

Psychologische Test-Theorie und der Zusammenhang zwischen physikalischer Schallenergie-Dosis und Belästigungswirkung

Karl Theodor Kalveram ¹

Zusammenfassung Ein Ziel der Lärmforschung ist, aus der (physikalischen) Lärmmessung die (psychologische) Belästigung vorherzusagen. Zur Beantwortung der Fragen, wie gut dies gelingen kann und welches Lärmmaß das geeignetste ist, werden die Begriffe 'Reliabilität', 'Validität' und 'Äquivalenz' aus der 'psychologischen Testtheorie' herangezogen. Dazu wird die Lärmmessung als 'Test', die Messung der Belästigung als 'Kriterium' aufgefaßt. Anhand einer Untersuchung über Fluglärmwirkungen in der Umgebung des Münchener Flughafens aus dem Jahre 1969, bei der sowohl verschiedene physikalische Lärmmaße als auch die Bevölkerungsreaktion an derselben Stichprobe von Probanden erhoben worden waren, kann gezeigt werden, daß die Maße L_{eq1} , L_s , L_{eq3} , L_{eq4} , NNI und $FB1$ die Äquivalenzkriterien der Testtheorie erfüllen, dh. sie können statistisch nicht unterschieden werden, weder hinsichtlich ihrer Reliabilitäts- noch Validitätskoeffizienten, und ihre Reliabilitätskoeffizienten sind praktisch gleich ihren Interkorrelationskoeffizienten (beide sind fast gleich 1). Diese sechs Maße können damit als 'Paralleltests' angesehen werden. Andere Lärmmaße, wie D_{10} , H_{81} , $\log N$, L_{eq10} , erweisen sich als nicht parallel (im Sinne von nicht äquivalent) zu den ersteren. Jedoch, auch die erstgenannten sechs Lärmmaße haben mit Bezug auf das Kriterium nur moderate Validitätskoeffizienten (ca 0.5). Die physikalische Lärmmenge ist daher aus Sicht der psychologischen Testtheorie ein zu grobes Maß, wenn die individuelle Belästigung vorhergesagt werden soll, doch reicht sie aus, wenn Belästigungsmittelwerte von Gruppen vorherzusagen sind. Ferner kann geschlossen werden, daß der Versuch, durch Modifikationen der physikalischen Meßprozeduren deren Validität zu verbessern, keinen Erfolg verspricht. In dieser Hinsicht dürfte es effektiver sein, z.B. die Reliabilität der psychologischen Meßinstrumente für die Belästigung zu verbessern.

Schlüsselwörter: Reliabilität, Validität und Äquivalenz von Lärmmessungen

The theory of mental testing and the correlation between physical noise level and annoyance.

Summary The aim of noise research is to predict (psychological) annoyance from (physical) noise measurements. In order to answer the questions, how precise this prediction is, and which of the noise measurement procedures is most suitable, the concepts of 'reliability', 'validity' and 'equivalence' defined in the 'theory of psychological testing' are applied. Thereby, the noise measurements procedures are regarded as 'tests' variables and the related annoyance as the criterion variable. Referring to data of an investigation on aircraft noise in the vicinity of Munich Airport in 1969, in which different physical as well as annoyance data were sampled from the same subjects, it turns out, that the measurements called L_{eq1} , L_s , L_{eq3} , L_{eq4} , NNI and $FB1$ meet the equivalence criteria of test theory, that is, they cannot statistically be distinguished, neither regarding their intercorrelations nor their correlations with the criterion, and their coefficients of reliability equal their intercorrelation coefficients (both are close to one). Therefore, these six measures can be considered as 'parallel'. Other measurements like D_{10} , H_{81} , $\log N$ and L_{eq10} , reveal that they are not parallel (in the sense of not equivalent) to the former six. However, even the former are only of moderate validity (about 0.5). In the framework of psychological testing, therefore, physical noise is only a poor measure when used to predict individual annoyance, but it suffices when used to predict group averages. Moreover, it can be concluded that the attempt to enhance validity by modification of the highly reliable physical measuring procedures cannot be successful. It would be more effective to enlarge the reliability of the psychological measurement procedures of annoyance.

Key words: Reliability, validity and equivalence of physical noise measures.

1 Einleitung

¹ Institut für Allgemeine Psychologie der Heinrich-Heine-Universität, Arbeitsrichtung Kybernetische Psychologie und Psychobiologie, Universitätsstr. 1, 40225 Düsseldorf, e-mail: kalveram@uni-duesseldorf.de. Die vorliegende Studie wurde gefördert vom Land Nordrhein-Westfalen im Rahmen des Projekts "Ökologische Lärmwirkungsforschung"

In der Akustik ist man bestrebt, Meßverfahren auf rein physikalischer Grundlage zu entwickeln, welche die "Lästigkeit" oder "Lärmigkeit" von Geräuschen durch einen einzigen Zahlenwert repräsentieren. Guski [12] führt aus, daß es mindestens 70 verschiedene solche Meßverfahren gäbe und daß immer noch welche hinzukämen, Meßverfahren, die meist "Lärmbewertungs- oder Lärmbewertungsverfahren" genannt werden. Diese Bezeichnungen jedoch führen häufig zu dem Mißverständnis, es handele sich bei ihnen um direkte Verfahren zur Erfassung etwa von Belästigungswirkungen. Tatsächlich aber erfassen diese Verfahren keine psychologischen Größen, sondern sind physikalische 'Lärm-Mengen-Maße', weil in diese Verfahren eben nur physikalisch definierte Größen wie Schallintensität, zeitliche Dauer und Häufigkeit eingehen, mit denen man dann psychologische Wirkungen, wie etwa die Belästigung, schätzen will, ohne sie unmittelbar zu messen.

Die verschiedenen physikalischen Maße für die Lärmmenge korrelieren meist sehr hoch miteinander, so daß viele Autoren ihre Zweifel haben, ob sich die Vielfalt der Maße rechtfertigen läßt. Bezugnehmend auf die Ergebnisse entsprechender Korrelationsstudien schreibt z.B. Jonckheere [14], daß "these results tend to prove that many noise descriptors are in fact interchangeable". Vallet et al. [36] kommen zu dem Schluß, daß solche "analyses...hardly allows one to suggest, that one acoustical indicator is better than another one." Auch Langdon & Griffith [22] schreiben, daß "all the inter-correlations between the various statistical noise indices are extremely high and this indicates that each of these indices is, in effect, measuring the same thing ... On the basis of the present data, it therefore seems hardly meaningful to speak of a 'best' index of noise, relative to dissatisfaction or 'bother'."

Man fragt sich, ob es unter diesen Umständen nicht sinnvoller wäre, sich auf ein Standardverfahren zu einigen, z.B. auf den 'energie-äquivalenten Dauerschallpegel' (in dieser Arbeit mit L_{eq3} bezeichnet), denn er korreliert in der Regel am besten - oder zumindest nicht schlechter - mit der Belästigungswirkung als andere Verfahren. Von dieser Erfahrung berichten viele Autoren, z.B. Vallet et al. [36], Rasmussen [30], Bullen & Hede [3], Öhrstöm, Björkman & Rylander [27], De Jong [6]. Für Düsseldorf wurden mir von der Flughafen-Umweltmeßtechnik für fünf Meßpunkte sowohl die L_{eq3} -Werte als auch die nach dem Fluglärngesetz von 1971 ermittelten L_{eq4} -Werte für die jeweils sechs verkehrsreichsten Monate der Jahre 1981-1993 zur Verfügung gestellt. Hieraus errechneten sich folgende Korrelationskoeffizienten (N=13, dh. über die Jahre hinweg) zwischen den beiden Pegeln:

Meßpunkt:	Lohausen	Tiefenbroich	Lintorf	Hösel	Ratingen-West
Korrelationskoeffizient:	0.997	0.992	0.996	0.991	0.998

Auch diese Korrelationen sind so hoch, daß man geneigt ist, den L_{eq4} als entbehrlich zu betrachten. Rice [32] bezeichnet denn auch den L_{eq3} vorsichtig als das beste "overall cumulative noise measure" oder die beste "commonly used noise scale" [31].

Diese sich anbahnende Einsicht hat inzwischen auch zu praktischen Konsequenzen geführt. So liegt z.B. eine offiziellen Empfehlung des englischen Department of Transport (DOT) vor, den NNI (s.u.) durch den L_{eq3} zu ersetzen, insbesondere bei der Erstellung von Lärmkonturen [28]. Was Deutschland betrifft, so wurde zwar 1971 im Fluglärngesetz der 'äquivalente Dauerschallpegel' (L_{eq4}) zur Ermittlung von Lärmkonturen festgeschrieben, aber für die Rechtsprechung bei Streitigkeiten in Fragen der Lärmbelästigung ist nach einem Urteil des Bundesgerichtshofes der L_{eq3} heranzuziehen [1].

Ein Problem bei der Verwendung des energie-äquivalenten Dauerschallpegels und davon abgeleiteter Maße scheint jedoch in gewissen Verständnisschwierigkeiten zu liegen, was Ollerhead [28] in die Worte kleidet, daß "the L_{eq} model ... requires more complicated logic than the NNI model". Diese Unanschaulichkeit veranlaßt Rice [32] zu der Feststellung, daß der L_{eq3} Beziehungen zwischen Lärmkomponenten ehe verschleierte, wie z.B. den 'trade-off' von Anzahl und Pegel. Hierzu sei angemerkt, daß der L_{eq3} nicht nur aus praktischen, sondern auch aus theoretischen Erwägungen ein bestes Maß darstellen würde, weil er sich physikalisch als Wirkungsgröße interpretieren läßt, der man einen biologischen Bezug unterstellen kann, welcher auf das Belästigungserlebnis führt [17]. Ferner sei darauf hingewiesen, daß sich der L_{eq3} durch einfache Umrechnungen in eine Form gebracht werden kann, die eine auch anschauliche Interpretation nahelegt [17].

