Inferior reliability of VAS scoring compared with International Society of the Knee reporting system for abstract assessment

Ole Rahbek[1], Steen L. Jensen[1], Martin Lind[1], Jeannette Ø. Penny[1], Thomas Kallemose[2], Thomas Jakobsen[1] & Anders Troelsen[1]

ORIGINAL ARTICLE

1) Scientific Committee of The Danish Orthopaedic Society
2) Department of Orthopaedic Surgery, Hvidovre Hospital, Denmark

Dan Med J 2017;64(4):A5346

ABSTRACT

INTRODUCTION: Knowledge of how abstracts may be selected for medical conferences in an efficient and reliable manner is sparse. To improve abstract selection, the Danish Orthopaedic Society implemented the International Society of the Knee (ISK) quality-of-reporting system and visual analogue scale (VAS) scoring for abstract evaluation at its 2014 Annual Congress. We sought to find out if a simple VAS score was more reliable than a multiple-question system for assessment of over-all abstract quality.

METHODS: A total of 214 abstracts were submitted for review. All abstracts were reviewed by 3 reviewers using a VAS score and the ISK score. Of the 214, 71 abstracts were reviewed again 6 months later to estimate intra-rater agreement.

RESULTS: The VAS and the ISK score were poorly correlated (r = 0.64), and the ISK score demonstrated a better intra- and interrater agreement ($p < 0.001$). The VAS scores of all abstracts were more widely distributed than the ISK scores, which clustered around values in the 50-70 range. Chronbach's alpha for the ISK score was 0.66 (95% confidence interval: 0.62-0.68).

CONCLUSIONS: The VAS score has a poorer intra- and inter-rater agreement than the ISK score, and the two scores do not correlate well. VAS scores were more widely distributed, which is beneficial when selecting a scientific programme, but the score is unreliable. We continue to use the ISK score, although its reliability may still be improved.

FUNDING: none.

TRIAL REGISTRATION: not relevant.

The abstract reviewing process greatly impacts the knowledge presented to us in journals and at scientific meetings. Bias in the dissemination of scientific results is a well-established phenomenon [1]. The literature is likely to have an over-representation of studies with positive outcomes [2] and clinical malpractice may be the ultimate consequence of a biased reviewing process, especially if evidence is not graded adequately by clinicians [3]. Despite these facts, the literature on the review process and on selection of abstracts before scientific meetings and congresses is sparse. The explanation may be that as medical experts we are quite confident in our ability to spot high scientific quality and therefore see no need to scrutinise the review process. However, subjective scores have the disadvantage that they do not require the reviewer to perform a standardised and systematic review compared with more objective scores. Furthermore, subjective and opinion-based criteria are more prone to having a low inter-rater agreement [4, 5]. Most objective scores require the reviewer to systematically grade the quality of the Introduction, Design, Material, Method, Results and Conclusion sections according to given criteria. However, this is more time consuming. Given the fact that reviewing of abstracts is often done by volunteers in their spare time, it is of interest to reduce reviewer's workload and make the process as efficient as possible. On this backdrop, simple subjective scores are more appealing.

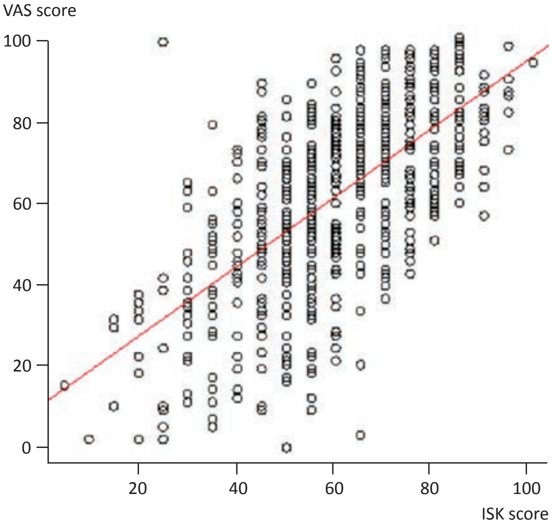
Important requirements to a score system include a high inter- and intra-rater reliability, an ability to differentiate between good and poor abstracts and the absence of any ceiling or floor effect. However, to our knowledge, the intra-rater reliability of abstract grading has never been reported, and it may be difficult to determine the ability of grading systems to select true high-quality abstracts from true poor-quality abstracts.

