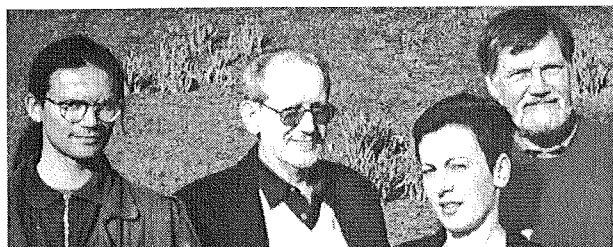


Internationale zur Erhebung von bereitende empirische



Dipl.-Psych. Ute Felscher-Suhr (3. v.l.)

Jahrgang 1966, Psychologiestudium in Bochum, seit 1994 am Fachbereich Kognitions- und Umweltpsychologie beschäftigt. Koordination des „Arbeitskreises Ökologische Lärmforschung“, der u.a. Studien zur Wirkung von Lärm auf den Alltag von Betroffenen sowie zur Entwicklung von Antwortskalen zur Erhebung von Lärmbelastigung durchführt.

Prof. Dr. Rainer Guski (4. v.l.)

Jahrgang 1941, Studium der Psychologie und Soziologie in Berlin und Hamburg, 1979 Professor für Sozialpsychologie an der TU Berlin, 1983 Professor für Psychologie in Bochum, leitet die Arbeitsgruppe Kognitions- und Umweltpsychologie der Fakultät für Psychologie, zahlreiche empirische Feld-Untersuchungen zu Auswirkungen des Lärms auf die Wohnbevölkerung.

Dr. Rudolf Schuemer (2. v.l.)

Jahrgang 1941, Studium der Psychologie in Hamburg bis 1968, Promotion an der Uni Mannheim 1971, dort wissenschaftlicher Angestellter von 1968 bis 1975, seit 1976 an der FernUniversität in Hagen mit den Arbeitsschwerpunkten Evaluation, Post Occupancy Evaluation und Lärmwirkungsforschung.

Jörg Schulte-Pelkum (1. v.l.)

Jahrgang 1973, seit 1993 Psychologiestudium in Bochum, seit 1997 studentische Hilfskraft im Arbeitskreis Ökologische Lärmforschung. Dort hauptsächlich mit statistischer Datenanalyse beschäftigt.

Zusammenfassung

Die Frage nach der Lästigkeit (annoyance) von Umwelt-Geräuschen ist häufig Bestandteil von Befragungsuntersuchungen. International lassen sich die Ergebnisse jedoch zum Teil schlecht vergleichen, weil sich die Formulierung der Frage, ihre Position innerhalb des Fragebogens sowie das Antwortformat von Land zu Land unterscheiden. Im Rahmen von Standardisierungsbestrebungen mit internationaler Reichweite (Fields, 1996) konnte über die Frageformulierung und die Frageposition unter den beteiligten Experten und Expertinnen weitgehend Konsens hergestellt werden. Jedoch entstand die Notwendigkeit, eine verbale Antwortskala zu entwickeln, die parametrische Auswertungen zulässt, sowohl in telefonischen als auch in persönlichen Befragungen einsetzbar ist und möglichst in verschiedenen Sprachen und Kulturkreisen gleich verstanden wird.

In Kooperation mit LärmexpertInnen aus zehn Ländern wurde eine Skalierungsuntersuchung über Adverbien durchgeführt, die unterschiedliche Intensitäten der Lärmbelastigung ausdrücken (Felscher-Suhr, Guski & Schuemer, 1998). Ziel der Untersuchung war es, den ExpertInnen der beteiligten Länder nach der Datenanalyse Skalen vorzuschlagen, die dann innerhalb der jeweiligen Länder auf linguistische Qualität und Tauglichkeit geprüft werden sollen, bevor sie als Skala für das jeweilige

Standardisierungsbestrebungen Lärmbelastigung – eine vor- Untersuchung in zehn Ländern

Land in internationale Empfehlungen aufgenommen wird. Unter Berücksichtigung der Ergebnisse aller beteiligten Länder hat sich gezeigt, daß sich überwiegend eine fünfstufige, annähernd äquidistante Antwortskala mit verbalen Markierungen besser als eine vierstufige realisieren läßt.

Abstract

International efforts towards a shared noise annoyance question – an empirical study with ten participating nations

One of the most frequently measured responses in community noise surveys is annoyance. The results of these studies are difficult to compare because of different wording, different positions of the questions in the questionnaire and different scales used for the answers. In international efforts (Fields, 1996) consensus was reached concerning wording and position of the question. But before experts reach consensus about the best scale for this purpose, a study should clarify several quality aspects which the scale has to comply with: It should be usable in telephone studies as well as in personal interviews, it should be applicable in different nations, and its metric qualities should allow for parametric statistical operations.

In cooperation with noise research experts from ten nations a scaling study about adverbs that describe different degrees of an-

noyance was performed (Felscher-Suhr, Guski & Schuemer, 1998). The aim was to propose a scale for each nation to the experts, who should then test the linguistic quality of the proposal in the country it will be applied to. Finally, international committees should recommend the use of the scale in future surveys. Taking the results from all participating nations into account, a five point scale with equal appearing distances and verbal qualifiers is easier to construct than a four point scale.

