Research Article

Modeling the reliability of the Freiburg monosyllabic speech test in quiet with the Poisson binomial distribution. Does the Freiburg monosyllabic speech test contain 29 words per list?

Abstract

Every speech test can be modeled as a Bernoulli experiment; this also applies to the Freiburg monosyllabic speech test. The model enables a quantitative calculation of the reliability based on the binomial distribution. Generally, the same probability is assumed for the recognition of each test word. Since the recognition of words within test lists of the Freiburg monosyllabic speech test differs, modeling with the Poisson binomial distribution is reasonable, and results in a narrower confidence interval than the simple binomial distribution. The variance of the Poisson binomial distribution for test lists of the Freiburg monosyllabic speech test with 20 words can be approximated using the variance of the simple binomial distribution based on test lists with 29 equally-recognizable words.

Keywords: Freiburg monosyllabic test, speech intelligibility, binomial distribution, reliability, confidence

Inga Holube[1,2]
Alexandra Winkler[1,2]
Ralph Nolte-Holube[1]

1 Institute of Hearing Technology and Audiology, Jade University of Applied Sciences, Oldenburg, Germany

2 Cluster of Excellence "Hearing4All", Oldenburg, Germany

Introduction

The Freiburg monosyllabic speech test (FBE) by Hahlbrock [1] is widely used in speech audiometry and hearing-aid fitting. The result of the speech test (score of correctly repeated words, in %) after performing a test list at a given level by a given person is often regarded as the actual or true value of the person's speech recognition at this level. However, every speech test is subject to a certain degree of uncertainty, so that the true value (mathematically: the expected value) of a speech-recognition score cannot be exactly determined by measuring it once with a test list [2]. Hagerman [3] suggested modeling speech tests as Bernoulli processes. In doing so, it is assumed that there is a certain probability p_{ji} for each word *i* from the list *j* that it will be correctly repeated by the listener. If the same probability p_j in % can be assumed for all words of the test list *j* at a given level, then the score for speech recognition in percent for this test list (ignoring learning effects and always assuming the same level of attention by the listeners) is subject to the standard deviation:

Equation 1

$$\sigma_j = \sqrt{\frac{p_j(100-p_j)}{n}}$$

Winkler and Holube [4] used this standard deviation to estimate the 95% confidence interval for the speech-recognition score of a single test list of the FBE. This is the interval around the true value at which 95% of the measurement results are expected. For the test lists of the FBE with 20 words, *n*=20. The width of the 95% confidence interval can be calculated directly using the binomial distribution. Alternatively, and more simply, the normal distribution can be used as an approximation. The width of the 95% confidence interval thus obtained is the standard deviation multiplied by *z*=1,96. The measurement result for a single test list would have to be outside this 95% confidence interval in order to be considered as significantly different.

In the context of an expert interview for the revised version of the guidelines for assistive devices including hearing aids in Germany, the question was raised as to how many test lists of the FBE are necessary in order to establish a distinguishable hearing ability with a probability of 95% [5]. One of the interviewed experts commented: "The approach of a binomial distribution for the Freiburg test is not easy to follow: the probability of correctly recognizing a single word depends on the degree of difficulty of each individual word and therefore cannot be set at 0.5. In the Freiburg test, 20 words with different degrees of difficulty are tested in a list." This statement is based on the everyday experience of working with the FBE, that some words are almost always – and others almost never – understood. Differences in word recognition within the test lists were also described by Hey et al. [6] for CI patients.

The assumption that all words within a list j have the probability p_j for correct recognition is thus apparently not true in the FBE. However, the number of correctly repeated words can be interpreted as a random variable subject to a Poisson binomial distribution. Hagerman [3] already used the Poisson binomial distribution to account for the variability in word recognition and to estimate its effect on reliability. He was able to show that a test list with 25 words of different recognition has the same reliability as a test list with 33 words with equal recognition. The reason for this effect is that the reliability is worst at a speech-recognition score of 50% and best at 0% and 100%. Hence, the 95% confidence interval is minimal at a speech-recognition score of 0 and 100% and maximal at 50% [4]. Thus if, for example, a word is recognized with a probability of 100%, then it is recognized again and again when it is repeated. However, if the probability is only 50%, then the word is sometimes recognized and sometimes not recognized when it is repeatedly presented. If a test list has a mean speech-recognition score of 50% and all words have the same probability of 50% of being recognized, then the 95% confidence interval for this test list is larger than if some words are well and others are not well recognized. This narrowing of the probability distribution of the number of correctly recognized words due to the word recognition variability can also be understood directly as a consequence of Equation 23 in the Appendix (Attachment 1).

In the current analysis, the single word recognition of all 400 words of FBE, which are grouped in 20 test lists, was used in two groups of participants (normal hearing: NH and hearing impaired: HI) to estimate the reliability of the test, taking into account the Poisson binomial distribution. As a measure of reliability, both the 95% confidence interval for the deviation of a measurement from the true value and the 95% confidence interval for the deviation of the true value from a given (measured) value were used.

Methods

Participants

In total, 120 individuals took part in the study. Table 1 gives an overview of the two groups of participants, who were all remunerated for their participation. The pure-tone audiograms according to DIN EN ISO 8253-1 [7] were measured with an audiometer (Siemens Unity 2) and headphones (Sennheiser HDA 200) in the frequency range from 125 Hz to 8 kHz with all octave and intermediate (average between octave) frequencies in both ears. The group NH met the criterion of normal hearing according to DIN EN ISO 8253-3 [8], i.e. the hearing threshold was at most two frequencies maximally 15 dB HL, and at all other frequencies maximally 10 dB HL. The mean hearing loss (pure tone average, PTA-4) for the frequencies 0.5, 1, 2, and 4 kHz is also included as a median for both groups of participants in Table 1. Figure 1 shows the hearing losses (mean and standard deviation) for both groups. Data from the NH group was already included in Baljic et al. [9].

