Analysis of the reliability of the evaluation and scoring process for abstract submissions for the “Journées Francophones de la Kinésithérapie 2023”

Jean-Philippe Deneuville , PT, PhD[1], **Matthieu Gallou-Guyot , PT, PhD**[2,3] **and Matthieu Guémann , PT, PhD**[4,5]
[1]PRISMATICS Lab (Predictive Research in Spine/Neuromodulation Management and Thoracic Innovation/Cardiac Surgery), Poitiers University Hospital, Poitiers, France, [2]Ochanomizu University, Department Center for Interdisciplinary AI and Data Science, Tokyo, Japon, [3]Laboratoire HAVAE, Université de Limoges, Limoges, France, [4]École Universitaire de Kinésithérapie du centre val de Loire -EUK-CVL, Université d’Orléans, France, [5]Structure Fédérative de Recherche SAPRéM, Université d’Orléans, France

received : 28 August 2023
accepted: 13 December 2023

ISSN: 2823-989X
DOI: 10.52057/erj.v4i1.46

ABSTRACT

Background: The French Society of Physiotherapy (SFP) organises biennial conferences known as the "Journées Francophones de Kinésithérapie" (JFK) since 2007. Abstracts are submitted and are evaluated for acceptance by two independent reviewers from the SFP using a predefined rating checklist. However, the reliability of this process has never been evaluated. **Objective**: This study aims to assess the inter-rater reliability and internal consistency of the JFK submission rating process conducted by the scientific committee. **Methods**: Blind reviewers evaluated each submission in pairs using a standardized 47-item rating checklist, categorised into five domains including background, method, results and relevance to physiotherapy. Reliability was assessed using an Intraclass Correlation Coefficient (ICC) and Cohen’s Kappa. Agreement was assessed using the Standard Error of Measurement (SEM), Coefficient of Variation, Bland-Altman analysis, and percentage of agreement. Internal consistency was assessed using Cronbach’s Alpha. **Results**: 36 reviewers assessed a total of 217 abstracts. The reliability, measured by ICC, was poor (0.39 [CI95% = 0.30; 0.49]) as was the agreement; SEM = 3.08 and Coefficient of Variation = 23.1%. All individual checklist items had a Cohen’s kappa coeff below 0.6. All but one domain had a Cronbach Alpha above 0.7, indicating good consistency. However, five domains had a Cronbach Alpha above 0.9, suggesting redundancy. **Conclusion**: The JFK 2023 submission rating process displayed poor reliability. These findings can guide improvements in creating the JFK 2025 checklist. This study may help future scientific committees to enhance their evaluation process.

KEYWORDS: Agreement, Clinimetry, Congress, Internal consistency, Reliability, SFP

Introduction

The “Journées Francophones de Kinésithérapie” (JFK) has been conducted biannually since 2007. This conference is organised by the French Society of Physiotherapy (SFP) and gathers nearly 2,000 professionals, researchers, educators, and students from the field of physiotherapy. Each edition has a Scientific Committee (SC), comprised of approximatively twenty professionals. The SC is responsible for the (i) development and (ii) organisation of the SFP’s scientific program, and (iii) the evaluation of submitted abstracts. This evaluation process employs a specific rating checklist provided to the reviewers for a standardised assessment.

From a general perspective, the evaluation process for submissions holds a crucial position within the academic framework for disseminating research findings [1]. As indicated by several reviews, only 50% of the abstracts submitted will progress to full article publication in the years following the conference [2, 3]. Hence, conference abstracts frequently serve as the sole means of sharing research data for a substantial portion

Corresponding author:
Jean-Philippe Deneuville, PRISMATICS Lab (Predictive Research in Spine/Neuromodulation Management and Thoracic Innovation/Cardiac Surgery), Poitiers University Hospital, Poitiers, France. e-mail: deneuvillejp@msksolution.fr

of research outcomes. They also represent the sole opportunity for some researchers to disseminate their findings, obtain feedback, and progress their career consequently [4]. Therefore, each conference's SC serves as a gatekeeper for a significant portion of the available research evidence through the process of abstract submission assessment. On the other hand, this assessment should strive to be as equitable as possible for researchers and professionals submitting their work. The reliability of this process is typically low, with considerable variability from one reviewer to another [5, 6, 7].

In the specific context of the JFK 2023, the SC revised and used the evaluation checklist from the previous edition (**Table 1**). Each abstract was assessed independently by two reviewers to derive a final score, which was the average of the two reviewers' scores. The pilots of the SC (JPD and MGG) used this score to make decisions regarding acceptation. Abstracts were either rejected or accepted for inclusion as a poster presentation or an oral communication. Each abstract accepted for oral presentation was subsequently published in the journal "Kinésithérapie, la Revue".

