

Kapitel 2

Messtheorie

Measurement began our might.

W. B. Yeats, zitiert nach Duncan (1984: xiii)

2.1 Formalisierung des Messvorgangs

Die klassische Messtheorie, eingeführt von Stevens (1946, 1951) und zur Blüte gebracht durch Krantz, Luce, Suppes und Tversky,¹ geht von einem einfachen mathematischen Modell des Messvorgangs aus. Eine Menge W , welche die Realität (Welt) repräsentiert, wird durch eine feststehende Vorschrift s auf eine Menge Z von Zahlen abgebildet. Die Vorstellung ist, dass dabei einem realen Objekt $a \in W$ eine Zahl $z = s(a)$, sein Messwert, zugeordnet wird.² Mathematisch gesehen ist s nichts anderes als eine *Abbildung* der Menge W in Z . Empirisch gesehen entspricht s einer Messvorschrift, welche determiniert, was wie gemessen wird. Das Ergebnis einer Messung ist ein numerischer Wert z , welcher via s auf der Realität (also a) basiert. Deshalb charakterisiert Stevens jede Messung mit den Worten “[...] the assignment of numerals to things so as to represent facts [...] about them” (siehe S. 14).

Eine gute Messung sollte auf W existierende (reale) Strukturen respektieren, also geeignet in die Menge der Zahlen transportieren. Man spricht auch von der repräsentationalen Theorie des Messens, da so in Z ein Abbild der realen Verhältnisse entsteht. Diese Idee formalisiert die Messtheorie mithilfe von Relationen: Man betrachte hierzu zunächst die Menge Z . Eine bekannte Relation ist „kleiner“, also „ $<$ “. Genauer gesagt handelt es sich bei „ $<$ “ um eine zweistellige Relation auf der Menge der reellen Zahlen. Der Ausdruck $x < y$ verwendet nämlich genau zwei reellen Zahlen x und y und lässt sich für beliebige reelle Zahlen x, y interpretieren. D.h., man kann immer sagen, ob $x < y$ gilt oder nicht.³

¹ Siehe insbesondere Krantz et al. (1971), Suppes et al. (1989) und Luce et al. (1990)

² Ganz ähnlich heißt es bei Menger (1955: 167): “Thus quantity = (object, value).”

³ Allgemein ist eine zweistellige Relation zwischen zwei Mengen A und B nichts anderes als eine Teilmenge des cartesischen Produktes $A \times B$. Wichtig ist dabei nur, dass man für jedes Element $(a, b) \in A \times B$ zweifelsfrei sagen kann, ob für dieses die Relation besteht oder nicht. Im Beispiel ist $A \times B$ gerade $\mathbb{R} \times \mathbb{R}$, also die Ebene $\mathbb{R}^2$.

Wie lässt sich die Relation „ $<$ “ charakterisieren? Offensichtlich dadurch, dass man sämtliche möglichen Tupel $(x, y) \in \mathbb{R}^2$, also jeden Punkt in der Ebene, betrachtet und aufschreibt, ob die Beziehung $x < y$ für dieses Zahlenpaar gerade gilt oder nicht. Es genügt sogar die Menge M all jener Punkte zu betrachten, wo die Relation gilt,⁴ Im gewählten Beispiel ist M einfach die Fläche *oberhalb* der Winkelhalbierenden.⁵ Der Schritt von zweistelligen zu i -stelligen Relationen ist offensichtlich. Auch eine solche Relation lässt sich völlig analog charakterisieren durch die Menge $M \subseteq \mathbb{R}^i$, auf der sie gilt.⁶

Auf Mengen von Zahlen Relationen einzuführen, stellt also kein Problem dar. Doch wie sieht es mit der realen Welt aus? Die Messtheorie betrachtet auch auf W Relationen, die per konstruktionem⁷ als *empirische* Beziehungen zu interpretieren sind. Typischerweise stellt man sich die Elemente $a \in W$ als Objekte vor, die in einem gewissen Verhältnis zu einander stehen. Das klassische Beispiel ist die empirische Größerbeziehung, welche gerne mit dem Zeichen „ $<$ “ abgekürzt wird. Formal gesehen lassen sich wieder alle Vergleiche $a < b$ bilden und die Menge $M_{<}$ charakterisiert alle Tupel, für die die Relation gilt, also z. B. die Körpergröße von Person a kleiner als jene von Person b ist.

Aus naheliegenden Gründen nennt man die Menge W samt der dort definierten Relation „ $<$ “ *empirischen Relativ* E , während man die Menge Z mit der dort definierten Relation „ $<$ “ als *numerischen Relativ* N bezeichnet. Sind E und N gegeben, so wird eine Abbildung $s : E \rightarrow N$ als Skala aufgefasst, wenn sie die auf den beiden Mengen definierten Relationen respektiert. Das heißt, s ist eine Skala, wenn gilt:

$$\begin{aligned} a, b \in W \text{ stehen in Relation } R_W(a, b) \\ \iff \\ s(a), s(b) \in \mathbb{R} \text{ stehen in Relation } R_Z(s(a), s(b)). \end{aligned} \tag{2.1}$$

Dabei deuten die Subskripte W bzw. Z an, wo die jeweilige Relation R_W bzw. R_Z definiert ist. Im Beispiel gilt die empirische Relation $a < b$ also genau dann, wenn auch die numerische Relation $s(a) < s(b)$ wahr ist. Mathematisch gesprochen ist s ein Homomorphismus.⁸

Man beachte, dass a und b reale Objekte sind, die in einer empirischen Beziehung, z. B. $a < b$, zueinander stehen. Die entscheidende Anforderung

⁴ Da M und ihr Komplement M^c (also die Menge aller Tupel, für die die Relation nicht gilt), disjunkt sind und ihre Vereinigung gleich $\mathbb{R}^2$ ist. „Disjunkt“ bedeutet, dass die beiden Mengen keine gemeinsamen Punkte haben, sich bildlich gesprochen nicht überschneiden.

⁵ Diese Gerade wird durch $x = y$, also alle Punkte der Gestalt (x, x) definiert. Bewegt man sich oberhalb von ihr, so heißt das, dass die zweite Komponente größer als die erste ist ($y > x$). Anders gesagt: Die Relation „ $<$ “ ist wahr für alle Punkte (x, y) mit $x < y$.

⁶ Allgemein ist eine i -stellige Relation $R(i)$ der Mengen $A_1, \dots, A_i$ eine Teilmenge des cartesischen Produktes $A_1 \times \dots \times A_i$, so dass sich für jedes Element $(a_1, \dots, a_i) \in A_1 \times \dots \times A_i$ sagen lässt, ob die Relation besteht oder nicht.

⁷ Denn genau so wurde das Modell ja gemacht.

⁸ Man konsultiere Pfanzagl (1968: 6ff) für Details.

an jede Skala ist also, dass dann auch die Messwerte $s(a)$ und $s(b)$ in Relation zueinander stehen, also $s(a) < s(b)$ gilt. Die Messung bzw. der Messvorgang ist insofern zuverlässig, als dass er in der realen Welt existierende Relationen korrekt wiedergibt. Ist das Objekt a „tatsächlich“ kleiner als das Objekt b , so spiegelt sich dies auch in den ihnen zugeordneten Messwerten wider. Aufgrund der Äquivalenz in (2.1), gilt hiervon auch die *Umkehrung*. Aus der Tatsache, dass der beobachtete Wert von a numerisch kleiner als der beobachtete Wert von b ist, lässt sich folgern, dass in der realen Welt $a \prec b$ der Fall ist. Man kann also - aufgrund dieser *Forderung* an Messvorgänge - bequem und zuverlässig von den Messergebnissen auf die reale Welt zurück schließen!⁹

In voller Allgemeinheit ist bzw. steht nun der empirische Relativ $E = (W, \Sigma)$ für einen Teil der Wirklichkeit, oft eine Menge von Objekten W , versehen mit einer Struktur Σ , modelliert als eine gewissen Anzahl j von i -stelligen Relationen. Völlig analog ist der numerische Relativ $N = (Z, \sigma)$ nichts anderes als eine Menge von Zahlen Z versehen mit derselben Anzahl i -stelliger Relationen. Bei gegebenen Strukturen Σ und σ werde die Menge aller Skalen mit $S = S(\Sigma, \sigma)$ bezeichnet.

Da jede zusätzliche Relation Anforderungen stellt, wird S umso kleiner, je mehr Relationen in Betracht gezogen werden, d.h. mit wachsendem j schrumpft S . In vielen praktisch relevanten Fällen lässt sich jedoch zeigen, dass S nicht leer ist. Narens (1985: 5) schreibt zusammenfassend:

The view that measurement consists in specifying homomorphisms of some qualitative (or empirical) structure into a numerical one is called *representational theory of measurement*, and since the late 1950s it has gained wide support among measurement theorists [...]. The *existence theorem* for S consists in showing $S \neq \emptyset$. The *uniqueness theorem* for S consists in describing how the elements in S are related to one another. (Hervorhebungen im Original.)

2.2 Äquivalente Messungen

Dem zuletzt genannte Problem wenden wir uns nun zu: Betrachtet man empirische Messungen, so stellt man nämlich nahezu unmittelbar fest, dass ein und dasselbe empirische Merkmal ganz verschieden gemessen werden kann. Zum Beispiel wird die Temperatur in Europa typischerweise in °C gemessen, während man in Nordamerika °F benutzt und Physiker gerne °K verwenden. Entfernungen werden standardmäßig in Metern, oft aber auch in cm, mm oder km angegeben. Darüber hinaus sind auch Inches, Feet, Meilen usw. gebräuchlich. Diese Unterschiede können eher oberflächlicher Natur sein, denn bei der Verwendung eines Lineals werden Entfernungen in cm oder mm auf ein und dieselbe Weise bestimmt. Hinter verschiedenen Einheiten - gerne auch „Dimensionen“ genannt - können sich jedoch auch völlig verschiedene

⁹ Dies stellt natürlich eine sehr starke Forderung dar, die in der Praxis nicht immer erfüllt sein wird. Man denke nur an unvermeidliche Messungenauigkeiten.

Messtechniken verbergen. Es ist ein fundamentaler Unterschied, ob eine Entfernung mithilfe der Laufzeit eines Signals ermittelt wird oder aber ob man einen Zollstock im Wortsinne als Maß-Stab „anlegt“. Leider müssen sich völlig verschiedene Messmethoden nicht in der Einheit niederschlagen. Abgesehen von wissenschaftlichen Arbeiten wird nicht ausgewiesen, wie z. B. eine Entfernung tatsächlich bestimmt worden ist.

Es bleibt also nichts anderes übrig, als zum einen verbal exakt zu beschreiben, auf welche Art und Weise eine Größe bestimmt worden ist. Dies könnte man als konkrete empirische Messtechnik bezeichnen. Zum anderen sollte jede mathematische Theorie des Messens in der Lage sein, mit einem Wechsel des Maßstabs umzugehen. Oder, anders gesagt, es ist eine Mindestanforderung an jede Theorie des Messens, dass sie eine Änderung der Einheit adäquat berücksichtigt.

