Author’s response to reviews

Title: SANRA – A Scale for the Quality Assessment of Narrative Review Articles

Authors:

Christopher Baethge (cbaethge@uni-koeln.de)

Sandra Goldbeck-Wood (goldbeckwood@doctors.org.uk)

Stephan Mertens (mertens@aerzteblatt.de)

Version: 1 Date: 05 Feb 2019

Author’s response to reviews:

As requested, we have pasted our point-by-point-letter in this box. For the sake of comprehensibility and ease of process we refer the editors to the corresponding document that we have uploaded as point-by-point-letter under personal cover.

Dear Professor Meerpohl,

Thank you very much for giving us the opportunity to revise our manuscript „SANRA– A Scale for the Quality Assessment of Narrative Review Articles“ . We are grateful to both referees, Dr. Gasparayan and Ms. Struthers, for the time and effort they have put in reviewing our manuscript and for their comments. In the following, we will, on a point-by-point basis, refer to all points raised by the referees and document all changes made in the revised manuscript. The changes are also visible in the manuscript’s „track changes“-version (line numbers refer to „track changes“-version). In addition, we submitted a „clear“ version of our paper. For clarity, we have numbered reviewer 2’s points.

We hope we have addressed all points sufficiently, either in following the recommendations by the reviewers or in explaining when we did not.

Reviewer 1

This is a well designed and properly conducted study. I would like to add a few comments.

1. Background. It would be correct to avoid referring to narrative review as "unreliable" (3rd para, 1st line). There are numerous redundant systematic reviews that add nothing to the global pool of evidence medicine (https://www.ncbi.nlm.nih.gov/pubmed/27620683). Whereas narrative reviews of experts in their fields, which are published by review journals (e.g., Nature Review series, Expert Opinion in series) are more helpful.
Authors’ response:

We thank reviewer 1 for referring us to the instructive paper by John Ioannidis. We now cite this source, and we have changed the paragraph accordingly. However, we would like to stick with our current statement that narrative reviews are widely considered to be unreliable. This is not only stated in the paper by Bastian et al. – one of the frequently cited texts in evidenced based medicine – and in the other reference we provide, it is also our experience in our discussions with authors, reviewers, editors, and other experts.

The paragraph now reads (line 108-111, changes in italics):

While systematic reviews are not per se a superior article genre and while certain systematic reviews have been criticized lately [Ref Ioannidis], unsystematic reviews, or narrative reviews are widely considered to be unreliable [1, 3]. Hence,…

and

Apparently, narrative reviews are very popular among authors and it is plausible to assume that they exercise an enormous influence among doctors in clinical practice and research. (line 93-95)

2. Discussion. The authors present a valid option for evaluating narrative reviews. However, it is advisable to amend the SANRA, particularly its Item 3 - Description of the literature search. Points on whether searches through globally recognized multidisciplinary (i.e. Scopus, Web of Science) and specialist databases (MEDLINE/PubMed, CINAHL, Global Health, PEDro, SPORTSDiscus etc.) were employed or not could be added to the SANRA.

Authors’ response:

In the explanations and instructions document we refer to databases as desirable specifications for item 3 (please see Figure 2, item 3). In our opinion, adding databases to the SANRA tool itself (Fig 1) would not go beyond what is already there: search terms and inclusion criteria. In our experience, it is likely that authors who state their search terms also name the database they searched.

Reviewer 2

Although I work for the EQUATOR Network I would class myself as a consumer reviewer as I am not a trained researcher or statistician. This manuscript was of interest to me because it is aimed at evaluating a tool designed to improve conduct and reporting of research, in this case narrative reviews. I am an extremely infrequent reviewer, so my review reveals a lack of knowledge of much of the statistical and research jargon used. I think these comments are valid as the target reader and user of the knowledge (journal editors) may also not be familiar with the technical terms.
Authors’ response:

This (readers and users are not familiar with technical terms) is an important assumption, but is it justified? Here are two points to consider:

1. The target readers of this manuscript as well as the target user of SANRA are authors, reviewers, researchers, and editors. In our experience, and many of those groups have at least some statistical training and would welcome, perhaps even require, brief mention of specific statistical techniques. We appreciate that this may not be the case for all readers, but suggest that by keeping these references brief, we will not overly encumber the reading experience for either group. We would point out that the terminology used in this manuscript is standard in papers introducing assessment tools, and note that reviewer 1 and the editor did not express their difficulties in understanding our methods, both recommending a minor revision. Still, most readers, the authors of the present paper included, cannot fully grasp the methods in most papers, be it the specifics of statistics or of advanced experimental methods, such as imaging, genetics or qualitative research. Explaining the methods of research papers to the full extent would lengthen articles to the point of being unreadable. In our view, the statistics in our paper, as well as the methodological descriptions in other papers, serve the purpose of informing the expert reader so that it is transparent to this particular group that what has been done in the study is not unsound.

