Reviewer agreement in scoring 419 abstracts for scientific orthopedics meetings

Rudolf W Poolman[1,4,5], Lucien C M Keijser[2], Maarten C de Waal Malefijt[3], Leendert Blankevoort[4], Forough Farrokhyar[1] and Mohit Bhandari[1]

On behalf of the Dutch Orthopedic Association Scientific Committee

[1]Orthopaedic Research Unit, Division of Orthopaedic Surgery, McMaster University, Hamilton Health Sciences-General Hospital, Hamilton, ON, Canada, Departments of Orthopaedic Surgery, [2]Medisch Centrum Alkmaar, the Netherlands, [3]Radboud University Nijmegen Medical Centre, Nijmegen, [4]Orthopaedic Research Centre Amsterdam, Academic Medical Center, University of Amsterdam, [5]Onze Lieve Vrouwe Gasthuis, Amsterdam, the Netherlands
Correspondence RWP: Poolman@trauma.nl
Submitted 06-07-03. Accepted 06-09-23

Background **The selection of presentations at orthopedic meetings is an important process. If the peer reviewers do not consistently agree on the quality score, the review process is arbitrary and open to bias. The aim of this study was: (1) to describe the inter-reviewer agreement of a previously designed scoring scheme to rate abstracts submitted for presentation at meetings arranged by the Dutch Orthopedic Association; (2) to test whether the quality of reporting of submitted abstracts increased in the years after the introduction of the scoring scheme; and (3) to examine whether a review process with a larger workload had lower inter-rater agreement.**

Methods **We calculated intraclass correlation coefficients (ICC) to measure the level of agreement among reviewers using the International Society of the Knee (ISK) quality-of-reporting system for abstracts. Acceptance rate and quality of the abstracts are described.**

Results **Of 419 abstracts, 229 (55%) were accepted. Inter-reviewer agreement to rate abstracts was substantial (0.68; 95% CI: 0.47–0.83) to almost perfect (0.95; 95% CI: 0.92–0.97) and did not change over the eligible time period. A smaller proportion of abstracts were accepted after 2004. The mean ISK abstract score (with a maximum of 100 points) for accepted abstracts ranged from 60 (95% CI: 58–63) to 64 (95% CI: 62–66). The mean ISK abstract score for rejected abstracts varied from 46 (95% CI: 40–51) to 51 (95% CI: 47–55). Average scores for accepted and rejected abstracts did not change with time. The degree of workload of the reviewers did not influence their level of agreement.**

Interpretation **The ISK abstract rating system has an excellent interobserver agreement. Other scientific orthopedic meetings should consider adopting this ISK rating system for further evaluation in a local or international setting.**

Several scoring systems designed to help in selecting abstracts for a scientific orthopedic meeting have been described in the literature, ranging from simple "accept–unsure–reject" systems to multiple-item scoring schemes (Bhandari et al. 2004, Rowe et al. 2006). These previous studies have reported variable inter-rater agreement in scoring abstracts for scientific meetings (ICC or Kappa agreement range –0.12 to 0.81) (van der Steen et al. 2003, Bhandari et al. 2004, Rowe et al. 2006). If peer reviewers do not consistently agree on the quality score, the review process is arbitrary and open to bias (Rowe et al. 2006). Furthermore, the scoring system must be feasible to use. Long and confusing systems are less likely to be used correctly by reviewers with competing time constraints.

The scientific meetings of the Dutch Orthopedic Association (Nederlandse Orthopaedische Vereniging, NOV) are held 2–3 times a year. The

DOI 10.1080/17453670710013807

number of abstracts submitted for presentation at the meeting exceeds the number for which there is time available during the meetings. It is the aim of NOV to maintain a high scientific standard. In the past, accepted abstracts were published in Acta Orthopaedica, formerly Acta Orthopaedica Scandinavica. Now they are accessible through the Acta Orthopaedica website. Two scoring schemes were introduced—one for clinical abstracts and one for basic science abstracts—to help reviewers score the abstracts submitted for presentation at one of the meetings, for subsequent acceptance or rejection. The schemes are thought to help in rating of the methodological quality of an abstract and its newsworthiness. If reviewers disagree on the quality of abstracts, however, this system might fail to select possible high-quality synopses of research projects and the quality of the meeting might suffer. Thus, the abstract scoring scheme should be concise, easy to use, able to identify abstracts of good and poor quality, should not be subject to reviewer disagreement, and should not be time consuming.