Mathematisch gesehen führt dies zu sogenannten Transformationen $\gamma : N \rightarrow N'$, die also einen Messwert $s \in N$ in einen transformierten Wert $s' = \gamma(s)$ überführen. Bei N' handelt es sich typischerweise ebenfalls um die Menge der reellen Zahlen, versehen mit denselben Relationen wie zuvor. Eine Transformation γ heißt zulässig, wenn $\gamma(s)$ ebenfalls wieder eine Skala ist, also die Beziehung (2.1) einhält. Ein Beispiel: Gilt $a \prec b$ für beliebige $a, b \in W$, so muss auch für die Messwerte $s(a) < s(b)$ gelten. Ist γ zulässig, so gilt darüber hinaus $s'(a) = \gamma(s(a)) < \gamma(s(b)) = s'(b)$. Zudem folgt bei einer zulässigen Transformation aus $s(a) < s'(b)$, dass in der Empirie $a \prec b$ gilt. Insgesamt hat man also die Beziehung

$$a \prec b \Leftrightarrow s(a) < s(b) \Leftrightarrow \gamma(s(a)) < \gamma(s(b)) \Leftrightarrow s'(a) < s'(b), \quad (2.2)$$

die prägnant zum Ausdruck bringt, dass die beiden Skalen s und s' völlig äquivalent sind. (2.2) zeigt auch, dass der vermeintliche Unterschied zwischen s und s' , nämlich s als „grundlegender“ Skala und $s' = \gamma(s)$ als „abgeleiteter“ Skala nicht wirklich existiert: Die empirische Beziehung $a \prec b$ gilt genau dann, wenn eine der beiden Skalen dies anzeigt. Es ist also völlig beliebig, welche der beiden Skalen man verwendet.

Um festzustellen, ob ein Stab länger ist als ein anderer, genügt es also, die beiden Stäbe nur einmal - z. B. mithilfe der Einheit cm - zu messen. (Dies ist die erste Äquivalenz.) Gemäß der obigen Theorie ist es auch völlig gleichwertig, das in cm gemessene Ergebnis in m umzurechnen (also eine zulässige Transformation γ anzuwenden, dies ist die zweite Äquivalenz) oder aber die beiden Stäbe unmittelbar mithilfe der Einheit m zu messen. (Dies ist die dritte Äquivalenz.)

Es sollte aufgrund des eben Gesagten nun auch nicht weiter verwunderlich sein, dass man sogar von einer *beliebigen* Skala $s_0 \in S$ ausgehen kann und mithilfe einer geeigneten, zulässigen Transformationen γ jede andere Skala $s \in S$ in der Form $s = \gamma(s_0)$ darstellen kann (siehe z. B. Pfanzagl (1968: 26ff)). Dieser Sachverhalt wird gerne so formuliert, dass in einer gegebenen Situation die Skala bis auf eine zulässige Transformation eindeutig bestimmt

ist.¹⁰ Auch die Aufgabe, zwei Skalen $s, s' \in S$, welche denselben empirischen Sachverhalt E messen, zu vergleichen, ist damit, nämlich mittels des Werkzeugs der zulässigen Transformationen, vollständig beantwortet worden. Nicht zuletzt deshalb kommt den Transformationen eine zentrale Bedeutung innerhalb der Messtheorie zu.

Anders gesagt: Die Objekte in W und Z legen zusammen mit den Relationen die Menge aller Skalen $S = S(\Sigma, \sigma)$ fest. Es sind die „Außenbedingungen“, also $E = (W, \Sigma)$ und $N = (Z, \sigma)$, welche über die Menge aller „vernünftigen“ Skalen bestimmen. Alle Skalen $s \in S$ sind als gleichwertig zu betrachten. Da man zudem in der Wahl der Relationen auf Z ziemlich frei ist, kommt es entscheidend auf den Gegenstandsbereich W und vor allem auf die dort (real!) existierenden Relationen (die wir in der Struktur Σ zusammenfassen), an, welche Skalen äquivalent sind. Es ist maßgeblich der empirische Sachverhalt selbst, der bestimmt, wie „gut“ er vermessen werden kann. Suppes und Zinnes (1968: 9) geben einige Beispiele:

In general, any empirical procedure for measuring mass does not determine the unit of mass. The choice of unit is an empirically arbitrary decision made by an individual or group of individuals [...] The measurement of distance is a second example of measurement of this sort [...] An empirical procedure for measuring temperature by use of a thermometer determines neither a unit nor an origin [...] Examples other than measurement of temperature are provided by the usual measurements of temporal dates, linear position, or cardinal utility.

Es sollte hinzugefügt werden, dass zwar der empirische Relativ bestimmt, wie genau (im obigen Sinne) eine Größe *im Prinzip* gemessen werden kann. Da das numerische Ergebnis jedoch immer das Resultat einer Abbildung der Realität auf eine Menge von Zahlen ist, hat auch die konkrete Messmethode einen ganz erheblichen Einfluss auf das „Niveau“ der Messung (siehe den nächsten Abschnitt). Etwa zeigt das Beispiel der Temperatur, dass diese absolut in °K gemessen werden kann. In diesem Fall liegt der Nullpunkt der Messung sehr wohl fest. Und auch wenn die Geschwindigkeitsmessung im Prinzip bis auf die Wahl der Einheiten präzise möglich ist, so kann ein schlechtes Messverfahren, z. B. das Urteil von Passanten, die angeben sollen, welches von zwei vorbeifahrenden Autos schneller war, zu einem sehr unzuverlässigen Ergebnis führen.

2.3 Skalenhierarchie

Historisch gesehen wurde die nachfolgende Einteilung bereits von Stevens ab den 1940er Jahren vorgeschlagen. Die detaillierte mathematische Ausarbeitung folgte erst einige Jahre später. Sie spielt bis heute bei jeder Diskussion und Anwendung der Messtheorie eine zentrale Rolle.

¹⁰ Daher auch die Bezeichnung „uniqueness theorem“ bei Narens, siehe S. 25.

Nominal-, Ordinal-, Intervall-, Verhältnis- und Absolutskalen sind Stevens' wichtigste Beispiele für Skalen. Sie alle werden durch eine zugehörige Klasse zulässiger Transformationen beschrieben:¹¹

Skala	Zulässige Transformationen
Nominalskala	Injektive Abbildungen: $f(x) = f(x') \Rightarrow x = x'$
Ordinalskala	Monotone Abbildungen: $x < x' \Rightarrow g(x) < g(x')$
Intervallskala	Affin-lineare Abbildungen: $h(x) = ax + b$ mit $a \neq 0$
Verhältnisskala	Lineare Abbildungen: $h(x) = ax$ mit $a \neq 0$
Absolutskala	Identische Abbildung: $h(x) = x$

Da die weiter unten stehenden Transformationen jeweils Spezialfälle der weiter oben stehenden sind, hat man es mit einer Skalen-Hierarchie zu tun. Zudem enthalten tiefer stehende Skalen mehr Information über die reale Welt als weiter oben stehende. Einige typische, in der Literatur häufig genannte Beispiele sind:¹²

Bei Nominalskalen denkt man zumeist an Kategorien, in die eine Reihe von Objekten eingeordnet werden. Objekte in einer Kategorie werden als gleich betrachtet, Objekte verschiedener Kategorien als ungleich. Das heißt, die Relation zwischen den Klassen von Objekten ist „=“. Beispiel: Bei einer Erhebung geben die befragten Personen ihr Geschlecht an. Das heißt mathematisch, dass jeder Person ihr Geschlecht zugeordnet wird. Der empirische Relativ besteht aus den Personen zusammen mit der zweistelligen Relation „Geschlecht“. Diese ist genau dann für zwei Personen erfüllt, wenn jene dasselbe biologische Geschlecht besitzen.

Eine Skala ordnet nun jeder Person eine (völlig willkürliche) Zahl zu, wobei nur darauf zu achten ist, dass Personen unterschiedlichen Geschlechts auch weiterhin unterschieden werden können. D.h. allen Frauen ist (ein und dieselbe) Zahl x zuzuordnen und allen Männer eine andere Zahl $x' \neq x$. Zulässige Transformationen f müssen diese Unterscheidung aufrecht erhalten, d.h. $f(x)$ muss ungleich $f(x')$ sein. Genau dies leisten die oben genannten injektiven Transformationen.

Ordinalskalen enthalten mehr Information: Die Objekte oder eine Reihe von Klassen von Objekten sind geordnet. Man kann also entscheiden, welches Objekt (in irgendeinem Sinne) größer als ein anderes ist. Genau diese Ordnung bleibt bei der Messung erhalten, und genau diese Ordnung müssen zulässige Transformationen auch respektieren. Dies leisten die in der Tabel-

¹¹ Unsere Aufstellung orientiert sich an Duncan (1984: 123).

¹² Für die von Stevens von 1951-1975 angegebenen Beispielfälle siehe Duncan (1984: 124). Schleichert (1966) beschäftigt sich ebenfalls intensiv mit dem Messproblem und zeigt, dass Carnap ganz ähnlich argumentiert (siehe insbesondere S. 38 ibd.).

le angegebenen monotonen Transformationen g , da für diese $a \prec b \Leftrightarrow x = s(a) < x' = s(b) \Leftrightarrow g(x) < g(x')$ gilt.¹³

Bislang wurde nur qualitative Information bei der Messung übermittelt. Die Intervallskala ermöglicht es, quantitative Aussagen zu machen. Beispielsweise sind die Längengrade der Geodäsie intervallskaliert: Bei den Objekten handelt es sich um Orte auf der Erdoberfläche, z. B. Städte.¹⁴ Die Messung ordnet diesen ihren Längengrad zu, also z.B. $s(\text{London}) = 0$, $s(\text{Dresden}) = 13$ und $s(\text{Lissabon}) = -9$.¹⁵

Offenkundig ist der Nullpunkt der Messung das Ergebnis einer historischen Zufälligkeit. Würde man heute den Nullmeridian festlegen, so liefe er wohl eher durch einen markanten geographischen Ort oder fiel mit der Datumsgrenze zusammen. Formal gesehen heißt das: Verschiebt man die Längengrade um eine feste Zahl b , so ändern sich die Distanzen (gemessen in Längengraden) zwischen den Orten nicht. Sei zum Beispiel $b = 9$, dann hat man die neuen Positionen $s'(y) = h(s(y)) = s(y) + 9 = 9$ für London, $s'(z) = h(s(z)) = s(z) + 9 = 22$ für Dresden und $s'(x) = h(s(x)) = s(x) + 9 = 0$ für Lissabon. s' entspricht also einer Längengradskala, deren Nullmeridian durch Lissabon läuft. Die Abstände sind bei beiden Mess-Systemen dieselben, z. B. ist nach wie vor die Distanz d zwischen Lissabon und Dresden

$$d(x, z) = s(x) - s(z) = 13 - (-9) = s'(x) - s'(z) = 22 - 0 = 22$$

Längengrade. Auch eine Streckung um a , zum Beispiel $a = 2$, ist unabhängig von der Skala, wird also von der Lage der Orte auf der Erdoberfläche bestimmt. Denn $s^*(x) = 2 \cdot s(x) = -18$, $s^*(y) = 2 \cdot s(y) = 0$, $s^*(z) = 2 \cdot s(z) = 26$. Scheinbar hat sich die Distanz zwischen Dresden und Lissabon auf $26 - (-18) = 44$ Längengrade verdoppelt, da sich auch alle übrigen Abstände verdoppelt haben, bleibt die Gültigkeit der Aussagen erhalten. Zum Beispiel ergibt sich mit Skala s

$$d(x, y) + d(y, z) = d(x, z), \quad \text{da } 9 + 13 = 22.$$

Mit der Skala s^* kommt man auf

$$d(x, y) + d(y, z) = d(x, z), \quad \text{da } 18 + 26 = 44 \quad \text{bzw.} \quad 2 \cdot (9 + 13) = 2 \cdot 22.$$

Das heißt, die Beziehung bleibt weiterhin gültig. Anschaulich gesprochen überzieht man die Erde mit einem doppelt so feinen Netz von Längengraden.

