OPEN ACCESS

This is the German version.
The English version starts at p. 14.

Modellierung und Verifizierung der Test-Retest-Reliabilität des Freiburger Einsilbertests in Ruhe mit der verallgemeinerten Binomialverteilung

Zusammenfassung

Die Test-Retest-Reliabilität des Freiburger Einsilbertests wurde mit verschiedenen Methoden modelliert und mit Messdaten von Probanden mit und ohne Hörbeeinträchtigung verglichen. Die Methoden bauen auf den Verfahren von Thornton und Raffin sowie Altman et al. auf. Sie berücksichtigen durch die Verwendung der verallgemeinerten Binomialverteilung die Unterschiede im Wortverstehen innerhalb der Testlisten und beinhalten die Varianz der Testlisten. Die Methoden ermöglichen die Bestimmung der Grenzen für die 90%- und 95%-Konfidenzintervalle bei Verwendung von Testlisten mit 20 Wörtern und von Doppellisten mit 40 Wörtern. Diese Grenzen wurden durch die Messdaten bestätigt. Bei einem Sprachverstehen von 50% sind die Konfidenzintervalle am breitesten. Dort hat für Testlisten mit 20 Wörtern das 90%-Konfidenzintervall eine Breite von ±20% bzw. ±6,0 dB und das 95%-Konfidenzintervall eine Breite von ±25% bzw. ±7,4 dB. Für die Hörgeräte-Anpasspraxis bedeutet dies, dass erst Unterschiede, die diese Spanne übersteigen, als signifikant unterschiedlich gewertet werden können.

Schlüsselwörter: Freiburger Einsilbertest, Sprachverstehen, Binomialverteilung, Test-Retest-Reliabilität, Konfidenz

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

1 Institut für Hörtechnik und Audiologie, Jade Hochschule, Oldenburg, Deutschland

2 Exzellenzcluster "Hearing4All", Oldenburg, Deutschland

Einleitung

In Heft 1/2018 wurde die Modellierung der Reliabilität des Freiburger Einsilbertests (FBE) [1] in Ruhe mit der verallgemeinerten Binomialverteilung vorgestellt [2], [3]. Die Verwendung dieser Verteilung ermöglicht die Berücksichtigung der Unterschiede im Wortverstehen innerhalb einer Testliste. Dies führt zu einem kleineren Konfidenzintervall für die Messwerte als die Verwendung der einfachen Binomialverteilung, die für jedes Wort einer Liste die gleiche Erkennungswahrscheinlichkeit annimmt. Die Varianz der verallgemeinerten Binomialverteilung für Testlisten mit 20 Wörtern konnte durch die Varianz einer einfachen Binomialverteilung angenähert werden, die auf Testlisten mit 29 Wörtern mit gleichem Wortverstehen beruht.

Die Untersuchungen bei Holube et al. [2], [3] beschränken sich auf die Berechnung des 95%-Konfidenzintervalls für die Abweichung des Messwertes für eine Testliste vom wahren Wert und alternativ für die Abweichung des wahren Wertes von dem Messwert für eine Testliste. Die publizierten Konfidenzintervalle sind jedoch nicht für die Abschätzung der Test-Retest-Reliabilität oder bei Untersuchungen mit zwei Testlisten zum Vergleich von zwei Messbedingungen anwendbar. Genau dieser Fall liegt jedoch bei der Überprüfung von Hörgeräten oder anderen Therapiemaßnahmen vor. Das Ergebnis zweier Messungen (z.B. mit und ohne Hörgeräte), d.h. zweier Trefferraten, wird verglichen, und aus der Differenz der beiden Trefferraten wird der Erfolg der Maßnahme abgeleitet. In der Hilfsmittelrichtlinie [4] wird z.B. mit dem FBE in Ruhe eine Verbesserung des Sprachverstehens von mindestens 20 Prozentpunkten mit Hörgeräten im Vergleich zur unversorgten Kondition gefordert.

Thornton und Raffin [5] berechneten das 95%-Konfidenzintervall für die Differenz zwischen zwei Messungen, indem sie die Trefferraten in eine Skala mit homogenen Varianzen für alle Testergebnisse transformierten und dann die Varianzen der zwei Testergebnisse addierten. Carney und Schlauch [6] bestätigten im Wesentlichen die Ergebnisse dieser Methode durch einen anderen Ansatz. Sie berechneten die Varianz der Differenz zweier Trefferraten unter der Annahme binomialverteilter Testergebnisse. Für jeden Wert für die Trefferrate aus der ersten Messung berücksichtigten sie dabei alle möglichen Werte für die zweite Messung. Die Ergebnisse der Methode von Thornton und Raffin [5], die gleiches Verstehen aller 20 Wörter einer Testliste voraussetzt, wurden von Winkler und Holube [7] basierend auf Steffens [8] angegeben und mit Ergebnissen wiederholter Messungen verglichen.

Dillon [9] legte einerseits dar, dass bei Annahme der gleichen Wahrscheinlichkeit für das Verstehen jedes Wortes die Breite des 95%-Konfidenzintervalls für die Test-Retest-Kondition durch die Verwendung der Methode von Thornton und Raffin [5] überschätzt wird, wenn die Testlisten gleich verständlich sind und sich die Probanden immer gleich verhalten. Diese Annahme wird durch die

Analyse in Winkler und Holube [7] gestützt, da nur 3,2% der Messdaten, d.h. weniger als die erwarteten 5% der Messdaten außerhalb des Konfidenzintervalls nach Thornton und Raffin [5] lagen. Andererseits wies Dillon [9] darauf hin, dass die Methode von Thornton und Raffin [5] trotzdem zur Abschätzung des 95%-Konfidenzintervalls verwendet werden kann, da sich zwei Effekte gegenseitig aufheben: Bei Berücksichtigung unterschiedlichen Wortverstehens und Anwendung der verallgemeinerten Binomialverteilung nach Hagerman [10] werden die 95%-Konfidenzintervalle schmaler. Durch intraindividuelle Variabilität (z.B. durch Aufmerksamkeitsschwankungen) vor allem bei einem größeren zeitlichen Abstand der Messungen werden sie jedoch wieder breiter. Als zusätzliche Varianzquelle weist Dillon [9] auf mögliche Unterschiede zwischen den Testlisten hin. In der Sprachaudiometrie werden, im Gegensatz zu Winkler und Holube [7], im Allgemeinen nicht die gleichen Listen bei wiederholten Messungen verwendet. Das 95%-Konfidenzintervall für die Test-Retest-Reliabilität verbreitert sich bei Verwendung unterschiedlicher Testlisten infolge der unterschiedlichen mittleren Trefferraten der Testlisten.

Für die vorliegende Analyse wurden die Messungen aus Baljic et al. [11] und Holube et al. [2], [3] die für jeden Probanden die Ergebnisse von fünf Testlisten bei jedem von vier Pegeln beinhalten, im Sinne eines Test-Retest-Experiments interpretiert und die Test-Retest-Reliabilität ausgewertet. Alle Messungen wurden innerhalb eines Termins durchgeführt, so dass lediglich die Kurzzeit-Test-Retest-Reliabilität untersucht wurde, nicht jedoch die Test-Retest-Reliablität über einen längeren Zeitraum, die nach Dillon [9] vermutlich zu breiteren Konfidenzintervallen führen würde. Zum Vergleich mit den Messdaten wurden die Grenzen für das 95%- und das 90%-Konfidenzintervall mit verschiedenen Methoden modelliert. Die Methoden bauen auf der in Holube et al. [2], [3] verwendeten verallgemeinerten Binomialverteilung auf und modellieren zusätzlich die Variabilität der Testlisten. Intraindividuelle Varianzen der Probanden wurden aufgrund der geringen zeitlichen Abstände zwischen den Messungen vernachlässigt.

Methoden

Experimentelle Daten

Die Messmethoden werden hier nur kurz zusammengefasst. Für eine ausführliche Beschreibung sei auf Holube et al. [2], [3] verwiesen.

Bei 80 jungen Probanden mit normalem Hörvermögen (im Folgenden als Normalhörende bezeichnet), wurde das Sprachverstehen als Trefferrate für die Freiburger Einsilber in Ruhe bei vier Pegeln (17,5, 23,5, 29,5 und 35,5 dB SPL) mit jeweils fünf Testlisten à 20 Wörtern (n=20) bestimmt. Bei 40 älteren Probanden mit Hörbeeinträchtigung (im Folgenden als Schwerhörige bezeichnet) wurden bei sonst gleichem Verfahren die Pegel 65, 80, 90 und 95 dB SPL verwendet. In die Analyse wurden jedoch nur die Pegel 65 und 80 dB SPL einbezogen, da bei den beiden höheren Pegeln viele Trefferraten bei 100% lagen. Alle Messungen eines Probanden wurden innerhalb eines Termins durchgeführt.

Die fünf Testlisten-Trefferraten bei festem Pegel für jeden Probanden wurden als Test-Retest-Kombinationen in Paaren interpretiert. Die Paare setzten sich jeweils aus einer präsentierten Testliste und einer der danach präsentierten weiteren Testliste zusammen, d.h. (1; 2), (1; 3), (1; 4), (1; 5), (2; 3), (2; 4), (2; 5), (3; 4), (3; 5), (4; 5). Dadurch ergaben sich 3.200 Test-Retest-Paare für die Normalhörenden und 800 Test-Retest-Paare für die Schwerhörigen. Die Anzahl der Test-Retest-Paare verringerte sich, wenn die bei Baljic et al. [11] auffälligen Testlisten ausgeschlossen wurden (siehe Tabelle 1). In einer weiteren Variante wurden jeweils aus zwei Testlisten Doppellisten mit n=40 Wörtern gebildet. Für die Analyse der Test-Retest-Reliabilität wurden alle Doppellisten zu Test-Retest-Paaren kombiniert, so dass keine Einzelliste doppelt vorkam, d.h. (1+2; 3+4), (1+2; 3+5), (1+2; 4+5), (1+3; 2+4), (1+3; 2+5), (1+3; 4+5), (1+4; 2+3), (1+4; 2+5), (1+4; 3+5), (1+5; 2+3), (1+5; 2+4), (1+5; 3+4), (2+3; 4+5), (2+4; 3+5) und (2+5; 3+4). Daraus ergaben sich bei Verwendung aller Testlisten 4.800 Test-Retest-Paare für die Normalhörenden und 1.200 Test-Retest-Paare für die Schwerhörigen. Auch für diese Doppellisten wurden die nach [11] auffälligen Testlisten als Variante ausgeschlossen (siehe Tabelle 1).

Berechnungsmethoden

Bei gegebener Trefferrate p_{mess1} (Test) ist die Frage, in welchem kritischen Bereich die Retest-Trefferrate p_{mess2} liegt, sodass die Differenz $p_{mess1}-p_{mess2}$ bei zweiseitiger Fragestellung auf dem α=5%-Niveau gerade noch nicht signifikant von Null verschieden ist. Eine zweiseitige Fragestellung bedeutet dabei, dass die Retest-Trefferrate kleiner oder größer als die erste Trefferrate sein kann und 2,5% der Retest-Trefferraten unterhalb sowie 2,5% der Retest-Trefferraten oberhalb des 95%-Konfidenzintervalls um die erste Trefferrate liegen. Zur Berechnung des 95%-Konfidenzintervalls existieren in der Literatur unterschiedliche Methoden, von denen zwei (Thornton und Raffin [5] und Altman et al. [12]) in der vorliegenden Arbeit betrachtet werden. Beide Methoden werden zunächst reproduziert und dann für die vorliegenden Messdaten mit n=20 bzw. n=40 Worten pro Testliste (d.h. einfache Testlisten und Doppellisten) angewendet. Danach werden Modifikationen dieser Methoden vorgestellt, die die Variabilität des Einzelwortverstehens sowie die Variabilität des mittleren Verstehens der unterschiedlichen Testlisten berücksichtigen.

Methode 1: Kritische Differenzen nach Thornton und Raffin

Thornton und Raffin [5] schlugen die Berechnung eines 95%-Konfidenzintervalls zur Beurteilung der Test-Retest-Reliabilität nach folgender Methode vor: Die Anzahl X

Tabelle 1: Anzahl der verwendeten Datenpunkte und Prozentsatz der Daten außerhalb des berechneten 95%-Konfidenzintervalls
Angaben für Normalhörende (NH) und Schwerhörige (SH) mit *n*=20 und *n*=40 Wörtern pro Liste für die Methoden 1–5

		Methoden						
		1	**2**	**3**	**4**	**4/16**	**5**	**5/16**
NH $n=20$	Datenpunkte	3.200	3.200	3.200	3.200	2.034	3.200	2.034
	%	4,8	8,7	9,4	5,4	4,5	5,4	4,4
SH $n=20$	Datenpunkte	800	800	800	800	507	800	507
	%	4,9	8,9	9,3	5,3	5,1	5,0	4,7
NH $n=40$	Datenpunkte	4.800	4.800	4.800	4.800	1.890	4.800	1.890
	%	4,6	9,2	9,0	5,1	4,3	5,1	4,2
SH $n=40$	Datenpunkte	1.200	1.200	1.200	1.200	465	1.200	465
	%	3,9	9,5	8,9	4,5	3,9	4,7	3,9

richtiger Antworten bei *n* angebotenen Worten einer Liste wird als Zufallsgröße angesehen. Sie wird als binomialverteilt nach *B(n,p,X=k)* angenommen. Dabei ist *p* die Wahrscheinlichkeit dafür, dass ein Wort der Liste richtig verstanden wird. Hier und im Folgenden werden Wahrscheinlichkeiten in Prozent angegeben. Der Erwartungswert von *X* ist somit $\mathrm{E}(X)=\frac{np}{100}$. Das Sprachverstehen in Prozent (Trefferrate) ist mit diesen Bezeichnungen die Zufallsgröße $p_{\mathrm{mess}}=\frac{100X}{n}$. Ihr Erwartungswert beträgt $\mathrm{E}(p_{mess})=p$, ihre Varianz ist $\mathrm{Var}(p_{\mathrm{mess}})=\sigma^2=\frac{p(100-p)}{n}$. Diese Varianz nimmt ihr Maximum bei *p*=50% an. An den Rändern bei *p*=0 und *p*=100% ist die Varianz Null.
Für die Test-Retest-Reliabilität ist die Abschätzung eines Konfidenzintervalls für die Differenz $p_{mess1}-p_{mess2}$ zweier Trefferraten von Interesse. Dazu werden die Zufallsgrößen X_1 und X_2 zunächst (nach Gleichung 3 in [5]) gemäß Gleichung 1 in einen Winkelbereich $\theta(X,n)$ transformiert.

Gleichung 1

$$\theta(X,n)=\sin^{-1}\sqrt{\frac{X}{n+1}}+\sin^{-1}\sqrt{\frac{X+1}{n+1}}$$

Die so definierte Zufallsgröße θ hat näherungsweise eine von *p* unabhängige Varianz Var(θ). Thornton und Raffin [5] wählten die Näherungen $\mathrm{Var}(\theta)=\frac{1}{n+\frac{1}{2}}$ für $n\geq 50$ bzw. $\mathrm{Var}(\theta)=\frac{1}{n+1}$ für $10<n<50$. Die beiden Zufallsgrößen $\theta_1=\theta(X_1,n)$ und $\theta_2=\theta(X_2,n)$ haben im Rahmen dieser Näherung die gleiche Varianz Var(θ). Unter der Annahme, dass θ_1 und θ_2 statistisch unabhängig sind, ist die Varianz der Zufallsgröße $\Delta\theta=\theta_1-\theta_2$ die Summe der Varianzen, also $\mathrm{Var}(\Delta\theta)=2\mathrm{Var}(\theta)$. Für $\Delta\theta$ wird nun eine Normalverteilung mit der Varianz 2Var(θ) angenommen. Das 95%-Konfidenzintervall für θ_2 bei einer Trefferrate p_{mess1} ergibt sich somit zu $\theta_1\pm 1{,}96\sqrt{\mathrm{Var}(\Delta\theta)}$. Die so berechneten θ_2-Grenzen des 95%-Konfidenzintervalls werden zu X_2-Grenzen zurücktransformiert, um dann die entsprechenden p_{mess2} -Grenzen zu erhalten.