Allerdings gilt, wie schon Rice[31] feststellt, daß es schwierig ist zu zeigen, daß irgendein anderes Maß besser ist als der L_{eq3} . Die eigentlich zu beantwortende Frage ist also die: Wie hoch muß die Korrelation zwischen zwei Lärm-Maßen sein und welche Kriterien müssen sonst noch erfüllt sein, um mit Sicherheit sagen zu können, daß die beiden Maße äquivalent sind?

Ziel der vorliegenden Arbeit ist es, die oben angedeutete Diskussion durch Rückgriff auf die psychologische Testtheorie auf eine rationale Basis zu stellen und Kriterien vorzuschlagen, die zu unterscheiden gestatten, wann physikalische Lärmengengemaße äquivalent sind und wann nicht, und mit welcher Genauigkeit mit ihnen Belästigungswirkungen vorhergesagt werden können.

Im folgenden wird zunächst die psychologische Testtheorie in den Grundzügen dargestellt. Für Einzelheiten, wie Herleitungen von Formeln, Definitionen und Voraussetzungen wird auf die einschlägige Literatur verwiesen [11, 24, 8, 37]. Es folgt die Anwendung auf die Lärmmessung und eine Abschätzung, welche der physikalischen Lärmengengmaße als äquivalent angesehen werden können.

2 Abriß der psychologischen Testtheorie

Mathematisch gesehen handelt es sich bei der psychologischen Testtheorie um die Anwendung der Korrelations- und Regressionsrechnung (vgl. [4]) auf psychologische Meßgrößen, um Gütekriterien für psychologische Testverfahren zu definieren. Wichtige Begriffe in diesem Zusammenhang sind Zuverlässigkeit (Reliabilität) und Meßfehler sowie Gültigkeit (Validität) und Schätzfehler. Die Testtheorie wurde ursprünglich entwickelt, um den aus den klassischen Naturwissenschaften bekannten Begriff des Meßfehlers auch unter den besonderen Bedingungen psychologischen Messens verfügbar zu haben (vgl. dazu auch [24]). Hier bestehen die Schwierigkeiten unter anderem darin, daß eine intendierte psychologische Variable oft nur schwer, ungenau oder nur sehr umständlich zu messen ist - in der Testtheorie bezeichnet man diese als "Kriteriumsvariable" - während andere Variablen erheblich leichter zugänglich sind und auch genauer gemessen werden können, aber das Kriterium nicht vollständig erfassen - diese werden dann als "Testvariablen" bezeichnet. Es liegt nahe, das jeweils verwendete physikalische Lärmengengmaß als 'Test' und die jeweils betrachtete psychologische Wirkung als 'Kriterium' aufzufassen, um dann den Formalismus der Testtheorie anwenden zu können.

2.1 Grundlegende Definitionen

In der - mittlerweile als klassisch bezeichneten - psychologischen Testtheorie setzt man eine Testvariable X als Zufallsvariable an, für die man einen konkreten Wert x (=Realisation von X) erhält, wenn man den Test auf einen Probanden bzw. Merkmalsträger anwendet. Mit μ_X bezeichnet man wie üblich den Erwartungswert (Mittelwert), mit σ^2_X die Varianz und mit σ_X die Standardabweichung von X . Typisch für die Testtheorie ist nun, daß man sich X aus zwei nicht direkt beobachtbaren stochastisch unabhängigen Zufallsvariablen T und E additiv als $X = T + E$ zusammengesetzt denkt. T bezeichnet man als die Variable, die den "True Score" (wahren Wert) repräsentiert, E als "Error-Variable" (Fehler-Variable).

Ist Y eine weitere Zufallsvariable, so wird mit σ_{XY} die Kovarianz und mit $r_{XY} = \sigma_{XY}/(\sigma_X\sigma_Y)$ der Korrelationskoeffizient zwischen X und Y bezeichnet. r_{XY} kann zwischen -1 und +1 variieren. Besteht zwischen X und Y kein Zusammenhang, so wird $r_{XY}=0$, während $r_{XY}=1$ den maximalen gleichsinnigen und $r_{XY}=-1$ den maximalen gegensinnigen Zusammenhang widerspiegelt. Angemerkt sei, daß bei linearen Transformationen der Art $x'=ax+b$ und $y'=cy+d$ sich zwar Varianz und Kovarianz verändern, der Korrelationskoeffizient zwischen den Variablen aber erhalten bleibt. Der Korrelationskoeffizient spiegelt damit die 'gemeinsame' Variation der Variablen wider, unabhängig von der absoluten Höhe der Mittelwerte und Standardabweichungen.

Die "Reliabilität" oder der "Zuverlässigkeitskoeffizient" des Tests X wird durch r_{XX^*} gekennzeichnet, wobei X^* einen zu X äquivalenten (parallelen) Test bedeutet, durchgeführt an derselben Stichprobe von Probanden. X^* steht im einfachsten Falle für die Wiederholung von X . Der Fall mit mehreren äquivalenten Tests wird weiter unten behandelt. Der "Standard-Meßfehler" $\sigma_X(E)$ ist dann definiert als

$$\sigma_X(E) = \sigma_X \cdot \sqrt{1 - r_{XX^*}} \quad (1)$$

Er ist ein Maß für die zu erwartende Abweichung zwischen den Meßwerten für dasselbe Objekt bei Meßwiederholung bzw. Applikation eines äquivalenten Tests. Ist x der erhaltene Testwert eines Probanden, so geht man davon aus, daß der "wahre" Wert des Probanden mit einer

Wahrscheinlichkeit von 95% im Intervall $x \pm 1.96 \sigma_X(E)$ liegt.

2.2 Individuenbezogene Regression

Stellt Y eine Kriteriumsvariable dar (das ist -formal gesehen- ebenfalls ein 'Test'), so bezeichnet man mit r_{XY} die "Gültigkeit" (Validität) des Tests X bezüglich des Kriteriums Y . Die (lineare) Regression von Y bezüglich X ist dann gegeben durch

$$y' = \mu_Y + r_{XY} \cdot \sigma_Y / \sigma_X \cdot (x - \mu_X) \quad (2)$$

Hierbei bedeutet y' den durch x mittels der Formel (2) geschätzten "wahren" Wert y hinsichtlich des Kriteriums Y . Man sagt auch: Regression von y auf (den Basiswert) x . Die durch die Gleichung (2) festgelegte Gerade bezeichnet man als Regressionsgerade oder -in anderen Zusammenhängen- auch als Dosis-Wirkungs-Kurve, wobei x die Dosis und y die Wirkung bedeuten.

Der "Standard-Schätzfehler" ist gegeben durch

$${}_x\sigma_Y(E) = \sigma_Y \sqrt{1 - r_{XY}^2} \quad (3)$$

und ist ein Maß für die zu erwartende Abweichung zwischen dem Schätzwert y' für ein Objekt aufgrund von (2) und seinem "wahren" Meßwert y . Mit einer Wahrscheinlichkeit von z.B. 95% liegt y im Intervall $y' \pm 1.96 \cdot {}_x\sigma_Y(E)$. Mit der Angabe $\pm 1.96 \cdot {}_x\sigma_Y(E)$ wird daher ein Streifen (Konfidenzintervall) um die Regressionsgerade gelegt, in dem mit 95%iger Sicherheit die tatsächlichen Meßwerte hinsichtlich des Kriteriums Y lokalisiert werden können. Der Gültigkeitskoeffizient r_{XY} gibt also an, wie gut sich das Kriterium durch den Test vorhersagen läßt (s. Abbildung 1).

X^* sei nun eine weitere Messung von X ("Paralleltest" von X) und Y^* eine weitere Messung von Y ("Paralleltest" von Y). Die jeweiligen Korrelationskoeffizienten r_{XX^*} und r_{YY^*} zwischen X und X^* bzw. Y und Y^* (also die "Zuverlässigkeitskoeffizienten") seien als bekannt vorausgesetzt, ebenso sei die Korrelation zwischen X^* und Y^* gegeben. Dann gilt folgende Formel ("Verdünnungsformel"):

$$r_{XY} = \frac{r_{X^*Y^*}}{\sqrt{r_{XX^*} \cdot r_{YY^*}}} \quad (4)$$

(4) gibt an, wie groß r_{XY} unter den mit der rechten Seite von (4) gegebenen Umständen bestenfalls werden kann. Hat also ein Test X bereits eine hohe Reliabilität, so kann der Gültigkeitskoeffizient durch eine weitere Verbesserung dieses Tests kaum noch angehoben werden.

