Reviewer's report

**Title:** Pitfalls in the statistical examination and interpretation of the correspondence between physician and patient satisfaction ratings and their relevance for shared decision making research

**Version:** 2  **Date:** 1 April 2011

**Reviewer:** Peter T Sawicki

**Reviewer's report:**

The authors have revised the manuscript and improved it. They removed the control group. But still the question remains why they have included a control group in the first place at all and what the prospective aim of this study was. Perhaps the original study protocol could resolve this question? If the authors changed the study objectives during the course of the study or after finishing it, they should say so and give reasons for this. The authors have described methodological difficulties in calculating the correspondence between patient and physician satisfaction ratings. They conclude from their results that only the Bland-Altman method for assessing agreement augmented by bar charts of differences was able to capture a clinically relevant appreciation of the physician-patient relationship. It remains unclear to us what is the rationale for this statement. What was the “gold standard” to capture the clinically relevant appreciation of the physician-patient relationship with which the results of the Bland-Altman method and the other methods were compared?

We also still have some statistical objections regarding methods and results: While „n“ is intuitively clear, „Z“ needs a specification in the definition of the effect size for the Wilcoxon test. We are aware of the ongoing debate about using Likert scales as metric scales and accept the view of the authors. But the most striking argument for using a Likert scale as a metric scale is the robustness of a parametric method. The authors should give a more detailed discussion of the robustness of the “Bland-Altman method” which justifies the hypothesis that it is robust against violation of parametric assumptions (e.g. a theoretical statistical examination if available). In the “Methods” section the authors write that Wirtz and Caspar recommend for the marginal homogeneity test an # level of 0.25. It is unclear why they write this because in the heading of table 2 (results of the marginal homogeneity test) # is 0.05. We would recommend using #=0.05 in every calculation and not switch between different levels within one publication.

**Level of interest:** An article of importance in its field

**Quality of written English:** Acceptable

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

We declare that we have no competing interests.