Table 1: Overview of groups of participants

Group	NH	HI
Number of participants	80 (58 ♀, 22 ♂)	40 (10 ♀, 30 ♂)
Age (Median)	22 years	74 years
PTA-4	0 dB HL	45 dB HL
Presentation level	17.5, 23.5, 29.5 and 35.5 dB SPL	65, 80, 90 and 95 dB SPL

Another requirement of DIN EN ISO 8253-3 [8] for the group NH is otological normality. For this purpose, the participants were questioned orally about noise exposure in the 24 hours preceding the test, about taking ototoxic drugs, about hereditary hearing loss, and about their health status. All participants answered these questions with “no” and there were no health restrictions.

The participants of the NH group had had no exposure to the FBE. Since 23 participants of the group HI were fitted with hearing aids, it can be assumed that these participants had performed the FBE several times as part of their hearing-aid fitting process. A training effect can therefore not be excluded for the HI group.

Speech material

The Freiburg monosyllables according to DIN 45621-1 [10] and DIN 45626-1 [11] were presented monaurally via headphones (Sennheiser HAD 200). The ear with the better PTA-4 was used as the measurement ear. For the same PTA-4 for right and left, the measurements were made with the ear typically used for telephoning. The recording of 1969 [12] as a digitalization on the Siemens CD (Item No. 7970155 HH 922) was used as speech material. The presentation order of the test words corresponded to the word lists specified in DIN 45621-1 [10]. The headphone was calibrated according to DIN EN ISO 60318-1 [13], taking into account the free-field equalization for the HDA 200 [14] with the calibration signal according to *Comité Consultatif International Télégraphique et Téléphonique* (CCITT noise according to ITU-T G.227, [15]). The test words were presented by the Oldenburg Measurement Application (OMA) research version 1.5.5.0 (Hörtech gGmbH). The levels and test lists were randomized.

Word recognition

All participants heard each test list, and thus each word, only once. The test lists were presented at four different levels (see Table 1). Each participant thus listened to five test lists at each level. The assignment of the lists to the levels was different for each participant, so that per level and word, the results of 20 participants of the group NH and 10 participants of the group SH were available. Due

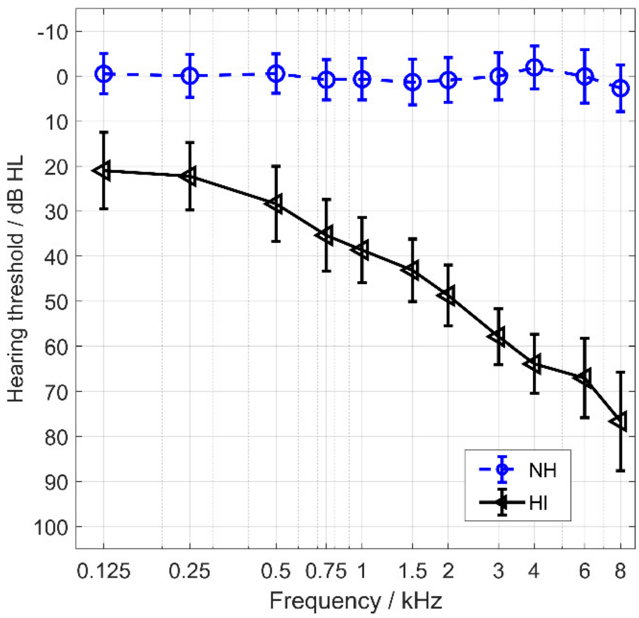

Figure 1: Pure tone thresholds (mean and standard deviation) for both groups of participants NH (blue circles) and HI (black triangles)

to data storage issues, all the test-list results were recorded, but the word-specific results were not stored for all measurements. The following data sets were used for the word-specific analysis:

- Group NH:
  - Test list 9: 77 data sets
  - Test lists 7, 11, 13, 18: 78 data sets
  - All other test lists: 79 data sets
- Group HI:
  - Test lists 3, 9, 13, 15: 39 data sets
  - All other test lists: 40 data sets

From the data sets, the word-recognition score in percent was calculated for each word and for each presentation level. Table 2 gives an example for the words “Aas” (carrion) and “Dorf” (village) of the group HI.

Table 2: Exemplary calculation of word recognition for the words “Aas” and “Dorf”

	„Aas“	„Dorf“
No. of presentations	10 at 80 dB SPL	9 at 80 dB SPL
No. of correct responses	1	7
Percentage word recongnitions	1/10 ·100%=10.0%	7/9·100%=77.8%

Results

List-specific word recognition

Figure 2 shows the participant-averaged speech-recognition results for each test list at each level for both groups of participants. The variability of the test lists for the group NH was already reported in detail in [9].

In the current contribution, the differences in word recognition within the test lists are of interest. When modeling with the Poisson binomial distribution, every word *i* in the test list *j* is assigned a recognition probability p_{ji}. Word recognition in percent can be taken as an approximation to the probability p_{ji}.

In Figure 3, for each of the 20 test lists for the group NH, the relative frequencies for percent word recognition at the four levels are shown as frequency polygons. For this purpose, the percentage word recognition was divided into classes with a width of 10% each. As a measure of the differences in word recognition, Table 3 shows the root mean square (RMS) of word recognition in percent for each of the *n*=20 test lists at each of the four levels according to:

Equation 2

$$s_{\vec{p}_j} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(p_{ji} - p_j)^2}$$

From Table 3 it can be seen that the percentage of word recognition varied within the test lists to varying degrees. The largest deviations on average over all four levels were shown in test list 16 with 29 percentage points, the smallest in test lists, 4, 9, 15, and 20 with 20 percentage points. On average across all test lists the RMS value was 23.5 percentage points. For the group HI (not in Table 3), the mean RMS value was 17.4 percentage points (14 percentage points for test lists 14 and 19, and up to 22 percentage points for test list 1). It should be noted, however, that for the group HI the presentation levels were chosen so that the speech recognition, and therefore also the word recognition, was often close to 100% (see Figure 2). In this range, the variance of the measurement results decreases according to Equation 1. Hence, it

