Auditiv-perzeptive Beurteilung stimmlicher Parameter

Ergebnisse einer Test-Retest-Studie zur Einschätzung der Heiserkeit – Vergleich von visueller Analogskala (VAS) und RBH-Verfahren

Peter Dicks

ZUSAMMENFASSUNG. Die Ergebnisse einer Test-Retest-Studie zur Reliabilität der Heiserkeitseinschätzung organischer Stimmstörungen beleuchten detailliert das Bedingungsgefüge von Stimmanalysen. Bei über alle Parameter sehr guter Test-Retest-Reliabilität und hoher interner Konsistenz der Gruppenurteile zeigt sich, dass Rauigkeit schwerer einzuschätzen ist als Behauchtheit und Heiserkeitsgesamtgrad. Referenzstimmen unterstützen reliablere Ergebnisse. Die Einschätzung per visueller Analogskala (VAS) ist genauer als mittels ordinalskalierter RBH-Klassifikation. Deutliche Lerneffekte zwischen Testhälfte 1 (Stimme Nr. 1-20) und Testhälfte 2 (Stimme Nr. 21-40) belegen die Notwendigkeit eines Trainings auditiv-perzeptiver Fähigkeiten. Klinisch erfahrenere Logopäden erzielen Reliabilitätsgrade im sehr guten Bereich. Die auditiv-perzeptive Beurteilung von Stimmstörungen erweist sich als ein zuverlässiger Baustein der Gesamtdiagnostik von Stimmstörungen.

Schlüsselwörter: Heiserkeit – auditiv-perzeptive Klassifikation – Stimmstörung – Stimmdiagnostik – Lerneffekte – Reliabilität

Peter Dicks ist seit 1997 nach langjähriger Klinik- und Praxistätigkeit u.a. in der Phoniatrie des Universitätsklinikums Düsseldorf als Lehrlogopäde tätig. Seine Ausbildung absolvierte er von 1987-90 an der IFBE-Schule in Köln. Zurzeit ist er für die Fachbereiche Stimmstörungen (inkl. Laryngektomie und Dysarthrie) und Dysphagie an der Schule für Logopädie Uniklinik/RWTH Aachen (dual-integrierender Bachelorstudiengang) zuständig. Im September 2011 beendete er sein Diplomstudium Lehr- und Forschungslogopädie mit der hier veröffentlichten Arbeit, die 2013 mit der Springorum-Denkmünze der RWTH Aachen ausgezeichnet wurde.

Einleitung

Die auditiv-perzeptive Beurteilung von Stimmstörungen ist ein wesentlicher, aber in Bezug auf seine Aussagefähigkeit vieldiskutierter Bereich der Beschreibung und Klassifizierung normaler und pathologischer Stimmen (*Oates* 2009, *Kreiman & Gerratt* 2010). Im europäischen Raum hat sich in Entwicklung aus dem GRBAS-System die modifizierte RBH-Klassifikation mittels Ordinalskalierung verbreitet (*Nawka* et al. 1996).

Die Europäische Laryngologische Gesellschaft (ELS) fordert in einem umfassenden Protokoll zur Stimmbefundung, die auditiv-perzeptive Beurteilung neben der akustisch-apparativen Analyse, aerodynamischen Messungen, laryngoskopisch-stroboskopisch bildgebenden Befunden, anamnestischen und selbsteinschätzenden Angaben miteinzubeziehen (*Friedrich & Dejonckere* 2005).

Dies entspricht auch den Anforderungen an Stimmuntersuchungen, die im angloamerikanischen Raum formuliert werden (*Sataloff* 2005, *Mehta & Hillmann* 2008, *Stemple* et al. 2010). Hier wurde in 2002 durch die ASHA als Protokoll der „Consensus Auditory Perceptual Evaluation of Voice" (CAPE-V) zur Etablierung vorgeschlagen (*Kempster* et al. 2009, *Kelchner* et al. 2010), der eine visuelle Analogskalierung für sechs Parameter nutzt. Stimmtherapie ist durch den Einsatz auditiv-perzeptiver Beurteilungsverfahren und akustischer Messungen in ihrer Effektivität adäquat beurteilbar und dokumentierbar (*McKenzie* et al. 2001, *Speyer* 2006).

Das in Europa als phoniatrisch-logopädischer Standard etablierte ELS-Protokoll erweist sich in keinem der modular erhobenen Diagnostikbereiche als isoliert valide, liefert aber in toto ein aussagefähiges Bild der Stimmproblematik mit guten Korrelationen der Einzelbereiche zueinander (*Gonnermann* 2007). Der Einsatz von Ankerstimmen bzw. -modalitäten als zu vergleichende externe Referenz zur inneren Repräsentation des Beurteilers wird von mehreren aktuellen Studien positiv evaluiert (*Chan & Yiu* 2006, *Eadie & Baylor* 2006, *Awan & Lawson* 2009, *Dicks & Nawka* in Vorbereitung).

Es stellen sich folgende Hauptforschungsfragen mit dieser Arbeit an den jetzigen Stand der auditiv-perzeptiven Stimmbeurteilung:

- Wie zuverlässig beurteilen das RBH-Verfahren und die visuelle Analogskalierung des CAPE-V die ausgewählten Stimmparameter?
- Fördern externe natürliche Ankerstimmen die Beurteilungsgenauigkeit?
- Welche Lerneffekte treten auf?
- Verbessert klinische Erfahrung die Beurteilungsgenauigkeit und
- Wie ist die Beurteilungsgenauigkeit bei unterschiedlichen Störungsgraden?