Bezeichnen also $p_{\mathrm{mess1}}=\frac{100X_1}{n}$ und $p_{\mathrm{mess2}}=\frac{100X_2}{n}$ die Trefferrate in der Test- und in der Retest-Messung, so kann diese Methode wie folgt zusammengefasst werden:

Gleichung 2

$$p_{\mathrm{mess2,\,Untergrenze}}=\frac{100}{n}\theta^{-1}\left(\theta(X_1,n)-1{,}96\cdot\sqrt{\mathrm{Var}(\Delta\theta)},n\right)$$

Gleichung 3

$$p_{\mathrm{mess2,\,Obergrenze}}=\frac{100}{n}\theta^{-1}\left(\theta(X_1,n)+1{,}96\cdot\sqrt{\mathrm{Var}(\Delta\theta)},n\right)$$

mit
Gleichung 4

$$\mathrm{Var}(\Delta\theta)=\frac{2}{n+1}.$$

Diese Grenzen wurden für alle interessierenden Trefferraten p_{mess1} zwischen 0 und 100% berechnet. Die Berechnung der Umkehrfunktion $X=\theta^{-1}(\theta, n)$ von Gleichung 1 erfolgte dabei numerisch.

Methode 2: Kritische Differenzen nach Thornton und Raffin mit variablem Einzelwortverstehen

Sind die einzelnen Worte einer Liste unterschiedlich gut zu verstehen, genügt die gleiche Trefferwahrscheinlichkeit *p* für jedes Wort nicht mehr zur Beschreibung. Jedes Wort hat eine eigene Trefferwahrscheinlichkeit, und die Binomialverteilung wird durch die verallgemeinerte Binomialverteilung ersetzt [10]. Um die Verschmälerung der Verteilung von *X* bei der verallgemeinerten Binomialverteilung gegenüber der einfachen Binomialverteilung zu berücksichtigen, soll nun in der Berechnung im θ-Bereich die Varianz von θ zu $\mathrm{Var}(\theta)=\frac{1}{n'+1}$ anstelle von $\frac{1}{n+1}$ angenommen werden. Der Wert für n' wurde aus [2], [3] übernommen, also $n'=29$ für $n=20$ und $n'=58$ für $n=40$. Die Methode 2 wird somit durch die Gleichung 2 und Gleichung 3 mit

Gleichung 5

$$\mathrm{Var}(\Delta\theta) = \frac{2}{n' + 1}$$

anstelle von Gleichung 4 beschrieben.

Methode 3: Kritische Differenzen nach Altmann et al. mit variablem Einzelwortverstehen

Altman et al. [12] empfehlen einen Ansatz, der der Methode 10 von [13] entspricht. Diese Methode wird hier zunächst unverändert vorgestellt. Danach wird sie modifiziert, um die Variabilität des Wortverstehens innerhalb einer Liste zu berücksichtigen.
Liegt eine Trefferrate p_{mess} für eine einzelne Testliste vor, kann nach dem 95%-Konfidenzintervall für den wahren Wert p gefragt werden. Wilson [14] machte dazu den folgenden Ansatz:

Gleichung 6

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

mit z=1,96. Dies ist eine quadratische Gleichung für p. Ihre beiden Lösungen u und o geben die untere bzw. die obere Grenze für das gesuchte Konfidenzintervall an (siehe [2], [3]). Liegen zwei Trefferraten p_{mess1} und p_{mess2} vor, so ergeben sich die zugehörigen Untergrenzen u_1 und u_2 sowie die Obergrenzen o_1 und o_2. Nach [12] wird die Signifikanz der Differenz p_{mess1}–p_{mess2} wie folgt beurteilt: Wenn die erste Trefferrate p_{mess1} größer ist als die zweite Trefferrate p_{mess2}, dann muss die Differenz p_{mess1}–p_{mess2} der beiden Trefferraten größer sein als

Gleichung 7

$$\delta_u = \sqrt{(p_{\mathrm{mess1}} - u_1)^2 + (p_{\mathrm{mess2}} - o_2)^2}\ ,$$

um auf dem 5%-Niveau signifikant unterschiedlich zu sein. Zur Berechnung des 95%-Konfidenzintervalls für die Differenz zwischen den beiden Trefferraten werden also die Varianz für die obere Trefferrate nach unten und die Varianz für die untere Trefferrate nach oben addiert. Für den anderen Fall, dass nämlich die zweite Trefferrate größer ist als die erste Trefferrate, muss die Differenz p_{mess2}–p_{mess1} entsprechend größer sein als

Gleichung 8

$$\delta_o = \sqrt{(p_{\mathrm{mess1}} - o_1)^2 + (p_{\mathrm{mess2}} - u_2)^2}\ .$$

Dieses Verfahren liefert für jeden der interessierenden Werte von p_{mess1} zwischen 0 und 100 % ein 95%-Konfidenzintervall für die Differenz p_{mess2}–p_{mess1}. Bei gegebenem p_{mess1} (Test) liegt p_{mess2} (Retest) mit einer Wahrscheinlichkeit von 95% zwischen p_{mess1}–δ_u und p_{mess1}+δ_o. Die sechs Gleichungen, d.h. die Gleichungen für u_1, u_2, o_1 und o_2 sowie die Gleichung 7 und Gleichung 8, müssen für gegebenes p_{mess1} gelöst werden. Geschlossene Lösungen lassen sich nicht angeben, daher wurden sie numerisch durch Fixpunktiteration gelöst.
Die bisher beschriebene Berechnungsmethode geht von gleichem Einzelwortverstehen innerhalb einer Testliste aus. Die Variabilität des Einzelwortverstehens führt wie schon für Methode 2 beschrieben zu einer Verkleinerung der Varianz $\frac{p(100-p)}{n}$ auf der rechten Seite von Gleichung 6. Hier soll dies durch die Ersetzung von n durch n' berücksichtigt werden. Dabei wird der Wert für n' wieder aus [2], [3] übernommen, also n'=29 anstelle von n=20 und n'=58 anstelle von n=40.

Methode 4: Kritische Differenzen nach Altmann et al. mit variablem Einzelwortverstehen und variablem Testlistenverstehen

Ausgehend von variablem Einzelwortverstehen unter gleichen Bedingungen variiert bei einem Sprachtest der Mittelwert zwischen den Testlisten aufgrund der unterschiedlichen Wortzusammensetzungen der Testlisten. Wäre für jede Testliste der Testlistenmittelwert unter gegebenen Messbedingungen genau ermittelbar, hätte dieser Mittelwert daher eine Varianz f_n^2. Diese hängt von der Anzahl n der Wörter pro Testliste sowie vom wahren Wert p ab. Die Varianz trägt zur Unsicherheit des wahren Wertes von p in Gleichung 6 bei. Wird also in dieser Gleichung sowohl das variable Einzelwortverstehen (Ersetzung von n durch n') als auch das variable Testlistenverstehen (Addition von f_n^2 zur Varianz von p) berücksichtigt, wird der Ansatz von Gleichung 6 zu:

Gleichung 9

$$(p - p_{\mathrm{mess}})^2 = z^2 \left(\frac{p(100 - p)}{n'} + f_n^2 \right)$$

mit z=1,96. Wenn die Varianz f_n^2 bekannt ist, können die weiteren Schritte der Methode nach [12], wie für Methode 3 beschrieben, durchgeführt werden.
Zur Ermittlung von f_n^2 wird die Stichprobenvarianz der gemessenen Testlistenmittelwerte berechnet. Betrachtet werden n_L Testlisten aus je n Wörtern mit dem Einzelwortverstehen p_{ji}, i=1...n, j=1...n_L. Die Trefferrate der Testliste j ist damit der Mittelwert $p_j = \frac{1}{n}\sum_{i=1}^{n} p_{ji}$. Mit den über alle Wörter in allen Testlisten gemittelten Trefferraten

Gleichung 10

$$\bar{p} = \frac{1}{n_L} \sum_{j=1}^{n_L} p_j$$

ist dann f_n^2 die Stichprobenvarianz der Testlistenmittelwerte gemäß:

Gleichung 11

$$f_n^2 = \frac{1}{n_L - 1}\sum_{j=1}^{n_L}(p_j - \bar{p})^2 \quad .$$

Die Varianz des Einzelwortverstehens ist

Gleichung 12

$$f_1^2 = \frac{1}{n \cdot n_L - 1}\sum_{j=1}^{n_L}\sum_{i=1}^{n}(p_{ji} - \bar{p})^2 \quad .$$

Zwischen der Varianz des Verstehens eines einzelnen Wortes und der Varianz der Mittelwerte aus *n* zufällig zu Testlisten zusammengestellten Einzelwörtern besteht die Beziehung

Gleichung 13

$$f_{n,\text{zufällig}}^2 = \frac{1}{n} f_1^2 \quad .$$

Die Abbildung 1 zeigt, dass diese Beziehung im Mittel für zufällig aus den Wörtern des FBE zusammengestellte Testlisten mit *n*=1, 20, 40 erfüllt ist. Die dargestellten Varianzen wurden aus 10^6 Realisierungen von zufällig zusammengestellten Testlisten gemittelt. Sie zeigt aber auch, dass die Varianzen der konkreten Testlisten des FBE deutlich von dem mittleren Ergebnis einer zufälligen Wortzusammenstellung abweichen. Darüber hinaus zeigt Abbildung 1 erwartungsgemäß, dass f_n^2 in der Nähe von *p*=0% (fast kein Wort wird verstanden) und *p*=100% (fast alle Wörter werden verstanden) kleiner ist als im mittleren Bereich um *p*=50%. Der genaue Verlauf von f_n^2 als Funktion von *p* ist nicht bekannt. Als Ansatz wird hier eine Parabel

Gleichung 14

$$f_n^2 = \frac{c^2}{n} p(100 - p) \quad ,$$

mit einem noch zu bestimmenden Parameter c^2 gewählt, so dass sich Gleichung 9 als

Gleichung 15

$$(p - p_{\text{mess}})^2 = z^2 \, \frac{p(100 - p)}{\tilde{n}}$$

mit

Gleichung 16

$$\frac{1}{\tilde{n}} = \frac{1}{n'} + \frac{c^2}{n}$$

schreiben lässt. Wird also in der Methode 3 die Gleichung 6 durch Gleichung 15 ersetzt, dann werden sowohl die Variabilität des Einzelwortverstehens als auch die Variabilität der Testlistenmittelwerte berücksichtigt.
Der Parameter c^2 wurde aus dem gemessenen Einzelwortverstehen p_{ji} wie folgt berechnet. Für jeden der vier verwendeten Pegel werden der Mittelwert $\bar{p}$ des Einzelwortverstehens nach Gleichung 10 und die Varianz f_n^2 nach Gleichung 11 berechnet. Die Werte der vier Paare $(\bar{p}, f_n^2)$ hängen von der Auswahl und von der Wortzusammenstellung der zugrunde liegenden Testlisten sowie von ihrer Länge *n* ab. An die vier Wertepaare $(\bar{p}, f_n^2)$ wird nach der Methode der kleinsten Quadrate eine Parabel $f_n^2 = \frac{c^2}{n} p(100 - p)$ angepasst. Dies liefert den gesuchten Wert für c^2. Drei der so resultierenden Parabeln sind in der Abbildung 1 eingezeichnet. Mit dem nun bekannten Wert für c^2 wird die effektive Listenlänge $\tilde{n}$ mit Hilfe von Gleichung 16 berechnet. Die Tabelle 2 zeigt die Ergebnisse für *n*=20 und für *n*=40. Da der FBE 20 Wörter pro Liste hat, wurden für die Berechnungen mit *n*=40 alle Kombinationen aus Paaren unterschiedlicher Listen berücksichtigt.

Methode 5: Kritische Differenzen nach Thornton und Raffin mit variablem Einzelwortverstehen und variablem Listenverstehen

Durch die Berücksichtigung der Einzelwortvariabilität verringert sich die Varianz von Gleichung 4 zu Gleichung 5. Es liegt also nahe, die Variabilität des Listenverstehens durch die Ersetzung von Gleichung 5 durch

Gleichung 17

$$\text{Var}(\Delta\theta) = \frac{2}{\tilde{n} + 1}$$

zu modellieren.

Kritische Differenzen bei einseitiger Fragestellung

Bisher wurde das 95%-Konfidenzintervall bei zweiseitiger Fragestellung betrachtet. Bei der Anwendung des FBE in der Hörgeräteanpassung wird jedoch vorausgesetzt, dass Hörgeräte das Sprachverstehen verbessern, dass also bei der zweiten Messung (mit Hörgerät) eine höhere Trefferrate erreicht wird als bei der ersten Messung (ohne Hörgerät). Der statistische Test zur Ermittlung eines signifikanten Unterschieds zwischen den beiden Trefferraten würde dann untersuchen, ob die Irrtumswahrscheinlichkeit für die Hypothese, dass die zweite Trefferrate größer als die erste Trefferrate ist, kleiner als 5% ist. Das entspricht der Grenze des 90%-Konfidenzintervalls. Dies kann mit den gleichen fünf Methoden berechnet werden, indem *z*=1,96 durch *z*=1,645 ersetzt wird. Obwohl die Fragestellung einseitig ist, werden die Grenzen des 90%-Konfidenzintervalls für die zweite Trefferrate der Vollstän-

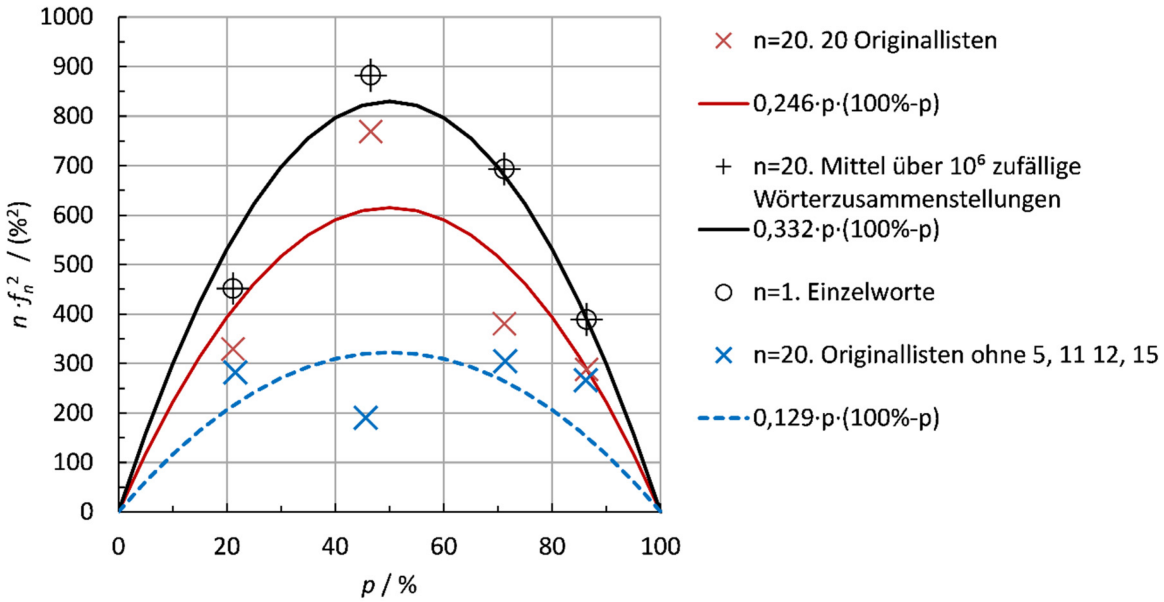

Abbildung 1: Varianz der Testlistenmittelwerte der gemessenen Trefferrate als Funktion der Trefferrate *p*. Die Varianzen wurden zur besseren Vergleichbarkeit jeweils mit der Wortanzahl *n* pro Liste multipliziert (siehe Gleichung 13). Die Symbole zeigen die Varianzen, die sich für unterschiedliche Zusammenstellungen der Wörter zu Listen ergeben. Für jede Zusammenstellung wurde der Mittelwert (10) der Trefferrate aller beteiligten Wörter als Abszissenwert *p* verwendet. Weil die Messwerte zu vier unterschiedlichen Pegeln gehören, gruppieren sich die Symbole um die vier entsprechenden Trefferraten *p*. Die eingezeichneten Linien zeigen die angepassten Parabeln nach Gleichung 14. Zur Zuordnung der Zahlenwerte für c^2 zu den Zusammenstellungen siehe Tabelle 2. Der Wert c^2=0,332 resultiert aus der Anpassung an die Varianz des Einzelwortverstehens (*n*=1) aller Wörter.