2.3 Regression bei gruppierten Stichproben

Bildet man Gruppen von Personen mit gleichem x -Wert (Basiswert), betrachtet also z.B. Personen, die physikalisch gleich hoch lärmbelastet sind, so lassen sich die Gruppenmittelwerte hinsichtlich der Variablen Y , also z.B. der Belästigung, aus den jeweiligen Basiswerten erheblich genauer vorhersagen. Dieses Vorgehen ist in Abbildung 1 angedeutet durch die vier hochgestellten ellipsenförmigen Konturen, die jeweils Teilmengen von Punkten aus der großen Punktwolke abgrenzen. Die Mittelwerte hinsichtlich X und Y dieser Teilmengen bilden nunmehr eine "Wolke" von nur noch vier Punkten (\bar{X}_i, \bar{Y}_i) , $i = 1, 2, 3, 4$, welche die Abhängigkeit jetzt der Gruppenmittelwerte im Merkmal Y von den Gruppenmittelwerten im Merkmal X wiedergibt. Die sich daraus ergebende zweite "Regressionsgrade" muß theoretisch natürlich mit der ersten Regressionsgeraden übereinstimmen. Geht man nun von einer Probandengruppe mit dem Mittelwert \bar{X}_i im Merkmal X aus, so erhält man für den per Regression analog (2) berechneten Schätzwert \bar{Y}'_i für den Mittelwert \bar{Y}_i im Merkmal Y dieser Gruppe einen erheblich kleineren Standard-Schätzfehler, nämlich

$${}_{\bar{x}}\sigma_{\bar{y}}(E) = \sigma_{\bar{y}} \sqrt{1 - r_{\bar{x}\bar{y}}^2} \quad (5)$$

wobei $r_{\bar{x}\bar{y}}$ jetzt die Korrelation zwischen den Basis-Werten \bar{X} , die die Gruppen kennzeichnen, und den zugehörigen Gruppenmittelwerten \bar{Y} bedeutet, und $\sigma_{\bar{y}}$ die Standardabweichung, gerechnet über die Gruppenmittelwerte im Merkmal Y , die in der Regel wegen der Minderung des Standardmeßfehlers beim Übergang von Y - auf \bar{Y} -Werte etwas kleiner als σ_Y ist.

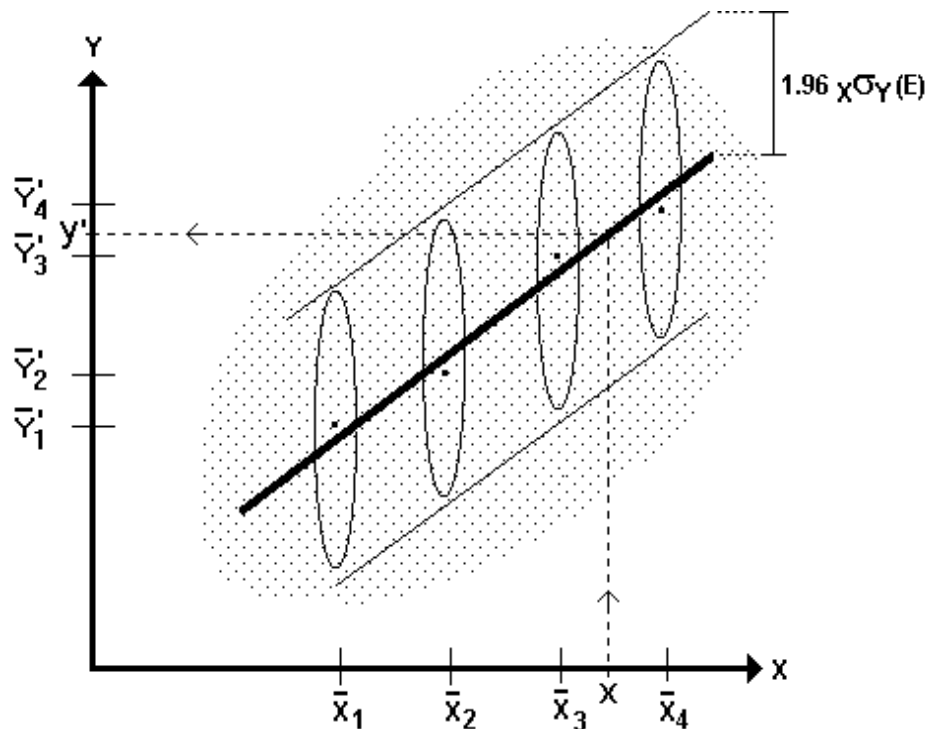


Abbildung 1: Regression von Y auf X. Jeder Punkt repräsentiert einen Probanden. Ausgehend von einem Basiswert x wird der y-Wert eines Probanden als Mittelwert y' der Gruppe von Probanden mit demselben Wert für x vorhergesagt. Durch die hochgestellten Ellipsen sollen Probandengruppen angedeutet werden, bei denen auf Basis der Gruppenmittelwerte \bar{X} die Gruppenmittelwerte \bar{Y} geschätzt werden sollen.

Ohne Beweis sei angeführt, daß für große Umfänge n der Teilstichproben, an Hand derer $r_{\bar{X}\bar{Y}}$ berechnet wurde, $r_{\bar{X}\bar{Y}}$ gegen eins tendieren muß, sofern nur der "wahre" Zusammenhang bzw. der Gültigkeitskoeffizient r_{XY} , ungleich Null und die Regression linear sind. Für genügend große n tendiert unter diesen Voraussetzungen der Standardschätzfehler für die Gruppenmittelwerte also in jedem Falle gegen Null. Daraus folgt, daß die Angabe der - in der Regel recht hohen - Korrelationskoeffizienten zwischen den Basiswerten und den Gruppenmittelwerten der beteiligten Variablen, wie dies häufig geschieht, keinen Informationswert hinsichtlich der für den Schätzfehler entscheidenden Größe, nämlich des (über die Individuen berechneten) Gültigkeitskoeffizienten, hat. Denn wie groß bzw. klein auch immer die Gültigkeit in Wirklichkeit ist - man wird für die Korrelation zwischen den Basiswerten und den Gruppenmittelwerten einen hohen Wert bekommen, sofern nur genügend Probanden rekrutiert wurden, der individuenbezogene Korrelationskoeffizient verschieden von Null und die Regression linear ist.

2.4 Äquivalente Tests

Um die Beziehungen zwischen äquivalenten Tests untereinander und zu einem Kriterium - genau um diese Fragen geht es in der vorliegenden Arbeit - klar darzustellen, wird auf eine modellhafte Darstellung mit drei Testvariablen X_1, X_2, X_3 und einer Kriteriumsvariablen Y , formal als vierte Variable X_4 geführt, zurückgegriffen, welche der sog. Faktorenanalyse entlehnt ist (vgl. auch [16]):

$$\begin{aligned}
 X_1 &= a_{11} \cdot F_1 + a_{12} \cdot F_2 + \dots + a_{16} \cdot F_6 + b_1 \cdot S_1 + c_1 \cdot E_1 \\
 X_2 &= a_{21} \cdot F_1 + a_{22} \cdot F_2 + \dots + a_{26} \cdot F_6 + b_2 \cdot S_2 + c_2 \cdot E_2 \\
 X_3 &= a_{31} \cdot F_1 + a_{32} \cdot F_2 + \dots + a_{36} \cdot F_6 + b_3 \cdot S_3 + c_3 \cdot E_3 \\
 Y = X_4 &= a_{41} \cdot F_1 + a_{42} \cdot F_2 + \dots + a_{46} \cdot F_6 + b_4 \cdot S_4 + c_4 \cdot E_4
 \end{aligned} \tag{6}$$

Hierbei werden die Zufallsvariablen F_1, F_2, \dots, F_6 *gemeinsame Faktoren*, $S_1 - S_4$ *spezifische Faktoren* und $E_1 - E_4$ *Fehlerfaktoren* genannt. Alle diese Faktoren werden als stochastisch voneinander unabhängig mit Mittelwert 0 und Standardabweichung 1 vorausgesetzt. Auch für $X_1 - X_4$ wird angenommen, daß der Mittelwert 0 und die Standardabweichung 1 ist. Dies bedeutet keine Einschränkung der Allgemeinheit, da jede Zufallsvariable durch eine geeignete lineare Transformation - nämlich durch die sog. Standardisierungsoperation $x' = (x - \mu_X) / \sigma_X$ - auf eine solche 'Standardform' gebracht werden kann, ohne daß die korrelativen Beziehungen sich ändern. Die Konstanten a_{ij}, b_i, c_i heißen Ladungen der Variablen $X_1 - X_4$ auf den betreffenden Faktoren. Ersichtlich kann das Modell auch auf mehr als drei Tests erweitert werden.

Aus (6) ergibt sich der Korrelationskoeffizient zwischen den Testvariablen X_i und X_j zu

$$r_{X_i X_j} = \sum_{k=1}^6 a_{ik} \cdot a_{jk}, \quad i, j = 1, 2, 3, 4 \text{ mit } i \neq j \tag{7}$$

und die Reliabilität der Tests zu

$$r_{X_i X_i^*} = \sum_{k=1}^6 a_{ik}^2 + b_i^2 \tag{8}$$

Hieran kann man gut erkennen, daß Korrelationen zwischen den Variablen nur von den gemeinsamen Varianzquellen $F_1 - F_6$ herrühren können, während für die Reliabilität noch die spezifische Varianz b_i^2 zu berücksichtigen ist. Die Fehler auch aufeinanderfolgender Messungen werden als unkorreliert betrachtet und tragen daher weder zu den Interkorrelationskoeffizienten noch zur Reliabilität bei. Als äquivalent - oder parallel- könnte man die Tests $X_1 - X_3$ z.B dann bezeichnen, wenn jeder Test in der gemeinsamen Variabilität, die er mit dem Kriterium Y teilt, mit den anderen Tests übereinstimmt. Die Tabelle 1 zeigt links einen unproblematischen Fall, bei dem man die Tests in diesem Sinne als äquivalent betrachten kann, und rechts ein Beispiel für eine "pathologische" Situation mit nur scheinbar äquivalenten Tests.