Klassifikationssysteme

Das RBH-System (GRBAS) ist ein aus untersuchungsökonomischen Gründen auf die Hauptparameter von Heiserkeit reduziertes Klassifizierungssystem mittels Ordinalskalierung (Werte von 0, 1, 2 und 3, von „keine Heiserkeit" bis „hochgradig heiser"). Als Nachteile des RBH-Systems werden angegeben, dass die Angabe von Zwischenstufen nicht möglich ist und weitere Stimmauffälligkeiten aufgrund fehlender Kategorien nicht bewertet werden können. Die RHB-Klassifikation ist aufgrund der aktuellen Studienlage und der Beschreibung in Standardwerken zur Stimmdiagnostik auch in der verbreiteten klinischen Anwendung als standardisiertes Verfahren bestätigt (*Bigenzahn & Schneider* 2007, *Hammer* 2009).

Der CAPE-V (Consensus Auditory Perceptual Evaluation of Voice, 2002 von der ASHA als Konsenspapier zur Beurteilung von Stimmstörungen vorgeschlagen) nutzt insgesamt sechs Parameter (Gesamtheiserkeit, Rauigkeit, Behauchtheit, Gepresstheit (Strain), Tonhöhe (Pitch) und Lautstärke (Loudness). Der Beurteiler muss auf einer 100 mm langen Linie das Merkmal in einem Kontinuum von nicht bis

hochgradig ausgeprägt einstufen. Auf diese Weise sind feinere Abstufungen bei der Bewertung auditiver Auffälligkeiten möglich. Das Protokoll sieht vor, dass angegeben wird, ob das Merkmal ständig (consistent: c) oder gelegentlich (intermittent: i) auftritt. Zusätzlich können zwei weitere, individuell wählbare Parameter beurteilt werden.

Sowohl das GRBAS- bzw. RBH-Verfahren und das seit 2002 als Konsens angestrebte visuelle Analogskalaverfahren des CAPE-V haben eine übergreifende klinische und durch Studien belegte Gültigkeit erworben, die beide als Modul einer profilorientierten Stimmdiagnostik als gut einsetzbar erscheinen lassen (*Evans* et al. 2004, 2005; *Karnell* et al. 2007).

Studiendesign und Hypothesen

Die experimentell-methodische Grundidee der vorliegenden Studie ist, als Ankerstimme denselben Lesetext als Referenzwert einer mittelgradig gestörten Stimme zum Hörvergleich nutzen können. Den Teilnehmern der Ankerstimmengruppe soll im Gegensatz zur Kontrollgruppe während des Trainings der Einsatz des externen Standards dieser Ankerstimme mit RBH 1,5/1,5/1,5 und VAS 50/50/50 mm verdeutlicht werden und dann im Test zusätzlich zehnmal zu den 40 Testitems als Hörbeispiel präsentiert werden. Für den Einsatz der Ankerstimmen ist mit beiden Skalenarten eine erkennbare Verbesserung der Reliabilitätswerte durch interne Referenzbildung im Vergleich zu den Kontrollgruppen zu erwarten (*Ptok* et al. 2006).

Der Einsatz von Ankerstimmen natürlicher oder synthetisch erzeugter Art in Verknüpfung mit einem vorgeschalteten Training wird in vielen aktuellen Studien empfohlen (*Gerratt & Kreiman* 2004, *Kreiman* et al. 2007). Im Sinne der internen Bildung eines „Summenwertes" der Stimmklangparameter aus dem zu beurteilenden Lesetext über eine längere Beurteilungsdauer (45-60 Sekunden) gilt diese Art der Einschätzung als sehr praxisrelevant und aussagefähig (*Bele* 2004).

Erwartet wurde ein leichter Vorteil des VAS-Verfahrens, weil mit diesem Verfahren eine kontinuierliche Ausprägung der Heiserkeit als Klangphänomen ähnlich der Bewertung einer Helligkeitsabstufung im visuellen Bereich mit Zwischenstufen adäquat bewertet werden kann. *Yui* et al. (2007) zeigten mittels der VAS-Skala eine verbesserte Interraterübereinstimmung, aber keinen Effekt auf die Intraraterreliabilität.

In der Auswertung sollten die unterschiedlichen Schweregrade der Störungen verglichen werden. Erwartet wurden im geringgradig bis mittleren Bereich gestörter Stimmen weniger zuverlässige Bewertungen. Der Parameter Rauigkeit erschien schwieriger interpretierbar als die beiden anderen Parameter. Der Einfluss der klinischen Expertise der Beurteiler (Stichprobe) wird unterschiedlich eingeschätzt *(Pützer & Barry* 2004).

Erwartet wurde eine hohe Motivationslage, aktuelle Kenntnisse bei diagnostischen Fertigkeiten und eine homogene Ausprägung des Kenntnisstandes der Gesamtgruppe bei den angeworbenen freiwilligen Teilnehmern, die im Mittel einem Berufsanfänger der Logopädie entsprechen und damit repräsentativ für diesen Stand der Expertise eines Fachbeurteilers ist.

Es ist übergreifend anzumerken, dass keine direkte Aussage über die Validität der Urteile zu den beurteilten Stimmparametern erfolgen kann (*Kreiman* et al. 1993). Die direkte Korrelation mit akustischen Verfahren ist sicher die adäquateste Möglichkeit, die Validität der perzeptiven Ergebnisse zu erhöhen (*Eadie & Doyle* 2005).

Untersuchungsmethodik

Studiendesign

Im Rahmen einer Test-Retest-Untersuchung mit 14-tägigem Intervall wurden die drei ausgewählten Parameter der Stimmqualität organischer Stimmstörungen Gesamtheiserkeitsgrad (H), Rauigkeit (R) und Behauchtheit (B) mit den zurzeit gebräuchlichsten Verfahren des Ordinalskala-Klassifikationssystems RBH (ORD) und der visuellen Analogskala (VAS) beurteilt.

