Kompendium

zur Durchführung von Hörversuchen in Wissenschaft und industrieller Praxis

erarbeitet von:

Wolfgang Ellermeier (Skalierung überschwelliger Größen)

Jürgen Hellbrück (Einführung und Experimentelle Grundlagen)

Armin Kohlrausch (Absolute Schwellen und Unterschiedsschwellen)

Alfred Zeitler (Mehrdimensionale Skalierung; Kontext und Bezugssysteme)

Leitung und redaktionelle Überarbeitung: Jürgen Hellbrück

September 2008

Herausgeber:

Deutsche Gesellschaft für Akustik e.V.

Geschäftsstelle: Voltastraße 5, Gebäude 10-6, 13355 Berlin

Tel.: (0)30 / 46 06 94-63, Fax: (0)30 / 46 06 94-70

dega@dega-akustik.de

www.dega-akustik.de

Mitarbeiter- und Beratungsteam:

Dipl.-Ing. Uwe Bergmann (AUDI AG, Ingolstadt)

Prof. Dr. Rudolf Bisping (SASS, Essen)

Prof. Dr. Wolfgang Ellermeier (TU Darmstadt)

Prof. Dr. Hugo Fastl (TU München)

Dr. Klaus Genuit (HeadAcoustics, Herzogenrath)

Dipl.-Ing. Hans-Peter Grabsch (Robert-Bosch GmbH, Stuttgart)

Prof. Dr. Jürgen Hellbrück (Kath. Universität Eichstätt-Ingolstadt)

Prof. Dr. Armin Kohlrausch (Philips Research Laboratories Eindhoven und Technische Universität Eindhoven, Niederlande)

Dr. Uwe Letens (Daimler AG, Stuttgart)

Dr. Angela Linow (Ford AG, Köln)

Prof. Dr. Joachim Scheuren (Müller-BBM, Planegg)

Prof. Dr. Brigitte Schulte-Fortkamp (TU Berlin)

Dr. Reinhard Weber (Universität Oldenburg)

Dr. Alfred Zeitler (BMW Group, München)

Inhalt

Teil 1: Einführung

Eine Initiative des Fachausschusses Hörakustik der Deutschen Gesellschaft für Akustik (DEGA)

Die Fachgruppe „Methoden der Hörakustik“ innerhalb des Fachausschusses „Hörakustik“ der Deutschen Gesellschaft für Akustik (DEGA) hat sich zum Ziel gesetzt, Qualitätsstandards zu setzen für die Methoden, die bei hörakustischen Versuchen in Wissenschaft und Praxis angewandt werden. Damit soll der Tatsache Rechnung getragen werden, dass sich die DEGA insgesamt und der Fachausschuss „Hörakustik“ insbesondere aus Mitgliedern vieler verschiedener Fachrichtungen (z.B. Physiker, Ingenieure, Psychologen) zusammensetzt, die häufig das Methodeninventar der jeweils anderen Fachrichtungen nur unzureichend kennen und aus dieser Unzulänglichkeit mehr oder weniger schwerwiegende Fehler bei der Planung, Durchführung und Auswertung von hörakustischen Versuchen sowie der Dateninterpretation begehen könnten. Transparenz der Methoden und kritische Einsicht in die jeweiligen Vorgehensweisen dienen nicht nur edukativen Zielen, sondern auch dem Zusammenhalt einer wissenschaftlich-technischen Fachgesellschaft, die wie kaum eine andere heterogen zusammengesetzt ist. Frauen und Männer aus Grundlagen- und Anwendungsforschung sowie aus der industriellen Praxis, kommend aus ganz unterschiedlichen Disziplinen, von Naturwissenschaften über Sozialwissenschaften bis zu Geisteswissenschaften, finden sich unter dem Dach der DEGA. In dieser Interdisziplinarität liegt ein großes Potenzial für Kreativität, aber auch die Gefahr unkritischer bzw. inadäquater Übernahme „fremder“ Methoden. Die Förderung des Verständnisses für die Lösungswege anderer Disziplinen kann beides, nämlich das Potenzial stärken und die Gefahren begrenzen.

Der erste Schritt in diese Richtung sind die hier vorgelegten Empfehlungen mit dem Titel *Hörversuche in Wissenschaft und industrieller Praxis*. Dieses Kompendium soll Wissenschaftlern und Anwendern verschiedener Fachrichtungen Orientierung bei der Planung, Durchführung und Auswertung hörakustischer Versuche bieten. Experten sollen unter Beachtung didaktischer Gesichtspunkte und unter Berücksichtigung der fachlich heterogen zusammengesetzten Zielgruppe über die adäquate methodische Bearbeitung hörakustischer Problemstellungen berichten sowie „Fallstricke“ aufzeigen und Lösungen vorschlagen, wie man sie vermeidet. Dieses Kompendium will insbesondere auch Anwendungen aus der industriellen Praxis berücksichtigen.

Sensorische Bewertung

In der Industrie werden sensorische Bewertungen in relativ großem Umfang durchgeführt. Sie dienen der Qualitätsbewertung eines Produktes im Hinblick auf das Verbraucherverhalten.

> Unter Sensorischer Bewertung (sensory evaluation) versteht man eine Disziplin, die dazu dient, Reaktionen auf Merkmale und Eigenschaften von Produkten oder Materialen auszulösen, zu messen, zu analysieren und zu interpretieren, die über den Gesichtssinn, den Geschmacks- und Geruchssinn sowie über den Tast- und Gehörsinn wahrgenommen werden. (nach Ruan & Zeng, 2004, 1).

Für den Verbraucher sind sensorische Eindrücke meist die einzige Möglichkeit, direkt die Qualität und Eigenschaften eines Produktes zu bewerten. Die Hersteller bemühen sich, ihre Produkte positiv erscheinen zu lassen, wohl wissend dass auch Eindrücke über Teilaspekte eines Produktes den Konsumenten häufig auf die gesamte Qualität des Produktes schließen lassen. Der Klang der ins Schloss fallenden Autotür kann für die Qualitätsbeurteilung des jeweiligen Fahrzeuges eine Leitfunktion ausüben, die sich beim Kunden auf das Fahrzeug insgesamt, ja auf die gesamte Marke ausweiten kann.

Es besteht in der Industrie ein großes Bedürfnis, sensorische Bewertungsverfahren möglichst objektiv und genau und nach Möglichkeit ohne Versuchspersonen, also rein instrumentell durchzuführen. Auch in der Hörakustik gibt es hochentwickelte instrumentelle psychoakustische Messverfahren, die auf physikalischen Messungen basieren, aber die Daten entsprechend dem gegenwärtigen Stand psychoakustischer Funktionsmodelle gewichten und integrieren. Dennoch sind Hörversuche mit Versuchspersonen heute immer noch unerlässlich, denn zum einen sind bestimmte wahrgenommene akustische Eigenschaften wie „scheppernd“ oder „knirschend“ instrumentell, d.h. basierend auf physikalischen Messungen nicht fassbar; zum anderen sind hörakustische Versuche unter Einbeziehung repräsentativer Personen im Hinblick auf das Verbraucherverhalten von höherer Validität.

Literaturempfehlung

Bech, S. & Zacharov, N. (2006). Perceptual audio evaluation. Chichester: Wiley.

Teil 2: Hörversuch – Experimentelle Grundlagen

Der Hörversuch

Definition

Unter dem Begriff „Hörversuch“ wird eine *planmäßige*, *wiederholbare* und unter *kontrollierten Bedingungen* durchgeführte Untersuchung mit Probanden verstanden, denen akustische Reize präsentiert werden, die von den Probanden nach entsprechender Instruktion beurteilt werden. Ein Hörversuch stellt somit eine experimentelle Untersuchung dar. Unsystematische, eher beiläufig erhobene Beobachtungen können unter Umständen wertvolle Anregungen liefern oder interessante Fragen aufwerfen, stellen jedoch keine Hörversuche dar.

Gegenstand eines Hörversuchs

Die Urteile in einem Hörversuch können verbal geäußert werden oder durch motorische Reaktion, z.B. Tastendruck erfolgen.

Die Urteile beziehen sich in der Regel darauf,

- ob ein akustischer Reiz hörbar ist oder nicht,
- ob sich zwei akustische Reize hörbar unterscheiden oder nicht,
- welchen Ausprägungsgrad ein akustischer Reiz auf einer Wahrnehmungsdimension (z.B. der wahrgenommenen Lautstärke) besitzt,
- welche Position ein akustischer Reiz in einem mehrdimensionalen Wahrnehmungsraum (z.B. Wohlklang) besitzt.

Ziele eines Hörversuchs

Hörversuche werden aus wissenschaftlichem Erkenntnisinteresse durchgeführt oder dienen der Qualitätsbewertung eines Produkts im Hinblick auf das Verbraucherverhalten.

Hörversuche in der Industrie fallen unter die Kategorie der sensorischen Bewertungen. Sie dienen der Qualitätsbewertung eines Produktes im Hinblick auf das Verbraucherverhalten (siehe oben). Produkte, die im Zusammenhang mit Hörversuchen von Relevanz sind, können beispielsweise Geräusch erzeugende Maschinen und Geräte sein, Geräte der Unterhaltungselektronik sowie Musikinstrumente oder auch die akustischen Eigenschaften eines Raumes.

Das Experiment

Fragestellung und Hypothesen

Fragestellung

Ein Hörversuch gründet auf einer Fragestellung, die prinzipiell empirisch beantwortbar sein muss. Spekulative Fragen (z.B. über die Existenz und Einfluss der Sphärenmusik) sind kein Gegenstand von Hörversuchen.

Man unterscheidet in einer empirischen Untersuchung ein- und zweiseitige Fragestellungen. Wenn wir z.B. wissen wollen, ob zwei Staubsauger unterschiedlich klingen (ob der Staubsauger A z.B. angenehmer klingt als der Staubsauger B oder ob umgekehrt Staubsauger B angenehmer klingt als Staubsauger A) dann handelt es sich um eine zweiseitige Fragestellung. Fragen wir uns jedoch, ob eine schalltechnische Innovation bei Staubsauger B zu einer bedeutenden Verbesserung des Klangerlebens gegenüber Staubsauger A führt, liegt eine einseitige Fragestellung vor.

Hypothesen: Nullhypothese H_0 und Alternativhypothese H_1

Hörversuche gehen wie alle Experimente von Hypothesen aus. Unter Hypothesen versteht man Antwortmöglichkeiten auf die gestellte Frage. Hypothesen müssen im Einklang mit dem vorhandenen Erfahrungsschatz und den gültigen Theorien sein bzw. sich logisch daraus ableiten lassen. Die Begründung muss dem aktuellen Stand der Erkenntnis entsprechen und die Erreichbarkeit des Untersuchungsziels erkennbar machen. Durch methodische, d.h. planmäßige Vorgehensweise wird eine statistisch abgesicherte Entscheidung über die möglichen Antworten (Hypothesen) gesucht.

Die statistische Beweisführung, die am Ende der Untersuchung steht, muss bereits bei der Hypothesenformulierung und -begründung mitbedacht werden. Die statistische Beweisführung geht von der Nullhypothese aus. Diese besagt beispielsweise: Zwei oder mehr Stichproben von Messwerten entstammen derselben Grundgesamtheit. Die Nullhypothese wird H_0 abgekürzt.

Der Nullhypothese wird eine Alternativhypothese H_1 gegenüber gestellt. Die Alternativhypothese ergibt sich aus der Untersuchungsfrage und stellt die Alternative zur Nullhypothese dar: Zwei (oder mehr) Stichproben von Messwerten stammen aus verschiedenen Grundgesamtheiten. Sie bezieht sich auf die gleichen Parameter wie die Nullhypothese. Bei der Frage, ob zwei (oder mehr) Stichproben aus derselben oder verschiedenen Grundgesamtheiten stammen, wird in der Regel auf einen bestimmten Verteilungsparameter bezogen, in der Regel der Mittelwert der Verteilung der Messwerte.

Man unterscheidet ungerichtete und gerichtete Alternativhypothesen. Wenn die Alternativhypothese besagt, dass die Stichproben aus Grundgesamtheiten mit unterschiedlichen Mittelwerten stammen, handelt es sich um eine ungerichtete (oder auch zweiseitige) Alternativhypothese; besagt sie beispielsweise jedoch, dass der Mittelwert aus Grundgesamtheit A größer ist als der von Grundgesamtheit B, haben wir eine gerichtete (oder einseitige) Alternativhypothese vor uns.

Auf den Untersuchungsgegenstand bezogen handelt es sich hierbei um die zwei- oder einseitige Fragestellung (siehe oben).

Nach Formulierung von Null- und Alternativhypothese wird die Untersuchung durchgeführt. Dabei stellt sich die Frage der Versuchsplanung.

Versuchsplanung

Systematik von Versuchsplänen

Ein Versuchsplan ist ein standardisiertes Untersuchungsschema, das Hypothesen, Versuchaufbau und –durchführung sowie die statistische Datenanalyse und statistische Entscheidung logisch miteinander verknüpft.

Es gibt eine Vielzahl von standardisierten Versuchsplänen. Eine wichtige Rolle bei allen Versuchsplänen spielt das Zufallsprinzip (Randomisierung).

Messwiederholungsplan und Zufallsgruppenplan

Es stellt sich häufig die Frage, ob jede Versuchsperson unter jeder experimentellen Bedingung getestet werden soll (verbundene Stichprobe) oder unter jeder Bedingung unterschiedliche Versuchspersonengruppen, die nach dem Zufallsprinzip aus einer Gesamtstichprobe zusammengestellt werden.

Im ersten Fall handelt es sich um einen Messwiederholungsplan, im zweiten Fall um einen Zufallsgruppenplan. Im Messwiederholungsplan (auch „Within-subject-Design“) spielt die natürliche Varianz zwischen den Versuchspersonen eine andere Rolle als im Zufallsgruppenplan (auch „Between-subject-Design“). Prinzipiell gilt, dass für einen Zufallsgruppenplan wesentlich mehr Versuchspersonen benötigt werden als für einen Messwiederholungsplan. Die natürlichen individuellen Unterschiede der Versuchspersonen bedingen einen Zufallsfehler, der nur durch Mittelung über eine ausreichend große Anzahl von Versuchspersonen reduziert werden kann. Der Vorteil des Zufallsgruppenplans ist die höhere Repräsentativität bzw. bessere Generalisierbarkeit der Ergebnisse.

Bei einem Messwiederholungsplan werden Unterschiede zwischen den experimentellen Bedingung leichter entdeckt (sofern sie vorhanden sind), da die

individuellen Unterschiede zwischen den Versuchspersonen als Fehlervarianz vernachlässigt werden können. In einem Messwiederholungsplan muss jedoch die Abfolge der experimentellen Bedingungen, denen jede Versuchsperson unterzogen wird, so variiert werden, dass keine systematischen Sequenzeffekte entstehen. Würde jede Versuchsperson zuerst unter der Bedingung A getestet, dann unter Bedingung B, wären die Versuchspersonen versuchstechnisch gesehen unter der Bedingung B nicht mehr identisch mit denen unter Bedingung A, auch wenn es sich um die selben Personen handelt; denn unter B haben sie alle eine Erfahrung im Gedächtnis gespeichert, die sie unter A noch nicht hatten. Diese Erfahrung könnte sich auf das Ergebnis in systematischer Weise auswirken.

In Hörversuchen ist es häufig so, dass mehrere akustische Stimuli, z.B. Produktschalle von jeder Versuchsperson je einzeln beurteilt werden, z.B. mittels Verfahren der direkten Skalierung (siehe unten). Akustische Reize können ja nicht gleichzeitig, sondern müssen sukzessive zur Beurteilung dargeboten werden, Hier handelt es sich dann um einen typischen Messwiederholungsplan (Between-subject-Design), bei dem sorgfältig darauf zu achten, dass sich Sequenzeffekte nicht systematisch auswirken.

Zufallsfehler und konstante Fehler

Ein Hörversuch gründet auf systematischen Beobachtungen und somit auf empirischen Tatbeständen. Wenn Relationen zwischen empirischen Tatbeständen in regelhafter Weise durch numerische Relationen abgebildet werden, spricht man von Messung (liberale Definition von Messung). Messungen beinhalten prinzipiell Messfehler. Daher werden Messungen wiederholt, um zu einer möglichst sicheren Schätzung des wahren Messwertes zu gelangen. Als beste Schätzwerte für den wahren Messwert gelten der Mittelwert (arithmetisches Mittel), sofern die einzelnen Messwerte intervallskaliert und normalverteilt sind, oder der Median, wenn die Daten nur rangskaliert sind oder keine Normalverteilung vorliegt.

Bei den Messfehlern sind *Zufallsfehler* und *konstante Fehler* zu unterscheiden. Erstere betreffen die Messgenauigkeit eines Verfahrens, letztere systematische Verzerrungen des Messergebnisses. (Eine systematische Verzerrung der Messung bezeichnet man auch als *Bias*.) Da bei Hörversuchen ein urteilender Mensch als „Messinstrument“ eingesetzt wird, sind hinsichtlich der Messfehler in einem Hörversuch - neben technisch bzw. apparativ bedingten Ungenauigkeiten und anderen äußeren zufälligen Veränderungen während eines Versuchs oder zwischen einzelnen Versuchen - auch psychische Funktionen in Rechnung zu stellen, welche die Zuverlässigkeit der Messwerte beeinflussen. Messungenauigkeiten können

beispielsweise durch unsystematische Aufmerksamkeitsschwankungen oder zufällige Aufmerksamkeitsablenkungen bedingt sein, konstante Fehler dagegen durch systematische Einflussfaktoren, beispielsweise durch bestimmte Gedächtniseffekte oder Übungseffekte. Letztere sind vor allem in einem Messwiederholungsplan als Sequenzeffekte zu beachten (siehe oben).