In 2014, the Scientific Committee of the The Danish Orthopaedic Society (DOS) decided to introduce a new grading system for abstracts submitted to the yearly congress of the Society. This provided the Scientific Committee with a unique opportunity to compare a subjective visual analogue scale (VAS) score with a more objective score. Several abstract grading systems have previously been validated in terms of inter-rater reliability. The inter-rater agreement of the International Society of the Knee (ISK) quality-of-reporting system was tested by the Dutch Orthopaedic Association and was found to be excellent [6].

To our knowledge, it has never been proven that an objective score is superior to a subjective score and the intra-rater reliability of scores has not previously been tested. The aim of the present study was to compare a subjective VAS score with the multiple-question ISK

FIGURE 1

Scatterplot of visual analogue scale (VAS) score plotted against International Society of the Knee (ISK) score (n = 625), correlation 0.64.



score in terms of inter- and intra-rater reliability, distribution of scores and floor & ceiling effects. Furthermore, the correlation of the scores was determined. Our hypothesis was that the VAS score could replace the ISK abstract score system for scientific abstract evaluation.

METHODS

In 2014, 214 abstracts were submitted for review at the Annual Congress of DOS. All abstracts were reviewed by three reviewers using both a VAS score and the ISK score. The ISK score was used to select the best abstracts for presentation at the Congress, and the VAS score was included as part of the design of the present study.

To avoid bias in the reviewing process, it was ensured that abstracts were not reviewed by reviewers from the same institutions as the abstract authors. The reviews were performed by reviewers who were blinded to the authors' names and institutions. The reviewers were asked to mark their overall impression of the abstract on a VAS line when the review of an abstract was initiated. This score was automatically converted into a number that remained unknown to the reviewer. The numbers ranged from 0 to 100. VAS scoring was a prerequisite before proceeding to the ISK score. When the ISK score was initiated, the reviewer had to decide if the study type was clinical or experimental. The category of experimental studies includes laboratory, anatomical, biomechanical and animal studies. Depending of the category, one of two abstract multiple choice scoring schemes (clinical or experimental) appeared.

Regardless of the category chosen (clinical or experimental), the baseline abstract score is 50 points, and points are added or subtracted from the baseline score depending on the choices made by the reviewer in each item. The maximum score is 100 and the minimum score is 0. The reviewers were blinded to the score changes and were not given the value of the final score to avoid the reviewers being biased by the score. Items concerning more objective methodological criteria are given more weight in the total score than more subjective items such as the significance of the results.

Members of the Scientific Committee scored a total of 71 abstracts as part of the abstract handling. These abstracts were reviewed again six months later to estimate the intra-rater agreement.

Statistics

Data are presented as scatterplots or histograms with quartiles. Chronbach's alpha was calculated with a bootstrap 95% confidence interval (CI).

Intra-rater agreement was assessed according to Bland-Altman [7] and plotted as the average versus the difference. Limits of agreement (LOA) were calculated as the interval within which 95% of the differences were expected to fall.

Inter-rater agreement was evaluated by calculating the largest difference in the score among all reviewers of an abstract, for each abstract. The distributions of these maximal differences for VAS and ISK were compared with the Wilcoxon rank sum test. p-values below 0.05 were considered significant. A mixed effect model for VAS, VAS clinical, VAS experimental, ISK, ISK clinical and ISK experimental with abstract as fixed effect and a random effect for reviewers was used to estimate the amount of variance from the reviewers. The inter-rater variance was calculated as the amount of the total variance (reviewer variance and the error term variance) that comes from reviewer variance, and intra-rater variance as the amount that comes from error term variance.

All analyses were done using R 3.0.2 (R Foundation for Statistical Computing, Vienna, Austria).

Trial registration: not relevant.