¹³ Siehe (2.2). Man beachte, dass für monotone Abbildungen auch $g(x) < g(x') \Rightarrow x < x'$ gilt, da $x' < x \Rightarrow g(x') < g(x)$, was im Widerspruch zu $g(x) < g(x')$ steht. Zudem ist $x = x' \Leftrightarrow g(x) = g(x')$.

¹⁴ Falls diese aufgrund ihrer Ausdehnung nicht als punktförmig bezeichnet werden können, gehe man stattdessen vom Stadtzentrum aus.

¹⁵ Im folgenden schreiben wir x für Lissabon, y für London und z für Dresden.

Auch wenn sowohl a als auch b ungleich Null sind, ändert sich daran nichts. Das heißt, die Abstände zwischen den Städten hängen nicht wesentlich von der verwendeten Skala ab, können also inhaltlich interpretiert, bzw. auf die realen Sachverhalte zurückgeführt werden. Mathematisch drückt man dies so aus, dass für vier Orte $x, y, u, v \in W$ bzw. je zwei Distanzen $d(x, y), d(u, v)$ immer gilt:

$$\begin{aligned} d(x, y) = d(u, v) &\Leftrightarrow s(x) - s(y) = s(u) - s(v) \\ &\Leftrightarrow a(s(x) - s(y)) + 2b = a(s(u) - s(v)) + 2b \\ &\Leftrightarrow (as(x) + b) - (as(y) + b) = (as(u) + b) - (as(v) + b) \\ &\Leftrightarrow s^*(x) - s^*(y) = s^*(u) - s^*(v) \end{aligned}$$

Aus dieser Eigenschaft rührt der Name „Intervallskala“ her. Äquivalent dazu ist die Sprechweise, dass durch die Messung weder Einheit noch Nullpunkt festgelegt werden, sondern frei gewählt werden können.

Bei Verhältnisskalen ist schließlich nur noch die Einheit frei wählbar. Wie der Name schon andeutet, lassen sich nun sogar Quotienten von Messwerten empirisch interpretieren, da gilt

$$\frac{s(x)}{s(y)} = \frac{s(v)}{s(w)} \Leftrightarrow \frac{a \cdot s(x)}{a \cdot s(y)} = \frac{a \cdot s(v)}{a \cdot s(w)} \Leftrightarrow \frac{s^*(x)}{s^*(y)} = \frac{s^*(v)}{s^*(w)}.$$

Bei einer Absolutskala liegt alles fest, auch die Einheit, mit der gemessen wird. Ein typisches Beispiel ist eine Anzahl von Objekten oder aber auch der Kontostand (bei einer festen Währung).

Wie am Ende des letzten Abschnitts festgehalten wurde, hängt das Skalenniveau einer empirischen Messung sowohl vom Gegenstandsbereich als auch von der Art der Messung ab. Die Struktur des Objektbereichs legt fest, wie gut eine Größe im Idealfall gemessen werden kann. Es kommt jedoch auf das konkrete Verfahren an, ob sich dieses prinzipiell erreichbare Niveau auch tatsächlich einstellt. Misst man z. B. die Temperatur mit einem üblichen Thermometer, so bewegt man sich auf Intervallskalenniveau, obwohl bestenfalls eine Verhältnisskala möglich wäre. Unterschiede im Skalenniveau ergeben sich erst recht, wenn man die historische Entwicklung der Temperaturmessung betrachtet, wie Duncan (1984: 147ff) es tut.

Man beachte darüber hinaus, dass *jedes* Skalenniveau mit nicht trivialen Forderungen an die empirische Struktur verknüpft ist. Es ist z. B. keinesfalls selbstverständlich, dass eine Menge von Objekten samt einer empirischen Relation angeordnet ist. Zirkuläre Strukturen treten nicht nur bei Wahrnehmungstäuschungen häufig auf. Sehr oft hat man es z. B. auch mit lediglich partiell geordneten Gegenstandsbereichen zu tun. Solche Phänomene genügen, um eine Messung auf Ordinalskalenniveau unmöglich zu machen.

Die entscheidende Trennung bei den Skalenniveaus verläuft zwischen Ordinal- und Intervallskala (siehe die gestrichelte Linie in der obigen Tabelle). Sie markiert im hierarchischen System der Messtheorie den Unterschied zwi-

schen *qualitativen und quantitativen* Aussagen. Wissenschaftlicher sind in der Tradition Galileis quantitativ orientiert und man kann sich anhand einfacher Beispiele leicht klar machen, warum. Kennt man z. B. nur die Rangfolge der Sieger in einem sportlichen Wettbewerb (Schnellster, Zweitschnellster, usw.), so lässt sich zwar entscheiden, wem die Medaillen zustehen, wichtige Informationen aber fehlen: Handelte es sich bei der Siegerzeit um einen neuen Weltrekord? War der Erfolg unangefochten oder setzte sich der Sieger nur um Haaresbreite durch? Fiel die Entscheidung auf den letzten Metern oder handelte es sich um einen nie gefährdeten Start-Ziel-Sieg? Noch dramatischer zeigt sich der Unterschied im Informationsgehalt beim Ergebnis einer politischen Wahl: Nur die wenigsten wären mit der Nennung der Reihenfolge der Parteien aufgrund der abgegebenen Stimmen (am meisten Stimmen, am zweitmeisten, usw.) zufrieden, lässt sich doch hieraus nicht ermitteln, wer die Wahl gewonnen hat, wie sich das neue Parlament zusammen setzt und wer wohl die neue Regierung stellen wird.

Im Politikerjargon würde man sagen, dass zwischen qualitativen und quantitativen Aussagen, ein „riesiger Quantensprung“ liegt. Deshalb ist es höchst verblüffend, wenn sich eine empirische Wissenschaft mit qualitativen Aussagen begnügt, etwa Aussagen über die Richtung eines Effekts oder die bloße Anordnung von Objekten. Über seine Bemerkung von 1969 (siehe S. 7) noch weit hinausgehend, formuliert Tukey (1991: 104) sein Unverständnis darüber in aller Deutlichkeit:

For pure intellectual curiosity - and perhaps for writing treatises for the intellectually curious - it may be that confident directions (up, uncertain, or down) can suffice. It would be good to understand why many psychologists, for example, seem to be content with only confident directions. Is this a desire for abstract knowledge? Or a sign of inability to make use of more quantitative results? Or an unwillingness to price (or a lack of experience in pricing) comparisons as a basis for real-world actions? Or a belief that *qualitative* knowledge is all that psychologists can hope for? Or what?¹⁶ (Hervorhebung im Original.)

2.4 Bedeutsamkeit

Führt man auf der Menge Z eine neue Relation ein, so sollte man aufgrund des bislang Gesagten darauf achten, dass diese nicht wesentlich von einer speziellen Skala abhängt. Formal: Eine Relation $R(n)$ auf der Menge Z heißt *bedeutungsvoll*, wenn für beliebige Objekte $a_1, \dots, a_n \in W$ gilt, dass die mit einer ersten Skala $s \in S$ erhobenen Messwerte $s(a_1), \dots, s(a_n)$ genau in dann in Relation $R(n)$ zueinander stehen, wenn sie dies auch für eine beliebige

¹⁶ Das Gegenteil ist hingegen umso bemerkenswerter: „Thus, while starting out from a purely qualitative system of axioms, one arrives at a quantitative measure of probability, and then at the theorem of total probability which permits the construction of the whole calculus of probabilities. (De Finetti 1937), zitiert nach Suppes (2009: 30).

andere Skala $s^* \in S$ tun. Beispiel: $s(a_1) < s(a_2)$ gilt genau dann, wenn auch $s^*(a_1) < s^*(a_2)$ wahr ist.

Anders gesagt, für die Gültigkeit der Relation $R(n)$ ist es egal, welche der zur Verfügung stehen Skalen man bemüht. Ist also Werkstück A auf der Kilogramm-Skala schwerer als Werkstück B , so gilt diese Relation auch für die Pfund-Skala und jede andere Skala in S . Darüber hinaus kann man aufgrund der grundlegenden Eigenschaft (2.1) sicher sein, dass Werkstück A auch „tatsächlich“ schwerer ist als Werkstück B .¹⁷

Nun will man nicht zuletzt Daten auswerten. Das heißt, formal gesehen, dass man aus einer Reihe von Messwerten $x_1, \dots, x_n$ eine Statistik t berechnet. Hat man z. B. mehrfach das Gewicht eines Brotes mit einer Waage gemessen und kam dabei nacheinander auf 500, 499 und 504 Gramm, so läge es nahe, das arithmetische Mittel, also $t = (500 + 499 + 504)/3 = 501$ Gramm als den „wahren Wert“ des Brotes anzusehen. Es sollte dabei keine Rolle spielen, ob wir das Brot mithilfe der Gramm-Skala oder aber mit der Einheit kg wiegen.

Dieser Gedanke führt auf die folgende Definition: Eine Statistik t ist eine Abbildung von Z^n nach $\mathcal{R}$. Die Statistik hängt also nur von den Messwerten ab. Sie heißt bedeutungstragend bzw. zulässig genau dann, wenn die von ihr auf dem numerischen Relativ N erzeugte $(2n)$ -stellige Relation $\{(x_1, \dots, x_n, x'_1, \dots, x'_n) \in Z^{2n} : t(x_1, \dots, x_n) = t(x'_1, \dots, x'_n)\}$ bedeutungsvoll ist. Es darf also nicht passieren, dass für eine Skala s die Gleichheit gilt, dass also, ausgehend von den Objekten $a_1, \dots, a_n; a'_1, \dots, a'_n$ die Gleichheit $t(s(a_1), \dots, s(a_n)) = t(s(a'_1), \dots, s(a'_n))$ besteht, während es für eine gewisse zweite Skala s^* zu einer Ungleichheit, also $t(s^*(a_1), \dots, s^*(a_n)) \neq t(s^*(a'_1), \dots, s^*(a'_n))$ kommt.¹⁸ Der Wahrheitswert der Aussage

$$t(s(a_1), \dots, s(a_n)) = t(s(a'_1), \dots, s(a'_n))$$

hinge dann nämlich von der speziellen Skala $s \in S$ ab.¹⁹ Aufgrund der Gleichheitsbedingung heißt diese Methode, die Bedeutsamkeit einer Statistik begrifflich zu fassen, auch *comparison invariance*: Die obige Aussage gilt also für beliebige Skalen s . Sie stellt gewissermaßen eine Minimalanforderung an bedeutungstragende Statistiken dar. Sie ist schwächer (siehe Pfanzagl (1968: 42ff)) und weniger gebräuchlich als die folgende Definition, welche sich direkt zulässiger Transformationen bedient:

Werden Objekte $a_1, \dots, a_n$ mit der Skala $s \in S$ gemessen, so erhält man die Messwerte $s(a_1), \dots, s(a_n)$. Darauf wendet man eine Statistik t an, z. B. das arithmetische Mittel, welches formal gesehen eine Abbildung von Z^n nach

¹⁷ Für weitere Details siehe Pfanzagl (1968: 34ff).

¹⁸ Wir hätten selbstverständlich wie in der Definition auch direkt von den Messwerten der Objekte, also $x_1 = s(a_1), \dots, x_n = s(a_n), x'_1 = s(a'_1), \dots, x'_n = s(a'_n)$ ausgehen können.