Tabelle 2: Aus dem Einzelwortverstehen ermittelte effektive Wörterzahlen für die Listenlängen *n*=20 und *n*=40.
Wie in der rechten Spalte angezeigt, wurde die Berechnung von $\tilde{n}$ je einmal für alle Listen des FBE durchgeführt und je einmal ohne Berücksichtigung der auffälligen Listen 5, 11, 12, 15. Der Parameter c^2 aus der Gleichung 16 ist zusätzlich angegeben.

n	n' [2]	$\tilde{n}$	c^2	Zusammenstellung der Listen für die Ermittlung von $\tilde{n}$
20	29	21,4	0,246	alle 20 Listen
		24,4	0,129	alle Listen ohne 5, 11, 12, 15 (16 Listen)
40	58	43,9	0,223	alle Paarkombinationen aus allen Listen
		49,8	0,114	Paarkombinationen aus Listen außer 5, 11, 12, 15

digkeit halber symmetrisch um die erste Trefferrate angegeben.

Kritische Differenzen im Pegelbereich

Mit dem FBE wird das Sprachverstehen für einen gegebenen Sprachpegel bestimmt und das Konfidenzintervall für die Trefferraten angegeben. Die adaptiven Verfahren wie der Oldenburger Satztest (OLSA, [15]) oder der Göttinger Satztest [16] ermitteln dagegen das Signal-Rausch-Verhältnis oder den Sprachpegel für ein gegebenes Sprachverstehen von zumeist 50% oder auch 80% (Speech Recognition Threshold, SRT). Die Genauigkeit der Satzteste beim SRT wird mit ca. ±1 dB ([17], [18]) angegeben. Zum Vergleich wurden die mit Methode 5 berechneten Konfidenzintervalle für die Trefferrate *p* in Konfidenzintervalle für den Sprachpegel *L* umgerechnet. Dazu wurde die in [18] gegebene Diskriminationsfunktion nach dem Sprachpegel aufgelöst:

Gleichung 18

$$L = L_{50} - \frac{\ln(\frac{100}{p} - 1)}{4 \cdot s_{50}}$$

Für den Pegel L_{50} bei einer Trefferrate von 50% und die Steigung s_{50} in diesem Punkt wurden die in [11] angegebenen medianen Werte L_{50}=24,7 dB und s_{50}=0,045/dB verwendet.

Ergebnisse

Vergleich der Berechnungsmethoden

Abbildung 2 zeigt einen Vergleich der Methoden 1–5 für das 95%-Konfidenzintervall der zweiten Trefferrate p_{mess2} bei gegebenem Ergebnis für die erste Trefferrate p_{mess1}. Die Grenzen nach Methode 1, die auf dem gleichen Wortverstehen für jedes Wort einer Liste beruht, liegen am weitesten außen, geben also das breiteste 95%-Konfidenzintervall an. Durch die Einbeziehung unterschiedlichen Wortverstehens in den Methoden 2 und 3 werden die 95%-Konfidenzintervalle schmaler, die Kurven liegen am weitesten innen. Im letzten Schritt wurde für die Methoden 4 und 5 die Varianz der Testlisten berücksichtigt, so dass die 95%-Konfidenzintervalle wieder weiter außen liegen und nahezu mit Methode 1 zur Deckung kommen. Zwischen den Ergebnissen der Berechnungsvarianten nach [5] und [12] bestehen nur geringe Unterschiede. Dies zeigen die Vergleiche der Grenzen aus den Methoden 2 und 3 sowie aus den Methoden 4 und 5.
Trefferraten des FBE sind bei 20 Wörtern pro Liste nur in Abständen von 5% möglich. Deshalb ist es sinnvoll, die Grenzen der 95%-Konfidenzintervalle konservativ auf Vielfache von 5% zu runden. Diese Grenzen für n=20 sind in Tabelle 3 angegeben. In Tab. A. 1 im Anhang 1 befinden sich die entsprechenden Grenzen für n=40. Durch die Rundungen werden die Unterschiede zwischen den Methoden z.T. vergrößert. Sie betragen jedoch sowohl für n=20 als auch für n=40 höchstens 5%. Die einzige Ausnahme davon ist die Differenz zwischen den Methoden 1 und 3 bei p=75% für die untere Grenze und p=25% für die obere Grenze bei n=20. Die Differenz nimmt hier einen Wert von 10% an.
Für die Methoden 4 und 5 sind in Tabelle 3 und Tab. A. 1 im Anhang 1 zwei Varianten angegeben. Bei der Einbeziehung von allen 20 Listen (Bezeichnungen 4 bzw. 5) wurde $\tilde{n}$=21,4 verwendet (siehe Tabelle 2). Durch Streichen der Listen 5, 11, 12 und 15, d.h. nur mit 16 Listen, erhöht sich die effektive Listenlänge auf $\tilde{n}$=24,4. Die entsprechenden Grenzen sind in den Spalten 4/16 bzw. 5/16 angegeben. Durch das Weglassen der vier Listen reduziert sich die Varianz der Testlisten, so dass die 95%-Konfidenzintervalle etwas schmaler werden.

Vergleich mit Messdaten

Die prozentualen Anteile der Messergebnisse außerhalb der 95%-Konfidenzintervalle sind in Tabelle 1 angegeben. Das Ziel, dass 5% der Messdaten außerhalb des Konfidenzintervalls liegen sollten, wird von Methode 1 sowohl für Normalhörende (NH) als auch für Schwerhörige (SH) und bei Verwendung von 20 oder 40 Wörtern pro Liste annähernd erreicht. Jedoch berücksichtigt Methode 1 weder die Unterschiede im Verstehen der Wörter noch diejenigen zwischen den Testlisten und überschätzt tendenziell die Breite des Konfidenzintervalls. Für die Methoden 2 und 3, die die Unterschiede im Wortverstehen berücksichtigen, liegen ca. 9% der Messwerte außerhalb des 95%-Konfidenzintervalls. Die angegebenen Grenzen sind also zu schmal. Die Methoden 4 und 5 berücksichtigen im Gegensatz zu den Methoden 2 und 3 die Variabilität der Testlisten und erreichen das 5%-Ziel in den verschiedenen Messdatenvarianten für alle 20 Testlisten bis auf eine maximale Abweichung von 0,5% und für die 16 Testlisten bis auf eine maximale Abweichung von 1,1% für Schwerhörige mit Doppellisten.
Abbildung 3 zeigt die Messdaten zusammen mit den kritischen Differenzen nach Methode 5. Für eine Trefferrate von 50% liegt das 95%-Konfidenzintervall zwischen 25% und 75% (siehe Tabelle 3, Spalten „5“). Bei Verwendung von Doppellisten (n=40) reduziert sich das 95%-Konfidenzintervall auf den Bereich zwischen 30% und 70% (siehe Anhang 1 Tab. A. 1, Spalten „5“).

Einseitige Fragestellung

Im Anhang 1 sind in Tab. A. 2 und Tab. A. 3 die gerundeten 90%-Konfidenzintervalle für n=20 und n=40 für alle Methoden angegeben. Den Prozentsatz der Daten außerhalb dieser Konfidenzintervalle für NH und SH für alle Varianten zeigt Tabelle 4. Das Kriterium für die Güte der Berechnungsmethode ist hierbei, dass 10% der Daten außerhalb des berechneten Konfidenzintervalls liegen. Die Ergebnisse entsprechen qualitativ denjenigen in Tabelle 1. Während die Grenzen nach Methode 1 tendenziell zu breit sind, so dass weniger als 10% der Daten außerhalb des 90%-Konfidenzintervalls liegen, fassen die Methoden 2 und 3 das Intervall zu eng. Mit den Methoden 4 und 5 können die Messergebnisse für Normalhörende und Schwerhörige besser als mit den Methoden 2 und 3 angenähert werden.
Abbildung 4 zeigt entsprechend die Messdaten zusammen mit dem 90%-Konfidenzintervall für Methode 5. Für n=20 umfasst das 90%-Konfidenzintervall bei einer Trefferrate von 50% nach Methode 5 den Bereich zwischen 30% und 70% (siehe Anhang 1 Tab. A. 2). Bei Verwendung von Doppellisten (n=40) reduziert sich das 90%-Konfidenzintervall an dieser Stelle auf den Bereich zwischen 35% und 65% (siehe Anhang 1 Tab. A. 3).

Kritische Differenzen im Pegelbereich

Zum Vergleich mit der Genauigkeit von Satztestverfahren sind in Tabelle 5 die Grenzen der Konfidenzintervalle im Pegelbereich bei einem Sprachverstehen von 50% und von 80% für einzelne Listen (n=20) und Doppellisten (n=40) angegeben. Die Konfidenzintervalle sind für n=40 schmaler im Vergleich zu n=20 und für das 90%-Konfidenzintervall schmaler im Vergleich zum 95%-Konfidenzintervall. Bei einem Sprachverstehen von 80% sind die Konfidenzintervalle breiter als bei einem Sprachverstehen von 50%. Die Breite der Konfidenzintervalle reicht von ±4,0 dB für n=40 bei einem Sprachverstehen von 50% (90%-Konfidenzintervall) bis zu ±11,3 dB für n=20 bei einem Sprachverstehen von 80% (95%-Konfidenzintervall).

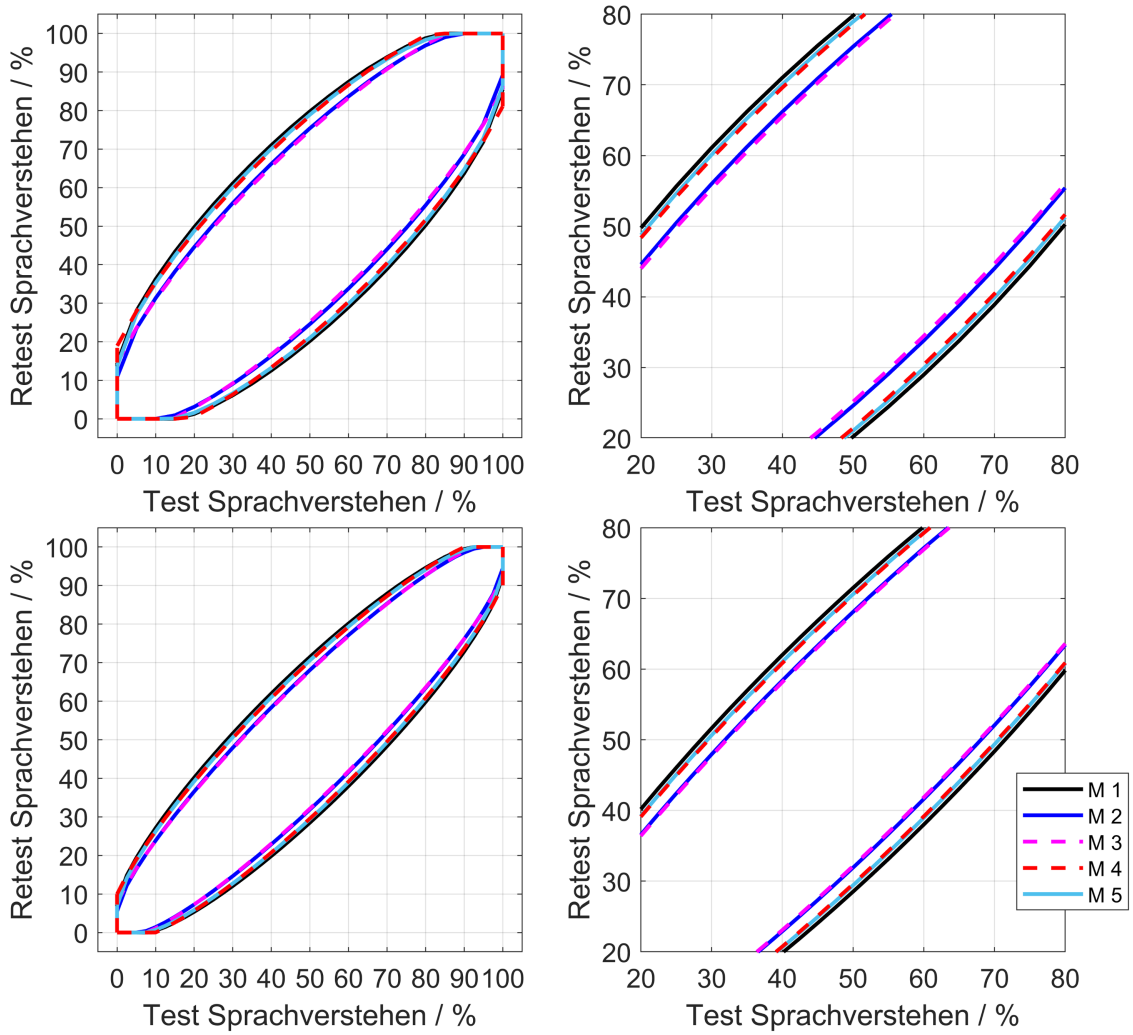

Abbildung 2: Vergleich der 95%-Konfidenzintervalle für die Test-Retest-Reliabilität für Testlisten mit 20 Wörtern (oben) und mit 40 Wörtern (unten) bei Verwendung der fünf Berechnungsmethoden. Rechts ist jeweils der mittlere Bereich der linken Abbildungen für einen besseren Vergleich vergrößert dargestellt. Schwarz: Methode 1, dunkelblau: Methode 2, magenta gestrichelt: Methode 3, rot gestrichelt: Methode 4, hellblau: Methode 5.

Tabelle 5: Mit Gleichung 18 berechnete 95%- und 90%-Konfidenzintervalle in dB SPL für Methode 5 für *n*=20 und *n*=40 bei Trefferraten von 50% und 80%

	$n = 20$		$n = 40$	
Sprachverstehen %	**unten**	**oben**	**unten**	**oben**
95%-Konfidenzintervall				
50	17,3	32,0	19,8	29,6
80	25,0	47,6	27,1	40,1
90%-Konfidenzintervall				
50	18,7	30,7	20,7	28,7
80	26,1	43,3	27,9	38,5

Diskussion

Mit der Annahme eines Bernoulli-Experiments für das Sprachverstehen mit unterschiedlichem Wortverstehen innerhalb der Testlisten wurden mit Hilfe der verallgemeinerten Binomialverteilung die 90%- und die 95%-Konfidenzintervalle modelliert. Die Methoden von Thornton und Raffin [5] und Altman et al. [12] führten dabei zu ähnlichen Ergebnissen. Diese beiden Methoden wurden durch zusätzliche Berücksichtigung der Testlistenvarianz erweitert. Damit erfüllen sie die Kriterien, dass ca. 5% bzw. 10% der Messdaten außerhalb der Grenzen der berechneten Konfidenzintervalle liegen.