Tabelle 1: Beispiele für Ladungsmatrizen von Test-Kriterium-Systemen gem. (6) mit gegebener (links) und fehlender (rechts) Äquivalenz. Ein Feld ohne Eintrag enthält eine Ladung von null. Ladungen auf nicht aufgeführten Faktoren sind ebenfalls zu null angenommen.

	F1	F2	F3	F4	F5	F6	E4
X_1, X_2, X_3 äquivalent:							
X_1	.71	.5	.5				
X_2	.71	.5	.5				
X_3	.71	.5	.5				
$X_4=Y$.5				.5	.5	.5
	$r_{12}=r_{13}=r_{23}=1, \text{ mit } r_{11}=r_{22}=r_{33}=1$ $r_{14}=r_{24}=r_{34}=0.36, \text{ mit } r_{44}=0.75$						
X_1, X_2, X_3 nicht äquivalent:							
X_1	.5	.5		.71			
X_2	.5		.5		.71		
X_3		.5	.5			.71	
$X_4=Y$.5	.5	.5	.5
	$r_{12}=r_{13}=r_{23}=0.5, \text{ mit } r_{11}=r_{22}=r_{33}=1$ $r_{14}=r_{24}=r_{34}=0.36, \text{ mit } r_{44}=0.75$						

Wie am rechten Beispiel in Tabelle 1 ersichtlich, kann durchaus der Fall eintreten, daß Tests deutlich untereinander und mit dem Kriterium korrelieren, aber so, daß die Gemeinsamkeiten der Tests nicht im Kriterium vertreten sind. Mit Bezug auf das Kriterium sind die Tests also tatsächlich nicht äquivalent, weil sie unterschiedliche Aspekte des Kriteriums erfassen, und untereinander sind sie nicht äquivalent, weil je zwei der Tests über unterschiedliche Gemeinsamkeiten zusammenhängen. Solches wird ausgeschlossen, wenn man zusätzlich fordert, daß die Testinterkorrelationskoeffizienten gleich der Reliabilität der Tests sein müssen.

Für den allgemeinen Fall können in Anlehnung an Lienert [24] S.348 ff.) vier Äquivalenzkriterien im Sinne von notwendigen Bedingungen für das Vorliegen von Äquivalenz von

Tests formuliert werden; sie betreffen Validität, Reliabilität, Verteilungskennwerte und Verteilungsformen der Tests:

- I. die Validitätskoeffizienten der Tests müssen übereinstimmen,
- II. a. die Reliabilitätskoeffizienten der Tests müssen übereinstimmen,
b. die Tests müssen untereinander gleich hoch korrelieren,
c. die Interkorrelationskoeffizienten der Tests sollten möglichst hoch sein - im Idealfall gleich der Reliabilität. Zusammen mit II.a. und II.b. bedeutet dies, daß die gemeinsame Varianz der Tests möglichst nahe an ihre reliable Varianz heranreichen sollen.
- III. die Mittelwerte und Standardabweichungen der Tests sollten gleich sein,
- IV. die Häufigkeitsverteilungen der Testwerte sollten übereinstimmen.

Hiervon läßt sich die Forderung III. insofern stets erfüllen, als man die Testwerte leicht in geeigneter Weise linear transformieren kann (s.o.), z.B. durch Standardisierung auf Mittelwert 0 und Standardabweichung 1. Die Forderung IV. nach Gleichheit der Verteilungen hat erst dann eine gewisse Bedeutung, wenn nicht-normale Verteilungen vorliegen, insbesondere mehrgipflige. Ob die Forderungen I. und II. erfüllt sind, muß dann im Einzelfall geprüft werden.

3 Anwendung auf die Lärmwirkungsmessung

In der Lärmwirkungsmessung schließt man von dem jeweils erhaltenen physikalischen Meßwert für den Lärm auf die zugehörige psychologische Lärmwirkung, also die Stärke des psychologischen Belästigungserlebnisses. Wendet man die vorangegangenen Überlegungen hierauf an, so ist das physikalische Maß der Lärmmenge (z.B. der L_{eq3}), der ein Proband unterworfen ist, mit der "Testvariablen X" und die subjektive Belästigung A, die er dabei erlebt, mit der "Kriteriumsvariablen Y" zu identifizieren. Über die Regressionsformel (2) erhält man dann für jeden physikalischen Meßwert x einen Schätzwert für die zugehörige psychologische Größe y. Man kann dabei davon ausgehen, daß die Messung des Lärms - wie meistens bei physikalischen Messungen - eine Reliabilität von nahezu eins hat. Aus den angestellten Überlegungen folgt dann, daß die hohe Genauigkeit der physikalischen Messung des Lärms sich nicht automatisch auch auf die eigentlich intendierte Messung der psychologischen Wirkung überträgt. Wie aus den Gleichungen (2) und (3) hervorgeht, ist für die Präzision, mit der die Belästigung aus der physikalischen Lärm-Messung vorhergesagt werden kann, die Größe des Standard-Schätzfehlers maßgebend, und der wiederum wird vom Validitätskoeffizienten bestimmt.

Die Lärm-Meßtechnik muß also sowohl auf physikalische als auch psychologische Verfahren zurückgreifen. Zunächst werden daher die physikalischen Verfahren und danach die psychologischen Verfahren behandelt. Anschließend werden Äquivalenz- und Gültigkeitsfragen erörtert.

3.1 Physikalische Verfahren zur Erfassung des Lärms

Die derzeit wohl allgemeinste Definition eines nur auf physikalischen Größen beruhenden Maßes für die Lärm-Menge ist der von Bürck et al. [2] vorgeschlagene 'Störindex' Q, auch 'äquivalenter Dauerschallpegel' genannt:

$$Q(k) = k \cdot \log \left\{ \frac{1}{T} \int_0^T 10^{1/k \cdot L(t)} dt \right\} \quad [\text{dB}], \quad \text{wobei}$$

$$L(t) = 10 \cdot \log (p(t)/p_0)^2 = 10 \cdot \log I(t)/I_0 \quad [\text{dB}], \quad \text{mit} \quad (9)$$

$$p_0 = 20 \cdot 10^{-6} \text{ Pa}$$

Hierbei ist T der Beobachtungszeitraum (häufig 16 oder 24 Stunden). L wird in der Regel als Schall(intensitäts)pegel oder Momentanpegel zum Zeitpunkt t bezeichnet. p(t) ist die zur Zeit t herrschende Schalldruckamplitude, $p_0 = 20 \mu\text{Pa}$ der Bezugsschalldruck. Die Druckwerte p bzw. Intensitätswerte I sind in der Regel als A-bewertet zu verstehen, so daß statt L eigentlich immer L_A bzw. statt dB immer dB_A oder $\text{dB}(A)$ geschrieben werden müßte. Der Buchstabe A wird aber im Folgenden meist weggelassen.

Manchmal wird der Momentanpegel auch als "perceived noise level", abgekürzt L_{pn} , angegeben [19, 20]). Bei der technischen Bestimmung dieses Pegels wird das betreffende Geräusch

durch eine Frequenzanalyse zunächst in Terzbänder zerlegt und dann die Energie in den einzelnen Terzbändern, mit geeigneten Gewichten versehen, addiert. Die Frequenzbereiche von 2000 bis 5000 Hz werden dabei stärker berücksichtigt. Überschlagsmäßig gilt $L_{pn}=L+13\text{dB}$, wobei L der A-bewertete Pegel nach (9) ist (vgl. [7, 34]).

Der Parameter k gibt, ohne daß dies aus (9) direkt schon erkennbar wäre, das Gewicht an, mit dem - im Vergleich zu den Momentanpegeln - Häufigkeit und Dauer von akustischen Ereignissen den Zahlenwert von Q(k) bestimmen (s.u.). Statt k wird oft auch der sog. Halbierungs- bzw. Verdoppelungsparameter q angegeben, um dieses Gewicht zu kennzeichnen. Die Beziehung zwischen beiden Parametern ist $k=q/\log 2$. Bei Verwendung von q schreibt man statt Q(k) meistens L_{eqx} . Den konkreten Zahlenwert von q hängt man dann als Indexzahl x an. Die Zahl x gibt an, um wieviel dB sich der betreffende Leq-Wert verkleinert bzw. vergrößert, wenn bei konstantem Pegel L Häufigkeit und/oder Dauer halbiert bzw. verdoppelt werden. L_{eq3} ist danach gleichbedeutend mit k=10 und bedeutet, daß z.B. bei Verdoppelung der Häufigkeit sich der L_{eq3} -oder Q(10)- um 3 dB erhöht. Der L_{eq3} repräsentiert also den energie-äquivalenten Dauerschallpegel. Für q=4 (entspricht k=13.3) ergibt sich der "äquivalente Dauerschallpegel" des Fluglärmgesetzes von 1971. Q trägt zwar den Namen 'Störindex', was eine psychologische Größe suggeriert, ist aber tatsächlich eine (quasi-) physikalische Größe, weil in die definierende Gleichung (9) nur physikalische Größen eingehen.

