Reviewer's report

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

Version: 1 Date: 18 November 2012

Reviewer: Jennifer Mary Weller

Reviewer's report:

This manuscript describes the development of a rating instrument for echocardiographic skills. The rationale for this is well described and in general the approach to establishing validity (which could be considered content validity and criterion validity) is reasonable. The results provide very useful information about the validity of the instrument when used to assess a normal patient. This is a useful starting point.

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.

Methods: The methodology is underspecified.

2. It’s not possible to understand the ratings without seeing the measurement tool. Please provide this.

3. As the authors point out, time in job may not be a predictor of TTE expertise. Please describe the demographics of the participants in more detail – e.g. how many TTE’s they had done, 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. This could be explained by their background – perhaps one had been an ultrasonographer in a past life. A discussion of this finding should be included in the discussion section.

4. Please provide a sample size calculation for both studies.

5. Randomisation and blinding for the scoring of the TTE images is not described, (though mentioned in the abstract) allowing for potential bias if the rater knew the participant, and randomisation process for rating the TTEs is not described. For example if the rater scored three TTEs from the same participant at the same time, they would expect them to be at the same level. Please describe how the rater was blinded to the identity of the person who did the TTE and how the TTE image order was presented to the rater.

6. Study 2 is seriously underpowered, produces little in the way of convincing results, and includes only two abnormal cases. I would consider this a pilot study. It could be mentioned as an area for future research in the discussion but should be removed from the main manuscript.

7. Only one rater was used, and with the fair to middling intra and inter rater reliability for global ratings this puts some doubt on the reliability of the scores.
Please explain why only one rater was used.

8. The authors refer to a paper (reference 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 needs to be better explained.

9. The limitations section refers only to areas of future research that were not part of the research question – i.e. a discussion on the utility of an assessment. The research question was about validity. The limitations of the study need to be mentioned here e.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 three images were from the same participant.

10. The abstract methodology differs from that described in the body of the manuscript; in particular, it claims that two experts blindly and anonymously rated all TTE images. 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 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.

Minor essential revisions

Background

11. The introduction is rather repetitive and needs to be more focussed.

12. The reference to the newest consensus standards on validity is 10 years old. Discussions on definitions of validity continue to be published though there is probably agreement that validity should be considered a unified concept that incorporates a range of different perspectives.


13. The last 2 paragraphs seem to include the aim and the approaches but it’s unclear if the final paragraph, where different approaches to validity are described, is describing what was addressed in the study or just general approaches. This should be clarified.

Methods

See comments under major essential revisions.

14. Page 7 – when reporting on the different groups, the data on years of experience of those recruited should go in the results.

15. What is the difference between mouth-to-mouth and face-to-face? Consider not using these terms. The first sounds like a component of basic life support and the latter is unclear – did you just happen to run into them or were they sought out, asked to volunteer at meetings etc?

16. What is IRB – please spell out.
17. Page 8. It seems that one of the researchers who subsequently scored the images was actually present when the novices were being tested and provided prompts. Please explain how this was controlled and how it wasn’t a source of bias.

18. I suggest the full checklist is provided in order for readers to understand what was done. It’s difficult to understand how the score was generated. What was the maximum score possible?

Results

19. There appears to be overlap of outcome measures between the three groups. Could 95% confidence intervals be reported? While level of expertise is likely to correlate with score, it would be interesting to know if there were significant differences between the three groups.

20. Sub study 2 – see above comments. Suggest this is removed.

21. Page 12 – a sentence is repeated.

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. An instrument can be reliable and not be valid, but an instrument cannot be valid if it isn’t reliable.

23. Results on sub study 2 – see above. While it is clearly important to see if this instrument is also valid in the context of disease, this study protocol is not designed to answer the question with only 2 different pathologies and 9 participants.

Discussion

24. The first few paragraphs are more or less a repeat of the introduction. I suggest the main results are summarised in the first paragraph, then they should be discussed in conjunction with the relevant literature.

25. The discussion on scores for global ratings could be interesting but one problem was they seemed less reliable than the checklist. Some parts of the discussion referring to study 2 treat non-significant results as if they were significant. If the result was not significant, not much more can be said, except perhaps there may be a trend and further work is required. Please revise.

26. Page 15. Strong correlations between global rating and checklist scores suggest they are measuring similar constructs but not the same construct, in which case the correlation would be 1. Please modify.

27. Whether or not a level of 0.6 is ‘acceptable’ depends on the purpose of the assessment. I suspect that for a validation study a higher level of agreement would be desirable. Please modify.

28. Page 16. 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, but it could be tested by collapsing the scale for the checklist to match the global rating scale. It could be that global rating scales are more reliable when the rater has a set of criterion against which to
make a judgement, which was not the case here. Can you provide this analysis to justify your claim?

29. The authors identify the issue of context specificity. In fact 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 the authors have collected some useful pilot data to do an interesting and definitive study looking at scores across a representative range of pathologies.

30. The acceptability and feasibility of the instrument were not part of the research question so I would not put them as limitations. As said previously, there are in fact some real limitations to the methodology re sampling, sample size, bias, etc which need to go here. Reference could be made on acceptability and feasibility elsewhere. If sufficient robust data has been collected on feasibility this should go in the aim, the methods and the results before appearing in the discussion.

Conclusion

31. The conclusions 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. This looks like a good starting point but I think they have over claimed the results.

Abstract.

32. This currently confuses study 1 and study 2 data and doesn’t explain how intra-rater reliability was determined. There are differences in the methodology reported. Firstly it states here that raters were blind and they rated the images anonymously (which is unclear – were the raters anonymous or was participant identity concealed from them) but blinding, anonymity are not mentioned in the methods. Also it states that two raters rated all the TTE images. In the main body of the manuscript it appears only a very small sample of the images was rated by more than one rater. This needs to be clarified.

Title

33. This strikes me as a validation study, not a feasibility study. I would change the title.

Ethical issues.

34. The authors state that the IRB (presumably an ethics board) felt the study didn’t need formal approval. However there are some ethical issues that should be addressed in the manuscript, mainly about confidentiality of the results of the rating. Some participant scores were very poor in comparison with their peer group. I would like to know how this data was managed to protect these individuals.

Level of interest: An article of importance in its field
Quality of written English: Acceptable

Statistical review: Yes, but I do not feel adequately qualified to assess the statistics.

Declaration of competing interests:
I declare that I have no competing interests