Je nach Variante (einzelne Listen oder Doppellisten, 90%- oder 95%-Konfidenzintervall, alle 20 oder nur 16 ausgewählte Testlisten) haben die Konfidenzintervalle bei einer Trefferrate für die erste Messung p_{mess1}=50% eine Breite von ±15% bis ±25%. Die Hilfsmittelrichtlinie [4] fordert eine Verbesserung von mindestens 20 Prozentpunkten für eine Hörgeräteversorgung im Vergleich zur unversorgten Messung. Bei einer Trefferrate von p_{mess1}=50% für die erste Messung ist eine Verbesserung um 20 Prozentpunkte in der zweiten Messung nur bei Nutzung von Doppellisten statistisch signifikant. Bei Verwendung von 20 Wörtern pro Liste ist eine Erhöhung der Trefferrate um 20 Prozentpunkte durch die Hörgeräte statistisch nicht signifikant, da die Irrtumswahrscheinlichkeit für die Entscheidung, dass durch die Hörgeräte das Sprachverstehen verbessert wird, bei mehr als 5% liegt. Damit aus einem Unterschied von 20 Prozentpunkten eine signifikante

Tabelle 3: Grenzen der 95%-Konfidenzintervalle für die Trefferrate der zweiten Testliste p_{mess2} bei gegebener Trefferrate für die erste Testliste p_{mess1} bei n=20. Die Grenzen ergeben sich nach den Methoden 1–5, siehe Text. Die genau berechneten Werte (siehe Abbildung 2) wurden in der Tabelle konservativ auf Vielfache von 5% gerundet.

	95%-Konfidenzintervall p_{mess2} in %													
Grenzen	unten							oben						
Methode p_{mess1}/%	1	2	3	4	4/16	5	5/16	1	2	3	4	4/16	5	5/16
0	0	0	0	0	0	0	0	10	10	10	15	15	10	10
5	0	0	0	0	0	0	0	25	20	20	25	25	25	25
10	0	0	0	0	0	0	0	35	30	30	35	30	35	30
15	5	5	5	0	0	5	5	40	35	35	40	40	40	40
20	5	5	5	5	5	5	5	45	40	40	45	45	45	45
25	5	10	10	5	5	5	5	55	50	45	50	50	50	50
30	10	10	10	10	10	10	10	60	55	55	55	55	60	55
35	10	15	15	10	15	10	15	65	60	60	60	60	65	60
40	15	20	20	15	15	15	15	70	65	65	65	65	70	65
45	20	25	25	20	20	20	20	75	70	70	70	70	70	70
50	25	25	30	25	25	25	25	75	75	70	75	75	75	75
55	25	30	30	30	30	30	30	80	75	75	80	80	80	80
60	30	35	35	35	35	30	35	85	80	80	85	85	85	85
65	35	40	40	40	40	35	40	90	85	85	90	85	90	85
70	40	45	45	45	45	40	45	90	90	90	90	90	90	90
75	45	50	55	50	50	50	50	95	90	90	95	95	95	95
80	55	60	60	55	55	55	55	95	95	95	95	95	95	95
85	60	65	65	60	60	60	60	100	95	95	100	100	95	100
90	65	70	70	65	70	65	70	100	100	100	100	100	100	100
95	75	80	80	75	75	75	75	100	100	100	100	100	100	100
100	90	90	90	85	85	90	90	100	100	100	100	100	100	100

Verbesserung gefolgert werden kann, müsste sowohl die unversorgte als auch die versorgte Kondition mit Doppellisten ermittelt werden. Bei Verwendung von Einzellisten kann erst ab einer Trefferrate für die erste Messung von 75% eine Verbesserung um 20 Prozentpunkte in der zweiten Messung als signifikant unterschiedlich angesehen werden.

Zur Reduktion der Konfidenzgrenzen könnte auf die Nutzung derjenigen vier Testlisten, die in Baljic et al. [11] auffällig waren, verzichtet werden, sodass sich die Testlistenvarianz verringert. Allerdings besteht keine Gewähr dafür, dass bei SH, in anderen deutschsprechenden Regionen oder in anderen Messkonfigurationen (z.B. im Störgeräusch), die gleichen vier Testlisten zu auffällig abweichenden Trefferraten führen. Ein Indiz für Abweichungen in den auffälligen Testlisten könnte sein, dass die aus den Messdaten der 16 ausgewählten Listen für die Gruppe der SH ermittelten Konfidenzintervallgrenzen tendenziell etwas zu weit gefasst sind, so dass geringfügig weniger als die angestrebten 5% bzw. 10% der Messdaten außerhalb der Konfidenzintervalle liegen. Auch bei Verwendung aller 20 Testlisten kann die Testlistenvarianz, die zur Modellierung aus den Messdaten der NH gewonnen wurde, bei verschiedenen Probandengruppen oder Messkonfigurationen unterschiedlich sein und zu schmaleren oder breiteren Konfidenzintervallen führen.

Für die Messdaten der SH konnte jedoch die Aussage von Dillon [9] bestätigt werden, dass SH die gleiche Test-Retest-Reliabilität aufweisen wie Normalhörende.

Der Vergleich der Messergebnisse mit den modellierten Konfidenzgrenzen bestätigt ebenfalls die Schlussfolgerung von Dillon [9], dass die Grenzen von Thornton und Raffin [5] nach Methode 1, also bei Verwendung der einfachen Binomialverteilung, relativ gut die gemessene Test-Retest-Reliabilität nachbilden können. Diese Grenzen wurden bereits für den FBE von Winkler und Holube [7] für n=20 angegeben. Durch die Verwendung der allgemeinen Binomialverteilung bei den Methoden 2 und 3 werden die Konfidenzintervalle schmaler, nach Berücksichtigung der Testlistenvarianz bei den Methoden 4 und 5 jedoch wieder breiter, so dass annähernd die Grenzen von Methode 1 erreicht werden. Dabei ist jedoch zu berücksichtigen, dass die von Dillon [9] diskutierte Variabilität zwischen den Probanden in der vorliegenden Untersuchung nicht integriert wurde. Ein möglicher Grund für die Vernachlässigbarkeit der Probandenvarianz könnte der Vergleich mit Wiederholungsmessungen zum gleichen Termin sein, so dass nur die Kurzzeit-Reliabilität für Test und Retest überprüft wurde. Diese vermutlich kleine intraindividuelle Varianz der Probanden innerhalb eines Termins liegt möglicherweise unterhalb der Testlistenvarianz, so dass sie hier vernachlässigt werden kann. Nicht unter-

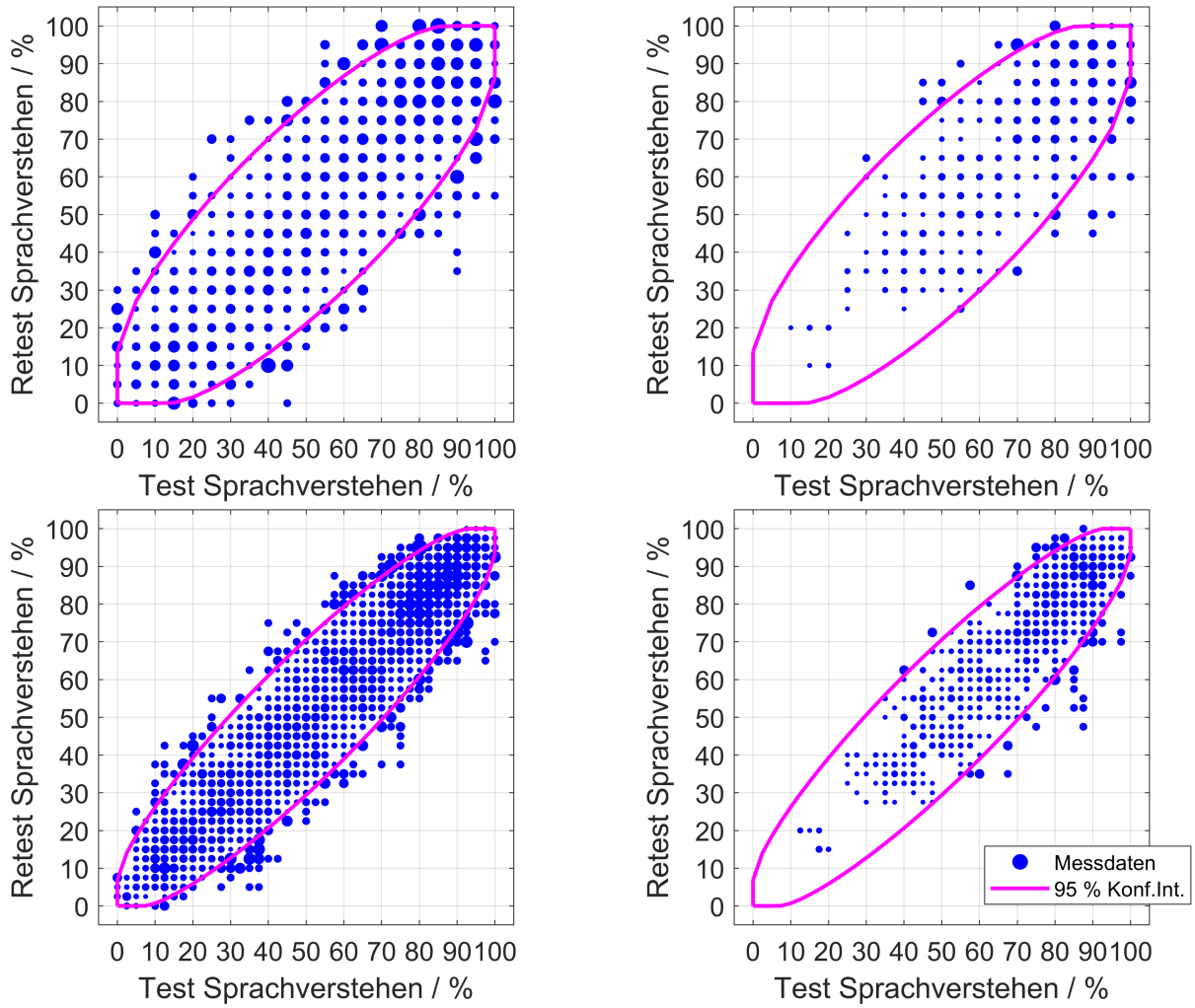

Abbildung 3: Datenpunkte (blau) und 95%-Konfidenzintervall (magenta) für Testlisten mit 20 Wörtern (oben) und mit 40 Wörtern (unten) für Normalhörende (links) und Schwerhörige (rechts) bei Verwendung von Methode 5 und zweiseitige Fragestellung.

Tabelle 4: Anzahl der verwendeten Datenpunkte und Prozentsatz der Daten außerhalb des 90%-Konfidenzintervalls
Angaben für Normalhörende (NH) und Schwerhörige (SH) mit *n*=20 und *n*=40 Wörtern pro Liste für die Methoden 1–5

		Methoden						
		1	**2**	**3**	**4**	**4/16**	**5**	**5/16**
NH $n = 20$	Datenpunkte	3.200	3.200	3.200	3.200	2.034	3.200	2.034
	%	8,9	17,1	17,0	11,0	9,5	9,4	8,5
SH $n = 20$	Datenpunkte	800	800	800	800	507	800	507
	%	9,1	16,6	16,6	9,9	8,9	9,3	8,1
NH $n = 40$	Datenpunkte	4.800	4.800	4.800	4.800	1.890	4.800	1.890
	%	8,8	16,3	16,4	11,3	9,0	10,7	8,4
SH $n = 40$	Datenpunkte	1.200	1.200	1.200	1.200	465	1.200	465
	%	9,3	15,3	15,4	11,6	11,6	11,1	11,4

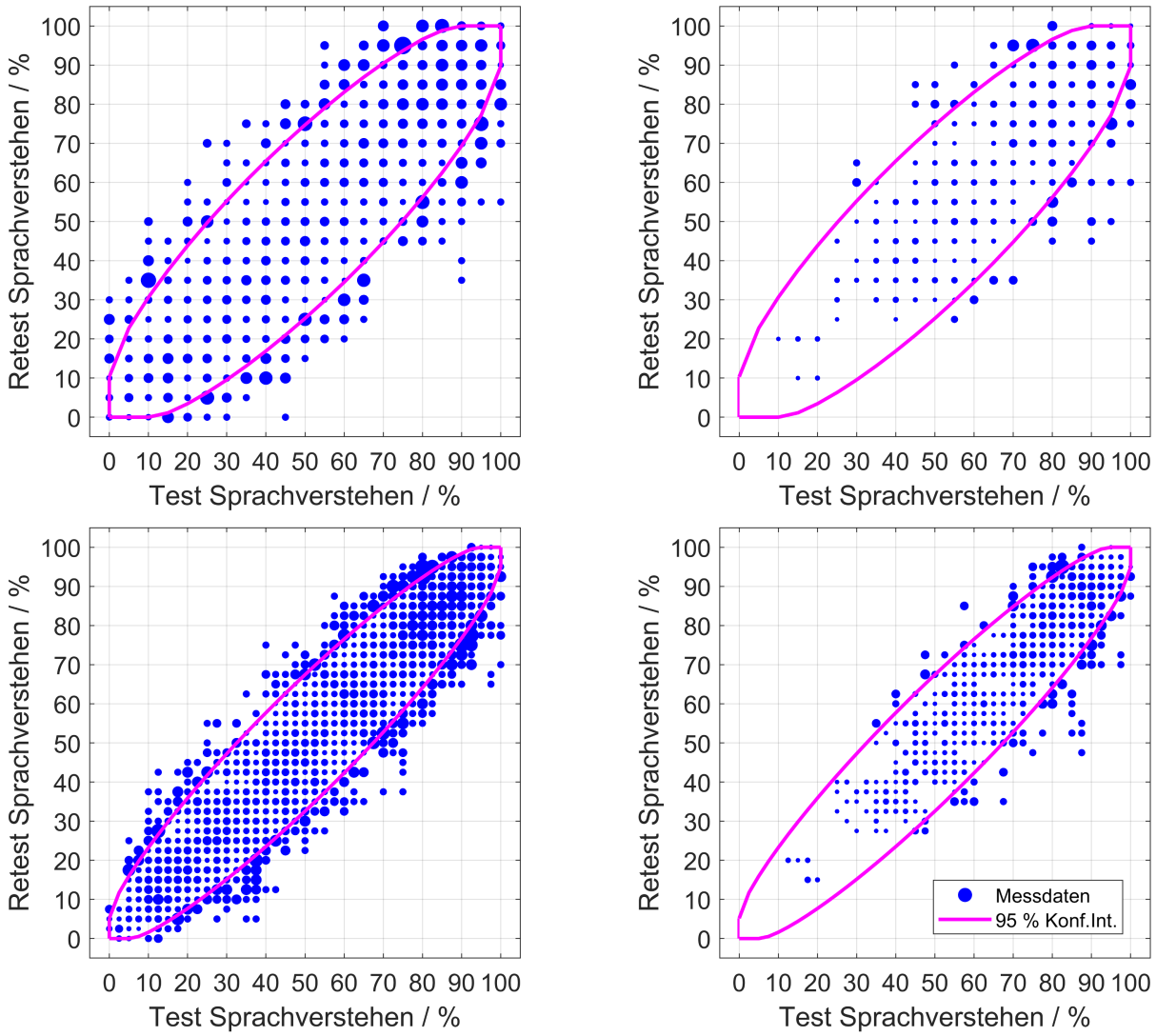

Abbildung 4: Datenpunkte (blau) und 90%-Konfidenzintervall (magenta) für Testlisten mit 20 Wörtern (oben) und mit 40 Wörtern (unten) für Normalhörende (links) und Schwerhörige (rechts) bei Verwendung von Methode 5 und einseitige Fragestellung.

sucht wurde die Reliabilität über einen längeren Zeitraum, d.h. über mehrere Termine, die sich durch die variable Tagesform der Probanden ändern könnte. Ein anderer Erklärungsansatz für die Vernachlässigbarkeit der Probandenvarianz könnte darin liegen, dass individuelle Unterschiede nicht genügend berücksichtigt wurden [9]: Zur Modellierung der verallgemeinerten Binomialverteilung wurden nur die Mittelwerte im Sprachverstehen für die einzelnen Wörter verwendet. Für einzelne Probanden kann sich das Sprachverstehen der Wörter noch deutlicher unterscheiden, so dass die Methoden 2 und 3 zu noch schmaleren Konfidenzintervallen führen würden. Dann wäre eine zusätzliche Varianzquelle, z.B. die intraindividuelle Varianz, notwendig, um die zu den Messdaten passenden Konfidenzintervalle zu modellieren.

