Kinesiologia Slovenica, 21, 2, 22–30 (2015), ISSN 1318-2269

Šuker Dragutin[1]
Grgantov Zoran[2]
Milić Mirjana[2]

INTRARATER AND INTERRATER RELIABILITY OF THE SPIKING EFFICIENCY ASSESSMENT IN TOP MEN'S VOLLEYBALL

ZANESLJIVOST OCENJEVANJA NAPADALNEGA UDARCA V VRHUNSKI MOŠKI ODBOJKI

ABSTRACT

On a sample of 11 volleyball sets played in the Men's Champions League, intrarater and interrater reliability was analyzed in the rating of 503 spikes in attack and counterattack. Three reliability components (rater association, rater bias and rater distribution) were analyzed separately. The reliability of the first rater (the author of the study) was analyzed by employment of the test-retest method. In the process, correlation and differences of the same rater were analyzed at two points of time. This way, the rater's consistency in spike efficacy rating was tested. Repeated measurement was conducted on the same group of sets, four weeks after the first one, in order to avoid any influence of the first measurement exerted on the second measurement. An additional check of the first rater's reliability was conducted by comparison of his rating to the second (expert) rater's rating on the same sample of 11 randomly selected sets at the first point of time. Pearson's correlation coefficient was applied to determine the raters' maximum level of concordance at two points of time, as well as a very high level of concordance between the rater and the expert. The dependent samples t-test disclosed no differences between the raters at two points of time, while the independent samples t-test showed very small differences between the rater and the expert. The very high level of similarity in rating distribution between the raters in repeated measurements and the experts substantiates the very high level of reliability in the rating of events analyzed.

Key words: Volleyball, spiking, rater association, rater bias, rating distribution

[1]Polytechnic of Rijeka, [2]University of Split, Faculty of Kinesiology

Corresponding author
Mirjana Milić, PhD
[2]University of Split, Faculty of Kinesiology, Teslina 6, Split – 21000, Croatia
Phone number: +385 98/989-7780
E-mail: mirjanam@kifst.hr

POVZETEK

Na vzorcu 11-ih odbojkarskih setov, odigranih v moški odbojkarski Ligi prvakov, so bile analizirane zanesljivosti v 503 udarcih v napadu in protinapadu. Tri komponente zanesljivosti so bile ločeno analizirane. Zanesljivost prvega ocenjevalca je bila analizirana z uporabo metode preizkušanja in ponovnega preizkušanja. V tem procesu so korelacije in razlike isti ocenjevalci analizirali na dveh časovnih točkah. Na ta način je bila testirana doslednost ocenjevalca v ocenjevanju učinkovitosti napadalnega udarca. Ponovne meritve smo izvedli na isti skupini setov, štiri tedne po prvih meritvah, v izogib vpliva prvih meritev na druge meritve. Dodatno preverjanje zanesljivosti prvega ocenjevalca je bila izvedena s primerjavo z drugim ocenjevalcem (strokovni ocenjevalec) na istem vzorcu 11 naključno izbranih setov v prvi časovni točki. Pearsonov koeficient korelacije je bil uporabljen za določitev najvišje ravni skladnosti ocenjevalcev na dveh časovnih točkah, kot tudi za zelo visoko stopnjo skladnosti med ocenjevalcem in strokovnim ocenjevalcem. Z uporabo T-testa za odvisne vzorce nismo odkrili nobenih razlik med ocenjevalcema na dveh časovnih točkah, medtem ko je uporaba T-testa za neodvisne vzorce pokazala zelo majhne razlike med ocenjevalcem in strokovnim ocenjevalcem. Zelo visoka stopnja podobnosti v razdelitvi ocenjevanja med ocenjevalcem v ponovljenih meritvah in strokovnim ocenjevalcem utemeljuje zelo visoko stopnjo zanesljivosti v oceni analiziranih dogodkov.

