

## Kapitel 2

# Grundlagen der Psychophysik und Psychometrie

In diesem Kapitel sollen einige Grundlagen der Messung mit menschlichen Versuchspersonen behandelt werden. Ziel ist es, eine quantitative Beschreibung von Wahrnehmungsgrößen zu bekommen, z.B. von Hörereignissen, Sehereignissen, oder von Qualitätsergebnissen. Da Qualität Ergebnis eines Wahrnehmungs- und Beurteilungsprozesses ist, sind solche Messungen unerlässlich, wenn die Qualität kommunikationstechnischer Systeme bestimmt werden soll. Das Kapitel orientiert sich in großen Teilen an den Ausführungen von Blauert (1994) und Jekosch (2000).

Das Gebiet der Psychophysik befasst sich mit den Zusammenhängen zwischen physikalischen Größen und deren Wahrnehmung durch den Menschen. Die hier interessierenden physikalischen Größen sind z.B. die Schallwelle, die das Ohr des Menschen erreicht, oder die elektromagnetische Welle, die auf das Auge des Betrachters trifft<sup>1</sup>. Diese physikalischen Größen sind **räumlich, zeitlich und eigenschaftlich** bestimmt; man bezeichnet sie deshalb auch als **physikalische Ereignisse** (z.B. Schallereignis).

Das physikalische Ereignis kann zu einem **Wahrnehmungsergebnis** führen, z.B. zu einem Hörereignis oder einem Sehereignis. Das Wahrnehmungsergebnis ist das Wahrgenommene; im Akustischen wird es auch als Hörgegenstand oder Hörempfindung bezeichnet. Es ist – wie alles Wahrgenommene – ebenfalls räumlich, zeitlich und eigenschaftlich bestimmt. Räumliche Merkmale sind z.B. die Entfernung, die Richtung oder die Ausdehnung von Hörereignissen; eigenschaftliche Merkmale sind z.B. die Farbe, die Klangfarbe, die Tonhöhe, die Lautheit oder die Rauigkeit. Ein Wahrnehmungsergebnis ist eindeutig mit einem physiologischen Zustand des wahrnehmenden Menschen verknüpft. Ein Wahrnehmungsergebnis ist auch mit einem physikalischen Ereignis verknüpft, und dieser Zusammenhang lässt sich z.B. durch eine Funktion beschreiben. Allerdings gibt es auch Wahrnehmungsergebnisse, die nicht durch physikalische Ereignisse hervorgerufen werden (z.B. beim Tinnitus).

---

<sup>1</sup> Wir beschränken uns in diesem Kapitel meist auf die Hör- und Sehwahrnehmung, weil sie für derzeitige informations- und kommunikationstechnische Systeme weitaus am relevantesten ist. Ähnliche Betrachtungen gelten aber auch für die taktile, die olfaktorische oder die gustatorische Wahrnehmung.

Wenn wir Wahrnehmungsergebnisse untersuchen wollen, müssen wir den Wahrnehmungsvorgang unter kontrollierten Bedingungen nachvollziehen. Dies kann z.B. in einem **Hörversuch** oder **Sehversuch** geschehen, bei dem Versuchspersonen mit physikalischen Ereignissen konfrontiert werden, und ihre Wahrnehmungsergebnisse beschreiben sollen. Die Kommunikation besteht aber nicht nur aus Hören und Sehen, sondern auch aus Sprechen bzw. Agieren, und dem Zusammenspiel zwischen Wahrnehmen und Agieren. Information über dieses Zusammenspiel lässt sich z.B. in **Konversationsversuchen** oder allgemein (bei Übertragung auf die Mensch-Maschine-Interaktion) in **Interaktionsversuchen** gewinnen, bei denen zwei (oder mehrere) Versuchspersonen unter kontrollierten Bedingungen interagieren und anschließend (oder währenddessen) eine Beurteilung der Interaktion liefern. Hör-, Seh- und auch Interaktionsversuche können als Selbstversuch (Versuchsperson und Beobachter sind identisch) durch sog. „Introspektion“ oder als Fremdversuch (Versuchsperson und Beobachter sind nicht identisch) durchgeführt werden. Im letzteren Fall ist der Beobachter auf eine Beschreibung der Versuchsperson oder seine eigene Beobachtung angewiesen.

Beide Arten von Versuchen sind **subjektiv**, d.h. sie bedienen sich menschlicher Versuchspersonen (Subjekte) als Messorgane. Subjektiv hat hier allerdings nicht die Bedeutung von „individuell“ oder gar „ungenau“. Es ist insbesondere nicht mit **objektiv** kontrastiert, in dem Sinne, dass eine physikalische Messung immer „objektiv“ und eine psychophysikalische Messung immer „subjektiv“ sei. Die Objektivität einer Messung ergibt sich aus ihrer Allgemeingültigkeit; dies kann sowohl für physikalische als auch für psychophysikalische Messungen der Fall sein. Wir sprechen deshalb im Folgenden von subjektiven Messungen, wenn daran menschliche Versuchspersonen als Messorgane beteiligt sind. Wenn wir ausdrücken wollen, dass eine Messung ohne Zutun von menschlichen Versuchspersonen zustande kommt (d.h. mit Messinstrumenten), so sprechen wir von einer **instrumentellen Messung**.

## 2.1 Eigenschaften von Messungen

Das Messen ist lt. Definition des Deutschen Instituts für Normung (DIN) „das Ausführen von geplanten Tätigkeiten zum quantitativen Vergleich der Messgröße mit einer Einheit“ (DIN 1319 Teil 1, 1995). Die Messgröße wird dabei als die „physikalische Größe, der die Messung gilt“ definiert. Teilaufgabe des Messvorgangs ist die Skalierung, d.h. **die Zuordnung von Zahlen zu Objekten nach festgelegten Regeln**.

Diese Definition von Messung lässt sich auch auf die Psychophysik übertragen. Hierbei will man Beziehungen zwischen physikalischen Phänomenen und Wahrnehmungsphänomenen quantitativ erfassen und verwendet dabei u.a. ebenfalls häufig eine Skala als Mittel der Beschreibung des Wahrgenommenen. Zu diesem Zweck werden im Folgenden die modifizierten Definitionen von Jekosch (2000) verwandt:

**Messung** (Jekosch, 2000): „Gesamtheit aller Tätigkeiten in der gesamten Messkette zur Bestimmung des Wertes einer Messgröße.“

**Messgröße** (Jekosch, 2000): „Merkmal des Messobjektes, welches im Zuge der Messung zahlenmäßig beschrieben wird.“

**Skalierung** (Jekosch, 2000): „Gesamtheit aller Tätigkeiten, die sich konkret auf den Vorgang der Zuordnung eines Wertes einer Messgröße, dessen Träger das Messobjekt ist, zu einem entsprechenden Skalenwert (Messwert) nach vorgegebenen Regeln beziehen.“

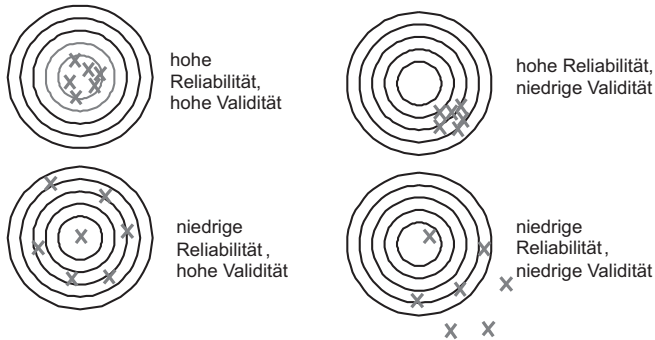
Die Skalierung ist also Teil des Messvorgangs. Hierauf wird in Kapitel 3 noch weiter eingegangen.

Damit Messungen „gute“ und sinnvolle Ergebnisse liefern müssen sie mehrere Kriterien erfüllen (Lienert, 1989):

- **Validität:** Die Validität gibt die Eignung eines Messverfahrens bzgl. seiner Zielsetzung an: Misst das Verfahren wirklich das, was es messen soll?
- **Reliabilität:** Die Reliabilität gibt die Zuverlässigkeit einer Messung an: Ist das Messergebnis bei erneuter Durchführung der Messung stabil? Man unterscheidet hierbei zwischen
  - Paralleltest-Reliabilität: Wie stark korrelieren die Ergebnisse, wenn mit einer Stichprobe von Versuchspersonen zwei streng miteinander vergleichbare Messungen durchgeführt werden?
  - Retest-Reliabilität: Wie stark korrelieren die Ergebnisse, wenn mit ein und derselbe Stichprobe von Versuchspersonen zweimal die gleiche Messung durchgeführt wird?
  - Innere Konsistenz: Diese kann z.B. ermittelt werden, indem man die Messwerte eines Tests aufteilt (splittet) und die Teilergebnisse miteinander vergleicht.
- **Objektivität:** Die Objektivität gibt den Grad der interpersonellen Übereinstimmung von Messungen an: Ist das Messergebnis abhängig von demjenigen, der die Messung durchführt?