Zum Vergleich mit den Satztestverfahren wurden die Konfidenzintervalle von Methode 5 für Trefferraten von 50% und 80% in Konfidenzintervalle für den Sprachpegel transformiert. Bei Verwendung von Einzellisten (n=20) bei einem Sprachverstehen von 50% hat das 90%-Konfidenzintervall eine Breite von ±6 dB. Das Konfidenzintervall für den FBE ist damit wesentlich breiter als die Konfidenzintervalle für die adaptiven Satzteste mit ca. ±1 dB ([17], [18]). Hörgeräte müssten den Sprachpegel für ein Sprachverstehen von 50% um mehr als 6 dB verbessern, um einen signifikanten Effekt zu erzielen. Wenn das Hörgerät den Sprachpegel z.B. nur um 3 dB verbessern würde, dann würden die Satzteste zwar zu einem signifikanten Unterschied und damit zu einem Wirksamkeitsunterschied führen, jedoch nicht der FBE. Dieses Ziel von einer Verbesserung um mehr als 6 dB erscheint für das Sprachverstehen in Ruhe leicht erreichbar. Ob jedoch diese Anforderung auf eine Verbesserung von 6 dB im Signal-Rausch-Verhältnis für den FBE im Störgeräusch übertragen werden kann, ist noch ungeklärt. Im Störgeräusch werden zwar die gleichen Listen mit n=20 bzw. n=40 Wörtern verwendet, die Varianz im Wortverstehen und im Listenverstehen kann sich jedoch von dem FBE in Ruhe unterscheiden, so dass sich abweichende Konfidenzgrenzen ergeben können.

Schlussfolgerungen

- Kritische Differenzen können allein aus der Anzahl der Messitems mit Methode 1 von Thornton und Raffin relativ gut abgeschätzt werden.
- Bei weiteren Kenntnissen über den Sprachtest zur Verteilung des Verstehens einzelner Items und der Varianz der Testlisten bieten die Methoden 4 und 5 eine genauere Modellierung der Test-Retest-Reliabilität.
- Bei Publikation von Sprachtestergebnissen sollten die Konfidenzintervallgrenzen immer mit angegeben werden. Dabei ist zu beachten, ob es sich um eine einseitige oder eine zweiseitige Fragestellung handelt.

Anmerkungen

Interessenkonflikte

Die Autoren erklären, dass sie keine Interessenkonflikte in Zusammenhang mit diesem Artikel haben.

Danksagung

Die Untersuchungen wurden vom Promotionsprogramm Jade2Pro der Jade Hochschule sowie aus dem Projekt VIBHear mit Mitteln des Europäischen Fonds für regionale Entwicklung (EFRE) und Mitteln des Landes Niedersachsen gefördert.

Anhänge

Verfügbar unter
https://www.egms.de/de/journals/zaud/2020-2/zaud000007.shtml

1. Anhang_1.pdf (183 KB)
   Anhang: Konfidenzintervalle

Literatur

1. Hahlbrock KH. Uber Sprachaudiometrie und neue Wörterteste [Speech audiometry and new word-tests]. Arch Ohren Nasen Kehlkopfheilkd. 1953;162(5):394-431. DOI: 10.1007/BF02105664
2. 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.
3. 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
4. Gemeinsamer Bundesausschuss. Richtlinie des gemeinsamen Bundesausschusses über die Verordnung von Hilfsmitteln in der vertragsärztlichen Versorgung. Hilfsmittelrichtlinie. 2018 [accessed 13. Dezember 2018]. Available from https://www.g-ba.de/downloads/62-492-1666/HilfsM-RL_2018-07-19_iK-2018-10-03.pdf
5. 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
6. Carney E, Schlauch RS. Critical difference table for word recognition testing derived using computer simulation. J Speech Lang Hear Res. 2007 Oct;50(5):1203-9. DOI: 10.1044/1092-4388(2007/084)
7. 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
8. Steffens T. Test-Retest-Differenz der Regensburger Variante des OLKI-Reimtests im sprachsimulierenden Störgeräusch bei Kindern mit Hörgeräten. Z Audiol. 2006;45(3):88-99.
9. Dillon H. A quantitative examination of the sources of speech discrimination test score variability. Ear Hear. 1982 Mar-Apr;3(2):51-8. DOI: 10.1097/00003446-198203000-00001
10. Hagerman B. Reliability in the determination of speech discrimination. Scand Audiol. 1976;5:219-28. DOI: 10.3109/01050397609044991
11. 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
12. Newcombe RG, Altman DG. Proportions and Their Differences. In: Altman DG, Machin D, Bryant TN, Gardner MJ, editors. Statistics with Confidence: Confidence Intervals and Statistical Guidelines. 2nd Edition. London: British Medical Journal Books; 2000. p. 45-56.
13. Newcombe RG. Interval estimation for the difference between independent proportions: comparison of eleven methods. Stat Med. 1998 Apr;17(8):873-90. DOI: 10.1002/(sici)1097-0258(19980430)17:8<873::aid-sim779>3.0.co;2-i
14. Wilson EB. Probable Inference, the Law of Succession, and Statistical Interference. J Am Stat Assoc. 1927;22(158):209-12. DOI: 10.1080/01621459.1927.10502953
15. Wagener KC, Kühnel V, Kollmeier B. Entwicklung und Evaluation eines Satztests für die deutsche Sprache I: Design des Oldenburger Satztests. Z Audiol. 1999a;38:4-15.
16. Kollmeier B, Wesselkamp M. Development and evaluation of a German sentence test for objective and subjective speech intelligibility assessment. J Acoust Soc Am. 1997 Oct;102(4):2412-21. DOI: 10.1121/1.419624
17. Wagener KC, Brand T. Sentence intelligibility in noise for listeners with normal hearing and hearing impairment: influence of measurement procedure and masking parameters. Int J Audiol. 2005 Mar;44(3):144-56. DOI: 10.1080/14992020500057517
18. Brand T, Kollmeier B. Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests. J Acoust Soc Am. 2002 Jun;111(6):2801-10. DOI: 10.1121/1.1479152

Korrespondenzadresse:
Prof. Dr. Inga Holube
Jade Hochschule, Institut für Hörtechnik und Audiologie, Ofener Str. 16/19, 26121 Oldenburg, Deutschland, Tel. +49-441-7708-3723
Inga.Holube@jade-hs.de

Bitte zitieren als
Holube I, Winkler A, Nolte-Holube R. Modellierung und Verifizierung der Test-Retest-Reliabilität des Freiburger Einsilbertests in Ruhe mit der verallgemeinerten Binomialverteilung. GMS Z Audiol (Audiol Acoust). 2020;2:Doc03.
DOI: 10.3205/zaud000007, URN: urn:nbn:de:0183-zaud0000070

Artikel online frei zugänglich unter
https://www.egms.de/en/journals/zaud/2020-2/zaud000007.shtml

Veröffentlicht: *27.03.2020*

OPEN ACCESS

This is the English version.
The German version starts at p. 1.

Research Article

Modeling and verifying the test-retest reliability of the Freiburg monosyllabic speech test in quiet with the Poisson binomial distribution

Abstract

The test-retest reliability of the Freiburg monosyllabic speech test was modeled using different methods. The results were compared to measurements from listeners with and without hearing impairment. The methods are based on the models of Thornton and Raffin as well as Altman et al. Both papers took into account differences in word recognition within the test lists by applying the Poisson binomial distribution and included the variance of the test-list results. The methods allow calculating the bounds of the 90% and 95% confidence intervals when using test lists with 20 words and double lists with 40 words. The data in the current report confirm these bounds. The confidence intervals are broadest for speech recognition scores of 50%. At this score and for test lists with 20 words, the 90% confidence interval has a width of ±20%, corresponding to ±6.0 dB, and the 95% confidence interval has a width of ±25%, corresponding to ±7.4 dB. Thus when evaluating hearing-aid fittings, only differences exceeding this range can be regarded as significantly different.

Keywords: Freiburg monosyllabic test, speech intelligibility, binomial distribution, test-retest 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

In issue 1/2018, modeling of the reliability of the Freiburg monosyllabic test (FBE) [1] in quiet with the Poisson binomial distribution was presented [2], [3]. The use of this distribution allows attention to differences in word recognition within a test list. This results in a smaller confidence interval for the measurement results than when using the simple binomial distribution that assumes the same probability of recognition for each word in a list. The variance of the Poisson binomial distribution for 20-word test lists could be approximated by the variance of a simple binomial distribution based on 29-word test lists with the same degree of word recognition.

The studies in Holube et al. [2], [3] were limited to the calculation of the 95% confidence interval for the deviation from the true value of the measured value for a test list and, alternatively, for the deviation of the true value from the measured value for a test list. However, the published confidence intervals are not applicable for estimating test-retest reliability or to studies with two test lists used to compare two measurement conditions. Exactly this case exists when verifying hearing aids or other therapeutic treatments. The results of two measurements (e.g., with and without hearing aids), i.e. two scores, are compared, and the success of the treatment is derived from the difference between the two scores. The guideline for assistive devices *(Hilfsmittelrichtlinie* in German) [4] requires, e.g., for the FBE in quiet, an improvement in speech recognition of at least 20 percentage points with hearing aids as compared to the condition without hearing aids.

Thornton and Raffin [5] calculated the 95% confidence interval for the difference between two measurements by transforming the scores into a scale with homogeneous variance for all test results and then adding the variances of the two test results. Carney and Schlauch [6] essentially confirmed the results of this method using a different approach. They calculated the variance of the difference between two scores assuming binomially distributed scores. For each value for the score from the first measurement, they considered all possible score values for the second measurement. Results using the method of Thornton and Raffin [5], which requires the same recognition probability for all 20 words of a test list, were given by Winkler and Holube [7] based on Steffens [8] and compared with results of repeated measurements.

On the one hand, Dillon [9] argued that if test lists are equally recognizable and the listeners always behave similarly, and assuming the same recognition probability for each word, the width of the 95% confidence interval for the test-retest condition is overestimated when using the method of Thornton and Raffin [5]. This assumption is supported by the analysis in Winkler and Holube [7] since only 3.2% of the measurement data, i.e. less than the expected 5%, were outside the confidence interval

according to Thornton and Raffin [5]. On the other hand, Dillon [9] pointed out that Thornton and Raffin's [5] method can nevertheless be used to estimate the 95% confidence interval, since two effects cancel each other out: Considering different word recognition and applying the Poisson binomial distribution according to Hagerman [10], the 95% confidence intervals become narrower. By intra-individual variability (e.g., by attention fluctuations), especially with a longer time interval between the measurements, they become wider again. As an additional source of variance, Dillon [9] pointed out possible differences among test lists. In speech audiometry, in contrast to Winkler and Holube [7], the same test lists are generally not used in repeated measurements. The 95% confidence interval for the test-retest reliability widens when using different test lists, due to the different mean scores of the test lists.

For the present analysis, measurements from Baljic et al. [11] and Holube et al. [2], [3], which for each subject included the results of five test lists at each of four levels, were interpreted in terms of a test-retest experiment, and the test-retest reliability was evaluated. All measurements were performed within one session. Therefore, only the short-term test-retest reliability was investigated, but not the test-retest reliability over a longer period of time that according to Dillon [9], would probably result in broader confidence intervals. For comparison with the measurement data, the bounds for the 95% and the 90% confidence intervals were modeled using different methods. The methods are based on the Poisson binomial distribution used in Holube et al. [2], [3]. Additionally, the variability of the test lists was modeled. Intra-individual variances of the participants were neglected due to the short time intervals between the measurements.

Methods

Experimental data

The measurement methods are summarized here only briefly. For a detailed description refer to Holube et al. [2], [3].

In 80 young participants having normal hearing abilities (hereinafter named normal-hearing participants), speech recognition was determined as the percentage score for the Freiburg monosyllables in quiet and at four levels (17.5, 23.5, 29.5, and 35.5 dB SPL), with each of five test lists comprising 20 words (n=20). In 40 older participants with hearing impairment (hereinafter named hearing-impaired participants), the levels 65, 80, 90, and 95 dB SPL were used in the same procedure. However, only 65 and 80 dB SPL were included in the analysis, because at the two higher levels, many scores achieved 100%. All measurements of a given participant were performed within one session.

The five fixed-level test-list hit rates for each participant were interpreted as test-retest combinations in pairs. The pairs each consisted of a presented test list and another, subsequently presented, list, i.e. (1; 2), (1; 3), (1; 4), (1; 5), (2; 3), (2; 4), (2; 5), (3; 4), (3; 5), (4; 5). This resulted in 3,200 test-retest pairs for the normal-hearing and 800 test-retest pairs for the hearing-impaired participants. The number of test-retest pairs decreased when the conspicuous test lists of Baljic et al. [11] were excluded (see Table 1). In another variant, two test lists each with double lists of n=40 words were formed. For the analysis of test-retest reliability, all double lists were combined into test-retest pairs so that no single list was duplicated, i.e. (1+2; 3+4), (1+2; 3+5), (1+2; 4+5), (1+3; 2+4), (1+3; 2+5), (1+3; 4+5), (1+4; 2+3), (1+4; 2+5), (1+4; 3+5), (1+5; 2+3), (1+5; 2+4), (1+5; 3+4), (2+3; 4+5), (2+4; 3+5), and (2+5; 3+4). This resulted in 4,800 test-retest pairs for the normal-hearing and 1,200 test-retest pairs for the hearing-impaired participants when all 20 test lists were used. As a variant, the conspicuous test lists of [11] were also excluded for these double lists (see Table 1).

Calculation methods

For a given percentage score p_{mess1} (test), the question was: In which critical range did the retest percentage score p_{mess2} lie, so that the difference $p_{mess1}-p_{mess2}$ for a two-sided test was not significantly different from zero at the α=5% level. A two-sided test means that the retest score may be less than or greater than the first score. Thus 2.5% of the retest scores are below and 2.5% of the retest scores are above the 95% confidence interval around the first score. Different methods exist in the literature for calculating the 95% confidence interval, two of which (Thornton and Raffin [5] and Altman et al. [12]) are considered in the current contribution. Both methods were first reproduced and then applied to the available measurement data with n=20 and n=40 words per test list (i.e. simple test lists and double lists). Afterwards, modifications of these methods are presented that took into account the variability of single word recognition, as well as the variability of the mean recognition of different test lists.