1 Hintergrund

Die Erforschung der Auswirkungen von Umweltgeräuschen auf den Menschen ist innerhalb der Umweltpsychologie ein etabliertes Forschungsfeld. Die so häufig geforderte Interdisziplinarität wird hier seit langer Zeit erfolgreich praktiziert. In Untersuchungen zur Lärmwirkung arbeiten PsychologInnen gemeinsam mit AkustikerInnen und MedizinerInnen im Feld oder im Labor. Im Mittelpunkt der Betrachtung stehen häufig lärm betroffene Personen, die anhand von Fragebögen zu ihrer Lebenssituation, ihrem Wohnumfeld und ihrer Zufriedenheit mit der Wohngegend im allgemeinen sowie der Belastung und Belästigung durch Lärm im besonderen befragt werden. Diese Befragungen finden vorwiegend durch persönlichen Kontakt mit einem Interviewer/einer Interviewerin oder am Telefon statt. Die Antworten der Befragten werden anhand

sogenannter Antwortskalen kodiert und später statistischen Prozeduren unterzogen.

Die Entscheidung für eine bestimmte Skala orientiert sich zum Teil an bestimmten methodischen Kriterien – u.a. hinsichtlich des Skalenniveaus, um bestimmte Transformationen und Berechnungen vornehmen zu können –, zum Teil aber auch an individuellen Präferenzen der betreffenden ForscherInnen. Zur Erhebung von Lärmbelastigung werden Skalen mit unterschiedlicher Anzahl von Skalenstufen sowie unterschiedlichen verbalen Markierungen eingesetzt. Es finden sich z.B. 3-stufige verbale Skalen in einer dänischen Untersuchung von Andersen, Kühl und Relster (1983), 4-stufige in englischen Studien (vgl. z.B. Fields & Walker 1982), 5-stufige in Deutschland bei Finke, Guski und Rohrmann (1980) oder in Deutsche

Insgesamt bleibt festzuhalten, daß es für die Feststellung des Ausmaßes der Belästigung durch Lärm keine einheitlichen Erhebungsinstrumente gibt.

Forschungsgemeinschaft (1974), 5-stufige in den USA bei Fidell und Tefte-teller (1981), 7-stufige Antwortformate in Deutschland bei Steinheider und Winneke (1993) und sogar 9-stufige Skalen in den USA bei Hall, Birnie, Taylor und Palmer (1981). Dabei wird die Belästigung teils direkt erfragt („Wie stark fühlen Sie sich hier bei Ihnen zuhause durch Lärm belästigt?“), teils indirekt, zum Beispiel aus der Anzahl genannter gestörter Aktivitäten, abgeleitet; in einigen Untersuchungen werden auch beide Erfassungsansätze realisiert (vgl. z.B. Möhler,

Schuemer, Knall & Schuemer-Kohrs, 1986). Im Falle der direkten Erhebung kommt neben den oben erwähnten verbalen Skalen zuweilen die sogenannte Thermometerskala zum Einsatz. Sie verläuft senkrecht, 10- oder 11-stufig, wobei am unteren Ende meist eine Verdickung eingezeichnet ist, die den Nullpunkt kennzeichnet. Das obere Ende stellt die maximale Belästigung dar (an den Endpunkten der Skala sind verbale Bezeichnungen angebracht, z.B. „0 = überhaupt nicht“, „10 = äußerst“). Die befragte Person soll ihre Belästigung innerhalb dieses „Thermometers“ angeben bzw. einzeichnen. Im Fall der indirekten Erfassung, also bei der Mitteilung von Einzelstörungen zur Gesamtgestörtheit, werden zumindest in Deutschland oft 5-stufige verbale Antwortformate benutzt, deren Abstände auf Äquidistanz überprüft wurden (Rohrmann, 1978).

Insgesamt bleibt festzuhalten, daß es für die Feststellung des Ausmaßes der Belästigung durch Lärm keine einheitlichen Erhebungsinstrumente gibt. Diese Situation ist aus mehreren Gründen unbefriedigend: (1) Direkte Vergleiche zwischen Untersuchungen aus verschiedenen Regionen oder Zeitabschnitten sind nicht möglich; (2) beim indirekten Vergleich der Resultate, die mit unterschiedlichen Instrumenten erhalten wurden, müssen zusätzliche (oft nicht prüfbare) Annahmen gemacht werden; (3) da die Mittel zur Finanzierung neuer Forschungsvorhaben immer stärker auch unter dem Gesichtspunkt des möglichen Ertrages für die Gesellschaft vergeben werden, müßte begründet werden, warum eine bestimmte

Fragestellung, die beispielsweise bereits in Skandinavien bearbeitet wurde, in einem anderen Land mit vergleichbarer Kultur erneut untersucht werden sollte.

