Author’s response to reviews

Title: Objective structured assessment of technical competence in transthoracic echocardiography: a feasiblity study

Authors:

Dorte Guldbbrand Nielsen DGN (dornis@rm.dk)
Ole Goetzsche OG (olegoetz@rm.dk)
Berit Eika BE (be@medu.au.dk)

Version: 2 Date: 19 December 2012

Author’s response to reviews: see over
Dear editor,

On behalf of all authors, I would like to thank both reviewers for some very useful comments on our manuscript; Objective structured assessment of technical competence in transthoracic echocardiography: a validity study in a standardised setting. Below, you will find our responses to all reviewer comments. We believe that the changes made to the manuscript based on these comments have improved the quality of the article and hope that you will consider our revised manuscript for publication.

Reviewer 1 – Alan Merry

The title makes explicit that this study is a feasibility study, but it isn’t written as a feasibility study.
- the title has been revised according to reviewer comments

The introduction should end with a clear statement of the objectives of the study
- the final paragraphs of the introduction have been rewritten to clearly state aim and objectives of the study

I would be sceptical about any stronger conclusion in relation to the instrument itself, and would be more convinced by a conclusion explaining how this study has informed the protocol of a larger study to establish the validity of your tool
- the conclusion has been modified according to reviewer comments. The conclusion now emphasizes that the study is just the beginning of a larger study of the assessment instrument in a clinical setting.

You explain that informed consent was obtained from participants, but you do not explain how the patients were recruited, what consent they gave, and whether or not they were paid.
- the recruitment of patients has now been clearly described in the methodology section

You talk about correlations between the level of expertise and particular results, but at no point do you actually explain how you rated expertise. Was this on a one to three scale? Do you consider it was linear? If so, do you think that that assumption was justified?
- expertise was rated by clinical positions – intern, cardiology resident and cardiology consultant. As previous studies have shown that there is not correlation between TTE competence and number of examinations performed or years of experience, it did not seem reasonable to use these demographics as base for level of expertise. I do not consider expertise to be linear, but since the true elements of expertise in echocardiography are not
known it is very difficult to describe exactly what makes the true echo expert. I know that our definition of expertise levels leaves much to wish as is the case with other novice-expertise studies. Particularly as echo training in Denmark is very unstructured it trainees at different levels will have very varying experience and supervised training. The selection of participants and definition of the three levels of expertise is described in the methodology section.

I was a little puzzled about your inter-observer reliability data. You appear to have had the same observer do the measurements twice, and the second observer repeat their measurements once. This seems to be a very small number of observers to base such an assessment on
- Due to the extensive time consumption needed for scoring all 45 examinations, we decided to let just one of the observers rate all examinations and then by random pick 10 examinations (3 novices, 3 intermediates, 3 experts, 1 unknown) to be evaluated by the same observer again and also to be evaluated by another observer.
This has been explained in the manuscript and mentioned as a limitation of the study.

Table two and three simply show P values. These are of very little value on their own.
- both tables have been omitted from the manuscript and the results written in the results section

In your study you have looked at the use of the tool to distinguish between levels of expertise and also to distinguish between easy cases and more difficult cases. This seems reasonable but I don’t think that objective was made clear at the outset at the end of the introduction, and it should be.
- the objectives have been more clearly stated in the introduction section. "Evidence based on response processes and relations to other variables is built on ensuring a study procedure close to everyday practice and testing of hypotheses of relations between three different levels of clinical experience and test scores as well as case complexity and test scores."

In the limitation section I think the primary limitation that needs emphasis is your small numbers.
- the limitation section has been modified to meet the suggestions of the reviewers with an emphasis on the small numbers of participants and cases studied.

I don’t think you have explored the potential of your tool to show relatively small differences in competency, whereas that may well be what one would be looking for in assessing trainees.
- as the discriminative features of the global rating and checklist are per se different, a paragraph discussing the instrument’s ability to discriminate between individuals has been added to the discussion section.
Reviewer 2 – Jennifer Mary Weller

Major compulsory revisions

1. The aim of the study is vague. Please clarify just what sort of evidence was being sought in the study to answer the question posed
   - we agree that the aim could be more accurate and have made a clearer statement of objectives and aim of the study including what sort of evidence we were looking for.