In a continuous effort to enhance the rating process, we implemented a quality approach within the SC of the JFK. This approach includes a reliability test of the rating process between two reviewers. In the present study, we assessed the reliability and agreement of the rating process for the submissions to the 2023 JFK conference.

Method

Design

In this reliability study, we assessed the inter-rater reliability and agreement of the rating process for abstract submission to the 2023 JFK conference. The present report was written according to the Guidelines for Reporting Reliability and Agreement Studies (GRRAS) [8].

Sampling methods

We included all consecutive abstracts submitted within the JFK submission process which were rated by two reviewers. We excluded abstracts which were only rated once or not at all.

The reviewers were either members of the JFK's SC, members of the SFP's "Collège des Sciences du Vivant (CSV)," or reviewers suggested by one of the SFP's partner associations.

The SC was assembled by the SFP prior to the conference. Initially, the board of the SFP elected two pilots for the future SC of the conference. These pilots then disseminated a call for entries. Candidates submitted their *Curriculum Vitae* and were subsequently included or excluded by the pilots based on their scientific expertise.

The CSV, a group within the SFP, is composed of researchers and academics with at least a Masters degree. The SC's pilots solicited this group for assistance in evaluating abstract submissions.

The 12 partner associations, working with the SFP to organise the JFK, could supply additional reviewers with academic expertise to support the SC in overseeing the evaluation process. During the submission process, authors had the option to indicate their membership in one of the specific partner associations. In such cases, the abstract was assessed by one reviewer put forth by the partner association and one reviewer from the SC.

As the present study is conducted retrospectively to the JFK submission process, we did not perform any sample size calculation for either the number of abstracts nor the number of reviewers.

The rating checklist and selection process

The checklist evaluated 47 items, categorised into five main domains (**Table 1**). Three of these domains applied to all research designs, while the remaining two were further subdivided into three subdomains, tailored for systematic reviews, interventional studies, and studies in the social and human sciences. Each item was scored as "yes", "no" or "not applicable".

The decision to accept or reject the abstract was based on an average **relative score**, derived from an **absolute score**, calculated as following:

- The denominator of the absolute score corresponded to the number of total items minus the number of items judged not applicable (example: 47 items in total - 23 not applicable, denominator of the relative score = 24).
- The numerator of the absolute score corresponded to the number of positive items (example: 16 "yes" items).
- The absolute score corresponded to the ratio between the numerator and the denominator (here: 16/24).
- To determine the **relative score**, we converted the absolute score into a score out of 20 (here: 13/20).
- The average relative score was calculated as the mean of the relative scores obtained from reviewer 1 and reviewer 2. This score was used by the pilot of SC to include or exclude the abstract.

Data management and analysis

We extracted data from the online platform used for the submission and evaluation process. We conducted the analyses using the R programming language on the RStudio interface version 2022.07.1 on a Mac computer with macOS Big Sur version 11.7. In line with the principles of transparent data analysis, we made no direct modifications to the original database. We coded all data manipulations in a computational document. The data and analysis code are available here DOI: 10.17605/OSF.IO/TYGFN. JPD conducted all the statistical analyses and carried out all graphical representations, while MGG subsequently cross-checked all of the analyses.

The main decision to include or exclude an abstract relied upon its relative score out of 20. Therefore, we defined the inter-rater reliability and the inter-rater agreement as the main objective of our research. For reliability, we calculated the Intraclass Coefficient Correlation (ICC one-way random effect [9]). As recommended by the GRRAS [8], we determined inter-rater agreement by calculating the Standard Error of Measurement (SEM), Coefficient of Variation (CV) and the Bland-Altman plots with the limit of agreements.

The interpretation of the ICC scores followed the Guidelines: "poor" below 0.50, "moderate" between 0.50 and 0.75, "good" between 0.75 and 0.90, and "excellent" above 0.90 [9]. The CV was calculated as the ratio (%) of the standard deviation to the mean. In a reliability study, it is used to assess the stability of measurements across repeated trials. The SEM indicates the extent to which measured test scores are spread around a 'true' score. Both low CV and low SEM indicates good agreement between reviewers.

We also determined several secondary objectives:

1. Inter-rater reliability and inter-rater agreement in calculating the absolute score for each submission.

2. Inter-rater reliability and inter-rater agreement in the selection of items to be used for each abstract, i.e. reliability and agreement of the "not applicable" mention by the reviewer.

3. Inter-rater reliability and inter-rater agreement in rating each of the 47 items on the evaluation checklist.

4. Internal consistency and redundancy of items in the checklist.

For the first secondary objective concerning a continuous variable, we calculated the ICC (one-way random effect) for reliability. For agreement, the SEM, the CV, and Bland-Altman plots with the limits of agreements were calculated.