Daneben gibt es noch weitere Nebengütekriterien, wie z.B. die Ökonomie, Normierbarkeit, Nützlichkeit und Vergleichbarkeit von Messungen. Der Unterschied zwischen Validität und Reliabilität ist in Abb. 2.1 skizziert.

Besonders bei der Messung mit Versuchspersonen stellt sich die Frage der Verallgemeinerbarkeit von Messungen – allerdings nicht in Bezug auf die Abhängigkeit vom Versuchsleiter, sondern in Bezug auf die Abhängigkeit vom Messorgan, d.h. von der Versuchspersonengruppe, die für die Messung verwendet wird. Hierbei ist der sog. **Analogieschluss** wichtig: Die physiologisch-psychologische Parallelität der Vorgänge, die ich an mir selbst beobachte, berechtigen mich zu dem Schluss, dass ein Mitmensch, dessen physiologischen und psychologischen Verhältnisse den meinen analog sind, bei den gleichen physiologischen Geschehen auch Analoges erlebt wie ich (Lorenz, 1963). Dieser Analogieschluss ist niemals beweisbar, da ich



**Abb. 2.1** Zur Reliabilität und Validität von Messungen

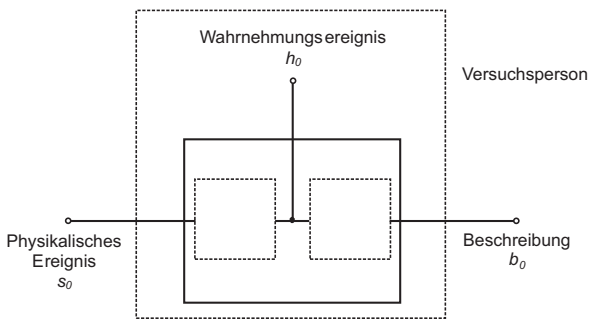
nicht die gleichen Wahrnehmungsereignisse haben kann wie ein Mitmensch. Über die Wahrnehmungsereignisse kann deshalb nur auf dem Umweg der Beschreibungen geschlossen werden.

Dennoch sollen viele Messungen **generalisierbar** sein, d.h. die Messergebnisse sollen nicht nur für die betrachtete Gruppe von Versuchspersonen gültig sein, sondern auch für eine andere Gruppe von Personen (z.B. die Nutzer eines Kommunikationstechnischen Systems, die aber nicht befragt werden können). Hierzu ist es notwendig, dass die Gruppe der Versuchspersonen **repräsentativ** ist, d.h. dass sie bzgl. aller Eigenschaften, die für das Messobjekt relevant sind (hier z.B. Hör- oder Sehvermögen, kommunikative Fähigkeit, Erfahrung mit dem betrachteten System, Motivation, etc.) möglichst gut mit dem interessierenden Personenkreis übereinstimmen, für den die Messung gelten soll.

Je nach dem zu bestimmenden Messobjekt und der Messgröße kann eine Messung als Beobachtungsverfahren, als Beurteilungsverfahren, als instrumentelles Verfahren, als Berechnungsverfahren oder als statistisches Schätzverfahren realisiert werden (Jekosch, 2000). Ein reines Beobachtungsverfahren liegt z.B. bei der Untersuchung der Reaktion einer Versuchsperson auf Hörereignisse vor, oder bei der Untersuchung des Verhaltens einer Versuchsperson in einer Konversationssituation. Ein Beurteilungsverfahren wird z.B. bei der Bestimmung der Sprachqualität im Hörversuch oder der Bildqualität im Sehversuch angewendet, bei dem die Versuchspersonen die Qualität dargebotener Proben quantitativ beschreiben müssen (z.B. auf einer Skala). Ein instrumentelles Verfahren wird z.B. zur Bestimmung physikalischer Messgrößen eingesetzt (z.B. Zeitmessung oder Längenmessung). Ein Berechnungsverfahren kann z.B. zur Berechnung des gewichteten Schalldruckpegels (in dB(A)) eingesetzt werden; hierbei ist die physikalische Messung (des Mikrophonsignals) mit einer Berechnung verbunden.

## 2.2 Psychophysikalische Messungen

Wir wollen nun genauer betrachten, was in einer Versuchsperson während einer psychophysikalischen Messung vorgeht, um daraus Anforderungen an den Messprozess abzuleiten. Dabei soll zunächst davon ausgegangen werden, dass die Versuchsperson sich in einer rein „passiven“ Wahrnehmungssituation befindet, d.h. nicht in einer Interaktion. In dieser Situation wird ihr vom Versuchsleiter ein physikalisches Ereignis  $s_0$ <sup>2</sup>, welches das interessierende, zu messende Merkmal aufweist. Dieses physikalische Ereignis führt zu einem Wahrnehmungsereignis  $h_0$ <sup>3</sup>, welches allerdings innerhalb der Versuchsperson liegt – nicht unbedingt im Sinne der räumlichen Eigenschaften des Wahrnehmungsereignisses, aber im Sinne der Zugänglichkeit. Das Wahrnehmungsereignis ist also der direkten Messung durch den Versuchsleiter nicht zugänglich. Die Versuchsperson wird nun aufgefordert, eine Beschreibung des Wahrnehmungsereignisses zu geben, und sie liefert ein Beschreibungsereignis  $b_0$ . Diese Situation ist in Abb. 2.2 gezeigt.



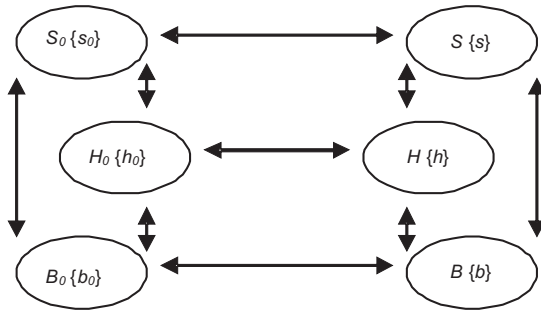
**Abb. 2.2** Schema einer Versuchsperson in einer psychophysikalischen Messung, in Anlehnung an Blauert (1997), S. 6

Die Boxen deuten an, dass zwischen dem Wahrnehmungsereignis und dem Beschreibungsereignis ein Umsetzungsprozess stattfindet. Wahrnehmungsereignis und Beschreibungsereignis sind also i. Allg. nicht identisch. Trotzdem ist man auf das Beschreibungsereignis angewiesen, um das Wahrnehmungsereignis qualitativ (bzgl. des Inhaltes) und quantitativ (bzgl. der mengenmäßigen Ausprägung der interessierenden Merkmale) zu erfassen.

Wenn man den Wahrnehmungs- und Beurteilungsvorgang wiederholt durchführt (z.B. mit verschiedenen Versuchspersonen) so erhält man Mengen von physikalischen Ereignissen, Wahrnehmungsereignissen und Beschreibungsereignissen. In Anlehnung an das zuvor gezeigte einfache Schema der Versuchsperson kann man diese Mengen wie in Abb. 2.3 angegeben darstellen.

<sup>2</sup> Definiert ursprünglich von Blauert (1997) für das Schallereignis

<sup>3</sup> Definiert ursprünglich von Blauert (1997) für das Hörereignis



**Abb. 2.3** Zusammenhang der bei psychophysikalischen Messung auftretenden Ereignisse und Skalen, in Anlehnung an Blauert (1997), S. 8

Auf der linken Seite sind die Grundmengen der Ereignisse dargestellt:  $S_0$  als die Menge der physikalischen Ereignisse  $s_0$ ,  $H_0$  als die Menge der Wahrnehmungseignisse  $h_0$ , und  $B_0$  als die Menge der Beschreibungseignisse  $b_0$ . Auf der rechten Seite befinden sich die Skalen: Physikalische Skala  $S$  mit den Elementen  $s$ , Wahrnehmungsskala  $H$  mit den Elementen  $h$ , und Beschreibungsskala  $B$  mit den Elementen  $b$ . Zwischen den Elementen der Grundmengen wie auch zwischen den Elementen der Skalen bestehen funktionale Zusammenhänge, die man als psychophysikalische Funktionen bezeichnet. Zwischen den Grundmengen und den jeweiligen Skalen bestehen ebenfalls Zusammenhänge, die man als Skalierungsfunktionen bezeichnet.

In einem psychophysikalischen Experiment interessieren wir uns meist für den Zusammenhang zwischen  $s$  und  $h$ , d.h. für die Funktion  $h = f(s)$ . Bei einem anliegenden physikalischen Ereignis  $s_0$  kann  $s$  z.B. mit einem physikalischen Messinstrument gemessen werden, d.h.  $s = f(s_0)$ . Wir gehen nun davon aus, dass das physikalische Ereignis  $s_0$  mit einem Wahrnehmungseignis  $h_0$  verknüpft ist. Da  $h_0$  nicht direkt messbar ist, instruiert man die Versuchsperson, eine entsprechende Beschreibung  $b_0$  zu geben, die eine zahlenmäßige Beschreibung des Wahrnehmungseignisses  $h_0$ , also  $h$  darstellt. Man kann deshalb sagen, dass  $b_0 = h$ . Somit lässt sich der Zusammenhang zwischen  $s$  und  $h$  indirekt mittels zweier Messungen bestimmen: einer physikalischen Messung  $s = f(s_0)$ , und einer psychophysikalischen Messung  $b_0 = f(s_0)$ , wobei die Versuchsperson gleichzeitig den Wahrnehmungszusammenhang  $h_0 = f(s_0)$  beinhaltet. Die Versuchsperson wirkt also gleichzeitig als wahrnehmendes und beurteilendes Messorgan.