¹⁹ Aufgrund der Darstellung $s^* = \gamma(s)$ hat man die Bedeutsamkeit von t gezeigt, falls aus $t(x_1, \dots, x_n) = t(x'_1, \dots, x'_n)$ die Aussage $t(\gamma(x_1), \dots, \gamma(x_n)) = t(\gamma(x'_1), \dots, \gamma(x'_n))$ für alle zulässigen γ folgt.

$\mathbb{R}$ ist. Das heißt, $t(s(a_1), \dots, s(a_n))$ ist eine reelle Zahl. Mit einer anderen Skala $s^* \in S$ wäre das Ergebnis $t(s^*(a_1), \dots, s^*(a_n))$. Diese beiden Resultate sollten sich nicht wesentlich unterscheiden, also „vergleichbar“ bzw. ineinander umrechenbar sein.

Nun wissen wir, dass es eine Darstellung $s^* = \gamma(s)$ mit einer zulässigen Transformation γ gibt. Beeinflussen s bzw. s^* nicht wesentlich das Ergebnis, so sollte es egal sein, ob man γ „zu Beginn der Berechnung“ oder zu deren „Ende“ verwendet, d.h. es sollte die Gleichheit

$$t(s^*(a_1), \dots, s^*(a_n)) = t(\gamma(s(a_1)), \dots, \gamma(s(a_n))) = \gamma(t(s(a_1), \dots, s(a_n)))$$

gelten. Schreibt man kurz $x_i = s(a_i)$ für die Messwerte, welche bei Verwendung der Skala s entstehen, so ist also die Bedingung

$$\gamma(t(x_1, \dots, x_n)) = t(\gamma(x_1), \dots, \gamma(x_n))$$

zu fordern.²⁰ Mit den Worten von Anderson (1961: 309): “It means that if a statistic is computed from a set of scale values and this statistic is then transformed, the identical result will be obtained as when the separate scale values are transformed and the statistic is then computed from these transformed scale values.” Eine prägnantere, mathematische Formulierung ist, dass die zulässige Transformation γ und die Statistik t kommutieren - also im o. g. Sinn vertauschbar sind.

Beispiele

Die häufigste Klasse bzw. der häufigste Wert, auch Modus genannt, ist eine bedeutsame Statistik auf Nominalskalenniveau. Der Grund: Hat man einmal erkannt, welche Klasse am meisten Objekte enthält, so „vergisst“ man dies nicht bei zulässigen Transformationen, welche lediglich eine andere Nummerierung der Klassen bewirken. Formal:

$$f(\text{Modus}(x_1, \dots, x_n)) = \text{Modus}(f(x_1), \dots, f(x_n))$$

mit einer injektiven Transformation f .

Bei geordneten Messwerten $x_{(1)} \leq \dots \leq x_{(2n-1)}$ ist es sinnvoll, den Median $\bar{x}_M = x_{(n)}$ zu berechnen, also jenen Messwert, der die Beobachtungsdaten in zwei gleich große Hälften vom Umfang $n-1$ teilt.²¹ Formal hat man zu zeigen, dass der Median dieser Werte, falls man ihn einer monotonen Transformation g unterwirft, gleich dem Median der zunächst transformierten Messwerte ist.

²⁰ Diese Definition von Bedeutsamkeit impliziert die vorgenannte, da für alle zulässigen γ gilt: $t(x_1, \dots, x_n) = t(x'_1, \dots, x'_n) \implies \gamma(t(x_1, \dots, x_n)) = \gamma(t(x'_1, \dots, x'_n)) \implies t(\gamma(x_1), \dots, \gamma(x_n)) = t(\gamma(x'_1), \dots, \gamma(x'_n))$

²¹ Nämlich die kleineren Werte $x_{(1)}, \dots, x_{(n-1)}$ und die größeren Werte $x_{(n+1)}, \dots, x_{(2n-1)}$.

Und genau dies ist aufgrund von

$$g(\text{Median}(x_{(1)}, \dots, x_{(2n-1)})) = g(x_{(n)}) = \text{Median}(g(x_{(1)}), \dots, g(x_{(2n-1)}))$$

der Fall.²²

Das arithmetische Mittel $\bar{x} = (x_1 + \dots + x_n)/n$ der Werte $x_1, \dots, x_n$ ist bei Intervallskalenniveau eine bedeutungstragende Statistik, da es lineare Transformationen $h(x) = ax + b$ mit $a \neq 0$ respektiert:

$$h(\bar{x}) = a\bar{x} + b = a \left(\sum_{i=1}^n \frac{x_i}{n} \right) + b = \frac{1}{n} \sum_{i=1}^n (ax_i + b) = \frac{\sum_{i=1}^n h(x_i)}{n} \quad (2.3)$$

Eine völlig analoge Rechnung zeigt für das geometrische Mittel $\sqrt[n]{x_1 \cdots x_n}$ der positiven Werte x_i , dass es auf Verhältnisskalenniveau zulässig ist.

Da es sich bei den genannten Skalen um eine Hierarchie handelt, ist eine bedeutungsvolle Statistik auf einem informationsreicheren „höheren“ Niveau auch automatisch bedeutungsvoll für ein informationsärmeres „niedrigeres“ Niveau, jedoch nicht umgekehrt.

Das wichtigste Beispiel einer nicht zulässigen Statistik ist das arithmetische Mittel, falls das Messniveau nur einer Ordinalskala entspricht. Gleichung (2.3) gilt nur bei linearen Transformationen h , deshalb lässt sich bei einer beliebigen monotonen Transformationen g unmittelbar ein Gegenbeispiel angeben: Seien $x_1 = 1, x_2 = 3$ und $g(x) = x^2$. Man berechnet

$$g(\bar{x}) = 2^2 = 4 \neq 5 = (1^2 + 3^2)/2 = \overline{g(x)} \quad (2.4)$$

Der Median ist auf Nominalskalenniveau nicht bedeutungstragend, weil formal gesehen mit den Werten $x_1 = 1, x_2 = -2, x_3 = 3$ und $f(x) = |x|$ gilt:

$$|\text{Median}(-2, 1, 3)| = 1 \neq 2 = \text{Median}(1, 2, 3).$$

Schließlich ist nahe liegend und völlig natürlich, nicht nur Relationen und Statistiken als bedeutungsvoll bzw. zulässig anzusehen, sondern die zentrale Idee auf *beliebige* Beziehungen zu verallgemeinern. Typischerweise nennt man eine numerische Aussage bedeutungsvoll, wenn sich ihr Wahrheitswert unter zulässigen Transformationen nicht ändert. Bei einer beliebigen numerischen Aussage kann es sich insbesondere um eine Gleichung oder Ungleichung (zwischen reellen Zahlen) handeln, in der eine Statistik vorkommt, womit der obige Fall von Statistiken mit abgedeckt ist. Darüber hinaus erinnere man sich insbesondere auch an das Beispiel der Physik (S. 20): Stoßen Physiker auf eine mathematische Beziehung, so studieren sie routinemäßig die Gruppe der Transformationen, die die entsprechende Aussage unverändert lassen. Man nennt dies die „Symmetrien“ der Aussage.

²² Analoges gilt im Fall einer geraden Anzahl von Messwerten.

Die Bedeutung der Bedeutsamkeit

Die Zulässigkeit einer Statistik (und vieler Relationen) lässt sich zumeist auch intuitiv leicht begründen. Wenn zum Beispiel die Daten nicht angeordnet sind - man also lediglich gleich und ungleich unterscheiden kann - so ist es nicht sinnvoll, eine Statistik wie den Median zu verwenden, der sich gerade auf die Anordnung stützt. Analoges gilt für das arithmetische Mittel: Sind die Messwerte lediglich angeordnet, ohne dass die „Distanz“ zwischen ihnen eine reale Bedeutung hätte, so ist es offenkundig fehlerhaft, das arithmetische Mittel zu berechnen, welches wesentlich auf dieser Information basiert.

Ganz allgemein gilt: Ignoriert man die Randbedingung „Skalenniveau“ so arbeitet man mit Informationen, die man de facto gar nicht hat, was zumindest zum Dissens zwischen Forschern führt oder aber einfach zu empirisch nicht zutreffenden Aussagen, also Artefakten und Fehlern. Die Messtheorie warnt nachdrücklich davor, mehr in Daten hineinzulesen, als sich de facto in ihnen befindet. Das Subtile dabei ist, dass Messwerte mathematisch gesehen nichts anderes als Zahlen sind, mit denen man nahezu beliebig rechnen kann. Die Messtheorie beschränkt diese Freiheit wesentlich, da sie zeigt, dass viele formal wohldefinierte Operationen inhaltlich gesehen nicht sinnvoll sind.

Eine Reihe von Autoren haben diesen normativen Aspekt der Messtheorie betont. Pfanzagl (1968: 34) verwendet die intersubjektive Übereinstimmung zwischen Forschern um die Idee zu illustrieren: “Assume that two experimenters measure temperature [...], one of them using centigrade, the other Fahrenheit. Comparing temperatures x, y [...] they will always agree whether $x < y$ is true or false, but never whether $y = 2x$, because $y = 2x$ in centigrade implies $y = 2x - 32$ in Fahrenheit. Thus contrary to the relation $x < y$, the relation $y = 2x$ [...] is meaningless *without specification of the scale*.” (Meine Hervorhebung.) Adams et al. (1965: 124) sprechen eine deutlichere Warnung aus: “The practice of ignoring scale type in making statistical tests could lead to the formulation of empirically meaningless hypotheses.” Noch drastischer äußert sich Luce (1959: 84):

If the interpretation of a particular statistic or statistical test is altered when admissible scale transformations are applied, then our *substantive conclusions will depend on which arbitrary representation* we have used in making our calculations. Most scientists, when they understand the problem, feel that they should shun such statistics and rely only upon those that exhibit the appropriate invariances for the scale type at hand. Both the geometric and the arithmetic means are legitimate in this sense for ratio scales (unit arbitrary), only the latter is legitimate for interval scales (unit and zero arbitrary), and neither for ordinal scales. (Meine Hervorhebung.)

Lauth und Sareiter (2005: 176) schließen kurz und bündig: „Die richtige Festlegung des Skalenniveaus einer Meßgröße ist wichtig für statistische Auswertungen von entsprechenden Meßergebnissen. So ist zum Beispiel bei nominal- und ordinalskalierten Größen die Berechnung von Mittelwerten und Varianzen völlig sinnlos: Es gibt keinen durchschnittlichen Familienstand.“

Ein Artefakt bleibt ein Artefakt auch dann, eine große „Forschergemeinde“ jahrelange ein bestimmtes Messverfahren angewandt und die Ergebnisse auf eine gewisse Art und Weise ausgewertet hat. Genausowenig wie ein Rechenfehler korrekt wird, wenn nur genügend viele Personen von seiner Richtigkeit überzeugt sind, genausowenig wird eine Auswertung, die ein zu hohes Skalenniveau vorausgesetzt hat, dadurch richtig, dass lange Jahre ein Konsens bestand, jenes sei erfüllt. Wie nicht anders zu erwarten, ist gerade die normative Komponente der Messtheorie, welche zwangsläufig eine zuweilen deutliche Einschränkung in den Forschungs- und Auswertungsmethoden mit sich bringt, nicht unwidersprochen hingenommen worden. Den Anfang machten wohl die bissigen Bemerkungen in Lord (1953) mit dem vielsagenden Titel „On the Statistical Treatment of Football Numbers.“

Wir beschäftigen uns mit diesem Problem ausführlich ab dem nächsten Abschnitt. Schon hier möchten wir jedoch darauf hinweisen, dass die Messtheorie auch relevant ist, wenn nur eine einzige Methode zur Messung eines Sachverhalts bekannt ist. Bei allen obigen Argumentationen zur Äquivalenz von Skalen kommen zwar zumindest zwei Skalen s, s^* vor, doch ist es, wie wir festgestellt haben, die Gestalt des empirischen Relativs, welche über das mögliche Skalenniveau entscheidet. Ins Positive gewendet, sind im Falle einer einzigen bekannten Skala messtheoretische Überlegungen die Methode der Wahl, um relevante, auf die Realität zurückgehende Informationen, von willkürlichen Setzungen zu unterscheiden. So genügt der Nachweis, dass eine Messung nur Ordinalskalenniveau erreichen kann, um das arithmetische Mittel zu desavouieren.