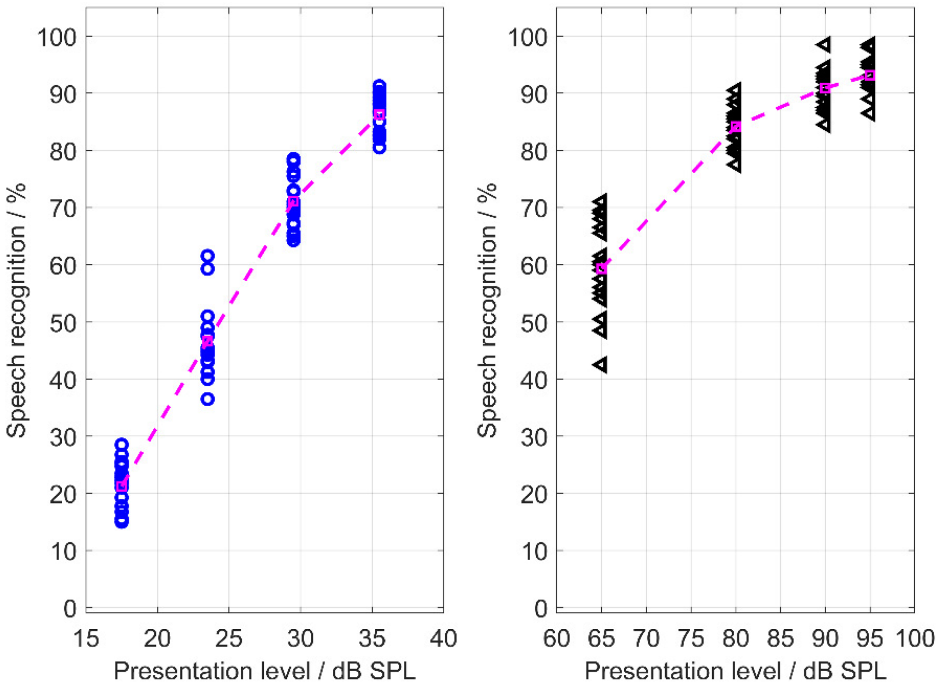

Figure 2: The symbols denote speech recognition per list, averaged across the listeners. The dashed lines link speech recognition averaged across all lists at the four presentation levels. Group NH (left) and group HI (right)

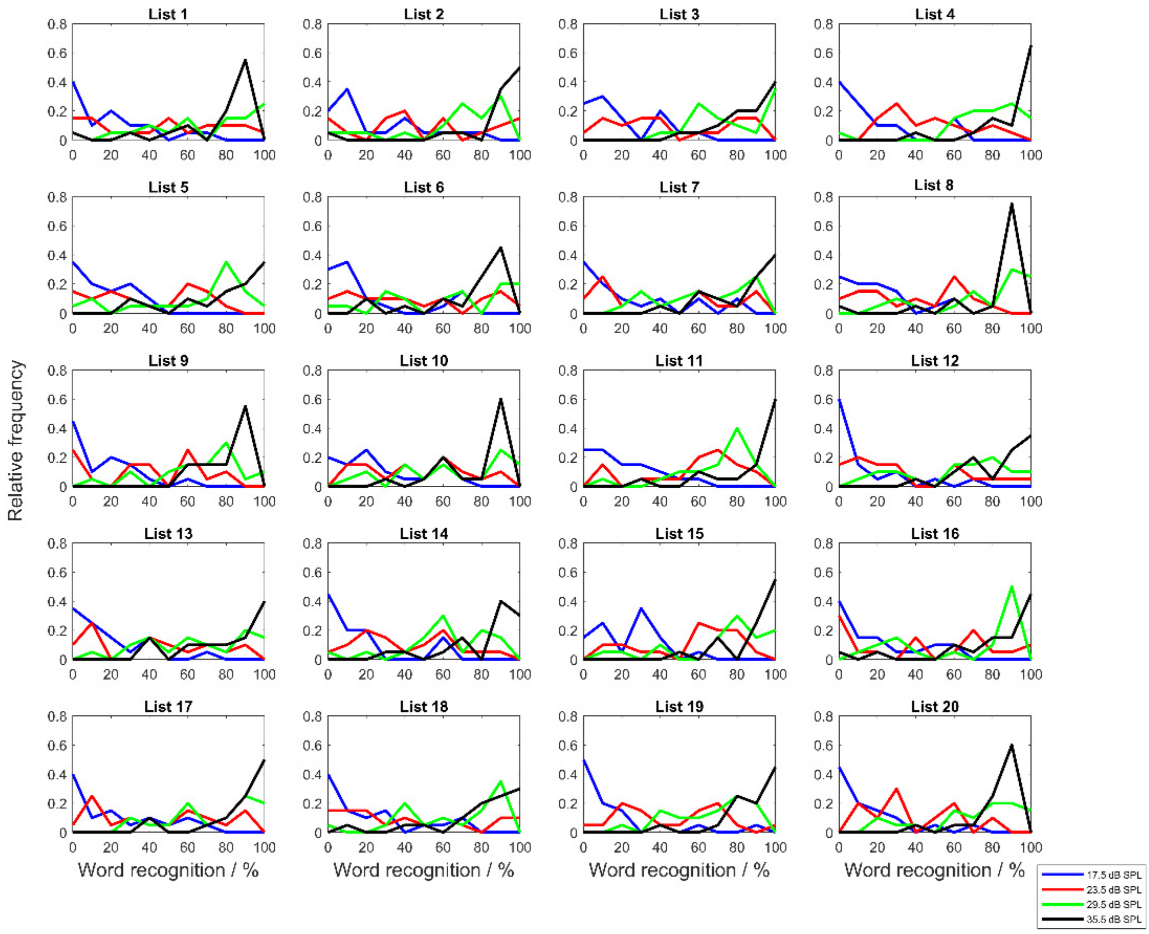

Figure 3: Frequency polygons for word recognition p_{ji} for all lists j (j=1...20) and for all presentation levels for group NH

Table 3: RMS-values of the measured word recognition p_{ji} in percentages, according to Equation 2; Group NH

Test list	17.5 dB SPL	23.5 dB SPL	29.5 dB SPL	35.5 dB SPL	Average of all 4 levels
1	20	31	30	25	27
2	25	34	30	22	28
3	20	30	19	15	21
4	20	24	22	15	20
5	14	27	29	23	23
6	25	32	32	24	28
7	26	32	25	16	25
8	21	26	24	24	24
9	17	29	22	14	20
10	22	26	28	20	24
11	17	27	22	18	21
12	21	30	27	15	23
13	20	30	26	21	24
14	19	25	25	18	22
15	16	25	25	14	20
16	20	36	33	26	29
17	23	31	23	17	23
18	25	34	28	22	27
19	23	26	21	14	21
20	19	21	24	14	20

cannot be concluded that a hearing impairment leads to a reduction of the variance.