Im Sinne eines Kontrollgruppendesigns schätzten randomisierte Beurteiler entweder a) mit 10 natürlichen Ankerstimmen (Gruppe A) während der Testung oder b) ohne diese Ankerstimmen (Gruppe K) 40 Stimmbeispiele des Lesetextes „Der Nordwind und die Sonne" (Tab. 1). Die Teilnehmer wurden mehrfach darauf hingewiesen, im Sinne des Test-Retest-Designs eine Wiederholung, Training oder jedwede übende Beschäftigung mit dem Aufgabenthema im Intervall zwischen Test 1 und 2 zu unterlassen.

Stimmbeispiele (Korpus)

Die Studie wurde im Rahmen eines Diplomprojekts des Studiengangs Lehr- und Forschungslogopädie der RWTH Aachen unter Betreuung von Prof. Tadeus Nawka (Charité Berlin) und Dr. Bruno Fimm (UK Aachen) von November 2009 bis März 2011 erstellt.

Der Gesamtkorpus von 127 Audioaufnahmen von Patienten mit typischen organischen Stimmstörungen der phoniatrischen Sprechstunde der Universität Greifswald setzt sich aus 46 männlichen (36 %) und 81 weiblichen (86 %) Stimmbeispielen aller Schweregrade zusammen (Alter: 10 bis 74 Jahre, Median 45 Jahre). Zu allen Stimmen gab es Gruppenurteile von 5 Beurteilern (deutsch) nach RBH-Klassifikation, 15 Beurteilern (europäisch) nach VAS-Klassifikation.

Alle Audioaufnahmen entstammen einer für Lernzwecke konzeptionierten multimedialinteraktiven CD-ROM (*Evans* & Nawka 2005). Diese beinhaltet 127 Aufnahmen des standardisierten Lesetextes „Der Nordwind und die Sonne". Die durchschnittliche Sprechdauer beträgt ca. 45-60 Sekunden.

Beurteiler (Stichprobe)

Die 99 Teilnehmer der Studie nahmen mit einem Gesamtzeitaufwand von 8 Stunden an vier Terminen freiwillig unter Erhalt einer Aufwandsentschädigung und Bescheinigung einer einführenden Schulung zur Anwendung auditiv-perzeptiver Verfahren mit Feedback an dieser Studie teil. Hörbeeinträchtigungen lagen nicht vor. Alter: 18-52 Jahre, Median 23 J. Die Geschlechterverteilung – 7 Männer, 92 Frauen – entspricht der typischen Repräsentation im logopädischen Arbeitsfeld.

69 (70 %) der Probanden waren Schüler der staatlichen Logopädieschulen Aachen und Münster, 25 Studierende (25 %) des Studiengangs Lehr- und Forschungslogopä-

Tab. 1: Studiendesign (n = 99 Rater, 95 % Schüler und Studierende der Logopädie)

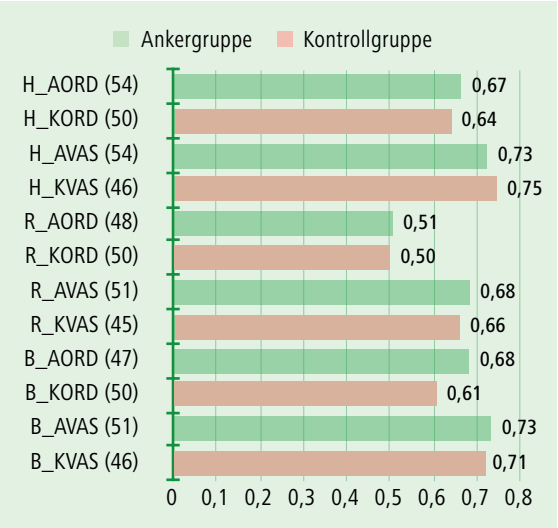
Training	Test 1 (T1)	2 Wochen Test-Retest-Intervall	Test 2 (T2)	Feedback
30 Items, 180 min	60min **AORD*** **KORD**** **AVAS*** **KVAS****	keine Rückmeldung zur Stimmbewertung, kein Training	ident. T1, (n Rater) **AORD** (26) **KORD** (25) **AVAS** (25) **KVAS** (23)	Evaluation

Test 1 und Test 2: 40 Urteile jeweils für die Parameter Gesamtheiserkeit (H), Rauigkeit (R), Behauchtheit (B)

*) **Ankerstimmengruppe (A)**: erhält Training und Testung mit 10 Referenzstimmbeispielen
) **Kontrollgruppe (K): erhält Training, aber keine Schulung zu Ankerstimmen

Verfahren: Ordinalskala (ORD) 0,1,2,3; Visuelle Analogskala (VAS) 0-100mm

Abb. 1: ICC Anker- vs. Kontrollgruppe



die RWTH Aachen und der Fachhochschule Heerlen (NL), 2 Lehrkräfte (2 %) und 3 (3 %) berufstätige Logopäden. Die Teilnehmer entschieden sich überwiegend nach eigenem Interesse und in freier Entscheidung für ihr dann fest zugeordnetes Verfahren während der Studie.