Da gesetzesmäßige Zusammenhänge formal gesehen nichts anderes als mathematische Aussagen sind, lassen sich die Einschränkungen der Messtheorie auch konstruktiv nutzen, indem man sie als Randbedingungen interpretiert. Das heißt, die Beschaffenheit des empirischen Relativs schränkt (insbesondere über das Skalenniveau) die Gestalt möglicher Gesetze ein, im bestmöglichen Fall legt der empirische Relativ sogar die Form eines Gesetzes fest. Luce (1959) hat diesen Ansatz bereits kurz nach Stevens Beiträgen ausgearbeitet. „On the possible psychophysical laws“ charakterisiert die Gestalt solcher Gesetze in Abhängigkeit vom jeweiligen Skalenniveau. Zum Beispiel sind Gesetze auf Verhältnisskalenniveau von der Form $y = ax^b$, bei Intervallskalen kommt man wenig verwunderlich auf lineare Zusammenhänge der Gestalt $y = ax + b$. Zwar wurde der Ansatz schnell (z. B. von Pfanzagl (1968: 50ff)) als zu eng kritisiert und erweitert, die zentrale Idee jedoch ist überzeugend: Liegt ein gewisser empirischer Relativ vor, so schränkt einen dies zwar bei der Verwendung formaler Methoden ein, andererseits hilft einem das Wissen aber auch bei der Formulierung allgemeiner Gesetzmäßigkeiten. Die auf Buckingham (1914) zurückgehende Dimensionsanalyse, „eine Methode zur Auffindung der Form von Naturgesetzen, sofern gewisse Voraussetzungen gegeben sind“ (Juhos und Schleichert 1966: 81) ist eine eng verwandte Idee.

2.5 Messtheorie in der Praxis

Die Tatsachen bilden stets den Archimedischen Punkt, von dem aus auch die gewichtigste Theorie aus den Angeln gehoben werden kann.

Planck (1913: 23)

2.5.1 Gegenbeispiele

Eine typische Anwendung der Messtheorie sollte die folgende sein. In den frühen 1950er Jahren führte die Kinderärztin Virginia Apgar (1953) den heute nach ihr benannten Test ein. Er wird benutzt, um kurz nach Geburt die Vitalität des Neugeborenen zu bestimmen:

Item	Sign	0 Points	1 Point	2 Points
1	Activity	absent	arms and legs flexed	active movement
2	Pulse	absent	below 100 bpm ²³	above 100 bpm
3	Grimace	no response	grimace	sneeze, cough, pulls away
4	Appearance	blue-gray, pale all over	normal, except for extremities	normal over entire body
5	Respiration	absent	slow, irregular	good, crying

Addiert man die Punkte, so erhält man den sogenannten APGAR-Score. Ein Wert von 7-10 wird als normal angesehen, Werte zwischen 4 und 7 deuten an, dass möglicherweise unterstützende Maßnahmen notwendig sind. Tieferere Werte sind besorgniserregend und erfordern sofortiges Handeln. Wie man mit unvoreingenommenem Blick leicht erkennt, misst jedes der Items einen anderen Aspekt von „Vitalität“. Es ist medizinisch gesehen deshalb nicht verwunderlich, dass die Summe der Punkte eine recht zuverlässige Aussage über den Gesundheitszustand des Neugeborenen erlaubt: Wenige Punkte bedeuten, dass die wichtigsten Vitalfunktionen kaum vorhanden sind, sehr viele Punkte hingegen zeigen, dass ein Baby atmet, strampelt und auf seine Umwelt reagiert.

Messtheoretisch lässt sich jedoch kaum begründen, dass die Items des Tests ihre jeweiligen Gegenstandsbereiche intervallskaliert messen. Alle Items kennen nur jeweils drei Unterscheidungen, diese werden hauptsächlich verbalungenau beschrieben und selbst beim dritten Item wird kein Versuch unternommen, die Anzahl der Herzschläge pro Minute so abzubilden, dass gleiche Differenzen in der Realität auch identischen Differenzen zwischen Zahlen entsprechen.

Wie das arithmetische Mittel, so ist auch die Summe einer Reihe von Messwerten nur sinnvoll im Sinne der Messtheorie, wenn mindestens Inter-

vallskalenniveau erreicht wurde.²⁴ Das ist auch intuitiv offensichtlich: Sind die Messwerte lediglich angeordnet, also nur bis auf monotone Transformationen festgelegt, so ist ihre Summe völlig beliebig. Die Tatsache, dass die Items des APGAR-Tests verschiedene Aspekte eines Sachverhaltes messen, kommt zudem noch erschwerend hinzu. Denn kombiniert man „ähnliche“ Ordinalskalen, also Ordinalskalen die verwandte Merkmale messen, so sollte die Summe auch wieder höchstens ordinalskaliert sein. Es ist nicht zu erkennen, woher die zusätzliche, metrische, also quantitative Information stammen sollte. Zusammenfassend lässt sich daraus nur der Schluss ziehen, dass der APGAR-Score ein Paradebeispiel für eine nicht sinnvolle Statistik ist.

Dessen ungeachtet wird der APGAR-Score seit Jahrzehnten weltweit mit großem Erfolg eingesetzt. Es ist unter Fachleuten völlig unumstritten, dass er ein wertvolles Instrument ist und das Leben vieler Kinder gerettet hat. Der immer wieder genannte Grund hierfür ist, dass er zum einen die Vitalität des Neugeborenen zufriedenstellend misst (der für uns relevante Punkt) und darüber hinaus mit seiner klaren, einfachen Interpretationsregel Geburtshelfern sagt, wann sie eingreifen müssen.

Der APGAR-Score ist kein Einzelfall. In vielen Bereichen der Wirtschafts- und Sozialwissenschaften wird mit Zahlen hantiert, als befänden jene sich auf (mindestens) Intervallskalenniveau. Es werden Mittelwerte berechnet, Varianzen analysiert, Steuungsmaße, Momente und eine Vielzahl anderer mathematischer Operationen angewandt. N. N. Cox (1996: 481) gibt ein sehr drastisches Beispiel: “Nominal scales are repeatedly said to allow only arbitrary numerical labels. Yet it is well known that assigning 0 and 1 to binary states leads to interpretable means and all manner of worthwhile analyses.”

Gleich mehrere Klassen von Gegenbeispielen werden von Velleman und Wilkinson (1993) genannt: Alternative Typologien von Skalen (welche auch nicht hierarchisch angeordnet sein müssen), multidimensionale Skalen, „robuste“ Statistiken, die sich nicht an eine strikte Skaleneinteilung halten. Ein schönes Beispiel sind getrimmte Mittel, welche insofern zwischen Median und arithmetischem Mittel stehen, als dass sie einen Teil der Daten bei der Mittelung nicht berücksichtigen. Hat man zum Beispiel die Daten $x_i = i^2$ für $i = 1, \dots, 5$ gegeben, so ist das arithmetische Mittel gleich $(1 + 4 + 9 + 16 + 25)/5 = 11$. Das 25%-getrimmte Mittel lässt den größten und den kleinsten Wert einfach weg und man erhält $(4 + 9 + 16)/3 = 9, \bar{6}$. Das 50%-getrimmte Mittel berücksichtigt die beiden größten und kleinsten Werte nicht und ist damit gleich dem Median, also 9. Auf welchem Niveau ist eine solche Vorgehensweise sinnvoll? Die Methode stammt bezeichnenderweise aus der Praxis, sie wird vor allem dann angewandt, wenn es Ausreißer gibt und hat sich gut bewährt. In der Stevensschen Typologie ist dafür jedoch kein Platz: Auf Intervallskalenniveau wertet das arithmetische Mittel die verfügbare metrische Information aus, für die „darunterliegende“ Ordinalskala ist der Median angemessen.

²⁴ Siehe Gleichung (2.3) und Ungleichung (2.4)

Nicht zuletzt kann es auch von der Fragestellung abhängen, wie Daten ausgewertet werden. Dann werden ein und dieselben Daten - je nach Verfahren - so behandelt, als befänden sie sich auf ganz unterschiedlichen Skalenniveaus. In der Statistik transformiert man darüber hinaus gerne Daten, bevor sie im eigentlichen Sinne ausgewertet werden. Die Rohwerte werden häufig zunächst logarithmiert, geglättet oder aber auch um sogenannte „Ausreißer“, also Beobachtungen, die sich erheblich von allen übrigen Daten unterscheiden, bereinigt. Bei alledem hofft man natürlich, durch die jeweilige Transformation besser interpretierbare Zahlenwerte zu erhalten.

Auf der Suche nach Alternativen

Wir stehen vor einem Paradoxon: Auf der einen Seite legt die Messtheorie überzeugend dar, wann bestimmte Berechnungen nicht sinnvoll sind. Auf der anderen Seite führt die Missachtung dieser Empfehlungen sehr häufig zu empirisch sinnvollen Ergebnissen. Diese Diskrepanz hat - wie leider nicht anders zu erwarten - zu einer Polarisierung der Meinungen geführt, zuweilen kam es sogar zu hitzigen Wortgefechten.²⁵ Dies ist nicht weiter verwunderlich, stehen doch letztlich mathematisch hergeleitete Sätze (samt ihren normativen Implikationen) empirischen Fakten gegenüber. Wird man vor diese unerquickliche Wahl gestellt, so hat man im Prinzip drei Möglichkeiten: Entweder man verteidigt die Mess-Theorie, lehnt sie ab, weil man der Empirie die Priorität einräumt, oder aber man sucht einen Kompromiss, der beiden Seiten gerecht wird. In dieser Reihenfolge gehen auch wir vor.