For the second and third secondary objectives related to a nominal variable, we calculated Cohen's Kappa for reliability and percentage of agreement for agreement. We interpreted the Kappa values as follows: values below 0.6 as unacceptable, and values above 0.6 as acceptable [10].

We evaluated the internal consistency and redundancy of the items by calculating the Cronbach's alpha coefficient for the entire checklist and for each domain/subdomain of the checklist (final secondary objectives). We set an acceptability threshold between 0.7 and 0.9. A score below 0.7 signalled a lack of consistency among the items, whereas a score above 0.9 suggested redundancies [11, 12].

Results

For the JFK, edition 2023, we received a total of 217 submissions. This corresponded to a theoretical total of 434 evaluations. Six evaluations were excluded, three because they were evaluated by a single reviewer and three because they were rejected without a score. For these three evaluations, the reviewers only provided a general comment, indicating that these abstracts should be excluded, without providing any score as they did not meet the submission requirements outlined. In total 428 evaluations were included in this study. Thirty-six reviewers participated in the evaluation process. Eighteen reviewers (50%) were from the SC, contributing 257 evaluations. Fourteen reviewers (39%) were from the CSV, contributing 153 evaluations. The remaining four reviewers (11%) were from member associations, contributing to 23 evaluations.

The inter-rater reliability of the relative score was found to be poor with an ICC value of 0.39 [CI95% = 0.30; 0.49]. In addition, we found a SEM of 3.08 and a CV value of 23.1%. The Bland-Altman analysis revealed a mean bias of measurement of -0.55 with an upper and lower limit of agreement of 7.98 and -9.08, respectively (**Figure 1**).

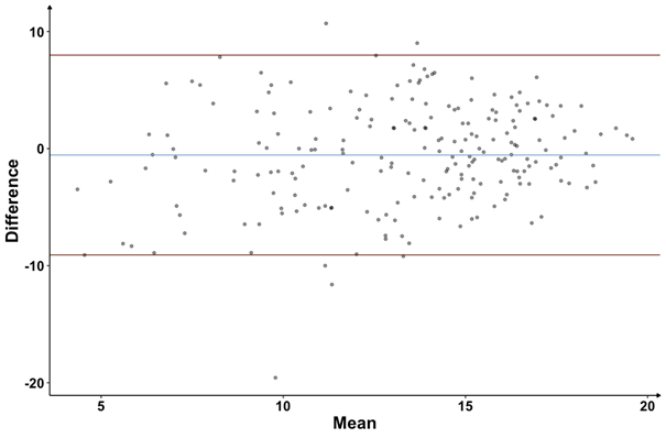

Figure 1 Graphical representation of the Bland-Altman analysis of the agreement of the relative score between the two reviewers. The blue line is the mean bias of the measurement, the red lines are the limits of agreement.

The inter-rater reliability of the process for determining the absolute score was poor too, with an ICC value of 0.25 [IC95 = 0.14; 0.35]. In addition, we found a SEM value of 5.78 and a CV value of 33%. The Bland-Altman analysis revealed a mean bias of measurement of -1.03 with an upper and lower limit of agreement of 14.99 and -17.06, respectively (**Figure 2**).

The reliability of the process for determining the number of relevant items in rating an abstract was also poor, with an ICC value of 0.12 [IC95% = 0.01; 0.23]. In addition, we found a SEM value of 6.68 and a CV value of 32.6%. The Bland-Altman analysis revealed a mean bias of measurement of 0.20 with an upper and lower limit of agreement of 18.75 and -18.35, respectively (**Figure 3**).

The results of the inter-rater reliability for each item of the checklist are presented in **Figure 4** and **Table 1**.

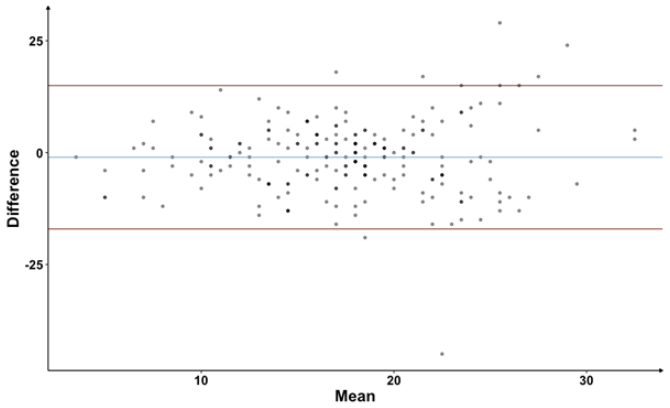

Figure 2 Graphical representation of the Bland-Altman analysis of the agreement of the relative score between the two reviewers. The blue line is the mean bias of measurement, the red lines are the limits of agreement.

