# REVIEWER A

## 1 General Comment

This paper is about a Rasch model for binary data when subjects decide not to answer after having seen the item(s). One may expect that this decision depends on ability giving rise to non-ignorable missingness. The subject of this paper is highly relevant for practice and I would welcome it's publication in *Psicologica*. However, the paper is not yet in a form suitable for Psicologica. If it would be published it would not have the impact it deserves. Thus, I recommend that it is improved and then re-submitted to be reviewed by the same reviewers.

The main issues are:

1. In its current state it is difficult to parse.

2. Conditional estimation is not treated in sufficient detail.

First, the problem is not clear enough. It is necessary to explain briefly why the answering process leads to non-ignorable missingness and discuss some of the approaches taken by earlier researchers in somewhat more detail.

I also believe that the model should be explained in more detail. The current presentation doesn't provide any conceptual argument. The compar- ison to existing models is too succinct and contains no (new or old) insights. The model is clearly related to the multi-dimensional Rasch model and the authors might want to have a look at the overview article by Rost (Chapter 2 in Essays on Item Response Theory Lecture Notes in Statistics, 2001, Volume 157, Edited by Anne Boomsma and others). My feeling is that when the response is coded as tri-chotome, the model presented is equivalent to a mul- tidimensional Rasch model presented by Stegelmann (1983 Psychometrika , 48(2), 259-267).

Second, I suggest the authors focus on CML and provide more than just the conditional likelihood. For MML a reference to CONQUEST will suffice: i.e., the current text can be kept. Special attention should be paid to the computation of elementary symmetric functions via recursion without which conditional estimation is not possible in practice. (Luckely, this problem is solved: unlike the issues of numerical integration that remain for MML.) Rather than Mathematica, I suggest the authors include a small R-script using the sum-algorithm or, better still, manage to trick an existing program to do the work. They might have a look at Rost and Carstensen (2002; Applied Psychological Measurement March 2002 vol. 26 no. 1 42-56) and check the MULTIRA software written by CH Carstensen.

Finally, what I miss is a (small) simulation showing parameter recovery and quality of asymptotic standard errors.

I wish to point out that *Psicologica* offers the possibility to publish articles that are somewhat longer and somewhat more expository than would be possible in Journals like *Psychometrika*. I suggest that you make use of this. Should the paper become too long it can be split into two: A paper about the problem of missingness and the RR-model and a second paper about the estimation of the model.

## 2 Small Comments

Page 2: Revise last phrase.

Page 5: Revise 1) first phrase after equation 1, 2) Phrase beginning with "In their taxonomy...". The vector notation introduced in equation 3 seems superfluous here.

On page 9: $\gamma$ are elementary symmetric functions. What is d$*$?

Timo Bechger
CITO (The Netherlands)

## REVIEWER B

The manuscript proposes an appealing two-dimensional Rash model for non-ignorable missing responses. It is simple, has the advantages of Rasch models and can be fitted with standard software when there is a concern about the absence of responses. However the manuscript does not make clear what the main contribution is, as compared to existing IRT models that also allow to model non-ignorable missing data. In addition I think information should be presented in a more clear and systematic way in order to improve the readability of the manuscript. These main concerns and other problems are listed below.

1. The types of nonignorable missing responses that can be of interest when applying IRT should be made explicit from the beginning (e.g. because of limited testing time, because the wording of the item poses difficulty for respondents, because there is a lack of information about the statement, because of lack of motivation, embarrament to provide an answer, etc), differentiating between ability tests (maximum performance) and attitude, personality, etc (typical performance). For example, from the introduction it is not clear if the model applicable for ability items or not. In addition, the manuscript seems to focus on one specific possible cause of missing data (the ones "originated from a respondent's choice"). The reasons to focus on this type of non-random missing data, as well as more examples of this type of missing data, should be provided.

2. It is not clear to me if, for the missing data under evaluation in the manuscript, the probability of the missing data depends on the ability level (as you are suggesting on page 3) or not (see Holman & Glas, 2005, p.2). Please provide clear information on this regard. It seems that the consequences of ignoring the non-randomness of missing data when using traditional estimation methods depend on whether the probability of missing data depends on the ability level measured by the test or not. For example, it has been shown that omitted responses can be ignored when using MML estimation methods if the probability of missing data does not depend on the ability level measured by the test (Bock & Aitkin, 1981).