Versuchen wir es zunächst mit der Verteidigung. Ein Protagonist könnte behaupten, die Messtheorie „habe doch recht“. Das heißt, man werde auf Dauer schon sehen, dass all jene, die sich nicht an die von ihr verordneten Vorsichtsmaßnahmen gehalten haben, letztlich Artefakten aufgesessen sind, auch wenn sie dies lange Zeit nicht wahrhaben wollten. In diesem Sinne schrieb mir ein anonymer Gutachter vor einigen Jahren: „Wenn eine Schutzmaßnahme (Beachtung des Skalenniveaus [...]) nicht realisiert wird, ohne daß die negativen Konsequenzen eintreten, vor denen man sich schützen will (sinnlose Ergebnisse [...]), dann widerlegt das nicht den Anspruch, daß die Schutzmaßnahme hinreichend zur Vermeidung dieser Konsequenzen ist.“

Dieses Argument ist umso überzeugender, je mehr empirische Beispiele zu seinen Gunsten angeführt werden können, d.h., wenn man in einer Reihe von Anwendungsfeldern zeigen kann, dass viele „unvorsichtige“ Untersuchungen zu nicht bedeutungsvollen Aussagen geführt haben. Da die Messtheorie seit gut 50 Jahren gelehrt und von einer ganzen Reihe von Gebieten zumindest zitiert wird, sollte es eigentlich nicht schwer fallen, solche Beispiele benennen. Jedoch - wirklich empirische Beispiele sind in der Literatur zur Messtheorie

²⁵ Siehe z. B. Velleman und Wilkinson (1993: 68) und Hand (1996: 447): [...] „most of the debate has taken the form of polemics favouring one or the other side of the debate [...]“

nicht allzu häufig, haben eher Lehrbuchcharakter,²⁶ und typischerweise kommen Autoren zum genau gegenteiligen Schluss, nämlich dass “[...] in spite of the importance of measurement theory [...], many of these prescriptions and proscriptions are either premature or, if strictly applied, would represent a major hindrance to the progress of empirical research [...]” (siehe das Abstract von Briand et al. (1996)).

Vorsichtsmaßnahmen, die sich permanent als überflüssig erweisen, können und werden in der Praxis ohne Schaden ignoriert werden. Man benötigt keine Dämme gegen Fluten, die nie kommen. Besonders schwer wiegen in dieser Hinsicht Beispiele wie jenes von N. Cox, in denen die Voraussetzungen massiv verletzt werden und trotzdem eine unzulässige Statistik - arithmetisches Mittel bei Nominalskalenniveau - sinnvolle Ergebnisse liefert. Schließlich kann der im Prinzip pessimistischen Haltung des weiter oben zitierten Gutachters auch eine optimistisch-zuversichtliche entgegengestellt werden, die noch dazu empirisch wesentlich besser fundiert ist:

[We] may prefer the freedom to do something ‘impermissible’ and to judge *a posteriori* whether it was warranted. [Some author] may analyse only the attributes that he understands *a priori* and may constrain his analyses by his assumptions about measurements. But why ask others to wear that strait-jacket? We should instead champion the freedom to learn from the data [... Whereas critics] will not even let us look! (Velleman (1996: 486), Hervorhebungen im Original.)

Aufgrund der immensen Evidenz²⁷ gegen die Messtheorie kann das Beibehalten der o.g. Position nur noch als *dogmatisch* bezeichnet werden. Sie gleicht einer Immunisierung (Lakatos 1974) gegen jegliche empirische Kritik. Es ist äußerst unplausibel, dass sich zehntausende von Wissenschaftlern auf Dutzenden von Gebieten über Jahrzehnte hinweg grundsätzlich getäuscht haben sollten, weil sie von der Messtheorie „verbotene“ Methoden und Statistiken verwendet haben.

2.5.2 Enge und Rigidität

Kommen wir zu den entschiedenen Kritikern der Messtheorie. Tukey (1986b: 246) führt als Beispiel die Temperaturmessung an: “How were temperatures measured? With one of any of several kinds of thermometers. [...] Would there be agreement between the different kinds of thermometers? Approximate agreement, yes; exact agreement, certainly not. Would any one kind have sufficient theoretical support to be chosen as *the* standard over all others. No. Clearly temperature was not measured on an interval scale in those days. But equally clear, it made good sense to [...] calculate the arithmetic mean

²⁶ Siehe z. B. Pfanzagl (1968), Bortz (2004), Narens (2002) sowie die Werke von Luce, Krantz, Suppes und Tversky

²⁷ Dieser Begriff wird im folgenden eine große Rolle spielen. Für eine genauere Begriffsbestimmung siehe S. 199.

of a group of temperatures. Temperature was not measured on a *mere* ordinal scale. It was measured on a scale which, though not an interval scale, was still quite well defined.” (Hervorhebungen im Original.) Er fährt fort (ibd., S. 247): “An oversimplified and overpurified view of what measurements are like cannot be allowed to dictate how data is to be analyzed.” Lucas Standpunkt (siehe S. 35) zeigt nach Tukey (1986b: 246) “A lack of adequate recognition that knowledge is approximate, not precise.” Er kommentiert ihn mit den Worten (ibd., S. 243)

The view thus summarized is a dangerous one. If generally adopted it would not only lead to inefficient analysis of data, but it would also lead to failure to give any answer at all to questions whose answers are perfectly good, though slightly approximate. All this loss for essentially no gain.

Schließlich fasst er seine Argumentation in einem *badmament*²⁸ (ibd., S. 201) zusammen: “BE EXACTLY WRONG, RATHER THAN APPROXIMATELY RIGHT.” (Hervorhebungen im Original.)

Die Approximationsidee von Tukey widerspricht in ihrem konzilianten Geist den strikten und absoluten Denkmustern der Messtheorie. Velleman (s.o.) spricht explizit von einer Zwangsjacke und Duncan (1984: 129) von einem Prokrustesbett, weil es auf der anderen Seite unmissverständlich heißt: “As is perhaps obvious, meaningfulness is an all-or-none concept. Thus a statement can not be almost meaningful” (Townsend und Ashby 1984: 394). Auch die enge Skalenhierarchie (mit lediglich fünf Skalen) wird nicht in Frage gestellt: “Many of the discussions of scale types, and virtually all of the mathematical results, treat them as absolute categories. Data are expected to fit into one or another of the categories. A failure to attain one level of measurement is taken as a demotion to the next level” (Velleman und Wilkinson 1993: 70). In der typischen Situation, dass eine Messung „zwischen“ dem Ordinal- und Intervallskalenniveau liegt, ist man deshalb im Zweifelsfall gezwungen, Daten auf dem vermeintlich sichereren bzw. angemessenen, also tieferen Niveau auszuwerten. Der Informationsgehalt ist dort, wie wir weiter oben deutlich gemacht haben, insbesondere beim Übergang von einer Intervall- zu einer Ordinalskala, jedoch deutlich geringer.

Wilkinson (1996: 487) hat hierfür nur Spott übrig: “And the popular device to those possessing ‘sloppy data’ (downgrade the measurement level to *ordinal* or *nominal*) can only be considered a parody of measurement theory.” (Hervorhebungen im Original, siehe z. B. Hartung et al. (2009: 408f).) Die Parodie wird vollends zur Farce, wenn Variablen „dichotomisiert“, also auf lediglich *zwei* mögliche Werte reduziert werden - eine genauso beliebte wie zugleich erschreckende Praxis (siehe MacCallum et al. (2002) oder Cohen (1983)). Schon Berkeley (1710) fand²⁹ für diese Strategie treffende Worte:

Upon the whole, I am inclined to think that the far greater part, if not all, of those difficulties which have hitherto amused philosophers, and blocked up the way to

²⁸ - also in einem absichtlich schlechten Ratschlag -

²⁹ Siehe die Einleitung zu seinem Werk

knowledge, are entirely owing to ourselves - *that we have first raised a dust and then complain we cannot see.* (Meine Hervorhebung)

Sich selbst Sand in die Augen zu streuen ist eine wirkungsvolle Möglichkeit, Daten gründlich zu missachten und potenziell wichtige Informationen zu ignorieren. Sie widerspricht diametral der ersten Maxime empirischer Wissenschaft, dass sich die Theorie den Fakten anzupassen habe und nicht umgekehrt.³⁰ Dieser Aufforderung widersetzt sich ein Wissenschaftler, wenn er, statt eine praxistaugliche Theorie zu entwickeln, bei der einfachen Klassifikation der Messtheorie bleibt, obwohl “the assertion, common to many traditional statistics texts, that ‘data values are nominal, ordinal, interval or ratio’ simplifies the matter so far as to be false” (Velleman und Wilkinson 1993: 69).

2.5.3 Verfeinerung und Mathematisierung

Protagonisten der Messtheorie haben die Kritik ernst genommen und konstruktiv darauf reagiert. Die hauptsächlichen Bestrebungen gingen dahin, die Messtheorie mathematisch zu verfeinern und zu erweitern. Die dabei erzielten erheblichen Fortschritte wurden etwa in Pfanzagl (1968), Falmagne (1985), Narens (1985, 2002) und dem schon öfters genannten fundamentalen dreibändigen Werk von Luce, Krantz, Suppes und Tversky (1971, 1989, 1990) eindrucksvoll dokumentiert.

Solche Bemühungen sind sehr sinnvoll und auf alle Fälle zu würdigen. Die entscheidende Frage ist und bleibt jedoch, ob es mithilfe dieser Maßnahmen gelingt, besser mit typischen, also in der Realität häufig auftretenden Situationen, zurecht zu kommen. Die Antwort auf diese Frage muss natürlich zum Teil *Ja* lauten, da ein Mehr an mathematisch erforschten Strukturen prinzipiell die Wahrscheinlichkeit erhöht, dass eine dieser Strukturen in einer speziellen Situation angemessen ist. Die Antwort lautet jedoch *Nein*, wenn man sich davon erhofft hätte, dass nun in der alltäglichen Forschungspraxis messtheoretische Überlegungen konsequent (und nicht nur als Parodie) angewandt würden bzw. werden könnten. Auch wenn in der Psychophysik (also dem Gebiet Stevens) oder einschlägigen Artikeln des *Journal of Mathematical Psychology* messtheoretisches Denken gepflegt wird, so ist das beim großen

³⁰ Mit den Worten von Galilei (1612): „Circa il qual particolare, io voglio solamente rappresentare a V. E. quello che egli sa molto meglio di me, et è che noi non doviamo desiderare che la natura si accomodi a quello che parrebbe meglio disposto et ordinato a noi, ma conviene che noi accomodiamo l'intelletto nostro a quello che ella ha fatto, sicuri tale esser l'ottimo et non altro; e perchè ella si è compiaciuta di far muover le stelle erranti circa centri diversi, possiamo esser sicuri che simile costituzione sia perfettissima et ammirabile, et che l'altra sarebbe priva d'ogni eleganza, incongrua e puerile.“ Die Übersetzung in Michell (1999: xiii.), lautet: “We must not ask nature to accommodate herself to what might seem to us the best disposition and order, but must adapt our intellect to what she has made, certain that such is the best and not something else.”

Mainstream nicht der Fall. Bücher wie Luce (2000) sind die Ausnahme. Die Messtheorie hat es bislang nicht geschafft, ihre konzeptionelle Enge wirklich zu überwinden. Was sind die Gründe hierfür?

Zunächst einmal sind auch alle wichtigen Erweiterungen der klassischen Messtheorie *deterministisch*.³¹ Sie verwendet im Wesentlichen algebraische Strukturen als empirischen Relativ, welche mittels die Struktur respektierenden Abbildungen auf die Welt der Zahlen übertragen werden. So reichhaltig deterministische mathematische Strukturen auch sind - sie sind bei weitem nicht so zahlreich und v.a. auch nicht so flexibel anpassbar wie Modelle, welche Wahrscheinlichkeiten und damit Unschärfen, Fehler und Schwankungen zulassen. Einem Objekt wird in einem fest gewählten messtheoretischen Modell immer dieselbe Zahl zugeordnet. Es ist nicht vorgesehen, dass ein und dasselbe Objekt, z. B. bei aufeinanderfolgenden Beobachtungen, etwas von einander abweichende Messwerte produziert. (Das ist auch eine Kritik von Duncan (1984: 121) und der dort genannten Literatur.) Im Prinzip genügt bei deterministischen Modellen sogar eine einzige Beobachtung, welche nicht mit dem gewählten Modell in Einklang zu bringen ist, um es zu falsifizieren.

