Reviewer’s report

Title: Development of burnout over time and the causal order of the three dimensions of burnout among male and female GPs. A three wave panel study.

Version: 2 Date: 7 January 2011

Reviewer: Jeremy Miles

Reviewer’s report:

This is an interesting paper that should be applauded for using a longitudinal sample over a long period of time, in a difficult population to observe and using sophisticated analysis. I have some suggestions to improve the paper.

Major Compulsory Revisions

I am not sure what is meant by “we did not randomize with respect to sex” (page 15). The authors did not randomize. Do they mean that the did not over or under sample with respect to sex?

The only mention of missing data in the paper is in a footnote to the table. How did the authors handle missing data in the SEM? LISREL calculates the full information maximum likelihood correlation matrix, so why wasn’t this presented? If, as the authors state on page 19, the dropout rate was higher for one gender than another, this means that the MAR (missing at random) assumption has been violated, and therefore listwise deletion is inappropriate is is likely to lead to biased results. If the data are MAR then full information maximum likelihood is appropriate as an estimator.

The SRMR is usually suggested as an appropriate measure of model fit (following Hu and Bentler’s seminal 1999 paper).

The authors write that competing models were compared with chi-square difference tests, but the AIC is described as being used to assess fit earlier - the only purpose of the AIC is to compare competing models.

It is not clear to me what “We did not specify the measurement models of the three burnout measures” means.

The sample size is small, and especially so when the sample is split by gender. Why not combine the samples and carry out multiple group tests to examine whether the parameters are different across genders.

The authors find stationarity not to hold – however, they did not explicitly test stationarity. Is there any reason that effects from time 1 to time 2 should be different to effects from time 2 to time 3?

Minor Essential Revisions
The estimator is not described (maximum likelihood? Robust maximum likelihood?)

In table 2, the authors present the F statistic to 2 decimal places of accuracy. The F statistic is essentially meaningless, it exists to give us the p-value. But they present the p-value only as < or > 0.05. It would be better to present the p-value, rather than the F statistic. In addition, an effect size would be useful.

**Level of interest:** An article whose findings are important to those with closely related research interests

**Quality of written English:** Acceptable

**Statistical review:** Yes, and I have assessed the statistics in my report.

**Declaration of competing interests:**

I declare that I have no competing interests’