is bolstered through the inclusion of players 1 and 8. The change of personnel has resulted in some shuffling amongst the ranks in order to maintain the defensive balance, with only five players retaining their original positions. There is once again enough depth in pitching talent such that the pitching score remains 2.9, despite the exclusion of players 4 and 13. The team details are summarized in Table 20.

Position	**P**	**C**	**1B**	**2B**	**3B**	**SS**	**LF**	**CF**	**RF**
Player	5	20	6	21	8	2	18	17	1
Pitching score		2.9	**Pitchers:**		5	21	2	8	
Batting score		4.948							

Table 20: Strongly offensive team selection and scores

4.2.4 Maximum Aggression

In the previous team selection the batting weights were basically pushed to the limit whilst still capturing enough information regarding the players' fielding ability. There is, however, one additional tactic that can be employed to strengthen the batting: the inclusion of a designated hitter (DH).

The use of a DH is not automatically allowed in amateur and lower level baseball and is subject either to both teams agreeing on the usage of the DH or on the practice of the home team, depending on the league specifications. Should the inclusion of a DH be allowed, this player replaces the starting pitcher and all subsequent pitchers in the batting line-up, but is not required to partake in fielding.

To add this position into the selection system developed in this research is as simple as adding another column to the matrix of weights. The full weight of this position is placed on the batting test, as can be seen in Table 21.

	Position									
Test	P	C	1B	2B	3B	SS	LF	CF	RF	DH
Hitting	0.2	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	1
Speed	0.1	0.1	0.15	0.15	0.15	0.15	0.2	0.2	0.2	0
Pitching	0.7	0	0	0	0	0	0	0	0	0
Base-Distance Throw	0	0.05	0.1	0.15	0.15	0.2	0	0	0	0
Long-Distance Throw	0	0	0	0	0.05	0	0.15	0.15	0.15	0
Base-Distance Catch	0	0.35	0.25	0.15	0.1	0.1	0	0	0	0
Fly Ball Catch	0	0	0	0	0	0	0.15	0.15	0.15	0
Ground Fielding	0	0	0.1	0.15	0.15	0.15	0.1	0.1	0.1	0
Reach	0	0.05	0	0	0	0	0	0	0	0
Reaction	0	0.05	0	0	0	0	0	0	0	0

Table 21: Weights for maximum aggression

We again compare the new team selection resulting from optimizing the efficiency matrix to the team that was originally selected as the best defensive unit. There will of course be the added position of DH, which is nevertheless shown as an added player in Table 22.

	Position									
Player	P	C	1B	2B	3B	SS	LF	CF	RF	DH
1	0.5849	0.6751	0.6478	0.6202	0.6256	0.5994	0.6562	0.6562	0.6562	0.6087
2	0.8995	0.8422	0.8729	0.8700	0.8405	0.8762	0.7883	0.7883	0.7883	0.9021
3	0.5725	0.5957	0.6038	0.5711	0.5432	0.5336	0.6146	0.6146	0.6146	0.4564
4	0.8009	0.6470	0.6052	0.5905	0.5816	0.5863	0.5886	0.5886	0.5886	0.5044
5	0.8987	0.7232	0.7111	0.6877	0.6725	0.6772	0.6946	0.6946	0.6946	0.6606
6	0.6169	0.7072	0.6778	0.6412	0.5960	0.6162	0.5841	0.5841	0.5841	0.4370
7	0.7551	0.4649	0.4954	0.5223	0.5336	0.5348	0.6400	0.6400	0.6400	0.5552
8	0.8634	0.5772	0.6302	0.6574	0.6744	0.6803	0.6374	0.6374	0.6374	0.6636
9	0.6262	0.2753	0.2771	0.2838	0.2871	0.2775	0.4213	0.4213	0.4213	0.3166
10	0.6572	0.1316	0.2279	0.2786	0.2857	0.2869	0.2749	0.2749	0.2749	0
11	0.2551	0.2188	0.2447	0.2626	0.2501	0.2585	0.2876	0.2876	0.2876	0
12	0.4744	0.3179	0.3104	0.3168	0.3043	0.3127	0.3513	0.3513	0.3513	0.4628
13	0.7901	0.5284	0.5747	0.5923	0.6063	0.6027	0.6550	0.6550	0.6550	0.2765
14	0.6660	0.4808	0.4587	0.4212	0.3846	0.3775	0.4580	0.4580	0.4580	0
15	0.1224	0.4363	0.4239	0.3922	0.3610	0.3610	0.4413	0.4413	0.4413	0.2259
16	0.2782	0.2913	0.2210	0.1960	0.1835	0.1835	0.2045	0.2045	0.2045	0.2765
17	0.5271	0.5613	0.5770	0.5693	0.5818	0.5568	0.6603	0.6603	0.6603	0.3687
18	0.7917	0.7360	0.7182	0.6813	0.6804	0.6459	0.7862	0.7862	0.7862	0.6963
19	0.7369	0.4450	0.4038	0.3772	0.3579	0.3626	0.3203	0.3203	0.3203	0.2497
20	0.7185	0.7245	0.6683	0.6212	0.5894	0.6025	0.6053	0.6053	0.6053	0.5470
21	0.9892	0.8622	0.8718	0.8439	0.8216	0.8251	0.8469	0.8469	0.8469	1