Statistische Auswertung

Die Datenauswertung erfolgte mit dem Programm PASW Statistics 18 Version 18.0.0 (SPSS 18, IBM 2010). Als Maß der Beurteilerübereinstimmung wurde der *Intraklassenkorrelationskoeffizient* (ICC, *Wirtz & Caspar* 2002) genutzt. Hierbei wird der „Anteil der Varianz der wahren Merkmalsausprägungen, der entweder durch die Urteile eines Raters oder durch den Mittelwert mehrerer Rater erklärt werden kann" ermittelt (ebd. 18).
Hier kann bei Werten über 0,9 von einer sehr guten Reliabilität ausgegangen werden, da Skala und Konsistenz der Gruppenurteile die „wahren" Merkmalsausprägungen ausgezeichnet voraussagen. Die Kennwerte für die Einzelurteile der ICC werden ab 0,7 als gut eingestuft (ebd. 25, 226), ab 0,3 sind diese signifikant. In Bezug auf *de Bodt & Wuyts* (1997) halten wir die in Tabelle 2 angegebenen Kennwertintervalle für sinnvoll. Die Ordinalskala ORD wurde aufgrund ihrer gestuften Werte auf absolute Übereinstimmung der ICC überprüft (model random), während die visuelle Analogskala VAS durch ihre metrischen Werte auf ihre Konsistenz überprüft werden (model mixed consistency) (*Wirtz & Caspar* 2002, 215). Signifikanztests zum Vergleich der ICC-Kennwerte wurden übergreifend bei den verschiedenen Vergleichen nicht durchgeführt.

Ergebnisse

Ankerstimme vs. Kontrollgruppe

Die Gruppenurteile beim Vergleich der Beurteilergruppen mit Ankerstimmen (A) und ohne Ankerstimmeinsatz (Kontrollgruppe K) liegen von 0,98-0,99 im sehr guten Bereich für die interne Konsistenz der Skala und Beurteilungen. Insgesamt bilden sich in der Analyse der ICC der 12 kleineren Untergruppen nach Anker- und Kontrollgruppen getrennt mäßige bis gute Reliabilitätskoeffizienten für das Einzelmaß ab (Bereich 0,50-0,75). In fünf der sechs Vergleiche ist das Ankerstimmverfahren tendenziell leicht genauer als die entsprechende Kontrollgruppe mit einer durchschnittlichen Differenz der ICC von +0,03 zugunsten des Ankerstimmverfahrens (Tab. 3, Abb. 1, keine Signifikanzprüfung).

Stimmparameter

Hier zeigten sich die besseren Werte für die Parameter H und B, die sich sehr gleichartig im guten Bereich ausprägen. Rauigkeit wird durch das ORD-Verfahren mit und ohne Ankerstimmen nur mäßig eingeschätzt.

Tab. 3: ICC Anker- vs. Kontrollgruppe

Verfahren / (n)	Test_gesamt	95 % Konfidenz-intervall	Cronbachs Alpha Gruppenmaße
H_AORD (54)	**0,665**	0,568-0,768	0,991
H_KORD (50)	**0,641**	0,539-0,749	0,992
H_AVAS (54)	**0,725**	0,637-0,814	0,993
H_KVAS (46)	**0,746**	0,661-0,829	0,993
R_AORD (48)	**0,506**	0,404-0,631	0,982
R_KORD (50)	**0,500**	0,397-0,621	0,984
R_AVAS (51)	**0,684**	0,590-0,782	0,991
R_KVAS (45)	**0,661**	0,564-0,764	0,989
B_AORD (47)	**0,683**	0,589-0,782	0,991
B_KORD (50)	**0,607**	0,504-0,721	0,99
B_AVAS (51)	**0,734**	0,648-0,821	0,993
B_KVAS (46)	**0,712**	0,621-0,804	0,991

ORD vs. VAS

Vergleicht man die sechs Untergruppen ORD und VAS unabhängig von Kontroll- oder Ankerstimmengruppe, so zeigt sich mit einem Mittelwert von 0,60 für ORD zu 0,71 für VAS eine deutliche Überlegenheit der visuellen Analogskala, die trotz der Anwendung von Zwischenskalierungen höhere Übereinstimmungsraten der Beurteiler ermöglicht (detailliertere Ergebnisse in *Dicks & Nawka* in Vorbereitung).

Tab. 2: Kennwertintervalle der Intraklassenkorrelationskoeffizienten zur Beurteilungsgenauigkeit (Einzelmaße) nach *de Bodt* et al. 1997, 77

< 0,20	sehr gering (poor)
0,21-0,40	gering (fair)
0,41-0,60	mäßig (moderate)
0,61-0,80	gut (good)
0,81-1,00	sehr gut (very good)

Testhälften und Lerneffekte

Beim Vergleich der Interraterreliabilität über die ICC wurden die beiden Testhälften (Stimme 1-20, Stimme 21-40) verglichen. Dabei zeigte sich ein sehr gleichmäßiger Lerneffekt über alle Hauptgruppen (Abb. 2) mit einem Mittelwert von +0,11. Am deutlichsten zeigte sich mit einer Differenz von + 0,18 die Verbesserung bei der Einschätzung der Rauigkeit mit R_AVAS (51), die fast einem Lerneffekt von einer Bewertungskategorie von gut nach sehr gut entspricht. Im Vergleich der ICC der Lernhälften T1 und Test 2 sind ORD und VAS gleichzusetzen. Die Mittelwerte der 3 ORD- bzw. 3 VAS-Hauptgruppen liegen bei 0,12 und 0,11 absolut auf gleichem Niveau. ORD profitiert im gleichen Maße wie VAS vom Lerneffekt der ersten Testhälfte.