Method 1: Critical differences according to Thornton and Raffin

Thornton and Raffin [5] proposed calculating a 95% confidence interval for the assessment of test-retest reliability by the following method: The number X of correct responses for n presented words in a test list is considered to be a random variable. It is assumed to be binomially distributed with $B(n,p,X=k)$. Here p is the probability that one word in the list will be correctly recognized. Here and below, probabilities are given in percent. The expected value of X is thus $\mathrm{E}(X)=\frac{np}{100}$. Speech recognition in percent (score) is the random variable $p_{\mathrm{mess}}=\frac{100X}{n}$. Its expected value is $\mathrm{E}(p_{mess})=p$ and its variance is $\mathrm{Var}(p_{\mathrm{mess}})=\sigma^2=\frac{p(100-p)}{n}$. This variance reaches its

Table 1: Number of data points and percentage of data outside the calculated 95%-confidence intervals
Results for normal-hearing (NH) and hearing-impaired (HI) participants with *n*=20 and **n**=40 words per test list for the methods 1–5

		Methods						
		1	**2**	**3**	**4**	**4/16**	**5**	**5/16**
NH $n = 20$	Data points	3,200	3,200	3,200	3,200	2,034	3,200	2,034
	%	4.8	8.7	9.4	5.4	4.5	5.4	4.4
HI $n = 20$	Data points	800	800	800	800	507	800	507
	%	4.9	8.9	9.3	5.3	5.1	5.0	4.7
NH $n = 40$	Data points	4,800	4,800	4,800	4,800	1,890	4,800	1,890
	%	4.6	9.2	9.0	5.1	4.3	5.1	4.2
HI $n = 40$	Data points	1,200	1,200	1,200	1,200	465	1,200	465
	%	3.9	9.5	8.9	4.5	3.9	4.7	3.9

maximum at p=50%. At the borders p=0 and p=100%, the variance is zero.
For the test-retest reliability, estimating a confidence interval for the difference p_{mess1}–p_{mess2} of two scores is of interest. For this purpose, the random variables X_1 and X_2 are first transformed (according to Equation 3 in [5]) using Equation 1 to an angle $\theta(X,n)$.

Equation 1

$$\theta(X,n) = \sin^{-1}\sqrt{\frac{X}{n+1}} + \sin^{-1}\sqrt{\frac{X+1}{n+1}}$$

The random variable θ thus defined has approximately a variance Var(θ) that is independent of p. Thornton and Raffin [5] chose the approximations $\mathrm{Var}(\theta) = \frac{1}{n+\frac{1}{2}}$ for n≥50 and $\mathrm{Var}(\theta) = \frac{1}{n+1}$ for 10<n<50. The two random variables $\theta_1=\theta(X_1,n)$ and $\theta_2=\theta(X_2,n)$ have the same variance Var(θ) within this approximation. Assuming that θ_1 and θ_2 are statistically independent, the variance of the random variable $\Delta\theta=\theta_1-\theta_2$ is the sum of the variances, i.e. Var($\Delta\theta$)=2Var(θ). For $\Delta\theta$, a normal distribution with the variance 2Var(θ) is assumed. The 95% confidence interval for θ_2 at the score p_{mess1} thus results in $\theta_1 \pm 1{,}96\sqrt{\mathrm{Var}(\Delta\theta)}$. The thus calculated θ_2 bounds of the 95% confidence interval are transformed back to X_2 bounds to obtain the p_{mess2} bounds. Thus, if $p_{\mathrm{mess1}} = \frac{100X_1}{n}$ and $p_{\mathrm{mess2}} = \frac{100X_2}{n}$ indicate the scores in the test and in the retest measurement, respectively, this method can be summarized as follows:

Equation 2

$$p_{\mathrm{mess2,\,lower\,bound}} = \frac{100}{n}\theta^{-1}\left(\theta(X_1,n) - 1{,}96\cdot\sqrt{\mathrm{Var}(\Delta\theta)},n\right)$$

Equation 3

$$p_{\mathrm{mess2,\,upper\,bound}} = \frac{100}{n}\theta^{-1}\left(\theta(X_1,n) + 1{,}96\cdot\sqrt{\mathrm{Var}(\Delta\theta)},n\right)$$

with

Equation 4

$$\mathrm{Var}(\Delta\theta) = \frac{2}{n+1}.$$

These bounds were calculated for all scores p_{mess1} of interest between 0 and 100%. The inverse function $X=\theta^{-1}(\theta,n)$ of Equation 1 was calculated numerically.

Method 2: Critical differences according to Thornton und Raffin, with variable word recognition

If the individual words of a test list are recognized differently, the same recognition probability p for each word is no longer sufficient for the description. Each word has its own recognition probability, and the binomial distribution is replaced by the Poisson binomial distribution [10]. In order to consider the narrowing of the distribution of X in the Poisson binomial distribution relative to the simple binomial distribution, the variance of θ is now set to $\mathrm{Var}(\theta) = \frac{1}{n'+1}$ instead of $\frac{1}{n+1}$. The value for n' is taken from [2], [3], hence, n'=29 for n=20 and n'=58 for n=40. Thus, method 2 is described by Equation 2 and Equation 3 together with

Equation 5

$$\mathrm{Var}(\Delta\theta) = \frac{2}{n'+1}$$

instead of Equation 4.

Method 3: Critical differences according to Altmann et al., with variable word recognition

Altman et al. [12] recommended an approach that corresponds to method 10 of [13]. This method will initially be presented unchanged. Then it is modified to take into account the variability of word recognition within a test list.

If a percentage score p_{mess} for a single test list was measured, the 95% confidence interval for the true value p is in question. Wilson [14] made the following approach:

Equation 6

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

with z=1,96. This is a quadratic equation for p. Its solutions u and o specify the lower and upper bounds, respectively, of the required confidence interval (see [2], [3]). If there are two hit rates p_{mess1} and p_{mess2}, then the associated lower bounds u_1 and u_2 and the upper bounds o_1 and o_2 result. According to [12], the significance of the difference p_{mess1}–p_{mess2} is assessed as follows: If the first score p_{mess1} is greater than the second score p_{mess2}, then to be significantly different at the 5% level, the difference p_{mess1}–p_{mess2} of the two scores must be larger than

Equation 7

$$\delta_u = \sqrt{(p_{\mathrm{mess1}} - u_1)^2 + (p_{\mathrm{mess2}} - o_2)^2}\ .$$

To calculate the 95% confidence interval for the difference between the two scores, the variance for the higher score is added downwards and the variance for the lower score is added upwards. For the other case, namely that the second score is larger than the first score, the difference p_{mess2}–p_{mess1} must be correspondingly larger than

Equation 8

$$\delta_o = \sqrt{(p_{\mathrm{mess1}} - o_1)^2 + (p_{\mathrm{mess2}} - u_2)^2}\ .$$

For each of the values of p_{mess1} of interest between 0 and 100%, this method provides a 95% confidence interval for the difference p_{mess2}–p_{mess1}. For a given p_{mess1} (test), the score p_{mess2} (retest) lies with a probability of 95% between p_{mess1}–δ_u and p_{mess1}+δ_o. The six equations, i.e. the equations for u_1, u_2, o_1, and o_2 and the Equation 7 and Equation 8, must be solved for a given p_{mess1}. There is no closed solution. Therefore, the equations were solved numerically by fixed point iteration.
The calculation method described so far is based on the same single-word recognition within a test list. The variability of the single-word recognition leads to a reduction of the variance $\frac{p(100-p)}{n}$ on the right side of Equation 6, as already described for method 2. This is now taken into account by replacing n by n'. The value for n' is taken again from [2], [3], i.e. n'=29 instead of n=20 and n'=58 instead of n=40.

Method 4: Critical differences according to Altmann et al., with variable word recognition and variable test list recognition

Starting from variable single-word recognition under the same conditions, in a speech test, the mean scores of the lists vary due to different word compositions of lists. If the test-list mean value for each test list could be exactly determined under given measurement conditions, this mean value would have a variance f_n^2. This depends on the number n of words per test list and on the true value p. This variance contributes to the uncertainty of the true value of p in Equation 6. Thus, taking into account both variable single-word recognition (replacing n by n') and variable test-list recognition (adding f_n^2 to the variance of p), Equation 6 becomes:

Equation 9

$$(p - p_{\mathrm{mess}})^2 = z^2 \left(\frac{p(100-p)}{n'} + f_n^2 \right)$$

with z=1,96. If the variance f_n^2 is known, the further steps of the method according to [12], as described for method 3, can be carried out.
To determine f_n^2, the sample variance of the measured test list mean values is calculated. n_L test lists of n words are considered with the single-word recognition p_{ji}, i=1...n, j=1...n_L. The percentage score of the test list j is thus the mean value $p_j = \frac{1}{n}\sum_{i=1}^{n} p_{ji}$. With the scores averaged over all words in all test lists

Equation 10

$$\bar{p} = \frac{1}{n_L}\sum_{j=1}^{n_L} p_j\ ,$$

the sample variance f_n^2 of the test list means is:

Equation 11

$$f_n^2 = \frac{1}{n_L - 1}\sum_{j=1}^{n_L} (p_j - \bar{p})^2\ .$$

The variance of single-word recognition is

Equation 12

$$f_1^2 = \frac{1}{n \cdot n_L - 1}\sum_{j=1}^{n_L}\sum_{i=1}^{n} (p_{ji} - \bar{p})^2\ .$$

The relationship between the variance of the recognition of a single word and the variance of the mean value of n words randomly assembled into test lists is

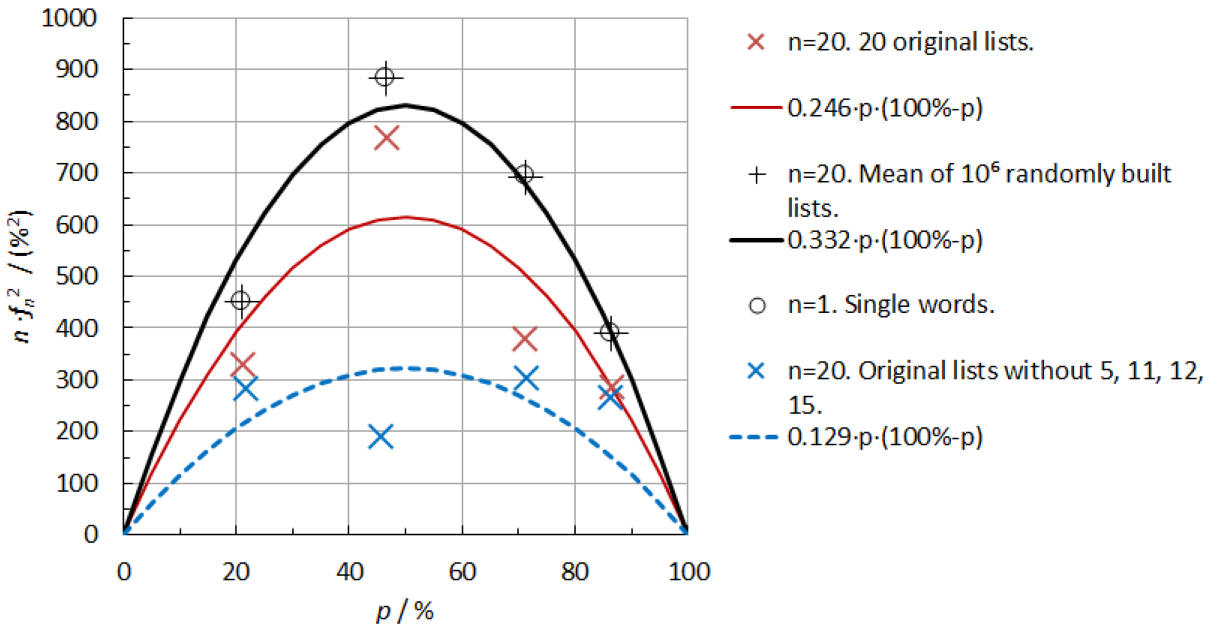

Figure 1: Variance of the test-list mean values as a function of the score p. To improve comparability, the variances were multiplied by the number of words *n* per test list (see Equation 13). The symbols show the variances for different combinations of words in test lists. For each combination, the mean score (10) of all words involved is used as the abscissa *p*. Since the scores belong to four different sound pressure levels, the symbols are grouped around the four respective mean scores *p*. The plotted lines show the fitted parabolas according to Equation 14 To relate the values for c^2 to the combinations, see Table 2. The value c^2=0.332 results from fitting to the variance of single-word recognition scores (*n*=1) of all words.

Equation 13

$$f_{n,\mathrm{random}}^2 = \frac{1}{n} f_1^2 .$$

Figure 1 shows that this relationship is satisfied for randomly composed test lists with *n*=1, 20, 40 from the words of the FBE. The variances shown were averaged out of 10^6 realizations of randomly assembled test lists. However, the variances of the specific test lists of the FBE deviate significantly from the average result of a random combination of words. In addition, Figure 1 shows, as expected, that f_n^2 is smaller in the vicinity of *p*=0 (almost no word is understood) and *p*=100% (almost all words are understood) than it is in the middle range around *p*=50%. The exact dependence of f_n^2 as a function of p is unknown. The approach chosen here is a parabola

Equation 14

$$f_n^2 = \frac{c^2}{n} p(100 - p),$$

with a parameter c^2 to be determined. Thus Equation 9 can be written as

Equation 15

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

with

Equation 16

$$\frac{1}{\tilde{n}} = \frac{1}{n'} + \frac{c^2}{n} .$$

If, in method 3, Equation 6 is replaced by Equation 15, then both the variability of the single-word recognition and the variability of the test-list mean values are taken into account.

The parameter c^2 was calculated from the measured single-word recognition p_{ji} as follows. For each of the four levels used, the average $\bar{p}$ of single-word recognition was calculated according to Equation 10 and the variance of f_n^2 according to Equation 11. The values of the four pairs $(\bar{p}, f_n^2)$ depend on the selection of the test lists and the word composition of the test lists and on the test-list length *n*. For the four pairs of values $(\bar{p}, f_n^2)$, a parabola $f_n^2 = \frac{c^2}{n} p(100 - p)$ was fitted according to the method of least squares. This yielded the value for c^2. Three of the resulting parabolas are shown in Figure 1. With the now-known value of c^2, the effective list length $\tilde{n}$ was calculated using Equation 16. Table 2 shows the results for *n*=20 and for *n*=40. Since the FBE has 20 words per list, for calculations with *n*=40, all combinations of pairs of different lists were considered.

Table 2: Effective number of words calculated from the recognition of single words for the test-list lengths *n*=20 and *n*=40
As shown in the right column, $\tilde{n}$ was calculated for all test lists of the Freiburg monosyllabic speech test and, additionally, with the exclusion of the outlier test lists 5, 11, 12, and 15. The parameter c^2 from Equation 16 is also given.

n	n' [2]	$\tilde{n}$	c^2	Compilation of the lists to determine $\tilde{n}$
20	29	21.4	0.246	all 20 test lists
		24.4	0.129	all test lists without 5, 11, 12, 15 (16 remaining test lists)
40	58	43.9	0.223	all paired combinations out of all test lists
		49.8	0.114	all paired combinations without test lists 5, 11, 12, 15

Method 5: Critical differences according to Thornton and Raffin, with variable word recognition and variable test-list recognition

To incorporate single-word variability, the variance in Equation 4 decreases to that in Equation 5. Consequently, the variability of test-list recognition is now included by replacing Equation 5 with

Equation 17

$$\mathrm{Var}(\Delta\theta) = \frac{2}{\tilde{n}+1}.$$

Critical differences in a one-sided test

So far, the 95% confidence interval has been considered for two-sided tests. However, when using the FBE in hearing-aid fitting, it is assumed that hearing aids improve speech recognition, i.e. that a higher score is achieved in the second measurement (with hearing aids) than in the first measurement (without hearing aids).
The statistical test for determining a significant difference between the two scores would then examine whether the error probability for the hypothesis that the second score is larger than the first score is less than 5%. This corresponds to the bounds of the 90% confidence interval and can be calculated using the same five methods by replacing z=1,96 with z=1,645. Although the problem is one-sided, for the sake of completeness, the limits of the 90% confidence interval for the second score are given symmetrically around the first score.