In the Poisson binomial distribution, the expected value of the proportion of correctly recognized words in a test list in percent is given by the mean of the percentage probabilities of the individual words:

Equation 3

$$p_j = \frac{1}{n}\sum_{i=1}^{n} p_{ji}$$

The standard deviation in percent of the measurement result for a test list j with n words of different probability is:

Equation 4

$$\tilde{\sigma}_j = \frac{1}{n}\sqrt{\sum_{i=1}^{n} p_{ji}\left(100 - p_{ji}\right)}$$

This replaces Equation 1. If the percentage word recognition shown in Figure 3 is used as an approximation for the actually unknown probabilities p_{ji}, then the standard deviation $_j$ can be calculated. The results are plotted for all 20 test lists at all four levels and for both groups of participants in Figure 4, depending on the speech recognition p_j of the test lists in %. The observed relative frequency (Figure 2) was used for the probability p_{ji}. It was calculated according to Equation 3 from the average speech recognition of the individual words of test list j.

Approximation with a simple binomial distribution

As in Hagerman [3], the standard deviations $_j$ calculated from the Poisson binomial distribution were approximated by the standard deviation of a simple binomial distribution with a different value n'_j instead of the number n. This number of words n'_j is chosen such that a fictitious test list with n'_j equally understood words of the probability p_j has the same standard deviation as the test list j with n=20 words that vary in recognition. This means with Equation 1 and Equation 4:

Equation 5

$$\sqrt{\frac{p_j\left(100 - p_j\right)}{n'_j}} = \frac{1}{n}\sqrt{\sum_{i=1}^{n} p_{ji}\left(100 - p_{ji}\right)}$$

and therefore

Equation 6

$$n'_j = \frac{n^2 p_j\left(100 - p_j\right)}{\sum_{i=1}^{n} p_{ji}\left(100 - p_{ji}\right)} = n \cdot \left(1 - \frac{s^2_{\tilde{p}_j}}{p_j\left(100 - p_j\right)}\right)^{-1}$$

This calculation was carried out for each individual test list j with a distribution in word recognition p_{ji} and its mean value p_j. Since 20 test lists at four levels were measured for two groups of participants, 160 different values for n'_j were available. These values for n'_j were in the range

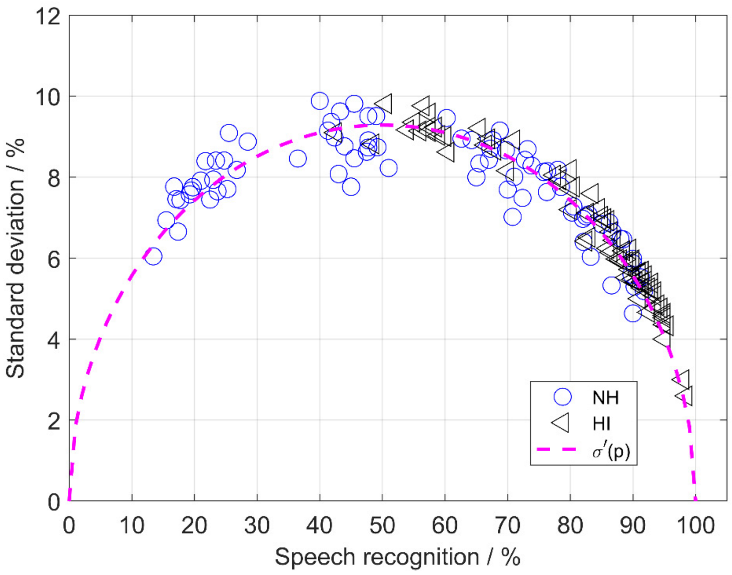

Figure 4: Standard deviation calculated using Equation 4 as a function of achieved speech recognition p_j in % for both groups (NH: blue circles, HI: black triangles). The dashed line gives the fitted function *σ'(p)* according to Equation 7 obtained from Equation 1 with *n'*=29 words (see Equation 9).

of 22–42 with an average of 28 (30 for NH and 27 for HI).

To obtain a common estimate of all conditions, instead of estimating the standard deviation from individual test lists and individual levels, a curve

Equation 7

$$\sigma'(p) = \sqrt{\frac{p(100-p)}{n'}}$$

was fitted to the standard deviations $_{j}$ plotted in Figure 4. The value *n'* is determined by the method of least squares:

Equation 8

$$\Delta(n') = \sum_{j=1}^{k} \left(\sigma'(p_j) - \tilde{\sigma}_j\right)^2$$

The calculation results in *dΔ/dn'*=0 for

Equation 9

$$n' = \left(\frac{\sum_{j=1}^{k} p_j\,(100-p_j)}{\sum_{j=1}^{k} \tilde{\sigma}_j \sqrt{p_j(100-p_j)}}\right)^2$$

Here *k* is the number of test-list measurement results used to calculate *n'*. If the results of both groups, NH and HI, are included, *k*=160 (20 test lists at four levels and two groups) and the result is *n'*=28.7. When fitted for each participant group separately, *k*=80. For NH alone, *n'*=29.5 results, and for HI alone *n'*=27.7. The standard deviation for test lists with 20 words in the FBE can therefore be modeled by the standard deviation *σ'* of test lists with approximately 29 words having the same word recognition. Using two test lists in the FBE (a total of 40 words) doubles not only *n*, but also *n'*. The standard deviation is inversely proportional to $\sqrt{n'}$. The relation *n'*>*n* is to be expected, because, according to Equation 23 in the Appendix (Attachment 1), the variance of the test-list score becomes smaller due to differences in word-recognition probability.