Je kleiner der Zufallsfehler ist, umso genauer ist die Schätzung des wahren Wertes. Als Maß dient der Standardfehler des Mittelwerts. Der Standardfehler ergibt sich durch Division der Standardabweichung durch die Quadratwurzel des Stichprobenumfangs. Während die Standardabweichung (also die gemessene Streuung) anzeigt, wie eng die Messwerte beieinander liegen, stellt der Standardfehler eine Normierung der Standardabweichung auf den Stichprobenumfang dar. Er ist ein Maß für die Güte (Vertrauenswürdigkeit) des Mittelwerts.

Auch durch Auswahl der Versuchspersonen sowie durch die Zuordnung der Versuchspersonen zu den Versuchsbedingungen können Fehler verursacht werden, und zwar in der Regel immer dann, wenn das Zufallsprinzip bei der Auswahl und der Zuordnung der Versuchpersonen verletzt wird. Bei der Auswahl der Versuchspersonen ist zu fragen, für welche Grundgesamtheit (Population) dieses Ergebnis des Versuchs repräsentativ sein soll. Dann muss aus dieser Population eine Zufallsstichprobe gezogen werden. Die Zuordnung der Versuchspersonen zu den experimentellen Bedingungen nach dem Zufallsprinzip ist ein wichtiges Kriterium für einen experimentellen Plan. Ist keine Zufallszuordnung möglich, weil beispielsweise die Zuordnung natürlicherweise vorgegeben ist, spricht man von einem quasi-experimentellen Plan.

Unabhängige und abhängige Variable

Mit dem Begriff „Variable“ meint man die Ausprägung eines Merkmals, die mindestens zwei Werte annehmen kann. Die in einem Experiment gesetzte Bedingung bezeichnet man als unabhängige Variable (UV). UV kann beispielsweise der Schallpegel sein, der mehrfach gestuft als experimentelle Bedingung eingesetzt wird. Die in Frage stehenden und gemessenen Reaktionen der Versuchspersonen stellen die abhängige Variable (AV) dar, z.B. das Urteil über die empfundene Lautstärke. Im Hörversuch ist die UV in der Regel eine physikalisch gemessene Größe, die AV dagegen eine psychologische Messgröße.

Versuchspersonen (Vpn)

Hörvermögen

Versuchspersonen, die für einen Hörversuch angeworben werden, sollen über ein normales Hörvermögen verfügen, das gegebenenfalls durch ein Tonschwellenaudiogramm oder durch Selbstauskunft (Fragebogen) überprüft wird.

Individuelle Unterschiede

Auch wenn gewährleistet ist, dass die Versuchspersonen hinsichtlich wichtiger untersuchungsrelevanter Eigenschaften keine wesentlichen Unterschiede aufweisen, unterscheiden sie sich doch in vielen anderen Eigenschaften, wie intellektuellen Eigenschaften und Persönlichkeitseigenschaften, die in gewissem Rahmen auch die Reaktionen in einem Hörversuch beeinflussen können. Damit sich solche Eigenschaften nicht systematisch auf das Ergebnis auswirken, müssen sie kontrolliert werden, in der Regel durch zufallsbedingte Auswahl und Zuordnung der Versuchspersonen. Bei zufälliger Zuordnung der Versuchspersonen zu den experimentellen Bedingungen kann davon ausgegangen werden, dass sich unerwünschte Wirkungen von untersuchungsirrelevanten Eigenschaften zu Null addieren.

Expert Panels und Consumer Panels

Bei sensorischen Produktbewertungen greift man in der Regel auf spezielle Stichproben (Panels) zurück: Man unterscheidet *Expert Panels* und *Consumer Panels*. Erstere bestehen aus Personen, die in der Beurteilung der jeweiligen Merkmale geübt bzw. trainiert sind. Expert Panels werden in der Regel bei dem Design und der Entwicklung eines Produktes eingesetzt. Sie beurteilen die in Frage stehenden Produktmerkmale im Wesentlichen neutral und nicht unter hedonischen Gesichtspunkten, d.h. nach Gefallen bzw. Nicht-Gefallen. Das Consumer Panel besteht in der Regel aus nicht-trainierten Personen, die die potenziellen Käufer repräsentieren und die Akzeptanz des Produktes hinsichtlich zukünftiger Kaufentscheidungen beurteilen. Das Expert Panel erarbeitet das sensorische Profil des Produktes, das Consumer Panel die hedonische Bewertung (Präferenzurteile).

Ethische Grundsätze

Allgemein gelten die ethischen Grundsätze für wissenschaftliche Untersuchungen, in denen Versuchspersonen beteiligt sind. Dazu zählen Freiwilligkeit der Versuchsteilnahme, Datenschutz sowie Gewährleistung der körperlichen und psychischen Unversehrtheit. Die jeweils zuständigen Ethikkommissionen sind gegebenenfalls zu informieren.

Man sollte sich in diesem Zusammenhang auch bewusst sein, dass die selbstverständlich zu garantierende Freiwilligkeit der Versuchsteilnahme auch ein Selektionskriterium sein könnte. So kann es bei bestimmten Untersuchungen schon eine bedenkenswerte Frage sein, warum manche Personen, die Teilnahme am Versuch ablehnen und andere sie bereitwillig akzeptieren, und ob dadurch das Versuchsergebnis systematisch beeinflusst werden könnte.

Statistische Datenanalyse

Skalenniveaus

Die Datenauswertung muss dem Skalenniveau der Daten angemessen sein. Wenn dies nicht der Fall ist, riskiert man Fehlinterpretationen. Man unterscheidet

1. Nominalskalen (nur qualitative Merkmale; Beispiel: Augenfarbe)
2. Rangskalen (Ordnungsrelationen interpretierbar; keine Aussagen über Abstände; Beispiel: Richterskala für Erdbebenstärke)
3. Intervallskalen (Abstände interpretierbar; keine Aussagen über Verhältnisse, da kein absoluter Nullpunkt gegeben; Beispiel: Celsiusskala)
4. Verhältnisskalen (absoluter Nullpunkt vorhanden, daher Verhältnisse interpretierbar; Beispiel: klassische Skalen des cm-g-s-Systems)

Bilden die Zahlen nur Rangfolgen zwischen den empirischen Gegebenheiten ab, können Abstände zwischen den Zahlen nicht interpretiert werden. Arithmetische Mittel und Standardabweichungen zu berechnen, wäre in diesem Fall nicht sinnvoll. Median und Interquartilabstände sind dann als repräsentative Maße für die Verteilung der Messdaten geeigneter. Ob subjektive Urteilsskalen beispielsweise Intervallskalenniveau (Gleichabständigkeit der Skaleneinheiten) oder gar Verhältnisskalenniveau besitzen, ist nicht immer leicht zu beantworten und wird in den Messtheorien zum Teil kontrovers diskutiert. Es ist ratsam, die Verteilung der Daten zu überprüfen. Handelt es sich um eine Normalverteilung fallen arithmetisches Mittel und Median zusammen.

Parametrische und non-parametrische Kenngrößen und Verfahren

Prinzipiell gilt, dass wenn Intervallskalenniveau gegeben ist und Normalverteilung der Messwerte in der Population (unbeschadet ihrer Verteilung in den Stichproben) vorausgesetzt werden kann, parametrische statistische Kenngrößen (z.B. arithmetisches Mittel und Standardabweichung) und statistische Verfahren (z.B. t-Test oder Produktmoment-Korrelation) verwendet werden sollten, da sie mehr Information ausnutzen. Wenn lediglich Rangskalenniveau vorliegt, scheiden parametrische Verfahren von vornherein aus. Auch wenn die Verteilung unbekannt ist, sollte auf non-

parametrische (verteilungsfreie) statistische Kenngrößen (z.B. Median und Interquartilabstände) und Verfahren (z.B. U-Test oder Spearmansche Rangkorrelation) zurückgegriffen werden.

Auch auf Maßskalen, die nicht linear, sondern z.B. geometrisch gestuft sind, kann das arithmetische Mittel nicht angewandt werden. Manchmal lässt sich jedoch eine Maßskala auch so transformieren, dass eine Normalverteilung der transformierten Daten resultiert (z.B. durch Logarithmierung). Dabei ist jedoch zu beachten, dass die Transformation sachlogisch und nicht nur zahlenlogisch begründet ist.

(Als Beispiel sei verwiesen auf Lautheitsurteile nach dem Größenschätzverfahren, die beispielsweise der sone-Skala der Lautheit zugrunde liegen; siehe unten).

Die statistische Auswertung von Daten, beispielsweise zur Frage, ob sich bestimmte Produktschalle in der direkten Beurteilung der Lautstärke unterscheiden, kann varianzanalytisch erfolgen. Dabei ist zu berücksichtigen, ob es sich um einen Messwiederholungsplan oder einen Zufallsgruppenplan handelt. Für beide Versuchspläne gibt es geeignete statistische Verfahren. Ebenso gibt es Varianzanalyse nicht nur für intervallskalierte Daten, sondern auch für Daten auf Rangskalenniveau.

Die statistische Entscheidung

Wenn Nullhypothese und Alternativhypothese klar formuliert, die Untersuchung stringent und sorgfältig durchgeführt und die Daten mit den geeigneten statistischen Verfahren analysiert wurden, erhält man als letzten Schritt der statistischen Analyse einen Wahrscheinlichkeitswert p. Dieser sagt uns – umgerechnet in Prozent – in wie vielen von 100 Untersuchungen dieser Art wir einen solchen oder höheren Stichprobenunterschied im Durchschnitt finden würden, unter der Voraussetzung, dass die Nullhypothese zutrifft. Wenn diese Wahrscheinlichkeit bzw. der Prozentwert sehr gering ist, wird die Nullhypothese aufgegeben und die Alternativhypothese angenommen. Das Risiko, das wir mit dieser Entscheidung eingehen, indem wir die Nullhypothese zu Unrecht verwerfen, nennt man Risiko I. Es wird mit α bezeichnet und kann Werte zwischen 0 und 1 annehmen. Die Ergebnisse, aufgrund deren wir die Nullhypothese aufgeben, nennt man „signifikant“ oder „statistisch gesichert“.

Ab welchem α darf man die Nullhypothese verwerfen? Es gibt im Wesentlichen drei (per Konvention festgelegte) Signifikanzniveaus, $\alpha = 0{,}05$, $\alpha = 0{,}01$ und $\alpha = 0{,}001$. Entsprechend nennt man ein Ergebnis z.B. signifikant auf dem 5 %-Niveau oder auf dem 1 %-Niveau bzw. 0,1 %-Niveau. Das statistische Signifikanzniveau ist vor einer Untersuchung festzulegen entsprechend der Fragestellung und anhand von

Überlegungen über die Wichtigkeit der Aussage, die aus den Untersuchungsergebnissen gefolgert wird. Wäre etwa mit der Annahme der Alternativhypothese – z.B. ein signifikanter Mittelwertsunterschied zwischen zwei Stichproben von Produktgeräuschen ist gegeben – ein großer technischer und finanzieller Aufwand verbunden, um beispielswiese den vermeintlichen Unterschied zu beseitigen, sollte auch ein entsprechend hohes Signifikanzniveau gewählt werden, 1 %- oder gar ein 0,1 %-Niveau, um mit einer hohen Wahrscheinlichkeit davon ausgehen zu können, dass ein solcher Unterschied auch wirklich vorhanden ist.

Interpretation der Ergebnisse

Statistische Ergebnisse stehen nicht für sich allein. Die numerischen Ergebnisse und statistischen Analysen bedürfen der Interpretation in Worten. Sie müssen logisch im Hinblick auf die Ausgangsfrage der Untersuchung interpretiert werden, sodass nach Möglichkeit eindeutige Schlussfolgerungen aus der Untersuchung gezogen werden können. Der Erkenntnisgewinn muss deutlich herausgestellt werden. Unerwartete, d.h. nicht hypothesenkonforme Ergebnisse können unter Umständen interessant und mitteilenswert sein. Sie können zu neuen Untersuchungen anregen.

Fallstricke

Der „texanische Scharfschütze“ und die Post-hoc-Erklärung

Auf keinen Fall dürfen Ausgangsfrage und die ursprünglichen Hypothesen aufgrund der Ergebnisse des Versuchs im Nachhinein (post-hoc) umformuliert werden. Das entspräche dem Beispiel des „texanischen Scharfschützen“, der auf ein Scheunentor schießt und danach um die Einschusslöcher die Zielscheibe malt.

„Pilze im Wald“ und das Problem der Nullhypothese

Besondere Vorsicht ist geboten, wenn ein Unterschied zwischen zwei oder mehreren Bedingungen erwartet wurde (Alternativhypothese H_1), aber statistisch nicht nachgewiesen werden konnte, sodass die Nullhypothese (H_0) beibehalten wurde. Hieraus darf nicht geschlossen werden, dass zwischen den Bedingungen kein Unterschied bestehe! Das Ergebnis darf nur dahingehend interpretiert werden, dass innerhalb der vorliegenden Untersuchung und der verwendeten Methode kein Unterschied statistisch signifikant nachgewiesen werden konnte. Analoges Beispiel: Wer in einem bestimmten Wald Pilze sucht und keine findet, kann nicht mit Recht behaupten, dass in diesem Wald keine Pilze sind. Die Nullhypothese – dass etwas nicht ist – ist streng genommen nicht beweisbar. Es könnte sein, dass – um im genannten Beispiel zu bleiben – wirklich keine Pilze in dem Wald sind; es kann aber auch sein, dass die Suchmethode zu nachlässig war, um die möglicherweise sogar

zahlreich vorhandenen Pilze zu entdecken. Es könnte auch sein, dass nur ganz wenige Pilze, unter Umständen sehr versteckt im Wald sind, sodass ein sehr hoher Aufwand betrieben werden müsste, um sie zu finden. Letzteres würde sich nur dann lohnen, wenn der Pilz sehr wertvoll wäre.

Der „kluge Hans“ und der Doppelblind-Versuch

Der „kluge Hans“ war ein Pferd, das laut seinem Besitzer rechnen konnte, wobei es die Lösungen der Rechenaufgaben durch Aufstampfen mit einem Huf anzeigte. Es konnte jedoch nachgewiesen werden, dass der kluge Hans auf das Mienenspiel und nahezu unmerkliche Reaktionen seines Herrn reagierte, wenn er sich mit dem Klopfen seines Hufes der Lösung näherte. Dies war seinem Besitzer selbst nicht bewusst, der davon ausging, dass der kluge Hans wirklich rechnen konnte.

Das Beispiel verweist auf die Bedeutung der Interaktion zwischen Versuchsleiter und Versuchsperson. Erwartungen des Versuchsleiters an das Ergebnis können der Versuchsperson auf verschiedene, oft sehr subtile Art kommuniziert werden. Wenn diese Gefahr besteht, sollte ein Doppelblind-Versuch durchgeführt werden, bei dem weder die Versuchsperson (Blindung) noch der direkte Versuchsleiter (Doppelblindung) über die dem Versuch zugrunde liegenden Hypothesen informiert sind, sodass sich Erwartungen nicht systematisch auswirken können. Hinweise, die bestimmte Erwartungen nahelegen könnten, sollten vermieden werden. So könnten beispielsweise bei der Beurteilung von Fahrzeuggeräuschen Hinweise auf die Automarke die Beurteilung systematisch beeinflussen. Hinweis: Dass Versuchspersonen mit bestimmten Erwartungen einen Versuch durchführen ist nicht ungewöhnlich, wichtig ist nur, dass ihnen nicht in einer systematischen Weise Erwartungen nahegelegt werden.

Die Instruktion zur Versuchsaufgabe ist wichtigster Teil der Interaktion zwischen Versuchsleiter und Versuchsperson. Auch wenn die Instruktion bei jeder Versuchsperson wörtlich gleich ist, könnten bei mündlichem Vortrag durch Intonation etc. bestimmte Tendenzen nahegelegt werden. Besser wäre, die Instruktion schriftlich zum Lesen vorzulegen.

Gütekriterien eines Hörversuchs

Objektivität

Objektivität im Zusammenhang mit einem Hörversuch meint den Grad der Unabhängigkeit der Reaktionen der Versuchsperson von den individuellen Eigenschaften und Verhaltensweisen des Versuchsleiters. In einem Hörversuch interagieren Versuchsleiter mit Versuchspersonen. Einflussmöglichkeiten ergeben

sich auf unterschiedlichen Ebenen, bei der Durchführung des Versuchs, bei der Auswertung der Daten und bei der Interpretation der Ergebnisse. Um Einflussmöglichkeiten zu vermeiden, sollte die Instruktion der Versuchsaufgabe schriftlich den Versuchspersonen zum Lesen präsentiert werden und die Versuchssteuerung sowie die Erfassung der Reaktionen der Versuchspersonen sollten nach Möglichkeit über ein Computerprogramm erfolgen. Dies verringert auch Fehler bei der Reizauslösung oder Datenübertragung, die sich versehentlich oder durch Nachlässigkeiten ergeben könnten.