Der Zusammenhang von Q bzw. L_{eqx} mit anderen Maßen für die Lärmmenge ergibt sich aus der Überführung von (9) in die Summenform, wobei auch einige Vereinfachungen vorgenommen werden:

$$Q(k) = k \cdot \log \left\{ \frac{1}{T} \cdot \sum_{i=1}^N 10^{1/k \cdot L_i} \cdot \Delta t_i \right\} \quad (10)$$

Hierbei ist angenommen, daß sich aus dem Geräuschbild während der Zeit T insgesamt N zeitlich getrennte Ereignisse perzeptiv abgrenzen lassen, z.B. Überflüge, zwischen denen der Momentanpegel auf vernachlässigbar kleine Werte abfällt. Mit Δt_i ($i=1,2,\dots,N$) sind dann die Dauern dieser Ereignisse bezeichnet, für die jeweils ein Pegelwert L_i festgesetzt wird, der charakteristisch für das jeweilige Ereignis ist. Man kann hierfür z.B. den für k=10 nach (9) berechneten äquivalenten Dauerschallpegel nehmen. Häufig werden hierfür auch die Spitzenpegel L_{max_i} (=Maximalwerte der Momentanpegel) der einzelnen Ereignisse eingesetzt, wenn angenommen werden kann, daß die Spitzenpegel ausschlaggebend sind, wovon im Folgenden ausgegangen werden soll. Nimmt man zunächst weiterhin an, daß die Ereignisdauern und Spitzenpegel konstant sind ($\Delta t_i=\Delta t$ und $L_{max_i}=L_{max}$ bzw. $I_{max}=I_{max}$, wobei mit den letzten beiden Symbolen die Intensitäten gemeint sind), wird aus (10)

$$Q(k) \approx k \cdot \log \left\{ \frac{\Delta t}{T} \cdot N \cdot (I_{max}/I_0)^{1/k} \right\} \\ = L_{max} + k \cdot \log N + k \cdot \log (\Delta t/T) \quad (11)$$

Bezeichnet man nun den arithmetischen Mittelwert der Ereignisdauern mit $\overline{\Delta t}$ und den Mittelungspegel der Spitzenpegel (sog. energetische Mittelung, s.u.) mit $\overline{L_{max}}$, so erhält man als Annäherung an Q(k)

$$Q(k) \approx \overline{L_{max}} + k \cdot \log N + k \cdot \log (\overline{\Delta t}/T) \quad (12)$$

wobei der Fehler von der Variabilität der Werte abhängt, aus denen die beiden Mittelwerte berechnet werden. Die hohe Korrelation zwischen den Q-Werten, die nach (9) und (12) mit verschiedenen k-Werten und Variabilitäten vom Autor berechnet wurden, zeigen jedoch, daß der Fehler in praktisch vorkommenden Fällen zu vernachlässigen ist (nicht veröffentlicht). (12) ist wiederum ein Spezialfall des Ausdrucks

$$Q(a,b,c) = a \cdot \overline{L_{max}} + b \cdot \log N + c \cdot \log (\overline{\Delta t}/T), \quad (13)$$

der von Kalveram[17] als verhaltensökologisch begründbares Maß für die Lärmmenge vorgeschlagen worden ist, wobei a, b, c geeignete Konstanten (Gewichte) darstellen.

3.2 In der vorliegenden Arbeit benutzte physikalische Lärm-Maße

Viele derzeit benutzte Maße für Lärm ergeben sich aus der Definitionsgleichung (9), ihrer Vereinfachung (12) oder ihrer Verallgemeinerung (13) durch spezielle Definition des ein Einzelereignis kennzeichnenden Pegels L_i , spezielle Wahl des Parameters k bzw. der Gewichte a, b, c , Vereinfachungen durch Weglassung von Komponenten, Hinzufügen oder Streichung von additiven Konstanten. Im folgenden sollen einige der Maße aufgelistet werden, um einen Eindruck von deren Vielfalt und Ähnlichkeit zu geben. Die Aufzählung ist jedoch nicht vollständig. Ausführlichere Angaben findet man z.B. bei Schick [34].

D_{10} - Ereignisdauer

Zeit, in der ein Pegelwert, der 10 dB unter dem Spitzenpegel des betreffenden Ereignisses liegt, überschritten wird.

D_{80} - Ereignisdauer (DFG)

Zeit, in welcher der Pegelwert des betreffenden Ereignisses 80 dB übersteigt.

H_{81} - Ereignishäufigkeit (DFG)

Anzahl der Ereignisse, bei denen Pegelwerte größer oder gleich 81 dB vorkommen.

L_{av} - Arithmetischer Mittelwert von Pegeln (DFG)

'Arithmetischer' Mittelwert von Einzelpegeln L_i :

$$L_{av} = 1/N \sum_{i=1}^N L_i$$

L_m - Mittelungspegel

"Energetischer" Mittelwert von Einzelpegeln L_i :

$$L_m = 10 \cdot \log \left(\frac{1}{N} \sum_{i=1}^N 10^{0.1 L_i} \right)$$

Für $L_i = L_{max_i}$ erhält man $L_m = L_{max}$.

L_s - Summenpegel (DFG)

'Energetische' Summe von Einzelpegeln L_i

$$L_s = 10 \cdot \log \left(\sum_{i=1}^N 10^{0.1 L_i} \right) = L_m + 10 \cdot \log N$$

L_{eq3} - Energie-äquivalenter Dauerschallpegel

Ergibt sich aus (9) bzw. (12) für $k=10$. Der L_{eq3} ist äquivalent einer Gesamtenergie, gleichgültig, wie sich die Schallintensität über die Zeit verteilt. Man sollte sich daher immer vor Augen halten, daß ein gleichmäßiges Geräusch, welches dasselbe Integral erzeugt wie eins, aus dem sich Einzelereignisse perceptiv abheben, subjektiv möglicherweise eine andere Wirkung hat.

L_{eq4} - Äquivalenter Dauerschallpegel des Fluglärmsgesetzes

Setzt man in (9) bzw. (12) $k=13.3$ (entspricht $q=4$), so erhält man den, bei Bürck et. al. (1965) auch als Störindex bezeichneten, äquivalenten Dauerschallpegel L_{eq4} , im Fluglärmsgesetz mit L_{eq} abgekürzt. Zur praktischen Berechnung wird D_{10} als Überflugdauer genommen.

NNI - Noise and Number Index (GB)

Ersetzt $\overline{L_{max}}$ durch den mittleren maximalen perceived noise level $\overline{L_{max_{pn}}}$, nimmt für $\overline{\Delta t}$ ungefähr $1/30000$ an und setzt $k=15$, erhält man aus (12) bei $T=24$ Stunden

$$NNI = \overline{L_{max_{pn}}} + 15 \log N - 80.$$

FB1 - Fluglärm-Bewertungsmaß 1 (DFG)

$$FB1 = L_m + 20 \log N - 50$$

CNR - Composite Noise Rating (USA)

Dieses Maß geht vom L_{pn} -Wert der einzelnen Überflüge aus, dem nach einem komplizierten Bewertungssystem verschiedene Korrekturwerte c_j hinzuaddiert werden, je nach Tonhaltigkeit, Tages- oder Nachtzeit, Flugzeugtyp usw. . Der CNR-Wert ergibt sich dann aus dem Mittelwert $\overline{L_{pn}}$ der L_{pn} -Werte, einer Art Mittelwert c der Korrekturwerte und der Ereigniszahl N in der Beobachtungszeit als $CNR = \overline{L_{pn}} + 10 \log N + c$

NEF - Noise Exposure Forecast (USA)

Weiterentwicklung des CNR. Hier werden Überflüge während der Nacht durch Multiplikation der nächtlichen Anzahlen mit dem Faktor 12 bewertet.

L_{dn} - Day-Night-Level

Entspricht dem L_{eq3} , nur daß Flüge in der Nacht um 10 dB höher bewertet werden, bevor sie in den Gesamtmittelwert einbezogen werden. Die genaue Definition lautet

$L_{dn} = 10 \log 1/24 \{15 \cdot 10^{L_d/10} + 9 \cdot 10^{(L_n+10)/10}\}$, wobei L_d und L_n die (A-bewerteten) L_{eq3} -Werte während der Tagzeit (7-22h) und Nachtzeit (22-7h) bedeuten.

3.3 Reliabilität und Äquivalenz der physikalischen Lärm-Mengen-Maße

Hinsichtlich der Frage, ob es sich bei den verschiedenen Verfahren zur Messung der Lärmmenge um (äquivalente) Paralleltests im Sinne der Testtheorie handelt, reicht es, wenn die Verteilungsformen als gleich angesehen werden, aus, die Äquivalenzkriterien I. und II. zu überprüfen. Eine Untersuchung, in der für diesen Zweck geeignete Daten erhoben wurden, wurde 1969 im Rahmen des DFG-Projekts "Fluglärmwirkungen" in München durchgeführt [7]. Eine Auswahl der im Untersuchungsbericht [7] mitgeteilten Korrelationskoeffizienten ist in Tabelle 2 zu finden. In die Korrelationsstudie gingen 357 Personen ein, die auf 32 Gruppen (Cluster) mit unterschiedlicher Belastung durch Fluglärm aufgeteilt waren. Die Korrelationskoeffizienten zwischen den physikalischen Maßen wurden über die 32 Cluster-Mittelwerte berechnet.