Erfahrungsgrad (Klinische Expertise)

Es wurden die ausgewiesenen Beurteiler mit einer Erfahrung von mindestens zwei Jahren klinischer Expertise in der Diagnostik und Therapie von Stimmstörungen einzeln bewertet (n=6, VAS-Verfahren: 3 AVAS, 3 KVAS). Es gab nur einen Experten, der mittels ORD beurteilt hatte, sodass hierzu keine Auswertung möglich war. Die Kontrollgruppe der Experten (ohne Ankerstimmeinsatz) schätzte mit einer ICC-Mittelwertdifferenz von -0,13 über alle drei Parameter die Stimmen deutlich schlechter ein als die vergleichbare Experten-Ankerstimmengruppe (Abb. 3).
Die Rauigkeit wurde von den Experten der Ankerstimmgruppe R_AVAS_Exp mit 0,80 sehr gut eingeschätzt und dies deutlich besser als die vergleichbare Gesamtteilnehmergruppe R_AVAS_alle (+0,11). Die Kontrollexpertengruppe (ohne Ankerstimme) H_KVAS_Exp erreichte nur bei der Beurteilung von B einen vergleichbaren sehr guten Wert von 0,80, d.h. Behauchtheit ist von Experten gut einzuschätzen auch ohne Ankerstimmeinsatz.
Diese Expertengruppe (R_KVAS) fällt bei der Reliabilität der Rauigkeitsbeurteilung mit 0,57 mit eher moderaten Werten stark auf. H wurde mit 0,69 gut eingeschätzt. Insgesamt kann für VAS gesagt werden, dass die ver-

kleinerten Expertengruppen mit einer guten bis sehr guten Zuverlässigkeit die Stimmparameter einschätzten.

Schweregrade

Die Schweregrade wurden über die ICC für die jeweiligen Untergruppen analysiert, wobei die wesentlich geringere Zahl an beurteilten Items (Angabe in Klammern) deutlich verringerte ICC-Maße bedingt. Die Einteilung der Stimmen in die jeweiligen Schweregrade erfolgte über die Berechnung der Mittelwerte der Urteile der einzelnen Rater zu den jeweiligen Stimmen von Test 1 und 2. Diese ICC-Ergebnisse sollen nicht mehr auf die Kennwerte (Tab. 1) bezogen werden, sondern untereinander verglichen und mithilfe der Verbesserung der ICC-Mittelwertdifferenzen beschrieben werden.

In der Detailanalyse der Schweregrade (SG) ist VAS zu ORD im Mittelwert der Differenzen der verringerten ICC mit +0,1 erkennbar überlegen (Abb. 4). Die gesteigerte Beurteilungsgenauigkeit wird bei den mittelgradig bis stärker gestörten Stimmen (SG_VAS_1 und _2) am deutlichsten. Die Behauchtheitseinschätzung durch VAS ist bei SG_VAS_2 deutlich SG_ORD_3 überlegen (0,54 zu 0,35). Die größte Mittelwertdifferenz zwischen VAS und ORD im Schweregrad SG_VAS_2 beträgt +0,12, die kleinste Mittelwertdifferenz zwischen VAS und ORD im Schweregrad SG_ORD_1 liegt bei +0,07.

ORD ist insgesamt homogener in der Beurteilungsgenauigkeit über die Schweregrade, aber SG_ORD_3 zeigt die höchste Beurteilungsgenauigkeit (beachte: Itemanzahl von 4!). VAS ist bei SG_VAS_2 und SG_VAS_3 deutlich stärker als ORD und erreicht bei Schweregrad 3 (severe) einen ICC von 0,54 bei aber hoher Itemanzahl von 17. Aufgrund der unterschiedlichen Itemanzahl (Stimmen) der Untergruppen ist dieser Vergleich zu relativieren und kann lediglich eine Tendenz beleuchten.

Diskussion

Ankerstimmeinsatz

Dem Einsatz von natürlichen Ankerstimmen, wie in dieser Studie spezifiziert (10 zusätzliche Stimmbeispiele mittlerer Ausprägung während der Testung, Zeitlimit Hören des Standardtextes und Beurteilung ca. 1 Minute, keine Verifizierung durch erneutes Hören), muss aufgrund des leichten, nicht durch Signifikanztests abgeklärten Effektes eine tendenzielle Förderung einer besseren Bewertung heiserer Stimmen zugesprochen werden, was der Tendenz aktueller Studien entspricht *(Chan & Yiu* 2002, *Ptok* et al. 2006, *Yui* et al. 2007, *Kreiman* et al. 2007, *Awan & Lawson* 2009). Unterstützend muss gewertet werden, dass die Expertengruppe mit Ankerstimmen sehr gute Reliabilitätswerte mit ICC von 0,8 (bis leicht darüber) im Vergleich mit der Kontrollgruppe und der Gesamtteilnehmergruppe erreichte. Außerdem kann der größte Lerneffekt zwischen den Testhälften 1 und 2 beim schwierigsten Stimmparameter Rauigkeit R auch bei der Ankerstimmgruppe (R_AVAS) beobachtet werden.

Eine Verbesserung bezogen auf den spezifischen Einsatz der Ankermodalität in dieser Studie ist durch zusätzlichen Einsatz textueller (erläuternder) oder visualisierender Hilfen *(Martens* et al. 2007) oder differenziertere Paarbildung sowie Abgleichung mit zugeordneten Stimmen vorstellbar. Kritisch ist zu werten, dass die Gültigkeit der Ankerstimmen nur auf die Vorbeurteilung der Expertengruppe und der Einschätzung des Autors beruht. Akustische Analysen oder größere Vorratings hätten deutlich mehr Sicherheit gebracht.

VAS vs. ORD

Die deutliche Überlegenheit des VAS-Verfahrens gegenüber dem ORD-Verfahren bei guten bis sehr guten Reliabilitätswerten bestätigt die Tendenz aktueller Studien *(Helou* et al. 2010, *Zraick* et al. 2010), die das VAS-Verfahren als Methode der Wahl zur Bestimmung der Stimmqualität heiserer Stimmen einschätzen. Diese bietet die Möglichkeit, mittels einer visuellen Orientierung den Maßstab eines Kontinuums von keiner vorliegenden Heiserkeit bis zur größten Ausprägung als Einschätzung mit sämtlichen Zwischenstufen anzugeben.