2. It’s not possible to understand the ratings without seeing the measurement tool. Please provide this.
   - The assessment tool has been included in the manuscript to clarify what we have done in the study.

3. Please describe the demographics of the participants in more detail
   - we have added information on how many TTEs participants have done and what training they had received.

The data suggest that some of the novices outperformed some of the cardiology residents and did as well as some of the consultants. A discussion of this finding should be included in the discussion section
   - we have included a discussion of the possibility of selection bias as well as the possibility of the assessment instrument not being able to discriminate between individuals.

4. Please provide a sample size calculation for both studies
   - in order to calculate sample size for the study we needed to be able to estimate the smallest relevant difference between two test scores and the variance of the scores. However, as we were working with a new assessment tool which had never been tested before, we did not have this information. Instead we turned to the literature where we found that the tradition was a sample size of 5-20 participants in each group. This has been explained in the manuscript. This topic was also discussed with a professional statistician at the university.

5. Randomisation and blinding for the scoring of the TTE images is not described
   - Each participant was assigned a ten-digit identification number unknown to the investigators. The participants typed the identification number as patient identity on the echo machine and covered the number on the screen by a post-it so that the identity of the participants remained anonymous to the investigators.
   Also, the observers who did the scoring of the TTE images were not present during image acquisition and did not know whether the images scored were performed by a novice or an expert. Hence, the scoring was performed both anonymously and blinded. This has now been described more clearly in the manuscript-

6. Study 2 is seriously underpowered, produces little in the way of convincing results, and includes only two abnormal cases. It could be mentioned as an area for future research in the discussion, but should be removed from the main manuscript.
   - As reviewer 1 has not suggested sub-study 2 removed, we have decided to keep it in the manuscript as a pilot study and have toned down the importance of this part of the study.
7. Please explain why only one rater was used
- Due to the extensive time consumption needed for scoring all 45 examinations, we decided to let just one of the observers rate all examinations and then by random pick 10 examinations (3 novices, 3 intermediates, 3 experts, 1 unknown) to be evaluated by the same observer again and also to be evaluated by another observer. This has been explained in the manuscript and mentioned as a limitation of the study.

8. The authors refer to a paper (ref. 8) that they report to be proven reliable for echocardiography. How does this study add to those results? And how does this fit with their statement that they don't know of any other instrument for ultrasound that has been validated? This need to be better explained.
- The previous study on TTE technical skills was performed as an OSCE with several short stations in a standard setting. In this study we aimed to develop an assessment tool where the procedure for use in a daily clinical setting. To our knowledge no such instrument for workplace-based assessment has been developed before. The difference between the previous study and our study has been emphasized more in the manuscript.

9. The limitations section refers only to areas of future research that were not part of the research question... ...The limitations of the study need to be mentioned here 3.g. low sample size, only one rater where ICC was subsequently found to be low for the global score, potential for bias with the rater potentially knowing the level of experience or that the three images were from the same participant.
- The limitation section has been rewritten to include potential bias including low sample size, one case, and one rater.

10. The abstract methodology differs from that described in the body of the manuscript
- unfortunately, this error occurred in the last revision of the manuscript before submission, we sincerely apologies! The abstract and methodology of the main manuscript has been better aligned.

It appears that the vast majority were rated only by one rater, the method of blinding and anonymising data is not described, and it appears for the the small sample of images that were rated by a second rater, this was in fact the researcher who had been present during the tests for the novices, which would put blinding into question. This needs to be clarified.
- please see answers to reviewer comment 5 and 7.