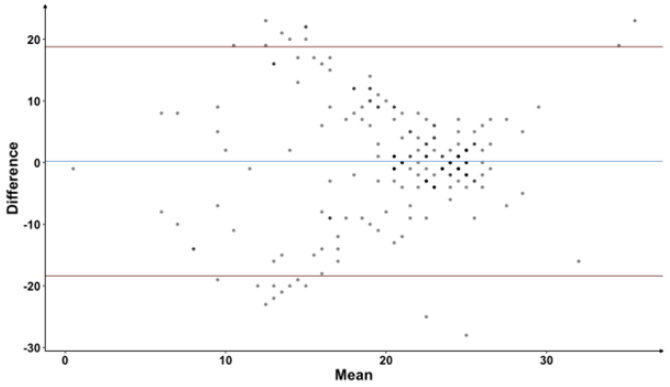

Figure 3 Graphical representation of the Bland-Altman analysis for the "non-applicable" mention for an item between two reviewers. The blue line is the mean bias of measurement, the red lines are the limits of agreement.

Table 1 Results for reliability and agreement for each item on the checklist.

Items[1]	Reliability[2]	Agreement[3]
A1	0.26 [0.11 - 0.41]	77.8 [72.2 - 83.4]
A2	0.03 [-0.05 - 0.11]	84.3 [79.4 - 89.2]
A3	0.08 [-0.05 - 0.21]	63.4 [57 - 69.9]
A4	0.12 [-0.04 - 0.27]	81.5 [76.3 - 86.7]
A5	0.57 [0.43 - 0.71]	88 [83.6 - 92.3]
B1	0.27 [0.12 - 0.41]	75 [69.2 - 80.8]
B2	0.21 [0.07 - 0.34]	69.4 [63.3 - 75.6]
B3	0.15 [0.02 - 0.28]	69.4 [63.3 - 75.6]
B4	0.12 [-0.02 - 0.25]	75.9 [70.2 - 81.7]
B5	0.07 [-0.06 - 0.21]	71.3 [65.2 - 77.4]
C1	0.39 [0.28 - 0.5]	66.7 [60.3 - 73]
C2	0.42 [0.32 - 0.52]	63 [56.5 - 69.5]

Table 1 Results for reliability and agreement for each item on the checklist (*Continued*)

Items[1]	Reliability[2]	Agreement[3]
C3	0.53 [0.44 - 0.63]	72.2 [66.2 - 78.2]
C4	0.32 [0.22 - 0.43]	62.5 [56 - 69]
C5	0.22 [0.1 - 0.34]	63.4 [57 - 69.9]
C6	0.41 [0.29 - 0.52]	71.8 [65.7 - 77.8]
C7	0.45 [0.35 - 0.54]	66.2 [59.8 - 72.6]
C8	0.51 [0.41 - 0.6]	69.9 [63.7 - 76.1]
C9	0.25 [0.15 - 0.36]	61.1 [54.6 - 67.7]
C10	0.42 [0.3 - 0.54]	78.7 [73.2 - 84.2]
C11	0.52 [0.39 - 0.64]	80.6 [75.2 - 85.9]
C12	0.57 [0.45 - 0.69]	83.3 [78.3 - 88.3]
C13	0.34 [0.21 - 0.47]	73.6 [67.7 - 79.5]
C14	0.37 [0.25 - 0.48]	67.6 [61.3 - 73.9]
C15	0.25 [0.14 - 0.37]	61.6 [55 - 68.1]
C16	0.29 [0.18 - 0.4]	61.6 [55 - 68.1]
C17	0.26 [0.15 - 0.37]	60.6 [54.1 - 67.2]
C18	0.38 [0.27 - 0.49]	67.6 [61.3 - 73.9]
D1	0.45 [0.35 - 0.55]	65.3 [58.9 - 71.7]
D2	0.45 [0.35 - 0.55]	65.7 [59.4 - 72.1]
D3	0.49 [0.39 - 0.58]	67.6 [61.3 - 73.9]
D4	0.38 [0.28 - 0.48]	61.6 [55 - 68.1]
D5	0.42 [0.32 - 0.52]	62.5 [56 - 69]
D6	0.33 [0.2 - 0.45]	71.8 [65.7 - 77.8]
D7	0.27 [0.16 - 0.38]	64.8 [58.4 - 71.2]
D8	0.32 [0.22 - 0.43]	66.2 [59.8 - 72.6]
D9	0.25 [0.14 - 0.36]	64.4 [57.9 - 70.8]
D10	0.27 [0.16 - 0.38]	67.1 [60.8 - 73.4]
D11	0.19 [0.09 - 0.29]	53.2 [46.5 - 59.9]
D12	0.19 [0.09 - 0.29]	51.4 [44.7 - 58.1]
D13	0.25 [0.14 - 0.35]	54.6 [47.9 - 61.3]
D14	0.3 [0.19 - 0.4]	58.3 [51.7 - 65]
D15	0.21 [0.1 - 0.32]	58.8 [52.2 - 65.4]
E1	0.05 [-0.08 - 0.17]	72.2 [66.2 - 78.2]
E2	0.17 [0.06 - 0.29]	62.5 [56 - 69]
E3	0 [-0.12 - 0.12]	80.6 [75.2 - 85.9]
E4	0.13 [0.01 - 0.25]	62 [55.5 - 68.6]

