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: 6 December 2010

Reviewer: Robert W Massof

Reviewer’s report:

Dr. van Nispen and her colleagues report the results of a sophisticated and expert psychometric analysis of a legacy vision-related quality of life questionnaire, the LVQOL. Like many other existing quality of life and visual function rating scale questionnaires, the LVQOL was developed and evaluated using traditional theory-free methods that were built on the assumption that sums of item response rank scores can be accepted at face value to be measures of the latent variable of interest. Axiomatic measurement theory says that sums of item response rank scores must be monotonic with the measure, if a valid measure can be estimated from the observed responses, however the sums of rank scores have a nonlinear, instrument-specific, and sample-specific relationship to the measure (see Massof RW, Likert and Guttman scaling of visual function rating scale questionnaires. Ophthal Epidemiol. 2004, 11:381-399). The estimation of measures from rating scale responses to items requires a model of the assumed relationship between the observations and the latent variable to be measured. There are a number of models that have been used for this purpose, which collectively are called “item response theory”, or IRT, models. Although different models generate similar estimates of measures of latent variables when valid measures can be estimated from the observations (see Massof RW, Application of stochastic measurement models to visual function rating scale questionnaires. Ophthal Epidemiol. 2005; 12:103-124), the mathematical differences between models are indicative of differences between models in their implicit assumptions.

The authors of this paper chose to use Samejima’s graded response model for their analyses, but they do not defend their choice. The three distinctive assumptions of the Samejima model are: 1) there is no variability between people in response category criteria, 2) response category criteria can vary across items, and 3) the variance of the error distribution can vary across items. These assumptions are arbitrary. The first assumption identifies the Samejima model as equivalent to a proportional odds regression model (McCullagh P. Regression models for ordinal data. J Roy Stat Soc Ser B, 1980;42:109-142.), so presumably it shares many of the same theoretical foundations. The second assumption essentially serves to provide the analyst with a large number of free bias parameters that, along with the third assumption, which provides the analyst with free variance parameters, can serve to improve the fit of the model to the data. Although not overdetermined for most data sets, the principle of parsimony has no part to play in constraining the Samejima model.
An alternative to the Samejima model is the Muraki rating scale model (Muraki E. Fitting a polytomous item response model to Likert-type data. Appl Psychol Measurement 1990;14:59-71.). The Muraki model is identical to the Samejima model except that response category thresholds are constrained to be the same for all items that use the same response categories. This assumption makes more sense than assuming that all respondents agree to use the same unique response category criteria for each item. But, the assumption that all respondents agree in response category criteria at all is still hard to accept. More likely, differences between respondents in their personal response criteria is the major source of random between person variability. If so, then the Andrich rating scale model (Andrich D. A rating formulation for ordered response categories. Psychometrika, 1978;43:561-573.), the Masters partial credit model (Masters GN. A Rasch model for partial credit scoring. Psychometrika 1982;47:149-174.), or the generalized partial credit model (Muraki E. A generalized partial credit model: Application of an EM algorithm. Appl Psychol Measurement. 1992;16:159-176.) would be a more appropriate choice.

Obviously, there are many more latent variable regression models from which to choose, each with its own set of implicit assumptions. However, the question to be asked in the analysis is not one of whether or not the observations can be fit with a regression model, but whether or not valid measurements can be estimated from the observations. In order to estimate valid measures, the observations must conform to the expectations of axiomatic measurement theory (Krantz DH, Luce RD, Suppes P, Tversky A. Foundations of Measurement, Vol I. New York: Academic Press, 1971.). Rasch models are constrained by axiomatic measurement theory and are used to test the hypothesis that valid measures can be estimated from a set of observations (Rasch G. On specific objectivity: An attempt at formalizing the request for generality and validity of scientific statements. Danish Yearbook of Philosophy 1977;14:58-93.). With this consideration, of the models mentioned above, the Andrich rating scale model would be the model of choice for the task defined by the authors.

An IRT vs Rasch debate has been going on for a long time in the field of educational testing and has almost become a cliché. Given the history of this issue, and given the opportunity for debate in this open review, I am not going to insist that the authors abandon their approach and substitute a Rasch model for the Samejima model. In all likelihood, the choice of model will not change the conclusions of the paper. But, from a scientific standpoint, the authors must be asked to state and defend the assumptions their choice of model implies. The authors also must be mindful of how their assumptions carry through the analyses and interpretation of the fit statistics. In particular, they need to explain what a misfit means in terms of assumptions about error distributions. Finally, the authors need to address the issue of unidimensionality of measures and what the consequences of their model assumptions are with respect to unidimensionality.

The authors’ study of DIF is laudatory. However, two parameters were not investigated that are potentially strong sources of DIF. One is whether or not the respondent had assistance with completing the questionnaire (referred to as
“proxy” in the paper, which I don’t think the authors mean) and the other is the effects of rehabilitation. Assistance with responses, at the very least, might be expected to produce non-uniform DIF. The other parameter is the effect of intervention. The legacy database has both baseline and post rehabilitation data. Do some item measures change as a consequence of rehabilitation (e.g., see Stelmack JA, Stelmack, TR, Massof RW. Measuring low vision rehabilitation outcomes with the NEI VFQ-25. Invest Ophthalmol Vis Sci 2002;43:2859-2868.)? Intervention-dependent DIF, if it exists, can create problems for using such instruments as outcome measures.

Overall, this is an excellent paper and I commend the authors on their rigorous approach. To me, the methodological issues are secondary to the scientific issues surrounding theory and the development and use of models (Massof RW. The measurement of vision disability. Optom Vis Sci. 2002;79:516-572.). I want to use the opportunity to review this fine paper as a means of opening an important debate in our field. I believe our field has moved beyond the traditional approach of applying classical test theory to visual function and quality of life questionnaires (Massof RW. Moving toward scientific measurements of quality of life. Ophthal Epidemiol. 2008;15:209-211.). But, we now have to understand our theories and models and seriously debate the science they represent.

**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:**

Non-financial competing interests:

1. I served as an outside member of Dr. van Nispen's Ph.D. committee in 2009, so the authors are well aware of the issues I raised in my review

2. Dr. van Nispen has invited me to collaborate on a project for which she is seeking research funding