Table 22: Maximal aggressive team compared with optimal defensive team

There is one extra factor to take into consideration before calculating the pitching and batting scores for this team, which has an effect on both these scores. The DH forfeits his batting privileges if he takes the field and is therefore not eligible for pitching duties. The selected pitcher on the other hand is replaced by the DH in the batting line-up and therefore the team's batting score is calculated as the aggregate of the other nine players' individual batting scores. The summary of this team, which is supposed to be the strongest batting unit, is given in Table 23.

Position	**P**	**C**	**1B**	**2B**	**3B**	**SS**	**LF**	**CF**	**RF**	**DH**
Player	5	20	6	2	1	8	18	17	13	21
Pitching score		2.85	**Pitchers:**		5	2	8	13		
Batting score		4.625								

Table 23: Maximally aggressive team selection and scores

The results are indeed baffling when compared to those of the strong offensive team obtained in the previous section. Not only did the pitching score drop from 2.9 to 2.85, but the batting score decreased quite dramatically from 4.948 to 4.625. This batting score is in fact even lower than that of the team selected as a moderately offensive unit. The causes of these stark contradictions are player 5 and player 21, both of whom excel in both batting and pitching.

Since the solitary weight for the DH position is on the batting test, the far-most right column in Table 23 is exactly the same as the relative scores in the batting test. The player who performed best in this test, player 21, is naturally selected as the DH. However, this player has been part of the four-pronged pitching attack in all the previous team selections and is now lost to the pitching staff. This explains the slight loss in pitching power from the previous team combinations. Similarly, player 5 had been a valuable contributor to the team in terms of batting ability and has now been forcibly replaced in the batting line-up by the inclusion of the DH.

This unexpected occurrence, if anything, teaches us the value case-specific consideration and understanding the finer points of any professional field. The obvious solution is to include player 5 in the batting line-up by letting one of the other three slightly weaker pitchers pitch. Since players 2, 8 and 13 each have a pitching score of 0.7, the one with the lowest batting score will be chosen as starting pitcher. Player 13, with a batting score of only

0.233, will therefore be the leading pitcher and will be replaced by player 21 in the batting line-up. This modification to the team causes the batting score to increase to 4.948 once more, as the batting line-up consists of the same nine players selected in the strongly offensive team.

Instead of adding an extra dimension to the team in terms of offence, the addition of the DH rather upsets the balance of this particular team. Even after correcting the resulting imbalance it cannot concluded that the team has been improved compared to the heavily offensive team selected in section 4.2.3.

One should also be aware of the fact that simply swapping a fielding player with a pitcher negates the careful placement of all the players resulting from applying the Hungarian algorithm. Each position's occupant is placed in that specific position entirely dependent on all the other field placements. When a specific player is fixed in a position, as is now the case with player 13 being forced to pitch, the fielding assignments need to be recalibrated. We will now discuss a method of achieving this with the introduction of relief pitchers.

4.3 Introducing Relief Pitchers

Professional pitchers rarely pitch through all nine innings in a match and are usually replaced after four to six innings or, depending on their performance, earlier if necessary. The substitute pitchers, referred to as relief pitchers, are not involved in the match until called into action by the manager.

The intended application of this research is to a large extent on amateur and novice baseball teams. One would not expect pitchers who receive little opportunity to pitch in a match situation to be able to pitch the entire nine innings. Furthermore, a small recreational team is unlikely to have an abundance of specialist pitchers at their disposal to replaced fatigued pitchers with. We will now illustrate a method of determining the optimal rotation of pitchers within a selected team of nine players and discover how the fielding assignments are affected.

The group of players participating in this research yielded several players with some pitching ability. This is evident when examining the pitching scores of the different team combinations identified so far in this chapter. The team constructed for strong offensive purposes in section 4.2.3 produced the same pitching score as the initial, purely defensive team. We will be using this selection, shown earlier in Table 20, to investigate the effect of relief pitchers.