RESULTS

A total of 67 reviewers were assigned abstracts. Six reviewers reviewed no abstracts before deadline, leaving 65 abstracts without a score. Another six abstracts were not scored because the reviewers considered that they had conflicts of interest. The 71 non-evaluated abstracts at deadline were distributed between and scored by the five members of the Scientific Committee. Thus, a total of 642 reviews were performed. Two reviewers (17 reviews) were excluded from data analysis as they scored 0 on the VAS score in all their reviews, leaving 625 reviews for analysis.

Agreement of scores

There was a poor correlation (r = 0.64) between the VAS and the ISK score (**Figure 1**). In particular, ISK scores ranging from 40 to 60 showed a wide range in the corresponding VAS score. When studying the differences between the average ISK score and VAS score for each abstract, only 60% of differences were within ± 10 points. Chronbach's alpha for the ISK score was 0.66 (95% CI: 0.62-0.68).

Distribution

The distribution of average scores based on three reviewers was not identical between the VAS and the ISK score (**Figure 2**). When using the VAS score, abstracts were more frequently given scores below 40. The ISK abstract scores had a more pronounced trend to cluster around values between 50 and 70. No ceiling or floor effects were observed for any of the scores.

Inter-rater agreement

The distributions of the maximal disagreement for each abstract scored with either VAS or ISK showed a better agreement when scoring with the ISK (p < 0.001) (**Figure 3**). A maximal disagreement exceeding 25 points was found for 64% of the abstracts which were scored with the VAS score. The use of the ISK reduced this to 26%. Disagreements exceeding 40 points was found in 32% of abstracts evaluations using VAS compared with 5% using the ISK. The inter-rater variances for both scores and for clinical and experimental abstracts, respectively, are presented in **Table 1**.

Intra-rater agreement

The mean difference was quite similar for VAS and ISK scores with 1.3 and 1.0, respectively. However, the limits of agreement were larger for the VAS score (–33-35) than for the ISK score (–5-27). The intra-rater variances for both scores and for clinical and experimental abstracts, respectively, are presented in Table 1.

DISCUSSION

To our knowledge, this is the first study to report the intra-rater agreement of abstract reviewing and to compare a subjective VAS to a more objective grading system.

Due to the variation in the likeliness of reviewers to accept abstracts [4], multiple reviewers were used. Appleton [8] found that three reviewers are almost as efficient in selecting the best abstracts as a panel of six or more. Therefore, three reviewers per abstract were used in the reviewing process of the present study.

Our data showed a considerably higher inter-rater disagreement when the VAS-based score was used compared with the ISK score. The VAS score disqualified itself with 64% of reviews having a maximal disagreement exceeding 25 points and 32% exceeding 40 points. The ISK system performed better. This may be explained by the fact that the reviewer is guided through a more systematic review and that objective items are included. However, considerable disagreement was found with the ISK scoring system as well. One explanation may be that not all items in the scoring system are fully objective. For example, reviewers were asked to rate "results" in one of the following categories: unique, new and important, existing knowledge, not important or not presented. The distinction between "unique" and "new and important" is somewhat subjective and may not be judged equally by the reviewers. In addition, it has been demonstrated that even highly esteemed surgeons with an interest in research can disagree on items that we consider to be very objective and easy to understand.

FIGURE 2

Histograms showing the distribution of abstract scores based on the average of three reviewers for visual analogue scale (VAS) (**A**) and International Society of the Knee (ISK) quality-of-reporting system (**B**), respectively (n = 214). The VAS score gives a wider distribution, and the ISK quality-of-reporting system score is distributed around values of 50-70.