We have therefore tried to find some middle ground between what reviewer 2 proposed and what we think is a reasonable degree of detail.

1. Introduction

1.a. Could maybe mention that AMSTAR, CASP etc. are complex and you wanted to develop something simpler for narrative reviews which are more numerous. I don't see the logic of arguing that narrative reviews are "unappreciated" because they are the largest text type. Or that they are influential for the same reason. The assertion that narrative reviews are widely known to be unreliable points to a 1987 reference. It is widely known that all the medical research literature is unreliable.

Authors’ response:

We have now tried to clarify the rationale behind the small number of items in SANRA in Introduction and in Discussion.

Introduction, line 132-134:

As a consequence, we have developed SANRA, the Scale for the Assessment of Narrative Review Articles, a brief critical appraisal tool for the assessment of non-systematic articles.
Discussion, line 308-310:

SANRA’s brevity is also in contrast to other tools to assess articles, such as AMSTAR 2, for systematic reviews, or, to a lesser extent, CASP for RCTs, with its 16 and 11 items, respectively.

Reviewer 2 finds our statements that narrative reviews are underappreciated and at the same time the most popular article genre in medicine, inconsistent. We have tried to clarify this seeming contradiction: narrative reviews are underappreciated by the EBM scene but popular among many authors and influential among doctors (Introduction, line 93-97):

Narrative reviews also appear popular among both, authors and readers, and it is plausible to assume that they exercise an enormous influence among doctors in clinical practice and research. However, because their quality varies widely, they have frequently been compared in blanket, negative terms with systematic reviews

We have cited Mulrow’s 1987 paper because it is the classic work with regard to this topic. We cannot follow reviewer 2’s statement that „all the medical literature is unreliable“. In our view, the whole ebm movement is largely about differentiating various degrees of validity and reliability, and it has succeeded in showing such differences. We submit that it is undisputed in medicine that there are important differences in reliability with regard to different study and publication types.

1.b. It would be good perhaps to phrase the last sentence of the background to explain what you mean by item-total correlation, internal consistency and criterion validity. These may not be understood easily by the target readership? I had to work them out by reading and re-reading the article. I'm still not quite sure…

Authors’ response:

We propose not to change the last sentence of the background section because it is intended to briefly summarize the aims of the study. Instead, we have added short definitions to the methods section. We welcome the editor’s guidance as to whether this really increases the readability of the paper or whether it comes across as a little awkward to explain terms like interrater reliability.

Methods, line 177-188:

The study aimed at testing SANRA’s internal consistency (Cronbach’s alpha) and the item-total correlation – indicating whether the items measure the same phenomenon, here different aspects of review paper quality – as well as SANRA’s inter-rater reliability with regard to its sum score. Inter-rater reliability, as a measure of the consistency among different raters, was expressed as the average measure intraclass correlation, ICC, using a two-way random effect model (consistency definition). As an approximation of SANRA’s criterion validity (Is the score predictive of other indicators of paper quality, e.g., acceptance and rejection or citations?) we
analyzed posthoc whether average sum scores of SANRA were associated with the decision to accept or reject the 30 manuscripts under study (pointbiserial correlation for the association between a dichotomous and a continuous variable).

2 Methods

2.a. Some of the language in this article is quite complicated/wordy with a lot of jargon - almost seems to be over-compensating for the relative simplicity of the study! Eg. "The maximal sum score is 12." "The sum score of the scale is intended to measure the construct…"

Authors’ response:

We propose to leave the sentence „The maximum sum score is 12“ as it is. We believe there is no harm in mentioning the sum score once because sum scores are key characteristic of scales. We also believe it prepares the reader for the results section.

