Reviewer’s report

Title: Re-evaluating a vision-related quality of life questionnaire with item response theory (IRT) and differential item functioning (DIF) analyses.

Version: 1 Date: 1 February 2011

Reviewer: Elizabeth Hahn

Reviewer’s report:

This manuscript addresses two important psychometric issues in the measurement of self-reported quality of life. Specifically, the manuscript explores the use of modern techniques (item response theory) for calibrating item responses, and the extent to which a multi-item questionnaire provides an unbiased measure across various subgroups. The manuscript is generally well-written and item non-response is very low (< 5%). However, the sample sizes are very small to be used with the IRT and DIF analyses that were conducted.

Major Compulsory Revisions:

One of the most important goals of modern psychometric testing is to accurately quantify respondents’ abilities on the latent trait (e.g., vision-related quality of life). In item response theory (IRT), accurate ability estimation is contingent on the proper identification of item characteristics. Errors in item parameter estimates result in inaccurate or unstable ability estimation (Hambleton, Jones, & Rogers, 1993). Thus, obtaining accurate item parameter estimates is a critical prerequisite in obtaining high-quality ability estimates. Sample size is one of the most important factors that affect the item calibration task (Hambleton & Jones, 1994). The minimum acceptable sample size for the graded response model is approximately 250, although many researchers recommend sample sizes of 500 (e.g., Baker, 1998) or even 1,000. The study described in this manuscript is based on a total sample size of 296, which may be sufficient for item calibration of the total group. However, the DIF analyses are based on small subsets of respondents (some as small as n=76). This means that the item calibrations for the subsets are not very stable for purposes of comparing them across groups. For example, this can easily be seen in Table 2 where the item threshold parameters exhibit a very large difference, e.g., 0.24 to 3.37 for thresholds 4 and 5 on the “Vision in general” item for females. This is likely due to very small sample sizes in the various response categories.

The Background section is very useful and provides good information. It would be useful to summarize which studies were conducted with the English version of the questionnaire and which were conducted with the Dutch translation.

It is unclear whether the authors’ previous studies of the LVQOL involved IRT analyses. Page 5 states that IRT was used, but that “a calibration-process was
not performed.” Page 8 mentions removal of two items based on previous analyses. This is not clear to the reader.

More details are needed about the factor analysis that was conducted for this study.

Page 10 mentions a very important issue in DIF; specifically, the importance of having subgroups that are comparable in other characteristics. Did the authors evaluate the characteristics of the various subgroups, to determine how similar they were?

The calculation of differences in expected scores at various levels of ability is an interesting way to summarize some of the DIF results. However, no reference citation was given for this technique, and it may not be the best way to summarize such analyses. In addition, some information is needed about why a 1-point difference is considered to be a large difference.

Many readers will be confused by use of terms such as “easier” (p.13) to describe the DIF results. In addition, it is difficult to interpret some of the results since there is no discussion about the direction of scoring for the LVQOL, e.g., is a high score considered to be a favorable score (better quality of life)?

Minor Essential Revisions:

Page 6 states that “variables that potentially lead to DIF are demographic variables,” yet the paragraph continues with a discussion of certain eye conditions as the focus of DIF analyses. “Demographic variables” are usually considered to be age, gender, etc., rather than health conditions.

Information about the Dutch translation and validation of the LVQOL should be added to the Methods section.

**Level of interest:** An article of insufficient interest to warrant publication in a scientific/medical journal

**Quality of written English:** Needs some language corrections before being published

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

**Declaration of competing interests:**

I declare that I have no competing interests.