[1] Items checklist

[2] Kappa of Cohen [CI95] for the item

[3] Percentage of agreement [CI95%]

None of the Cohen's kappa values reached the threshold of 0.6 set as acceptable *a priori*. Percentage of agreement fluctuated from 51.4% to 88% (**Figure 5** and **Table 1**).

Internal consistency was found to be good to very good with a Cronbach's alpha score for the domains exceeding the predefined acceptable threshold (0.6), except for domain A. The results are presented in **Figure 6**. However, the Cronbach's alpha exceeded 0.9 for each of the sub-domain C1, C2, C3, D1, D2, and D3 indicating probable redundancy.

Discussion

In this study, we assessed the reliability of the rating and selection process for abstracts submitted to the 2023 edition of the JFK. This represents the first analysis of its kind conducted by the SFP. Our findings indicate that the existing abstract selection process lacks reliability and, therefore, needs improvement.

Firstly, the average relative score, used as a reference by the SC to determine the inclusion or exclusion of submitted abstracts, exhibits an ICC of 0.39 [CI95 = 0.30; 0.49], a SEM value of 3.08, and a CV value of 23.1%. This high CV indicates significant variability between the measurements from Reviewer 1 and Reviewer 2 [13, 14]. In our results, the SEM was also high, with 3.98 points representing 15.4% of the total relative score (out of 20). The Bland-Altamn analysis revealed significant data dispersion with upper and lower 95% limits of 7.98 and -9.08 (out of a total score of 20). These results indicate that scores between two reviewers could vary to a considerable extent. Such variability can easily result in a change in status from accepted to rejected for an abstract, without necessarily reflecting its intrinsic quality. Thereby, this aspect necessities improvement for subsequent editions.

Upon detailed examination of the individual items, none of the Cohen's kappa values reached the predefined acceptable threshold of 0.6. Therefore, the reliability of rating individual items can be considered poor. A potential explanation for the substantial heterogeneity in ratings could be attributed to the significant freedom afforded to reviewers in selecting items they deemed relevant. Out of the 47 available items available, reviewers were free to choose the "not applicable" option if they deemed an item irrelevant for assessing the abstract. This discretionary choice could vary from one reviewer to another and appears to reduce reliability to an unacceptable level. The ICC value to determine the number of relevant items in rating an abstract reflected poor results too, with ICC = 0.12 [IC95% = 0.01 ; 0.23], SEM = 6.68 (out of a total of 47 items), and a CV value of 32.6%. This factor could also explain the low reliability and agreement results for the absolute score. The ICC for the absolute score is even lower than that of the relative scores: 0.25 [IC95% = 0.14; 0.35] compared to 0.39 [CI95% = 0.30; 0.49], respectively. Consequently, reviewers seemed to diverge on the items applicable for each reviewed study. This observation implies a potential lack of clarity in the items or overall checklist, a lack of expertise among reviewers, or excessive item redundancy.

The results reveal that the Cronbach's alpha coefficients are high (see **Figure 6**), indicating strong internal consistency. The only exception is Domain A, related to "Quality of Delivery," which has a Cronbach's alpha below 0.7 (Alpha = 0.42). This outcome initially appears puzzling, as all items within Domain A pertain to the delivery of the abstract and seem to be congruent with the writing. Upon closer examination of the data, a more significant divergence in reviewer responses for item A4 appeared compared to the other four items. Items A1, A2, A3, and A5 were used (scored either "yes" or "no") by the reviewers 81%, 92%, 73%, and 82% of the time, respectively, whereas item A4 was utilized in only 3% of cases. Item A4 relates to the format of tables and figures. However, due to editorial issues, authors were instructed not to submit tables or figures. Consequently, considering this instruction, item A4 should have been removed prior to the rating process. We recalculated the Cronbach's