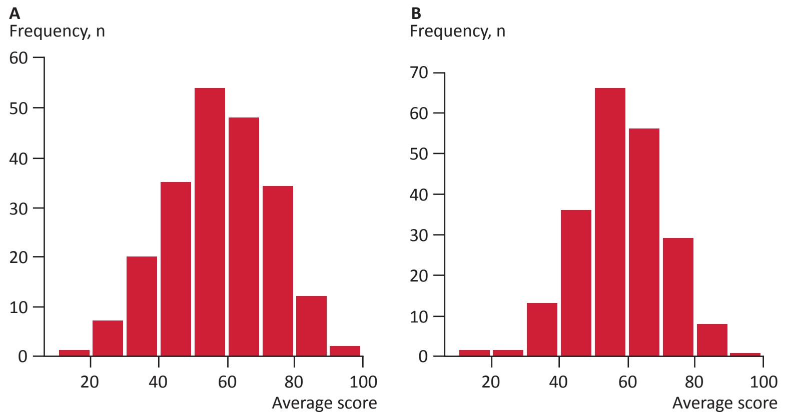


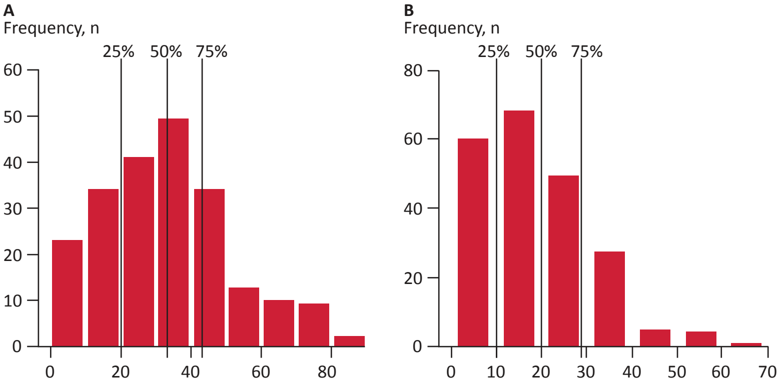
TABLE 1

The inter-rater and intra-rater variance. Data for both visual analogue scale (VAS) and International Society of the Knee (ISK) scores of all abstracts are presented as well as for the subgroups clinical and experimental abstracts. The values are mean (± standard deviation).

Score	Inter-rater	Intra-rater
VAS		
All abstracts	96 (± 10)	276 (± 17)
Clinical abstracts	118 (± 11)	270 (± 16)
Experimental abstracts	18 (± 4)	203 (± 14)
ISK		
All abstracts	35 (± 6)	126 (± 11)
Clinical abstracts	34 (± 6)	120 (± 11)
Experimental abstracts	57 (± 8)	158 (± 13)

FIGURE 3

Histograms showing the distribution of the maximal differences among the three reviewers in the scoring of each abstract (n = 214) for visual analogue scale scores (**A**) and International Society of the Knee quality-of-reporting system scores (**B**), respectively.



Bhandari et al tested the reliability of surgeons from academic centres in classifying studies according to study type and level of evidence (I-V). Absolute agreement ranged from 67% to 82%. Interestingly, they found that there was perfect agreement among reviewers trained in epidemiology, which may indicate that training can improve reliability in the assessment of what we call objective parameters [9].

We found that the intra-rater agreement of the ISK was superior to that of the VAS score. The mean difference for the two systems was similar, but the LOA were larger for the VAS score. Nonetheless, the LOA interval for the ISK score was disturbingly high. The intra-observer reliability was tested by the members of the Scientific Committee. All members hold scientific degrees and have extensive research experience. Even so, reliability was not impressive, which may underline the need for an improvement of the score.

In selecting a scientific programme, it may be preferable that abstracts are scored using the whole scale from 0 to 100. This makes it easier to select abstracts for rejection or to be nominated for awards. Having all abstracts scored in the 50-70 range is not helpful at all. In that respect, we found that the VAS was superior to the ISK score. Average VAS scores had a wider distribution and scores below 40 points were given more frequently. However, the VAS score was more imprecise and therefore less valid. We found a poor correlation between the VAS and the ISK score.

The scientific programme would have been selected differently using the VAS score because of the poor correlation and agreement between the VAS and the ISK score. The present study has demonstrated that the ISK score is more precise in selecting the programme, and the ISK seems to be superior for scientific abstract evaluation. However, one question remains unanswered: Is ISK actually selecting the best abstracts?