Critical differences in the level domain

The FBE determines speech recognition for a given speech level. Its accuracy is provided by the corresponding confidence interval for percentage scores. In contrast, adaptive methods such as the Oldenburg sentence test (OLSA, [15]) or the Göttingen sentence test [16] determine the signal-to-noise ratio or speech level for a given speech recognition score of (mostly) 50%, or even 80% (Speech Recognition Threshold, SRT). The accuracy of the sentence tests in the SRT is given as approx. ±1 dB ([17], [18]). For comparison, the confidence intervals for the percentage score p obtained from method 5 were converted into confidence intervals for the speech level L. For this purpose, the discrimination function given in [18] was solved for the speech level:

Equation 18

$$L = L_{50} - \frac{\ln(\frac{100}{p} - 1)}{4 \cdot s_{50}}.$$

For the level L_{50} at p=50% and the slope s_{50} at this point, the median values L_{50}=24.7 dB and s_{50}=0.045/dB given in [11] were used.

Results

Comparison of calculation methods

In a comparison of the results from methods 1–5, Figure 2 shows the 95% confidence interval of the second percentage score p_{mess2}, given the value for the first percentage score p_{mess1}. The bounds from method 1, which are based on the same word recognition for each word in a test list, are farthest out, indicating the widest 95% confidence interval. By including variable word recognition in methods 2 and 3, the 95% confidence intervals become narrower, the curves are closest to the center. In methods 4 and 5, the variability of the test lists was taken into account. Thus, the 95% confidence intervals are again farther outside and almost coincide with those of method 1. There are only minor differences between the results of [5] and [12]. This is reflected in Figure 2, in which the bounds from methods 2 and 3, and from methods 4 and 5, lie close together.
Percentage scores of the FBE with 20 words per test list are possible only at intervals of 5%. Therefore, it is useful to conservatively round the bounds of the calculated 95% confidence intervals to multiples of 5%. These bounds for n=20 are given in Table 3. Tab. A. 1 in the Attachment 1, contains the corresponding bounds for n=40. The rounding partially increases the differences between the methods. However, the differences are at most 5% for both n=20 and n=40. The only exception is the difference between methods 1 and 3 at p=75% for the lower bound and p=25% for the upper bound at n=20, where the difference is 10%.
For methods 4 and 5, two variants are given in Table 3 and Tab. A. 1 in the Attachment 1. When including all 20 test lists (designations 4 and 5), $\tilde{n}$=21.4 was used (see

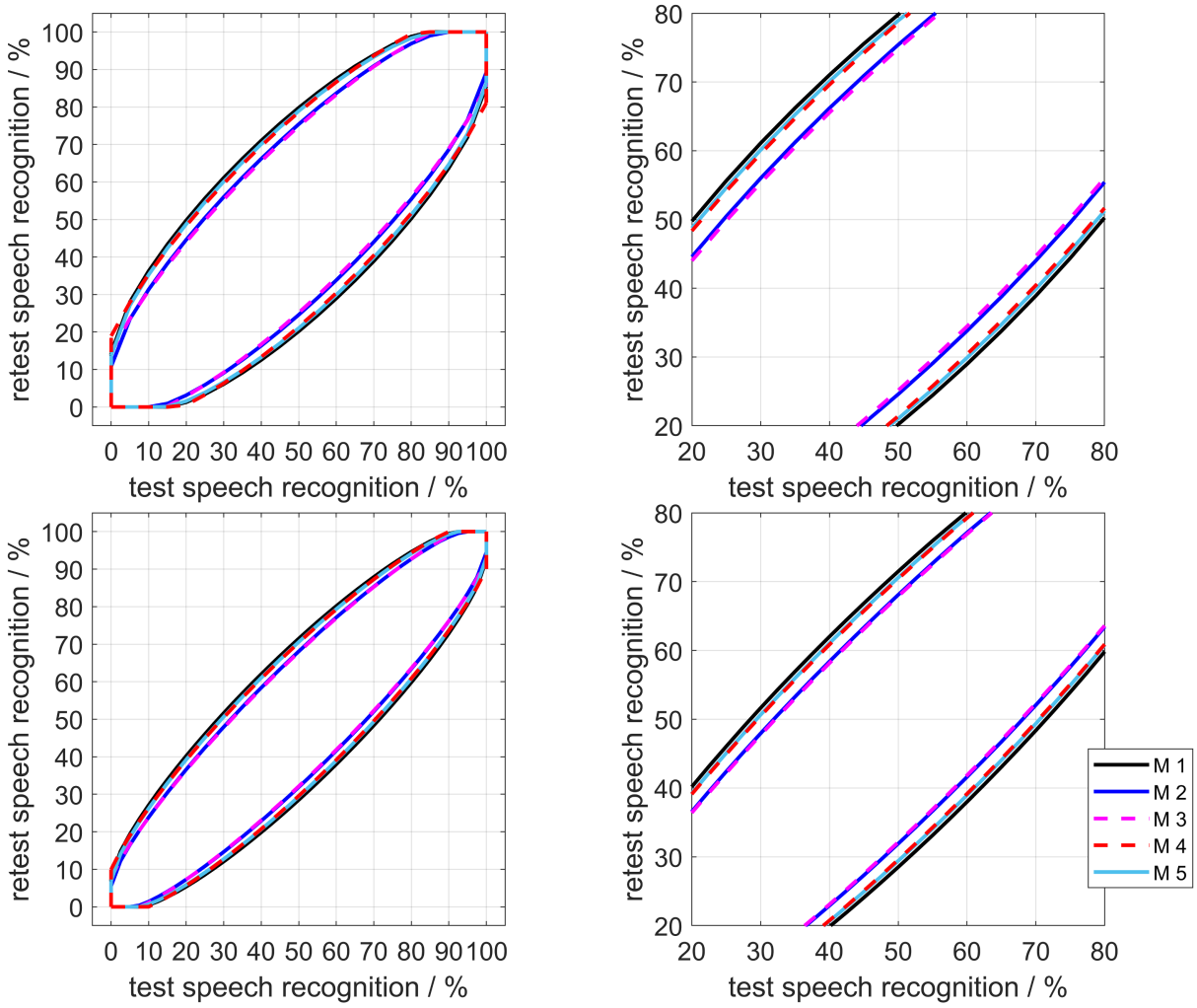

Figure 2: Comparison of the 95%-confidence intervals for test-retest reliability for test lists with 20 words (top) and with 40 words (bottom) when applying the five calculation methods. To improve comparability, the right side shows an enlargement of the central sector of the left figures, respectively. Black: method 1, dark blue: method 2, magenta dashed: method 3, red dashed: method 4, light blue: method 5.

Table 2). By omitting lists 5, 11, 12, and 15, i.e. with only 16 test lists, the effective list length increases to $\tilde{n}$=24.4. The corresponding bounds are given in columns 4/16 and 5/16. Omitting these four test lists reduces the variance of the test lists, and, consequently, the 95% confidence intervals become somewhat narrower.

Comparison with measurement data

The percentages of the measurements outside the 95% confidence intervals are given in Table 1. The goal that 5% of the measurement data should be outside the confidence interval is closely approached by method 1 for both normal hearing (NH) and hearing-impaired (HI) participants and when using 20 or 40 words per test list. However, method 1 does not take into account the differences in word recognition, nor those among test lists, and thus tends to overestimate the width of the confidence interval. For methods 2 and 3, which account for differences in word recognition, approximately 9% of the measurements are outside the 95% confidence interval. The specified bounds are therefore too narrow. Methods 4 and 5, in contrast to methods 2 and 3, take the variability of the test lists into account and achieve the 5% target in the various measurement data variants for all 20 test lists up to a maximum deviation of 0.5%, and for the 16 test lists up to a maximum deviation of 1.1% for the hearing-impaired participants with double test lists.
Figure 3 shows the measurement data, together with the critical differences according to method 5. For a percentage score of 50%, the 95% confidence interval lies between 25% and 75% (see Table 3, columns „5“). When double test lists are used (n=40), the 95% confidence interval is reduced to 30 % and 70 % (see Tab. A. 1, columns „5“ in the Attachment 1).

One-sided test

In the Attachment 1, Tab. A. 2 and Tab. A. 3 show the rounded 90% confidence intervals for n=20 and n=40 for all methods. The percentage of data outside of these confidence intervals for NH and HI for all variants is shown in Table 4. The criterion for the quality of the calculation method is that 10% of the data lies outside the calculated

Table 3: Bounds of the 95%-confidence interval for n=20 for the score of the second test list p_{mess2} when the score of the first test list p_{mess1} is given. The bounds were calculated with the methods 1–5 (see text). The precise values (see Figure 2) were conservatively rounded to multiples of 5%.

	95% confidence interval p_{mess2} in %													
Bounds	lower							upper						
Method p_{mess1}/%	**1**	**2**	**3**	**4**	**4/16**	**5**	**5/16**	**1**	**2**	**3**	**4**	**4/16**	**5**	**5/16**
0	0	0	0	0	0	0	0	10	10	10	15	15	10	10
5	0	0	0	0	0	0	0	25	20	20	25	25	25	25
10	0	0	0	0	0	0	0	35	30	30	35	30	35	30
15	5	5	5	0	0	5	5	40	35	35	40	40	40	40
20	5	5	5	5	5	5	5	45	40	40	45	45	45	45
25	5	10	10	5	5	5	5	55	50	45	50	50	50	50
30	10	10	10	10	10	10	10	60	55	55	55	55	60	55
35	10	15	15	10	15	10	15	65	60	60	60	60	65	60
40	15	20	20	15	15	15	15	70	65	65	65	65	70	65
45	20	25	25	20	20	20	20	75	70	70	70	70	70	70
50	25	25	30	25	25	25	25	75	75	70	75	75	75	75
55	25	30	30	30	30	30	30	80	75	75	80	80	80	80
60	30	35	35	35	35	30	35	85	80	80	85	85	85	85
65	35	40	40	40	40	35	40	90	85	85	90	85	90	85
70	40	45	45	45	45	40	45	90	90	90	90	90	90	90
75	45	50	55	50	50	50	50	95	90	90	95	95	95	95
80	55	60	60	55	55	55	55	95	95	95	95	95	95	95
85	60	65	65	60	60	60	60	100	95	95	100	100	95	100
90	65	70	70	65	70	65	70	100	100	100	100	100	100	100
95	75	80	80	75	75	75	75	100	100	100	100	100	100	100
100	90	90	90	85	85	90	90	100	100	100	100	100	100	100

confidence interval. The results are qualitatively similar to those in Table 1. While the bounds according to method 1 tend to be too wide, leading to less than 10% of the data outside the 90% confidence interval, methods 2 and 3 make the interval too narrow. The measurement results for normal hearing and hearing impaired participants can be better approximated using the methods 4 and 5 than when using the methods 2 and 3. Figure 4 shows the measured data together with the 90% confidence interval for method 5. According to method 5 and for n=20, the 90% confidence interval at a hit rate of 50% covers the range between 30% and 70% (see Tab. A. 2 in the Attachment 1). When using double test lists (n=40), the 90% confidence interval at this point is reduced and ranges between 35% and 65% (see Tab. A. 3 in the Attachment 1).

Critical differences in the level domain

For comparison with the accuracy of sentence tests, Table 5 shows the limits of the confidence intervals transformed to the level domain with a speech recognition score of 50% and of 80% for single test lists (n=20) and for double test lists (n=40). The confidence intervals are narrower for n=40 compared to n=20, and narrower for the 90% confidence interval compared to the 95% confidence interval. With 80% speech-recognition rate, confidence intervals are wider than for a speech recognition rate of 50%. The width of the confidence intervals ranges from ±4.0 dB for n=40 with a speech recognition rate of 50% (90% confidence interval) to ±11.3 dB for n=20 with a speech recognition rate of 80% (95% confidence interval).

Table 5: 95% and 90% confidence intervals in dB SPL from method 5 for n=20 and n=40 for scores 50% and 80%, calculated with Equation 18

	$n = 20$		$n = 40$	
Speech recognition %	**lower**	**upper**	**lower**	**upper**
95% confidence interval				
50	17.3	32.0	19.8	29.6
80	25.0	47.6	27.1	40.1
90% confidence interval				
50	18.7	30.7	20.7	28.7
80	26.1	43.3	27.9	38.5

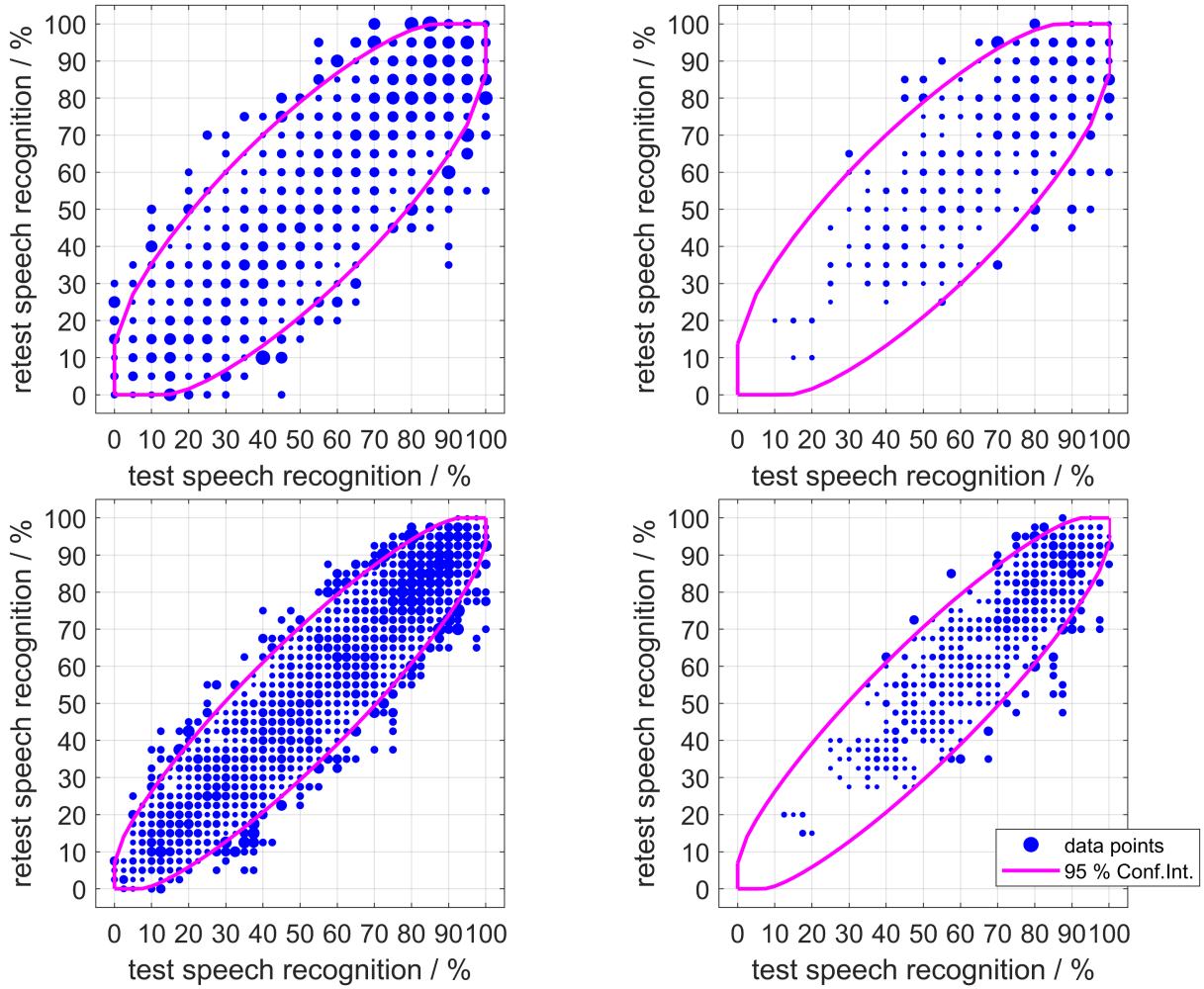

Figure 3: Data points (blue) and 95%-confidence interval (magenta) for test lists with 20 words (top) and with 40 words (bottom) for normal-hearing (left) and hearing-impaired (right) participants using method 5 and two-sided test.

