Modellierung der Test-Retest-Reliabilität von Sprachtests

Inga Holube, Alexandra Winkler, Ralph Nolte-Holube

Institut für Hörtechnik und Audiologie, Jade Hochschule, Oldenburg

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

Einleitung

Der Freiburger Einsilbertest (FBE, Hahlbrock, 1953) wird zur Messung des Sprachverstehens in Ruhe u.a. bei der Indikationsstellung für eine Hörgeräteversorgung und zur Überprüfung nach der Hörgeräteanpassung eingesetzt. Für die Indikationsstellung ist das Ergebnis einer Einzelmessung des Sprachverstehens mit einer Testliste relevant. Bei der abschließenden Überprüfung des Hörhilfenversorgungs-Ergebnisses in Ruhe werden zwei Testlisten-Ergebnisse (ohne und mit Hörgeräten) miteinander verglichen. Die Verbesserung durch die Hörgeräteversorgung muss mindestens 20 Prozentpunkte betragen. Jedoch ist die Reliabilität des FBE noch immer Gegenstand von Untersuchungen. Die Reliabilität kann abgeschätzt werden, indem der Sprachtest als Bernoulli-Experiment modelliert wird, dessen Ergebnisse einer Binomialverteilung folgen. Dabei wird im einfachsten Fall für die Erkennung jedes Testwortes die gleiche Wahrscheinlichkeit angenommen. Die Wörter des FBE sind innerhalb der Testlisten jedoch unterschiedlich gut oder schlecht zu verstehen. Deshalb wurde in Holube et al. (2018) wie von Hagerman (1976) vorgeschlagen der FBE mit der verallgemeinerten Binomialverteilung modelliert, die die unterschiedliche Wahrscheinlichkeiten für die Erkennung der Testwörter berücksichtigt. Die verallgemeinerte Binomialverteilung führt zu einem kleineren 95 %-Konfidenzintervall für das Messergebnis einer Testliste als die Verwendung der einfachen Binomialverteilung. In Holube et al. (2018) wurde die Varianz einer verallgemeinerten Binomialverteilung für den FBE mit $n = 20$ Wörtern pro Testliste durch diejenige Varianz einer einfachen Binomialverteilung angenähert, die auf Testlisten mit einer effektiven Anzahl von $n' = 29$ Wörtern mit gleichem Wortverstehen beruht. Im vorliegenden Beitrag wurde die Modellierung so erweitert, dass die Test-Retest-Reliabilität bei Durchführung von zwei Messungen berechnet werden kann. Die Fragestellung ist äquivalent zur Bestimmung des 95 %-Konfidenzintervalls der Differenz der zwei Messungen. Eine ausführlichere Beschreibung der Berechnungsmethoden und Ergebnisse wurde zur Publikation eingereicht (Holube et al., 2019).

Methoden

Bei der Berechnung des 95 %-Konfidenzintervalls der Differenz der zwei Messungen wird angenommen, dass sich die Varianzen der beiden Messwerte, d.h. die Varianzen der ermittelten Trefferraten addieren. Die Varianz der Trefferraten ist jedoch aufgrund der Kurvenverläufe der Diskriminationsfunktionen (und wie bei jeder binomialverteilten Zufallsgröße) von der Trefferrate abhängig. Deshalb schlugen Thornton und Raffin (1978) eine Transformation der Trefferraten zu einer Variablen mit näherungsweise konstanter Varianz vor. Diese Varianz ist lediglich von der Anzahl n der Wörter abhängig. Um das unterschiedliche Wortverstehen zu berücksichtigen, kann bei der Berechnungsmethode von Thornton und Raffin (1978) n durch n' aus Holube et al. (2018) ersetzt werden.

Bei der Verwendung des FBE zum Vergleich der Trefferrate ohne und mit Hörgeräteversorgung werden unterschiedliche Testlisten verwendet. Deshalb müssen bei der Berechnung des 95 %-Konfidenzintervalls zusätzlich noch die Unterschiede zwischen den Testlisten berücksichtigt werden. Das Verfahren von Altman et al. (2000) ermöglicht die Berechnung der Varianz der Differenz zweier Trefferraten aus der Binomialverteilung. Dieses Verfahren wurde verändert, um eine zusätzliche Varianz infolge der Variabilität der Testlisten zu berücksichtigen. Diese zusätzliche Varianz wurde aus den gemessenen Wortverständlichkeiten abgeschätzt. Die Modellierung führt zu einer gegenüber n' modifizierten Anzahl von $\tilde{n} = 21{,}4$ Wörtern pro Liste.

Die berechneten Grenzen der 95 %-Konfidenzintervalle wurden mit Testergebnissen von Probanden mit normalem Hörvermögen (im Folgenden Normalhörende genannt) verglichen, die bereits in Baljic et al. (2016) und Holube et al. (2018) verwendet wurden. Bei allen Probanden wurde das Sprachverstehen bei vier verschiedenen Pegeln mit allen 20 Testlisten bestimmt (fünf Testlisten pro Pegel). Die Messungen beim gleichen Pegel wurden für jeden Probanden als Test-Retest-Paare interpretiert.