The four identified pitchers and their pitching scores are shown in Table 24. The optimal team assignment has player 5 pitching.

Player	Pitching score
5	0.75
21	0.75
2	0.7
8	0.7

Table 24: The pitching staff

To determine which of the relief pitchers upsets the balance of the team the least we first need a benchmark to compare the altered teams with. Since we are using the same efficiency matrix for this entire exercise we can again use the team efficiency (E) as a measurement.

The efficiency matrix and placement of the team being considered are displayed in Table 25.

	Position								
Player	**P**	**C**	**1B**	**2B**	**3B**	**SS**	**LF**	**CF**	**RF**
1	0.5849	0.6751	0.6478	0.6202	0.6256	0.5994	0.6562	0.6562	0.6562
2	0.8995	0.8422	0.8729	0.8700	0.8405	0.8762	0.7883	0.7883	0.7883
5	0.8987	0.7232	0.7111	0.6877	0.6725	0.6772	0.6946	0.6946	0.6946
6	0.6169	0.7072	0.6778	0.6412	0.5960	0.6162	0.5841	0.5841	0.5841
8	0.8634	0.5772	0.6302	0.6574	0.6744	0.6803	0.6374	0.6374	0.6374
17	0.5271	0.5613	0.5770	0.5693	0.5818	0.5568	0.6603	0.6603	0.6603
18	0.7917	0.7360	0.7182	0.6813	0.6804	0.6459	0.7862	0.7862	0.7862
20	0.7185	0.7245	0.6683	0.6212	0.5894	0.6025	0.6053	0.6053	0.6053
21	0.9892	0.8622	0.8718	0.8439	0.8216	0.8251	0.8469	0.8469	0.8469

Table 25: Isolated efficiency matrix with player 5 pitching

The fielding efficiency, which is at a maximum in this arrangement is:

$$E = \max_{Y} Z(Y) = \sum_{i=1}^{n} \sum_{j=1}^{k} y_{ij}\, Q_{ij} = 6.7982$$

Again, expressed as a percentage: $E_p = {}^{6.7982}/_{0.09} = 75.54\%$. This is the benchmark we will be using.

Now to obtain the optimal fielding placements with one of the other possible pitchers on the mound, we fix the pitching position by removing the pitching column and the relief pitcher's row of scores from the efficiency matrix. The remaining eight players, including player 5, are now shuffled around in a way to yield maximum overall team efficiency. One example is shown in Table 26, where player 21 is fixed in the pitching position.

	Position								
Player	**P**	**C**	**1B**	**2B**	**3B**	**SS**	**LF**	**CF**	**RF**
1	0.5849	0.6751	0.6478	0.6202	0.6256	0.5994	0.6562	0.6562	0.6562
2	0.8995	0.8422	0.8729	0.8700	0.8405	0.8762	0.7883	0.7883	0.7883
5	0.8987	0.7232	0.7111	0.6877	0.6725	0.6772	0.6946	0.6946	0.6946
6	0.6169	0.7072	0.6778	0.6412	0.5960	0.6162	0.5841	0.5841	0.5841
8	0.8634	0.5772	0.6302	0.6574	0.6744	0.6803	0.6374	0.6374	0.6374
17	0.5271	0.5613	0.5770	0.5693	0.5818	0.5568	0.6603	0.6603	0.6603
18	0.7917	0.7360	0.7182	0.6813	0.6804	0.6459	0.7862	0.7862	0.7862
20	0.7185	0.7245	0.6683	0.6212	0.5894	0.6025	0.6053	0.6053	0.6053
21	0.9892	0.8622	0.8718	0.8439	0.8216	0.8251	0.8469	0.8469	0.8469

Table 26: Isolated efficiency matrix with player 21 pitching

This team combination yields a team fielding efficiency of 74.81%.

The process is repeated by fixing player 2 and player 8 in the pitching position and then calculating the team efficiency after optimally re-assigning the rest of the team. The results are summarized in Table 27.

Pitcher	**Team efficiency**
5	75.54%
8	75.12%
21	74.81%
2	73.35%

Table 27: Team efficiency with different pitchers

The choice of order in which to let these four players pitch is left to the captain or coach in charge. The two logical alternatives are:

- Rank the relief pitchers first according to their pitching score. If there is a tie, base the decision on the team efficiencies. The order in which the relief pitchers will then be called upon is: player 21; player 8; player 2.
- Rank the relief pitchers solely based on the resulting team fielding efficiency. The order of relief pitchers is then: player 8; player 21; player 2.