3. Since different IRT models for dealing with nonignorable missing data have been previously proposed, the differences and possible advantages of the new proposed model should be strengthened from the beginning to clearly show the contributions of the manuscript. For example Holman and Glas (2005) and Rose, Von Davier & Xu

(2010) proposed and evaluated models which are generalizations of the Rash model for the analysis of nonignorable missing data. In fact some of these models can be fitted to polytomous items, so they are more suitable for the type of items analyzed in the example of the manuscript. If one of the main contributions of the present study is that conditional maximum likelihood estimation procedures are used, this should be clearly stated from the beginning. When choosing among possible existing models, apart from some theoretical or practical reasons, the key question is how well each model represents the data, so I encourage the authors to prove the adequacy of RRM and to compare the results with those obtained with previous IRT models for non-ignorable missing data by using simulated data (apart from keeping the empirical example).

4. Please provide additional clear information about the reparameterization of equation 2 and the importance of considering this reparameterization and the 2PL model if the discrimination parameters are not modeled at the end. Also when reformulating equation 2 into equation 4, it is not completely clear the specific change made on the left hand side of the first formula of equation 4 and the differences in the first step of parameterizations (probability of "saying no" or "disagreeing" rather than not answering the item). Please provide detailed information to improve readability.

5. Because chi-square indices are not valid with sparse data, it would be interesting that, in order to test fit, apart from using Andersen's test (for CML) and MNSQ (for MML), the information was complemented by bootstrap methods applicable in both cases (in order to increase comparability of results)

FORMAT:

- Citations in the text do not follow APA rules (check first citations, for which all authors' names must be cited, check citations within brackets, etc)

- Table 2 does not clearly show the criteria for the formed groups (there are not different shaded areas)

- On page3 (first paragraph) when the authors write "patients with a higher proficiency level […] to impress the nurses" they quoting Holman and Glass (2005, so the page number for the quotation should be provided.

- For equation 2, check notations and sub-indices. State the meaning of d and different subindices, and include the missing bracket in a(q-d)

MINOR:

• Since the authors focus on an example on racial prejudices, it would be better to talk about item location or attractiveness (instead of item difficulty) and positive or negative responses (instead of correct answers)

• Please clarify whether there were response categories "don't know" and "not answered", as it seems from the description of the scoring system on page10, or people that did not provide an answer was categorized as "don't know"

# REVIEWER C

1. The introduction of this IRT model for Binary data with Nonignorable Nonresponses is not new as claimed by the authors in the introduction " we introduce a new IRT model – that belongs to the Rasch family of models – for the analysis of dichotomously scored items in the presence of nonignorable nonresponses, called Rasch-Rasch Model (RRM)". Since this model was already introduced as a special cases in the work of Pimentel (2005, chapter 2 Equation 2.1 and 2.6) and Glas and Pimentel (2006).

2. As mentioned by the authors, MML method constitutes the standard estimation technique for the models presented in this paper while they used CML for estimating item parameters in the joint likelihood using RRM. It is true that since all the models introduced in the paper are all Rasch models, CML is an advantage since it does not require distributional assumptions about the ability and missing data process. However, it is only true for RM, because when we consider a multidimensional IRT models, consistency of item parameter estimates may be in question especially maximizing joint likelihood.This is related to the fact the number of person's parameter grows proportional with the number of observations and in general lead to inconsistency (Neyman & Scott, 1948). Further, Kiefer and Wolfowitz (1956) have shown that MML estimates are consistent in IRT models. Hence MML method is more preferred as other have already done it.

3. The paper introduced an illustration on racial prejudices application that uses such method and then compared to the usual method of estimating item parameters which I think motivates other researchers to explore more on incorporating nonignorable noresponses in the general with the use of IRT models.

4. Lastly, the paper did not present any value (especially in the application) between the parameters of the person latent trait and the missing data process as to signifiy that the two parameters are related and nonresponses committed by the subject are indeed nonignorable.

Jonald L. Pimentel
University of Southern Mindanao (The Philippines)