Interne Validität eines Hörversuchs

Der Begriff „interne Validität“ bezieht sich auf die Strenge, mit der ein Hörversuch kontrolliert wird. Kontrolle muss in einem Experiment über potenzielle Störvariablen ausgeübt werden, die die Verlässlichkeit der experimentellen Ergebnisse in Frage stellen könnten. (Was als eine Störvariable in Betracht zu ziehen ist, muss theoretisch begründet sein.) Je strenger ein Experiment kontrolliert wird, umso geringer ist die Fehlervarianz und umso deutlicher können die durch die experimentellen Bedingungen verursachten systematischen Effekte hervortreten (sofern sie vorhanden sind).

Externe Validität eines Hörversuchs

Externe Validität meint die Übertragbarkeit der experimentellen Ergebnisse auf die Wirklichkeit außerhalb des experimentellen Settings. Wenn ein Experiment intern sehr streng kontrolliert wird, und damit über eine hohe interne Validität verfügt, ist die Möglichkeit relativ groß, dass es wirklichkeitsfremd ist und Gültigkeit (Validität) mehr oder weniger nur für die spezielle Laborsituation besitzt. Bei Hörversuchen zu psychoakustischen Phänomenen im engeren Sinn bzw. bei physiologienahen Experimenten kann eine streng kontrollierte Laborsituation auch eine hohe externe Validität besitzen, denn sie bezieht sich auf Funktionen des Organismus, der im Labor der gleiche ist wie außerhalb des Labors; bei Hörversuchen beispielsweise zur Präferenz von Beschleunigungsgeräuschen von Automobilen ist möglicherweise eine strenge Isolation der Bedingungen im laborexperimentellen Setting der externen Validität abträglich, etwa dann, wenn beispielsweise kinästhetische Empfindungen als eine wichtige Kontextbedingung für die Beurteilung der Geräusche erachtet werden. Hier könnte sich ein Feldversuch als sinnvoller als ein Laborversuch erweisen, auch wenn im Feldversuch eine höhere Fehlervarianz einzukalkulieren wäre.

Über das Verhältnis zwischen interner und externer Validität muss je nach Untersuchungsziel diskutiert und befunden werden.

Ökonomie und Poweranalyse

Ein Versuch sollte so sparsam wie möglich durchgeführt werden. Eine häufig gestellte Frage betrifft die nötige Anzahl der Versuchspersonen bzw. der Messwiederholungen. Generell gilt, je höher die Anzahl der Messwiederholungen umso kleiner wird der Messfehler (Standardfehler) und umso besser die Schätzung des wahren Mittelwerts. Konsequenterweise bedeutet dies, dass die Nullhypothese theoretisch gesehen bei ausreichend vielen Messwiederholungen immer widerlegbar ist, wobei die Mittelwertsunterschiede auch extrem klein ausfallen können und damit praktisch gesehen keine Relevanz mehr besitzen.

Folgende Kriterien entscheiden darüber, wie hoch ein Messaufwand zu betreiben ist,

1. das festzulegende statistische Signifikanzniveau
2. die Größe des Mittelwertsunterschiedes, die man für praktisch relevant hält, und
3. die zu erwartende oder anzunehmende Streuung der Messwerte.

Auf der Basis dieser Werte kann man im Rahmen einer sogenannten Power-Analyse die Anzahl der Messwiederholungen bzw. die Anzahl der benötigten Versuchspersonen berechnen. Das statistische Signifikanzniveau wird entsprechend der Fragestellung festgelegt (s.o.); wie groß z.B. ein Mittelwertsunterschied sein sollte, ist meist von praktischen Überlegungen abhängig; mit welcher Streuung der Messwerte zu rechnen ist, weiß man häufig von früheren oder von ähnlichen Messungen (vgl. z.B. Sedlmeier & Renkewitz, 2008, 383f.).

Nützlichkeit

Ein Hörversuch sollte auch nützlich sein, d.h. er sollte eine für Theorie oder Praxis relevante Frage zufriedenstellend beantworten. Dies bedeutet konkret, dass vor der Durchführung eines Hörversuchs sorgfältig geprüft werden sollte, ob die jeweilige Frage relevant ist, ob zu ihrer Beantwortung ein Hörversuch notwendig ist und ob die Frage nicht schon durch frühere Untersuchungen beantwortet ist. Dies setzt voraus, dass vor der Planung und Durchführung eines Hörversuchs ausführlich über die Fragestellung nachgedacht und im Kollegenkreis diskutiert wird, und dass die einschlägige Literatur sorgfältig recherchiert und studiert wird.

Literaturempfehlung

Als Einstieg in das Experiment

Sarris, V. & Reiss, S. (2005). Kurzer Leitfaden der Experimentalpsychologie. München: Pearson

Als Überblick und Vertiefung

Sedlmeier, P. & Renkewitz, F. (2008). Forschungsmethoden und Statistik in der Psychologie. München: Pearson.

Der Klassiker (nur noch in Bibliotheken oder antiquarisch)

Campbell, D.T. & Stanley, J.C. (1963). Experimental and quasi-experimental designs for research. Chicago: Rand McNally.

TEIL 3: MESSMETHODEN

Psychophysik

Grundlegende Begriffe

Psychophysik

Die Psychophysik ist die Wissenschaft von den Beziehungen zwischen Reizen (physikalischen Tatbeständen) und Empfindungen (psychischen Tatbeständen). Die Psychophysik sucht nach quantitativen Gesetzmäßigkeiten. Zwei wichtige Gesetze der Psychophysik sind das Fechnersche Gesetz, das eine logarithmische Beziehung zwischen Reizgröße und Empfindungsgröße postuliert, und das Stevenssche Potenzgesetz, das eine Potenzfunktion annimmt. Die Psychophysik hat wichtige Methoden entwickelt, um Reiz- und Unterschiedsschwellen sowie Empfindungsgrößen zu messen.

Reiz

Unter *Reiz* bzw. *Stimulus* wird in der Hörakustik eine energetische Veränderung verstanden, die das Gehörsystem anregt, zu Empfindungen führt und im Verlauf der Informationsverarbeitung weiterverarbeitet wird und Perzepte (Wahrnehmungen) und Assoziationen auslöst. Der Reizbegriff ist problematischer als es auf den ersten Blick erscheint. In der Psychophysik wird der Reiz in physikalischen Termen beschrieben (objektiver Reiz). Man kann unterscheiden zwischen *nominalem* Reiz, als dem objektiv definierten Reiz und dem *funktionalen* Reiz, als dem vom menschlichen Informationsverarbeitungssystem aufgenommenen, erkannten und weiterverarbeiteten Reiz. Der funktionale Reiz ist abhängig von physiologischen Voraussetzungen (Reizschwelle, Adaptation der Rezeptoren etc.) und psychologischen Voraussetzungen (Aufmerksamkeit, Assoziationen etc.).

Empfindung und Wahrnehmung

Unter Empfindung versteht man einen einfachen subjektiven Vorgang, der sich unmittelbar an die isolierte Reizung eines Sinnesorgans anschließt. Empfindungen können nach vier Grunddimensionen differenziert werden, nämlich Räumlichkeit, Zeitlichkeit, Qualität und Intensität. Die durch eine Stimmgabel ausgelöste Schwingung mit bestimmter Frequenz und Amplitude regt eine bestimmte Stelle auf der Basilarmembran an und ruft die Empfindung eines Tones mit bestimmter Tonhöhe

und Lautstärke hervor. Dieser Ton kann im Raum (mehr oder weniger gut) geortet werden und besitzt eine bestimmte Dauer. Empfindungen sind im Gegensatz zu Vorstellungen „ich-fremd“; d.h. sie werden als außerhalb vom eigenen Ich erlebt und mit einem Gegenstand verbunden. (Die Glocke beispielsweise klingt hoch und laut; man erlebt nicht im Inneren die Empfindung eines Glockentons).

Unter *Wahrnehmung* versteht man einen komplexen aktiven Informationsverarbeitungsvorgang, an dessen Beginn Selektionsprozesse stehen, die – geleitet von Bedürfnissen und impliziten Hypothesen – aus der Vielzahl der Reize bestimmte auswählen, und an dessen Ende eine anschaulich mentale Repräsentation eines Weltausschnittes mit Erkenntnischarakter steht. Dazwischen finden Assoziationsprozesse mit Gedächtnisinhalten statt. In den Wahrnehmungsprozess gehen neben den Reizfaktoren somit auch Erfahrungen ein, die im Wesentlichen von allen Menschen gleichermaßen geteilt werden, im Einzelnen aber auch individuell und situationsbedingt geprägt sein können.

Zwischen Empfindung und Wahrnehmung wird nicht immer begrifflich scharf unterschieden. Es sei hier auf die englische Terminologie verwiesen. Dort wird zwischen „Sensation“ und „Perception“ unterschieden, wobei Sensation im Sinne unseres Begriffes Empfindung im engeren Zusammenhang mit dem Sinnesorgan steht. Am Geruch soll ebenfalls der Unterschied erläutert werden, da dieses Beispiel für viele Menschen aus dem Alltag nachvollziehbar ist: Bei geringer Duftkonzentration haben wir eine (undifferenzierte) Geruchsempfindung, aber erst bei höherer Konzentration erkennen wir wonach es riecht. Im ersten Fall handelt es sich um die Empfindlichkeit des Sinnesorgans, im zweiten Fall um eine Wahrnehmung mit Erkennenscharakter.

Bei einfachsten Signalparametern eines akustischen Reizes, z.B. Frequenz oder Amplitude eines Sinustons, macht es keinen Sinn begrifflich zwischen Empfindung und Wahrnehmung zu unterscheiden.

Urteil und Entscheidung

Urteil ist das Ergebnis eines auf Vergleichsprozessen beruhenden Entscheidungsverhaltens. Die Aufgabe einer Versuchsperson kann beispielsweise darin bestehen zu entscheiden, ob in einem Geräusch ein Ton enthalten ist oder nicht, zwei Geräusche gleich oder verschieden sind, das eine lauter als das andere, ob es doppelt, dreimal oder x-mal so laut wie das andere ist.

Beobachtbar für den Versuchsleiter ist nur das geäußerte Urteil. Das Urteil bzw. der das Urteil repräsentierende Zahlenwert geht auch in die Datenauswertung ein. Interpretiert werden jedoch die den Urteilen zugrunde liegenden Empfindungen und

Wahrnehmungen. Es stellt sich daher die wichtige Frage, ob und welche Einflüsse das Urteil systematisch beeinflussen könnten, sodass zwischen Urteil und Empfindung bzw. Wahrnehmung kein unmittelbarer Zusammenhang besteht und ein Bias in Betracht gezogen werden muss.

Absolute Schwellen und Unterschiedsschwellen

Definitionen

Schwelle

Der Begriff der Schwelle definiert einen Zusammenhang zwischen einem physikalisch bzw. akustisch definierten Parameter (Pegel, Frequenz, Dauer, Modulationstiefe etc.) und der Empfindung einer Versuchsperson. Die Reizschwelle bezeichnet den minimalen Wert des Parameters, bei dem der durch diesen Parameter variierte Aspekt des Schallsignals gerade wahrnehmbar wird. Man bezeichnet diesen Reizwert auch als absolute Schwelle. Im Gegensatz dazu beschreibt die Unterschiedsschwelle, wie stark ein Parameter verändert werden muss, damit diese Änderung gerade eben wahrgenommen wird.

Eine Schwelle gibt somit den Wert der physikalischen Größe des zu messenden Signalparameters an, bei dem sich die Wahrnehmung – sei es absolut oder differenziell – gerade eben ändert. Es ist allerdings eine falsche Vorstellung des Begriffs Schwelle, wenn man diese als einen unendlich scharfen Übergang zwischen zwei verschiedenen Wahrnehmungsqualitäten versteht, etwa: Bei einem Wert des Pegels gerade unterhalb der Schwelle wird nichts wahrgenommen, bei einem Wert gerade oberhalb der Schwelle wird das Schallsignal deutlich wahrgenommen. Nach heutigem Verständnis verläuft der Übergang graduell. Im Bereich der Schwelle besteht, bei wiederholter Darbietung desselben Stimulus, eine bestimmte Wahrscheinlichkeit, dass das Signal wahrgenommen wird, und diese Wahrscheinlichkeit steigt monoton mit der physikalischen Reizstärke an. Aufgrund dieses statistischen Verhältnisses zwischen physikalischer Reizstärke und Stärke der Wahrnehmung sind Absprachen darüber nötig, bei welchem Wahrscheinlichkeitswert man die Schwelle legt. Typische Beispiele für solche Absprachen werden bei den einzelnen Messverfahren besprochen.

Hörschwelle

Die Hörschwelle bezeichnet den Pegel eines akustischen Signals, bei dem dieses gerade hörbar ist. Eine solche Angabe macht nur Sinn, wenn eine Reihe zusätzlicher Stimulusparameter definiert und bekannt ist, vor allem die Art des Signals (Sinuston,

Rauschsignal, komplexes Geräusch). Bei Sinustönen kommen dabei die Frequenz, die Dauer, eventuell die Art der Einschaltflanken, sowie die räumliche Darbietungsform (Kopfhörer, Lautsprecher, eventuelle Richtung der Schallquelle) in Frage. Bei breitbandigen Geräuschen sollte die spektrale Bandbreite bzw. der spektrale Verlauf bekannt sein, zusätzlich die beim Sinuston genannten Parameter, soweit sie relevant sind.

Man kann auch von einer Schwelle sprechen, wenn nur die Wahrnehmbarkeit eines Teils eines Signals gemessen wird, z.B. Sprache in Hintergrundgeräusch, Hörbarkeit von Artefakten digitaler Kodierung oder auch die Wahrnehmung des tonalen Charakters einer Signalkomponente.

Die Unterschiedsschwelle gibt – am Beispiel der Größe Pegel – an, um wie viel der Pegel verändert werden muss, sodass diese Änderung wahrgenommen werden kann. Bei einer solchen Messung muss zusätzlich auch noch der Ausgangspegel dokumentiert werden, da Unterschiedsschwellen im Allgemeinen nicht über den ganzen Wahrnehmungsbereich konstant sind.

Die absolute Schwelle kann auch als Unterschiedsschwelle zwischen einem unhörbaren sowie einem gerade wahrnehmbaren Schall verstanden werden kann. Im Folgenden werden daher die Messmethoden für beide Formen der Schwellenmessungen gemeinsam besprochen, und nur wo nötig auf Unterschiede hingewiesen.

Schwellenmessverfahren

Das Entscheidungskriterium

Ein grundlegendes Unterscheidungskriterium für den Vergleich verschiedener Schwellenmessverfahren besteht darin, inwieweit das Messverfahren eine Kontrolle über das von der Vp verwendete Entscheidungskriterium bzw. die Antwortneigung erlaubt. In Verfahren der klassischen Psychophysik wird das Antwortverhalten der Vp nicht berücksichtigt. Sie werden üblicherweise als Verfahren der Kategorie I bezeichnet. Wenn keine Kontrolle über das Antwortverhalten der Vp ausgeübt wird, kann es schwierig sein, zwischen der sensorischen Empfindlichkeit einerseits sowie der Antwortneigung andererseits zu differenzieren. Zwei Personen könnten beispielsweise die gleiche sensorische Empfindlichkeit aufweisen, die eine urteilt jedoch sehr konservativ, indem sie nur dann mit „ja“ bzw. „gehört“ antwortet, wenn sie ganz sicher ist, die andere eher liberal, indem sie auch dann mit „ja“ bzw. „gehört“ antwortet, wenn sie nur halbwegs glaubt, das Signal gehört zu haben.

Verfahren, die das Antwortverhalten der Vp berücksichtigen, fallen unter Kategorie II, und werden vor allem in der wissenschaftlichen Forschung verwendet. Sie erlauben

eine größere Aussagekraft der gemessenen Schwellenwerte, haben aber den Nachteil eines deutlich größeren Zeitaufwandes. In der Praxis (z.B. Audiologie) werden traditionell Messverfahren aus der Kategorie I verwendet, obwohl auch hier Bestrebungen bestehen, Verfahren aus der Kategorie II für den klinischen Einsatz zeitlich zu optimieren. In der folgenden Übersicht werden wir zunächst Methoden aus der Kategorie I besprechen, die alle den Nachteil aufweisen, dass nicht entschieden werden kann, inwieweit das Entscheidungskriterium das Messergebnis beeinflusst. Sie besitzen jedoch den Vorteil, der relativ unaufwändigen und schnellen Durchführung.

Verfahren ohne Kontrolle über das Entscheidungskriterium

Einregelverfahren.

Beschreibung: Bei diesem Verfahren hat die Vp Kontrolle über den Stimulusparameter (durch Drehknopf, Keyboardtasten oder andere Eingabegeräte). Weiterhin kann sie angeben, wann die Einstellung abgeschlossen ist.

Aufgabenstellung für die Messung der Absolutschwelle: Stelle den Parameter (hier Pegel) so ein, dass der Stimulus gerade wahrnehmbar ist.