Inwiefern dies dem validen Identifizieren und Quantifizieren des jeweiligen Parameters durch die entsprechende Skalenart und deren Auflösung entspricht, ist erst durch Verschränkung akustischer und auditiv-perzeptiver Verfahren zu bestätigen. Die Skalierung mit der visuellen Analogskala zeigt in allen Parametern sehr gute Analysemöglichkeiten in Gruppen- und Einzelurteil, die durch die Verbesserung der Werte im Expertenstatus bestätigt werden. Vergleichbare Studien mit kleineren Expertengruppen zeigen für das ORD-Verfahren ähnliche Ergebnisse, sodass auch das ökonomische, aber auch gut austauschbare ORD-Verfahren, das sich im europäischen Raum längst etabliert hat, nicht an Gültigkeit verliert.

Der deutliche Abfall des ORD-Verfahrens zu VAS beim Parameter Rauigkeit lässt darauf schließen, dass mit ihrer beruflichen Tätigkeit beginnende LogopädInnen zum Parameter R stärker geschult werden müssen. Mit der Überlagerung durch andere Aspekte der Stimmqualität, wie z.B. Knarren, Angestrengtheit (Strain), Tonhöhen- wie Resonanzphänomene wird damit die Hypothese bestätigt, dass das Stimmphänomen Rauigkeit zwar als ergänzendes polares Phänomen von Behauchtheit zu betrachten ist (Irregularität der Schwingung bei R vs. Rauschanteil durch fehlenden Glottisschluss bei B), aber insgesamt deutlich schwieriger perzeptiv abzugreifen ist.

In den Schweregraden VAS_1 und VAS_2 zeigt sich deutlich die Überlegenheit des VAS-Verfahrens und dessen Stärke; bei gering bis leichter gestörten Stimmen ist es dem ORD-

Abb. 2: ICC Testhälften ORD vs. VAS

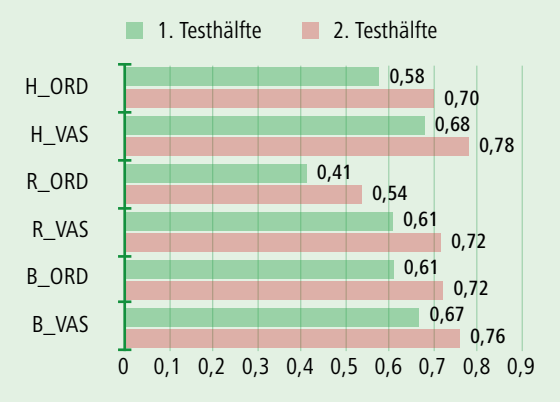


Abb. 3: ICC Experten Anker- (n = 3) vs. Kontrollgruppe (n = 3)

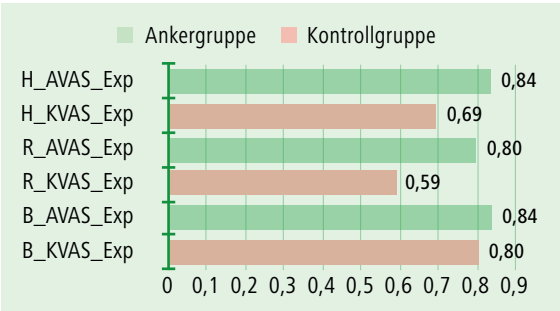
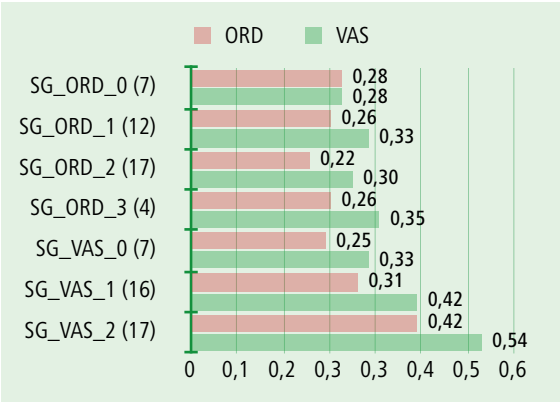


Abb. 4: ICC Mittelwerte (Einzelmaße) Schweregrade ORD vs. VAS



Verfahren nicht überlegen. Interessant und unterstützend für VAS spricht, dass, obwohl das VAS-Verfahren von keinem der Teilnehmer (ORD immerhin 51 %) vorher angewandt wurde, sich ein schneller Lerneffekt und gute Reliabilität einstellten. Inwieweit die hohe Motivation, etwas „Neues" anzuwenden oder gar ein Versuchsleitereffekt, das Verfahren als aktuellste Methode vorzustellen, ein Einflussfaktor waren, kann nicht beantwortet werden. In den Lerneffekten zwischen den Testhälften ist kein bedeutsamer Unterschied zwischen ORD und VAS zu beobachten. Das ORD-Verfahren ist für stark gestörte Stimmen SG_ORD_3 am höchsten reliabel.

Ausblick

Trotz der Maßgabe, dass kein einzelner diagnostischer Baustein der Komplexität der Beschreibung und Quantifizierung stimmlicher Daten gerecht werden kann, bestätigt die vorliegende Studie die gute Beurteilbarkeit der untersuchten Parameter mittels der durch den CAPE-V vorgeschlagenen Beurteilungsart durch eine visuelle Analogskala. Die auditiv-perzeptive Beurteilung ist aufgrund des von diversen Autoren konstatierten Bedingungsgefüges „Stimmklangbeurteilung" Schwankungen und Beurteilungsartefakten unterworfen, die von der allgemeinen Expertise der Rater und ihrer Vorbereitung durch ein Training, den zu beurteilenden Stimmitems, dem angewandten Ratingverfahren und der jeweiligen Nutzung interner und externer Referenzen bestimmt wird (*Kreiman* et al. 2007).