Ključne besede: odbojka, napadalni udarec, zanesljivost

INTRODUCTION

By successful performance of a spike in attack and counterattack, more points are scored than in all other elements of the volleyball game together (Voight and Veter, 2003; Stanganeli, Dourado, Oncken, Mançan and Da Costa, 2008). Studies conducted to analyze correlation of various phases of the volleyball match to success at the tournament point out that the spike in attack and spike in counterattack phases were significant predictors of success at the tournament (Zang, 2000; Marelić et al., 2004; Grgantov, 2005; Zetou et al., 2007; Monteiro et al. 2009; Rodriquez-Ruiz et al., 2011; Inkinen et al., 2013). For that matter, it is important that game actions be recorded in a precise and reliable way, by means of notational analysis.

In legacy studies, spiking efficacy was most often rated on a three to five-point scale. For instance, Oliveira et al. (2005); Ramos, Nascimento and Collet (2009) are rating spike efficacy on a three-point scale (point scored, ball remained in play and performance error). Matias and Greco (2011) used a four-point spike efficacy scale. The lowest (1) and the highest (4) rating denote an error or, respectively, a point scored. Rating 2 is assigned to all spikes after which the ball remained in play in a way to prevent a varied counterattack by the opposing team or to enable the team that has performed the spike to launch a varied repeat attack. Rating 3 denotes situations where after the spike the ball remained in play in a way to enable the opposing team to launch a varied counterattack or to prevent the team that has performed the spike to launch a varied repeat attack. Bergeles et al. (2009) use a five-point scale. As in the previous scales, the lowest rating (1) denotes a performance error, while the highest rating (5) denotes a point scored. Rating 2 is a poor performance after which the opposing team has an advantage in the further course of the rally, Rating 3 is an average, neutral performance after which neither team has any advantage in the further course of the rally, and Rating 4 is a good performance after which the team that has made contact with the ball has an advantage in the further course of the rally.

In some studies, each point of the scale is treated as a separate variable (for instance Marelić, 1994; Marcelino et al., 2010; Bergeles and Nikolaidou, 2011; Claver et al., 2013), while in other studies, these points were used to derive various efficacy coefficients (Marcelino et al., 2008; Drikos et al., 2009; Zadražnik et al., 2009). Marcelino et al. (2008) have found that spike performance has the highest influence on a team's ranking at a tournament, but only when variables are stated in relative values, i.e. when efficacy coefficients and percentage were calculated. Drikos et al. (2009) also proved that in individual phases of the volleyball game efficacy coefficients are better predictors of the team's performance as compared to individual variables that served as a base for calculation (such as number of points scored, number of errors). For that matter, the methods used for calculation of performance efficacy coefficients for the individual phases of the game will vary. In the majority of cases, the efficacy coefficient will be calculated by applying the Coleman (1975) method, i.e. by adding up all performances in specific categories (quality levels), multiplying them by the rating (number) related to the respective category and dividing them by the total number of performances in all categories. Such calculation method proved to be of higher quality in other studies as well (Marelić et al., 2004; Grgantov, 2005; Zadražnik et al., 2009).

Regardless of which spike quality rating scale is applied, the important thing is that the individuals who are rating performance quality during the tournament are well trained and focused while making observations and recording data. Otherwise, substantial mistakes could be made during tournament data collection that might impair reliability of the entire analysis, or even call it into

question. Nevertheless, Hughes et al. (2002), having analyzed scientific papers (N=72) published in the field of notational analysis, established that almost 70% of authors had not performed any efficacy and performance indicator reliability analysis whatsoever. The reliability of data collected in notational analysis shows to what extent such collected data give a clear picture of what was actually going on during the match.

Different causes may result in rater disagreement on a given case. With interval-level data, these various causes have effects that can be broadly grouped into three categories: effects on the correlation or association of raters' ratings, rater bias, and rater differences in the distribution of ratings (Uebersax, 2006).

In reliability analyses, very often reliability indices are applied (for instance, the intraclass correlation or coefficient of concordance). Such indices summarize all components of disagreement into a single number. This results in loss of important information about the exact causes of rater differences, therefore researchers are unable to identify the steps necessary to improve agreement.

Hence, we are of the opinion that only separate analysis of different components of agreement can enable us to precisely specify the procedures that can improve rater reliability.

