Questionnaires for hearing aid evaluation-useful tools or wasting time?

Matthias Latzel, Sabine Blab**, Marty Auer***, Heike Heuermann**

*Siemens Audiologische Technik GmbH, Erlangen
**University of Applied Science, Oldenburg
***Academy of Psychotherapy, Wiesbaden

Introduction

For the evaluation of hearing aids, several tools are available and useful. On one hand there are *objective tests* (e.g. speech tests, categorical loudness scaling…). These tests provide mostly accurate results with high reliability but are lacking because of the limited relation to the benefit in real life and individual performance. Additionally, the results are referenced to average data so that the individual perception is not taken into account sufficiently. On the other hand, *subjective tests* (e.g. questionnaires, paired comparison…) have results with direct relevance to individual benefit in real life. The disadvantages of these types of tests are inaccurate results with low reliability, potential bias from the experimenter and high effort in time and concentration for both the experimenter and the subject. Consequently, the results are hard to analyze and leave a big question mark with respect to validity.

But as subjective tests - especially questionnaires - are powerful tools to obtain impressions about how a hearing aid as a whole is perceived by the user, the optimization of the set of questionnaires is always a major issue.

Therefore the following **research questions** have been identified:

- How accurate is the *short-term* Test-Retest Reliability?
- Does the test-retest reliability depend on the "complexity" of the item?
- Do experienced listeners yield results with higher test-retest reliability than inexperienced listeners?
- Does test-retest reliability increase when reducing categories?
- Is the variance of test-retest results larger than the loss of accuracy when changing from a questionnaire with a higher number of categories to one using fewer categories?

As the expected results of the research questions depend on the type of questions (absolute versus relative), this paper concentrates on questionnaires with *absolute questions* only.

Subjects

47 normal hearing and 10 hearing impaired subjects took part in four studies whose data have been pooled in this paper. The hearing impaired subjects were experienced hearing aid users with moderate to moderately severe hearing loss.

Methods

The major task of the subjects during the study was the completion of the relevant questionnaires under controlled conditions to check the validity. For this reason 3 different hearing aid settings were defined and a commercial hearing aid was programmed accordingly:

- Setting #1: Firstfit setting based on the individual audiological data
- Setting #2: Modification of setting #2 to excite a *metallic* sound perception (Gabrielsson et al, 1990)
- Setting #3: Modification of setting #1 to excite a *dull* sound perception (Gabrielsson et al, 1990)

Setting#1 is predicted to be the most difficult as it has no predefined sound as in setting#2 and setting#3.

Three different sound samples (speech in quiet, speech in traffic noise, music) were presented for each of the three hearing aid settings. The output for each setting was recorded and replayed to the subjects via headphones.

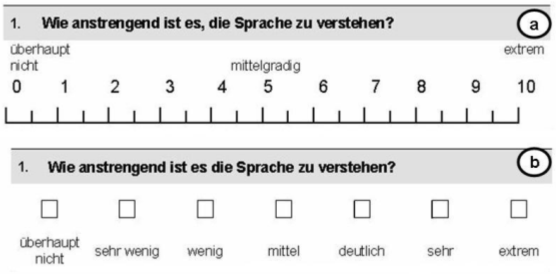
1. Wie anstrengend ist es, die Sprache zu verstehen?
a
überhaupt nicht
mittelgradig
extrem
0
1
2
3
4
5
6
7
8
9
10
1. Wie anstrengend ist es die Sprache zu verstehen?
b
überhaupt nicht
sehr wenig
wenig
mittel
deutlich
sehr
extrem

Fig.1 Illustration of the two questionnaires in test for item "sound quality".

a: questionnaire with the continuous scale, b: questionnaire with discrete scale and verbal categories

In the first session the subjects had to fill in one questionnaire with a continuous scale (CS) from 0 to 10 and one with a discrete (7-point) scale (DS) with verbal categories (see figure1). Both questionnaires contained the same items. In the second session the subjects executed the same procedure to check the long-term test-retest. But as the questionnaires with same items were conducted within one session as well, the results also provide access to the short term test-retest reliability. At the end of the study, the subjects were asked to indicate their preference for one of the questionnaires.