Messgröße ist der Wert des eingestellten Parameters. Der Stimulus kann hierbei kontinuierlich oder auch gepulst angeboten werden. Gepulste Darbietung macht es für die Vp etwas einfacher, den Übergang zwischen *hörbar* und *nicht hörbar* zu erfassen.

Aufgabenstellung Unterschiedsschwelle: Stelle den Parameter so ein, dass sich der Testschall gerade wahrnehmbar von einem alternierend angebotenen Referenzschall unterscheidet. Mit dieser Aufgabenstellung kann auch die Hörschwelle eines Signals in einem Maskierer gemessen werden. Dazu wird das maskierende Signal kontinuierlich angeboten und das Testsignal gepulst, z.B. 500 ms an, 500 ms aus. Messgröße ist wiederum der eingestellte Parameterwert.

Ablauf der Messung: Die Vp regelt den Stimulus entsprechend der Fragestellung ein, und gibt, z.B. durch Tastendruck, an wenn sie die Schwelle gefunden hat. Der eingestellte Parameterwert wird aufgezeichnet und als Messwert verwendet.

Vorteil: Relativ schnelles Verfahren, Versuchsperson wird aktiv an der Messung beteiligt, sie kann während der Messung die Stimulusstärke (Stimulusunterschied) so groß wählen, dass sie eine deutliche Wahrnehmung hat (und dann in den Schwellenbereich zurückregeln).

Fallstricke: Es sollte darauf geachtet werden, dass die Steuerung des Parameters keine Informationen über den absoluten Wert des eingestellten Parameters gibt (z.B. durch Markierungen am Drehknopf).

Békésy-Verfahren

Beschreibung: Bei diesem Verfahren wird der Stimulusparameter (meist der Pegel) kontinuierlich in einer von zwei möglichen Richtungen (größer, kleiner) variiert. Die Versuchsperson kann durch Knopfdruck die Änderungsrichtung beeinflussen. Ziel ist es, den Parameter um den Schwellenwert herum pendeln zu lassen.

Aufgabenstellung: Beeinflusse den Parameterwert so, dass der Stimulus von hörbar nach unhörbar wechselt, verändere dann die Änderungsrichtung durch Knopfdruck. Sobald der Stimulus wieder hörbar ist, verändere die Änderungsrichtung wiederum durch Knopfdruck.

Die Auswertung basiert auf der Mittelung der oberen sowie unteren Umkehrpunkte (klassisch aus einer Grafik der aufgezeichneten Parameterwerte über der Zeit, moderner durch Computeraufzeichnung und Mittelung dieser Werte).

Variante: Stimulus kann in der Form aufeinander folgender Pulse angeboten werden, der Parameter wird dann zwischen zwei Pulsen um einen vorgegebenen Wert verändert. Bei dieser Variante kann man durch zufällige Variation der Größe der Änderung die Antizipation der Vp verringern.

Bemerkungen: Dieses Verfahren eignet sich nur zum Messen von Wahrnehmungsschwellen, sowohl absolute als auch Wahrnehmungsschwellen in Hintergrundgeräuschen, nicht jedoch zur Bestimmung von Unterschiedsschwellen.

Vorteil: Relativ schnell, Vp hat eine gewisse Kontrolle über die Stärke des Parameters

Nachteil: Es besteht das Problem der Antizipation: Wenn die Größe der Änderung konstant bleibt, kann die Vp die Stimulusstärke bei der nächsten Darbietung (oder in der nächsten Sekunde) vorhersehen, und ihr Urteil aufgrund der antizipierten Stärke fällen.

Eine Variation des Békésy-Verfahrens ist das Eingabelverfahren.

Eingabelverfahren

Beschreibung: Beginnend mit einem deutlich überschwelligen Stimulus bietet der/die Versuchsleiter(in) den Stimulus mit stufenweise abnehmender Stärke an, bis die Versuchsperson angibt, dass sie den Stimulus nicht mehr wahrnimmt (absteigende Darbietung). Anschließend wird, beginnend mit einem noch weiter verringerten Stimulusparameter, eine aufsteigende Reihe angeboten, bis die Vp erstmals angibt, den Stimulus wahrzunehmen. Aufgezeichnet werden die Parameterwerte, bei denen sich die Antwort ändert, und diese Werte werden über eine gleiche Anzahl aufsteigender sowie absteigender Sequenzen gemittelt.

Es gelten dieselben Anmerkungen wie beim Békésy-Verfahren. Bei diesem Verfahren ist die Gefahr der Antizipation (bei absteigenden Sequenzen) besonders groß, da im Allgemeinen nicht nur die Größe der Parameteränderung, sondern auch der Darbietungsrhythmus fest liegt.

Dieses Verfahren kann ebenfalls nicht zur Messung von Unterschiedsschwellen verwendet werden.

Konstanzverfahren

Beschreibung: Bei diesem Verfahren wird der Stimulus mit einem bestimmten Wert des Parameters wiederholt angeboten, und die Vp hat zwei Antwortalternativen: „gehört“, oder „nicht gehört“. Gemessen wird die Häufigkeit der Antworten in einer dieser Kategorien, basierend auf 20 - 100 wiederholten Darbietungen. Diese Messung wird für mehrere Werte des Stimulusparameters wiederholt. Dabei muss der Stimulusparameter so gewählt werden, dass er ausreichend dicht an der zu erwartenden Schwelle liegt, bei zu großen Abweichungen treten Saturierungseffekte auf (die Vp hört entweder nie etwas, oder kann den Stimulus deutlich bei allen Darbietungen wahrnehmen). Wenn die Häufigkeit der Antworten als Funktion des Parameters aufgetragen wird, ergibt sich eine sogenannte psychometrische Funktion. Als Schwellenwert wird typischerweise der Wert des Parameters abgelesen (eventuell nach Interpolation), bei dem die Vp in 50 % der Fälle antwortet: „gehört“.

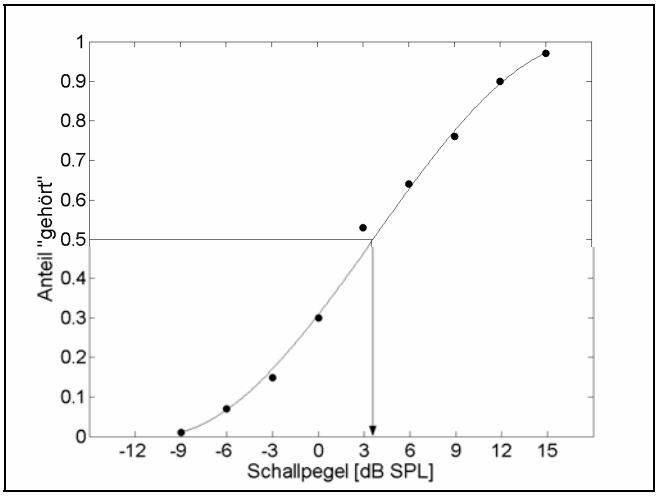


Abbildung 01. Psychometrische Funktion zur Bestimmung der Absolutschwelle (aus: Hellbrück & Ellermeier, 2004, 219)

Nachteile: Wie alle der vorher beschriebenen Verfahren fehlt auch hier eine Kontrolle über das von der Vp verwendete Kriterium. Weiterhin setzt dieses Verfahren voraus, dass man bereits ungefähr weiß, wie die psychometrische Kurve verläuft, da sonst die Gefahr besteht, im Extremfall Stimulusparameter zu wählen, die zu 0 oder 100 % Antworten in derselben Kategorie führen.

Bemerkungen: In der hier beschriebenen Form ist das Verfahren nur für Messung der Absolutschwelle geeignet. Um Unterschiedsschwellen zu messen, muss neben dem Teststimulus immer auch noch ein Referenzstimulus angeboten werden, und die Antwortalternativen der Vp sind dann: „Unterschied hörbar“, oder „kein Unterschied hörbar“.

Verfahren mit Kontrolle über das Kriterium der Vp

Einleitung

Die hier zu beschreibenden Verfahren sind eng mit der Signalentdeckungstheorie verknüpft. Sie haben zum Ziel, die Diskriminationsfähigkeit zu messen (wobei auch hier gilt, dass die absolute Schwelle auch als Diskriminationsschwelle anzusehen ist). Während bei den bisher beschriebenen Methoden die Versuchsperson sich selbst eine deutliche Vorstellung bilden muss, wie das zu beurteilende Schallattribut klingt und sich ein Entscheidungskriterium suchen muss, wird in den im folgenden beschriebenen Verfahren die Fähigkeit gemessen, verschiedene Stimulusversionen unterscheiden zu können. Wenn der akustische Unterschied zwischen den Stimuli unterhalb der Unterschiedsschwelle liegt, kann die Vp nur noch raten, aber nicht mehr konsistent richtig antworten. Dadurch wird eine deutlich bessere Kontrolle darüber möglich, inwieweit die Antworten der Vp in der Tat auf ihrem sensorischen Unterscheidungsvermögen basieren.

Die Verfahren lassen sich wiederum unterscheiden nach dem Gesichtspunkt, ob der Stimulusparameter variiert wird oder nicht (adaptive oder nichtadaptive Verfahren), sowie nach der Anzahl der angebotenen Vergleichsintervalle, zwischen denen sich die Vp entscheiden muss. Wegen dem Zwang, eine Entscheidung fällen zu müssen, heißen diese Verfahren auch *Forced-choice*-Verfahren. Dabei kommt der Version mit nur einem Intervall, die dem oben genannten Konstanzverfahren ähnelt, in Hinblick auf die Antwortalternativen eine Sonderstellung zu. Anhand dieses Verfahrens, das nur in der Version mit konstantem Parameterwert verwendet wird, werden kurz einige Ideen der Signalentdeckungstheorie beschrieben.

Ja/Nein Verfahren (Forced-choice-Verfahren mit 1 Darbietungsintervall):

Beschreibung: Dieses Verfahren ist in Hinblick auf die Antwortalternativen „gehört“ bzw. „nicht gehört“ dem obigen Konstanzverfahren ähnlich. Der Hauptunterschied liegt in der Stimulusdarbietung. Während beim Konstanzverfahren in jedem Beobachtungsintervall der Stimulus angeboten wird, gibt es beim Ja/Nein-Verfahren zwei verschiedene Darbietungsalternativen: Im Beobachtungsintervall wird der Stimulus nur mit einer bestimmten Wahrscheinlichkeit p angeboten. Mit der Wahrscheinlichkeit (1-p) wird kein Stimulus präsentiert. Bei der Messung von Unterschiedsschwellen

werden entsprechend zwei verschiedene Versionen des Stimulus mit den Wahrscheinlichkeiten p und (1-p) angeboten. In der Regel ist p = 0.5.

Auswertung: Da es zwei verschiene Darbietungsformen (Stimulus anwesend oder nicht) sowie zwei Antwortalternativen (gehört, nicht gehört) gibt, lässt sich das Antwortverhalten in Form einer 2x2-Matrix beschreiben und die prozentuale Häufigkeit für jede Matrixzelle berechnen. Da die Vp nach jedem Beobachtungsintervall eine Antwort geben muss – und die Gesamtzahl der Intervalle mit und ohne Signal bekannt ist – sind unter den vier Matrixzellen nur zwei unabhängig, so dass wir uns beschränken auf:

- die relative Häufigkeit von Treffern, das ist die relative Anzahl von Antworten „gehört“ für diejenigen Intervalle, in denen der Stimulus auch tatsächlich vorhanden war;
- die relative Häufigkeit von falschem Alarm, das ist die relative Zahl der Antworten „gehört“ für diejenigen Intervalle, in denen der Stimulus nicht angeboten wurde.

Wenn die Vp den Stimulus deutlich wahrnehmen kann, wird sie einerseits eine hohe Trefferrate haben, anderseits eine niedrige Rate falscher Alarme. Wenn sie dagegen den Stimulus nicht wahrnehmen kann, werden diese beiden Raten ähnlich hoch liegen. Dabei ist es eine Frage der Antwortneigung (Bias), wie hoch diese Rate liegt. Wenn die Vp dazu neigt, häufig „gehört“ zu sagen, werden beide Raten hoch sein, wenn sie häufig „nicht gehört“ sagt, werden sie niedrig liegen. Hiermit wird also eine Trennung der Diskriminationsfähigkeit von dem Antwortkriterium erreicht.

Aus diesen beiden Raten wird der Diskriminationsindex d´ berechnet, indem für die beiden relativen Häufigkeitswerte der Wert der z-Verteilung (kumulative Gaußverteilung) berechnet wird, und die Differenz dieses z-Wertes zwischen Trefferrate und der Rate der falschen Alarme gebildet wird. Als Schwellenwert wird typischerweise der Parameterwert verwendet, bei dem der Diskriminationsindex d´ den Wert 1 hat. Diese Messung muss deshalb mit verschiedenen Parameterwerten im Bereich der Schwelle wiederholt werden (typischerweise 5 Werte), um eine gute Schätzung von d´ zu ermöglichen.

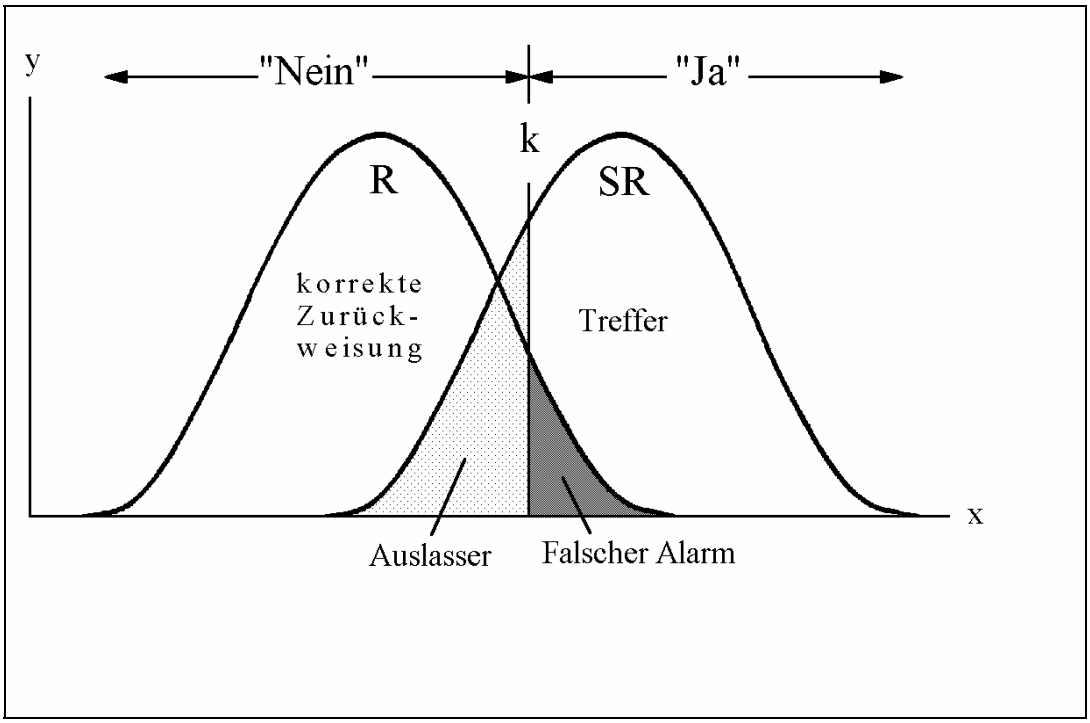


Abb. 02. Wahrscheinlichkeitsverteilungen R (Rauschen allein) und SR (Signal und Rauschen). Entlang der Entscheidungsachse x. Wenn die Verteilungen weit voneinander entfernt sind, bedeutet dies, dass sich das Signal deutlich vom Rauschen abhebt (nach Hellbrück & Ellermeier, 2004, 233)

Forced-choice-Verfahren mit 2 Beobachtungsintervallen (2IFC)

Beschreibung : Bei diesem Verfahren werden nacheinander in getrennten Beobachtungsintervallen, die durch eine kurze Pause von einigen hundert ms getrennt sind, zwei verschiedene Stimuli angeboten, zum einen der Teststimulus, zum anderen der Vergleichsstimulus. Bei der absoluten Wahrnehmungsmessung besteht der Vergleichsstimulus aus Stille, deshalb ist es nötig, die Beobachtungsintervalle durch optische Signale zu markieren. Die Auswahl, ob der Teststimulus im ersten oder zweiten Intervall angeboten wird, erfolgt zufällig und mit gleicher Wahrscheinlichkeit. Die Versuchsperson erhält die Aufgabe anzugeben, in welchem der beiden Intervalle sie den Teststimulus gehört hat. Dies kann über ein Computerkeyboard oder andere Antwortinterfaces, die mindestens zwei verschiedene Antworten registrieren können, erfolgen. Diese Antwort wird als richtig bewertet, wenn der Teststimulus tatsächlich in diesem Intervall angeboten wurde, ansonsten gilt sie als falsch. Es ist deshalb möglich, nach jeder Antwort die Vp zu informieren, ob ihre Antwort richtig oder falsch war. Ein solches Feedback kann es der Vp erleichtern, ein Gefühl dafür zu entwickeln, welches Perzept mit der Anwesenheit des Teststimulus verbunden ist. Wenn der Stimulus deutlich überschwellig ist, wird die Häufigkeit richtiger Antworten nahe bei 100 % liegen, bei unterschwelliger Darbietung wird die Vp nur raten können, was bei zwei Alternativen im Mittel 50 % richtige Antworten ergibt.