Therefore, the purpose of this study is to determine intrarater and interrater reliability of spike efficacy assessment and, in doing so, analyze each component of rater agreement (association, bias, and distributional differences) separately.

MATERIALS AND METHODS

The entities in this study are the teams in a volleyball set. 11 sets or, respectively, 22 teams were analyzed. The sets were randomly selected from the overall sample of Champions League matches in the period from the year 2008 to 2012. A total of 503 actions (spikes) were analyzed. During the matches, spike efficacy was rated on the basis of a four-point scale: Rating 4 denotes a spike that resulted in a point scored. These are, for instance, spiked balls which hit the opposing team's court or which, after an opponent's block or defensive player's action, were deflected to the floor out of bounds.

Rating 3 is assigned to all spike performances after which the spiking team keeps a dominant position in the further course of the rally. An example for that would be when, after spiking and blocking, the ball is again deflected to the spiking team's playing area, but in such a way that the setter can organize a counterattack with a larger number of options. Another possibility is when the spiked ball passes to the opposing team's possession, but they are unable to organize a good counterattack (the defensive player has played an imprecise first ball that caused the setting or spike in counterattack to be performed under more difficult conditions, with a very limited number of options for the setter).

The first possibility to assign a Rating 2 to a spike performance is when the spiked ball is deflected by the opponent's block to the spiking team's playing area, but in such a way that counterattack action is made substantially more difficult. Another possibility would be that the opposing team has managed to play the spiked ball in their playing area in a way enabling them to organize a successful counterattack.

Rating 1 is assigned when an error is made in spike performance (the ball was driven into the net or lands outside the court, the ball rebounds from an opponent's block and lands in the spiking team's playing area, or the spiker has violated the rules of the game – he has made contact with the net, prolonged his palm's contact with the ball and the like).

The collected data served as a basis for calculation of the spike efficacy coefficient in a set. The coefficient may have a value from one to four; it is calculated in the following manner: *ideal* spikes are multiplied by four, *good* spikes by three, *inadequate* spikes by two and performance errors by one. After that, the obtained value is divided by the total number of spikes performed in the respective set.

Data collection or, respectively, spike performance rating was conducted by means of a laptop computer and a specialized computer software, the *Data Volley Professional 3.2.1.* program (Data Project, Salerno Italy).

Data were collected from video recordings made by a video camera that was placed in such a position to clearly cover the entire court and all players in the court.

Based on the video recordings, the spike quality was independently rated by two raters: by one of the authors of this study who has long-standing experience as a volleyball player and coach (Rater 1) and by an expert who is a long-standing official statistician with Slovenia's Bled Team that competed in the Men's Volleyball Champions League (Rater 2).

In order to establish test-retest reliability, Rater 1 has been rating events twice on the same group of sets, with a time difference of four weeks.

Upon completion of data collection, the data were copied from the Data Volley application to *Microsoft Excel* files where they were prepared for further analysis.

After that, the data were further processed in the *Statistica for Windows* statistics software package.

Data processing methods were applied in order to calculate three separate reliability components – rater agreement, rater bias and rating distribution. For calculation of the rater agreement reliability component, Pearson's correlation coefficient was applied. This coefficient was calculated to compare ratings performed by Rater 1 at two points of time (test-retest reliability) and to determine the concordance between Rater 1 and Rater 2.

Rater bias was tested by dependent samples t-test for assessment of Rater 1 test-retest reliability in the first and second measurement, while the independent samples t-test was used to analyze differences between Rater 1 and Rater 2.

RESULTS

Table 1 shows descriptive indicators (arithmetic means and standard deviation) of the spike efficacy coefficient, separately for Rater 1 in the first and second measurement and for the expert (Rater 2) in the first measurement. It is quite evident that in both measurements Rater 1 has the same average values and the same variance in results in terms of average values of the spike efficacy coefficient. The descriptive indicator values of Rater 2 were also very similar to the values of Rater 1. Pearson's correlation coefficient was used to assess the rater agreement reliability

Table 1. Descriptive indicators of spike efficacy indicators and intrarater and interrater association and bias in spike efficacy assessment