Results and discussion

Short term test-retest reliability:

Figure2 shows the result of the same questionnaire for all sound samples and for the item "sound quality" filled in within appr. 20 minutes. The difference of scales (use of two different questionnaires with discrete ⇔ continuous scale) has been equalized for comparison. The results are quite surprising as the same questions for the same setting lead to different results within a 20 minute period. The median for speech in quiet and noise differ by about 1.5. The median for music is even more reduced as the quartiles of conditions have no overlap implying statistical significance So, as there should be little to no change of the sound perception of the hearing aid and the condition of the subject within the 20 minute period, the only possible reason for the observed deviation has to be the inaccuracy of the questionnaire or the inaccuracy of the individual subject due to the lack of internal sound level reference.

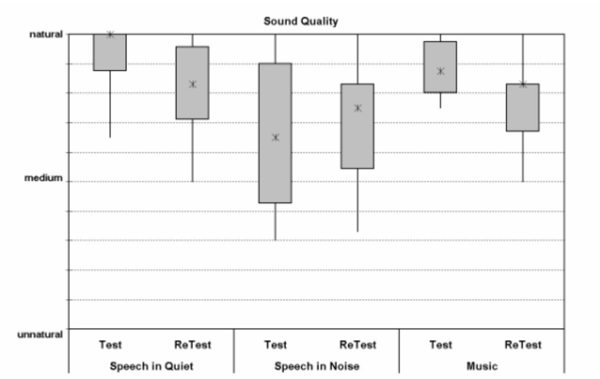

Fig.2 Inter-individual data (median, minimum, maximum, 25-quartile, 75-quartile) of ratings of setting#1 for item "sound quality" separated regarding sound samples (appr. 20 minutes later). Parameter: test time (test ⇔ retest)

Dependency on complexity of the item:

A comparison of the results in figure 2 and 3 confirm the argument that the complexity of the items of a questionnaire vary. The answering of the "sound quality" question seems to be difficult as the test-retest variability is rather high. However, the item "loudness" appears to be much easier as the median of the judgment is almost the same for all sound samples, implying stable answer patterns.

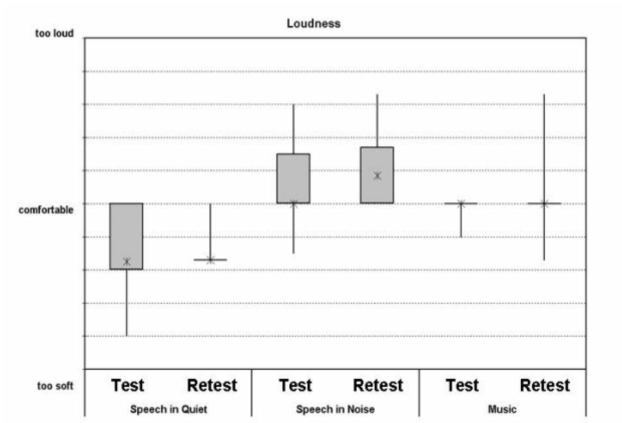

Fig.3 Inter-individual data (median, minimum, maximum, 25-quartile, 75-quartile) of ratings of setting#1 for item "loudness" separated regarding sound samples (appr. 20 minutes later). Parameter: test time (test ⇔ retest)

Dependency on the experience of the subject:

Figure4 illustrates the consequence on the results if experienced or inexperienced listeners fill in the questionnaire. In this figure relative values display the average difference between test and retest values. The difference between test and retest values for all items are smaller for experienced listeners than for the inexperienced ones which supports the hypothesis that experienced subjects provide more reliable judgments of sound perception. This is especially valid for the item "loudness" where the difference between test and retest ratings is only about one half category.

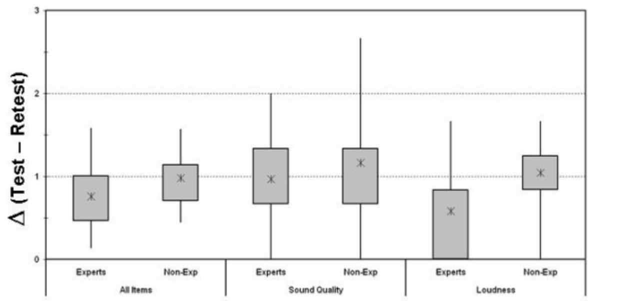

Fig.4 Inter-individual data (median, minimum, maximum, 25-quartile, 75-quartile) of ratings of setting#1 for all items and separated regarding items "loudness" and "sound quality". Visualized results are relative data with test data related to retest data. Parameter: experience of subjects

Improvement of test-retest reliability with less categories:

A comparison of the differences between test and retest results of a continuous scale and a discrete scale is shown in figure 5. If all items are pooled and averaged, the improvement using a discrete scale is about 0.5, a doubling of the reliability. Considering only the item "sound quality," the results show that the reliability could be doubled, an improvement of one scale from 2 to 1. In contrast, the *easy* item "loudness" is not sensitive for this scale because the originally high accuracy of the continuous scale could not be improved upon using a questionnaire with a discrete scale.