Rauigkeit erscheint den Autoren aufgrund der unterschiedlichen Ausprägungsarten und Überlagerung mit den anderen Stimmeigenschaften als komplexestes Merkmal und verständlicherweise ausführlicher zu trainieren. Die Möglichkeit zur kontinuierlichen Abstufung mittels VAS scheint der „Königsweg" zur besseren Quantifizierung im Gegensatz zur eingeschränkt abgestuften Abbildbarkeit mittels ORD zu sein. Der Autor sieht in der Anwendung des CAPE-V-Protokolls einen notwendigen internationalen Konsens vorgegeben und durch viele aktuelle Arbeiten bestätigt. Es zeigt sich deutlich, dass für Forschungsaspekte kleinere Expertengruppen die Genauigkeit der Ergebnisse eindeutig erhöhen und somit größeren Gruppen vorzuziehen sind.

Die Untersuchung der vom CAPE-V vorgeschlagenen weiteren Stimmparameter wie Tonhöhe, Lautstärke und Anstrengungsgrad in klinischen und Forschungsdesigns steht an (*Kelchner* et al. 2010). Software-gestützte akustische Analyseverfahren könnten in Kombination mit auditiv-perzeptiver Analyse dazu dienen, diese vergleichend auf ihre akustischen Parameter zu analysieren und über Korrelationsanalysen die Validität der Stimmbewertung weiter zu steigern (*Batalla* et al. 2004).

Eine Integration des Trainings auditiv-perzeptiver Fähigkeiten der Stimmdiagnostik in die Ausbildung logopädischer Berufsgruppen erscheint uns unabdingbar und sollte mit mindestens acht Stunden veranschlagt werden. Dies ist nur mit der Eigenerfahrung von Stimmrezeption und -produktion im Sinne einer ausführlichen individuellen praktisch-stimmentwickelnden Übungsbehandlung jedes Logopädiestudierenden zu erreichen, die durch das Erfahren eigener „phonatorischer" Schwierigkeiten eine über die reine Stimmdiagnostik hinausgehende Akzeptanz und respektvolle Haltung vor dem Gesamtphänomen stimmgestörter Mensch im Sinne kommunikativer und psychosomatischer Fragestellungen (*Deuster* et al. 2006) wachhält.

LITERATUR

Awan, S.N. & Lawson, L.L. (2009). The effect of anchor modality on the reliability of vocal severity ratings. Journal of Voice 23, 341-352

Batalla, N., Corte Santos, P., Sequeriros Santiago, G., Senaris Gonzalez, B. & Suarez Nieto, C. (2004). Perceptual evaluation of dysphonia: correlation with acoustic parameters and reliability. *Acta Otorhinolaryngologica Esp.* 55, 282-287

Bele, I.V. (2004). Reliability in perceptual analysis of voice quality. *Journal of Voice* 19, 555-573

Bigenzahn, W. & Schneider, B. (2007). *Diagnostik der Stimme.* Wien: Springer

Chan, K.M.K. & Yiu, E. (2002). The effect of anchors and training on the reliability of perceptual voice evaluation. *Journal of Speech, Language, and Hearing Research* 45, 111-126

Chan, K.M.K. & Yiu, E.M.L. (2006). A comparison of two perceptual voice evaluation training programs for native listeners. *Journal of Voice* 20, 229-241

de Bodt, M.S. & Wuyts, F.L. (1997). Test-retest study of the GRBAS scale: influence of experience and professional background on perceptual rating of voice quality. *Journal of Voice* 11, 74-80

Deuster, D., Knief, A., Schmidt, M., Hübner, R. & am Zehnhoff-Dinnessen, A. (2006). *Stimmstörungen in der phoniatrischen Klinik*. www.egms.de/static/de/meetings/dgpp2006/06dgpp71.shtml (25.04.2014)

Dicks, P. & Nawka, T. (in Vorbereitung). Reliabilität der auditiv-perzeptiven Beurteilung der Heiserkeit organischer Stimmstörungen mittels visueller Analogskala und Ordinalskala unter Einsatz natürlicher Ankerstimmen. *Sprache – Stimme – Gehör*

Eadie, T.L. & Baylor, C.R. (2006). The effect of perceptual training on inexperienced listeners judgements of dysphonic voice. *Journal of Voice* 20, 527-544

Eadie, T.L. & Doyle, P.C. (2005). Classification of dysphonic voice: acoustic and auditory-perceptual measures. *Journal of Voice* 19, 1-14

Evans, R., Nawka, T., Gong, Y. & Gluud, C. (2004). *Auditive Stimmbeurteilung nach dem CAPE-V-Protokoll in einer multizentrischen Studie.* www.egms.de/static/en/meetings/dgpp2004/04dgpp75.shtml (25.04.2014)

Evans, R. & Nawka, T. (2005). *Auditive Stimmbeurteilung nach einer visuellen Analogskala und einer Ordinalskala*. www.e-gms.de/static/de/meetings/dgpp2005/05dgpp060.shtml (25.04.2014)

Friedrich, G. & Dejonckere, P.H. (2005). The voice evaluation protocol of the European Laryngological Society (ELS). *Laryngorhinootologie* 84, 744-752

Gerrat, B.R. & Kreiman, J. (2004). Perceptual evaluation of voice quality. In: Kent, R.D. (Hrsg.), *The MIT encyclopedia of communication disorders* (78-80). Cambridge: Bradford Book MIT Press.