Als nächstes ist der grundlegende Homomorphismus (siehe S. 24, insbesondere Gleichung (2.1)) zu nennen. So natürlich er mathematisch ist, de facto stellt er sehr weitreichende Anforderungen an jede Messung. Es kommt einer *naiv-realistischen* Position gefährlich nahe, wenn man voraussetzt, dass eine Relation zwischen Zahlen *genau dann* gilt, wenn sie auch in der Realität zwischen Objekten besteht. Neben allen hochgradig philosophischen Problemen, die man sich damit einhandelt,³² bleiben auch ganz schlichte, jedoch allenthalben zu erwartende *Messfehler* unberücksichtigt; stattdessen modelliert man die perfekte Entsprechung empirischer und numerischer Strukturen.

Beunruhigender noch ist, dass die Entwicklung der klassischen und mehr noch der aktuellen Messtheorie primär *mathematisch-formalen* Gesichtspunkten folgt. Schon die klassischen Skalen der Messtheorie wurden mithilfe grundlegender mathematischer Transformationen definiert. An keiner Stelle spielen empirische Überlegungen wirklich eine Rolle. Die Theorie stellt eine Skalen-Hierarchie bereit, weil die berücksichtigten Transformationen Teilmengen voneinander sind. Man vergleiche das mit dem Vorgehen in der Physik: Ist man dort auf eine interessante Gleichung gestoßen, so wird routinemäßig die Gruppe der *zu ihr gehörigen* Transformationen ermittelt. Es ist nicht nur völlig alltäglich, Invarianzbetrachtungen anzustellen, wichtiger noch ist, dass die Invarianzüberlegungen von einer konkreten Relation ausgehen und die Lösung auf den gerade betrachteten Fall zugeschnitten ist.

Wir haben uns in der Einleitung vehement für die Mathematisierung ausgesprochen, wenn diese einem *kristallisierten Verständnis* (siehe S. 8) realer Sachverhalte entspricht. Was wir im Falle der Messtheorie jedoch sehen, ist ein Beispiel schlechter Mathematisierung, also von Abstraktion um der

³¹ Die einzige mir bekannte Ausnahme ist Katz (1981).

³² Sollten wir bei Zutreffen des Modells die Welt nicht exakt und fehlerfrei erkennen können?

Abstraktion willen, wobei der Kontakt zu den konkreten Problemstellungen mehr und mehr verloren geht. Schlägt man die genannten Standardwerke zur Messtheorie, aber auch viele andere Bücher zum Thema, auf,³³ so stellt man unschwer fest, dass man im Wesentlichen ein Mathematikbuch in Händen hält. Man erwartet eine Theorie des Messens - Ausführungen über konkrete Messvorgänge, experimentelle Prozeduren, Apparate, konkrete Beispiele gemessener Größen, Fehlerquellen usw. - und findet sich nach wenigen Seiten in reiner Mathematik wieder. Formale Definitionen, Sätze und Beweise, viele der modernen Algebra eng verwandt oder sogar entlehnt, folgen im Stakka-to aufeinander, und nur selten verirrt sich ein wirklich empirisches Beispiel zwischen die Zeilen. Duncan (1984: 121) sagt dazu:

The earlier remarks on a certain deficiency of realism in philosophical discussions of measurement could be applied here as well. The idealization of the measurement process is carried even further in this fully mathematical approach [...]

Krantz et al. (1971: 13) erklären zwar ausdrücklich “What is invariant, and so is not a matter of convention, is the empirical relational structure and its empirical properties, some of which are formulated as axioms. [...] We shall make an effort to point out the status of various axioms or classes of axioms as empirical laws.” Aber lassen die Autoren ihrer Ankündigung wirklich Taten folgen? Kaum 10 von über 500 Seiten dieses Hauptwerkes der Messtheorie beschäftigen sich mit der Verknüpfung von empirischen Fakten und ihren numerischen Repräsentationen. Der Rest des Buches widmet sich v.a. mathematischen Strukturen und ihren homomorphen Abbildungen. Dies provozierte schon früh Kritik. Estes (1975: 273), zitiert nach (Schönemann 1994: 149f), schreibt:

One reason for the relative paucity of connections between measurement theory and substantive theory in psychology may arise from the fact that models for measurement have largely been developed independently as a body of abstract formal theory with empirical interpretations being left to a later stage. The difficulty with this approach is that the later stage often fails to materialize.

In den folgenden Bänden (Suppes et al. (1989), Luce et al. (1990)) setzt sich diese Entwicklung nicht nur fort, darüber hinaus wird auch, wie Schönemann (1994) überzeugend darlegt, die Theorie belastende Evidenz konsequent ausgeblendet, so dass er (ibd., S. 156 und 150) kommentiert:

If one persistently ignores inconvenient empirical facts, it is hardly surprising if empirical connections fail to materialize in the long run [...] In hindsight it seems fairly plain that the virtually perfect absence of empirical support must have been one reason why the ‘measurement revolution’ (Cliff 1992) never happened.

Jedenfalls ist es außerordentlich treffend, dass viele einschlägige Bücher das Adjektiv *abstract* im Titel oder Untertitel verwenden. Sie demonstrieren eindrucklich, was geschieht, wenn mathematische Überlegungen eine Eigendynamik entfalten, die eine Theorie konsequent weg von konkreten empirischen

³³ Ein typisches Beispiel ist Klein (1994).

Fragestellungen führen. Wohin gelangt man auf dem mathematischen Weg? Einerseits zu mathematisch „glatteren“ Strukturen, da mit deren Hilfe zwar weiter reichende Resultate bewiesen werden können. Zum Beispiel ersetzt Narens (1985: 7) den Homomorphismus (siehe (2.1), S. 24) durch einen Isomorphismus und schreibt dazu explizit: “The primary reason for this is that isomorphisms preserve truth whereas homomorphisms do not. (The choice of homomorphisms as the basis for the representational theory of measurement has never been adequately justified, and in theoretical considerations about measurement it often leads to serious difficulties.)”

Eine elegantere Mathematik bedeutet jedoch andererseits auch zumeist, dass die Situation stärker idealisiert wird als am Ausgangspunkt der Untersuchungen. Ein alltägliches Beispiel hierfür sind die reellen Zahlen. Man benutzt sie weit häufiger als die rationalen Zahlen, und auch als numerischer Relativ sind sie sehr beliebt. Warum? Es sind letztlich innermathematische Gründe - mit reellen Zahlen lässt sich besser arbeiten als mit rationalen. Doch eigentlich lehrt jede reale Messung, die immer zu einer Zahl mit endlich vielen Nachkommastellen, also einer rationalen Zahl führt, dass die Menge der rationalen Zahlen ein angemesseneres Modell für reale Messungen mit endlicher Messgenauigkeit wäre (siehe Barnard (1996a: 480)).

Der Gipfelpunkt einer solchen Entwicklung ist, wenn Begriffe, die gerade die Verknüpfung zwischen Theorie und Praxis herstellen, formalisiert werden. In einem jüngeren Buch widmet sich Narens (2002) v.a. dem Bedeutsamkeitsproblem. Der Leser wird bereits ahnen, wie die dort entwickelte Lösung aussieht: Eine streng-formale Zwei-Ebenen-Theorie, entsprechend der klassischen Unterscheidung zwischen empirischem und numerischem Relativ, innerhalb derer sich Bedeutsamkeit exakt fassen lässt. Die zentralen Kapitel sind: “3. Axiomatic Set Theory, 4. Axiomatic Generalizations of the Erlanger Programm, 5. Representational Theory of Measurement.” Er trennt dabei konsequent zwischen rein formaler „Bedeutsamkeit“ und “real-world‘empiricalness”. Ersterer gilt sein Augenmerk, letztere wird lediglich und ziemlich spät³⁴ eingeführt “to distinguish empiricalness from meaningfulness [...] Because of this concern’s limited nature, a detailed development of ‘empiricalness’ is not needed [...]”³⁵

Ein weiteres Resultat dieser Entwicklung ist, dass die aktuelle Messtheorie nur noch für einen mathematisch wirklich versierten Wissenschaftler zugänglich ist. Sie ist beeindruckend, wie jeder Berg, den man nur mit äußerster Anstrengung bezwingen kann, zugleich aber ist sie genauso schroff und unfruchtbar wie ein Hochgebirge. Davor hatte schon vor über vier Jahrzehnten Pfanzagl (1968: 9) ausdrücklich gewarnt: “A close relationship to applications is necessary, because otherwise the theory of measurement becomes rather fruitless [...]” Ganz allgemein schreibt Jaynes (2003: 674ff):

³⁴ Narens (2002: 409)

³⁵ Man vergleiche dies mit Krantz (1991): “Numerical statements are meaningful insofar as they can be translated, using the mapping conventions, into statements about the original qualitative [empirical] structure.”

[...] nothing could be more pathetically mistaken than the prefatory claim [...] that mathematical rigor ‘guarantees the correctness of the results’. On the contrary, much experience teaches us that the more one concentrates on the appearance of mathematical rigor, the less attention one pays to the validity of the premises in the real world, and the more likely one is to reach final conclusions that are absurdly wrong in the real world.

Aktuelle Mathematik (Exkurs)

Am Rande sei bemerkt, dass zahlreiche Stimmen aus den empirischen Wissenschaften, aber auch der Mathematik selbst, Pfanzagl's Warnung mit Blick auf die reine Mathematik wiederholen und vertiefen. Ausführlich, wenn auch ziemlich polemisch, geht Jaynes (2003) in seinem Anhang B auf den aktuellen mathematischen Mainstream ein. Dieser wird tatsächlich „seit Bourbaki“, also seit gut fünfzig Jahren, dezidiert von formal-abstrakten und eben nicht (mehr) echten empirischen Problemen geprägt. Auch der angewandte Mathematiker Kline (1980) widmet in seinem einflussreichen Buch ein ganzes Kapitel (XIII) der bedauerlichen „Isolation der Mathematik“. Ganz ähnlich äußert sich Feynman (2005a: 484ff). Schon J. von Neumann (1947: 191) ging auf das Thema mit einem besorgten Unterton ein:

The symbolism of algebra was invented for domestic, mathematical use, but it may be reasonably asserted that it had strong empirical ties. However, modern, ‘abstract’ algebra has more and more developed into directions which have even fewer empirical connections. The same may be said about topology. And in all these fields the mathematician’s subjective criterion of success, of the worth-whileness of his effort, is very much self-contained and aesthetical and free (or nearly free) of empirical connections.

Er fährt fort (ibid., S. 195): ‘I think it is a relatively good approximation to the truth [...] that mathematical ideas originate in empirics [...]. But, once they are so conceived, the subject begins to live a peculiar life of its own and is better compared to a creative one, governed by almost entirely aesthetical motivations, than to anything else and, in particular, to an empirical science.’ Genau dies beunruhigt ihn, denn er schließt seinen Aufsatz eine Seite später mit dem Absatz:

As a mathematical discipline travels far from its empirical source, or still more, if it is a second and third generation only indirectly inspired by ideas coming from ‘reality’, it is beset with very grave dangers. It becomes more and more purely aestheticizing, more and more *l’art pour l’art*. This need not be bad, if the field is surrounded by correlated subjects, which still have closer empirical connections, or if the discipline is under the influence of men with exceptionally well-developed taste. But there is a grave danger that the subject will develop along the line of least resistance, that the stream, so far from its source, will separate into a multitude of insignificant branches, and that the discipline will become a disorganized mass of details and complexities. In other words, at a great distance from its empirical source, or after much ‘abstract’ inbreeding, a mathematical subject is in danger of degeneration. At the inception the style is classical; when it shows signs of becoming baroque, then

the danger signal is up. It would be easy to give examples, to trace specific evolutions into the baroque and the very high baroque [...] In any event, whenever this stage is reached, the only remedy seems to me to be the rejuvenating return to the source: the reinjection of more or less directly empirical ideas. I am convinced that this was a necessary condition to conserve the freshness and the vitality of the subject and that this will remain equally true in the future. (Hervorhebung im Original.)