The aim of our study was threefold: (1) to describe the level of inter-reviewer agreement on the quality and newsworthiness in schemes to rate clinical and basic science abstracts submitted for presentation at the Dutch Orthopedic Association; (2) to test whether quality of reporting of submitted abstracts has increased over recent years; (3) to determine whether reviewers with a larger workload, i.e. more abstracts to score, had lower inter-rater agreement.

We hypothesized that reviewers would show good agreement in scoring abstracts submitted for presentation at the NOV meetings between 2001 and 2006. Also, we hypothesized that the quality of abstracts would improve over the period 2001–2006. Finally, we hypothesized that the number of abstracts to be scored would influence inter-rater agreement.

Material and methods

Eligibility

For this study, the "submitted abstracts" database of NOV was used. This database consists of accepted and rejected abstracts for NOV's scientific meetings as well as the grading score of the abstracts. The data were retrieved from the existing administrative database used for abstract submission from 2000 to 2006. We concentrated on abstracts of the meetings with subsequent publication on the website of Acta Orthopaedica. Online publication data were not yet available for the 2005 and 2006 meetings during the analysis of the data. The first year the abstract scoring scheme was introduced (2000) was not included in the analysis.

Abstract grading—checklist

Since 2000, the NOV has used a scoring system for evaluation of abstracts that was originally introduced by the International Society of the Knee (ISK), which became part of the International Society of Arthroscopy, Knee Surgery and Orthopaedic Sports Medicine (ISAKOS) (Appendix 1) Contact with the ISAKOS office and a thorough Medline and internet search did not reveal any previous report describing the scoring system. This scoring system involves a score for clinical research abstracts and a score for basic science abstracts. For clinical trials, the score is based on 7 criteria. The score starts with a baseline score of +50 points. The 7 criteria are weighted differently by different adding or subtracting possibilities. Objective criteria (methodological safeguards) have more weight in the total score than subjective criteria (newsworthiness). The scoring system for basic science abstracts consists of 5 items. As in the score for clinical abstracts, the basic science score weighs objective items more strongly than subjective items in calculating the total score. Also, this score starts at +50. The total maximum score is 100 and the minimum score is 0. The scoring system was used in its original form without translation or modification of items.

Abstracts requirements for submission

Abstracts should be submitted in both Dutch and English. The abstract should not exceed 250 words and should consist of 4 parts: introduction and aim of the study, methods, results, and conclusion, i.e. have a structured format.

Criteria for acceptance and rejection of submitted abstracts

The criteria for acceptance were an ISK score of at least 50 points, that the work should not have been

published as a full text manuscript more than 1 year before the meeting, and that the abstract should fit the specific topic of the meeting (for the January meetings only). Finally, in a consensus meeting of the scientific committee's reviewers, the abstracts were accepted or rejected.

Abstract grading—review process

All abstracts were coded in terms of authors and institutions. A minimum of 6 reviewers scored each abstract in a blind fashion (regarding their origin). The reviewers in the group changed gradually, with a maximum change of 2 per year. From 2001 onward, the authors of abstracts were informed about the use of the scoring system in the selection process prior to the deadline for submission. The reviewers did not receive any formal training in the use of the ISK abstract scoring system. Only the scheme as presented in Appendix 1 was given.

Statistics

Descriptive statistics were used to report ISK abstract scores for rejected and accepted abstracts. Categorical variables are reported as percentages and were compared using the chi-squared test. Continuous variables are reported as mean and 95% confidence intervals (CI) and were compared using one-way analysis of variance (ANOVA). We used intraclass correlation coefficients (ICC) described as Cronbach's alpha (95% CI) to measure the level of agreement among the reviewers (van der Steen et al. 2003, Bhandari et al. 2004). Landis and Koch (1997) suggested criteria for interpretation of agreement: 0 to 0.2 representing slight agreement, 0.21–0.40 fair agreement, 0.41–0.60 moderate agreement, and 0.61–0.80 representing substantial agreement. A value above 0.80 is considered almost perfect agreement. In our study, each abstract was reviewed by at least 6 reviewers. The reviewers were conceived as being a random selection from all possible reviewers and abstracts were conceived as being a random factor as well. In each year, reviewers rated all of the abstracts submitted to the meetings and it is known how each reviewer rated each abstract; therefore, the two-way random-effect model for interclass correlation seemed to be the most appropriate for our data analysis. The ICC is interpreted as being generalizable to all possible reviewers. Spearman's correlation coefficient analysis was used to determine the correlation between the number of abstracts submitted and the ICC of agreement between reviewers or the percentage of accepted abstracts.