In diesem Messsystem können dreierlei Arten von Messfehlern auftreten, vgl. Blauert (1997):

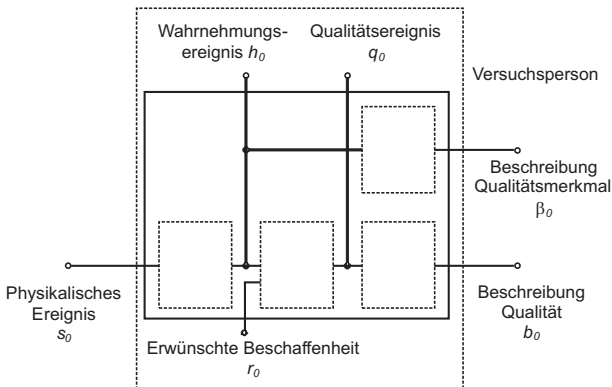
- die Messungenauigkeit des physikalischen Messgerätes
- die Messungenauigkeit des psychophysikalischen Messorgans
- Schwankungen im wahrnehmenden Messorgan

Unter der Annahme, dass das physikalische Ereignis mit recht hoher Genauigkeit erzeugt und gemessen werden kann (also konstant ist), und dass das beschreibende Systemelement invariant gegenüber Versuchswiederholungen ist (was man z.B.

durch eine gute Planung des Experimentes und eine genaue Instruktion der Versuchspersonen erreichen kann), können die Änderungen des Wahrnehmungsereignisses und des damit verbundenen Beschreibungsereignisses vorwiegend mit Änderungen des wahrnehmenden Elementes erklärt werden. Diese Schwankungen im Wahrnehmungsprozess sind jedoch einer Kontrolle oder Vorhersage nicht zugänglich; man muss deshalb von einer Zufallsvariablen ausgehen, die die Messergebnisse beeinflussen kann. Voraussetzung dafür ist allerdings, dass die psychophysikalischen Methoden und die Instruktionen der Versuchspersonen derart gestaltet sind, dass sich hierbei keine Messfehler ergeben.

## 2.3 Messung von Qualität und Usability

Die Wahrnehmungs- und Beurteilungsprozesse, die bei der Bildung von Qualität eine Rolle spielen, wurden bereits in Abschnitt 1.3 erläutert. Dort wurde insbesondere gezeigt, dass ein Vergleich zwischen dem Wahrgenommenen und dem Erwarteten stattfindet, wenn Qualität bestimmt werden soll. Diesen Vergleich kann man in einem erweiterten Schema der Versuchsperson berücksichtigen, wie es in folgender Abb. 2.4 dargestellt ist.



**Abb. 2.4** Erweitertes Schema einer Versuchsperson in einer psychophysikalischen Messung, in Anlehnung an Raake (2006)

Bis zum Wahrnehmungsereignis sind beide Schemata zunächst identisch. Aufbauend auf das Wahrnehmungsereignis findet nun die Bestimmung der Qualität statt. Dabei wird zunächst eine Unterscheidung zwischen Qualitätsmerkmalen, d.h. einzelnen erkannten und benennbaren Eigenschaften von Qualität, sowie der Qualität als Ganzes (Qualitätsereignis) gemacht. Beide sind wiederum durch Prozesse vom Wahrnehmungsereignis getrennt, die sich im Innern der Versuchsperson abspielen.

Neben dem Wahrnehmungsprozess ist für die Bewertung von Qualität vor allem die interne Referenz der Versuchsperson (erwünschte Beschaffenheit) wichtig, vgl. Abb. 2.4. In dieser Referenz werden alle Aspekte der „individuellen Erwartungen, sachgerechten Erfordernisse oder gesellschaftlichen Forderungen“ (vgl. die Definition von Qualität in Abschnitt 1.2) abgebildet. Die Referenz umfasst insbesondere individuelle Präferenzen der wahrnehmenden Person, besondere Fähigkeiten oder Wissen, Emotionen, aufgabenbezogene Aspekte und Funktionalität, sowie Gewöhnung und Tradition. Aufgrund der Individualität dieser Referenz muss die Messung von Qualität mit Versuchspersonen gleichen Hintergrunds durchgeführt werden, wenn man zu validen Ergebnissen kommen möchte. Dieses Ziel ist oft nur schwer zu erreichen, insbesondere, wenn die Eigenschaften der Referenz nicht komplett bekannt sind. Offensichtlich ist aber, dass bspw. trainierte Experten eine andere Referenz aufweisen können als „normale“ Benutzer eines Systems oder Dienstes. Aus diesem Grunde liefern Evaluierungen mit Experten, wie sie häufig in der Entwicklungsphase von neuen Diensten und Systemen durchgeführt werden, meist nur begrenzt valide Ergebnisse. Im Gegenzug können die Ergebnisse aber sehr analytisch sein, da die Experten darauf trainiert werden können, einzelne Qualitätsmerkmale zu erfassen und zu unterscheiden.

Je nach interessierendem Merkmal kann der Vergleich zwischen Wahrnehmungsereignis und Referenz auf verschiedenen Ebenen stattfinden. Interessiert man sich nur für die Eigenschaften des physikalischen Ereignisses (bspw. für den Lautstärkepegel eines Geräusches), so kann man physikalische Messgeräte zur Hilfe nehmen. Mit ihrer Hilfe kann man bestimmen, ob der Lautstärkepegel unterhalb eines bestimmten Grenzwertes bleibt. Interessiert man sich hingegen für die formbezogenen Merkmale des Wahrnehmungsereignisses (bspw. eines Hörereignisses), so können psychophysikalische Merkmale zur Hilfe genommen werden, wie die Lautheit, die Rauigkeit, etc. Auch diese Merkmale lassen sich zu einem Teil in „neutralisierter Form“ betrachten. Bspw. kann die Lautheit von Schallen in einem standardisierten Hörversuch ermittelt werden. Ob ein Hörereignis dann tatsächlich als „zu laut“ wahrgenommen wird hängt daneben aber auch von der Bedeutung des Geräusches sowie von der Hörsituation ab, die in der Referenz mit erfasst werden. Man hat etwa herausgefunden, dass Schienenfahrzeuglärm i. Allg. als angenehmer empfunden wird als Fluglärm gleicher subjektiver Lautheit. Tritt der Lärm während der Ruhephase (Schlafzeit) auf, so wird er als deutlich unangenehmer empfunden als während der Aktivitätsphasen.

Neben der Qualität des physikalischen und des Wahrnehmungsereignisses liegt das Interesse aber meist in der Qualität des Systems oder des darauf beruhenden Dienstes. Um diese Qualität adäquat zu erfassen, muss der Vergleich auf psychologische, semantische und funktionale Aspekte erweitert werden. Die Bestimmung eines dazu passenden Anforderungsprofils ist nicht einfach, da dieses Profil die interne Referenz der Versuchsperson vollständig abdecken müsste. Eine weitgehende Bestimmung ist aber notwendig, um im Rahmen einer System-Qualitätsmessung valide Ergebnisse zu erzielen.

Bislang wurde von „passiven“ Versuchspersonen ausgegangen, die uns nach erfolgter Selbst-Introspektion eine Auskunft über das Wahrnehmungsereignis bzw.



das Qualitätsereignis liefern. Versuchspersonen sind aber in den seltensten Fällen wirklich passiv. So zeigen sie körperliche Veränderungen bspw. des Pulsschlages oder des Hautleitwertes, und sie reagieren auf Stimuli bspw. durch Blickbewegungen. Solche Reaktionen können ebenfalls als Indikatoren für Wahrnehmungsereignisse oder Qualitätsereignisse herangezogen werden. Darüber hinaus kann versucht werden, Hirnaktivitäten auf verschiedenen Ebenen zu bestimmen, und diese Informationen ebenfalls mit Qualitätsaspekten wie Komfort, Stress oder *Joy-of-Use* in Verbindung zu bringen. Diese **indirekten Messungen** von Wahrnehmungsereignissen stehen noch am Anfang ihrer Entwicklung; sie könnten jedoch von großem Nutzen sein, da der Reflexionsprozess weitgehend übergangen wird, und unmittelbare Reaktionen vielleicht auf tieferliegende Wahrnehmungsprozesse schließen lassen.