Since the team efficiency values include the pitching score one might be inclined to suggest that the chosen team should operate with the second of these policies. Note that whilst this pitching rotation schedule might be most effective in this specific case it should not be assumed as such for teams in general.

4.4 An Overview of the Team Selections

The different team selections calculated in this chapter are the result of attempting to create teams with certain specific strengths, especially in terms of balance between offence and defence. We have also suggested descriptive statistics with which to measure and compare this balance. A summary of the different team statistics are given in Table 28.

Team strategic description	Batting score	Pitching score
Optimal defence	4.535	2.9
Medium offence	4.669	2.9
Strong offence	4.948	2.9
Strong offence including DH	4.625	2.85
Strong offence including DH with player 13 as pitcher	4.948	2.85

Table 28: Teams summary

The team that stands out above the others is the one optimized for a strongly offensive strategy. This team yields the highest batting and pitching scores of all combinations tested whilst still maintaining reasonable weights on the fielding tests.

It is worth noting again the unexpected drop in the batting score when a specialist DH is played. It cannot be overemphasized that, although the system is designed to optimally select and place a team for overall performance, the output must be understood in order to adjust for unexpected results. Furthermore, the input also requires careful deliberation and a fundamental understanding of what is to be achieved, not only when considering the team as a whole, but also with regards to specific tactics or plans.

Of the 21 players attending the trials, only 11 were included in any of the teams. Once again it is the user's responsibility to have an in-depth look at the data and not just read the results. There is, from personal opinion, at least one player not selected in any of the teams who definitely has the talent and ability to be playing alongside the other players in the group. Further research on the subject would do well to devise a way of identifying these fringe players.

There is, as is the case with any new system, room for improvement. Some of the skills tests need refining and participants should perhaps be given more attempts in some of the tests, especially the pitching test.

Another useful tool of analysis would be a method of objectively comparing the fielding ability of different teams. Seeing as how the skills test will inevitably be used as measure of ability, a way needs to be found in which subjective weights need not be used in order to calculate a total fielding score.

Chapter 5 - Conclusion

This research set out to develop a system with which to assess the skill levels of players with regards to different aspects of baseball and to not only select the optimal team for a predetermined strategy, but also optimally assign them to fielding positions. The ideal system is one which can easily be applied to players in various levels of baseball expertise, although the novice baseball clubs in South Africa are the primary target for this research.

The basic concept, taken from Baeva *et al.* (2008), is customised and the results thoroughly analysed. The methodology *first* of all entails a series of skills tests specific to baseball application. *Secondly,* after applying the appropriate transformation on the data a set of weights relating the tests to the different defensive positions is multiplied with this modified data set. *Finally,* the resulting matrix is then optimised with the use of the Hungarian algorithm, a mathematical solution to the assignment problem formally defined in 1955 by H. W. Kuhn.

The ultimate goal of a baseball team is to maximise the difference between runs scored and runs allowed. It is quite an ambitious task to find the correct, balanced combinations of weights and variables with which to conclusively achieve this task. Although the combinations used in this research and, therefore, the results obtained are open to interpretation, insightful discoveries regarding a group of baseball players can nevertheless be made courtesy of the procedure.

5.1 Further Research

There are several aspects of the process which can either be improved or expanded upon. The first of these is regarding the skills tests. Future application of the system should involve experimentation with different tests or perhaps just modified versions of the tests applied in this research until the method is refined to the ideal. Another characteristic that might be quite influential is decisiveness, *i.e.* the ability of making good decision quickly.

As more data is gathered on a much larger set of players, the effect of outliers or exceptional players within a group must also be investigated. These players would have a large impact on the entire group's scores when calculating the relative scores. Methods of smoothing these ranked scores might prove useful.

Other possible solutions to the problem of outliers might be to introduce either a ranking or a rating system. A rating is understood as a fixed value that is allocated to all players whose test scores fall in a certain interval. A ranking, on the other hand, is simply a value indicating which position a player achieved in the order of merit for each test.

Application of the methodology to other team sports with different, perhaps more diverse positional requirements, should be encouraged. Apart from pitching, which has developed into an art form in itself, all of the other defensive roles in baseball require the same basic abilities of reliable catching and accurate throwing. Baseball offense is a quite one-dimensional concept as well.

In a sport such as American football, one would expect to see a greater variation in each player's innate characteristics, as defensive and offensive roles both require a wide range of abilities. In general the skills tests conducted would test aptitude for a specific defensive or offensive role, resulting in the weights assigned to different positions being distributed across fewer tests. The choice of weights is, therefore, a less extensive task than is the case when aiming to optimize a baseball team for a specific goal.