We also propose to leave the sentence „The sum score of the scale is intended to measure the construct “quality of a narrative review article”“ as it is now. It should be explicitly stated once what exactly the scale is intended to measure. In this vein, the term „construct“ means that „quality of a narrative review“ is a cultural construct rather than a natural phenomenon that can be exactly measured, such as weight or temperature.

In addition, the statement is important because it clarifies that it is the sum score that is of interest (and psychometrically measured in this study) not single items (this is a frequent misuse of rating scales: using single items or subscores when the reliability has been studied with regard to the total score only). Please see also our answer to reviewer 2’s point 3.a. below.

2.b. In describing the statistics I find it that I am expected to know about things like Cronbach's alpha and ICC. I also needed to know what your pre-specified criteria for "success" in consistency and reliability scores, based on tests of comparable scales such as AMSTAR. I have no idea what criterion validity or pointbiserial correlation is.

Authors’ response:

Regarding the meaning of criterion validity and pointbiserial correlation we have now explained the terms (see our answer to reviewer 2’s point 2.b.).

Within the space constraints of a scientific article we have explained the concepts essential to the understanding of our paper (see our answer to reviewer 2’s point 1.b. above). For example, we have explained what Cronbach’s alpha measures.

We are hesitant to go into the details of ICC, because, similar to coefficient alpha, we now explain what ICC measures but deliberately omit detailed explanation as to how it is derived, which would need a lot of text. Of note, it is standard to use ICC for measuring intrarater
reliability, the specific information for methodologists of rating scales here is that ICC was arrived at by using the average measure calculation, but should we really delve into that?

As for pre-specified criteria for success we doubt that there are universally agreed upon thresholds for rating scales. For example, it is often said that Cronbach’s alpha should be at least 0.7. However, since, statistically, alpha depends on the number of items 0.7 would not be very striking in a scale of 20 items whereas a value of 0.68 in a six-item scale (as in our study) is reasonably good. We have explained this (discussion, line 255-259).

It should be noted that because coefficient alpha increases with the number of items [12], simplifying a scale by reducing the number of items – as we did – may decrease internal consistency. However, this needs to be balanced against the practical need for brevity. In fact, the earlier seven-item versions of SANRA had higher values of alpha: 0.80 and 0.84, respectively [9].

It also follows, other than reviewer 2 assumes, that AMSTAR 2 is not comparable to SANRA. Discussion, line 308-310:

SANRA’s brevity is also in contrast to other tools to assess articles, such as AMSTAR 2, for systematic reviews, or, to a lesser extent, CASP for RCTs, with its 16 and 11 items, respectively.

The same applies for ICC or item-total correlation: Some say an ICC of 0.5 is desirable, others favor 0.6 (in our study 0.77), similar to well-known thresholds for other correlation coefficients. AMSTAR’s ICC, for example, was 0.61, in a 2017 study. Item-total correlation should be above 0.2 (in our study all items above 0.32). Like coefficient alpha, inter-rater reliability increases with the number of items, which makes comparisons among scales difficult. If anything, the comparisons add to the impression that SANRA is quite consistent and reliable. We discuss that the correlations in our study indicate sufficient consistency and reliability (discussion).

Finally, it is important to realize that the psychometric results we present are specific to our study, not necessarily to SANRA in general. Therefore, it is possible, even likely, that in a different setting, researchers would arrive at different figures. Our study only demonstrates that SANRA, in principle, can be used with sufficient feasibility, validity, consistency, and reliability. This is an underappreciated fact in the discussion of scales, and we have emphasized it twice in discussion (discussion, line 279-282, no change; discussion – limitations, line 340).

2.c. It's great that you followed the GRRAS reporting guidelines. It would be good herefore to have a copy of the checklist submitted with the ms so reviewers (including me) could see where you have reported each of the items, and an explanation of why some items weren't appropriate.

Authors’ response:

While some journals require authors to fill out tables of reporting guidelines like CONSORT or PRISMA on submission of a manuscript this is not a requirement for submissions to Research Integrity and Peer Review. We therefore did not hand in a GRASS-based table.
However, we followed GRASS to the letter except for the recommendation to present a sample size calculation because we deliberately decided against it. We have discussed this limitation in discussion (line 328). Finally, we also did not follow GRRAS in mentioning in the title what exactly we measured because this would make the title too long, and interested readers (or researchers) will easily find the abstract, anyway. The often issued recommendation for titles regarding the method is caused by the fear that without comprehensive titles papers may be lost in searches for systematic reviews, but, today, every decent systematic review is based on searches of both titles and abstracts.