Reaktionen von Versuchspersonen können natürlich auch in Interaktionssituationen zur Messung von Wahrnehmungs- und Qualitätsereignissen herangezogen werden. Dabei sind zwei Fälle zu unterscheiden:

- Innerhalb der „normalen“ Benutzung eines interaktiven Systems oder Dienstes agiert die Versuchsperson, und ihre Aktionen beeinflussen die Interaktion – und somit das Wahrnehmungsereignis und die Qualität. Die Versuchsperson wird also – neben ihrer Rolle als Messorgan – selbst zum handelnden Subjekt innerhalb des Qualitätsmessvorganges. Dabei werden die Aktionen der Versuchsperson von mindestens drei Dingen abhängen: Der Persönlichkeit der Versuchsperson (z.B. introvertiert vs. extrovertiert), dem Ziel, das die Versuchsperson mit der Interaktion verfolgt, sowie von den Reaktionen des Systems. Man kann nun versuchen, die Handlungen der Versuchsperson quantitativ zu erfassen, und aus diesen Metriken Indikatoren für die Qualität abzuleiten. Entsprechende Methoden sind in Kapitel 7 bis 9 beschrieben. Bei der Interpretation der Ergebnisse muss dann allerdings unterschieden werden zwischen den oben genannten Einflüssen.
- Der Versuchsperson können speziell auf den Beurteilungsprozess zugeschnittene Aufgaben gestellt werden. Bspw. kann die Versuchsperson aufgefordert werden, bestimmte Aufgaben schnell mit dem System zu lösen, oder ihr werden parallel zur Interaktion weitere Aufgaben (sog. *Parallel Tasks*) gestellt, vgl. z.B. Chateau et al. (2006). Der Erfolg bei der Lösung dieser Aufgaben kann dann als Indikator für die Qualität herangezogen werden. Bspw. könnte man zwei Versuchspersonen auffordern, bestimmte Informationen so schnell wie möglich über eine gestörte Telefonverbindung auszutauschen; die Effizienz des Informationsaustausches (quantifiziert z.B. durch die Anzahl der Einzelinformationen pro Zeiteinheit) könnte als Indikator für die Qualität der Verbindung herangezogen werden. Es ist offensichtlich, dass dabei auch die Persönlichkeit der Versuchsperson sowie ihre (im Test u.U. künstlich gegebene) Motivation eine Rolle spielt.

Mit Hilfe solcher Messungen erhält man allerdings nur indirekte Indikatoren für die Qualität und Gebrauchstauglichkeit eines interaktiven Dienstes; direkte Qualitätsmesswerte verlangen das Urteil der Versuchsperson.

Eine genaue Kenntnis der Einflussfaktoren auf die vom Menschen erfahrene Qualität ist aus zweierlei Gründen wichtig. Zum einen müssen die Einflussfaktoren bei der Definition eines geeigneten Messaufbaus berücksichtigt werden. Will

man bspw. den Einfluss von Sprachkodierern und Paketverlusten auf die Übertragungsqualität einer Voice-over-IP-Verbindung messen, so könnte man einen Hörversuch mit naiven Versuchspersonen der interessierenden Zielgruppe durchführen. Je nach Realismus der Hörsituation würde man einen Qualitätsmesswert (z.B. in Abhängigkeit von der IP-Konfiguration) erhalten, der mehr oder weniger repräsentativ für die Gesamtqualität in der Hörsituation ist. Möchte man hingegen eine genaue Aufschlüsselung der verschiedenen Störungen (Rauschhaftigkeit, Klangverfärbung, Unterbrochenheit) haben, so sind u.U. trainierte Versuchspersonen besser geeignet, denen man speziell ausgewähltes Sprachmaterial in einer idealisierten (z.B. extrem ruhigen) Situation vorspielt. Hierdurch kann man sehr analytische Aussagen über einzelne Störquellen erhalten. Möchte man hingegen den Einfluss der Paketverluste auf die Effizienz der Verbindung testen, so sollte man einen Konversationsversuch in einer realistischen Nutzungssituation (z.B. am Rechner) durchführen, bei dem Versuchspersonen z.B. bestimmte Aufgaben mit Hilfe der VoIP-Verbindung lösen müssen. In solchen Versuchen lassen sich allerdings keine analytischen Informationen über die Quelle der Störungen abfragen.

Neben der Bestimmung des Messaufbaus ist Wissen über die bei der Qualitätsmessung beteiligten Prozesse und Einflussfaktoren auch zur Definition von geeigneten Vorhersagemodellen wichtig. So kann man versuchen, im Idealfall alle beteiligten Prozesse individuell algorithmisch zu beschreiben, und daraus einen Schätzwert für das Qualitätsurteil zu berechnen. Auch wenn ein solches komplettes Modell bislang nicht vorliegt, so beruhen doch viele der in Kapitel 9 vorgestellten Verfahren darauf, einzelne Prozesse mehr oder weniger detailliert nachzuvollziehen. Aufgrund der Komplexität der beteiligten Vorgänge erzielt man leider nicht immer optimale Ergebnisse, wenn man einzelne Prozesse möglichst exakt abbildet; häufig kann man mittels einfacher Interpolation kurzfristig zu Lösungen kommen, welche bessere Schätzwerte liefern. Solche kurzfristigen Lösungen sind jedoch meist sehr speziell und wenig verallgemeinerbar. Wissen über die Wahrnehmungs- und Beurteilungsprozesse zählt sich also langfristig meist aus.

## 2.4 Nutzertypen

Offensichtlich spielt die Versuchsperson eine große Rolle bei einer psychophysikalischen Messung. Leider steht für eine Qualitätsmessaufgabe praktisch nie die gesamte Zielgruppe vollständig zur Verfügung (mit Ausnahme von sehr speziellen Nutzungskontexten und Systemen). Aus diesem Grunde müssen alle für die Messung relevanten Eigenschaften der Versuchspersonen bekannt sein, sodass im Sinne eines validen und reliablen Ergebnisses eine optimale Auswahl von Versuchspersonen getroffen werden kann. Für kommunikations- und informationstechnische Systeme sind dabei die folgenden Eigenschaften interessant:

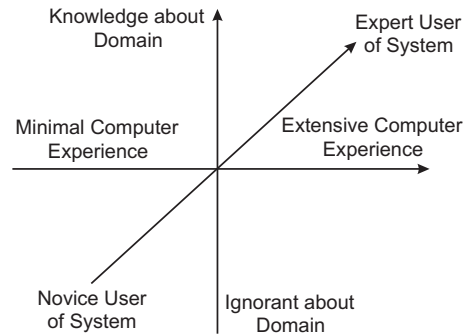
- **Wahrnehmungseigenschaften:** In Abschnitt 2.2 sind wir bislang von zufälligen Schwankungen des wahrnehmenden Systemelementes ausgegangen. Dies trifft aber nicht für alle möglichen Nutzer zu. Bspw. nehmen die Wahrneh-

mungsleistungen normalerweise mit dem Alter ab (Einschränkung des Hör- und Sehvermögens, ersteres insbes. bei hohen Frequenzen), oder spezielle Gruppen (bspw. Jugendliche) weisen verhaltensbedingte Wahrnehmungseinschränkungen auf (Hörverluste durch laute Diskothekbeschallung). Es kann sinnvoll sein, Versuchspersonen nach solchen Kriterien auszuwählen, oder ihre Wahrnehmungsleistungen zumindest vor dem Versuch zu überprüfen.

- **Verhaltenseigenschaften:** Aufgrund von Erfahrungen, aus genetischen wie auch aus anderen (teilweise nicht erforschten) Gründen weisen bestimmte Versuchsgruppen besondere Verhaltensweisen auf, die für die Qualität eine Rolle spielen können. So kann es bspw. für ein interaktives System von Bedeutung sein, ob es von Rechts- oder Linkshändern bedient wird. Der regionale und soziale Hintergrund wird bspw. die sprachliche Ausdrucksweise beeinflussen. Sowohl Wahrnehmung als auch Verhalten ändern sich mit dem Alter und können darüber hinaus geschlechtsspezifisch ausgeprägt sein; dabei kann es notwendig sein, zwischen dem biologischen (*Sex*) und dem (sozialen/psychologischen) Identitätsgeschlecht (*Gender*) zu unterscheiden.
- **Erfahrung:** Bei der Bildung der Referenz spielen individuelle Erfahrungen eine besondere Rolle. So bilden sich durch Gewöhnung spezielle Erfahrungen heraus, die dann als Referenz im Qualitätsbeurteilungsprozess dienen können. Die Erfahrungen können sich auf das betrachtete System (Messobjekt) oder auf andere vergleichbare Systeme beziehen, die eine ähnliche Funktionalität besitzen. Darüber hinaus sind auch Erfahrungen mit der Domäne des Systems (bspw. Erfahrung mit dem öffentlichen Verkehr bei einer Bahnauskunft) wichtig für die Beurteilung eines Systems. Erfahrungen entwickeln sich bei der längeren Benutzung eines Systems. Daher ist es wichtig zu wissen, ob es sich bei den Versuchspersonen um erfahrene Benutzer **des betrachteten Systems**, um **mit ähnlichen Systemen erfahrene Benutzer** oder um **mit der Domäne erfahrene Benutzer** handelt. Man bezeichnet einzelne dieser Gruppen häufig ungenau als „Experten“ (im Gegensatz zu „Novizen“), leider ohne jedoch anzugeben, um welche Expertise es sich dabei handelt.
- **Motivation:** Wichtig für die Beurteilung der Qualität eines informations- oder kommunikationstechnischen Systems ist es, aus welchem Grunde es benutzt wird. Dabei kann unterschieden werden zwischen beruflicher oder privater Nutzung, sporadischer oder regelmäßiger Nutzung, und es kann die Wichtigkeit der Benutzung klassifiziert werden (bspw. bei Notrufen oder bei finanziellen Transaktionen).
- **Individuelle Präferenzen, Fähigkeiten und Wissen:** Diese können ebenfalls einen Einfluss auf die Qualitätsbeurteilung haben, lassen sich aber in den wenigsten Fällen gut klassifizieren und bei der Auswahl der Versuchspersonen berücksichtigen. Ein Beispiel hierfür ist die Auswahl von Musikern oder Tonmeistern zur analytischen Beurteilung von Hörproben; hierbei wird neben einer höheren Erfahrung und Wissen über Harmonien auch davon ausgegangen, dass diese Personen eine besondere Affinität zum analytischen Hören besitzen, u.U. sogar ein sog. „absolutes Gehör“, welches für manche Beurteilungsaufgaben von Vorteil sein kann.