Bemerkung: Dieses Verfahren setzt voraus, dass der perzeptive Effekt des zu messenden Stimulusaspektes eindeutig für die Vp zu erkennen ist. Bei der Messung der Hörbarkeit eines Sinussignals in Rauschen ist das z.B. einfach, weil die Anwesenheit der Sinuskomponente ein deutlich verändertes Klangbild erzeugt. Wenn aber eine tonale Komponente in einem tonalen Hintergrundgeräusch angeboten wird, kann es schwierig sein, eine bestimmte Klangqualität mit der tonalen Komponente zu identifizieren. In solchen Fällen kann ein sog. „Labeling“-Problem eintreten: Die Vp ist zwar in der Lage, zwischen den beiden Beobachtungsintervallen zu unterscheiden, sie kann also diskriminieren, aber sie hat Schwierigkeiten zu entscheiden, welches der beiden Intervalle das Testsignal enthält. Hierdurch wird die Häufigkeit richtiger Antworten verringert, es kann sogar dazu kommen, dass diese Häufigkeit signifikant niedriger liegt als die Ratewahrscheinlichkeit, und zwar dann wenn sie konsistent das falsche Intervall angibt. Bei solchen Stimuluskonfigurationen empfiehlt es sich, ein Verfahren mit drei Beobachtungsintervallen zu verwenden.

Nicht-adaptives Verfahren – fester Parameterwert

Bei dieser Variante muss zunächst ein Parameterwert im Bereich der zu erwartenden Schwelle ausgewählt werden, für den dann 20 bis 100 Trials (Kombination beider Beobachtungsintervalle plus Antwort der Vp) angeboten werden. Der Messwert ist direkt gegeben durch die Prozentzahl richtiger Antworten. Durch Wiederholung mit ca. 5 verschiedenen Parameterwerten entsteht eine (aufgrund der Ratewahrscheinlichkeit von 50 % transformierte) psychometrische Funktion, deren Richtig-Antwortrate zwischen 50 und 100 % liegt. Als Messwert wird hieraus meist der Wert bei 75 % abgelesen (ein Wert von 76 % entspricht einem d´ von 1).

Adaptives Verfahren – variabler Parameterwert

Bei diesem Verfahren wird der Parameterwert im Verlaufe einer Messung automatisch so variiert, dass er häufig im Bereich der Schwelle angeboten wird. Dies geschieht durch eine Auswertung der Antworten in den zurückliegenden Intervallen. Weit verbreitet, und einfach als Computerprogramm zu implementieren, ist das sogenannte „2-down/1-up“-Verfahren. Immer, wenn die Vp zweimal hintereinander beim selben Parameterwert, eine richtige Antwort gegeben hat, wird der Parameterwert verkleinert. Andererseits wird nach jeder falschen Antwort der Parameterwert erhöht. Bei einer solchen Parametersteuerung wird dieser sich so einpendeln, dass die Vp mit 70.7 % Wahrscheinlichkeit eine richtige Antwort gibt.

Vorteil: Bei diesem Verfahren ist es nicht nötig, eine genaue Vorstellung vom erwarteten Schwellenwert zu haben. Man muss allerdings schon wissen, wie steil die psychometrische Kurve verläuft, da sich daran die Schrittgröße der Parameter-

änderung in der Messphase orientiert. Diese wird bei Messungen des Pegels üblicherweise auf 1 oder 2 dB gesetzt.

Durchführung: Um der Vp einen deutlichen Eindruck vom Testintervall zu geben, wird der Stimulus am Beginn des Versuchs deutlich überschwellig und mit einer großen Änderungsschrittgröße angeboten. Dadurch kann der Parameterwert in wenigen Intervallen in die Nähe der Schwelle geregelt werden. Die Schrittgröße wird dann sukzessiv verkleinert, bis sie die Messschrittgröße erreicht. Diese Beginnphase des Parametertrackings wird bei der Auswertung im Allgemeinen ignoriert. Die darauf folgende Messphase wird beendet, wenn eine bestimmte Anzahl von Beobachtungsintervallen (z.B. 50), oder eine bestimmte, gerade Zahl von Umkehrpunkten im Parametertrack (z.B. 8) erreicht ist. Der Messwert kann berechnet werden durch Mittelung aller Parameterwerte aus der Messphase, oder indem man aus den Parameterwerten an den Umkehrpunkten den Mittelwert bzw. Medianwert berechnet.

Varianten: Der asymptotisch erreichbare Prozentwert kann durch andere Steuerungsverfahren verändert werden (3-down, 1-up). Die Parametersteuerung kann auch durch Auswertung einer größeren Zahl zurückliegender Antworten realisiert werden.

Forced-choice-Verfahren mit drei Beobachtungsintervallen (3IFC)

Beschreibung: Dieses Verfahren unterscheidet sich vom 2IFC Verfahren dadurch, dass es drei Beobachtungsintervalle gibt, weil neben dem Testintervall zwei Referenzintervalle angeboten werden. Die Auswahl des Testintervalls erfolgt wiederum zufällig. Bei diesem Verfahren ist es ausreichend, wenn die Vp dasjenige Intervall angeben kann, das sich von den zwei anderen unterscheidet (daher auch Odd-ball-Verfahren), wodurch das oben genannte Labelingproblem umgangen wird. Ansonsten kann dieses Verfahren genauso (adaptiv, nichtadaptiv) angewendet werden wie das 2IFC Verfahren

Nachteil: Aufgrund der größeren Zahl von Beobachtungsintervallen dauert eine Messung ca. 50 % länger als bei einem vergleichbaren 2IFC Verfahren.

Skalierung überschwelliger Größen

Eindimensionale und mehrdimensionale Skalierung

Bei der eindimensionalen Skalierung handelt es sich – hier im Rahmen der Hörakustik – um die Skalierung von Schalleigenschaften, die nur durch sich selbst bestimmt sind und keine zusammengesetzten Größen darstellen. Bei eindimensionaler Skalierung besteht das Ziel in der Regel darin, den funktionalen Zusammenhang zwischen

Empfindungsstärke und auslösenden physikalischen Reizgröße herauszufinden (Psychophysik). Die empfundene Lautstärke eines Tones, seine Tonhöhe oder Rauhigkeit sind beispielsweise eindimensionale Größen. Für diese drei Beispiele gibt es auch standardisierte Skalen mit Größenbezeichnung und Größenzeichen, Skaleneinheit und Skalenzeichen. Allgemein versteht man unter psychophysischer Skalierung die Quantifizierung von Empfindungsgrößen im Sinne des Messens; d.h. die Zuordnung von Zahlen zu Empfindungsstärken erfolgt derart, dass sich die Eigenschaften der Empfindungsgrößen in den Eigenschaften der Zahlen, auf die sie abgebildet werden, widerspiegeln.

Im Gegensatz zu eindimensionalen Größen ist die Wahrnehmung komplexer Schalleigenschaften, z.B die Klangfarbe von mehreren Faktoren beeinflusst. Es handelt sich hierbei um mehrdimensionale Größen. Oftmals sind diese Faktoren im Einzelnen gar nicht oder nur zum Teil bekannt und müssen erst über die Ähnlichkeitsstruktur vergleichbarer Schalle (Beispiel: Beschleunigungsgeräusche von Autos) herausgefunden werden. Dazu dienen Verfahren der multidimensionalen Skalierung. Hierzu zählen die Ähnlichkeitsskalierung und das Semantische Differenzial (auch Eindrucksdifferenzial genannt).

Eindimensionale Skalierung

Indirekte und direkte Skalierung

Es gibt zwei sehr unterschiedliche Verfahrensweisen der eindimensionalen Skalierung, die man als *indirekte* und *direkte* Skalierung bezeichnet. Beide werden im Folgenden dargestellt. Die indirekten Verfahren eignen sich wegen des großen Aufwands und der Notwendigkeit unmittelbarer Vergleiche eher für Laboruntersuchungen; die direkten Verfahren sind wegen ihrer Ökonomie und des „absoluten“ Urteilscharakters auch in Feldversuchen, mit langen Intervallen zwischen Beurteilungen und mit kleinen Teilnehmerzahlen einsetzbar.

Indirekte Skalierung – Thurstone-Skalierung

Einführung

Ausgehend von der Annahme, dass identische Reize nicht immer die gleiche Empfindungsstärke auslösen, sondern eine Verteilung von Empfindungsstärken, lässt sich die Diskriminierbarkeit von Reizen aufgrund der Zuordnung zu ordinalen Kategorien (law of categorical judgment) oder aufgrund von Dominanz-Paarvergleichen (law of comparative judgment) bestimmen. Nimmt man an, dass die durch einen Reiz ausgelösten Empfindungsstärken normalverteilt sind, so lassen sich aus den relativen Häufigkeiten – der Kategorienzuordnung oder der Dominanz – z-Werte

berechnen, die die Abstände der Reize auf der Empfindungsskala repräsentieren. Thurstone's Verfahren ist eine von mehreren Methoden, intervallskalierte Werte indirekt aus Daten über die Diskriminierbarkeit von Reizen zu erschließen.

Durchführung

Voraussetzungen: Die Reize müssen große Varianz (d.h. überlappende Verteilungen) auf der Empfindungsdimension aufweisen. Sonst ist weder die Kategorisierungsmethode nach Thurstone geeignet, noch die Paarvergleichsmethode.

Urteile: (a) Methode der sukzessiven Kategorien: Zuordnung von Einzelreizen zu ordinalen Kategorien (z.B. auf 5 Stufen von "wenig lästig" bis "sehr lästig"); (b) Paarvergleichsmethode: vollständiger Dominanz-Paarvergleich; bei n Reizen n(n-1)/2 Paarvergleiche bezüglich eines Attributs (z.B. Lästigkeit).

Auswertung: Von den Rohdaten wird nur die ordinale Information benutzt. Aufgrund der oben skizzierten theoretischen Annahmen werden durch z-Transformation der kumulierten Häufigkeiten Skalenwerte berechnet. Diese haben dann Intervallskalenniveau. Statistisch kann die Gültigkeit des Thurstone-Modells mittels des χ^2-Tests geprüft werden.

Bewertung

Ökonomie: Sehr aufwändig, wegen des vollständigen Paarvergleichs.

Vorteile: Theoretische Annahmen (Normalverteilung; Additivität) sind z.T. überprüfbar.

Indikation: Skalierung ähnlicher (bezüglich des untersuchten Attributs verwechselbarer) Reize; Paarvergleichsmethode insbesondere dort angezeigt, wo mittels kategorialer Urteile schlecht differenziert werden kann (z.B. subtile Unterschiede in Maschinengeräuschen).

BTL-Skalierung und verwandte Methoden

Voraussetzungen

Das Bradley-Terry-Luce (BTL)-Modell postuliert eine einfache Beziehung zwischen Präferenzwahrscheinlichkeiten und Skalenwerten:

$$p_{ab} = \frac{v(a)}{v(a) + v(b)}$$

wobei p_{ab} für die Wahrscheinlichkeit steht, Reiz a dem Reiz b "vorzuziehen" (d.h. ihn als "lauter", "angenehmer" o.ä. zu beurteilen, je nachdem, was die Aufgabe vorsieht); v(a) und v(b) bezeichnen die (zu schätzenden) Skalenwerte dieser Objekte. Ist

Gleichung 1 für alle Zellen der vollständigen Paarvergleichsmatrix erfüllt, so haben die geschätzten v-Skalenwerte Verhältnisskalenniveau, d.h. Aussagen der Art "Reiz c klingt dreimal so lästig wie b" sind gerechtfertigt.

Allerdings ist das BTL-Modell nur dann gültig, wenn das Urteilsverhalten quasi "eindimensional" erfolgt, d.h. wenn für alle Paarvergleiche stets die gleichen Reizattribute (mit gleicher Gewichtung) herangezogen werden. Wechselt dagegen die Gewichtung der Reizattribute, indem etwa in einem Vergleich Tonhaltigkeit, im nächsten Lautheit das Urteil dominiert, so muss das Modell verworfen werden. Für diese Fälle wurden weniger restriktive Wahlmodelle (Präferenzbäume, Eliminierung nach Aspekten) entwickelt, die ebenfalls zu Verhältnisskalen führen, aber auch mehrdimensionale Urteilsstrategien und damit "Aspektwechsel" erlauben. Diese Modelle können auch dazu dienen, die dimensionale Struktur in den Daten aufzudecken.

Durchführung

Voraussetzungen: Die Präferenzwahrscheinlichkeiten müssen von Null und Eins verschieden sein ($0 < p_{ab} < 1$), d.h. kein Reiz darf sicher stärker als ein anderer bezüglich der beurteilten Dimension sein.

Anzahl der Versuchspersonen: Etwa 6-mal soviel wie Stimuli, bei kleineren Zahlen kein befriedigender Modelltest möglich.

Urteile: vollständiger Dominanz-Paarvergleich; bei n Reizen n(n-1)/2 Paarvergleiche bezüglich eines Attributs (z.B. Tonhaltigkeit).

Auswertung: Überprüfung der Rohdaten (Paarvergleichsmatrix) auf Transitivität; liegen zu viele "zirkuläre Triaden" vor, ist das BTL-Modell nicht geeignet zur Beschreibung der Daten. Modellschätzung: Likelihood-Quotienten-Test auf Güte der Anpassung des BTL-Modells. Schätzung der Modellparameter (d.h. der Skalenwerte). Evtl. Schätzung von Vertrauensintervallen für die Skalenwerte. Detaillierte Angaben einschließlich Schätzprogrammen finden sich in der Literatur im Anhang.

Bewertung

Ökonomie: Sehr aufwändiges Verfahren: Bei 10 Reizen z.B. sind n*(n-1)/2 = 45 Paarvergleiche pro Vp zu erheben und ca. 60 Vpn vonnöten, um das Modell adäquat statistisch testen zu können.

Vorteile: Theoretisch fundierte Methode, die – wenn das Modell gilt – erlaubt, Verhältnisskalen aus (ordinalen) Paarvergleichen abzuleiten.

Indikation: Skalierung von auditiven Attributen, deren (Ein-)dimensionalität geprüft werden soll.

Direkte Skalierung

Einführung

Unter direkten Skalierungsverfahren versteht man ganz allgemein alle Verfahren, die ohne Zwischenschritte unmittelbar zu einer Skala führen. Genügt es beispielweise, Reize unterschiedlicher Ausprägung in eine Rangordnung zu bringen, kann die Vp aufgefordert werden, die Reize direkt in eine Rangordnung zu bringen. Die mittleren Rangordnungen etablieren dann eine Ordinalskala der betreffenden Reize. Diese Vorgehensweise wird bei Sound-quality-Untersuchungen relativ häufig angewandt, da die Aufgabe von der Vp einfach durchzuführen ist, sofern die Anzahl der Reize relativ niedrig ist (< 10) (vgl. hierzu „Ranking procedure – Random Access“ nach Fastl, 2005, 140f.).

Bei höherem Skalenniveau (Intervallskala und Verhältnisskala) etablieren die Verfahren der direkten Skalierung die in der Instruktion bzw. in den Antworten der Versuchspersonen geäußerten Zahlworte ohne weitere Transformationen direkt die Skalenwerte. Dahinter steht die Annahme, dass Menschen in der Lage sind, Empfindungsstärken direkt auf einer Skala zu quantifizieren. Bezeichnet eine Person beispielsweise einen Ton als mittelhoch, geht man davon aus, dass dieser Ton auf der subjektiven Bewertungsskala in der Mitte zwischen „hoch“ und „tief“ bzw. in der Mitte zwischen „sehr hoch“ und „sehr tief“ liegt, und dass diese Skala Intervallskalenniveau besitzt. Würde man diesen verbalen Bezeichnungen Zahlen zuordnen, könnte man diese Zahlen als intervallskaliert ansehen und z.B. arithmetische Mittel und Standard-abweichungen berechnen und interpretieren. Dies gilt z.B. für die Kategorien- bzw. Ratingskalen, die in den Sozialwissenschaften sehr häufig verwendet werden.

In vergleichbarer Weise werden Zahlen, denen Urteile zugrunde liegen, die Verhältnisse zwischen Empfindungsstärken widerspiegeln sollen – beispielsweise: dieser Ton ist dreimal so laut wie ein anderer – auch im Sinne von Verhältnissen interpretiert. In diesem Sinne verwendete Verhältnisskalen finden sich sehr häufig in der Psychophysik und selten in den Sozialwissenschaften.

Mit anderen Worten: Die in der Instruktion bzw. in den Antworten geäußerten Zahlworte werden für wahr genommen, und das Skalenniveau wird in der üblichen Forschungspraxis meist ungeprüft als gegeben angenommen („Messung per fiat“). Dies ist unter axiomatisch orientierten Messtheoretikern nicht unumstritten, wird jedoch von eher pragmatisch eingestellten Skalentheoretikern akzeptiert. Allerdings gibt es hier, grob gesagt, zwei Fraktionen, nämlich bestehend aus denjenigen Anwendern, die eher Kategorienskalen favorisieren und solchen, die eher den Verhältnisskalen den Vorzug geben.