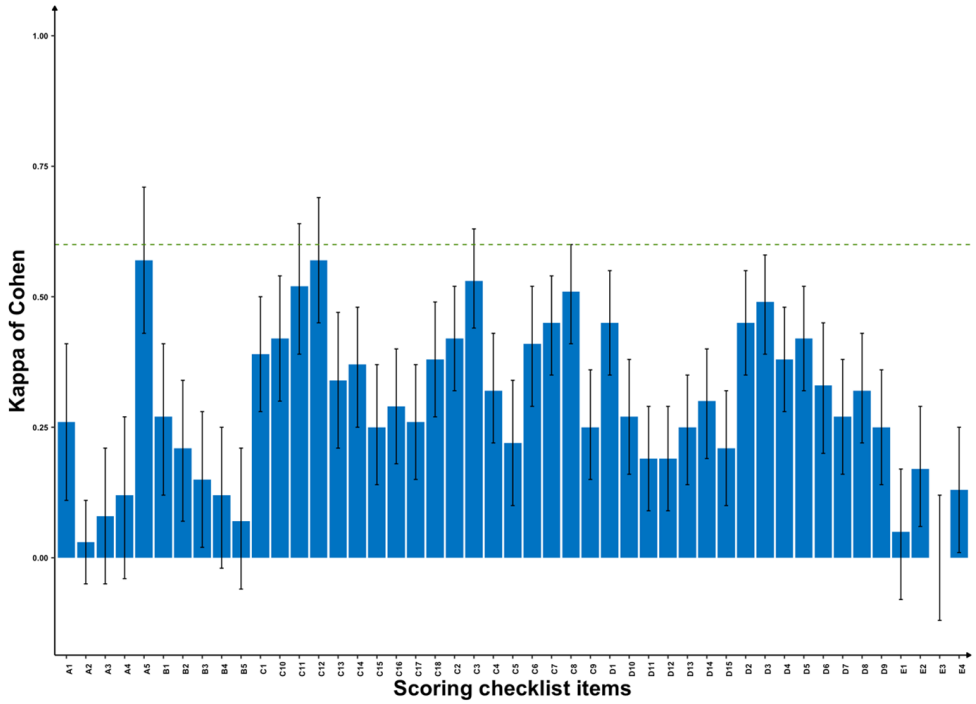

Figure 4 Graphical representation of Cohen's Kappa scores for each of the 47 items on the evaluation checklist, along with 95% confidence intervals. The items are plotted on the x-axis, while the Cohen's Kappa scores are shown on the y-axis. The green line represents the predetermined threshold of acceptability.

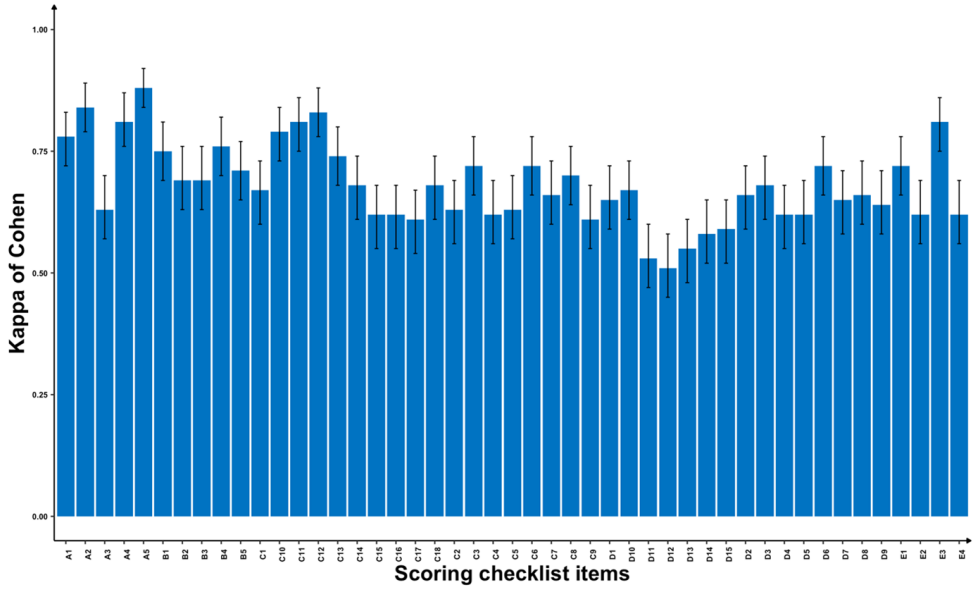

Figure 5 Graphical representation of the percentage agreement between two reviewers, with the 95% confidence interval, for each item.

alpha for Domain A, excluding item A4. The result improved, but still fell below the 0.7 threshold (0.52), indicating the necessity to rewrite this domain.

For six out of the 10 domains in the checklist, we observed a Cronbach's alpha coefficient higher than 0.9. Such a score indicate redundancy among items. For example, we can identify redundancy in the wording between item C3 "There is a comparator group present" and item C6 "Procedures that could be randomised have been randomised".

To the best of our knowledge, the reliability of the conference abstract rating process is rarely addressed in the literature. Among the scarce studies we have identified, our findings align with those previously published, revealing an overall low reliability of the process across various fields of health science, including internal medicine [6], paediatrics [5], and hepatology [7]. Altogether, these results underscore the necessity for conference organisers to replicate such reliability analyses and establish robust quality enhancement processes.