Tabelle 2: Korrelationskoeffizienten zwischen verschiedenen Lärm-Expositionsmaßen und zwischen den Expositionsmaßen und der Belästigung (untere Zeile) (nach [7]). Zur Bedeutung der einzelnen Maße s. Kapitel 3.2. Die als äquivalent anzusehenden Maße (s. 1. Spalte) und die zugehörigen Korrelationskoeffizienten (s. Spalten 8-13) sind fett gedruckt. Der Korrelationskoeffizient zwischen L_{eq1} und L_{eq4} fällt jedoch bereits schon aus diesem Rahmen.

Physikal. Maß	D10	D80	10 log N	H81	Lav	Lm k=0	Leq1 k=3.3	Ls k=10	Leq3 k=10	Leq4 k=13.3	NNI k=15	FB1 k=20	Leq10 k=33.3
-D10		.106	.614	.747	.827	.823	.817	.819	.774	.747	.805	.793	.566
D80			.262	.395	.444	.441	.434	.437	.486	.482	.416	.409	.474
10 log N				.926	.754	.770	.830	.845	.858	.882	.876	.900	.951
H81					.991	.925	.948	.960	.964	.973	.973	.980	.955
Lav						.991	.981	.982	.972	.955	.972	.960	.839
Lm							.990	.990	.981	.967	.980	.969	.853
Leq1								.996	.994	.983	.994	.987	.892
Ls									.995	.987	.997	.993	.904
Leq3										.996	.995	.992	.929
Leq4											.991	.991	.953
NNI												.998	.924
FB1													.940
-R1U Belästig.	.493	.257	.525	.569	.550	.558	.563	.576	.567	.562	.574	.579	.534

Die subjektive Belästigung (Betroffenheit) wurde in der DFG-Studie am besten durch die "Globalreaktion R1U" (s. letzte Zeile in Tabelle 2) wiedergegeben, die als gewichtete Summe verschiedener Einzelmaße gewonnen wurde. R1U wies die höchsten Korrelationskoeffizienten mit den gebräuchlichen physikalischen Maßen auf, jetzt berechnet über die 357 Probanden. Das Minuszeichen bei R1U ergibt sich aus der Art der zugrundeliegenden Skalen, die so konstruiert waren, daß die

Reaktionen um so niedriger ausfielen, je höher der Lärmpegel war. Das Minuszeichen bei D_{10} ergibt sich daraus, daß wegen der spezifischen Bedingungen der Schallausbreitung die durch D_{10} definierte Überflugdauer im Durchschnitt um so kürzer ist, je näher der Vorbeiflug erfolgt, je lauter also der betreffende Pegel ist. Die ungewöhnliche Höhe der meisten Korrelationskoeffizienten zwischen den verschiedenen Lärmengengmaßen in Tabelle 2 stimmt mit den Ergebnissen anderer Autoren überein, die ähnlich hohe Werte gefunden haben [14, 13]. Dieses liegt zum Teil auch daran, daß diese Korrelationskoeffizienten unter "natürlichen", dh. im Feld vorkommenden, Bedingungen berechnet worden sind. Unter solchen Bedingungen korrelieren schon die verschiedenen Komponenten, nämlich Spitzenpegel, Ereignishäufigkeit und Ereignisdauer, die in die Formeln eingehen, untereinander relativ hoch. Somit ist es leicht erklärlich, daß man beim Weglassen von einer oder sogar zwei dieser Komponenten, wie dies bei manchen physikalischen Maßen geschieht, immer noch hohe Korrelationskoeffizienten erhält. Unter anderen Bedingungen können daher durchaus auch andere Korrelationskoeffizienten resultieren.

Zur Überprüfung des Äquivalenzkriteriums I. wurden die 13 Validitätskoeffizienten in der letzten Zeile von Tabelle 2 zunächst hinsichtlich ihres Betrages in eine Rangreihe gebracht. Sodann wurden sie mittels aufeinanderfolgender χ^2 -Quadrat-Tests auf Gleichheit getestet. Die betreffende Prüfgröße (s. [9] Formel (721)) lautet

$$\chi^2 = \sum_{i=1}^m (n_i - 3) (z_i - \bar{z})^2, \quad (\text{df}_{\chi^2} = m - 1), \text{ wobei}$$

$$z_i = 0.5 \cdot \ln \frac{(1 + r_i)}{(1 - r_i)}, \quad \bar{z} = \frac{\sum_{i=1}^m (n_i - 3) \cdot z_i}{\sum_{i=1}^m (n_i - 3)} \quad (14)$$

Hierbei bedeuten m die Anzahl der jeweils zum Vergleich anstehenden Korrelationskoeffizienten und n_i die Anzahl der Meßwertpaare, aus denen die Korrelationskoeffizienten r_i berechnet werden. Begonnen wurde mit den beiden größten Korrelationskoeffizienten. Bei jedem folgenden χ^2 -Test wurde der nächst kleinere Korrelationskoeffizient mit einbezogen. Die Irrtumswahrscheinlichkeit wurde, da der Stichprobenumfang n_i hierbei 357 betrug, gemäß Lienert ([24] S.348) zu 1% angenommen. Für $m=13$ (also 13 zu testende Koeffizienten) überstieg $\chi^2=47.815$ erstmals die Signifikanzschranke ($\chi^2_{1\%}=26.217$, $\text{df}=12$). Damit können alle Gültigkeitsskoeffizienten in der letzten Zeile von Tabelle 2 bis auf den kleinsten (0.257) als gleich angesehen werden.

Zur Überprüfung des Äquivalenzkriteriums II. a. ist zu sagen, daß die Reliabilität (Wiederholungszuverlässigkeit) aller aufgeführten physikalischen Lärmengengmaße, wie immer bei physikalischen Meßverfahren, praktisch gleich eins sein dürfte. Hinsichtlich der Forderung II. b. wird bei der Überprüfung, welche der in Tabelle 2 angegebenen Interkorrelationskoeffizienten als gleich zu betrachten sind, ähnlich wie eben vorgegangen: Die Interkorrelationskoeffizienten werden, natürlich mit Ausnahme derjenigen in der letzten Zeile, wieder dem Betrage nach in eine Rangreihe gebracht, die jetzt 78 Glieder hat. Der Stichprobenumfang beträgt jetzt formal 32, die Irrtumswahrscheinlichkeit wird wie vorhin zu 1% angenommen. Für $m=19$ wird $\chi^2=38.420$ und übersteigt damit erstmals die Signifikanzschranke ($\chi^2_{1\%}=34.805$, $\text{df}=18$). Damit können die 18 Interkorrelationskoeffizienten, die zwischen 0.998 und 0.987 -Grenzen eingeschlossen- liegen, als gleich angenommen werden. Ein Blick auf Tabelle 2 zeigt, daß nur drei dieser Koeffizienten außerhalb der Interkorrelationsmatrix liegen, die durch die sechs Variablen Leq_1 , L_S , Leq_3 , Leq_4 , NNI und $FB1$ gebildet wird.

Nimmt man nun die Reliabilität der physikalischen Maße zu 1 an, so bleibt für den relativen Anteil der spezifischen Varianz an der Gesamtvarianz dieser sechs physikalischen Maße höchstens ein Betrag von 0.026 übrig. Damit kann auch die Forderung II. c. als erfüllt gelten und man kann von der Äquivalenz der sechs Variablen Leq_1 , L_S , Leq_3 , Leq_4 , NNI und $FB1$ ausgehen. Ein Vorbehalt muß jedoch bei der Variablen Leq_1 gemacht werden, weil bei ihr ein Korrelationskoeffizient (0.983) vorkommt, der signifikant von den übrigen Interkorrelationskoeffizienten dieser Sechsergruppe abweicht.

In Tabelle 2 sind die Korrelationskoeffizienten zwischen den als äquivalent anzusehenden physikalischen Maßen fett geschrieben. Es sei jedoch hinzugesetzt, daß diese Äquivalenz korrelationsstatistisch definiert ist. Sie bedeutet, daß solchermaßen äquivalente (parallele) Verfahren insofern ununterscheidbar sind, als sie, informationstheoretisch gesehen, dieselbe Information liefern.

Mittelwerte und Standardabweichungen können sich, wenn die Tests nicht in Standardform gebracht sind, dabei durchaus unterscheiden.

3.4 Psychologische Verfahren zur Messung der Belästigung durch Lärm

Das Belästigungserlebnis A (annoyance) wird in der Regel durch Fragebögen erfaßt, meist durch Ankreuzen einer Stelle auf einer sog. Kategorien-Skala. Dabei sind in der Regel zwei- bis 10-kategoriale Skalen in Gebrauch. Typische Beispiele für solche Skalen sind:

1. **Ich fühle mich zu Hause durch Lärm sehr stark belästigt** : ja - nein
2. **Bitte kreuzen Sie an, wie stark Sie sich durch die Geräusche beeinträchtigt gefühlt haben:**
nicht (0), sehr schwach (1), schwach (2), deutlich (3), stark (4), sehr stark (5), unerträglich stark (6)
3. "Thermometer-Skala" mit 10 Kategorien zwischen "kalt" und "heiß", deren Grenzen mit 0-10 durchnummeriert sind, wobei der Proband eine Kategorie entsprechend der subjektiven Belästigung ankreuzen soll.