3. Results

3.a. Surely the total number of ratings was 3 (number of raters) x 30 (number of mss) x 6 (number of items) = 560. It would be good to make the raw score data available in an excel spreadsheet as well as presenting summary statistics which could then be checked. In the methods you said you would present medians where appropriate. I think natural frequencies might be more appropriate than means for a 3 point scale? I didn't see any medians …...were there any?

Authors’ response:

When we refer to ratings, we refer to the ratings of the sum scores (as stated in methods, please see also point 2.a.), and surely this number is 90.

As previously stated in the manuscript, we will provide the data upon request.

We are uncertain as to what reviewer 2 means by requesting summary statistics, because we did provide summary statistics in abstract and results.

We do not agree that means are not appropriate for a three-point-item (which are not a scale, the scale is the group of six items, and its range is 12, see above). This applies even in slightly positively skewed distributions as in some of the items. Means allow to order items. It is, however, true that we did not present medians so far, and we thank reviewer 2 for this constructive point. For completeness, we have now added medians (Results, line 207-212):

All 90 ratings (3 raters x 30 manuscripts) were used for statistical analysis. The mean sum score across all 30 manuscripts (N=90) was 6.0 out of 12 possible points (SD: 2.6, range: 1-12, median: 6). Highest scores were rated for item 4 (mean: 1.25; SD: 0.70), item 2 (mean: 1.14; SD: 0.84), and item 1 (mean: 1.1; SD: 0.69) whereas items 6, 5, and 3 had the lowest scores (means of 0.81 (SD: 0.65), 0.83 (SD: 0.67), and 0.84 (SD: 0.60), respectively) (all single-item medians: 1).

3.b. Disagreements most often occurred with item 1 and 4. Would be useful to know how often the disagreement was total (ie. 0 and 2) and how often 0 and 1 or 1 and 2.
Authors’ response:

We thank reviewer 2 for this constructive recommendation and we have added the frequencies to the results section (results, 222-225):

Across 180 (6 x 30) single-item ratings (6 items x 30 manuscripts) the maximum difference among the three raters was 2 in 12.8% (n=23; most often in items 1, 2, and 4), in 56.7% (n=102) the raters differed by no more than 1 point, and in 30.6% (n=55) they entirely agreed (most often in items 2 and 3).

3.c. Again the language is quite complicated…eg. "the lower level of the confidence interval (0.57) indicates that we cannot exclude a degree of reliability that would be difficult to accept in most settings" I have no idea what this sentence means. Where does the 0.57 come from? This doesn't look like a confidence interval

Authors’ response:

The figure 0.57 is the lower level of the confidence interval of the intraclass coefficient (ICC) as reported in results (line 226). Normally, confidence intervals’ lower (and upper) levels are interpreted as a reasonable range of uncertainty of a measure, in this case ICC. The lower level is too low, by any standard, to be called satisfactory, so that in most applications of SANRA such a degree of interrater reliability would not be enough. This is a limitation of the study and thus we mention it in discussion. However, in order to half the CI we would have needed 120 articles. As a result, this is an uncertainty we have to live with until someone carries out a study with 120 manuscripts (a number rarely achieved in scale development). In order to make the sentence easier to understand we have changed it (discussion, line 270-272):

While the ICC suggests sufficient reliability the lower confidence limit (0.57) does not preclude a level of reliability normally deemed unacceptable in most applications of critical appraisal tools.

3.d. The result that all raters confirmed that completing the scale is feasible in everyday editorial work is unsurprising given that the raters were the editors developed the scale specifically for that purpose.

Authors’ response:

In our experience, it is tempting for scale developers to overload scales with items so that they are difficult to use. Since our scale comprises of just 6 items, we believe it is suitable in everyday editorial work. We propose to leave the sentence as it is.

4. Discussion
4.a. It would be good to have included discussion of why there was more disagreement with items 1 and 4. They have since been edited along with 5 and 6, but you didn't report what changes have been made.

Authors’ response:

We are not sure why ICC is lower with items 1 and 4, it would be great if we (or any scale developer) knew exactly. And we have not described it in detail because the subsequent edits were only minor (as are the revisions in the explanations and instructions document).