Wenn also nationale und internationale Vergleiche durchgeführt werden sollen, sollten einige Standards für die Messung eingehalten werden, wie dies im Bereich der Akustik übrigens schon lange üblich ist. Fields (1996) schlägt für das ICBEN Team 6¹ vor, weltweit in Feldstudien zur Lärmbelastigung eine Frage zur allgemeinen Lärmbelastigung aufzunehmen, an derselben Stelle im Fragebogen zu platzieren und mit derselben Skala beantworten zu lassen. Während über die Formulierung der Frage („Thinking about the last 12 month or so, when you are at home, how much would you say the noise from (...noise source...) bothers or annoys you...?“) sowie die Positionierung im Fragebogen schnell Einigung erzielt werden konnte, konnten sich die eingeladenen ExpertInnen² auf einem After Congress Workshop anlässlich der Inter-Noise in Budapest im Sommer 1997 nicht auf ein gemeinsames Antwortformat verständigen. Unter Berücksichtigung bisheriger Forschungsge-wohnheiten sowie Argumenten für die Einsetzbarkeit der Skala in persönlichen und telefonischen Interviews wurde vorgeschlagen, eine vier- oder fünfstufige Skala zu entwickeln. Um diese Entscheidung zu erleichtern, schlug das AutorInnenteam die Durchführung einer – internationalen – Untersuchung zur Entwicklung gleichabständiger Skalen für die ver-

schiedenen Sprachen und zur Klärung der Frage nach einer 4- oder einer 5-stufigen Skala vor. Es erklärte sich bereit, ein Untersuchungsdesign zu entwickeln, Material zu erstellen und via Internet zu versenden sowie die Koordination der Untersuchung und die Auswertung der Daten zu übernehmen (Felscher-Suhr, Guski & Schuemer, 1998, Guski, Felscher-Suhr & Schuemer, 1998).

2 Durchführung der Studie

2.1 Auswahl der zu skalierenden Begriffe und Festlegung der Beurteilungsaufgabe

Zunächst wurden ausgewählte ExpertInnen³ in den teilnehmenden Ländern gebeten, bis zu 21 Adverbien in der jeweiligen Landessprache zu bestimmen, die in unterschiedlichem Ausmaß die Intensität der Belästigung ausdrücken. Dann sollten in jedem Land 2 x 30 Untersuchungspersonen fünf verschiedene Aufgaben bearbeiten:

1. Alle Adverbien sollten jeweils in eine von 9 Kategorien eingeordnet werden, die eine zunehmende Belästigung durch Lärm repräsentieren sollen.
2. Es sollte unter allen Adverbien dasjenige herausgesucht werden, das die höchste Stufe der Belästigung beschreibt.
3. Unter den verbleibenden Adverbien sollten diejenigen drei herausgesucht werden, die im Kontext mit einem vorgegebenen „niedrigsten“ Adverb (im

² Als ExpertInnen werden hier regelmäßige TeilnehmerInnen an lärmrelevanten internationalen Kongressen sowie die Mitglieder des oben erläuterten ICBEN Team 6 verstanden.

³ Die Ermittlung der ExpertInnen pro Land erfolgte über die LandesvertreterInnen des ICBEN Teams 6. Ein(e) ExpertIn ist in diesem Fall eine Person, die im jeweiligen Land Untersuchungen zur Lärmwirkung durchführt.

¹ ICBEN = International Commission on the Biological Effects of Noise, die Arbeitsgruppe Team 6 trägt den Titel „Community response to noise“ und beschäftigt sich in den letzten Jahren mit den Möglichkeiten der Standardisierung in der Lärmwirkungsforschung.

- Deutschen: „überhaupt nicht“; im Englischen: „not at all“) und dem selbstbestimmten „höchsten“ Adverb eine äquidistant erscheinende 5-Stufen-Skala der Belästigung ergeben.
4. Unter den nach (2) verbleibenden Adverbien sollten diejenigen zwei herausgesucht werden, die im Kontext mit dem vorgegebenen niedrigsten und dem selbstbestimmten höchsten Adverb eine äquidistant erscheinende 4-Stufen-Skala der Belästigung ergeben.
 5. Alle Adverbien sollten auf einer grafischen Skala im Belästigungskontext einzeln hinsichtlich ihrer Aussageintensität skaliert werden.

In der Auswahl der Aufgaben sollten zwei Aspekte Berücksichtigung finden. Zum einen sollte eine weitgehende Vergleichbarkeit mit bereits existierenden Studien

Zur Auswahl der für die Skalenkonstruktion am besten geeigneten Adverbien wurden verschiedene Kriterien bestimmt...

zur Entwicklung von Lärmbelastigungsskalen gewährleistet sein, zum anderen sollte die Bearbeitungsdauer der Aufgaben möglichst gering bleiben. Unser Vorgehen orientiert sich an Rohrmann (1978) sowie an Levine (1981); in diesen beiden Studien wurde eine ähnliche Aufgabe wie die oben beschriebene Aufgabe 1 verwendet, die als Variante einer Thurstone-Skalierung (vgl. Thurstone, 1931; Thurstone & Chave, 1929; s. auch Sixtl 1967, S. 145ff) interpretiert werden kann. Für den deutschen Sprachraum wurden folgende Adverbien für die Untersuchung ausgewählt:

überhaupt nicht, nicht, kaum, wenig, etwas, ein wenig, teilweise, einigermaßen, mittelmäßig, ziemlich, besonders, sehr, beträchtlich, erheblich, stark, äußerst, völlig, total, enorm, ungemein, außerordentlich.