Um Versuchspersonen bzgl. der o.g. Eigenschaften einzuordnen, bedient man sich verschiedener Klassifikationsschemata. Gebräuchlich sind bspw. Klassifikationen nach:

- **Nutzerexpertise:** Nielsen (1993) unterscheidet bspw. 3 Arten von Expertise in seinem *User Cube*: Erfahrung mit dem System, mit Computern im Allgemeinen, sowie mit der Aufgaben-Domäne, vgl. Abb. 2.5.

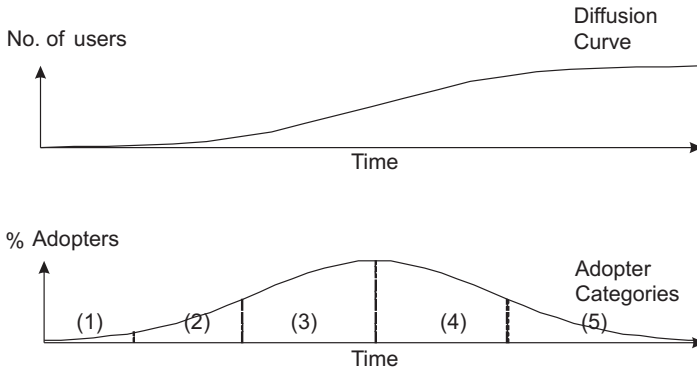


**Abb. 2.5** Klassifikation von Nutzern nach Expertise, nach Nielsen (1993), basierend auf Cotterman und Kumar (1989)

- **Annahmefähigkeit von Innovationen:** In der Innovationsforschung bedient man sich dabei häufig eines Diffusionsmodells. Man geht davon aus, dass sich eine Innovation in Form einer s-förmigen Kurve in einem Nutzerkreis durchsetzt. Dabei nehmen unterschiedliche Gruppen von Nutzern die Innovation allerdings zu unterschiedlichen Zeitpunkten an. Man unterscheidet deshalb zwischen
  1. *Innovators*, einer kleinen Gruppe von Nutzern, die neue Produkte schnell kaufen und neue Technologien bereitwillig annehmen; hierbei handelt es sich häufig um Menschen mit höherem Einkommen, höhere beruflicher Stellung, und um sozial mobile Menschen;
  2. *Early Adopters*, d.h. eine größere Gruppe von Nutzern, die den Innovatoren folgen; auch sie kaufen ein Produkt recht schnell, sind aber stärker in ihren sozialen Gruppen verankert und den darin bestehenden Normen verhaftet;
  3. *Early Majority*, d.h. eine größere Mehrheit (ca. 1/3 der gesamten Nutzer), die erst danach in den Markt eintreten und weniger bereit sind, Risiken in Kauf zu nehmen;
  4. *Late Majority*, eine größere Gruppe von Nutzern, die die Nutzung erst beginnen, wenn die Neuheit eines Produktes bereits wieder abnimmt; sie sind weniger durch Gruppennormen beeinflusst und können bspw. durch Werbung überzeugt werden; sowie die
  5. *Laggards*, d.h. die Nachzügler, die ein Produkt erst annehmen wenn es bereits vollständig am Markt etabliert ist.

Diese Gruppen sind in Abb. 2.6 skizziert.

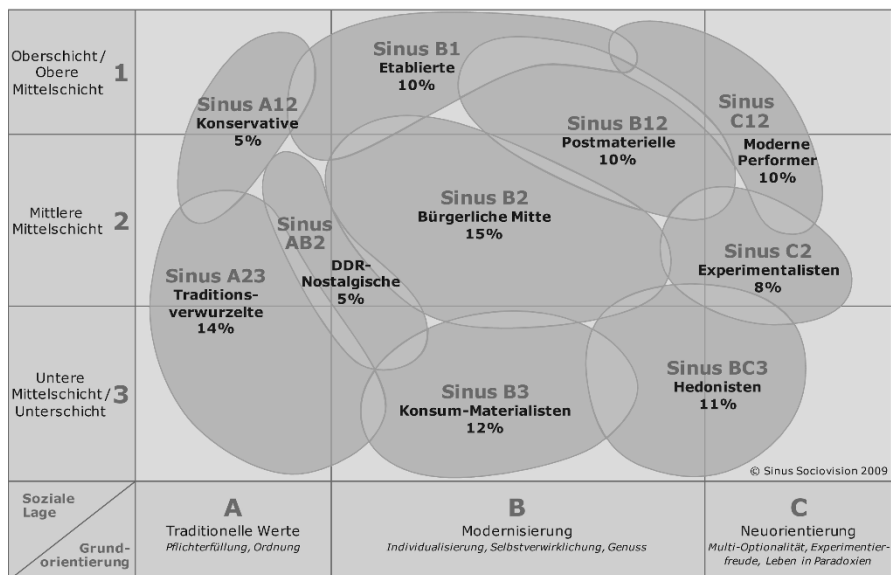
- **Kaufverhalten:** Je nach interessierendem System lassen sich Nutzergruppen als Käufergruppen identifizieren. Diese Klassifikation ist naturgemäß sehr stark



**Abb. 2.6** Diffusionsmodell für die Annahmefähigkeit von Innovationen, vgl. z.B. Rogers (1983)

domänenabhängig. Für die Telekommunikation hat sich bspw. eine Klassifikation bewährt, bei der Nutzer zunächst eingeschätzt werden nach ihrem sozialen Status und ihren Werten (traditionelle Werte, Modernität und Experimentierfreude). Aus dieser Einschätzung lassen sich ca. 10 verschiedene Nutzergruppen ableiten, die sich in Bezug auf ihr Alter, Bildung, Einkommen, zur Verfügung stehendes Budget, Affinität zu Innovationen und zu Technik im Allgemeinen, sowie in ihrem Kaufverhalten unterscheiden. Solche Klassifikationen sind i. Allg. von wirtschaftlicher Bedeutung und deshalb nicht öffentlich zugänglich; ein öffentlich zugängliches Beispiel ist die Klassifikation der sog. Sinus-Milieus<sup>®</sup> der Fa. Sinus Sociovision, welche in Abb. 2.7 gezeigt ist. Für bestimmte Untersuchungen ist es notwendig, diese Gruppen weiter zu unterteilen; Feinstrukturierungen bis hin zu 40–50 Nutzergruppen sind keine Seltenheit.

- Verhalten im Umgang mit Systemen:** Für die Bestimmung von Qualität und Gebrauchstauglichkeit ist insbesondere eine Klassifikation nützlich, bei der Nutzer nach ihrem Verhalten im Umgang mit einem informations- oder kommunikationstechnischen System unterschieden werden. Hierzu stellten Naumann et al. (2008) einen Klassifikationsansatz vor. Auf Basis eines Expertenworkshops, einer Literaturrecherche und einer sog. *User Clinic* wurden zunächst Faktoren identifiziert und gewichtet, die voraussichtlich wichtig für den Umgang eines Nutzers mit informations- und kommunikationstechnischen Systemen (IKT-Systemen) sind. Hierbei wurden Erfahrung mit und Affinität zu IKT-Systemen, sowie generelle Arbeitsmethoden und Fähigkeiten (kognitive Fähigkeiten, Problemlösungsstrategien, Zielstrebigkeit, etc.) als wichtigste Einflussfaktoren identifiziert, daneben (weniger wichtig) auch das Domänenwissen, Sprachkompetenz, Alter, und die Orientierung nach sozialen Normen. Auf Basis dieser Eigenschaften wurden 7 Nutzertypen definiert, die durch ihre Grundeinstellung zu sowie ihre Erfahrung mit IKT-Systemen charakterisiert sind, vgl. Tabelle 2.1. Obwohl diese Typisierung bislang nicht vollständig validiert wurde scheint sie insbesondere für die Usability-Forschung von besonderer Bedeutung zu sein.