Ergebnisse

Ein Vergleich der berechneten 95 %-Konfidenzintervalle mit den gemessenen Trefferraten der Normalhörenden zeigt für die Berechnungsmethode von Thornton und Raffin (1978) je nach verwendeter Wortanzahl pro Liste eine unterschiedliche Anzahl von Messdaten außerhalb der Konfidenzintervalle (siehe Abb. 1). Für $n = 20$ liegen 4,8 % der Datenpunkte, d.h. in etwa die Zielgröße von 5 % der Datenpunkte, außerhalb des Konfidenzintervalls (Abb. 1, links). Die Ersetzung von $n = 20$ durch $n' = 29$ durch Berücksichtigung des unterschiedlichen Wortverstehens führt zu schmaleren Konfidenzgrenzen und damit zu einem Anstieg des Anteils der Datenpunkte außerhalb des 95 %-Konfidenzintervalls auf 8,7 % (Abb. 1, Mitte). Die Erweiterung um die Testlistenvarianz, d.h. die Ersetzung von $n = 20$ durch $\tilde{n} = 21{,}4$ führt zu einer Verbreiterung des 95 %-Konfidenzintervalls auf nahezu die ursprüngliche Größe, so dass 5,4 % der Datenpunkte außerhalb liegen (Abb. 1, rechts). Damit konnte die Anwendbarkeit dieser Berechnungsmethode zur quantitativen Angabe des 95 %-Konfidenzintervalls gezeigt werden. Bei einer Trefferrate von 50 % ist das 95 %-Konfidenzintervall mit +/- 25 % am breitesten. Für die Hörgeräte-Anpasspraxis bedeutet dies, dass erst Unterschiede, die diese Spanne übersteigen, als signifikant unterschiedlich gewertet werden können. Die Berechnungsmethode kann auf andere Sprachteste übertragen werden, wenn die Erkennungsraten der einzelnen Test-Items bekannt sind.

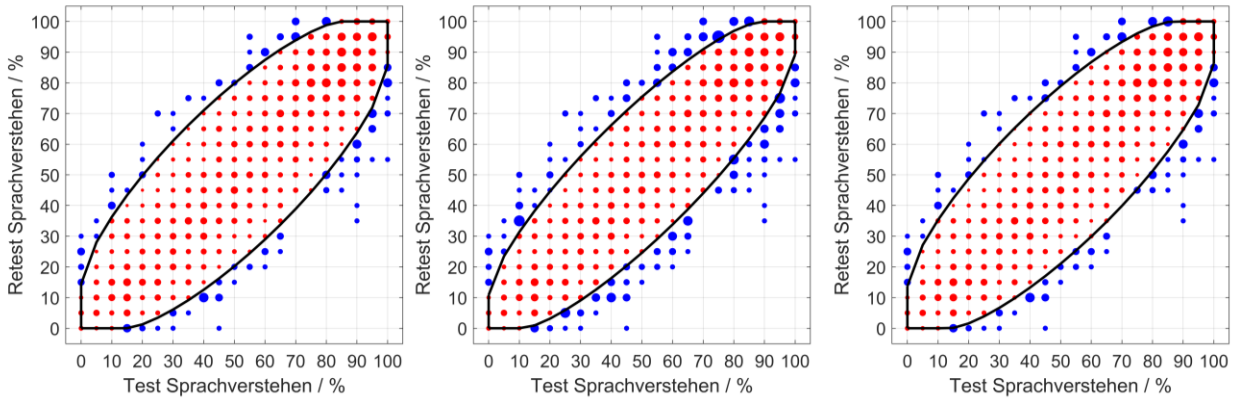

Abbildung 1: Datenpunkte (blau) und 95 %-Konfidenzintervalle (schwarz) für die Berechnungsmethode nach Thornton und Raffin (1978) mit $n = 20$ *(links),* $n' = 29$ *(Mitte) und* $\tilde{n} = 21{,}4$ *(rechts) Wörtern pro Testliste.*

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.

Literatur

Altman D G, Machin D, Bryant T N, Gardner M J (2000) Statistics with confidence. British Medical Journal Books, Kapitel 6, 2. Auflage.

Baljic I, Winkler A, Schmidt T, Holube I (2016) Untersuchungen zur perzeptiven Äquivalenz der Testlisten im Freiburger Einsilbertest. HNO 64, 572-583.

Hagerman B (1976) Reliability in the determination of speech discrimination. Scand. Audiol. 5, 219-228.

Hahlbrock K H (1953) Über Sprachaudiometrie und neue Wörterteste. Archiv Ohr- usw. Heilk. u. Z. Hals- usw. Heilk. 162, 394-431.

Holube I, Winkler A, Nolte-Holube R (2018) Modellierung der Reliabilität des Freiburger Einsilbertests in Ruhe mit der verallgemeinerten Binomialverteilung. Zeitschrift für Audiologie 57(1), 6-17.

Holube I, Winkler A, Nolte-Holube R (2019) Modellierung der Test-Retest-Reliabilität des Freiburger Einsilbertests in Ruhe mit der verallgemeinerten Binomialverteilung. Eingereicht bei Zeitschrift für Audiologie.

Thornton A R, Raffin M J M (1978) Speech-discrimination scores modeled as a binomial variable. J. Speech Hear. Res. 21, 507-518.