In Equation 7 this is achieved by increasing *n'* relative to *n*.

Confidence interval for test-list results

Table 4 gives the bounds of the 95% confidence interval for the result of individual test lists around the true value of speech recognition for *n'*=29 (one test list with 20 words) and *n'*=58 (two test lists with, together, 40 words). The bounds were calculated by multiplying *σ'* by *z*=1.96. In addition, the 95% confidence interval was determined directly from the Poisson binomial distribution. The distribution is explicitly known for each condition (each test list *j* at all four levels and for both groups of participants), see Equation 20 in the Appendix (Attachment 1). Thus, the confidence interval can be determined symmetrically for each score, starting from the two boundary values (0 words recognized, *n* words recognized). Figure 5 shows a good agreement between the two methods. A comparison of the bounds in Table 4 with Winkler and Holube [4] shows a shift of the bounds by a maximum of 5 percentage points when using one test list and by a maximum of 2.5 percentage points when using two test lists. Table 4 shows that for a speech recognition of 50%, doubling the word count from 20 to 40 causes the bounds to shift by 2.5 percentage points each. The 95% confidence interval is thus narrowed by 5 percentage points when the number of words is doubled. Thus, for a true speech recognition of 50%, the result of one test list must deviate by at least 20 percentage points to be significantly different, i.e. maximum 30% or at least 70%. In terms of

Table 4: 95% confidence interval for deviations from true values when using one or two test lists. The data was based on the standard deviation *σ'(p)* according to Equation 7; all values in %; intermediate values for two test lists were omitted for better clarity.

	95 % confidence interval in % for the measurement result of one test list		95 % confidence interval in % for the measurement result of two test lists	
Speech recognition, %	lower	upper	lower	upper
0	0	0	0.0	0.0
5	0	10	0.0	10.0
10	0	20	2.5	17.5
15	5	25	7.5	22.5
20	10	30	10.0	30.0
25	10	40	15.0	35.0
30	15	45	20.0	40.0
35	20	50	25.0	45.0
40	25	55	27.5	52.5
45	30	60	32.5	57.5
50	35	65	37.5	62.5
55	40	70	42.5	67.5
60	45	75	47.5	72.5
65	50	80	55.0	75.0
70	55	85	60.0	80.0
75	60	90	65.0	85.0
80	70	90	70.0	90.0
85	75	95	77.5	92.5
90	80	100	82.5	97.5
95	90	100	90.0	100.0
100	100	100	100.0	100.0

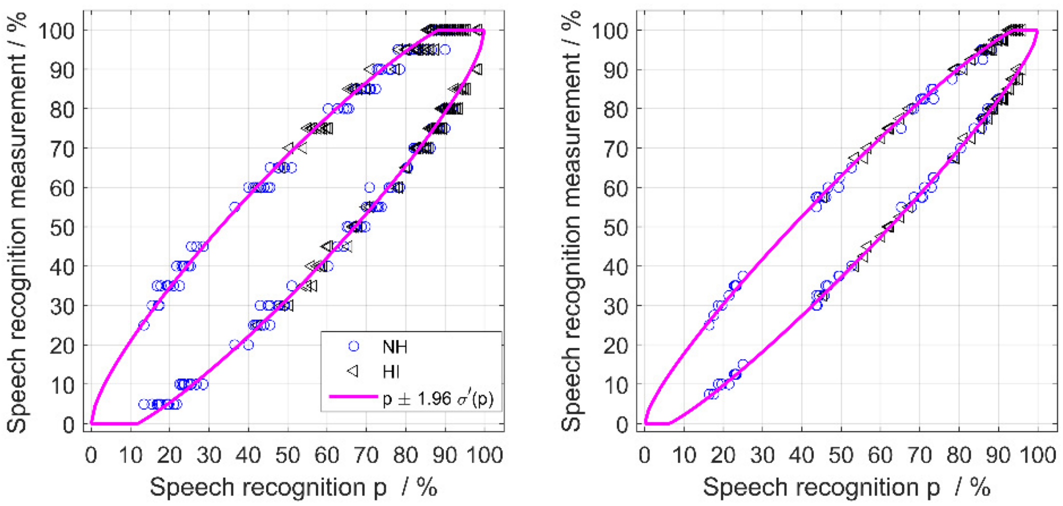

Figure 5: 95% confidence intervals for measured speech recognition with single test lists as a function of the true recognition score. The bounds for *p*±1.96 · *σ'(p)* with *n'*=29 (left) and *n'*=58 (right) are given as magenta lines. The symbols give the bounds directly obtained from the empirical Poisson binomial distribution for the groups (NH: blue circles, HI: black triangles) when using one test list (left) and two test lists (right).

statistics, it can then be concluded that the test-list result comes from a different population (with its own true value in speech recognition). Using two test lists, speech-recognition scores of 35% or 65%, i.e. a deviation of 15 percentage points, are significantly different from the assumed true value.

In order to verify the estimate of the 95% confidence interval from the binomial distribution, Figure 6 (left) shows the measurement results of group NH for single test lists in addition to the curves $p \pm 1.96 \cdot \sigma'(p)$ from Figure 5 (left). The symbols indicate the measurement result for each participant, each test list and each level. These values are given as a function of speech recognition for the test list averaged for all participants at the respective level. These average values represent an approximation of the true value of speech recognition of the respective test list at the given level. If all participants of the group NH had the same characteristics and abilities, the measurement results would be scattered according to the binomial distribution, and 95% of the results would be within the confidence interval (including bounds). However, of the 1,600 test-list results, 261, or 16.31%, are outside the confidence interval. In order to eliminate the variance due to the diversity of the participants, the average over all 20 test lists was used as a simple approach for each participants. The difference between the participant-specific average and the total mean of all measurement result was subtracted from results of the respective participant. This participant-specific difference is a measure of the characteristics and abilities of each participant relative to the other participants. Corrected measurement results below 0% were limited to 0% and above 100% were limited to 100% (necessary for a total of 7 measurement results). The corrected measurement results are shown in Figure 6 (right). After correction, there are 101 values, i.e. 6.3%, outside the 95% confidence interval.