**Abb. 2.7** Unterteilung von Kundengruppen in Deutschland 2009 – Soziale Lage und Grundorientierung. Kategorisierung der Fa. Sinus Sociovision

## 2.5 Psychometrische Methoden

In diesem Abschnitt sollen nun einige generelle Eigenschaften von psychometrischen Methoden diskutiert werden – also Methoden, mit deren Hilfe man quantitative Aussagen zur Wahrnehmung von Versuchspersonen erhalten kann. Beispiele für die Anwendung einzelner Methoden bei der Beurteilung der Qualität kommunikationstechnischer Systeme werden in den Kapiteln 5 bis 8 gegeben.

**Tabelle 2.1** Klassifizierung von Nutzertypen nach ihrem Verhalten im Umgang mit IKT-Systemen, vgl. Naumann et al. (2008).

Benutzertyp	Einstellung gegenüber IKT	Erfahrung mit IKT
Der IKT-Ängstliche	gering	gering
Der vertrauende IKT-Benutzer	gering bis mittel	gering
Der interessierte Amateur-IKT-Benutzer	gering bis mittel	mittel
Der erfahrene IKT-Benutzer	hoch	hoch
Der pragmatische, inspirierte IKT-Nutzer	mittel	mittel
Der spielerische IKT-Benutzer	hoch	hoch
Der funktionsliebende IKT-Benutzer	mittel bis hoch	mittel

Psychometrische Methoden lassen sich nach einer Reihe von Kriterien klassifizieren (Blauert, 1994). Gebräuchlich sind insbesondere folgende Kriterien:

1. Nach der **Skalierungsmethode** und dem sich daraus ergebenden Skalenniveau:

- Methoden der **Ratio-Skalierung**: *Magnitude Estimation* (ME), *Ratio Estimation*, *Magnitude Production*, *Ratio Production*, etc.
- Methoden der **Ordinal- oder Intervallskalierung**: Kategorienzuordnung, Herstellung von Kategorien, Paarvergleich, Ähnlichkeitsskalierung, etc.
- Methoden der **Nominalskalierung**: Bestimmung von Wahrnehmbarkeitsschwellen, Identifikationstests wie z.B. Verständlichkeitstests, etc.

Zur Skalierung und den hier erwähnten Skalen vgl. auch Kapitel 3.

2. Nach der **Präsentationsmethode**: Herstellungsmethoden vs. Konstanzmethoden.

- **Herstellungsmethode**: Voraussetzung hierfür ist, dass sich das interessierende Stimulusmerkmal kontinuierlich verändern lässt. Dabei wird das Stimulusmerkmal solange eingeregelt, bis eine bestimmte Bedingung erfüllt ist (bspw. etwas ist genauso laut oder genauso hell wie ein Referenzstimulus). Es spielt dabei keine Rolle, ob der Versuchsleiter oder die Versuchsperson selbst die Regelung vornimmt.
- **Konstanzmethode**: Hierbei ist das Stimulusmerkmal während der Darbietungsdauer konstant, und die Versuchsperson wird aufgefordert, aus einem Vorrat von Urteilen das jeweils passendste herauszusuchen – bspw. mittels einer Skala. Der Versuch wird mit einer Vielzahl von Stimuli wiederholt.

3. Nach der „**Modalität**“ des Versuchs: Hörversuche, Sprechversuche, Sehversuche, Konversationsversuche, Interaktionsversuche, etc.:

- **Hörversuche, Sehversuche**: Hierbei bekommt die Versuchsperson Stimuli auditiv oder visuell dargeboten und beurteilt das interessierende Stimulusmerkmal.
- **Hör- und Sehversuche**: Hierbei werden den Versuchspersonen audio-visuelle Stimuli angeboten, die anschließend beurteilt werden sollen.
- **Sprechversuche**: Hierbei wird die Versuchsperson aufgefordert, selbst zu sprechen (z.B. einer anderen Person ins Wort zu fallen oder gegen ein Geräusch anzusprechen), und anschließend die Sprecherfahrung und Merkmale des dabei wahrgenommenen Hörereignisses zu beurteilen.
- **Hör- und Sprechversuche**: Hierbei werden die Versuchspersonen wiederum in eine kontrollierte Situation gebracht, in der sie Sprechen und Hören müssen, um anschließend die dabei gemachte Erfahrung zu beurteilen.
- **Konversationsversuche**: Hierbei werden zwei oder mehr Versuchspersonen in eine Konversationssituation gebracht, entweder in kontrollierter Form (z.B. durch Verteilen einer Aufgabe, eines sog. Konversationsszenarios) oder in einer freien Konversation. Nach Abschluss der Konversation sollen dann Aspekte der Konversation beurteilt werden. Die Konversation kann entweder direkt

(unvermittelt) ablaufen, oder über eine Sprach- oder audiovisuelle Verbindung (technikvermittelt).

- **Interaktionsversuche:** Hierbei interagieren die Versuchspersonen mit einem technischen System, im Sinne einer Mensch-Maschine-Interaktion. Die Interaktionen können wiederum szenariobasiert oder frei sein, und es können verschiedene Aspekte der Interaktion beurteilt werden.
4. Nach der **Mittelbarkeit** der Messung: Direkte vs. indirekte Messungen (Blauert, 1994).
- **Indirekte Messungen:** Diese bestimmen zunächst Schwellen bzw. Punkte gleicher Wahrnehmung. Sie lassen allerdings ohne zusätzliche Annahmen keine Punkt-zu-Punkt-Zuordnung zwischen den physikalischen und den Wahrnehmungsereignissen zu. Dieses Verfahren wird bei Jekosch (2000) als mittelbare Messung bezeichnet.
  - **Direkte Messungen:** Dabei werden direkte – also unmittelbare – Zuordnungen zwischen physikalischen und Wahrnehmungs-Ereignisskalen gefordert. Jekosch (2000) bezeichnet dies als unmittelbare Messung.

Neben der Mittelbarkeit über Wahrnehmungsschwellen und Punkte gleicher Wahrnehmung lässt sich aber noch eine weitere Mittelbarkeit in Betracht ziehen. Beispielsweise ist es denkbar, dass die Versuchsperson kein direktes Urteil über den Wahrnehmungsgegenstand fällt, sondern dass nach (instrumentell) messbaren Korrelaten des Wahrnehmungsereignisses gesucht wird. Die Korrelate könnten z.B. physiologische Parameter wie die Pulsfrequenz, der Blutdruck, die Atmungsfrequenz oder der Widerstand der Hautoberfläche sein, oder (z.B. bei Konversationsversuchen) kann als Korrelat der Erfolg einer Haupt- oder Nebenaufgabe verwendet werden, vgl. Abschnitt 2.3.

In der sogenannten „klassischen Psychophysik“ kommen meist Messungen auf Nominal- bzw. Ordinalniveau zum Einsatz. Damit können z.B. Wahrnehmbarkeitsschwellen, Unterschiedsschwellen oder Punkte gleicher Wahrnehmung bestimmt werden.

Wahrnehmbarkeitsschwellen werden meist so definiert, dass 50% der Versuchspersonen sagen, das betreffende Merkmal sei da (wahrnehmbar), und die andere Hälfte sagt, es sei nicht da (nicht wahrnehmbar); das Urteil hat also Nominalniveau. Unterschiedsschwellen können entweder auf Nominalniveau bestimmt werden (50% der Versuchspersonen sagen, die Stimuli sind gleich, 50% sagen, sie sind ungleich) oder als Ordinalurteile (Punkt A, bei dem 75% sagen, Stimulus 1 sei größer und Stimulus 2 sei kleiner, vs. Punkt B, bei dem 75% sagen, Stimulus 2 sei größer und Stimulus 1 sei kleiner; die Unterschiedsschwelle befindet sich in der Mitte zwischen diesen beiden Punkten). Punkte gleicher Wahrnehmung werden auf Ordinalniveau bestimmt: 50% der Versuchspersonen sagen, die Merkmalsausprägung ist größer bei Stimulus 1, 50% sagen, sie sei größer bei Stimulus 2.

Die genannten Verfahren können als Herstellungs- oder Konstanzmethoden implementiert werden. Bei der Herstellungsmethode ist zu beachten, dass die Richtung, aus der eingeregelt wird, das Ergebnis beeinflussen kann. Deshalb lässt man



das Merkmal meist in beide Richtungen variieren und bildet einen Mittelwert aus beiden Ergebnissen. Bei der Konstanzmethode ist zu beachten, dass die Reihenfolge, mit der die Stimuli vorgespielt werden, das Ergebnis beeinflussen kann. Man verwendet deshalb entweder eine randomisierte Präsentationsreihenfolge, oder die Reihenfolge der Stimuli im Test wird so ausbalanciert, dass alle Einflussfaktoren (z.B. Sprach- oder Bildmaterial, Übertragungsweg, Umgebungssituation, etc.) in möglichst allen Positionen vorkommen, vgl. den folgenden Abschnitt.