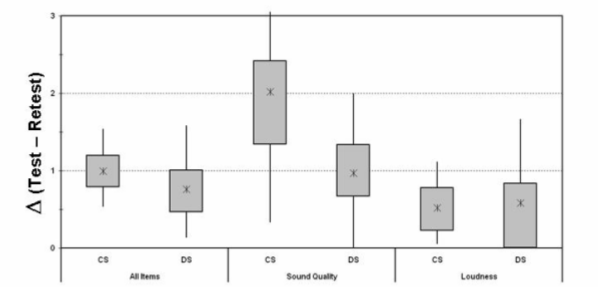

Fig.5 Inter-individual data (median, minimum, maximum, 25-quartile, 75-quartile) of ratings of setting#1 for all items and separated for items "loudness" and "sound quality". Visualized results are relative data with test data related to retest data. <u>Parameter</u>: continuous scale (CS) ⇔ discrete scale (DS)

<u>Loss of accuracy with fewer categories:</u>

The scatter plot in figure 6a illustrates a comparison of the individual results for the item "sound quality" for test and retest case pooled for all sound samples. If the test-retest reliability is optimal, all dots must lie on the bisecting line, which is not the case. The dots are distributed over the whole diagram reflecting a large variance. Reducing the number of categories from 10 to 7 (using only the discrete scale) the bisection line broadens as identified by the red area in figure 6b. If the dots were to lie within the shaded area, this would indicate that the variance is equal to/smaller than 1.7 and the reduction of the categories would reduce the accuracy of the questionnaire. However, this is not the case. A majority, 55%, of the dots fall outside of the shaded area, confirming the hypothesis that the reduction of categories does not essentially affect the accuracy!

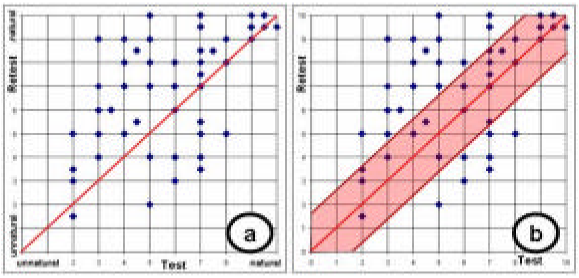

Fig.6 Scatter plot of individual data for test (x-axis) and retest (y-axis) of ratings of setting#1 for item "sound quality". a: bisection line shows area for optimal test-retest stability, b: red area symbolizes tolerance area if categories are reduced from 10 to 7.

Conclusions

The study showed that the test-retest reliability of the investigated questionnaires with a continuous scale is not sufficient. Even within a short time period, the data for some of the same tasks deviate dramatically. There are various parameters determining the test-retest variability such as experience with listening tasks, complexity of item and the number of categories. The continuous scale was found to contribute significantly to the test-retest variance, suggesting that the use of a "discrete" scale with explicit anchors for absolute ratings improves reliability of the data.

<u>But</u>: the loss of accuracy when switching to a scale with fewer categories is lower than the variance of the answers.

In a direct comparison, the subjects confirmed that completing a questionnaire with a discrete scale is an easier task as it is less challenging, less time consuming and less exhausting.

The results of the study confirm the necessity for an adequate design for questionnaires to be a useful tool for hearing aid evaluation to collect *ABSOLUTE* subjective data.

This means:

- limiting the number of categories
- careful instruction of the subjects
- awareness of the influence of experience with listening tasks
- auditory training for subjects prior to a study in order to generate homogeneous groups

In the near future, the study will be repeated using questionnaires with relative answers to to relate the test conditions between each other..

References

Gabrielsson A, Hagerman B, Bech-Kristensen T, Lundberg G (1990) Perceived sound quality of reproductions with different frequency responses and sound levels. JASA 88: 1359-1366.