This may not be a straight forward question to answer. Before accuracy can be evaluated, we need to know the true quality of the abstract. A measure could be later publication in peer reviewed high-profile journals. Jackson et al [10] showed that abstracts selected for podium presentation had a significantly higher publication rate than those not accepted for presentation (53% versus 38%). This indicates that the abstract scoring and reviewing is predictive for later publication. Nevertheless, only 38-70% [10-13] of abstracts from orthopaedic meetings get published, and other factors than the quality of the scientific work may influence the odds of publication. Sprague et al [14] showed that among investigators with unpublished abstracts six years after the meeting, nearly 50% of cases had not been published due to lack of sufficient time for scientific work. Only 16% stated rejection from a scientific journal as a reason for not publishing. Thus, subsequent publication rates may not be a good indicator for abstract quality as the reasons for not publishing are more complex. In addition, publishing in orthopaedic journals may be biased [2, 15]. Abstracts presenting significant, positive results are more likely to be published regardless of sample size, study design and sponsorship. We are, therefore, left without unbiased measures of quality.

CONCLUSIONS

The VAS score has a poorer intra- and inter-rater agreement than the ISK score, and the two scores do not correlate well. The ISK score cannot be replaced by a simple VAS score in the selection scientific abstracts. Further studies should focus on improvement of the ISK score to improve its reliability. Based on the present study DOS will continue to use the ISK score for abstract evaluation, but modifications of the score may be needed.

CORRESPONDENCE: *Ole Rahbek.* E-mail: ole0rahbek@gmail.com

ACCEPTED: 19 January 2017

CONFLICTS OF INTEREST: Disclosure forms provided by the authors are available with the full text of this article at www.danmedj.dk

LITERATURE

1. Song F, Parekh S, Hooper L et al. Dissemination and publication of research findings: an updated review of related biases. Health Technol Assess 2010;14:1-193.
2. Harris IA, Mourad M, Kadir A et al. Publication bias in abstracts presented to the annual meeting of the American Academy of Orthopaedic Surgeons. J Orthop Surg (Hong Kong) 2007;15:62-6.
3. Guyatt GH, Oxman AD, Vist GE et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008; 336:924-6.
4. Bhandari M, Templeman D, Tornetta P 3rd. Interrater reliability in grading abstracts for the orthopaedic trauma association. Clin Orthop Relat Res 2004;(423):217-21.
5. Rowe BH, Strome TL, Spooner C et al. Reviewer agreement trends from four years of electronic submissions of conference abstract. BMC Med Res Methodol 2006;6:14.
6. Poolman RW, Keijser LC, de Waal Malefijt MC et al. Reviewer agreement in scoring 419 abstracts for scientific orthopedics meetings. Acta Orthop 2007;78:278-84.

7. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;1:307-10.
8. Appleton DR, Kerr DN. Choosing the programme for an international congress. Br Med J 1978;1:421-3.
9. Bhandari M, Swiontkowski MF, Einhorn TA et al. Interobserver agreement in the application of levels of evidence to scientific papers in the American volume of the Journal of Bone and Joint Surgery. J Bone Joint Surg Am 2004;86-A:1717-20.
10. Jackson KR, Daluiski A, Kay RM. Publication of abstracts submitted to the annual meeting of the Pediatric Orthopaedic Society of North America. J Pediatr Orthop 2000;20:2-6.
11. Daluiski A, Kuhns CA, Jackson KR et al. Publication rate of abstracts presented at the annual meeting of the Orthopaedic Research Society. J Orthop Res 1998;16:645-9.
12. Nguyen V, Tornetta P 3rd, Bkaric M. Publication rates for the scientific sessions of the OTA. Orthopaedic Trauma Association. J Orthop Trauma 1998;12:457-9,discussion 456.
13. Hamlet WP, Fletcher A, Meals RA. Publication patterns of papers presented at the Annual Meeting of The American Academy of Orthopaedic Surgeons. J Bone Joint Surg Am 1997;79:1138-43.
14. Sprague S, Bhandari M, Devereaux PJ et al. Barriers to full-text publication following presentation of abstracts at annual orthopaedic meetings. J Bone Joint Surg Am 2003;85-A:158-63.
15. Harris IA, Mourad MS, Kadir A et al. Publication bias in papers presented to the Australian Orthopaedic Association Annual Scientific Meeting. ANZ J Surg 2006;76:427-31.