Gonnermann, U. (2007). *Quantifizierbare Verfahren zur Bewertung von Dysphonien.* Frankfurt/M.: Lang

Hammer, S. (2009). *Stimmtherapie mit Erwachsenen.* Heidelberg: Springer

Helou, L.B., Henry, L.R. & Coppit, G.L. (2010). The role of listener experience on Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) Ratings of Postthyroidectomy Voice. *American Journal of Speech Language Pathology* 19, 248-58

Karnell, M.P., Melton, S.D., Childes, J.M., Coleman, T.C., Dailey, S.A. & Hoffman, H.T. (2007). Reliability of clinician-based (GRBAS and CAPE-V) and patient-based (V-RQOL and IPVI) documentation of voice disorders. *Journal of Voice* 2, 576-590

Kelchner, L.N., Brehm, S.B., Weinrich, B., Middendorf, J., de Alarcon, A., Levin, L. & Elluru, R. (2010). Perceptual evaluation of severe pediatric disorders: rater reliability using the Consensus Auditory Perceptual Evaluation of Voice. *Journal of Voice* 24, 441-449

Kempster, G.B., Gerratt, B.R., Verdolini, A.K., Barkmeier-Kraemer, J. & Hillman, R.E. (2009). Consensus Auditory-Perceptual Evaluation of Voice: Development of a standardized clinical protocol. *American Journal of Speech Language Pathology* 18, 124-132

Kreiman, J, Gerratt, B.R., Kempster, G.B., Erman, A. & Berke, G.S. (1993). Perceptual evaluation of voice quality – review, tutorial and a framework for future research. *Journal of Speech Hearing Research* 36, 21-40

Kreiman, J., Gerratt, B.R. & Ito, M. (2007). When and why listeners disagree in voice quality assessment tasks. *Journal of the Acoustical Society of America* 122, 2354-2364

Kreiman, J. & Gerratt, B.R. (2010). Perceptual assessment of voice quality: Past, present, and future. *Perspective on Voice and Voice Disorders* 20, 62-67

MacKenzie, K., Millar, A., Wilson, J.A., Sellars, C. & Deary, I. (2001). Is voice therapy an effective treatment for dysphonia? A randomized controlled trial. *BMJ* 323, 658-661

Martens, J.W., Versnel, H. & Dejonckere, P.H. (2007). The effect of visible speech in the perceptual rating

of pathological voices. *Archives of Otolaryngology – Head and Neck Surgery* 133, 178-185

Mehta, D.D. & Hillmann, R.E. (2008). Voice assessment: updates on perceptual, acoustic, aerodynamic and endoscopic imaging methods. *Current Opinion in Otolaryngology & Head and Neck Surgery* 16, 211-215

Nawka, T. & Evans, R. (2006). *RBH-Training und Diagnostik* (Multimedia CD-Rom, unveröffentlichtes Manual). Forchheim: Wevos

Nawka, T. & Anders, L.C. (1996). Die auditive Beurteilung heiserer Stimmen nach dem RBH-System. (Doppel-CD mit Stimmbeispielen, Manual). Stuttgart: Thieme

Oates, J. (2009). Auditory-perceptual evaluation of disordered voice quality – Pros, cons and future directions. *Folia Phoniatrica et Logopaedica* 61, 49-56

Ptok, M., Schwemmle, C., Iven, C., Jessen, M. & Nawka, T. (2006). On the auditory evaluation of voice quality. *HNO* 54, 793-802

Pützer, M. & Barry, W.J. (2004). *Methodische Aspekte der auditiven Beurteilung von Stimmqualität.* Sprache – Stimme – Gehör 28, 188-197

Sataloff, R.T. (2005). *Clinical assessment of voice.* San Diego: Plural Publishing

Speyer, R. (2006). Effects of voice therapy: a systematic review. *Journal of Voice* 22, 565-580

Stemple, J.C., Glaze, L. & Klaben, B. (2010). *Clinical voice pathology – theory and management.* San Diego: Plural Publishing

Wirtz, M. & Caspar, F. (2002). *Beurteilerübereinstimmung und Beurteilerreliabilität.* Göttingen: Hogrefe

Yui, E.M.L., Chan, K.M. & Mok, R.S. (2007). Reliability and confidence in using a paired comparison paradigm in perceptual voice quality evaluation. *Clinical Linguist & Phonetics* 21, 129-145

Zraick, R.I., Kempster, G.B., Connor, N.P., Thibeault, S. Klaben, B.K., Bursac, Z., Thrush, C.R. & Glaze, L.E. (2011). Establishing validity of the consensus auditory perceptual evaluation of voice disorders (CAPE-V). *American Journal of Speech-Language Pathology* 20 (1), 14-22

SUMMARY. Auditory perceptual evaluation of voice parameters: Results of a test-retest study concerning the assessment of hoarseness – A comparison between Visual Analogue Scale (VAS) and RBH-method

The results of a test-retest study about the reliability of rating the hoarseness in organic dysphonias account in detail for the conditions of vocal assessment. A very good test-retest reliability and a high internal consistency of group judgements were found in all parameters, but roughness proved to be be more difficult to rate than the overall grade of hoarseness and breathiness. Anchor voices helped to get more reliable results. The rating through Visual Analogue Scale is more precise than through ordinal-scaled RBH-classification. The necessity to train auditory perceptual skills was evidenced by considerable learning effects between the first half of the test (voice 1-20) and the second (voice 21-40). Experienced speech therapists achieved high degrees of reliability. The auditory perceptual rating of voice disorders proved to be a reliable element in the general assessment of dysphonias.

KEYWORDS: Hoarseness – auditory-perceptual classification – dysphonia – voice assessment – learning effects – reliability

DOI dieses Beitrags (www.doi.org)
10.2443/skv-s-2014-53020140401

Korrespondenzadresse
Peter Dicks
Diplom-Logopäde
RWTH/Uniklinik Aachen
Schule für Logopädie
Pauwelstr. 30
52074 Aachen
pdicks@ukaachen.de