	$M_1 \pm SD_1$	$M_2 \pm SD_2$	r	t-test	p	$M_1 \pm SD_1$	$M_E \pm SD_E$	r	t-test	p
Sp. coef.	2,85±1,21	2,85±1,21	1,00	0,00	1,00	2,85±1,21	2,84±1,20	0,98	0,05	0,96

Legend: Sp. coef. – spike efficacy coefficient; **r** – Pearson's coefficient of correlation between measurement lots, $\mathbf{M_1}$ – arithmetic means of Rater 1 in the first measurement, $\mathbf{M_2}$ - arithmetic means of Rater 1 in the second measurement, $\mathbf{SD_1}$ – standard deviation of Rater 1 in the first measurement, $\mathbf{SD_2}$ - standard deviation of Rater 1 in the second measurement, $\mathbf{M_E}$ – arithmetic means of the expert rater's rating (Rater 2), $\mathbf{SD_E}$ – standard deviation of the expert rater's rating (Rater 2), **t-test** – test value in testing of significance of differences between average values of the first and second measurement and between two raters * - significant difference on the level p≤0,05.

component. A maximum value of this coefficient was obtained by analysis of correlation of Rater 1 results in the first and second measurement. A comparison of the two raters' results also gave an almost maximum correlation coefficient value.

Zero values of dependent samples t-test were obtained in rater bias analysis, i.e. in the analysis of Rater 1 result variances in the first and second measurement. In addition, very small differences were found in the independent t-test that was applied to analyze differences in spike efficacy coefficients (rater bias) between Rater 1 and Rater 2.

The third reliability component, rating distribution, was assessed on the basis of diagrams (histograms), i.e. based on frequency of recording of individual spike categories on a scale from 1 (spike performance error) to 4 (point scored by spike).

The histograms 1-3 show the rating distribution of Rater 1 in both measurements, as well as the rating distribution of Rater 2. A comparison of the histograms shows almost identical distribution of the raters' ratings, which points to a very high level of this reliability component.