The overall results of the abstract reliability scoring process at JFK indicate a clear need for process improvement. We propose some avenues

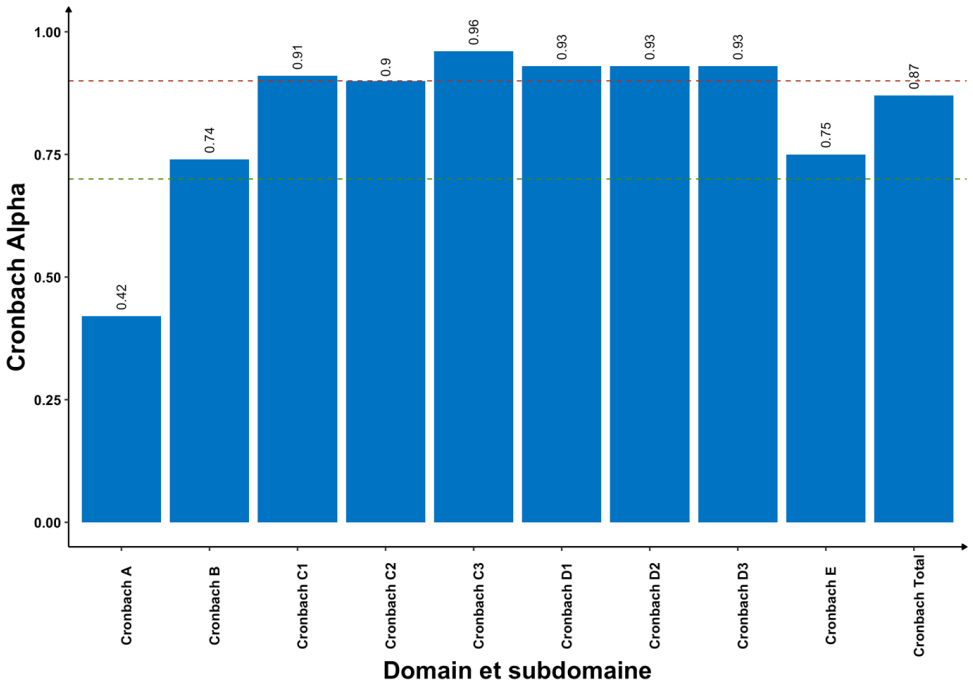

Figure 6 Graphical representation of the Cronbach's alpha for each domain/subdomain of the checklist and for the total checklist. The green and red lines represent the predetermined threshold of acceptability.

for improvement. The fields in rehabilitation are numerous, as are the types of possible studies. Thus, 47 items may not be enough to cover such diversity. However, it would be unreasonable to include additional items. Such an approach would increase the complexity of using the checklist and could further decrease the reliability of the process. An alternative approach could involve inviting authors to select the category of their abstract upon submission. Based on the chosen category for the abstract, specific items (aligned with that category) could be presented to reviewers for scoring, thereby negating the necessity for reviewers to identify the items that are "not applicable". In this scenario, the SC would need to design a checklist for each study design. Another strategy would consist in simplifying the scoring checklist. To this end, rather than employing three disparate grids types for three different study designs, we suggest the establishment of common criteria. This task will be undertaken by the subsequent SC of a future edition of the JFK. This approach addresses additional identified issues with the existing checklist:

a) The reviewers were choosing the items in the grid that they deemed relevant to assess the current abstract based on its design and methods. As the ICC for the choice of relevant items is very low (ICC = 0.12; [IC95% = 0.01; 0.23]) this relative freedom seems to be one of the factors that decrease the reliability of the whole process. With a general list of items that could apply to all designs and methods, this problem no longer exists.

b) The wording of items specific to the study designs. In the current checklist, Domains C and D are subdivided into three subdomains specific to study designs. From our perspective, the wording of these three subdomains lacks clarity. The current checklist has:

   - Interventional studies.
   - Reviews.
   - Social and human sciences.

   These items evaluate quantitative, qualitative, and review studies. The wording "social and human sciences" implies that a sociology study cannot employ quantitative methods, and likewise, a musculoskeletal study cannot use qualitative methods. We believe this to be confusing. Again, a common checklist to all study designs eliminates this issue.

