The Amblyopia and Strabismus Questionnaire: Rasch Analysis

We read with interest the article by Vianya-Estopa et al., 1 “An Evaluation of the Amblyopia and Strabismus Questionnaire Using Rasch Analysis.” It criticized our Amblyopia and Strabismus Questionnaire (A&SQ) by highlighting “the limitations of the A&SQ instrument in the assessment of VR-QoL in subjects with strabismus and especially in those with amblyopia alone.” We would like to comment on the selection of subjects used in their appraisal.

The A&SQ was presented to 102 patients (mean age, 48 years) from ophthalmic, orthoptic, and optometric clinics and practices. Eleven had strabismus without amblyopia, 33 had anisometropic amblyopia, 39 had strabismic amblyopia, and 19 had combined-mechanism amblyopia. The sample seems representative of amblyopia of various causes in patients age 48. However, average visual acuity is worse than would be expected in 48-year-old U.K. citizens with amblyopia. Visual acuity of the amblyopic eye was 0.38 logMAR in anisometropic amblyopes, 0.6 logMAR in strabismic amblyopes, and 0.70 logMAR in combined-mechanism amblyopes. Accordingly, almost all had amblyopia with a visual acuity worse than 0.3 logMAR.

The prevalence of adults with a visual acuity of worse than 0.3 logMAR is approximately 2.5% in untreated populations,2–4 but 0.8% to 1.1% in treated populations.5–8 Treatment of amblyopia was well established in the United Kingdom the 1900s. Hence, the sample of patients consisted primarily of severe or insufficiently treated amblyopia, whereas there were only 11 patients with strabismus without amblyopia. The large category in the middle of the spectrum that runs from strabismus without amblyopia, through strabismus with mild or treated amblyopia, to more severe amblyopia is underrepresented in the sample of Vianya-Estopa et al.1

Although the Rasch model does not assume a population distribution from a certain class, in practice, a sample that extends over the whole range of the spectrum provides more balanced statistical information about the functioning of the different response categories. This may well be the reason that they found that merging five-point scales into three-point scales did not make the A&SQ less informative for their sample. It could also explain the ceiling effect that they found. The relatively small sample size, 102, may have aggravates the problem.

In previous studies,9–11 in an almost nonselective sample of amblyopia and/or strabismus patients treated in the Waterland area 35 years earlier, we found that the two conditions strabismus and amblyopia are heavily interwoven. We agree that, in principle, separate instruments would be preferable to assess VR-QoL in individuals with strabismus, individuals with amblyopia, individuals with both amblyopia and strabismus, and individuals with loss of binocular vision. However, this conclusion cannot be drawn from the study by Vianya et al.1 in a sample dominated by amblyopia with very low visual acuity.

Elisabeth S. Van de Graaf
Huibert J. Simonsz
Department of Ophthalmology, Erasmus Medical Center, University Medical Center Rotterdam, Rotterdam, The Netherlands.
E-mail: e.vandegraaf@erasmusmc.nl

References

Citation: Invest Ophthalmol Vis Sci. 2010;51:6898. doi:10.1167/iovs.10-5761

Rasch Analysis for QoL Questionnaires

Vianya-Estopa et al.1 have criticized the Amblyopia and Strabismus Questionnaire (A&SQ), in contrast with the very favorable...
outcome of our recent factor analysis. We would like to comment on the measurement model used in their appraisal.

In clinimetrics, test administration is seen as a probability experiment in which the essential attributes of interest, here quality of life, as well as incidental effects such as errors due to the measurement experiment itself, are thought to influence the item responses. The purpose of a measurement model is to describe this experiment and provide estimates of the latent variable of interest and the variance of its measurement errors. The authors analyzed A&SQ data with Andrich’s rating scale model which assumes that (1) the subject is making a series of consecutive choices between neighboring categories, (2) the items measure in the same scale units, and (3) a subject’s responses uniquely indicate the attribute of interest and do not depend on each other. None of these assumptions, however, seem to adequately describe the measurement experiment or have much relevance for the meaningfulness of statements based on A&SQ scores. Lack of fit of this model should, however, not be interpreted as evidence for multidimensionality but as a lack of fit to all the assumptions of the rating scale model. Therefore, the second dimension found by the authors may well have been the result of a shortcoming in the description of the measurement experiment.

In our analysis we took a pragmatic approach and used principal component analysis (PCA) to assess whether a substantial amount of the response variance can be explained by a small number of factors (preferably one), instead of detecting deviations from a restrictive one-dimensional model. PCA allows items to be measured on different interval scales, controls for correlated residuals by introducing additional factors that explain negligible amounts of response variance. We searched for effect size rather than adherence to a (virtually nonattainable) normative standard such as Rasch homogeneity. For this reason PCA is generally considered the model of choice in assessment of measurement instruments for attitudes and personality like the Big Five.

If there is reason to believe that our methods have been too crude, the usual method is to compare the fit of a one-dimensional discrete item response model against a well-fitting multidimensional model. Given the number of parameters to be estimated, multidimensional models require a large sample of subjects. We predict, however, that even when a two-dimensional model is preferred over a one-dimensional one, the correlation between the factors will be too high in most populations to be of much practical interest. If uncorrelated, the second dimension may be an effect of the measurement experiment and picks up only small factors due to artifacts such as priming, which will generally not have a systematic influence on the A&SQ score. Summarizing, we do not think the analysis of Vianya et al. underlines the practical usefulness of the A&SQ or threatens the meaningfulness of statements based on A&SQ scores.

References


Author Response: Amblyopia and Strabismus Questionnaire

We thank Drs. van de Graaf, Simonsz, Kelderman, Felius, and Passchier for their comments and interest in our paper which evaluated their Amblyopia and Strabismus Questionnaire (A&SQ).

Drs. Van de Graaf and Simonsz commented on the average visual acuity of our sample and suggested that almost all had amblyopia with a visual acuity worse than 0.30 logMAR. In fact, only 61% had worse visual acuity. The general tone of the criticisms raised by Van de Graaf and Simonsz is that our amblyopic sample is similar to many previously published investigations. For example, Attebo et al. (whom van de Graaf and Simonsz cite) found that, in an adult population, 58% had an acuity of worse than 0.60 logMAR and 90% had an acuity of worse than 0.18 logMAR. Rahi et al. found that 3% of their amblyopes at age 16 had acuity of 0.50 logMAR or worse, and in the sample in Brown et al. of ~4700 adults (mean age very similar to ours at 59 years), 54% of amblyopic eyes had VA worse than 0.30 logMAR. We are therefore strongly of the opinion that our sample is representative of adult amblyopes in developed countries.

Unfortunately, Van de Graaf and Simonsz also appear to misunderstand our findings in regard to the sensitivity of the A&SQ. They suggest that the ceiling effect that we found in our study was due to “a sample dominated by amblyopia with very low visual acuity” and a relatively small sample size. In our paper, we acknowledged the limitations extending from our modest sample size. However, the ceiling effect that we report is attributable to the finding that most of the subjects had little or no difficulty with many of the functional activities that the questionnaire addresses. If it is true that our sample was dom-