Bei der Auswahl der Adverbien wurde angestrebt, den Bereich von „keine/geringste Ausprägung der Belästigung/Störung“ bis „höchste Ausprägung der Belästigung/Störung“ gleichmäßig abzudecken und dabei auch Ausdrücke aus dem aktuellen Sprachjargon zu berücksichtigen. Zur Auswahl der für die Skalenkonstruktion am besten geeigneten Adverbien wurden verschiedene Kriterien bestimmt:

- Häufigkeit der Auswahl bei der Skalenkonstruktion durch die Untersuchungspersonen,
- Distanz zum gewünschten Mittelwert (für die 5-er Skala sollten also Begriffe ausgewählt werden, deren Skalenwerte auf der 10cm-Skala den Werten 2.5, 5.0, 7.5 und 10.0 möglichst nahe sind, für die 4er Skala den Werten 3.33, 6.67 und 10.0),
- Hohe Übereinstimmung der Skalenwerte: geringe Standardabweichung, hohe Übereinstimmung zwischen den einzelnen BeurteilerInnen,
- Hohe Übereinstimmung zwischen den beiden Skalen (grafische Skala und Kategorienskala),
- Hohe Übereinstimmung zwischen den beiden Untersuchungsgebieten.

Weiterhin sollten Kriterien berücksichtigt werden, die unabhängig von den metrischen Eigenschaften der Adverbien sind:

- Expertenurteil über die Eignung der konstruierten Skala,
- „Linguistische Qualität“ der Skala (Kombinationsfähigkeit der ausgewählten Adverbien mit „gestört/belästigt“, „Zueinanderpassen“ der Adverbien der einzelnen Skalenstufen),

- Tauglichkeit der Skala für die nächste Generation befragter Personen (Zukunftsfähigkeit).

2.2 Erhebungsorte und Untersuchungspersonen

Es beteiligten sich folgende zehn Nationen an unserer Untersuchung: Australien, Deutschland, England, Frankreich, Japan, Norwegen, Spanien, Türkei, Ungarn und die USA⁴. Zur Verminderung ortsspezifischer Effekte sollte die Untersuchung pro Land bzw. Sprache an jeweils zwei getrennten Orten durchgeführt werden (Kriterium 1), und pro Ort sollten mindestens 30 Personen untersucht werden (Kriterium 2). Als Untersuchungspersonen sollten BeurteilerInnen im Alter bis zu 40 Jahren ausgewählt werden. Da die zu konstruierende Skala vermutlich über eine längere Zeit verwendet werden soll, schien es geboten, für die Beurteilungsaufgabe vorwiegend jüngere Personen auszuwählen. Wegen des großen Zeitdrucks fungierten an den meisten Erhebungsorten Studierende als Untersuchungspersonen.

Die Hälfte der Länder (Australien, Deutschland, Japan, Spanien und die USA) erfüllten beide o.g. Kriterien, bei den restlichen Ländern wurde die Erhebung nur in einem Ort durchgeführt oder wurden pro Ort weniger als 30 Untersuchungspersonen untersucht. Für Deutschland wurden die Erhebungen in Bochum und in Oldenburg durchgeführt.

⁴Die Untersuchung wurde in den beteiligten Ländern unterstützt von: M. Arana, M. Burgess, P. Champelovier, L. Ensel, H. Ferenc, J.M. Fields, I. Flindell, A. Garcia, A.M. Garcia, T. Gjestland, H. Hoegel, S. Job, Z. Karabiber, S. Kurra, S. Kuwano, A. Montag, T. Sato, A. Schick, B. Schulte-Fortkamp, K. Shepherd, M. Vallet, A. Vela und T. Yano, denen wir an dieser Stelle noch einmal herzlich danken.

2.3 Auswertung

Die Auswertung hatte zwei Hauptziele:

- Bestimmung von Maßen für die Konsistenzen der Urteile sowie
- Bestimmung der statistischen Kennwerte für die Auswahl der Adverbien in Hinblick auf die zu konstruierenden Skalen.

2.3.1 Konsistenz

Die Konsistenz der Beurteilung kann mit unterschiedlichen Methoden überprüft werden. Pro Ort und Aufgabe (Kategorienskalisierung oder Skalierung auf der grafischen Skala) kann eine zwei-faktorielle Varianzanalyse (Faktor 1: die 21 Adverbien, Faktor 2: BeurteilerInnen) durchgeführt und daraus Interrater-Konsistenzmaße nach Ebel (1951) bestimmt werden. Die Konsistenz sinkt in dem Maß, in dem eine Rater x Adverbien-Wechselwirkung vorliegt (also z.B. dadurch, daß die Relationen zwischen den Skalenwerten der Adverbien bei den verschiedenen RaterInnen unterschiedlich sind). Ebel schlägt sowohl Maße für die Konsistenz der durchschnittlichen Urteile als auch für die durchschnittliche Konsistenz der EinzelraterInnen vor.

Bei Ländern mit Erhebungen in zwei Orten können zudem pro Land weitere Analysen (nach dem ALM) durchgeführt werden, so u.a. eine vier-faktorielle Analyse: Faktor 1: Ort; Faktor 2: Upn („nested“ unter Ort); Faktor 3: Aufgabe (Kategorien - vs. grafische Skala)⁵ und Faktor 4: Adverbien, wobei Meßwiederholung über die

Faktoren 3 und 4 vorliegt. In Ländern, in denen nur an einem Ort erhoben wurde, wurden entsprechende drei-faktorielle Analysen – ohne den Faktor 1 „Ort“ – durchgeführt. Mit diesen Analysen kann u.a. untersucht werden, ob sich die Beurteilungen in den beiden Erhebungsorten unterscheiden (Haupteffekt Faktor 1) oder ob es Wechselwirkungen zwischen den zu beurteilenden Adverbien und den Skalen oder den Orten gibt. Eine signifikante Ort x Adverb- bzw. Skala x Adverb-Wechselwirkung würde beispielsweise darauf hinweisen, daß die Relationen zwischen den Skalenwerten der Adverbien in den beiden Erhebungsorten bzw. bei den beiden Skalen (Kategorien- und grafische Skala) unterschiedlich sind. Zudem können pro Ort die Skalenwerte für die Kategorien- und die grafische Skala über die 21 Adverbien hinweg korreliert werden (Kongruenz der Skalen).