Table 1. Interclass correlation coefficients and mean ISK abstract scores (N = 419)

Year: month	n_a	n_r	ICC (95% CI)	ISK score mean (95% CI)
2001: V	62	8	0.94 (0.92–0.96)	55 (51–59)
2001: X	11	8	0.90 (0.77–0.97)	53 (47–59)
2002: V	10	6	0.83 (0.60–0.95)	58 (50–66)
2002: X	14	7	0.91 (0.79–0.97)	59 (52–66)
2003: I	45	7	0.87 (0.79–0.93)	60 (57–63)
2003: V	15	6	0.82 (0.64–0.93)	55 (49–60)
2003: X	14	6	0.83 (0.63–0.94)	55 (49–62)
2004: X	39	7	0.68 (0.47–0.83)	53 (50–57)
2005: I	46	8	0.96 (0.93–0.98)	59 (55–63)
2005: V	14	7	0.81 (0.59–0.94)	58 (53–63)
2005: X	37	7	0.95 (0.92–0.97)	55 (49–60)
2006: I	92	7	0.79 (0.68–0.87)	56 (54–59)
2006: V	20	8	0.87 (0.75–0.95)	53 (49–57)

n_a – no of abstracts; n_r – no of reviewers

Sample size

We calculated the sample size where the level of agreement between the reviewers was measured using the ICC (Walter et al. 1998). Assuming an increase of 0.1 in the ICC for all combinations of the hypothesized value of ICC, with an alpha level of 0.05 and a beta level of 0.2, a total of 99 abstracts would be needed to have a minimum of 80% power to calculate ICCs among 13 different reviewers.

Results

Overall agreement in the review process

Of the 426 abstracts 7 (1.6%) had missing data, leaving 419 that were suitable for analysis. Table 1 shows the number of abstracts, number of reviewers, ICC, mean ISK abstract score and their relative 95% CIs for each NOV scientific meeting. Across 13 review periods, inter-reviewer agreement to rate abstracts was substantial 0.68 (95% CI: 0.47–0.83) to almost perfect 0.95 (95% CI: 0.92–0.97) and did not change appreciably over the time period. The overall mean ISK abstract score did not change

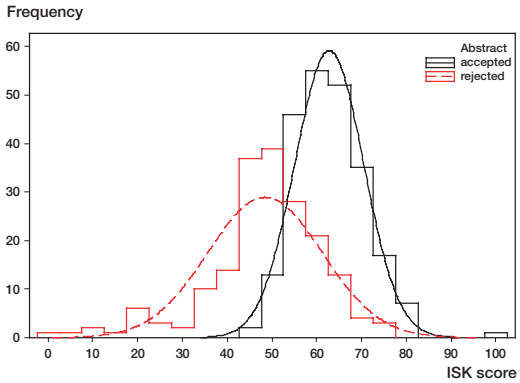

Figure 1. Histogram of ISK abstract score for accepted and rejected abstracts.

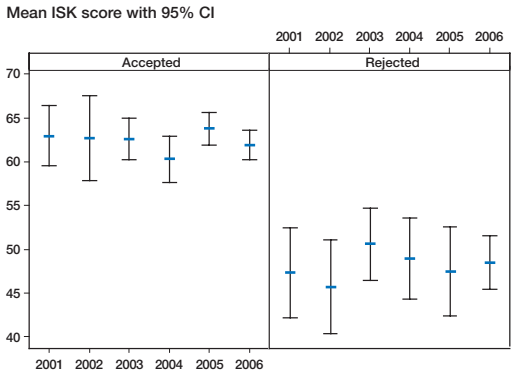

Figure 2. Mean ISK abstract scores with 95% CI, by abstract acceptance for NOV scientific meetings.