4.b. The conclusion that the internal consistency and item total correlation are sufficient and the inter-rater reliability is satisfactory seems based on the authors' personal judgement rather than evidence. As I mentioned before there is no comparison with reliability tests of other similar scales such as AMSTAR which I presume exist.

Authors’ response:

We have referred to this at length in our answer to reviewer 2’s point 2.b. As an aside: It is a strong assumption by reviewer 2 that authors have no reasons other than their personal judgment for their statements just because they do not refer to each and every source.

4.c. I read up on Cronbach's alpha here https://stats.idre.ucla.edu/spss/faq/what-does-cronbachs-alpha-mean/and it says that a reliability coefficient of .70 or higher is considered "acceptable" in most social science research situations. In this study the figure was 0.68. The earlier version of the scale with seven items had 0.80 and 0.84. I think the argument that the manuscripts which were included in the two earlier studies might have been easier to rate is weak.

Authors’ response:

On coefficient alpha: see above.

On the difference in comparison to earlier versions of the scale: One important difference is that these versions contained one more item (as discussed).

4.d. You conclude that because the intraclass correlation is .77 SANRA can be used by different raters. Again it would be good to have a comparison with the correlation achieved by other scales. And what intraclass correlation actually means.

Authors’ response:

See above.
4.e. I think rather than recommending training, it would be better to test the further tweaked SANRA again pragmatically - perhaps in a randomised trial with a larger set of mss's (sampled from the published literature) and with a larger number of raters independent of the authors of this study.

Authors’ response:

In our opinion, it is desirable to practice filling out SANRA in a set of manuscripts (training) before using it for its various purposes. Obviously, the extent of the training depends on its intended use, be it for editorial decision making, for improving one’s own manuscript or paper review, or for research purposes. Here, SANRA is not different from most other scales in medicine: Filling out scales is no trivial exercise. An example is the Hamilton Rating Scale for Depression (HRSD) that is used in both clinical decision making and in research projects, with rater training as the norm in research projects and, to a lesser extent, in clinical work.

Also, we started to develop SANRA in 2010, and this is already the third version of SANRA. It is likely that even a fourth version would result in less than perfect results. We are not aware of any RCTs in scale development (as opposed to their later use), and we think SANRA is now widely tested and ready for use.

4.f. Another idea would be to collaborate with others to develop a tool to help researchers plan, conduct and report narrative reviews. They are indeed a very important type of study which form the foundation of all primary research so it would be more constructive to work with other stakeholders (eg. the authors referenced 12-14 as having written editorials.), interested in improving both research conduct and reporting. I find it frustrating that a lot of research on improving conduct, reporting and quality appraisal is not joined up.

Authors’ response:

In fact, we had input from other editors. However, we tried to win other editors for a joint testing of SANRA, and we failed. We are certainly open to all collaborations imaginable, but we now feel that in order for such projects to happen SANRA has to be out and attract a minimum amount of attention.

4.g. I think it's impractical to expect journals or anyone else to instigate rater training on the use of such a tool for editors and peer reviewers of one particular type of study however common it is. Having said that, I can see the value of using SANRA for training early-career researchers in how to conduct better quality narrative reviews in the first place whether for publication, or as foundation of their primary research.

Authors’ response:

See above. We have made different experiences. We are confident, however, that in a study employing SANRA a lack of rater training would be considered a major limitation.
4.h. I am totally in agreement that simplicity is key but I still think the simplicity should preclude the need for training which in the real world will not happen - at least not for one study design. More development and testing needs to be done, and if the tool is proved reliable and useful, it should be made easy to find and use for authors, editors and peer reviewers.

Authors’ response:

(See above)

Changes made independently from the reviewers’ comments:

In order to emphasize the need for a critical appraisal tool for narrative reviews in comparison to other fields, we slightly change the following sentence.

Background, line 124-125:

For narrative reviews, in contrast, no critical appraisal, or quality assessment tool is available.

Dear Professor Meerpohl, we hope the manuscript is now ready for acceptance in Research Integrity and Peer Review, but should you think more work needs to be done we are open to discussion.

Sincerely,

Christopher Baethge, MD,
Sandy Goldbeck-Wood, MD,
Stephan Mertens, PhD