2.3.2 Bestimmung von Kennwerten für die Adverbienauswahl

In Hinblick auf die Auswahl von Adverbien für die zu konstruierenden Skalen können u.a. folgende Kennwerte bestimmt werden: Mittelwerte und Mediane sowie Streuungen und Interquartilabstände pro Adverb auf jeder der beiden Skalen; Häufigkeit der Wahl eines Adverbs für eine der Skalenpositionen für die 5- bzw. 4-stufige Skala (vgl. Aufgaben 3. und 4. oben). Weitere Kennwerte und Kriterien sind in Fields et al. (1998) beschrieben.

⁵Für diese Analysen wurden die Skalenwerte der 9-stufigen Kategorien so transformiert, daß sie denselben Range wie die grafische Skala (von „0“ bis „10“) aufweisen.

⁶Die vollständigen Analyseergebnisse können bei der Erstautorin eingesehen werden.

3 Ergebnisse

Die Ergebnisse können hier nur ausschnittsweise wiedergegeben werden⁶.

3.1 Konsistenz

Pro Land, Ort und Skala wurde zunächst die Interrater-Konsistenz nach Ebel (1951) bestimmt. Die Koeffizienten für die interne Konsistenz der Durchschnittsurteile liegt zwischen $r = .992$ und $r = .999$ bei der Kategorienskala und zwischen $r = .993$ und $r = .999$ bei der grafischen Skala. Auch die Koeffizienten für die durchschnittlichen Einzelrater liegen – mit einer Ausnahme – alle über .85 (und vielfach über .9); die Ausnahme ist ein Koeffizient von nur .49 bei der Kategorienskala in Canberra. Die Inspektion der Daten zeigte, daß hier vor allem der Begriff „rather“ äußerst unterschiedlich beurteilt worden ist: Während einige der Upn „rather“ im oberen Skalenbereich platziert haben, stuften andere diesen Begriff als Ausdruck einer eher geringen Intensität ein.

Trotz einzelner solcher Unterschiede und trotz einiger instruktionsbedingten Probleme bei der Kategorienskala (s.u.) stimmen pro Erhebungsort die relativen Positionen der Skalenwerte (Mediane bzw. Mittelwerte) der Adverbien zwischen den beiden Skalen insgesamt gut überein: Die Rangkorrelation nach Kendall variiert zwischen $\tau = .91$ und $\tau = .98$.

Zu den mehrfaktoriellen Analyseergebnissen: In allen Ländern erwies sich (trivialerweise) der Faktor Adverb als signi-

fikant. Die Adverbien sind also in ihren Skalenwerten deutlich unterscheidbar. In vier der zehn Länder findet sich ein signifikanter Unterschied zwischen den Skalen: Dieser Effekt spiegelt wider, daß die Upn in mehreren Ländern die oberen Kategorien der 9-stufigen Kategorienskala nicht voll ausgenutzt haben (besonders deutlich in Norwegen). Zusätzlich zu diesen Haupteffekten sind in einigen Ländern auch verschiedene Wechselwirkungen zu verzeichnen. So findet sich in sieben Ländern eine Adverb x Skala-Wechselwirkung: Trägt man die Skalenwerte beider Skalen gegeneinander auf, so zeigt sich, daß bei mehreren dieser Länder eine Abweichung insbesondere im unteren Skalenbereich festzustellen ist; die Adverbien mit eher niedrigen Skalenwerten werden jeweils im Durchschnitt auf der 9-stu-

figen Kategorienskala (trotz Transformation der Urteile auf denselben Skalenrange wie bei der grafischen Skala) tendenziell höher als auf der grafischen Skala eingestuft. Dies ist in Abbildung 1 am Beispiel der USA-Daten dargestellt. Wir können keine plausible post hoc-Erklärung für diesen Effekt anbieten.

Aus den statistisch bedeutsamen Effekten unter Einbeziehung des Faktors „Skala“ wie auch aus nachträglich festgestellten Schwächen der Instruktion für Aufgabe 1, die in einigen der Länder aus der Übertragung der Aufgabeninstruktion in die jeweilige Landessprache resultierten, wurde die Konsequenz gezogen, für die Auswahl der Adverbien bzw. für die Skalenkonstruktion nur die Daten für die grafische Skala (Aufgabe 5) heranzuziehen.