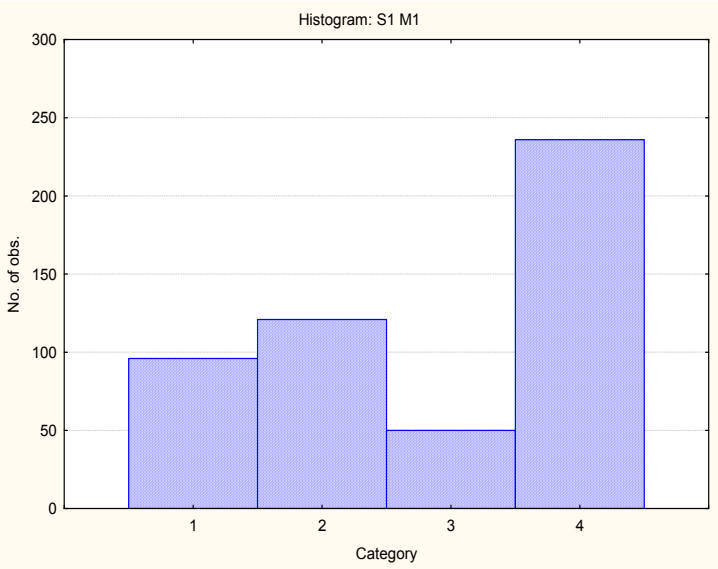

Diagram 1 - Rating distribution of Rater 1 in the first measurement

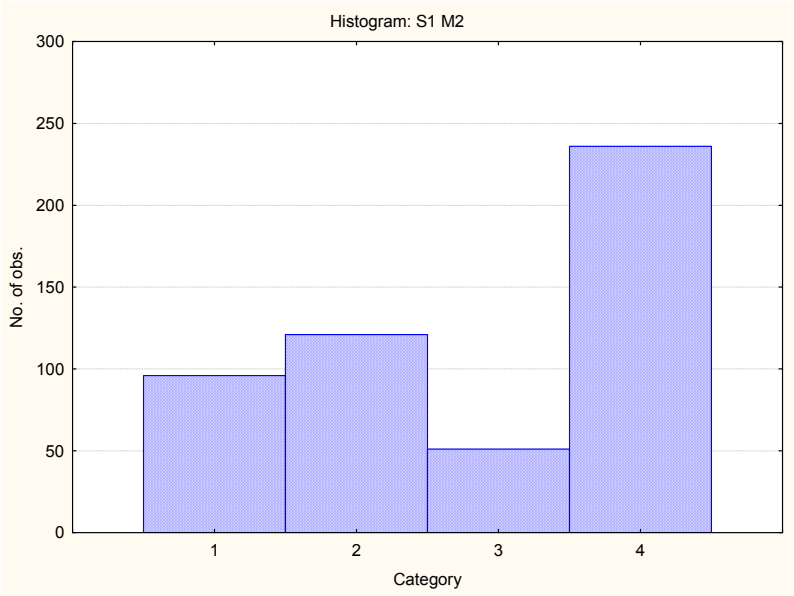

Diagram 2 - Rating distribution of Rater 1 in the second measurement

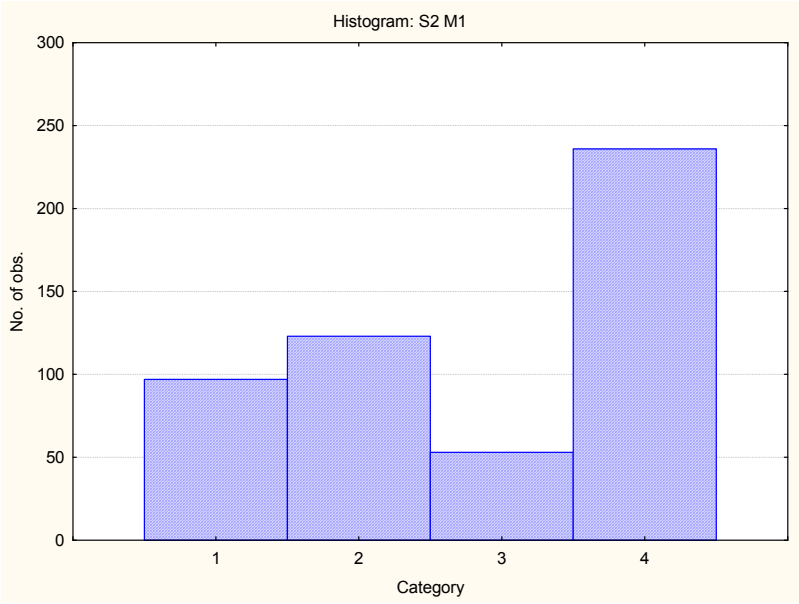

Diagram 3 - Rating distribution of Rater 2 (expert)

DISCUSSION AND CONCLUSION

Analysis of the distribution of raters' ratings leads to the conclusion that, most often, top teams score direct points by spikes (just slightly below 50% of situations), while slightly less than 20% of spikes result in an performance error. This is in line with the results of Oliveira et al. (2005)

who, during the Volleyball World League 2003, established that the teams score points in 47% of attacks on average, while 18% of attacks result in an error.

Separate analysis of the three reliability components showed an almost maximum level of raters' reliability in spike assessment. There are several causes that may have contributed to such a high level of agreement of Rater 1 at two points of time, as well as agreement between Rater 1 and Rater 2. Both raters are experienced volleyball experts who have been working as coach or, respectively, statistician for many years. The applied scale for spike quality assessment is common in volleyball (Marelić et al., 2004; Grgantov, 2005; Zadražnik et al., 2009) and frequently used by coaches, in cooperation with statisticians, for evaluation of the players' performance in various technical and tactical elements of the volleyball game, thus also for assessment of spike efficacy at the tournament. Therefore, it can be concluded that both raters were familiar with the rating scale and that they have often applied the same in their work with volleyball players. What's more, the scale is simple to use and comprehensible, and experienced experts should have no difficulties in assigning spike performances to the respective categories in terms of quality. Identification of errors and ideal performances (points scored) is very simple and eventually defined by the referee's call. After such performances, the team's rally ends and all one has to do is to check which team got the right to serve next in order to come to a conclusion whether the previous spike resulted in a point scored for the team, or in an error. A somewhat more demanding task is to assess the quality of the spike after which the rally continues. In such situations it must be assessed which team has better prospects to score a point after the spike or, respectively, which team has an advantage in the further course of the rally.