Other sports in which this paper's application might prove useful include basketball, netball, rugby (careful consideration would be required) and perhaps even paintball.

Multi-agent systems of course do not necessarily refer just to sports teams. Future researchers are to free their thoughts to the possibility of application to any collaborative task, whether it is man or machine, recreation or science.

References

- Baeva, S. Komarevska, L. Nedeva, C. & Trenev, L., 2008. Multi-criterial Decision Making for Selection and Assignment of Sportsmen in Team-games. *Applications of Mathematics in Engineering and Economics*, 34, pp. 451-57.
- Baker, J. Côté, J. & Abernethy, B., 2003. Sport-Specific Practice and the Development of Expert Decision-Making in Team Ball Sports. *Journal of Applied Sport Psychology*, 15, pp. 12-25.
- Bergan, R., 1982. *Sports in the Movies*. New York: Proteus Books.
- Berger, J.O. & Deely, J., 1988. A Bayesian Approach to Ranking and Selection of Related Means with Alternatives to Analysis-of-Variance Methodology. *Journal of the American Statistical Association*, 83 (402), pp. 364-373.
- Castello, B., 2010. *The Hungarian Algorithm.* Baltimore: Johns Hopkins University. [Course notes]
- Cao, Y., 2008. *Munkres Assignment Algorithm.* [Online] (Updated 27 June 2010) Available at: http://www.mathworks.com/matlabcentral/fileexchange/20328-munkres-assignment-algorithm
  [Accessed 12 November 2010]
- Faurot, D.J., 2001. Equilibrium Explanation of Bargaining and Arbitration in Major League Baseball. *Journal of Sports Economics*, 2 (1), pp. 22-34.
- Frank, A., 2004. On Kuhn's Hungarian Method—A tribute from Hungary. *Naval Research Logistics*, 52 (1), pp. 2-5.
- Kuhn, H. W., 1955. The Hungarian Method for the Assignment Problem. *Naval Research Logistic Quarterly*, 2 (1-2), pp. 83-97.
- Martello, S., 2010. Jeno Egerváry: From the Origins of the Hungarian Algorithm to Satellite Communication. *Central European Journal of Operations Research*, 18, pp. 47–58.
- MATLAB, 2009b. *MATLAB 7.9*. Natick, MA: The MathWorks Inc.
- MLB Advanced Media, L.P., 2010. [Online] (Updated 29 December 2010). Available at:
  http://mlb.mlb.com/stats/sortable_player_stats.jsp?c_id=mlb&tcid=mm_mlb_stats

[Accessed 29 December 2010]

- Munkres, J., 1957. Algorithms for the Assignment and Transportation Problems. *Journal of the Society for Industrial and Applied Mathematics*, 5 (1), pp. 32-38
- Raimondo, H., 1983. Free Agents' Impact on the Labor Market for Baseball Players. *Journal of Labor Research*, 4 (2), pp. 183-193.
- Randolph Youth Softball and Baseball Association, 2006. [electronic print] (Updated 4 December 2006).
  Available at: http://www.rysba.com/TipsAndDrills/strike.zone.htm
  [Accessed 7 December 2010]
- Rencher, A.C., 2002. *Methods of Multivariate Analysis*. 2nd ed.
  New Jersey: John Wiley & Sons, Inc.
- Schlegel, J., 2009. *Classic an international celebration.* [Online] (Updated 24 March 2009) Available at:
  http://mlb.mlb.com/wbc/2009/news/article.jsp?ymd=20090324&content_id=4058588&vkey=wbc&team=
  [Accessed 25 February 2010]
- Schrijver, A., 2003. *Combinatorial Optimization: Polyhedra and Efficiency (Algorithms and Combinatorics, Vol. 24)*. Springer.
- StataCorp, 2007. *Stata Statistical Software: Release 10*. College Station, TX: StataCorp LP.
- Tichavský, P. & Koldovský, Z., 2004. Optimal Pairing of Signal Components Separated by Blind Techniques. *IEEE Signal Processing Letters*, 11 (2), pp. 119-122.
- Tidhar, G. Rao, A.S. & Sonenberg, E.A., 1996. Guided Team Selection. *In*: 'Proceedings of the Second International Conference on Multiagent Systems' (ICMAS), Kyoto, Japan, pp. 369-376.
- Washingtonprepsports.com, 2010. *Baseball as Prep Sports*. [Online] (Continually updated). Available at: http://www.washingtonprepsports.com/a-brief-description-on-baseball-as.html
  [Accessed 29 November 2010]