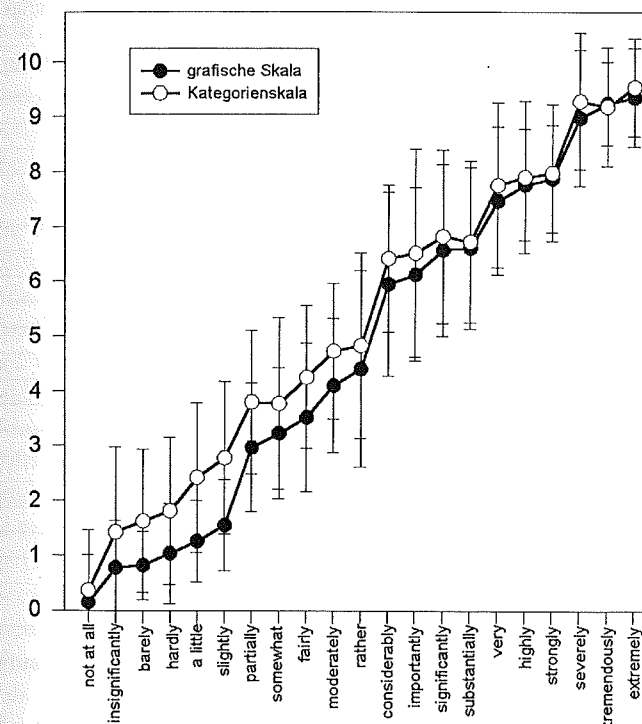


Abbildung 1: Mittelwerte und Standardabweichungen der Skalierungsaufgaben (Datensatz USA).

In zwei Ländern (USA und Spanien) erwiesen sich zudem Wechselwirkungen unter Einbeziehung des Faktors „Ort“ als signifikant. Beispielsweise ist die Ort x Adverb-Wechselwirkung in Spanien signifikant: Einige der Begriffe werden also in Pamplona anders als in Valencia eingestuft; dies könnte – einmal abgesehen von möglichen Stichprobeneffekten – darauf hindeuten, daß zumindest einige der Begriffe für die Upn in Pamplona eine etwas andere Bedeutung

Schwerpunktthema

als in Valencia haben bzw. etwas andere Intensitätsgrade repräsentieren. Unglücklicherweise wurde eins dieser problematischen Adverbien von vielen Personen in den Aufgaben 3 und 4, in denen von den Untersuchungspersonen selbst Skalen hergestellt werden sollten, besonders häufig benutzt.

3.2 Auswahl von Adverbien

Tabelle 1 enthält einige der statistischen Kennwerte für die deutschen Daten (Bochum und Oldenburg zusammen). Wie ersichtlich, sind die Skalenwerte (Mittelwerte⁷ auf der grafischen Skala) ziemlich gleichmäßig über das Kontinuum verteilt,

Begriff	Mittelwert	Standardabweichung	Differenz zum Idealwert*	Häufigkeit der Auswahl für eine 5-stufige Skala in %	Häufigkeit der Auswahl für eine 4-stufige Skala in %
überhaupt nicht	0,05	0,059	0,05	100	100
nicht	0,13	0,139			
kaum	0,93	0,526			
ein wenig	1,31	1,310			
wenig	1,39	0,973			
etwas	1,86	0,999	-0,64	37,7	
teilweise	3,53	1,334			27,8
einigermaßen	3,63	1,457			
mittelmäßig	4,78	0,690	-0,22	57,3	
ziemlich	6,01	1,613			
beträchtlich	7,02	1,358			19,6
besonders	7,28	1,215			
stark	7,62	1,319	0,12	19,6	
sehr	7,67	1,312			
erheblich	7,73	1,224			
ungemein	7,95	1,549			
enorm	8,45	1,197			
außerordentlich	8,77	1,098			
äußerst	9,07	0,993			
völlig	9,43	0,734			
total	9,54	0,804	-0,46	40,9	39,3

* Differenz zwischen dem empirischen Mittelwert und dem jeweiligen nächstgelegenen Idealwert für eine äquidistante 5-stufige Skala (0; 2,5; 5; 7,5; 10)

Tabelle 1: Kennwerte für die Begriffseinstufungen auf der grafischen Skala in Oldenburg und Bochum (N=61).

⁷ Die Mittelwerte und Mediane für die Begriffe liegen jeweils nahe beieinander; daher werden in Tabelle 1 nur die Mittelwerte dargestellt.

es gibt also auf den ersten Blick keine größeren „Lücken“. Bei näherer Betrachtung zeigt sich jedoch, daß nicht für jeden der Idealpunkte (bei der zu konstruierenden 5-er Skala also für die Positionen 0, 2,5, 5, 7,5 und 10) auch jeweils ein passendes Adverb mit entsprechendem Skalenwert vorhanden ist. Auch für die anderen Länder ergeben sich jeweils mehr oder weniger große Abstände zwischen den empirischen Skalenwerten und den Idealpunkten in Hinblick auf die zu konstruierenden 4-er oder 5-er Skalen; bei der Mehrzahl der Länder sind diese Abweichungen von den Idealwerten jedoch für eine 5-er Skala geringer als für eine 4-er Skala.

Desweiteren sind in Tabelle 1 deutliche Unterschiede in den Streuungen – selbst bei Adverbien mit sehr ähnlichen durchschnittlichen Skalenwerten – festzustellen. Beispielsweise wird „ziemlich“ offensichtlich uneinheitlicher als das benachbarte „mittelmäßig“ beurteilt.