In addition to the simplicity and comprehensibility of the measuring scale, reliability may be impaired for numerous other reasons. For instance, raters may differ by circumstances or technical and tactical details which they are taking into consideration when assessing the individual event (spike). Furthermore, different raters may assess the same spike performance differently, or combine the same circumstances (the post-spike situation on the field) in different ways, aiming to assign a final rating. In addition, random errors may be made in assessment. For instance, the raters may be obstructed or tired while making the assessment, which will certainly increase the chance of error. Due to such circumstances, it is even uncertain that the same rater will assess the same event in the same way at two points of time. This is why the test-retest reliability coefficient is calculated. The combined impact of the above-mentioned factors would in all likelihood impair the values of Pearson's correlation coefficient. As this did not happen, it can be concluded that the raters have performed their task in a focused and routine manner. However, it must also be pointed out that the Pearson correlation only assesses certain types of disagreement. For instance, it is possible that a rater is constantly assessing events in a way to arrive at a rating that differs equally, by always the same value, from another rater's assessment. In this case, although the raters are assessing each event differently, Pearson's correlation coefficient would be maximal. This is why this coefficient was also combined with the dependent and independent samples t-test, for calculation of the rater bias component (Uebersax 2006).

Rater bias refers to the tendency of a rater to make ratings generally higher or lower than those of other raters. Bias may occur for several reasons. For example, some raters may simply interpret the calibration of the rating scale differently so as to make generally higher or lower ratings. Taking into consideration the zero values of the dependent samples t-test and the very small differences of the independent samples t-test, it can be said that Rater 1 had the same rating criteria in repeated measurements and that his rating criterion was in line with the criterion of Rater 2.

It is also possible to calculate statistical indices that reflect the similarity of one rater's ratings distribution with that of another, or between each rater's distribution and the distribution for all ratings. However, such indices usually do not characterize precisely how two distributions differ - merely whether or not they do differ. Therefore, if this is of interest, it is probably more useful to rely on graphical methods (Uebersax 2006). Analysis of the diagrams or, respectively, histograms 1-3 show almost identical rating distribution of individual events by Rater 1 in both measurements and Rater 2.

This study showed that experienced volleyball experts are very reliable in spike quality assessment. A facilitating factor was certainly the fact that performances were analyzed from video recordings, so that recordings of individual performances could be replayed several times, if required, in order to arrive at a higher quality rating. However, one must bear in mind that sometimes statisticians are required to assess performance of individual technical and tactical elements in real time (in the course of the match). In this case, the reliability of their rating should also be analyzed under such conditions.

In future studies, it would be recommendable to also analyze the reliability of coaches and statisticians who are beginners in assessment of performance of individual volleyball elements. This should be performed, as done in this study, by analysis of the agreement level of the rater's own ratings at two points of time, and by comparison of their rating to the rating of an expert.

REFERENCES

Bergeles, N., Barzouka, K., & Nikolaidou, M. (2009). Performance of male and female setters and attackers on Olympic level volleyball teams. *International Journal of Performance Analysis in Sport, 9*(1), 141-148.

Bergeles, N., & Nikolaidou, M. (2011). Setter's performance and attack tempo as determinants of attack efficacy in Olympic-level male volleyball teams. International *Journal of Performance Analysis in Sport, 11*(3), 535-544.

Claver F, Jiménez R, Gil A, Moreno A, Moreno MP. (2013). Relationship between performance in game actions and the match result. A study in volleyball training stages. *Journal of Human Sport & Exercise,8*(3), 651-659.

Coleman, J. E. (1975). *A statistical evaluation of selected volleyball techniques at the 1974 world's volleyball championships.* Unpublished doctoral dissertation, Brigham Young University, Provo, Utah.

Drikos, S., Kountouris, P., Laios, A., & Laios, Y. (2009). Correlates of team performance in volleyball. *International Journal of Performance Analysis in Sport, 9*(2), 149-156.