## 2.6 Versuchsplanung und Versuchsdesign

Möchte man die Qualität und Gebrauchstauglichkeit eines informations- oder kommunikationstechnischen Systems quantitativ bestimmen, so ist die Durchführung eines Versuches mit menschlichen Versuchspersonen meist unumgänglich. In den folgenden Absätzen sollen deshalb einige praktische Hinweise zur Planung und Durchführung solcher Versuche gegeben werden. Allerdings kann das Thema hier nicht erschöpfend behandelt werden; zu einzelnen Themen gibt es umfangreiche Literatur, in der Details beschrieben sind.

Vor der Planung eines Versuches sollte zunächst das **Ziel der Messung** genau festgelegt werden. Dabei reicht es nicht aus, dass man allgemein „die Qualität eines Systems“ messen möchte. Eine erste Einschränkung ergibt sich meist schon aus der Anwendungssituation. Man unterscheidet hier in der englischen Literatur zwischen (Jekosch, 2000; Hirschman und Thompson, 1997):

- *Assessment* (oder auch *Performance Evaluation*): Messung der Leistungen des Systems oder einzelner Komponenten bezüglich eines oder mehrerer festgelegter Kriterien. Wird verwendet, wenn man unterschiedliche Implementierungen eines Systems (oder einzelner Komponenten) miteinander vergleichen möchte.
- *Evaluation* (oder auch *Adequacy Evaluation*): Untersucht, ob ein System die Anforderungen eines bestimmten Nutzungskontextes erfüllt. Hierbei wird typischerweise mit zukünftigen Nutzern in realistischen Situationen getestet.
- *Diagnosis* (oder auch *Diagnostic Evaluation*): Hierbei werden die Leistungen eines Systems diagnostisch und systematisch bezüglich eines Profils erfasst. Ziel ist es, Systemeigenheiten und Probleme aufzudecken und ihre Ursachen zu ergründen.

Neben dem Messziel muss auch das **Messobjekt** genau spezifiziert werden. Dabei kann es sich um ein interaktives System oder ein Übertragungssystem handeln, einzelne Komponenten solcher Systeme, oder auch Kombinationen solcher Systeme (bspw. eine Telefonauskunft umfasst meist eine Datenbank, ein Dialogsystem und ein Übertragungssystem). Während der Systementwicklung sind meist noch nicht alle Komponenten verfügbar; in diesem Fall kann man sich mit Ersatzkomponenten, Simulationen oder sog. Wizard-of-Oz-Systemen (vgl. Kapitel 7) behelfen, muss allerdings die Abweichungen vom „realen“ Fall in Kauf nehmen und bei der Versuchsplanung und der Analyse der Ergebnisse berücksichtigen. Auch ist von Be-

deutung, ob das System in „Echtzeit“ (*online*) verfügbar sein muss, oder ob Offline-Varianten ausreichen; letztere sind meist einfacher zu erhalten, schränken aber die Möglichkeit der Interaktion beim Test weitgehend ein.

Sobald das Messziel und das Messobjekt feststehen kann man die interessierenden **Messgrößen** festlegen. Hierbei können Taxonomien von Leistungs- und Qualitätsaspekten, wie sie in Abschnitt 1.4 vorgestellt wurden, helfen. Die Taxonomien zeigen auch Einflussfaktoren (*Quality Factors*), welche bei der Auswahl der Messmethode berücksichtigt werden sollten. Die Messgrößen hängen darüber hinaus von der Zugänglichkeit des Systems sowie seiner Komponenten ab. Man unterscheidet hier i. Allg. zwischen sog. **Glass-Box-Tests**, bei denen das Innenleben des Systems bekannt und (zumindest teilweise) zugänglich ist, und **Black-Box-Tests**, bei denen das Innenleben unbekannt und/oder unzugänglich ist.

Ein besonderer Einflussfaktor ist die **Messumgebung**. So muss in Abhängigkeit vom Messziel, von der Messgröße und von den äußeren Rahmenbedingungen zunächst entschieden werden, ob eine Messung im Labor oder in Feld durchgeführt werden soll. Laborversuche zeichnen sich i. Allg. durch eine bessere Kontrollierbarkeit der Versuchsbedingungen aus, was die Messung im Prinzip zuverlässiger (Kriterium Reliabilität) macht. Allerdings ist die Motivation und u.U. auch das Verhalten der Versuchspersonen nicht realistisch, was das Messziel (Kriterium Validität) verfälschen kann. Einige Qualitätsaspekte (z.B. die Akzeptanz) lassen sich überhaupt nur im realen Anwendungszusammenhang (d.h. im Feld) messen.

Auf Basis dieser umfangreichen Analyse kann nun eine **Messmethode** ausgewählt werden. Hierzu können die Kriterien aus Abschnitt 2.5 zu Rate gezogen werden, wie auch die Taxonomien aus Abschnitt 1.4. Gern verwendet man instrumentelle Messungen, allerdings lassen sich damit allenfalls Leistungsindikatoren erfassen, keine Qualitätsaspekte. Aus diesem Grunde werden häufig Kombinationen unterschiedlicher Methoden im gleichen Versuch parallel angewandt. Bspw. kann man während eines Interaktionsversuches mit einem Sprachdialogsystem den Nutzer mit einem System interagieren lassen, und ihn nach jeder Interaktion mittels eines Fragebogens über die wahrgenommenen Qualitätsaspekte befragen. Parallel kann man das Benutzerverhalten aufzeichnen (Audio- Video-, Logdateien), und daraus quantitative Indices für sein Verhalten bestimmen, welche – u.U. nach erfolgter Transkription und Annotation – mit einzelnen Systemleistungen in Verbindung gebracht werden können (bspw. Wortfehlerrate). Aus der Vielzahl der so erhaltenen Resultate lässt sich ein relativ genaues diagnostisches Profil des Systems ableiten, welches sowohl zur Systemoptimierung verwendet werden kann als auch ein detailliertes Bild der vom Benutzer erfahrenen Qualität zeichnet.

Je nach ausgewählter Messmethode müssen Details des Versuches genauer bestimmt werden. Entscheidet man sich bspw. für einen Hörversuch, um die Qualität eines Übertragungssystems zu bestimmen, so müssen zunächst Stimuli ausgewählt werden, welche repräsentativ für das System (Messobjekt) sind und die realen Nutzungssituationen (bspw. unterschiedliche Sprecher) widerspiegeln. Die Anzahl der Stimuli wird durch die Anzahl der zu testenden Systemkonfigurationen bestimmt, wobei allerdings jede Konfiguration mit unterschiedlichem Sprachmaterial getestet werden sollte, da letzteres einen Einfluss auf die Übertragungsqualität haben wird.

Auch müssen die Versuchspersonen (Messorgane) ausgewählt werden, z.B. nach den in Abschnitt 2.4 aufgelisteten Kriterien. Je nach erforderlicher Teststärke (statistische Signifikanz der Ergebnisse) muss die Anzahl der notwendigen Versuchspersonen bestimmt werden. Diese ist jedoch meist stark durch den möglichen Aufwand begrenzt; Tests mit Versuchspersonen sind teuer und zeitaufwändig, sodass meist nur eine minimale Anzahl von ihnen zum Test eingeladen wird.

Der genaue **Ablauf des Versuches** muss geplant werden. Er besteht normalerweise aus vier bis fünf Phasen:

- **Vorbereitung:** Umfasst die Vorbereitung für jede einzelne Versuchsperson, Akquise der Versuchspersonen, Versuchsleiter, Testräume, Bereitstellung von Fragebögen, etc.
- **Einführung:** Begrüßung und Instruktion der Versuchsperson(en), Aufklärung über mögliche Risiken, Verwendung der Daten, Abbruchmöglichkeiten, etc.
- Optionale **Trainingsphase:** Wenn ein Training des Benutzers mit dem betrachteten System oder Testaufbau für notwendig oder wünschenswert erachtet wird, so kann dies zu Beginn des Tests oder auch zwischen einzelnen Testphasen eingebaut werden.
- **Durchführung** des eigentlichen Tests, z.B. Präsentation und Bewertung der Stimuli, Durchführung der Interaktionen, etc.
- **Testabschluss:** Abfrage von auf den gesamten Test bezogenen Urteilen, Interview, Nachfragen zu einzelnen Ereignissen des Testablaufes, Sicherung der Ergebnisse, etc.

Jede dieser Phasen muss genau vorbereitet und geplant werden.

So muss z.B. die Aufteilung der Testobjekte (Stimuli, Systemkonfigurationen, Testaufgaben) auf die Versuchspersonen im **Versuchsplan** möglichst so vorgenommen werden, dass keine systematischen Beeinflussungen entstehen. Dies kann im Rahmen eines *Between-Subjects* oder eines *Within-Subjects Design* geschehen. Bei *Between Subjects* testet jede Versuchsperson nur ein System (eine Systemvariante, eine Gruppe von Varianten oder Stimuli). Für jedes System (Variante, Stimulus, etc.) wird eine neue Gruppe von Versuchspersonen benötigt. Die Zuordnung der Systeme zu den Versuchspersonen sollte zufällig oder balanciert bezüglich aller möglichen Einflussfaktoren der Versuchspersonen erfolgen. Bei *Within Subjects* testet jede Versuchsperson alle zur Verfügung stehenden Systeme (Varianten, Stimuli). Dies hat den Vorteil, dass sich individuelle Unterschiede zwischen den Versuchspersonen herausmitteln. Allerdings ist die Versuchsperson bei wiederholten Tests nicht mehr „naiv“ in dem Sinne, dass sie eine Erfahrung mit dem Testgegenstand und -ablauf aufbaut. Daher muss die Reihenfolge der Präsentation variiert werden, um diesen Einflussfaktor herauszumitteln. Detailliertere Hinweise zur Gestaltung von Versuchsplänen findet man z.B. bei Bortz (2005) und bei Bortz und Döring (2002).