Meistens wird auf getrennten Skalen nach der Lautheit, Lästigkeit oder Zumutbarkeit oder ähnlicher Beurteilungen gefragt, welche die Anwohner hinsichtlich ihrer Wohnsituation (Felduntersuchung) oder Versuchspersonen hinsichtlich eines vorgespielten Geräusches (Laboruntersuchung) abgeben. Über die Zuverlässigkeitskoeffizienten der Skalen finden sich nur spärliche Berichte, offenbar liegen sie deutlich unter 1. In einer unpublizierten Feldstudie etwa fand man eine Paralleltest-Zuverlässigkeit von 0.7. Wenn man bedenkt, daß es sich hierbei um zwei "Tests" mit nur je einem Item gehandelt hat, ist dies ein erstaunlich hoher Wert. Nach Lienert [24] reicht ein solcher Wert auch für den vorgesehenen Zweck, nämlich die Vorhersage von Gruppenmittelwerten, aus.

3.5 Validität der physikalischen Lärm-Mengen-Maße

Es wird hier wiederum zwischen individuums- und gruppenbezogener Validität unterschieden, je nachdem man ob man aus einer physikalischen Lärmmessung auf die Belästigungswirkung eines Individuums oder auf die mittlere Belästigung einer Gruppe von Individuen mit gleicher physikalischer Belastung schließen will.

3.5.1 Individuumsbezogene Gültigkeitskoeffizienten

Maßgebend für diesen Anwendungsfall ist nach (2) die Präzision der Regression der psychologischen (Y) auf die physikalischen Meßwerte (X). Die aber wird nach (3) vom Standard-Schätzfehler und der wiederum vom Gültigkeitskoeffizienten bestimmt. Nach Lienert ([24] S.310) sollten Gültigkeitskoeffizienten bei Tests, die zur individuellen Begutachtung von Personen benutzt werden, nicht kleiner als 0.7 sein. Gültigkeitskoeffizienten in dieser Höhe findet man z.B. bei Intelligenztests, wenn als Kriterium der Schulerfolg herangezogen wird. Die Korrelationskoeffizienten zwischen physikalischen Lärm-Maßen und (individuenbezogenen!) subjektiven Belästigungsmaßen erreichen diese Größe jedoch praktisch nie. In Tabelle 3 sind in Spalte 4 solche Gültigkeitskoeffizienten, soweit sie in der Literatur zu finden sind, zusammengetragen. Sie variieren zwischen 0.25 und 0.68 mit einem Mittelwert von 0,474. Häufig werden jedoch in den einschlägigen Arbeiten nur die gruppenbezogenen Korrelationskoeffizienten zwischen physikalischen und psychologischen Lärmmaßen angegeben. Diese sind erwartungsgemäß in der Regel recht, groß gelegentlich sogar fast gleich eins, lassen aber eine Aussage über die Gültigkeit aus weiter oben schon genannten Gründen nicht zu.

Tabelle 3: Individuumsbezogene Validitätskoeffizienten aus verschiedenen Untersuchungen.
*) : Zitiert nach DFG [7].

1	2	3	4
Autor/ Studie/ Jahr	Lärmart	Physikal. Lärmmaß	Validität (Individuell)
McKinnell (1963) [25] *)	Fluglärm (Heathrow I)	L_{PN} exc. by 10% by 50%	0.37 0.38
MIL (1971) [26]	Fluglärm (Heathrow II)	"NNI" (k=12) "NNI" (k= 4)	0.432 (w) 0.411 (t)
DFG (1974) [7]	Fluglärm (München-Riem)	FB1	0.58
Grandjean et al. (1973), [10] *)	Fluglärm (Zürich, Genf, Basel)	NNI	0.53 0.68 0.53
TRACOR 7 cities Patterson et al. (1973), [29] *)	Fluglärm (Dallas, Los Angeles, Chicago, Denver, Miami, New York Chattanooga)	CNR	0.49
TRACOR 2 cities Connor et al. (1972), [5] *)	Fluglärm (Boston, Reno)	CNR	0.25
Bullen & Hede (1986) [3]	Fluglärm (Sydney, Adelaide Perth, Melbourne Richmond)	NEF	0.36
Hazard (1971) [13]	Fluglärm (USA)	CNR	0.35
Jonckheere (1989) [15]	Fluglärm (Brüssel)	FB1	0.474
Leonard & Borsky (1973), [23] *)	Fluglärm (New York)	CNR	0.38
Labiale (1983) [21]	LKW (Labor)	Leq3	0.274

Aus der geringen Größe des Gültigkeitskoeffizienten folgt, daß die *individuelle* Belästigung über physikalische Maße der Lärmmenge nicht hinreichend genau ermittelt werden kann. Ein Rechenbeispiel kann diese Aussage näher erläutern: Auch unter günstigen Umständen ($r_{XY}=0.6$) erreicht man nach (2.3) nur eine Reduktion des Verhältnisses von Standardschätzfehler $\chi\sigma_Y(E)$ zur Standardabweichung σ_Y im Merkmal Y um den Faktor 0.8. In einer Untersuchung über Fluglärmwirkungen betrug z.B. die Standardabweichung σ_Y der erhobenen Belästigungsdaten ca 1/4 der maximal möglichen Skalendifferenz, übertragen auf eine 10-stufige Skala also 2.5 Skalenteile. Bei der Vorhersage der Belästigung Y mit Hilfe der physikalischen Lärmexposition X wird damit der Standard-Schätzfehler $2.5 \times 0.8 = 2.0$ Skalenteile. Das 95%-Vertrauensintervall erstreckt sich also immer noch über $\pm 1.96 \cdot 2.0 \approx \pm 4$ Skalenteile, also immer noch fast über die gesamte Skala.

3.5.1 Gruppenbezogene Gültigkeit

Bildet man Gruppen von physikalisch gleich hoch belasteten Personen, so lassen sich die Gruppenmittelwerte hinsichtlich der Belästigung aus den jeweiligen physikalischen Maßen der Lärmmenge erheblich genauer, dh. mit erheblich geringerem Standardschätzfehler, vorhersagen (s. (5)). Sofern also nur Aussagen über Gruppenmittelwerte erforderlich sind, reichen u.U. schon Gültigkeitskoeffizienten aus, die um 0.3 liegen ([24] S. 312). Bei Gültigkeitskoeffizienten, die um 0.5 liegen, sind daher der L_{eq3} und seine Derivate zur Beurteilung der durchschnittlichen Belästigung größerer Gruppen von gleich stark physikalisch belasteten Personen durchaus geeignet. An dieser Stelle sei darauf hingewiesen, daß man sich bei der Feststellung der Wirkung von Arzneimitteln oder der Toxizität von Substanzen mit erheblich geringeren Gültigkeitskoeffizienten begnügt. In diesem Bereich wird häufig der Wirkungsnachweis mit dem "BESD" (Binomial Effect Size Display) geführt, welches sich leicht in einen Korrelationskoeffizienten umrechnen läßt. Rosenthal [33] nennt als Beispiele $r=0.04$ für die Unterbindung von Herzattacken durch Aspirin, oder $r=0.23$ für die Wirkung von AZT bei der Behandlung von AIDS. In beiden Fällen wurde das betreffende

Kontrollgruppenexperiment zur Wirkungsprüfung aus ethischen Gründen abgebrochen, nachdem Zwischenauswertungen die erwähnten Zusammenhänge offengelegt hatten.

4 Diskussion

Die Ausführungen zeigen, daß die psychologische Testtheorie auf die Lärmmeßtechnik angewendet werden kann. Physikalische Lärmexpositionsmaße, wie der energie-äquivalente Dauerschallpegel L_{eq3} , die energetische Pegelsumme L_s , die äquivalenten Dauerschallpegel L_{eq1} und L_{eq4} , der Noise and Number Index NNI oder das Fluglärmbewertungsmaß FB1 entsprechen dabei verschiedenen 'Testvariablen', die auf Skalen erfaßten subjektiven Reaktionen der 'Kriteriumsvariablen'. Die Aufgabe der Lärmmeßtechnik sollte dann darin bestehen, aus der Lärmexposition die subjektive Reaktion vorherzusagen.

Anhand von Daten aus der DFG-Fluglärmuntersuchung [7] bei der sowohl verschiedene Lärmexpositionsmaße wie auch subjektive Reaktionen bei derselben Stichprobe erfaßt worden waren, konnte statistisch 'herausgefiltert' werden, daß z.B. die o.g. sechs Expositionsmaße als äquivalent (also als Paralleltests) anzusehen sind, diese Eigenschaft aber nicht mit Maßen wie D_{10} (mittlere Überflugdauer, berechnet aus der 10dB-Down-Time der Einzelereignisse), H_{81} (Anzahl der Überflüge mit Pegelwerten über 80 dB(A)), L_m (aus Einzelpegeln berechneter Mittelungspegel), L_{eq10} (ähnlich wie L_{eq4} , nur eine noch stärkere Gewichtung der Anzahl der Ereignisse als beim FB1) oder $10 \cdot \log N$ (N =Anzahl der Überflüge) teilen.

Die Paralleltest-Reliabilität der o.g. sechs Expositionsmaße, definiert durch ihre Interkorrelationskoeffizienten, ist nahezu gleich eins und bedeutet einen Meßfehler von fast null, wenn man ein Verfahren gegen ein anderes austauscht. Die genannten physikalischen Meßverfahren sind daher im wesentlichen ununterscheidbar. Das heißt, daß die unter Verwendung dieser Verfahren gewonnenen Meßwerte untereinander praktisch in einem linearen Zusammenhang stehen und unter Verwendung einer Formel der Art $x' = ax + b$ ohne signifikanten Informationsverlust ineinander umgerechnet werden können.