Table 2. Frequency and mean ISK abstract score for accepted (n = 229) and rejected (n = 190) abstracts

	Accepted			Rejected		
			ISK score			ISK score
Year	n	%	mean (95%CI)	n	%	mean (95%CI)
2001	37	51	63 (60–66)	36	49	47 (42–53)
2002	18	75	63 (58–66)	6	25	46 (40–51)
2003	45	61	63 (60–65)	29	39	51 (47–55)
2004	14	36	60 (58–63)	25	64	49 (44–54)
2005	56	58	64 (62–66)	41	42	48 (42–53)
2006	59	53	62 (60–64)	53	47	49 (46–52)

significantly over time between scientific meetings (p = 0.3). This was also true after adjusting for multiple comparisons using Bonferroni test with the smallest p-value of 1.0. The number of abstracts for a meeting, representing the workload of the reviewers, did not influence the level of agreement between them (ICC) (p = 0.7). Thus, reviewers with greater numbers of abstracts to review did not have scores that differed significantly from those with few abstracts to review.

Acceptance rates

Of the 419 abstracts, 229 (55%) were accepted. The mean ISK score was significantly greater in accepted abstracts than in rejected abstracts (p < 0.001). The distribution of ISK abstract score for accepted and rejected abstracts is presented in Figure 1. There were some overlaps of ISK abstract score between accepted and rejected abstracts. The acceptance rate did change between 2001 and 2006 (p = 0.039) and a smaller proportion of abstracts were accepted after 2004 (Table 2). The mean ISK abstract score for accepted abstracts ranged from 60 (95% CI: 58–63) to 64 (95% CI: 62–66). The mean ISK abstract score for rejected abstracts varied from 46 (95% CI: 40–51) to 51 (95% CI: 47–55). There was no significant change in mean ISK abstract scores for accepted or rejected abstracts over time, with p values of p = 0.7 and 0.9, respectively (Table 2; Figure 2).

Discussion

We report the following: (1) reviewers using the ISK abstract score scheme had substantial to almost perfect agreement; (2) the quality of abstracts was unchanged from 2001 to 2006; and (3) having a larger number of abstracts to score did not influence agreement among reviewers.

Strengths and limitations

The methodology of our study is strengthened by the following: (1) the sample size of the abstracts was large and the study was sufficiently powered to answer the research questions; (2) attempts were made to blind the reviewers as to the names of the authors and the names of the institutions; and (3) the group of reviewers changed in a smooth fashion with a change of no more than 2 per year over the 6 years. However, our study had certain limitations. 1) Scoring schemes have limitations, especially if thresholds are used. Thresholds are

arbitrary and have resulted in methodological flaws when used in meta-analyses (Juni et al. 1999). The scheme scores methodological quality as well as newsworthiness. If an abstract is scored poorly on methodological soundness and is scored as being good regarding newsworthiness, it still gets a relatively high ranking. In contrast, an abstract reporting a trial with a rigorous methodological construct may score poorly if the topic is not in vogue. 2) The Annual Dutch Orthopedic Association's meetings, held in January, are dedicated to a specific topic. The spring and fall meetings were not designated a specific topic. Abstracts with higher scores were sometimes rejected if they were not within the scope of the meeting dealing with a specific topic. Final acceptance for a meeting was decided on through consensus by the scientific committee and, at times, may have been subjective. This may have influenced our results. 3) Our study was of one National Orthopedic Association. The results cannot be generalized for other countries, associations, or (sub-) specialties. 4) We concentrated on the meetings with subsequent publication of abstracts on the website of Acta Orthopaedica. In 2004, the abstracts of only one meeting were published online. The abstracts from the 2006 meeting are not yet available online. The scientific meetings were usually held 3 times a year. In 2001, there was no specific-topic meeting. Unfortunately, the NOV abstract scoring database did not have information on three meetings: January 2002, and January and May 2004. For unknown reasons, abstracts of these meetings were not published or scoring data were not saved.

Previous literature

A recent report describing a 9-item scale had a moderate agreement for peer review of abstracts for the Canadian Association of Emergency Physicians (Rowe et al. 2006). Rowe and co-workers wanted to find criteria associated with poor agreement among the reviewers. The authors found greater agreement in the more specific and objective criteria. The ISK system presented in our study weighs objective criteria accordingly; this may have contributed to the excellent agreement among our reviewers. Since our study reports excellent agreement, identification of factors associated with low agreement became less relevant.