Minor essential revisions

Background
11. The introduction is rather repetitive and needs to be more focused.
- the introduction has been rewritten to be more focused

12. The reference to the newest consensus standards on validity is 10 years old.
- the reviewer’s suggestion of a new reference has been met
13. The last 2 paragraphs seem to include the aim and the approaches but it's unclear...  
- the last two paragraphs have been reformulated to make clear what is the aim and that the different approaches to validity has been addressed in the study

Methods
14. Page 7 – when reporting on the different groups, the data on years of experience should go in the results  
- this information has been removed to the results section

15. What is the difference between mouth-to-mouth and face-to-face? Consider not using these terms.  
- the description of how participants were recruited has been rephrased

16. What is IRB – please spell out  
- IRB is the ethical board. This has been spelled out

17. It seems that one of the researchers who subsequently scored the images was actually present when novices were being tested.  
- DGN who was present at the scanning did not participate in scoring the participants, only in the overall discussion of how to use the scale. However, the identification number of the participants was also unknown to DGN.

18. I suggest the full checklist I provided. What was the maximum score possible?  
- the checklist and global rating scale has been provided as an appendix to the manuscript. The total score of 440 has been stated clearly in the text.

Results
19. There appears to be overlap of outcome measures between the three groups. Could 95% CI be reported? It would be interesting to know if there were significant differences between the three groups.  
- as we had to use non-parametric statistic analyses, unfortunately CI cannot be calculated. Calculations are presented that there is significant differences between the three levels.

20. Sub-study 2  
- please see answer to reviewer comment No. 6

21. Page 12 – a sentence is repeated  
- this has been corrected

22. Inter- and intraobserver reliability is not good for the global score. This suggests more raters were required to generate reliable scores on which to make judgements of validity.  
- the issue about reliability of global rating scores is extensively discussed in the discussion section.

23. Results on sub study 2  
- it has been emphasized that this study is indeed a pilot study
Discussion

24. The first few paragraphs are more or less a repeat of the introduction.
- the beginning of the discussion has been rewritten to include the findings of our study.

25. The discussion on score for global rating should be revised
- the scores on the global rating are discussed based on the validity study alone and only the pilot study is included in a perspectives role.

26. Strong correlations suggest that the global rating and checklist are measuring similar constructs, but not the same.
- this has been revised

27. Whether or not a level 0.6 is acceptable depends on the purpose of the assessment. Please modify.
- the ICC values are now described as moderate

28. The difference between the ICC for the global rating and checklist score may be due to a number of things – I’m not sure if the proposal that it could be due to the scale is reasonable. It could be tested by collapsing the scale for the checklist to match the global rating scale.
- this has been done and shows that the ICC of the checklist decreases to levels close to ICC of the global rating scale which supports the proposal that part of the explanation for the low ICC of global rating is because of the scale itself

It could be that global rating scales are more reliable when the rater has a set of criterion against which to make a judgement.
- the observers did rate the TTE examinations based on predefined criteria. This has been emphasized in methodology section of the manuscript

29. They have only two difficult cases and very few participants. I don’t think anything can be said about the validity of this instrument except for rating normal patient images at this stage, but they have collected some useful pilot data to do an interesting and definite study.
- we have made clear in the manuscript that we only collected validity based on the normal case and that the pilot study alone provided us with information for further studies.

30. The acceptability and feasibility of the instrument were not part of the research question, so I would not put them as limitations.
- acceptability and feasibility issues have been withdrawn from the manuscript

Conclusion

31. The conclusion suggest a valid instrument has been developed. This isn’t really what has been done. They have developed a measurement tool in a way that addresses content validity and they have produced some reasonable evidence to support criterion validity in normal TTE images.
- the conclusion has been revised based on the comments from reviewer 2

Abstract

32. Please see answer to reviewer comment No. 10
Title
33. This strikes me as a validation study, not a feasibility study. I would change the title.
- The title has been changed according to reviewer 2’s suggestion

Ethical issues
34. The authors stated that the ethical board felt the study didn’t need formal approval. However, there are some ethical issues that should be addressed, mainly about confidentiality of the results of the ratings. I would like to know how this data was managed to protect these individuals.
- as described earlier, all data was handled anonymously for all investigators.