Author's response to reviews

Title: Evaluation of the Edinburgh Post Natal Depression Scale using Rasch analysis

Authors:

Julie F Pallant (jpallant@swin.edu.au)
Renee L Miller (r.miller@bigpond.net.au)
Alan Tennant (alantennant@compuserve.com)

Version: 2 Date: 3 April 2006

Author's response to reviews: see over
Dear Biomed,

We thank the reviewers for their time and efforts with this paper. Below are the detailed replies to each issue raised.

Kind Regards
Alan Tennant

Reviewer 1.

1. From a Rasch analysis point-of-view the study is performed technically correct. However, it is important that the authors indicate much more clearly that the Rasch analysis is a validation procedure, e.g. in the measurement process, to avoid item bias, but the Rasch analysis is not a test of clinical validity.

Yes, agreed. We have emphasised this in the discussion.

2. The Edinburgh Post-Natal Depression Scale has clinical validity for depression, but the validation analysis by means of the Rasch model has identified item 7 as being biased when measuring the dimension of depressive states. It is, however, difficult to understand why item 8 (I have felt sad and miserable) also was found problematic. This is counterintuitive. The authors should treat this problem more clearly in the discussion section.

Agreed, we have added material in the discussion to explain why the ‘felt sad and miserable’ item may have misfit to the model, despite its apparent face validity for the construct

Reviewer 2

This study concludes that a revised 8-item version of the Edinburgh Postnatal Depression Scale would be psychometrically more robust. To shorten a scale while improving its psychometrics properties is useful. However, I would have prefer a more convincing demonstration.

1. This article is based on Item Response Theory (IRT) which is a modern test theory. It is difficult to evaluate the relevance of the methodology given that accepted procedure constantly changes. There is no consensus among psychometricians.

There is little consensus, as the reviewer, suggests, within the IRT community. The principal divide seems to be between those who advocate the Rasch model as a measurement model which satisfies the axioms of additive conjoint measurement, and other IRT models, which fail to satisfy those axioms. However, we do not feel that this should be brought up in the article as it would only tend to confuse readers, most of whom will not be familiar with any IRT procedure.

2. In the present study, the sample size can be considered as adequate. IRT typically requires large sample size. However, the IRT literature usually don’t recommend sample size as there are so many dependencies. Some experts suggest at least 250 people or more per group or greater. If you use the Rasch model, as in the present study, smaller samples size can be used.
Certainly IRT models like the 2-PL model require much larger sample sizes
than the Rasch model, and there are other issues such as convergence which
are problematic in other IRT models. For Rasch analysis sample size is
primarily about the precision of the parameter estimates. We have added a
short paragraph which shows the sample size requirement for a given level of
precision.

3. I suggest the authors mention the reservations about IRT and Rasch model
expressed by many psychometricians.

This is difficult to do without extending the discussion to a considerable
length. For example, there are considerable problems with classical
approaches using factor analytic techniques on ordinal data. Thus there are
reservations about classical approaches as well, and we feel it is not the task
of this article to address these issues.

4. The reason why the misfitting item 5 "I have felt scared or panicky for no very good
reason" was not deleted is not clear to me. To delete item 8 "I have felt sad or
miserable" seems to me to reduce the face validity of the scale.

We have made this clearer in the text. Removing item 8 resolved the misfit
issue, and thus item 5 was retained.

At first sight this may seem a strange omission, in that the item appears to
have face validity. The high negative residual misfit indicates is that it adds
nothing to the information gained by other items, and its removal significantly
improved the fit of those remaining items. In some respects it is more like a
summary item with a high item-total correlation (0.8).

5. I think it is not enough to show that revised cut points maintain the case
identification of the original scale. It has to be demonstrated that the revised scale
performed better than the original scale in the identification of clinical cases using a
standardized diagnostic interview.

This is the matter of further work and we have included this in the discussion.
As for the original scale, the summed score is invalid, and thus the
identification of caseness for depression is confounded.