Die Validität, definiert durch die Korrelationskoeffizienten mit der Bevölkerungsreaktion, ist hingegen moderat, sie liegt nur zwischen 0.5 und 0.6. Diese reicht nach üblicher Ansicht für individuelle Vorhersagen nicht aus, weil der Schätzfehler zu groß ist. Die hohe Präzision der physikalischen Lärmengemessung überträgt sich also keinesfalls auch auf die Erfassung der Belästigungsreaktion durch die physikalische Messung. Mit anderen Worten: Die physikalische Lärmengemessung ist ein relativ grobes Maß für die Belästigungsreaktion. Zur Vorhersage von Gruppenmittelwerten reichen Validitätskoeffizienten in der oben mitgeteilten Höhe aber völlig aus, in vielen Anwendungsfällen begnügt man sich mit noch erheblich kleineren Validitäten.

Dennoch sollten Anstrengungen unternommen werden, die Validität der physikalischen Lärmengemessung zu erhöhen. Der Versuch aber, dies allein über eine weitere Verfeinerung der physikalischen Meßverfahren zu erreichen, ist nicht erfolgversprechend. Denn es ist vorhersehbar, daß ein neues Verfahren, wenn es effektiv ist, mit bereits eingeführten Verfahren - z.B. dem L_{eq3} - wiederum zu fast eins korreliert. Nach der (testtheoretischen) Verdünnungsformel (4) kann daher die Validität durch Einführung eines solchen Verfahrens nicht weiter verbessert werden. Erfolgversprechender dürfte es hingegen sein, an der Verbesserung des Kriteriums anzusetzen. Dies kann man in einfacher Weise wahrscheinlich schon durch die Bildung von (gewichteten) Summenscores aus mehreren Items erreichen, wie noch in der DFG-Untersuchung 1974 [7] geschehen. Allerdings wurde diese Anregung dann kaum mehr aufgegriffen, möglicherweise eine Folge der Bemerkung von Schultz ([35] S. 378), "that a person's degree of annoyance can be more simply and more reliably determined from his response to a direct question, asking how he is annoyed by the noise under investigation". Zurzeit jedenfalls wird auf die Erfassung der Bevölkerungsreaktion oft nur noch vergleichsweise wenig Sorgfalt gelegt. Oft werden den Probanden, ganz im Stil von Meinungsumfragen, einige wenige für einschlägig gehaltene Skalen oder Fragen vorgelegt, eingebettet in eine mehr oder weniger große Zahl anderer Fragen, die möglicherweise ebenfalls einen Bezug zum Belästigungserlebnis haben, ohne daß aber Methoden, die bei der Konstruktion psychologischer Tests normalerweise 'Stand der Technik' sind, zur Anwendung kämen, wie z.B. die Berechnung von Trennschärfe und Schwierigkeit der Items und eine darauf gründende Itemselektion. Diese Vorgehensweise verschwand offenbar ab Mitte der siebziger Jahre wieder aus dem Repertoire der Lärmforschung.

Literatur

1. BGH Urteil vom 25.3.1993 - III ZR 60/91 - NJW 93 (1993) 1700
2. Bürck, W.; Grützmaker, M.; Meister, F. J.; Müller, E. A.; Matschat, K.: Fluglärm. Seine Messung und Bewertung, seine Berücksichtigung bei der Siedlungsplanung, Maßnahmen zu seiner Minderung. Gutachten, erstattet im Auftrage des Bundesministers für Gesundheit. Göttingen 1965.
3. Bullen, R. B.; Hede, A. J.: Comparison of the effectiveness of measures of aircraft noise exposure by using social survey data. *J Sound Vibration*, 108 (1986) 227-245
4. Bronstein, I. N.; Semendjajew, K. A.: Taschenbuch der Mathematik (23. Auflage). Thun und Frankfurt: Harri Deutsch 1979
5. Connor, W. K.; Patterson, H. P.: Community reactions to aircraft noise around smaller city airports. NASA Report No CR-2104. Washington DC: National Aeronautics and Space Administration 1972
6. De Jong, R. G.: Review of research developments in community response to noise. In: B. Berglund, T. Lindvall (Eds): *Noise '88. Proceedings of the 5th International Congress on Noise as a Public Health Problem*, Vol. 5 (Part 2). 1988, 99-113.
7. DFG Forschungsbericht Fluglärmwirkungen I-III. Bonn-Bad Godesberg: Deutsche Forschungsgemeinschaft 1974
8. Fischer, G.: Einführung in die Theorie psychologischer Tests. Bern usw.: Huber 1974
9. Geigy: *Documenta Geigy. Wissenschaftliche Tabellen*. 6. Auflage. Basel: J.R. Geigy A.G. 1960
10. Grandjean, E.; Graf, P.; Lauber, A.; Meier, H. P.; Müller, R.: A survey of aircraft noise in Switzerland. *Proceedings of the International Congress on noise as a Public Health Problem: Dubrovnik 1973*, pp.645-659
11. Gulliksen, H.: *Theory of mental tests*. New York 1950
12. Guski, R.: *Lärm. Wirkung unerwünschter Geräusche*. Bern usw.: Hans Huber 1987
13. Hazard, W. R.: Predictions of noise disturbance near large airports. *Journal Sound Vibration* 15 (1971) 425-445.
14. Jonkheere, R. E.: Noise exposure, annoyance, pollution and defensive behavior correlations in relation with aircraft operations. In: B. Berglund, U. Berglund, J. Karlsson, T. Lindvall (Eds): *Noise '88. Proceedings of the 5th International Congress on Noise as a Public Health Problem*, Vol.3: Performance, behavior, animal, combined agents and community responses. Stockholm: Swedish Council for Building Research 1988, 327-332.
15. Jonkheere, R. E.: Evaluation of noise exposure and annoyance around Brussel's airport. *Noise Control Engineering Journal* 32 (1989) 93-98
16. Kalveram, K. Th.: Über Faktorenanalyse. Kritik eines theoretischen Konzepts und seine mathematische Neuformulierung. *Archiv für Psychologie* 122 (1970) 92-118
17. Kalveram, K. Th.: Zur Evolution des Belästigungserlebnisses. Ökopsychologische und verhaltensbiologische Betrachtungen über die Wirkung von Lärm. *Psychologische Beiträge* 38 (1997), 315-230
19. Kryter, K. D.: Scaling human reactions to the sound from aircraft. *J. Acoust. Soc. Am.* 31 (1959) 1415-1429
20. Kryter, K. D.: A note on the quantity (effective) perceived noisiness and limits of perceived noise levels. *J. Sound Vibration* 25 (1972) 383-393
21. Labiale, G.: Laboratory study of the influence of noise level and vehicle number on annoyance. *J. Sound Vibration*, 90 (3) (1983) 361-371
22. Langdon, F.; Griffiths, I.: Subjective effects of traffic noise exposure II: Comparisons of noise indices; response scales, and the effects of change in noise levels. *Journal of Sound and Vibration*, 83(2) (1982) 171-180
23. Leonard, S.; Borsky, P. N.: A causal model for relating noise exposure, psycho-social variables and aircraft noise annoyance. *Proceedings of the International Congress on Noise as a Public Health Problem*. Dubrovnik 1973, pp. 13-18
24. Lienert, G. A.: *Testaufbau und Testanalyse*. Weinheim usw.: Beltz 1969
25. McKennell, A. C.: *Aircraft noise annoyance around London (Heathrow) Airport*. London: Central Office of Information 1993 (Heathrow I)
26. MIL (Market Investigations Ltd.): *Second survey of aircraft noise annoyance around London (Heathrow) airport*. London: Her Majesty's Stationery Office 1971 (Heathrow II)
27. Ohrström, E.; Björkman, M.; Rylander, R.: Laboratory annoyance and different traffic noise sources. *J. Sound Vibration* 70 (1980) 333-341.
28. Ollerhead, J. B.: DORA Report 9120. The CAA aircraft noise contour model: ANCON version 1. London: Civil Aviation Authority/Dep. of Transport 1992

29. Patterson, H. P.; Connor, W. K.: Community responses to aircraft noise in large and small cities in the USA. Proceedings of the International Congress on noise as a Public Health Problem. Dubrovnik 1973, pp. 707-818
30. Rasmussen, K. B.: Annoyance from simulated road traffic noise. Journal of Sound and Vibration, 65(2) (1979) 203-214
31. Rice, C. G.: Trade-off effects of aircraft noise and number of events. In: Proceedings of the Third International Congress on Noise as a Public Health Problem. American Speech-Language-Hearing Association (ASHA) Report 10 (1980) 495-510
32. Rice, C. G.: Development of cumulative noise measure for the prediction of general annoyance in an average population. Journal of Sound and Vibration, 52(3) (1977) 345-364
33. Rosenthal, R.: How are we doing in soft psychology. American Psychologist, June 1990, 775-776
34. Schick, A.: Schallbewertung. Berlin usw.: Springer 1990
35. Schultz, T. J.: Synthesis of social surveys on noise annoyance. J. Acoust. Soc. Am. 64 (1978) 377-405
36. Vallet, M.; Pachiaudi, G.; Depitre, A.; Tanguy, Y.; Francois, J.: Community reactions to aircraft and residual noise. In: Berglund B; Lindvall T (Eds): Noise '88. Proceedings of the 5th International Congress on Noise as a Public Health Problem, Vol.3. Stockholm: Swedish Council of Building Research 1988, 289-294
37. Wottawa, H.: Psychologische Methodenlehre. München: Juventa 1977