Magnitude Estimation (Größenschätzung)

Die Verfahren der Magnitude Estimation gehören zur Klasse der Verhältnisskalierung. Sie gehen im Wesentlichen auf S.S. Stevens zurück. Man unterscheidet mehrere Vorgehensweisen. Beim klassischen Verfahren werden ein Standardreiz und ein Vergleichsreiz präsentiert. Der Standardreiz erhält die Bezeichnung „10“, dem Vergleichsreiz sollen Zahlen so zugeordnet werden, dass sich das Verhältnis der Empfindungsstärken in den Zahlenwerten widerspiegelt. Ist der Vergleichsreiz doppelt so laut wie der Standard, erhält er den Zahlenwert 20, ist er dreimal so laut, dann 30, ist er halb so laut, dann 5 usw.. Heute wird häufig kein Standardreiz mehr vorgegeben, sondern den einzeln präsentierten Reizen sollen direkt Zahlen so zugeordnet werden, dass die Verhältnisse der Zahlen die Verhältnisse der Empfindungsstärken reflektieren.

Es gibt auch "Herstellungs"-Varianten dieses Verfahrens (Magnitude production) Hierbei werden (umgekehrt) Reizintensitäten zu vorgegebenen numerischen Skalenwerten eingestellt.

Beim "Cross-modality matching", einer weiteren Variante, wird die Empfindungsstärke (z.B. der empfundenen Lautstärke) durch Einstellung auf einer anderen Reizdimension (etwa der Leuchtdichte einer Lichtquelle) angegeben. Die numerischen Ergebnisse dieser Verfahren werden als Messwerte auf dem Zahlenstrahl interpretiert, d.h. ihnen werden – in der Regel ungeprüft – die Eigenschaften von Verhältnisskalen zugesprochen.

Eine beliebte Variante des Cross-modality Matching ist die Angabe der Empfindungsstärke durch Markierung auf einer nicht unterteilten Linie (visuelle Analogskala), wobei das eine Ende für schwache Empfindungsintensitäten, das andere für hohe Empfindungsintensitäten steht. Hier ist jedoch nicht klar, ob diese Variante noch im Sinne einer Verhältnisskala interpretiert werden darf, oder nicht vielmehr eine Intervallskala widerspiegelt. Solchermaßen etablierte Skalen werden daher häufig unter Kategorien- bzw. Ratingskalen geführt (vgl. unten).

Durchführung

Anzahl der Versuchspersonen: In der Regel werden 3-5 Wiederholungen pro Reiz benötigt, um reliable Messwerte zu erhalten. Da ein Einzelurteil nur wenige Sekunden verlangt, ist das Verfahren recht ökonomisch. Replizierbare Kollektivdaten können erfahrungsgemäß mit ca. 15-30 Versuchspersonen erhoben werden.

Urteile: (a) Größenschätzung (magnitude estimation): Darbietung von Einzelreizen oder von Reizpaaren aus Standardreiz und Vergleichsreiz; numerisches Schätzurteil aus dem Bereich der positiven rationalen Zahlen; (b) Größenherstellung (magnitude

production /fractionation): Darbietung von Zahlenwerten oder Zahlenverhältnissen (in der Regel positiv ganzzahlig oder Brüche), Einstellung eines Reizparameters (Schallpegel, Tonfrequenz) durch die Versuchsperson. Beispiel für eine typische Instruktion im Anhang.

Auswertung: Geometrische Mittelung der Einzelurteile/-einstellungen bzw. arithmetische Mittelung der logarithmierten Einzelwerten; Angabe von Standard-abweichungen oder -fehlern, eventuell auch Interquartilabständen. Als theoretisches Modell für eine Kurvenanpassung der Einzelwerte dient die mathematische Potenzfunktion.

Zur Bewertung der statistischen Signifikanz von Unterschieden in der Einschätzung von Schallen werden in der Regel Methoden der Varianzanalyse verwendet.

Bewertung

Vorteile: Ökonomisches Verfahren zur Datensammlung, bei dem Skalenbereich und Feinunterteilung (fast) vollständig unter der Kontrolle der Versuchsperson sind. Vorhandensein einer breiten Forschungsliteratur.

Nachteile: Die Interpretation der untransformierten Daten als Verhältnisskala ist wahrscheinlich nicht haltbar; ebenso wenig die Interpretation eines Potenzfunktions-exponenten als Kennwert der individuellen sensorischen Empfindlichkeit. Vielmehr scheinen individuelle Besonderheiten im Zahlengebrauch zu der großen Variabilität in den Skalenwerten zu führen.

Kategorienskalierung stationärer Schalle

Der Begriff Kategorienskala (category scale) wird in im Bereich der Psychophysik verwendet, in den Sozialwissenschaften verwendet man eher den Begriff Ratingskala. Die Kategorienskalierung stellt ein Sammelbecken für eine Reihe von Verfahren unterschiedlicher Herkunft dar. Gemeinsam ist allen, dass den Urteilen mehr als nur ordinale Eigenschaften zugesprochen werden, und zwar in der Regel die einer Intervallskala. Manche dieser Verfahren sind lediglich pragmatisch-intuitiv begründet (Beispiel: Visuelle Analogskala; vgl. oben), andere haben sich vor dem Hintergrund experimenteller Forschung zu Kontexteffekten entwickelt, insbesondere im Rahmen der sog. Bezugssystemforschung, das Kategorienunterteilungsverfahren.

Durchführung

Voraussetzungen: Bei Untersuchungen zu "Sound quality" ist oftmals das Auffinden relevanter Attribute (attribute elicitation) aufwändiger als die eigentliche "Attribut-Quantifizierung". Hierzu sind besondere Methoden, v.a. für Experten-Jurys entwickelt worden, die unter Stichworten wie "Descriptive Analysis" firmieren.

Anzahl der Versuchspersonen: Nur wenige erforderlich, um reliable Daten zu erhalten, da Kategorisierung im Alltag hoch geübt ist.

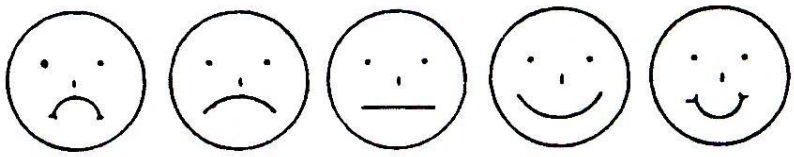
Urteile: Darbietung von Einzelreizen, Beurteilung verbal, numerisch, oder graphisch, beim Kategorienunterteilungsverfahren in der Regel 2-stufig: zuerst über die verbale Grobkategorisierung, dann mittels numerischer Feinabstufung (vgl. Heller, 1985; Hellbrück, 1996). Dieses Verfahren ist aufgrund seiner hohen Auflösung besonders dann geeignet, wenn kleine Unterschiede wiedergegeben werden sollen, aber nur wenige Messwiederholungen möglich sind.

Im Allgemeinen werden bei Kategorienverfahren folgende Antwortformate empfohlen:

- Häufigkeit
    - nie – selten – gelegentlich – oft – immer
- Intensität
    - gar nicht – kaum – mittelmäßig – ziemlich – außerordentlich
- Wahrscheinlichkeit
    - keinesfalls – wahrscheinlich nicht – vielleicht – ziemlich wahrscheinlich – ganz sicher
- Bewertung (stimmen Sie überein?)
    - völlig falsch – ziemlich falsch – unentschieden – ziemlich richtig – völlig richtig

Weitere Varianten von Antwortformaten:

- Symbolische Marken: Smileys



- Graphisches Rating: _____|_____
- Numerische Skala

1	2	3	4	5

- Bipolare Skalen:

- 2	- 1	0	1	2

- Kategorienunterteilungsskala (am Beispiel Lautstärkenempfindung; nach Heller, 1985; s. auch Hellbrück, 1996)

	...
	53
schmerzhaft	52
	51
	50
	49
	48
	47
	46
sehr laut	45
	44
	43
	42
	41
	40
	39
	38
	37
	36
laut	35
	34
	33
	32
	31
	30
	29
	28
	27
	26
mittel	25
	24
	23
	22
	21
	20
	19

leise	18 17 16 15 14 13 12 11
sehr leise	10 9 8 7 6 5 4 3 2 1
nicht gehört	0

Orientierung: Entscheidend ist eine adäquate Entsprechung der Reizmenge (bzw. der durch sie ausgelösten Empfindungen) mit den Urteilskategorien. Wird etwa eine Skala, die von "sehr leise" bis "sehr laut" reicht, auf Reize angewandt, die sich nur um wenige sone unterscheiden, so werden Kontexteffekte und individuelle Unterschiede verstärkt. In der Regel wird ein adäquater Umgang mit Kategorienskalen durch 2 Maßnahmen gefördert: (1) Orientierung über den Umfang der Reizserie vor der eigentlichen Datensammlung und (2) verbale oder operationale Verankerung der Urteilskategorien.

Bisektion/Äquisektion: Herstellungsverfahren zur Bestimmung von Mitten oder subjektiv gleichabständigen Reizintensitäten.

Auswertung: Arithmetische Mittelung der Einzelurteile/-einstellungen; Angabe von Standardabweichungen oder –fehlern; Varianzanalyse zur Absicherung von Mittelwertsunterschieden.

Bewertung

Vorteile: Ökonomisches, alltagsnahes Verfahren, bei dem die Urteilsskalen standardisiert sind. Breiter Erfahrungsschatz in Untersuchungen zur Geräusch-bewertung und audiologischen Fragestellungen (Hörgeräteanpassung).

Indikation: Erhebung von vielen Reizattributen in einer Untersuchung (vgl. auch Semantisches Differenzial, unten). Beurteilung von Effekten in Bezug auf eine

semantische Interpretation (z.B. "higly annoyed"). Messung von individuellen Unterschieden (z.B. Hörfeldaudiometrie).

Kategorienskalierung zeitvarianter Schalle

Besonders bei langdauernden, über die Zeit variierenden Schallen, ist es oftmals von Interesse, die zeitlichen Veränderungen in einem subjektiven Attribut (v.a. der Lautheit, aber andere sind denkbar) von Augenblick zu Augenblick zu verfolgen. Zu diesem Zweck sind Methoden der fortlaufenden Beurteilung (continuous judgment by category) entwickelt worden. Das kann in diskreten Kategorien erfolgen, etwa über 7 Tasten, die für ebenso viele Lautheitskategorien stehen und nur gedrückt werden, wenn sich die Lautheit merklich ändert. Ebenso ist aber auch kontinuierliches Nachfahren mittels eines Schiebereglers denkbar. Von Interesse ist bei solchen Untersuchungen entweder, den Zusammenhang zwischen den „instantanen“ Lautheitseinschätzungen und der „globalen“ (Gesamt-) Lautheit zu bestimmen, oder die Beziehung des subjektiven Maßes zum physikalisch gemessenen Pegelverlauf (vgl. z.B. Kuwano, 2008).

Durchführung

Anzahl der Versuchspersonen: Die Anforderungen an die Zahl der Probanden entsprechen denjenigen der Kategorienskalierung.

Urteile: Der Beurteiler erhält die Instruktion kontinuierlich während der Präsentation des experimentellen Schalls, z.B. ein Computerkeyboard zu bedienen, bei dem einzelne Tasten (vorzugsweise die Tasten „0 – 9“) per Beschriftung den zu bewertenden Kategorien zugeordnet sind. Sowohl die momentan gewählte Kategorie als auch der zugehörige Zeitwert werden vom Rechner erfasst und gespeichert. Gleichzeitig kann auf dem Bildschirm des Rechners als Rückmeldung eine Darstellung des kontinuierlichen Beurteilungsverlaufs erfolgen. Alternativ ist auch die Verwendung eines speziellen Eingabegeräts üblich, das nur eine diskrete Anzahl von Tasten oder einen analogen Schieberegler enthält. Häufig verfügt bei dieser Lösung das Eingabegerät selbst über ein Display für die optische Rückmeldung. Bei analoger Eingabe (Schieberegler) sollte die Taktrate des einlesenden A/D-Wandlers 20 Hz (entsprechend einer Auflösung von 10 Hz) nicht unterschreiten, um auch rasche Reaktionen, z.B. beim Auftreten impulsartiger Schalle, erfassen zu können. Bei diskreter Erfassung (Tastatur) hängt die zeitliche Auflösung insbesondere von den Einstellungen in der Systemsteuerung des Rechners ab (typischerweise im Bereich von 2-10 ms, abhängig von den Systemeinstellungen des Tastaturtreibers). Bei Verwendung spezieller Tastaturinterfaces können deutlich höhere zeitliche Auflösungen (< 100 µs) erreicht werden.

Auswertung: Für die Auswertung kommen sowohl globale statistische Kenngrößen (Mittelwert, Standardabweichung, etc.) als auch zeitlich spezifische, z.B. ereigniskorrelierte, Kenngrößen in Frage. Da die akustischen Parameter des präsentierten Schalls ebenfalls eine zeitliche Verlaufstruktur aufweisen (Pegelverläufe, Ordnungsverläufe, Zeit-Frequenz-Verläufe etc.) besteht die Möglichkeit, die gemessenen Beurteilungsverläufe mit den akustischen Verlaufsgrößen in Beziehung zu setzen, z.B. mit Hilfe des Maximums der Kreuzkorrelationsfunktion. Dadurch können zeitliche Abhängigkeiten des Urteils von Variationen der akustischen Parameter bestimmt werden.

Bewertung

Vorteile: Der wesentliche Vorteil der fortlaufenden Beurteilung gegenüber den bisherigen (zeitlich globalen) Skalierungsverfahren liegt in der Möglichkeit begründet, statistische Beziehungen zwischen zeitlich variierenden akustischen Merkmalen und den Urteilsvariationen sehr viel unmittelbarer messen zu können. Hypothesen über die kausale Verknüpfung von Reiz und Reaktion lassen sich dadurch leichter und besser überprüfbar ableiten. Ein weiterer Vorteil ist durch die vergleichsweise hohen Anzahl der Messwerte gegeben, die der Zuverlässigkeit und Stabilität des Messverfahrens zu gute kommt.

Indikation: Langdauernde Reize wie Fahrzeuginnengeräusche, Beschleunigungsgeräusche oder Verkehrslärm. Bestimmung der subjektiv relevanten Geräuschanteile. Evaluation instrumenteller Verfahren zur Integration psychoakustischer Größen über die Zeit.

Mehrdimensionale Verfahren

Ähnlichkeitsskalierung

Es ist oft wichtig, die subjektive Wahrnehmung eines Produktgeräusches von Konsumenten bzw. Nutzern dieses Produktes zu bestimmen. Komplexe Geräusche werden in der Regel im Hinblick auf mehrere Eigenschaftsdimensionen wahrgenommen. Diese Dimensionen bestimmen einen Wahrnehmungsraum. Häufig sind diese Dimensionen nicht bekannt. Um sie herauszufinden, bietet sich die multidimensionale Skalierung auf der Grundlage der Ähnlichkeitsskalierung an.

Durchführung: Die Beurteilungsobjekte, z.B. Fahrgeräusche von Fahrzeugen verschiedener Hersteller werden beispielsweise anhand einer 5-stufigen Ratingskala paarweise nach ihrer Ähnlichkeit beurteilt. Die Skala ist gleichabständig unterteilt, ihre Endpunkte sind mit „vollkommen unähnlich“ und „vollkommen ähnlich“ bezeichnet. Es werden k(k-1)/2 Beurteilungen durchgeführt (k = Anzahl der Objekte).

Auswertung: Das Ziel ist die Ermittlung von Urteilsdimensionen auf der Basis dieses Datenmaterials. Dazu werden die Ähnlichkeiten in Distanzen transformiert und in einem euklidischen Raum dargestellt. In diesem Raum werden den untersuchten Objekten Positionen so zugewiesen, dass die räumlichen Distanzen möglichst gut mit den empirisch bestimmten übereinstimmen. In einem Auswertungsverfahren, das der Faktorenanalyse gleicht, werden die Dimensionen der Ähnlichkeit ermittelt. Aus der in Abbildung 03 wiedergegebenen Konfiguration von Fahrzeugmarken können die Dimensionen interpretativ erschlossen werden, in diesem Beispiel „Sportlichkeit“ auf der horizontalen Achse sowie „Prestige“ auf der vertikalen Achse.