Confidence interval for the true value

So far, in this contribution, the 95% confidence interval for measured values was calculated around the true value p of a test list using an assumed probability distribution. However, this true value is unknown. Another question is, in which range would this true value p lie with a probability of 95%, if only the measurement result p_{meas} for a single test list were available. Frequently, Equation 1 is also used to calculate this 95% confidence interval. Wilson [16] used the following approach for the limits of the 95% confidence interval for the true value with z=1.96:

Equation 10

$$(p - p_{\mathrm{meas}})^2 = z^2 \, \frac{p(100 - p)}{n}$$

Instead of n, the value n' was used in the present investigation (n'=29 for n=20, n'=58 for n=40). This takes into account the smaller width of the Poisson binomial distribution compared to the simple binomial distribution. Solving for p yields the lower limit u

Equation 11

$$u = \frac{p_{\mathrm{meas}} + 100\frac{z^2}{2n'} - \sqrt{p_{\mathrm{meas}}(100 - p_{\mathrm{meas}})\frac{z^2}{n'} + 100^2\frac{z^4}{4n'^2}}}{1 + \frac{z^2}{n'}}$$

and the upper limit o

Equation 12

$$o = \frac{p_{\mathrm{meas}} + 100\frac{z^2}{2n'} + \sqrt{p_{\mathrm{meas}}(100 - p_{\mathrm{meas}})\frac{z^2}{n'} + 100^2\frac{z^4}{4n'^2}}}{1 + \frac{z^2}{n'}}$$

The method for calculating the 95% confidence intervals for the true value around the measurements according to Equation 11 and Equation 12 was recommended by Altman et al. [17] and Brown et al. [18]. The results are shown in Figure 7 for n'=29 and n'=58.

Table 5 indicates the bounds thus obtained as numerical values. Since the true value in speech recognition is not limited by any measurement resolution, i.e. 5% steps when using 20 words and 2.5% steps when using 40 words, a corresponding rounding was omitted here. The largest differences compared to Table 4 are found at speech-recognition scores of 0% and 100%. The width of the confidence interval for the true value is larger than zero at these positions. For a test-list score of 80% using a 20-word test list, the true value of speech recognition lies in the range from 62.4% to 90.6% is with a probability of 95%. In addition, with a test list score of 90%, the true speech-recognition value can also be less than 80% with considerable probability. Even with the use of two test lists, the lower 95% confidence limit for the true value is still just under 80% for a measured speech recognition of 90%.

Discussion

In the current contribution, the FBE was modeled with a Poisson binomial distribution to account for the varying word recognition within the test lists. The modeling allowed the Poisson binomial distribution of the 20-word FBE test lists to be approximated by a simple 29-word binomial distribution. Thus, the results of Hagerman [3] were qualitatively confirmed. He derived a similar increase in the number of test items from n=25 to n'=33. When applied to the underlying measurement data, after elimination of participant's variability by a global, participant-specific correction value, 6.3% of the measurement results were outside the 95% confidence interval. This proportion is surprisingly close to the 5% expected theoretically to be outside the confidence interval. In doing so, other sources of variance, such as the fluctuating attention of the participants and of the examiner from test list to test list were not considered.

Equation 23 (see Attachment 1), which describes the relationship between the variances of a simple and a Pois-

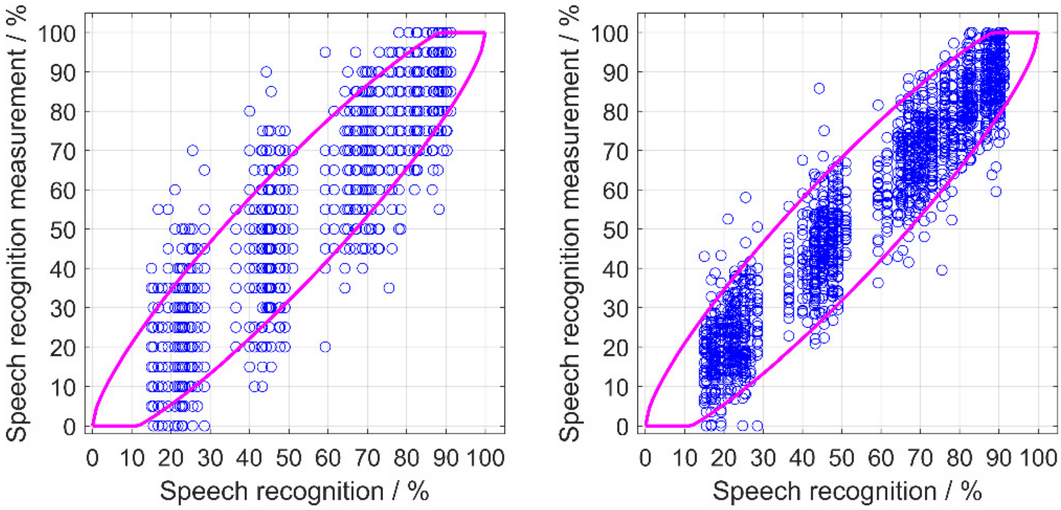

Figure 6: The circle symbols show the speech-recognition score for single test lists versus the average of the test list for all participants of group NH at a given level. The variance due to the differences between the participants was eliminated in the right side of the figure. As a comparison, the magenta lines give the 95%-confidence interval according to Figure 5(left).