Aus der Tabelle ergibt sich ferner, daß kaum ein Adverb bzw. Begriff jeweils von einer größeren Anzahl der Upn für eine bestimmte Skalenposition in Hinblick auf die zu konstruierenden 4-er oder 5-er Skalen ausgewählt wurde. Beispielsweise wird „mittelmäßig“ von nur etwas über der Hälfte der Upn für die mittlere Position auf der 5-er Skala ausgewählt; bei den anderen Positionen sind die relativen Wahlhäufigkeiten für die verschiedenen Adverbien jeweils noch geringer (triviale Ausnahme: „überhaupt nicht“; dieser Begriff wurde zur Skalenverankerung am unteren Skalenende benutzt). D.h., es gibt also für die verschiedenen Skalenpositionen jeweils keine sich offensichtlich anbietenden „Favoriten“. Im Datensatz für die anderen Länder sehen die Ergebnisse für die Wahlhäufigkeiten nicht wesentlich anders aus. Die Voraussetzungen für die Bildung von Skalen sind also alles andere als ideal. Wir haben dennoch versucht, aus den vorliegenden Daten Vorschläge für zu bildende Skalen zu entwickeln. Dazu hat Team 6 ('Community Response') der IC-BEN-Gruppe – via Internet – gemeinsam einen Kriteriumskatalog ausgearbeitet, der es erlaubt, aus den Kennwerten Vorschläge für 4- bzw. 5-stufige Skalen für jedes der beteiligten Länder zu gewinnen (vgl. Fields et al., 1998). Bei Anwendung dieser Kriterien zeigte es sich, daß für die Mehrzahl der Länder eine 5-er Skala einfacher als eine 4-er Skala zu formulieren ist: Die Abweichungen zwischen den empirischen Werten und den Idealpunkten waren

für die 5-er Skala etwas geringer als für die 4-er Skala; s.o.. Eine 5-er Skala erscheint uns auch aus anderen Gründen vorteilhafter: Sie erlaubt die differenziertere Erfassung der Belästigung und ist dennoch von den Befragten – auch bei telefonischer Erhebung – noch gut handhabbar.

öko:journal

Zeitschrift für Ökologie und >Dritte Welt<

Themen in Heft Nr. 16

Nicht nur reden, handeln! Ein internationaler Tourismusrat könnte Maßstäbe setzen ♦ Patagonien - Land des Windes und der Naturwunder ♦ Teutonengrill, Ökologie und Weihnachtsmarkt - Der Tourismuspolitik fehlen die Leitbilder ♦ Neue Pfade im Dschungel - naturverträgliche Nutzung in Guatemala ♦ Eine harte Nuß? Ansätze nachhaltiger Entwicklung im Südpazifik ♦ Eine Perspektive im Chaos - Ökologische Landwirtschaft in Bulgarien

Das öko:journal erscheint zweimal im Jahr.

➤ Jahresabonnement 18 DM
➤ Probeheft gegen 5 DM in Briefmarken ♦ Herausgeber:



Postfach 100 466
33504 Bielefeld
Tel. 0521 - 65943
Fax 0521 - 64975

Schwerpunktthema

Für die deutschen Daten würde bei Anwendung der Kriterien folgende vorläufige Liste für eine 5-stufige Antwortskala resultieren:

„überhaupt nicht/etwas/mittelmäßig/stark/total“.

Die Vorläufigkeit dieser Liste muß aus zumindest zwei Gründen betont werden: Die linguistische Tauglichkeit der Skala (z.B. das „Zueinanderpassen“ der Adverbien) ist noch nicht geprüft. Gegenwärtig läuft dazu u.a. auch noch eine kleine Untersuchung, in der insbesondere die Kontextabhängigkeit einiger Begriffe geklärt werden soll. Zudem sollen hier auch nicht-studentische Upn befragt werden. Für zumindest einige der verwendeten Begriffe soll zudem die Skalierung der Begriffe auf der grafischen Skala wiederholt werden. Auch die Verwendung eines Mode-Ausdrucks wie „total“ ist nicht unproblematisch. Da die Skala im gesamt deutschen Sprachraum anwendbar sein soll, steht noch eine Abstimmung mit unseren österreichischen und Schweizer KollegInnen aus. Wenn diese zusätzlichen Prüfungen und Abstimmungen in den verschiedenen Ländern abgeschlossen sind, können für die jeweils ausgewählten verbleibenden fünf Begriffe einige der oben skizzierten Analysen wiederholt werden – etwa um zu überprüfen, ob bestimmte Effekte (wie z.B. die Adverb x Ort-Wechselwirkung) auch bei Betrachtung nur der ausgewählten Begriffe noch bedeutsam sind.

4 Diskussion und Ausblick

Aus der vorstehenden Ergebnisdarstellung wird bereits ein Dilemma deutlich: Einerseits sind die vorliegenden Daten in mehrfacher Hinsicht unbefriedigend (u.a.

mehr oder weniger große Abweichungen zu den Idealpositionen, Daten vorwiegend nur von Studierenden, zweifelhafte Repräsentativität für das jeweilige Land, usw.). Andererseits bieten diese Daten trotz aller Unzulänglichkeit bessere Voraussetzungen als bisher für international vergleichende Studien: Anstatt einfach Begriffe aus einer Sprache in die andere zu übersetzen und die Vergleichbarkeit der damit ausgedrückten Intensitäten ungeprüft zu unterstellen, liegen hier zum erstenmal Daten aus einer größeren Anzahl von Ländern vor, die auf vergleichbare Weise erhoben und nach identischen Kriterien ausgewertet worden sind. Damit sind also sehr viel bessere Voraussetzungen für die Bestimmung integrativer, nationenübergreifender Dosis-Wirkungskurven gegeben als bisher (vgl. die Studien von Schultz, 1978, oder Miedema, 1993, 1998). Und so wünschenswert es wäre, für jedes der Länder ergänzende Studien (z.B. mit nicht-studentischen Upn in unterschiedlichen Regionen und unter Berücksichtigung möglicher Kontextabhängigkeiten) durchzuführen, so klar ist auch, daß sich eine vergleichbare international abgestimmte Anstrengung wie bei der Studie, über die hier berichtet wurde, so schnell nicht wiederholen läßt. Zudem stehen in mehreren der beteiligten Ländern größere Untersuchungen an; daher ist es dringend geboten, sich möglichst schnell auf eine bestimmte Antwortskala festzulegen. Es bleibt also kaum Zeit, an sich wünschenswerte ergänzende Skalierungsstudien durchzuführen.