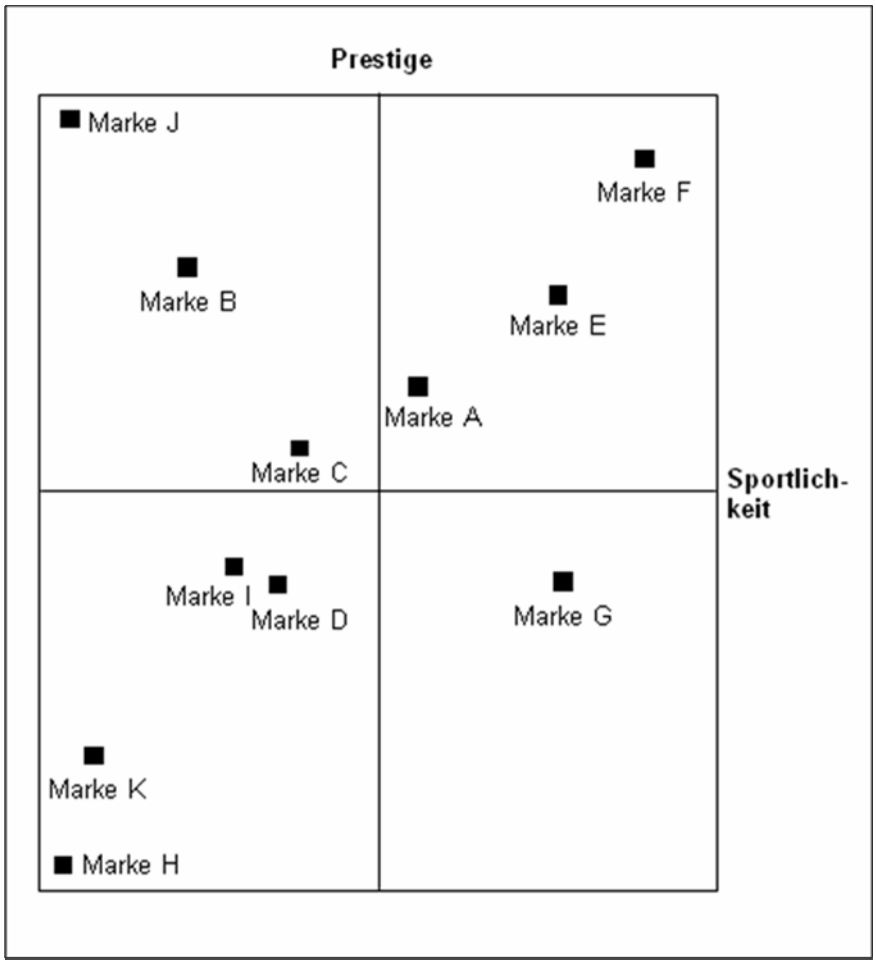


Abb. 03. Anordnung von Fahrzeugmarken aufgrund von Ähnlichkeitsskalierung entlang der interpretativ erschlossenen Dimensionen Prestige und Sportlichkeit.

Die Verfahren der MDS sind theoretisch wie auch vom statistischen Auswertungsverfahren anspruchsvoll und aufwändig. Die gängigen Statistik-Softwarepakete stellen jedoch Programme zur Verfügung, mit deren Hilfe der Aufwand relativ einfach zu bewältigen ist.

Bewertung: Ähnlichkeitsskalierung ist dann von Vorteil, wenn die Eigenschaftsdimensionen unbekannt sind, oder wenn Gefahr besteht, mit hypothetisch angenommenen Eigenschaften und Verbalisierungen das Ergebnis zu beeinflussen. Ein

Nachteil kann in der gelegentlich schwierigen Interpretation der Ergebnisse gesehen werden. Eine Vielzahl von Methoden kann jedoch auch bei schwierigen Interpretationen zusätzlich zu Hilfe gezogen werden.

Semantisches Differenzial

Einführung

Ein weiteres Verfahren zur mehrdimensionalen Bewertung von Geräuschen ist die Semantische Differenzial Technik, kurz: Semantisches Differenzial (SD), welches auch als „Eindrucksdifferenzial“ und „Polaritätenprofil“ bezeichnet wurde. Im Gegensatz zu der oben dargestellten Methode der Ähnlichkeitsskalierung werden beim SD den Probanden die zu beurteilenden Attribute vorgegeben und deren Ausprägung direkt erfasst. Hierzu werden bipolare Ratingskalen mit meist sieben Stufen verwendet, die durch gegensätzliche Adjektive verankert sind (z.B. „glatt-rau“, „schwach-kraftvoll“, „tief-hoch“; vgl. Abbildung 4).

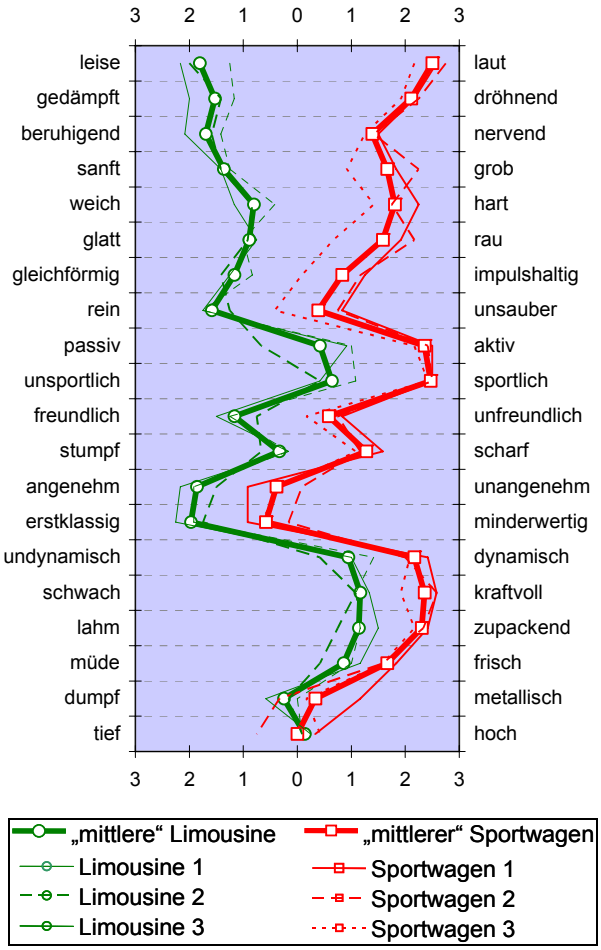


Abb. 04. Polaritätenprofile für das Innengeräusch in Luxuslimousinen und Sportwagen.

Das SD hat seine Wurzeln in der Psycholinguistik und wurde ursprünglich von Charles Osgood zur Erfassung der affektiven Bedeutung beliebiger Begriffe entwickelt. Übertragen auf die Hörakustik wird angenommen, dass die Wahrnehmung eines Geräuschereignisses stets auch emotionale und evaluative Reaktionen hervorruft, welche mittels Konditionierung durch den Schall ausgelöst und auf entsprechenden Adjektivskalen abgebildet werden können (z.B. „nervend - beruhigend“, „unangenehm - angenehm“). Entgegen der ursprünglichen Intention in der Psycholinguistik werden in der Hörakustik mit dem SD jedoch auch sogenannte denotative Komponenten erfasst. Hierbei handelt es sich häufig um physikalisch objektivierbare, dem konventionellen Sprachgebrauch der Adjektive entsprechenden Geräuscheigenschaften (z.B. „leise - laut“, „tief - hoch“), einschließlich der Attribute aus der Expertensprache (z.B. „tonhaltig – rauschhaft“; „impulshaltig-glatt“).

Anwendungsbereiche

Mit dem SD können prinzipiell zwei unterschiedliche Zielrichtungen verfolgt werden. Zum einen können die Urteile auf den Adjektivskalen dazu verwendet werden auf statistischem Wege den Wahrnehmungsraum für eine bestimmte Geräuschklasse zu untersuchen, soweit er durch die experimentellen Schalle repräsentiert wird. Ziel ist es, auf einer abstrakten übergeordneten Ebene eine Reihe von Faktoren zu bestimmen, welche die Urteile der Probanden auf den Adjektivskalen determinieren. Gleichzeitig soll damit die Anzahl der Variablen reduziert und die in den Adjektivskalen mehr oder weniger enthaltene Bedeutungsredundanz eliminiert werden. Dies erreicht man durch statistische Analyse mit der sogenannten Faktorenanalyse, welche aus den korrelativen Zusammenhängen zwischen den Variablen (Adjektivskalen) eine aus wenigen Komponenten bestehende Struktur generiert. Man versucht etwa drei bis sieben meist orthogonale Faktoren (auch als Komponenten bezeichnet) zu extrahieren, deren Bedeutung wie bei der Ähnlichskalierung jedoch erst durch Interpretation erschlossen werden muss (vgl. Tabelle 2). Aus dieser Vorgehensweise ergibt sich, dass im Gegensatz zur Anwendung der MDS bei der Ähnlichkeitsskalierung nur solche Dimensionen entdeckt werden können, die im Hörversuch auch durch entsprechende Adjektivskalen im SD repräsentiert waren.

Hat man auf diesem Weg eine Ordnung für die in der Wahrnehmung relevanten Hörattribute gefunden, kann für die künftige Beurteilung ähnlicher Schalle ein verkürztes SD konstruiert werden, das nur noch aus jenen Adjektivskalen besteht, welche die Faktorenstruktur am besten repräsentieren. Das Ziel besteht dann im Sinne eines deskriptiven Ansatzes darin, konkret eine Auswahl von Geräuschen

mehrdimensional zu bewerten und die mittleren Urteile in Form eines sensorischen Profils darzustellen.

	Timbre	Power	Evaluation
stumpf-scharf	.70		
gedämpft-dröhnend	.61		
gleichmäßig-ungleichmäßig	.60		
gleichförmig-stetig	.58		
dumpf-grell	.56		
weich-hart	.55		
leicht-schwer		.77	
schwach-stark		.76	
tief-hoch		-.55	
glatt-rau		.52	
undynamisch-dynamisch			.80
unangenehm-angenehm			.64
unsportlich-sportlich			.64
aufgeklärte Varianz (in % der Gesamtvarianz)	20.15	15.48	13.30

Tab. 02. Korrelationen (Faktorladungen) der Adjektivskalen aus dem Semantischen Differenzial für Motorgeräusche mit den extrahierten Faktoren „Timbre“, „Power“ und „Evaluation“.

Auswahl der Adjektivpaare

Da die Auswahl der Adjektivpaare beim SD von zentraler Bedeutung ist, sollte deren Eignung für die aktuell zu beurteilende Geräuschserie stets empirisch überprüft werden. Sofern noch keine Daten aus Untersuchungen mit vergleichbaren Schallen vorliegen, empfiehlt sich daher vor dem ersten Einsatz der Skalen im Hörversuch ein Pretest. In diesem Vorversuch wird ein Pool mit möglichst vielen verbalen Deskriptoren von einer Probandenstichprobe auf einer Skala (z.B. „sehr geeignet“, „geeignet“, „möglich“, „ungeeignet“, „sehr ungeeignet“) bewertet um auf dieser Grundlage anschließend die am besten geeigneten Adjektive für das SD auswählen zu können. Dabei ist besonders darauf zu achten, dass die Begriffe im Sprachgebrauch der späteren Probandenstichprobe vorkommen.

Ferner sollten insbesondere die zur Erfassung konnotativer Bedeutungen vorgesehenen Adjektivpaare möglichst ein bipolares Kontinuum aufspannen. Dies ist in der Regel gegeben, wenn es sich um wahre Antonyme (z.B. „schwach - stark“) handelt und nicht um kontradiktorische Begriffe (z.B. „unbehaglich - behaglich“). Schließlich wird empfohlen, die Eignung der Adjektivskalen nach der Datenerhebung im Hörversuch noch mittels Varianzanalyse hinsichtlich ihrer Sensitivität zu analysieren. Als Entscheidungskriterium können sogenannte Sensitivitätskoeffizienten berechnet

werden, welche ein Maß dafür sind, wie gut die Adjektivskalen die Unterschiede zwischen den dargebotenen Schallen erfassen konnten (vgl. Abschnitt „Datenanalyse“).

Hinsichtlich der Anzahl der Adjektivpaare gilt als Faustregel, dass für eine explorative Untersuchung zur Bestimmung der Dimensionalität des Wahrnehmungsraumes jede der vermuteten Dimensionen mindestens durch drei Adjektivpaare repräsentiert sein soll. Häufig besteht ein SD mit dem Ziel einer anschließenden Faktorenanalyse daher aus 20 bis 30 Adjektivskalen.

Durchführung

Bei den Adjektivskalen in einem SD handelt es sich um eine Form der Kategorienskalierung, somit können die oben angeführten Hinweise hier direkt übertragen werden (vgl. Abschnitt „Kategorienskala stationärer Schalle“). Da beim SD sowohl Stimuli als auch Antwortskalen variieren, sind im experimentellen Ablauf grundsätzlich zwei Darbietungsmodi möglich. Führt man die Methode in traditioneller Art und Weise durch, wird jedes Geräusch der Reihe nach auf allen Adjektivskalen beurteilt bevor das nächste dargeboten wird. Bei stationären Schallen bietet es sich dann an, die Geräuschwiedergabe solange auszudehnen bis alle Skalen abgearbeitet wurden, bei instationären Signalverläufen empfiehlt sich die wiederholte Darbietung. Alternativ können die Schalle jedoch auch skalenweise beurteilt werden, d.h. es wird immer nur eine Adjektivskala vorgelegt, und alle Geräusche werden darauf beurteilt. Diese Vorgehensweise dauert zwar meist länger, ist jedoch aus bezugssystemtheoretischer Sicht sehr vorteilhaft. Die Merkmalsunterschiede können genauer auf der Skala abgebildet werden, weil die Bezugssysteme für die verschiedenen Skalen nicht gewechselt werden müssen und die Urteile damit einen stärkeren Gedächtnisbezug haben.

Aus experimentalpsychologischer Sicht ist auch beim SD das Prinzip der Randomisierung anzuwenden. Dies betrifft sowohl die Reihenfolge der Adjektivpaare als die der Schalle. Auf die zufällige Umkehrung der Pole einer Adjektivskala zur Vermeidung einer Urteilstendenz in eine bestimmte Richtung sollte jedoch verzichtet werden, da dies die Aufgabe häufig zu schwierig macht und fehlerhafte Antworten hervorgerufen werden können.

Auswertung

Die mittels SD erhobenen Daten können je nach Anwendungsbereich in vielfältiger Weise analysiert und dargestellt werden. Die Darstellung der Urteile in Form von Mittelwertsprofilen (vgl. Abbildung 4) ist für das SD grundlegend und veranschaulicht Bedeutungsunterschiede zwischen Geräuschen auf intuitiv verstehbare Weise. Zur statistischen Absicherung von Mittelwertsunterschieden kann auf interferenz-

statistische Verfahren zurückgegriffen werden. Neben den üblicherweise für Mittelwertsvergleiche verwendeten t-Tests existieren spezielle statistische Verfahren zur Prüfung von Profilunterschieden (siehe Literaturliste).
In der Datenanalyse sollten jedoch nicht nur die Geräuschunterschiede im Fokus stehen, sondern auch die Frage geklärt werden, inwieweit die verschiedenen Adjektivskalen die Unterschiede überhaupt diskriminieren konnten. Hierzu empfiehlt sich die Analyse der Sensitivität der Skalen, die in Analogie zu einer univariaten Varianzanalyse mit Messwiederholungen durchgeführt werden kann. Man erhält so Aufschluss darüber, in welchem Ausmaß die Urteilsstreuung die Unterschiede zwischen den Geräuschen widerspiegelt bzw. in welcher Relation dazu die Gruppenstreuung steht.
Häufig stellt das SD nur den ersten Schritt in der Untersuchung des Wahrnehmungsraumes einer bestimmten Geräuschklasse dar und liefert die Datengrundlage für eine Faktorenanalyse. Hierbei wird eine ordnende Struktur erzeugt, welche die Lokalisation der Geräusche in einem n-dimensionalen Raum erlaubt. Auf diese Weise können die in den Adjektivskalen enthaltenen Informationen zur Beschreibung der Ähnlichkeiten und Unterschiede in hohem Maße verdichtet werden. Darauf aufbauend bietet sich ferner die Bestimmung von Euklidschen Distanzmaßen an, wodurch die Profilunterschiede global durch einen Einzahlwert ausgedrückt werden können.

Bewertung

Das SD hat einen hohen Stellenwert in explorativen Untersuchungen zur Untersuchung von Klangbildwahrnehmung und Sound Quality. Die Methode ist sowohl für akustische Laien als Experten geeignet und liefert auf ökonomische Weise eine wertvolle Informationsbasis für weiterführende Untersuchungen zu einzelnen Hörattributen. Häufig dienen die mit dem SD entdeckten Geräuschunterschiede dazu, korrelative Zusammenhänge zu instrumentellen Signalparametern zu erschließen und die mathematische Modellierung psychoakustischer Kenngrößen voranzutreiben.
Sofern beabsichtigt ist, den Wahrnehmungsraum für eine bestimmte Geräuschklasse grundlegend zu untersuchen, besteht aufgrund der selektiven Zusammenstellung der Adjektivskalen naturgemäß das Risiko weitere Eigenschaften zu übersehen, die im Wahrnehmungsraum eine Rolle spielen. Dies gilt insbesondere auch für die Interpretation der Ergebnisse aus der Faktorenanalyse.

Kontext und Bezugssystem

Definitionen

Mit dem Begriff *Kontext* wird in der Psychophysik allgemein der Umstand zum Ausdruck gebracht, dass Wahrnehmung und Urteilsbildung einer gewissen Relativität unterliegen. In einem Hörversuch versteht man unter *Kontextfaktoren* alle Einflüsse, die sich neben den experimentell variierten Schallparametern noch zusätzlich in den Reaktionen der Probanden systematisch niederschlagen. Diese können gegeben sein durch die Versuchsumgebung (z.B. Natürlichkeit der Situation, Schallfeld), die Erhebungsmethode (z.B. Typ und Konstruktionsmerkmale der Urteilsskala) und die Schallserie (z.B. Reihenfolge, Umfang, Verteilung der Merkmalsausprägung). Von derartigen Kontextfaktoren zu unterscheiden sind die sogenannten interindividuellen (z.B. Geschlecht, Alter, Expertenstatus) und intraindividuellen Variablen (z.B. Tageszeit, Stimmung), welche ebenfalls zur Varianz der unabhängigen Variablen beitragen können.

