**PMET-604: Optimal Bayesian Adaptive Design for Test-Item Calibration**
**Referee Report**

# General Comments

This ms. proposes a new design approach for item calibration in an adaptive test. The usual design approach is to select examinees (design points) to optimize item parameter estimation for a particular item or items. Instead the approach of this ms. is, for each examinee, to select item(s) from the set of un(der)-calibrated items in the item pool for which this examinee is an optimal design point according to the usual optimal design criteria such as D-, E- and A-optimality.

The approach has certain practical and computational advantages. Practically, the optimization cannot fail to select an item, because it is always optimizing over the remaining items to be calibrated. Computationally, the approach alternates between estimating a single examinee parameter $\theta$ at a time with many items, or a single item parameter $\eta$ at a time with many examinees. It is possible to organize the required computations so that even relatively slow methods—Bayesian MCMC is the preferred choice in the ms.—run at approximately operational speeds.

However, no evidence is provided, beyond a single simulation study at the end of the ms., that the alternating estimation procedure will converge, or converge to the right answer, over a range of data and parameter values. More broadly, although the algorithm can be applied in "any" adaptive testing setting, the authors do not discuss (beyond a negative speculation about putting all calibration items at the end of the test) possible interactions between the calibration design criteria and other (technical or practical) design features of the adaptive test, that might undermine either item calibration or examinee proficiency estimation.

For the MCMC calculations in the ms., a "Metropolis-Hastings within Gibbs" approach is used. M-H within Gibbs is usually burdened by the need to hand-tune proposal distributions to optimize each MCMC run; the authors spend a fair amount of effort automating an adaptive tuning approach, with an eye toward unsupervised use in operational adaptive testing. This works well; it might also be pointed out that in practice the general approach in the ms. could be adapted to other estimation methods without great difficulty.

The optimal design criteria here all depend on functionals of a suitable information matrix for estimating item parameters. The information matrix used is classical (likelihood-based) Fisher information, averaged "Bayesianly" over missing elements.

I see two possible problems here. First, asymptotic standard error and information functions reflect the curvature in the objective function used for estimation; classical approaches require the curvature in the likelihood and Bayesian approaches require the curvature in the posterior. Therefore I was surprised not to see information computed for the posterior. Second, the averaging process described in the ms. is very much a marginal information calculation. However, computation of marginal information usually requires a between/within calculation usually associated with the name "missing information principle", not unlike calculating a marginal variance or covari-

ance (see e.g. Tanner, *Tools for Statistical Inference*, Springer). I was surprised to see neither issue addressed in the ms.

Two simulation studies were conducted, to assess the convergence of the MCMC algorithm and to illustrate adaptive item calibration under A-optimality. The speed of the MCMC algorithm was impressive, even running in R on a medocre PC, and recovery of item parameters was good, with some very mild exceptions quite fairly noted by the authors.

I think there is a good idea here, one that is worth pursuing theoretically and practically. However, I question whether the ms. in its current form is suitable for *Psychometrika*.

Essentially, the ms. describes in great detail an algorithm that works in a single simulation study. Although much of the ms. is very sensible from a practical implementation p.o.v., no theoretical results are provided about the operating characteristics of the algorithm, nor are any larger lessons drawn about adaptive calibration (e.g. in a broader sequential design/estimation setting), in IRT or other settings.

A few miscellaneous comments follow…

## Specific Corrections and Suggestions

**p 9.** Equation (2) cannot follow from eq (1). Perhaps a reference to Mislevy & Chang is needed instead.

**p 9.** I find it awkward to use the same name $g$ with different index subscript names ($i$ and $j$) to represent different distributions. It makes it difficult to have simple notation when referring to the posterior density of several items, or the prior density of several test-takers, for example.

**p 15.** As mentioned in the general comments above, I'm not convinced that (a) we don't really want posterior information rather than Fisher information; and (b) this is the correct way to marginalize (without applying the missing information principle).

**p 24, para before Fig 6.** It seems like the same explanation applies to both items 3 and 4. So I didn't understand what you meant by not finding an explanation "for the slightly aberrant behavior of the former".