We are aware of one report in the orthopedic literature describing poor agreement (ICC 0.23–0.27) among reviewers scoring abstracts with a simple "accept–unsure–reject" system (Bhandari et al. 2004). 8 reviewers scored 440 abstracts in 2001 and 9 reviewers scored 438 abstracts in 2002. This large number of abstracts to be scored in a short period of time may have led to a less vigorous review process. Our study reports 419 abstracts scored by a total of 13 reviewers in a 6-year period; this workload was less intense. Interestingly, we found no influence of a larger number of abstracts to score on the reviewer agreement. In our report, the number of abstracts per reviewer did not exceed 92. We do not know whether the amount of agreement would become reduced with larger numbers of abstracts, as described previously in a report covering more than 400 abstracts per meeting (Bhandari et al. 2004). Furthermore, we do not know the exact time it takes to score an abstract with the ISK system. This will be the subject of future research. In meetings where nearly all abstracts are to be accepted for presentation, either in oral or poster form, then the "accept–unsure–reject" system is sufficient to identify the few really poor abstracts that need to be rejected. In cases where a more rigorous selection is required, i.e. where the number of submissions exceeds the number of presentation slots, then an objective and reproducible scoring system is required to ensure just selection, i.e. giving each submission an equal chance to be selected without bias.

Equally important is the gray literature generated by the accepted abstracts that are, for example, published online at the website of a scientific journal. The gray literature comprises scientific and technical reports, patent documents, conference papers, internal reports, government documents, newsletters, fact sheets and theses, which are not readily available through commercial or library channels. In contrast, it does not include normal scientific journals, books, or popular publications that are available through traditional commercial publication channels. Gray literature is often the only published manuscript from a trial (Sprague et al. 2003, Rowe et al. 2006). Thus, the quality of the abstract must be high to facilitate selection—for example, for future meta-analyses (McAuley et al.

2000, Martin et al. 2005). Cochrane reviews often include proceedings of meetings in their search strategy (Poolman et al. 2005) whereas many journals do not accept them as references. Our scoring scheme helps in maintaining a high standard of reporting.

We chose not to include the data of 2000 in our analysis since this was the first time the scoring system was used. In 2000, fewer abstracts were submitted, making it easier to select from them and fill the program. Furthermore, in that year the policy of the committee was to fill the program. In the following years (2001 and onwards), the policy of the committee changed, to filling the program with good quality presentations only. Thus, more weight was assigned to the score as a quality measure and it was accepted that the program might not be fully filled. This resulted in a rather constant average score for accepted abstracts over the years. Our theory that educational efforts on the practice of evidence-based orthopedics over the past years improved the quality of reporting in abstracts could not be confirmed.

Blinding of reviewers was attempted, but in our small orthopedic community the origin of the coded abstract is not difficult to guess. It is therefore questionable whether reviewers can be properly blinded as to author and institution. Furthermore, blinding in the peer review process has come under scrutiny after two randomized controlled trials gave similar results in the peer review process with or without blinding for the identity of authors (Justice et al. 1998, Smith, Jr. et al. 2002).

Relevance of our findings

Our study has shown that the ISK abstract scoring system gave excellent inter-reviewer agreement over a 6-year period. This is the first system reported to show a constantly good agreement. The reviewers did not receive training in the use of the scoring system. This further strengthens the rationale for implementation of the ISK abstract rating system since it is easy to use, reliable, and easy to learn.

Conclusion

The ISK abstract rating system has shown excellent interobserver agreement. Organizers of other scientific orthopedics meetings should consider adopting the ISK rating system for further evaluation in local or international settings.

Contributions of authors

RWP: initiated and designed the study, co-conducted the analyses, and drafted and revised the manuscript. LCMK and MCdWM: maintained the database, reviewed and edited the manuscript. LB: introduced the ISK score in the review process of abstracts for meetings of the Dutch Orthopaedic Association, assisted in the conception and design of the study and analysis of the data, and reviewed and revised the manuscript. FF: calculated the power analysis, designed and conducted the statistical analyses, and reviewed and edited the manuscript. MB: was co-initiator, designed the study, and revised the manuscript. All authors agreed on the final version of the manuscript.

Dr. Bhandari was supported in part by a Canada Research Chair from the Canadian Institutes of Health Research. Dr. Poolman was supported in part by a Stichting Wetenschappelijk Onderzoek Orthopaedische Chirurgie Fellowship, and by Biomet (the Netherlands), Anna Fonds, Zimmer (the Netherlands), MSD (the Netherlands), Stryker (the Netherlands) and by a De Nederlandse Vereniging voor Orthopedische Traumatologie Fellowship.