Unter *Bezugssystem* versteht man das bei einem Probanden im Hörversuch wirksame perzeptive Ordnungssystem, auf das sich seine Eindrücke bzw. Antworten beziehen („Maßstab“). Im Gegensatz zu dem per Konvention festgelegten cm-g-s-Bezugssystem der Physik entwickelt sich das Bezugssystem für Wahrnehmungsurteile vor dem Hintergrund der Erfahrung. Demnach können prinzipiell zwei Arten von Bezugssystemen unterschieden werden. Zum einen stellt der *individuelle Erfahrungshintergrund* eines Probanden ein Bezugssystem dar, das sich über die Zeit im Langzeitgedächtnis stabilisiert hat und sogenannte *absolute Urteile* (z.B. „leise“, „laut“, „sehr laut“) ermöglicht. Absolute Urteile, die wir in der alltäglichen Kommunikation problemlos verwenden, wären nicht verständlich, wenn nicht die kommunizierenden Personen über das gleiche Bezugssystem verfügten. Zum anderen wird in einem Hörversuch durch die aktuell dargebotenen Schallreize ein *Bezugssystem* induziert, welches mit den im Gedächtnis verankerten Erfahrungen konkurrieren kann. Für die Kategorienskalierung bedeutet dies, dass sich die Probanden zunächst entscheiden müssen, welches Bezugssystem sie ihren Urteilen zugrunde legen. Je stärker sich nun die Probanden in ihren Bezugssystemen unterscheiden, desto weniger kann die Urteilsstreuung durch Empfindungsunterschiede erklärt werden, sondern durch die unterschiedliche Verwendung der Urteilsskala. Dadurch kann die Reliabilität der Mittelwerte erheblich reduziert werden. Dieses Problem ist umso gravierender, je größer der Unterschied zwischen den früheren und den aktuell im Experiment gemachten Erfahrungen ist und kann daher bei Expertenbeurteilungen in besonders starker Ausprägung auftreten.

Kontrolle von Kontextfaktoren

Dieses bei Kategorienskalierungen vielfach untersuchte Bezugssystemproblem kann gemindert werden, indem versucht wird allen Probanden möglichst ein einheitliches Bezugssystem für die Beurteilung vorzugeben. Dies erreicht man, indem die Probanden zu Beginn des Versuchs durch Darbietung entweder der gesamten Schallserie oder deren Extreme über den Umfang des Merkmalskontinuums, den die Schallserie darauf einnimmt, orientiert werden (vgl. Abschnitt „Kategorienskalierung stationärer Schalle“).

Ein weiterer Beitrag zur Reduzierung von Kontexteffekten bei Kategorienskalierungen kann durch die adäquate Bezeichnung der Urteilskategorien geleistet werden. Die verbalen Bezeichnungen der Skala sollen zu den aktuellen Merkmalsausprägungen der dargebotenen Schalle passen, da sonst die Gefahr besteht, dass die Probanden verschieden große Bereiche auf der Urteilsskala ausnutzen (vgl. Abschnitt „Kategorienskalierung stationärer Schalle“): Während manche Probanden ihre Urteile über den gesamten Bereich der Skala verteilen, ziehen andere möglicherweise nur die Kategorien heran, deren verbale Bezeichnungen auch im alltäglichen Sprachgebrauch für die jeweiligen Merkmalsausprägungen verwendet werden.

Wie am Beispiel der Kategorienskalierung erläutert, können viele Kontextfaktoren durch sachgemäße Anwendung der Erhebungsmethoden wirksam kontrolliert werden. Von wesentlicher Bedeutung sind hier ferner auch die im Abschnitt „Versuchsplanung“ beschriebenen Techniken, welche im Prinzip ebenfalls darauf abzielen, den durch die interessierenden Variablen erklärten Anteil der Varianz in den abhängigen Variablen (Reaktionen) möglichst zu vergrößern.

Schließlich soll noch auf die Notwendigkeit einer umfassenden Versuchsdokumentation hingewiesen werden um die Vergleichbarkeit der Kontextbedingungen zwischen verschiedenen Hörversuchen zu einem späteren Zeitpunkt noch möglichst genau bewerten zu können.

Literaturempfehlungen

Schwellenmessungen und eindimensionale Skalierung

Gescheider, G. A. (1997). Psychophysics: The fundamentals (3rd ed.). Mahwah NJ: Erlbaum.

Methoden der Signalentdeckungstheorie

Macmillan, N. A. & Creelman, C. D. (2005). Detection theory: a user's guide. New York: Cambridge University Press.

Mehrdimensionale Skalierung und Ähnlichkeitsskalierung

Backhaus / Erichson /Plinke / Weiber (2006). Multivariate Analysemethoden. Eine anwendungsorientierte Einführung. Berlin: Springer (s. S. 499 – 563).

Semantisches Differenzial

Diehl, J. M. & Schäfer, A. (1975). Techniken der Datenanalyse beim Eindrucks-differential. In R. Bergler (Hrsg.), Das Eindrucksdifferential (S. 157-211). Bern: Huber.

Literatur nach Sachgebieten geordnet (Literatur z.T. mehrfach aufgeführt)

Hören und Hörversuche allgemein

Bech, S. & Zacharov, N. (2006). Perceptual audio evaluation. Chichester: Wiley.

Blauert, J. & Bodden, M. (1994). Gütebeurteilung von Geräuschen – Warum ein Problem? In Q.-H. Vo (Ed.), *Soundengineering. Kundenbezogene Akustikentwicklung in der Fahrzeugtechnik* (S. 1-9). Renningen-Malmsheim: expert Verlag.

Blauert, J. & Jekosch, U. (1997). Sound-quality evaluation – a multi-layered problem. *Acta Acustica, 83,* 747-753.

Choisel, S. & Wickelmaier, F. (2007). Evaluation of multichannel reproduced sound: Scaling auditory attributes underlying listener preference. *Journal of the Acoustical Society of America*, 121(1), 388-400.

Fastl, H. (2008). Psychoacoustics and product sound quality. In S. Kuwano (Ed.), Recent topics in environmental psychoacoustics (pp. 63 – 87). Osaka: Osaka University Press.

Fastl, H. (2005). Psycho-acoustics and sound quality. In: J. Blauert (Ed.), Communication acoustics (pp. 139 – 162). Berlin: Springer.

Fastl, H., Zwicker E. (2007) Psychoacoustics – Facts and Models, 3rd Ed. Berlin: Springer.

Gelfand, S. A. (2004). Hearing. An introduction to psychological and physiological acoustics (4rd ed.). New York: Marcel Dekker.

Gelfand, S. A. (2001). Essentials of audiology. Stuttgart: Thieme.

Hellbrück, J. & Ellermeier, W. (2004). Hören. Physiologie, Psychologie und Pathologie (2. akt. u. erw. Auflage). Göttingen: Hogrefe.

Meilgaard, M, Civille, G. V. & Carr, B. T. (1999). Sensory Evaluation Techniques. Boca Raton: CRC Press.

Ruan, D. & Zeng, X. (2004). Intelligent sensory evaluation. Berlin: Springer.

Schick, A. (1995). Geräusche im Fahrgastraum des PKW: Die Bedeutung der Sprache – erläutert am Beispiel der japanischen Klangfarbenforschung. In S. R. Ahmed (Hrsg.), *Akustik und Aerodynamik des Kraftfahrzeuges. Grundlagen - Optimierungsmethoden - Meß- und Versuchstechnik* (S. 18-31). Renningen-Malsheim: expert-Verlag.

Schick, A. (1995). Geräusche im Fahrgastraum des PKW: ein geschichtlicher Abriß. In S. R. Ahmed (Hrsg.), *Akustik und Aerodynamik des Kraftfahrzeuges. Grundlagen - Optimierungsmethoden - Meß- und Versuchstechnik* (S. 32-47). Renningen-Malsheim: expert-Verlag.

Namba, S. (2008). The evaluation of sound environment and psychological methods. In S. Kuwano (Ed.), Recent topics in environmental psychoacoustics (pp. 1 – 32). Osaka: Osaka University Press.

Forschungsmethoden allgemein, experimentelle Versuchsplanung im Besonderen

Bortz/Döring (2006). Forschungsmethoden und Evaluation (4. Aufl.). Heidelberg: Springer.

Campbell, D.T. & Stanley, J.C. (1963). Experimental and quasi-experimental designs for research. Chicago: Rand McNally.

Huber, O. (2005). Das psychologische Experiment. Eine Einführung (4. Aufl.). Bern: Huber.

Massaro, D. (1989). Experimental Psychology. San Diego: Harcourt.

Sarris, V. & Reiss, S. (2005). Kurzer Leitfaden der Experimentalpsychologie. München: Pearson

Sedlmeier, P. & Renkewitz, F. (2008). Forschungsmethoden und Statistik in der Psychologie. München: Pearson.

Snodgrass, J.G., Levy-Berger, G. & Haydon, M. (1985). Human Experimental Psychology. New York: Oxford University Press.

Statistische Analyseverfahren

Backhaus / Erichson /Plinke / Weiber (2006). Multivariate Analysemethoden. Eine anwendungsorientierte Einführung. Berlin: Springer.

Bortz, J. (2005). Statistik (6. Aufl.). Berlin: Springer.

Bortz, J. Lienert, G. & Böhnke, K. (2000). Verteilungsfreie Verfahren der Biostatistik (S. 449-502). Berlin: Springer.

Kendall, M. G. & Gibbons, J. D. (1990). Rank Correlation Methods. London: Arnold.

Sedlmeier, P. & Renkewitz, F. (2008). Forschungsmethoden und Statistik in der Psychologie. München: Pearson.

Psychophysik, Messtheorie und Schwellenmessverfahren

Baird, J. C. & Noma, E. (1978). Fundamentals of scaling and psychophysics. New York: Wiley.

Gescheider, G. A. (1997). Psychophysics: The fundamentals (3rd ed.). Mahwah NJ: Erlbaum.

Gigerenzer, G. (1981). Messung und Modellbildung in der Psychologie. München: Ernst Reinhardt.

Green D. M. & Swets, J. A. (1966). Signal detection theory and psychophysics. New York: Wiley.

Guilford, J.P. (1954). Psychometric methods (2nd Ed.). New York: McGraw-Hill.

Kaernbach, C. (1991). Simple adaptive testing with the weighted up-down method. Perception & Psychophysics, 49, 227-229.

Kaernbach, C. (2001). Adaptive threshold estimation with unforced-choice tasks. Perception & Psychophysics, 63, 1377-1388.

Levitt, H. (1971). Transformed up-down methods in psychoacoustics. Journal of the Acoustical Society of America. 49(2), 467-477.

Luce, R.D. & Suppes, P. (2002). Representational measurement theory. In (H. Pashler & J.Wixted, Eds.) Stevens’ Handbook of Experimental Psychology, 3rd Edition, Vol 4 (pp. 1-41). New York: Wiley.

Macmillan, N. A. & Creelman, C. D. (2005). Detection theory: a user's guide. New York: Cambridge University Press.

Mosteller, F. (1951). Remarks on the method of paired comparisons: III. A test of significance for paired comparisons when equal standard deviations and equal correlations are assumed. Psychometrika, 16, 207-218.

Sixtl, F. (1982). Meßmethoden der Psychologie: Theoretische Grundlagen und Probleme (S. 176-221). Weinheim: Beltz.

Stevens, S.S. (1975). Psychophysics. New York: Wiley.

Wickens, T. D. (2002). Elementary signal detection theory. New York: Oxford University Press.

Swets, J. A. (1961/1988). Is there a sensory threshold? Science, 134, 168-177. (reprinted in Swets, J. A. (1988). Signal detection and recognition by human observers: Contemporary readings. New York: Wiley.)

Swets, J. A., Tanner, W. P. & Birdsall, T. G. (1961). Decision processes in perception. psychological Review, 68, 301-340.

Tack, W. H. (1983). Psychophysische Methoden. In H. Feger & J Bredenkamp (Hrsg.) Enzyklopädie der Psychologie. Forschungsmethoden der Psychologie. Messen und Testen, Bd. 3, (S.346-426). Göttingen: Hogrefe.

Torgerson, W., S. (1958). Theory and Methods of Scaling (pp. 155-204). New York: Wiley & Sons.

Zwislocki, J., Maire, F., Feldman, A. S. & Rubin, H. (1957). On the effect of practice and motivation on the threshold of audability. Journal of the Acoustical Society of America, 30, 254-262.

Skalierung

Backhaus / Erichson / Plinke / Weiber (2006). Multivariate Analysemethoden. Eine anwendungsorientierte Einführung. Berlin: Springer.

Borg, I. & Staufenbiel, T. (1997). Theorien und Methoden der Skalierung: eine Einführung (S. 53-69). Bern: Huber.

Ellermeier, W. & Faulhammer, G. (2000). Empirical evaluation of axioms fundamental to Stevens' ratio-scaling approach: I. Loudness production. Perception & Psychophysics, 62, 1505-1511. Diehl, J. M. & Schäfer, A. (1975). Techniken der Datenanalyse beim Eindrucksdifferential. In R. Bergler (Hrsg.), Das Eindrucksdifferential (S. 157-211). Bern: Huber.

Garner, W. R. (1954). Context effects and the validity of loudness scales. Journal of Experimental Psychology, 48, 218-224.

Guilford, J.P. (1954). Psychometric methods (2nd Ed.). New York: McGraw-Hill.

Hellbrück, J. (1996). Category-subdivision scaling – a powerful tool in audiometry and noise assessment. In H. Fastl, S. Kuwano & A. Schick (Eds.), Recent trends in hearing research. Festschrift for Seiichiro Namba (pp. 317-336). Oldenburg: BIS.

Heller, O. (1985). Hörfeldaudiometrie mit dem Verfahren der Kategorienunterteilung (KU). Psychologische Beiträge, 27, 478-493.

Kuwano, S. (2008). Psychological evaluation of environmental noise from the viewpoint of temporal aspects. In S. Kuwano (Ed.), Recent topics in environmental psychoacoustics (pp. 33 – 61). Osaka: Osaka University Press.

Osgood, C. E. (1952). The nature and measurement of meaning. Psychological Bulletin, 49(3), 197-237.

Osgood, C. E., Suci, G. J. & Tannenbaum, P. H. (1957). The measurement of meaning. Urbana: University Press of Illinois.

Poulton, E.C. (1989). Bias in quantifying judgments. Hove: Erlbaum

Schäfer, B. & Fuchs, A. (1975). Kriterien und Techniken der Merkmalsselektion bei der Konstruktion eines Eindrucksdifferentials. In R. Bergler (Hrsg.), Das Eindrucksdifferential (S. 119-137). Bern: Huber.

Sixtl, F. (1982). Meßmethoden der Psychologie: Theoretische Grundlagen und Probleme (S. 176-221). Weinheim: Beltz.

Stevens, S.S. (1975). Psychophysics. New York: Wiley.

Swets, J. A. (1961/1988). Is there a sensory threshold? Science, 134, 168-177. (reprinted in Swets, J. A. (1988). Signal detection and recognition by human observers: Contemporary readings. New York: Wiley.)

Swets, J. A., Tanner, W. P. & Birdsall, T. G. (1961). Decision processes in perception. psychological Review, 68, 301-340.

Tack, W. H. (1983). Psychophysische Methoden. In H. Feger & J Bredenkamp (Hrsg.) Enzyklopädie der Psychologie. Forschungsmethoden der Psychologie. Messen und Testen, Bd. 3, (S.346-426). Göttingen: Hogrefe.

Torgerson, W., S. (1958). Theory and Methods of Scaling (pp. 155-204). New York: Wiley & Sons.

Wickelmaier, F. & Ellermeier, W. (2007). Deriving auditory features from triadic comparisons. Perception & Psychophysics (in press).

Wickelmaier, F. & Schmid, C. (2004). A Matlab function to estimate choice model parameters from paired-comparison data. Behavior Research Methods, Instruments, and Computers, 36(1), 29-40.

Zimmer, K. & Ellermeier, W. (2003). Deriving ratio-scale measures of sound quality from preference judgments. Noise Control Engineering Journal, 51(4), 210-215.

Zimmer, K., Ellermeier, W. & Schmid, C. (2004). Using probabilistic choice models to investigate auditory unpleasantness. ACUSTICA - Acta Acustica, 90(6), 1019-1028.

Skalierung und Kontexteffekte

Gescheider, G. A. (1997). Psychophysics: The fundamentals (3rd ed.). Mahwah NJ: Erlbaum (S. 255 – 262).

Lockhead, G.R. (1995). Psychophysical scaling methods reveal and measure context effects. Behavioral and Brain Sciences, 18, 601-612.

Lockhead, G.R. (2004). Absolute judgments are relative: A reinterpretation of some psychophysical ideas. Review of General Psychology, 8 (4), 265–272

Zeitler, A., Hellbrück, J., Ellermeier, W., Fastl, H., Thoma, G. & Zeller, P. (2006). Methodological approaches to investigate the effects of meaning, expectations and context in listening experiments. Internoise (03-06 December 2006), Honolulu,Hawaii, USA.