In addition to clarifying the checklist, it is imperative to ensure that reviewers are proficient in utilising the checklist. Consistent with previous studies, we proposed a video to explain the scoring and use of the checklist [4, 15], albeit with limited effect. A prospective pathway for enhancement involves refining the scoring explanations for subsequent editions. Additionally, we posit that instituting an initial testing phase, wherein reviewers employ the checklist on a fictitious abstract would foster familiarity and allow for the addressing of any emergent queries or concerns. While this alternative seems prudent, it conjures questions related to the workload imposed on volunteer reviewers. The task of reviewing abstracts already constitutes a considerable endeavour, and it is not our intention to amplify this burden. However, the necessity for executing the task with utmost efficiency remains paramount.
Finally, beyond reliability, validity assessment of the checklist and the whole process of scoring should be assessed in the future.

Conclusion

In this study, we assessed the reliability of the abstract scoring process used in the 2023 JFK conference. While the overall internal consistency is good, the scoring process lacks reliability and agreement, *i.e.* ICC = 0.39 [CI95% = 0.30; 0.49], SEM = 3.08 and a CV = 23.1%. Based on the obtained results, we have put forward proposals for modifications that could be applied to future JFK submissions as well as to all conferences organised by scientific societies within the SFP network.

Statement and declaration

Competing Interests

All authors declare they have neither financial nor non-financial interests.

Disclosure statement

All authors declare they have neither financial nor non-financial interests.

References

[1] David Ross Appleton and David N Kerr. Choosing the programme for an international congress. *BMJ*, 1(6110):421–423, 1978. doi: 10.1136/bmj.1.6110.421.

[2] Roberta W Scherer, Kay Dickersin, and Patrica Langenberg. Full publication of results initially presented in abstracts: a meta-analysis. *JAMA*, 272(2):158–162, 1994.

[3] Roberta W Scherer, Joerg J Meerpohl, Nadine Pfeifer, Christine Schmucker, Guido Schwarzer, and Erik von Elm. Full publication of results initially presented in abstracts. *Cochrane Database of Systematic Reviews*, 20(11):1–565, 2018. doi: 10.1002/14651858.mr000005.pub4.

[4] Jaime Jordan, Laura R Hopson, Caroline Molins, Suzanne K Bentley, Nicole M Deiorio, Sally A Santen, Lalena M Yarris, Wendy C Coates, and Michael A Gisondi. Leveling the field: Development of reliable scoring rubrics for quantitative and qualitative medical education research abstracts. *AEM Education and Training*, 5(4):e10654, 2021. doi: 10.1002/aet2.10654.

[5] Kathi J Kemper, Paul L McCarthy, and Domenic V Cicchetti. Improving participation and interrater agreement in scoring ambulatory pediatric association abstracts: How well have we succeeded? *Archives of Pediatrics & Adolescent Medicine*, 150(4):380–383, 1996. doi: 10.1001/archpedi.1996.02170290046007.

[6] Haya R Rubin, Donald A Redelmeier, Albert W Wu, and Earl P Steinberg. How reliable is peer review of scientific abstracts?: Looking back at the 1991 annual meeting of the society of general internal medicine. *Journal of General Internal Medicine*, 8(5):255–258, 1993. doi: 10.1007/bf02600092.

[7] Hendrik Vilstrup and Henrik Toft Sørensen. A comparative study of scientific evaluation of abstracts submitted to the 1995 european association for the study of the liver copenhagen meeting. *Danish Medical Bulletin*, 45(3):317–-319, 1998.

[8] Jan Kottner, Laurent Audigé, Stig Brorson, Allan Donner, Byron J Gajewski, Asbjørn Hróbjartsson, Chris Roberts, Mohamed Shoukri, and David L Streiner. Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. *Journal of Clinical Epidemiology*, 64(1):96–106, 2011. doi: 10.1016/j.jclinepi.2010.03.002.

[9] Terry K Koo and Mae Y Li. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2):155–163, 2016. doi: 10.1016/j.jcm.2016.02.012.

[10] J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.

[11] David L Streiner. Starting at the beginning: An introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment*, 80(1):99–103, 2003. doi: 10.1207/s15327752jpa8001_18.

[12] Mohsen Tavakol and Reg Dennick. Making sense of cronbach's alpha. *International Journal of Medical Education*, 2:53–55, 2011. doi: 10.5116/ijme.4dfb.8dfd.

[13] Joseph L Fleiss, Bruce Levin, and Myunghee Cho Paik. *Statistical methods for rates and proportions*, volume Wiley series in probability and statistics. John Wiley & Sons, Hoboken, NJ, 3rd edition, 2003.

[14] Michael J Campbell, Stephen J Walters, and David Machin. *Medical statistics: a textbook for the health sciences*. Wiley-Blackwell, Hoboken, NJ, 5th edition, 2020.

[15] Nia S Mitchell, Kelly Stolzmann, Lauren V Benning, Jolie B Wormwood, and Amy M Linsky. Effect of a scoring rubric on the review of scientific meeting abstracts. *Journal of General Internal Medicine*, 36(8):2483–2485, 2020. doi: 10.1007/s11606-020-05960-6.