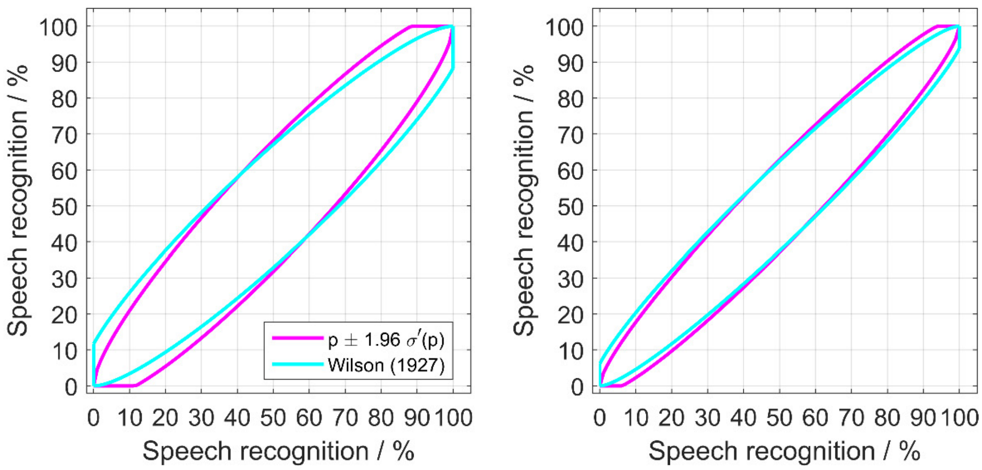

Figure 7: 95% confidence intervals for the true score as a function of a measured score, based on Wilson [16], Equation 11 and Equation 12 in comparison to the 95% confidence interval for measured scores relative to the true speech recognition according to $p \pm 1.96 \cdot \sigma'(p)$, as shown in Figure 5.

son binomial distribution with the same expectation value, leads to the conclusion that the reliability of a speech test improves with increasing inequality of test items in speech recognition. In an extreme case, a test list could consist only of words that are understood either always (i.e. with a probability of 100%) or never (i.e. with a probability of 0%). This measurement result could be reproduced with certainty. Here it becomes clear that narrow confidence intervals or a high reliability alone are not sufficient for evaluating a speech test. The aim of a speech test is to establish speech recognition as a function of the presentation level, to determine the success of a rehabilitation approach, or to compare different provisions with technical hearing devices. This requires the measurement of the course of the discrimination function, or specific points on the discrimination function, as accurately as possible. These goals cannot be achieved with test items that are either not recognized at all or are always recognized. A good speech test is not only characterized by high reliability or narrow 95% confidence intervals. It should also have a high sensitivity to level changes, for disability due to hearing loss, and for the effects of rehabilitation or care. These criteria were not quantitatively investigated in the present study. It should be noted, however, that a higher variation in word recognition within a test list leads to a flatter slope of the discrimination function for this list [19]. In [9] the test list-specific discrimination functions for the data of the NH group were given. The slope was only 4.5 percentage points per dB. In this sense, the variability of word recognition within a test list makes it possible to increase reliability and to decrease sensitivity. In order to increase the measurement accuracy of the FBE, the modified guidelines for assistive devices describe the use of two test lists. Doubling the number of words reduces the 95% confidence intervals. However, even with the Poisson binomial distribution, the reduction is not linear with the number of words n, but linear with $\sqrt{n}$ and thus does not change as much as would be desirable for a doubling of the measurement effort.

Table 5: 95% confidence interval for deviations of the true value from the measurement result when using one or two test lists. Given are confidence intervals based on Wilson [16]; Equation 11 and Equation 12. All values in %. Intermediate values for two test lists were omitted for better clarity.

	95 % confidence interval in % for the true value of one test list		95 % confidence interval in % for the true value of two test lists	
Speech recognition, %	lower	upper	lower	upper
0	0.0	11.7	0.0	6.2
5	1.1	19.4	1.7	13.9
10	3.4	26.0	4.6	20.4
15	6.2	32.0	8.0	26.3
20	9.4	37.6	11.7	32.0
25	12.8	43.0	15.6	37.5
30	16.5	48.2	19.8	42.7
35	20.3	53.2	24.0	47.9
40	24.4	58.0	28.4	52.8
45	28.6	62.6	32.9	57.7
50	32.9	67.1	37.5	62.5
55	37.4	71.4	42.3	67.1
60	42.0	75.6	47.2	71.6
65	46.8	79.7	52.1	76.0
70	51.8	83.5	57.3	80.2
75	57.0	87.2	62.5	84.4
80	62.4	90.6	68.0	88.3
85	68.0	93.8	73.7	92.0
90	74.0	96.6	79.6	95.4
95	80.6	98.9	86.1	98.3
100	88.3	100.0	93.8	100.0

Regarding the application of the analysis with regard to the guidelines for assistive devices, it has to be taken into account that the 95% confidence intervals calculated with the Poisson binomial distribution apply only to the FBE in quiet. The distribution of word recognition within the test lists for the FBE in noise is not yet known and may lead to a different reliability. If the distribution is wider, it will result in a reduction of the 95% confidence interval; if narrower, the 95% confidence interval will be increased. Furthermore, it should be noted that the analyses only show the 95% confidence intervals for the deviations of measured values from the true value, and for the deviations of the true value from a measured value. However, 95% confidence intervals of the difference of two measured scores are different variables that are needed to obtain the test-retest reliability. Here, the variances of the two individual measurements add up [20]. The test-retest reliability is relevant for the assessment of the comparison of the conditions with and without hearing aids, or of two hearing aids, or their settings. Therefore, the confidence intervals reported in this article cannot be used to compare with the requirements in the guideline (improvement by 20% in quiet and 10% in noise). This will be covered in a future contribution.

Notes

Publication note

This contribution was previously published in German as: Holube I, Winkler A, Nolte-Holube R. Modellierung der Reliabilität des Freiburger Einsilbertests in Ruhe mit der verallgemeinerten Binomialverteilung. Z Audiol. 2018;57(1):6-17.

Acknowledgement

This study was funded by the Ph.D. program Jade2Pro of Jade University of Applied Sciences. The authors thank Daniel Berg (HörTech gGmbH) for technical support as well as Sascha Bilert, Lena Haverkamp, Miriam Kropp, and Florian Wiese for their support in data acquisition. English language services were provided by stels-ol.de.