Table 4: Number of data points and percentage of data outside the calculated 90%-confidence intervals
Results for normal-hearing (NH) and hearing-impaired (HI) participants with *n*=20 and *n*=40 words per test list for the methods 1–5

		Methods						
		1	**2**	**3**	**4**	**4/16**	**5**	**5/16**
NH $n = 20$	Data points	3,200	3,200	3,200	3,200	2,034	3,200	2,034
	%	8.9	17.1	17.0	11.0	9.5	9.4	8.5
HI $n = 20$	Data points	800	800	800	800	507	800	507
	%	9.1	16.6	16.6	9.9	8.9	9.3	8.1
NH $n = 40$	Data points	4,800	4,800	4,800	4,800	1,890	4,800	1,890
	%	8.8	16.3	16.4	11.3	9.0	10.7	8.4
HI $n = 40$	Data points	1,200	1,200	1,200	1,200	465	1,200	465
	%	9.3	15.3	15.4	11.6	11.6	11.1	11.4

Discussion

Modeling speech recognition as a Bernoulli experiment, with different word recognition scores within the test lists, the Poisson binomial distribution was used to calculate the 90% and 95% confidence intervals using different methods. The methods of Thornton and Raffin [5] and Altman et al. [12] led to similar results. These two methods were extended by additional consideration of the test-list variance. With this approach, the methods met the criteria that approximately 5% and 10% of the measured data are outside the limits of the calculated confidence intervals.

Depending on the variant (single or double test lists, 90% or 95% confidence interval, all 20 or only 16 selected test lists), the confidence intervals at a percentage score p_{mess1}=50% for the first measurement have a width of ±15% to ±25%. The guideline for assistive devices [4]

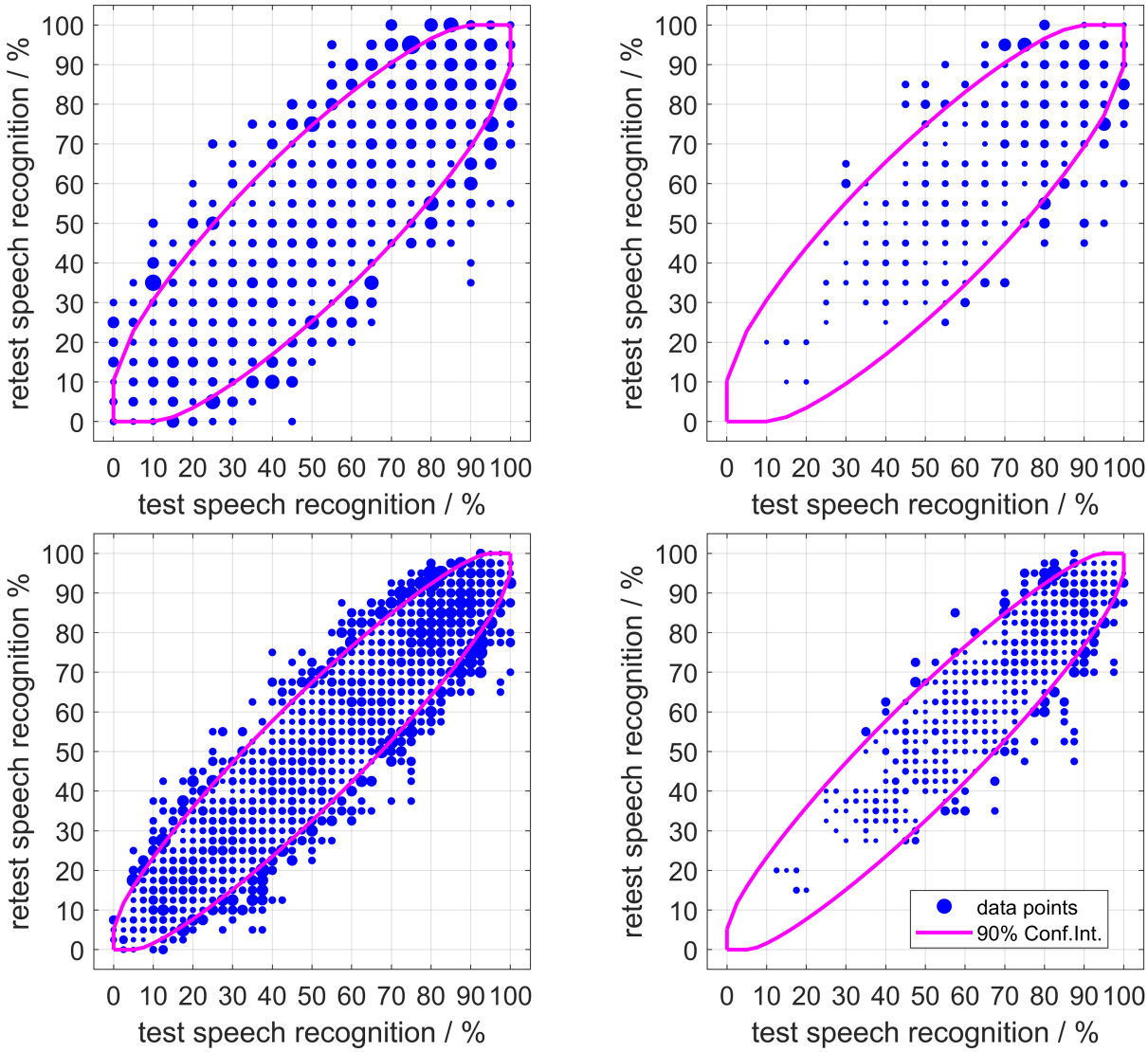

Figure 4: Data points (blue) and 90%-confidence interval (magenta) for test lists with 20 words (top) and with 40 words (bottom) for normal-hearing (left) and hearing-impaired (right) participants using method 5 and one-sided test.

requires an improvement of at least 20 percentage points for a hearing-aid fitting compared to the unaided measurement. At a percentage score of p_{mess1}=50% for the first measurement, an improvement of 20 percentage points in the second measurement is only statistically significant if double test lists are used. When using 20 words per test list, an increase of the percentage score of 20 percentage points by hearing aids is not statistically significant, because the error probability for the decision that the hearing aids improve speech recognition is more than 5%. For a significant improvement to be inferred from a difference of 20 percentage points, both the unaided and the aided condition would have to be determined using double test lists. When using single test lists, an improvement of 20 percentage points in the second measurement can only be regarded as significantly different for a percentage score of 75% or above for the first measurement. To narrow the confidence bounds, the four test lists that were conspicuous in Baljic et al. [11] may be omitted. Thus, the test-list variance would be reduced. However, there is no guarantee that for HI, in other German-speaking regions, or in other measurement configurations (e.g., in background noise), the same four test lists would still be outliers. An indication for deviations in conspicuous test lists could be that the confidence-interval bounds determined from the measurement data of the 16 selected test lists for the group of HI tended to be too broad. Thus, slightly less than the targeted 5% or 10% of the measurement data lay outside the confidence intervals. Even if all 20 test lists were used, the test list variance obtained from NH measurement data for modeling may be different for different groups of listeners or measurement conditions, leading to narrower or wider confidence intervals. For the measurement data of HI, however, the statement of Dillon [9], that HI have the same test-retest reliability as NH, was confirmed.

A comparison of the measurement results with the modeled confidence bounds also confirmed the conclusion of Dillon [9] that the bounds of Thornton und Raffin [5] according to method 1, i.e. when using the simple binomial distribution, can mimic the measured test-retest reliability relatively well. These bounds had already been

specified for the FBE by Winkler and Holube [7] for n=20. By using the Poisson binomial distribution in methods 2 and 3, however, the confidence intervals became narrower. After considering test-list variance in methods 4 and 5, widths became wider again, so that the limits of method 1 are approached. It should be noted, however, that the variability between the participants discussed by Dillon [9] was not incorporated in the present study. A possible reason for the negligible variability of the participants could be that two measurements within the same session were compared. Therefore, only the short-term reliability for test and retest was examined. The probably small intra-individual variance of participants within one session may be below test-list variance and might have been negligible here. Reliability has not been studied over an extended period, i.e., over several sessions, so that changes due to the variables "physical and mental state" of the participants were not measured. Another explanation for the negligibile variability between the participants could be that individual differences were not sufficiently considered [9]: To apply the Poisson binomial distribution, only the mean speech-recognition values of each single word were used. In individual participants, speech recognition of the words may have differed even more clearly, and methods 2 and 3 would have led to even narrower confidence intervals. In that case, an additional source of variance, e.g., the intra-individual variance, would be necessary to model the confidence intervals matching the measurement data.

For comparison with the sentence tests, the confidence intervals obtained from method 5 for percentage scores of 50% and 80% were transformed to confidence intervals for speech level. Using single test lists (n=20) with a speech recognition score of 50%, the 90% confidence interval has a width of ±6 dB. The confidence interval for the FBE is thus considerably wider than the confidence intervals for the adaptive sentence tests of about ±1 dB ([17], [18]). Hearing aids would need to improve the speech level by more than 6 dB at a speech recognition score of 50% in order to achieve a significant effect. For example, if the hearing aid only improved the speech level by 3 dB, then the sentence tests would result in a significant difference, and thus in a difference in efficacy, but not the FBE. The goal of an improvement of more than 6 dB appears to be easily achievable for speech recognition tests in quiet. However, whether this requirement can be transferred to an improvement by 6 dB in signal-to-noise ratio for FBE in noise is still unclear. Even if the same lists with n=20 and n=40 words would be used in noise, the variance in word recognition and in test-list recognition may differ from the FBE in quiet, and, therefore, deviating confidence bounds may result.

Conclusions

- Critical differences can be estimated relatively well solely from the number of measurement items, using method 1 proposed by Thornton und Raffin.
- With further knowlege about the speech test, i.e. the distribution of recognition of single items and the variance of test lists, methods 4 and 5 provide a more accurate model of the test-retest reliability.
- Confidence intervals should always be stated when publishing speech test results. It should also be noted whether a one-sided or a two-sided test was considered.

Notes

Competing interests

The authors declare that they have no competing interests.

Acknowledgement

This analysis was funded by the doctoral program Jade2Pro of Jade University of Applied Sciences. Additional funds were provided by the European Regional Development Fund (ERDF-Project Innovation network for integrated, binaural hearing system technology [VIBHear]), together with funds from the State of Lower Saxony. Manuscript language services were provided by http://stels-ol.de/.

Attachments

Available from
https://www.egms.de/en/journals/zaud/2020-2/zaud000007.shtml

1. Attachment_1.pdf (182 KB)
   Appendix: Confidence intervals

References

1. Hahlbrock KH. Uber Sprachaudiometrie und neue Wörterteste [Speech audiometry and new word-tests]. Arch Ohren Nasen Kehlkopfheilkd. 1953;162(5):394-431. DOI: 10.1007/BF02105664
2. 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.
3. 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
4. Gemeinsamer Bundesausschuss. Richtlinie des gemeinsamen Bundesausschusses über die Verordnung von Hilfsmitteln in der vertragsärztlichen Versorgung. Hilfsmittelrichtlinie. 2018 [accessed 13. Dezember 2018]. Available from https://www.g-ba.de/downloads/62-492-1666/HilfsM-RL_2018-07-19_iK-2018-10-03.pdf
5. 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

6. Carney E, Schlauch RS. Critical difference table for word recognition testing derived using computer simulation. J Speech Lang Hear Res. 2007 Oct;50(5):1203-9. DOI: 10.1044/1092-4388(2007/084)

7. 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

8. Steffens T. Test-Retest-Differenz der Regensburger Variante des OLKI-Reimtests im sprachsimulierenden Störgeräusch bei Kindern mit Hörgeräten. Z Audiol. 2006;45(3):88-99.

9. Dillon H. A quantitative examination of the sources of speech discrimination test score variability. Ear Hear. 1982 Mar-Apr;3(2):51-8. DOI: 10.1097/00003446-198203000-00001

10. Hagerman B. Reliability in the determination of speech discrimination. Scand Audiol. 1976;5:219-28. DOI: 10.3109/01050397609044991

11. 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

12. Newcombe RG, Altman DG. Proportions and Their Differences. In: Altman DG, Machin D, Bryant TN, Gardner MJ, editors. Statistics with Confidence: Confidence Intervals and Statistical Guidelines. 2nd Edition. London: British Medical Journal Books; 2000. p. 45-56.

13. Newcombe RG. Interval estimation for the difference between independent proportions: comparison of eleven methods. Stat Med. 1998 Apr;17(8):873-90. DOI: 10.1002/(sici)1097-0258(19980430)17:8<873::aid-sim779>3.0.co;2-i

14. Wilson EB. Probable Inference, the Law of Succession, and Statistical Interference. J Am Stat Assoc. 1927;22(158):209-12. DOI: 10.1080/01621459.1927.10502953

15. Wagener KC, Kühnel V, Kollmeier B. Entwicklung und Evaluation eines Satztests für die deutsche Sprache I: Design des Oldenburger Satztests. Z Audiol. 1999a;38:4-15.

16. Kollmeier B, Wesselkamp M. Development and evaluation of a German sentence test for objective and subjective speech intelligibility assessment. J Acoust Soc Am. 1997 Oct;102(4):2412-21. DOI: 10.1121/1.419624

17. Wagener KC, Brand T. Sentence intelligibility in noise for listeners with normal hearing and hearing impairment: influence of measurement procedure and masking parameters. Int J Audiol. 2005 Mar;44(3):144-56. DOI: 10.1080/14992020500057517

18. Brand T, Kollmeier B. Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests. J Acoust Soc Am. 2002 Jun;111(6):2801-10. DOI: 10.1121/1.1479152

Corresponding author:
Prof. Dr. Inga Holube
Jade University of Applied Sciences, Institute of Hearing Technology and Audiology, Ofener Str. 16/19, 26121 Oldenburg, Germany, Phone. +49-441-7708-3723
Inga.Holube@jade-hs.de

Please cite as
Holube I, Winkler A, Nolte-Holube R. Modellierung und Verifizierung der Test-Retest-Reliabilität des Freiburger Einsilbertests in Ruhe mit der verallgemeinerten Binomialverteilung. GMS Z Audiol (Audiol Acoust). 2020;2:Doc03.
DOI: 10.3205/zaud000007, URN: urn:nbn:de:0183-zaud0000070

This article is freely available from
https://www.egms.de/en/journals/zaud/2020-2/zaud000007.shtml

Published: *2020-03-27*