Kontaktadresse

Fakultät für Psychologie, Ruhr-Universität Bochum, D-44780 Bochum.

Literatur

- Anderson, T.V., Kühl, H. & Relster, E. (1983). Reactions to railway noise in Denmark. *Journal of Sound and Vibration*, 87, 311-314.
- Deutsche Forschungsgemeinschaft (1974). *Fluglärmwirkungen – eine interdisziplinäre Untersuchung über die Auswirkung des Fluglärms auf den Menschen*. Bonn, Bad Godesberg: Boldt.
- Ebel, R.L. (1951). Estimation of the reliability of ratings. *Psychometrika*, 16, 407-424.
- Felscher-Suhr, U., Guski, R. & Schuemer, R. (1998). Constructing equidistant annoyance scales – an international study. In *Inter-Noise Proceedings 1998*, CD-Rom, paper 112.
- Felscher-Suhr, U., Guski, R. & Schuemer, R. (1998). Some results of an international scaling study and their implications on noise research. In *7th International Congress on Noise as a public health problem – Noise Effects '98, Congress Proceedings*, Vol. 2, 733 - 736.
- Fidell, S. & Teffeteller, S. (1981). Scaling the annoyance of intrusive sounds. *Journal of Sound and Vibration*, 78, 291-298.
- Fields, J.M. (1996). Progress toward the use of shared noise reaction questions. In *Inter-Noise Proceedings 1996*, Vol. 5, 2389-2394.
- Fields, J.M., de Jong, R.G., Flindell, I.H., Gjestland, T., Job, R.F.S., Kurra, S., Schuemer-Kohrs, A., Lercher, P., Vallet, M., Yano, T. (1998). Recommendation for shared annoyance questions in noise annoyance surveys. In *7th International Congress on Noise as a public health problem – Noise Effects '98, Congress Proceedings*, Vol. 2, 481-486.
- Fields, J.M. & Walker, J.G. (1982). The response to railway noise in residential areas in Great Britain. *Journal of Sound and Vibration*, 85, 177-255.
- Finke, H.O., Guski, R. & Rohrmann, B. (1980). *Betroffenheit einer Stadt durch Lärm. Bericht über eine interdisziplinäre Untersuchung*. Berlin: Umweltbundesamt / Braunschweig: PTB.
- Guski, R., Felscher-Suhr, U. & Schuemer, R. (1998). Entwicklung einer international vergleichbaren verbalen Belästigungsskala. In *Fortschritte der Akustik, DAGA 1998, Zürich*, 80-81.
- Hall, F.L., Birnie, S.E., Taylor, S.M. & Palmer, J.E. (1981). A direct comparison of community response to road traffic noise and to aircraft noise. *Journal of the Acoustical Society of America*, 70, 1690-1698.
- Levine, N. (1981). The development of an annoyance scale for community noise assessment. *Journal of Sound and Vibration*, 74, 265-279.
- Miedema, H.M.E. (1993). Response functions for environmental noise. In *Proceedings of the 6th International Congress, Noise & Man '93, Noise as Public Health Problem, Nice, France, 5-9 Juillet 1993*, Vol. 3, 428-433.
- Miedema, H. (1998). Revised DNL-annoyance curves for transportation noise. In *7th International Congress on Noise as a public health problem – Noise Effects '98, Congress Proceedings*, Vol. 2, 491-496.
- Möhler, U., Schuemer, R., Knall, V. & Schuemer-Kohrs, A. (1986). Vergleich der Lästigkeit von Schienen- und Straßenverkehrslärm. *Zeitschrift für Lärmbekämpfung*, 33, 132-142.
- Rohrmann, B. (1978). Empirische Studien zur Entwicklung von Antwortskalen für die sozialwissenschaftliche Forschung. *Zeitschrift für Sozialpsychologie*, 9, 222-245.
- Schultz, T.J. (1978). Synthesis of social surveys on noise annoyance. *Journal of the Acoustical Society of America*, 64, 377-405.
- Sixtl, F. (1967). *Meßmethoden der Psychologie. Theoretische Grundlagen und Probleme*. Weinheim: Beltz.
- Steinheider, B. & Winneke, G. (1993). Industrial odours as environmental stressors: Exposure-annoyance associations and their modification by coping, age and perceived health. *Journal of Environmental Psychology*, 13, 353-363.
- Thurstone, L. L. (1931). The measurement of social attitudes. *Journal of Abnormal and Social Psychology*, 26, 249-269.
- Thurstone, L. L. & Chave, E.J. (1929). *The measurement of attitudes*. Chicago, Ill.: University of Chicago Press.