Competing interests

The authors declare that they have no competing interests.

Attachments

Available from
https://www.egms.de/en/journals/zaud/2020-2/zaud000005.shtml

1. Attachment1_zaud000005.pdf (169 KB)
   Attachment1: Poisson binomial distribution

References

1. Hahlbrock K. Über Sprachaudiometrie und neue Wörterteste. Archiv f. Ohren-, Nasen- u. Kehlkopfheilkunde. 1953;162:394–431. DOI: 10.1007/BF02105664
2. Egan JP. Articulation testing methods. Laryngoscope. 1948 Sep;58(9):955-91. DOI: 10.1288/00005537-194809000-00002
3. Hagerman B. Reliability in the determination of speech discrimination. Scand Audiol. 1976;5:219-28. DOI: 10.3109/01050397609044991
4. Winkler A, Holube I. Test-Retest-Reliabilität des Freiburger Einsilbertests [Test-retest reliability of the Freiburg monosyllabic speech test]. HNO. 2016 Aug;64(8):564-71. DOI: 10.1007/s00106-016-0166-2
5. Gemeinsamer Bundesausschuss. Tragende Gründe zum Beschluss des Gemeinsamen Bundesausschusses über eine Änderung der Hilfsmittel-Richtlinie: Freiburger Einsilbertest im Störschall. 24. November 2016. [accessed 05.06.2017]. Verfügbar unter: https://www.g-ba.de/informationen/beschluesse/2758/
6. Hey M, Brademann G, Ambrosch P. Der Freiburger Einsilbertest in der postoperativen CI-Diagnostik [The Freiburg monosyllable word test in postoperative cochlear implant diagnostics]. HNO. 2016 Aug;64(8):601-7. DOI: 10.1007/s00106-016-0194-y
7. DIN EN ISO 8253-1. Akustik – Audiometrische Prüfverfahren – Teil 1: Grundlegende Verfahren der Luft- und Knochenleitungs-Schwellenaudiometrie mit reinen Tönen. Berlin: Beuth Verlag; 2011.
8. DIN EN ISO 8253-3. Akustik – Audiometrische Prüfverfahren – Teil 3: Sprachaudiometrie. Berlin: Beuth Verlag; 2012.
9. Baljić I, Winkler A, Schmidt T, Holube I. Untersuchungen zur perzeptiven Äquivalenz der Testlisten im Freiburger Einsilbertest [Evaluation of the perceptual equivalence of test lists in the Freiburg monosyllabic speech test]. HNO. 2016 Aug;64(8):572-83. DOI: 10.1007/s00106-016-0192-0
10. DIN 45621-1. Sprache für Gehörprüfung. Teil 1: Ein- und mehrsilbige Wörter. Berlin: Beuth Verlag; 1995.
11. DIN 45626-1. Tonträger mit Sprache für Gehörprüfung, Teil 1: Tonträger mit Wörtern nach DIN 45621-1. Berlin: Beuth Verlag; 1995.
12. Brinkmann K. Die Neuaufnahme der „Wörter für Gehörprüfung mit Sprache“. Z Hörgeräteakustik. 1974;13:12-40.
13. DIN EN 60318-1. Akustik – Simulatoren des menschlichen Kopfes und Ohres – Teil 1: Ohrsimulator zur Kalibrierung von supra-auralen und circumauralen Kopfhörern. Berlin: Beuth Verlag; 1999.
14. DIN EN ISO 389-8. Akustik – Standard-Bezugsschwellenpegel für die Kalibrierung audiometrischer Geräte – Teil 8: Äquivalente BezugsSchwellenschalldruckpegel für reine Töne und circumaurale Kopfhörer. Berlin: Beuth Verlag; 2004.
15. ITU. ITU-T Recommendation G.227. Conventional telephone signal. Genf: ITU; 1993.
16. Wilson EB. Probable inference, the law of succession, and statistical interference. Journal of the American Statistical Association. 1927;22(158):209-212. DOI: 10.1080/01621459.1927.10502953
17. Altman DG, Machin D, Bryant TN, Gardner MJ. Statistics with confidence. BMJ Books. 2nd ed. 2000.
18. Brown LD, Cai TT, DasGupta A. Interval estimation for a binomial proportion. Statistical Science. 2001;16(2):101-117. DOI: 10.1214/ss/1009213286
19. Kollmeier B, Warzybok A, Hochmuth S, Zokoll MA, Uslar V, Brand T, Wagener KC. The multilingual matrix test: Principles, applications, and comparison across languages: A review. Int J Audiol. 2015;54 Suppl 2:3-16. DOI: 10.3109/14992027.2015.1020971
20. Thornton AR, Raffin MJ. Speech-discrimination scores modeled as a binomial variable. J Speech Hear Res. 1978 Sep;21(3):507-18. DOI: 10.1044/jshr.2103.507
21. Wang YH. On the number of successes in independent trials. Statistica Sinica. 1993;3:295-312.
22. Hong Y. On computing the distribution function for the Poisson binomial distribution. Computational Statistics & Data Analysis. 2013;59:41-51. DOI: 10.1016/j.csda.2012.10.006

Corresponding author:

Prof. Dr. Inga Holube
Jade Hochschule, Institut für Hörtechnik und Audiologie, Ofener Str. 16/19, D-26121 Oldenburg
Inga.Holube@jade-hs.de

Please cite as
Holube I, Winkler A, Nolte-Holube R. Modeling the reliability of the Freiburg monosyllabic speech test in quiet with the Poisson binomial distribution. Does the Freiburg monosyllabic speech test contain 29 words per list? GMS Z Audiol (Audiol Acoust). 2020;2:Doc01.
DOI: 10.3205/zaud000005, URN: urn:nbn:de:0183-zaud0000059

This article is freely available from
https://www.egms.de/en/journals/zaud/2020-2/zaud000005.shtml

Published: *2020-03-23*