In 2006 the members of the Dutch Orthopaedic Association Scientific Committee were: J.W.K. Louwerens, N.J.A. Tulp, W.J.A. Dhert, B.J. van Royen, J.J.A.M. van Raaij, L.C.M. Keijser, M.C. de Waal Malefijt, J.B.A. van Mourik and B.W. Schreurs.

Bhandari M, Templeman D, Tornetta P, III. Interrater reliability in grading abstracts for the orthopaedic trauma association. Clin Orthop 2004; (423): 217-21.

Juni P, Witschi A, Bloch R, Egger M. The hazards of scoring the quality of clinical trials for meta-analysis. JAMA 1999; 282: 1054-60.

Justice A C, Cho M K, Winker M A, Berlin J A, Rennie D. Does masking author identity improve peer review quality? A randomized controlled trial. PEER Investigators. JAMA 1998; 280: 240-2.

Landis J R, Koch G G. The measurement of observer agreement for categorical data. Biometrics 1977; 33: 159-74.

Martin J L, Perez V, Sacristan M, Alvarez E. Is grey literature essential for a better control of publication bias in psychiatry? An example from three meta-analyses of schizophrenia. Eur Psychiatry 2005; 20: 550-3.

McAuley L, Pham B, Tugwell P, Moher D. Does the inclusion of grey literature influence estimates of intervention effectiveness reported in meta-analyses? Lancet 2000; 356: 1228-31.

Poolman R W, Goslings J, Lee J, Statius M M, Steller E, Struijs P. Conservative treatment for closed fifth (small finger) metacarpal neck fractures. Cochrane Database Syst Rev 2005;CD003210.

Rowe B H, Strome T L, Spooner C, Blitz S, Grafstein E, Worster A. Reviewer agreement trends from four years of electronic submissions of conference abstracts. BMC Med Res Methodol 2006; 6: 14.

Smith J, Jr., Nixon R, Bueschen A J, Venable D D, Henry H H. Impact of blinded versus unblinded abstract review on scientific program content. J Urol 2002; 168: 2123-5.

Sprague S, Bhandari M, Devereaux P J, Swiontkowski M F, Tornetta P, III, Cook D J, Dirschl D, Schemitsch E H, Guyatt G H. Barriers to full-text publication following presentation of abstracts at annual orthopaedic meetings. J Bone Joint Surg (Am) 2003; 85: 158-63.

van der Steen L P, Hage J J, Kon M, Mazzola R. Reliability of a structured method of selecting abstracts for a plastic surgical scientific meeting. Plast Reconstr Surg 2003; 111: 2215-22.

Walter S D, Eliasziw M, Donner A. Sample size and optimal designs for reliability studies. Stat Med 1998; 17: 101-10.

Appendix. ISK abstract scoring scheme

Scoring for abstracts of clinical studies (maximum 100 points)	
Baseline score	+50
1. Problem description	
clear	+5
vague	0
non-existent	–5
2. Design	
prospective	+5
retrospective	0
questionnaire only	–5
not specified	–10
3. Control group	
yes	0
multiple tests, one group	+5
matched	+5
randomized	+10
matched and randomized	+15
no control	–5
4. Material	
unique	+5
adequate (size, length of follow-up)	0
insufficient (or not described)	–5
5. Methods	
objective and valid	+10
described	0
not described	–5
6. Results	
unique	+10
new and important	+5
existing knowledge	0
not important or incoherent	–5
not presented	–10
7. Conclusions	
valid	0
not supported by results	–5
non-existent	–10

Scoring for abstracts of experimental studies [a] (maximum 100 points)	
Baseline score	+50
1. Problem description	
important problem	+5
clear	0
vague	–5
2. Animal experiments only:	
Design	
Control group	
yes	0
multiple, tests, same group	+5
matched	+5
randomized	+10
matched and randomized	+15
no control	–10
Methods	
objective and valid	+10
described	0
not described	–10
3. Anatomy and biomechanics only:	
Material (samples, specimens)	
unique	+10
adequate	0
insufficient	–10
Methods	
objective and valid	+15
well described (accuracy and technique	+5
described	0
not described	–10
4. Results	
unique	+10
new and important	+5
existing knowledge	0
not important	–5
incoherent with methods or wrong	–10
not presented	–15
5. Conclusions	
valid	+5
not supported by results	–5
non described	–10

[a] Laboratory experiments, anatomy, biomechanics, animal studies