Grgantov, Z., Katić, R., & Marelić, N. (2005). Effect of new rules on the correlation between situation parameters and performance in beach volleyball. *Collegium antropologicum, 29*(2), 717-722.

Hughes, M.D. & Bartlett, R.M. (2002). The use of performance indicators in performance analysis. *Journal of Sports Sciences, 20*(10), 739-754.

Inkinen, V., Häyrinen, M., & Linnamo V. (2013). Technical and tactical analysis of women's volleyball. *Biomedical Human Kinetics, 5*(1), 43-50.

Marcelino, R., Mesquita, I., & Afonso, J. (2008). The weight of terminal actions in volleyball. Contributions of the spike, serve and block for the teams' rankings in the World League'2005. *International Journal of Performance Analysis in Sport, 8*(2), 1-7.

Marcelino, R., Mesquita, I., Sampaio, J., & Moraes, C. (2010). Estudo dos indicadores de rendimentoem voleibol em função do resultado do set. *Revista Brasileira de Educação Física e Esporte. 24*(1), 69-78.

Marelić, N. (1994). Utjecaj situacionih parametara u odbojci na rezultat u odbojkaškom setu. *Hrvatski športskomedicinski vjesnik, 9*(2-3), 70-76

Marelić, N., Rešetar, T., & Janković, V. (2004). Discriminant analysis of the sets won and the sets lost by one team in A1 Italian volleyball league–A case study. *Kineziologija, 36*(1), 75-82.

Matias, C.J., & Greco, P.J. (2011). Análise da organização ofensiva dos levantadores da Superliga de Voleibol. *Revista Brasileira de Ciências do Esporte, 33*(4), 1007-1028.

Monteiro, R, Mesquita, I., & Marcelino, R. (2009). Relationship between the set outcome and the dig and attack efficacy in elite male volleyball game. *Journal of Performance Analysis in Sport, 9*(3), 294-305.

Oliveira, R., Mesquita, I., & Oliveira, M. (2005). Caracterização da eficácia do ataque no voleibol de elevado rendimento competitivo: estudo aplicado em equipas masculinas participantes na Liga Mundial 2003. In: Pinto, J. (Ed.). *Estudos 5.* (pp. 156-166). Porto: CEJD/FCDEF-UP.

Ramos, M.H.K.P., Nascimento, J.V., & Collet, C. (2009). Avaliação do desempenho das habilidades técnico-táticas em equipes de voleibol infantil masculino. *Revista Brasileira de Cineantropometria e Desempenho Humano, Florianópolis, 11*(2), 181-189.

Rodriguez–Ruiz, D., Quiroga, M.E., Miralles J.A., Sarmiento, S., de Saá, Y., & García–Manso J.M. (2011). Study of the Technical and Tactical Variables Determining Set Win or Loss in Top-Level European Men's Volleyball. *Journal of Quantitative Analysis in Sports. 7*(1).

Stanganelli, L.C.R., Dourado, A.C., Oncken, P., Mançan, S., & Da Costa, S.C. (2008). Adaptations on Jump Capacity in Brazilian Volleyball Players Prior to the Under–19 World Championship. *Journal of Strength and Conditioning Research, 22*(3), 741-749.

Uebersax J S. 2006. Agreement on Interval-Level Ratings Statistical Methods for Rater Agreement web site, Available at: http://john-uebersax.com/stat/agree.htm. Accessed May 30, 2015.

Voight, H.F., & Veter, K. (2003). The Value of Strength-Diagnostic for the Structure of Jump Training in Volleyball. *European Journal of Sport Science, 3*(3), 1-10.

Zadražnik, M., Marelić, N., Rešetar, T. (2009). Differences in rotations between the winning and losing teams at the youth European volleyball championships for girls. *Acta Universitatis Palackianae Olomucensis, Gymnastic, 39*(4), 33-40

Zetou, E., Moustakidis, A., Tsigilis N., & Komninakidou, A. (2007). Does effectiveness of skill in Complex I predict win in men's Olympic volleyball games? *Journal of Quantitative Analysis in Sports, 3*(4), 1-9.