Darüber hinaus gibt es weitere Einflussfaktoren, die ebenfalls möglichst herausgemittelt werden sollen. So ist bei einem Interaktionstest bspw. die Versuchsaufgabe entscheidend, oder bei einem Hörversuch das zur Übertragung verwendete Sprachmaterial. Um keine zu starke Erfahrung aufzubauen und eine gewisse Allgemeingültigkeit zu erlangen, verwendet man üblicherweise **unterschiedliche** Auf-

gaben und Sprachmaterialien. Die Aufteilung der Aufgaben / Sprachmaterialien zu den einzelnen Systemkonfigurationen sollte zwischen den Versuchspersonen variieren, ebenso wie ihre Reihenfolge im Test. Wenn die Anzahl der Stimuli und Einflussfaktoren gering ist kann man versuchen, ein *Full-Factorial Design* umzusetzen, bei dem alle Einflussfaktoren (inkl. Position im Test) miteinander kombiniert werden. Man geht davon aus, dass sich diese Einflüsse dann herausmitteln. Abb. 2.8 zeigt ein sogenanntes griechisch-lateinisches Quadrat, was hierzu verwendet werden kann. Wo dies nicht möglich ist kann man ein *Partial-Factorial Design* verwenden, bei denen nur einzelne Faktoren komplett kombiniert werden; andere Faktoren werden nicht betrachtet, oder nur vereinfacht (z.B. in Gruppen) gemittelt.

	1	2	3	4	5	6	7	8	9	10	11	12
I	A $\alpha$	B $\mu$	C $\zeta$	D $\eta$	I $\varepsilon$	J $\delta$	K $\kappa$	L $\lambda$	E $\iota$	F $\theta$	G $\beta$	H $\gamma$
II	B $\beta$	A $\lambda$	D $\varepsilon$	C $\theta$	J $\zeta$	I $\gamma$	L $\iota$	K $\mu$	F $\kappa$	E $\eta$	H $\alpha$	G $\delta$
III	C $\gamma$	D $\kappa$	A $\theta$	B $\varepsilon$	K $\eta$	L $\beta$	I $\mu$	J $\iota$	G $\lambda$	H $\zeta$	E $\delta$	F $\alpha$
IV	D $\delta$	C $\iota$	B $\eta$	A $\zeta$	L $\theta$	K $\alpha$	J $\lambda$	I $\kappa$	H $\mu$	G $\varepsilon$	F $\gamma$	E $\beta$
V	E $\varepsilon$	F $\delta$	G $\kappa$	H $\lambda$	A $\iota$	B $\theta$	C $\beta$	D $\gamma$	I $\alpha$	J $\mu$	K $\zeta$	L $\eta$
VI	F $\zeta$	E $\gamma$	H $\iota$	G $\mu$	B $\kappa$	A $\eta$	D $\alpha$	C $\delta$	J $\beta$	I $\lambda$	L $\varepsilon$	K $\theta$
VII	G $\eta$	H $\beta$	E $\mu$	F $\iota$	C $\lambda$	D $\zeta$	A $\delta$	B $\alpha$	K $\gamma$	L $\kappa$	I $\theta$	J $\varepsilon$
VIII	H $\theta$	G $\alpha$	F $\lambda$	E $\kappa$	D $\mu$	C $\varepsilon$	B $\gamma$	A $\beta$	L $\delta$	K $\iota$	J $\eta$	I $\zeta$
IX	I $\iota$	J $\theta$	K $\beta$	L $\gamma$	E $\alpha$	F $\mu$	G $\zeta$	H $\eta$	A $\varepsilon$	B $\delta$	C $\kappa$	D $\lambda$
X	J $\kappa$	I $\eta$	L $\alpha$	K $\delta$	F $\beta$	E $\lambda$	H $\varepsilon$	G $\theta$	B $\zeta$	A $\gamma$	D $\iota$	C $\mu$
XI	K $\lambda$	L $\zeta$	I $\delta$	J $\alpha$	G $\gamma$	H $\kappa$	E $\theta$	F $\varepsilon$	C $\eta$	D $\beta$	A $\mu$	B $\iota$
XII	L $\mu$	K $\varepsilon$	J $\gamma$	I $\beta$	H $\delta$	G $\iota$	F $\eta$	E $\zeta$	D $\theta$	C $\alpha$	B $\lambda$	A $\kappa$

**Abb. 2.8** Griechisch-lateinisches Quadrat der Ordnung 12 zum Design eines Hörversuches. Römische Ziffern: Versuchsperson; arabische Ziffern: Position innerhalb des Versuches; lateinische Buchstaben: System (-konfiguration); griechische Buchstaben: Sprachmaterial

Um eine gleichmäßige Instruktion aller Versuchspersonen zu gewährleisten, sollten schriftliche Anleitungen verwendet werden. Darüber hinaus mag aber auch eine mündliche Instruktion notwendig sein. Die Instruktion sollte darauf zielen, dass den Versuchspersonen die Aufgabenstellung vollständig klar ist, ohne sie allerdings in ihrem Verhalten oder ihrem Urteil zu beeinflussen (außerhalb dedizierter Trainingsphasen).

Die im Versuch gesammelten Ergebnisse müssen anschließend analysiert werden. Hierzu sind verschiedene statistische Verfahren verfügbar. Einen kurzen Über-

blick über die statistische Auswertung skaliertter Urteile oder Indices findet sich in Abschnitt 3.6. Nicht skalierte Daten werden meist aggregiert (z.B. durch manuelle Klassifikation) und dann bzgl. ihrer Häufigkeit und Wichtigkeit für das Evaluationsziel ausgewertet. So können bspw. die Ergebnisse eines strukturierten Interviews anhand der Interviewfragen zu Klassen ähnlicher Versuchspersonen-Aussagen zusammengefasst und mit einer Nennungshäufigkeit versehen werden. Die am häufigsten genannten Aspekte werden dann als besonders auffällig oder als besonders wichtig eingestuft, und daraus Konsequenzen abgeleitet.

## Literaturverzeichnis

- Blauert J (1994) Kommunikationsakustik 2. Skriptum (unveröffentlicht) zur Vorlesung and der Ruhr-Universität, Bochum
- Blauert J (1997) Spatial Hearing: The Psychophysics of Human Sound Localization. The MIT Press, Cambridge MA
- Bortz J (2005) Statistik für Sozialwissenschaftler. Springer, Berlin
- Bortz J, Döring N (2002) Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler. Springer, Heidelberg
- Chateau N, Gros L, Durin V, Macé A (2006) Redrawing the link between customer satisfaction and speech quality. In: Möller S, Raake A, Jekosch U, Hanisch M (Hrsg) Proc. 2nd ISCA/DEGA Tutorial and Research Workshop on Perceptual Quality of Systems, Berlin, S 88–94
- DIN 1319 Teil 1 (1995) Grundlagen der Meßtechnik. Teil 1: Grundbegriffe. Deutsches Institut für Normung, Beuth Verlag, Berlin
- Hirschman L, Thompson HS (1997) Survey of the State of the Art in Human Language Technology, Cambridge University Press and Giardini Editori, Pisa, Kapitel Overview of Evaluation in Speech and Natural Language Processing, S 409–414
- Jekosch U (2000) Sprache hören und beurteilen: Ein Ansatz zur Grundlegung der Sprachqualitätsbeurteilung. Habilitationsschrift (unveröffentlicht), Universität/Gesamthochschule, Essen
- Lienert GA (1989) Testaufbau und Testanalyse. Verlag Julius Beltz, Weinheim
- Lorenz K (1963) Das sogenannte Böse. Borotha-Schoeler, Wien
- Naumann AB, Hermann F, Peissner M, Henke K (2008) Interaktion mit Informations- und Kommunikationstechnologie: Eine Klassifikation von Benutzertypen. In: Herczeg M, Kindschmüller MC (Hrsg) Mensch & Computer 2008: Viel Mehr Interaktion, Oldenbourg Wissenschaftsverlag, München, S 37–45
- Nielsen J (1993) Usability Engineering. Academic Press, Boston MA
- Raake A (2006) Speech Quality of VoIP: Assessment and Prediction. John Wiley & Sons Ltd., Chichester, West Sussex
- Rogers EM (1983) Diffusion of Innovations. Free Press, New York NY

Quality Engineering

Qualität kommunikationstechnischer Systeme

Möller, S.

2017, XI, 200 S. 87 Abb. Book + eBook., Hardcover

ISBN: 978-3-662-56045-7