Man kann durchaus noch weiter gehen und nach „tieferen“ Gründen für die Distanzierung der Mathematik von empirischen Problemen suchen. Arnold, zitiert nach Beutelspacher (2008: 107), sieht die Ursache auch in der dominierenden, auf Hilbert zurückgehenden formalistischen Philosophie der Mathematik:

Zu Beginn dieses [des 20.] Jahrhunderts wurde ein selbstzerstörerisches demokratisches Prinzip in die Mathematik eingeführt (vor allem durch Hilbert), nach dem alle Axiomensysteme das gleiche Recht auf Analyse haben und der Wert einer mathematischen Leistung nicht durch seine Bedeutung und seinen Nutzen für andere Disziplinen bestimmt wird, sondern allein durch seine Schwierigkeit, wie beim Bergsteigen. Dieses Prinzip führte schnell dazu, daß die Mathematiker mit der Physik brachen und sich von allen anderen Wissenschaften abschotteten. In den Augen aller normalen Leute verwandelten sie sich in eine obskure priesterliche Kaste... Merkwürdige Fragen wie Fermats Problem oder Summen von Primzahlen wurden zu angeblich zentralen Problemen der Mathematik erhoben.

Wir gehen auf dieses Thema nochmals ab S. 548 ein.

2.5.4 Distanzierung durch begriffliche Differenzierung

Aufgrund der offensichtlichen Schwierigkeit, den empirischen Relativ und die Bedeutsamkeit von Relationen oder Statistiken *unmittelbar* empirisch interpretieren zu können, hat man nach weiteren Wegen gesucht, dieses Problem zu entschärfen.

Ein klassischer, von Adams et al. (1965)³⁶ ausgeführter Ansatz, ist, zwischen „fundamentalen“ und „sonstigen“ Messungen zu unterscheiden: “Critics of Stevens [...] do not distinguish between measurements such as [test scores] and fundamental measurement systems for which there are clearly defined sets of permissible transformations.”

Wieder steht die Physik Pate: “In physics, most measurement is ‘fundamental’, in the sense that it is not derived from other measurement” (Schönemann 1994: 152). Da nur bei fundamentalen Messungen ein empirischer Relativ zweifelsfrei zu erkennen ist und der grundlegende Homomorphismus (2.1) dann eine Analogie zwischen empirischen und formalen Beziehungen herstellt, kann man auch nur dort problemlos von zulässigen Transformationen usw. sprechen. In vielen anderen Fällen, insbesondere wenn unklar ist, was überhaupt gemessen wird, ist Stevens’ Theorie hingegen nicht anwendbar. In den

³⁶ Zitiert nach Duncan (1984: 146)

Augen der Verteidiger der Messtheorie ist diese häufige Nicht-Anwendbarkeit jedoch kein Makel der Theorie. Vielmehr schreiben Adams et al. (1965): “[...] the imprecision of the concept in these applications is inherent, and no deeper analysis can discover what the permissible transformations ‘really are’ in these cases.” Gleichwohl reicht es noch zu einer einschlägigen Warnung: “However, this is not to say that questions of appropriate statistics and significance cannot legitimately be raised about uses of statistics in connection with these scales [...]”

Eine etwas subtilere Strategie „theoretisiert“ die problematischen Konzepte. D.h., anstatt die Anwendbarkeit der ganzen Theorie in Frage zu stellen, werden lediglich einige wenige, wenn auch zentrale Begriffe wie „empirischer Relativ“ und insbesondere „Bedeutsamkeit“ nicht mehr unmittelbar auf die Realität bezogen. Bei der Ausarbeitung dieser Alternative haben sich nicht zuletzt deutschsprachige Autoren hervorgetan. In einer vielbeachteten Buchrezension unterscheidet Niederée (1994) zwischen mehreren prinzipiell verschiedenen Arten von Bedeutsamkeit:

- “Roughly, type-1 concepts of meaningfulness are understood here to be concerned with a semantically sound, i.e., unambiguous, reference to quantitative models (and other entities) [...]” (Ibd., S. 543.)
- Wie zu erwarten, widmet er einen eigenen Abschnitt (Ibd., S. 545-554) den “Type-2 Concepts of Meaningfulness: Invariance.”
- Schließlich führt er sogar “Type-3 Meaningfulness” ein: “The rationale of this interpretation is [...] that it is not the above statistical hypotheses themselves which [...] may be considered as being either true or false with respect to certain scales *X*. What often is doubtful is *their significance in regard to other, more substantive, scientific hypotheses* [...] However, to preclude misunderstanding, one had perhaps better simply use expressions like *scientific significance* (or *practical relevance*, respectively) [...]” (Hervorhebungen im Original.)

Genauso entschieden differenzieren Steyer und Eid (1993) zwischen dem Bedeutsamkeitsbegriff der Messtheorie und der semantischen Interpretation des Begriffs. Sie schreiben (S. 19): „Dieses Beispiel macht klar, dass ‚bedeutsam‘ ein *Terminus technicus* ist und nicht mit dem umgangssprachlichen Begriff gleichen Namens verwechselt werden darf [...] Tatsächlich ist es manchmal sinnvoll und notwendig, auch die Mittelwerte einer Nominalskala zu berechnen. Richtig ist aber, daß die Wahrheitswerte der Aussagen über diese Kennwerte *nicht invariant* sind [...]“ Dies ist nur ein Beispiel von vielen im genannten Buch der Autoren. Fast wortgleich sagt Narens (2002: 56): “In the above and subsequent examples, ‘meaningfulness’ is used in a technical manner, and as such, it and its negation ‘meaningless’ should not be given the ordinary English interpretations of ‘having meaning’ and ‘having no meaning.’ As a technical term, it is always defined throughout the chapter with respect to some measurement scale.” Am Ende seines Buches kontrastiert

er ausdrücklich (theoretisch-formale) „meaningfulness“ mit der tatsächlichen empirischen Bedeutsamkeit (Abschnitt 7.11.2 Empiricalness).

Scheinbar stellt diese Differenzierung eine überzeugende Lösung des Problems dar: Innerhalb der Messtheorie wird weiterhin von Bedeutsamkeit geredet, man verzichtet jedoch auf eine empirische Interpretation des Begriffes. Ein Resultat der Messtheorie kann streng-deduktiv beweisbar, also (Typ 2-)Invarianz-bedeutsam sein, ohne dass es in der Realität eine Entsprechung haben müsste. Und umgekehrt: eine Statistik kann eine große (Typ 1-)empirische Bedeutsamkeit erlangen, ohne dass sie deshalb messtheoretisch zulässig sein müsste. Das heißt, durch die Aufspaltung des Begriffes der „Bedeutsamkeit“ in zwei separate Begriffe entkoppelt man die empirisch-semantic und theoretisch-formale Bedeutung desselben. Es bleibt dann eigentlich nur das Problem (jeder Theorie), zwischen vielen verschiedenen Nuancen ein und desselben Konzepts wählen zu müssen. Zum Beispiel sieht sich Klein (1994: 43) vor die Aufgabe gestellt, zwischen neun subtil verschiedenen Definitionen von (Typ 2-)Bedeutsamkeit zu entscheiden, Narens (1985: 168) hat drei Möglichkeiten zur Auswahl.

Wir sprechen von einer Scheinlösung, da die Messtheorie durch diesen Schachzug zu einer rein mathematischen Unternehmung wird. Hier die rein formalen Zusammenhänge und Begriffe innerhalb der Theorie, da die empirischen Bedeutungen. Jede wissenschaftliche Theorie muss sich aber, das ist seit Galilei unbestritten, an der Wirklichkeit messen lassen. Huxley (1893) fand die heute wohl am häufigsten zitierte Formulierung:

The great tragedy of science – the slaying of a beautiful hypothesis by an ugly fact.

Popper sprach prägnant vom Kriterium der „Falsifizierbarkeit“, welches jede wissenschaftliche Theorie erfüllen müsse. Und nach Tukey (1962: 397) ist „Reliance upon the test of experience as the ultimate standard of validity“ entscheidend für jede wissenschaftliche Theorie.³⁷

Selbstredend sind solche Forderungen gerade auch an jede Theorie des *Messens* zu richten, welche den Anspruch erhebt, wesentliche Aspekte empirischer Messvorgänge formal zu fassen. Zudem sollte eine solche Theorie in aller Regel anwendbar sein, denn was nutzt eine Mess-Theorie, die sich in den meisten praktischen relevanten Fällen für „nicht zuständig“ erklärt? Sie darf sich auch nicht durch eine Immunisierungsstrategie ihrer Falsifikation entziehen, was insbesondere dann geschieht, wenn man den syntaktischen vom semantischen Teil des Begriffes „Bedeutsamkeit“ entkoppelt. Die Konsequenz ist klar: „Axiomatic measurement theory is mathematics rather than science“ (Velleman und Wilkinson 1993: 68), also unangreifbar und zugleich oftmals empirisch irrelevant.

³⁷ Als weitere Prüfsteine für eine empirische Theorie nennt er *intellectual content* und *organization into an understandable form*, und er fährt fort: „By these tests mathematics is not a science, since its ultimate standard of validity is an agreed-upon sort of logical consistency and provability.“

2.5.5 *Instrumentalismus*

Die Messtheorie hat sich nicht nur auf den gerade beschriebenen Pfaden von der Realität entfernt. Darüber hinaus hat ein weiterer Faktor ganz erheblich zu einer Distanzierung der Messtheorie von der Empirie beigetragen: Viele Autoren im Bereich der Messtheorie vertreten dezidiert nicht-repräsentative Theorien des Messens.

Während eine repräsentative Theorie von einer von uns weitgehend unabhängigen Realität ausgeht und betont, dass deren Eigenschaften ebenfalls weitgehend unabhängig von unseren Messinstrumenten erfassbar sind, stellt sich der Messvorgang für einen Operationalisten völlig anders dar: Messinstrumente und Operationalisierungen sind nicht randständiger Natur, sondern von entscheidender Bedeutung für unser Verständnis der Realität. Genau diese philosophische Strömung wurde von Stevens vertreten. Bis heute hat sie eher an Bedeutung gewonnen als verloren.

Die operationale Theorie des Messens ist noch weit besser als die bislang genannten Strategien geeignet, die „Realität auf Distanz“ zu halten. Sie wird für gewöhnliche Bridgman (1927) zugeschrieben. Hand (1996: 453) fasst die zentrale Überzeugung des Operationalismus in einem Satz zusammen: „Thus, an attribute is defined by its measuring procedure, no more and no less, and has no ‘real’ existence beyond that.“ Zuvor zitiert er Dingle (1950), welcher ausführt:

Formerly science was regarded as the study of an external world, independent of the observer whose experiments and observations were simply means of finding out how the world was constructed and by what laws its behaviour was governed. The emphasis has now shifted from the nature of the world to the operations of experiment and observations. These are no longer regarded as more or less arbitrary means of discovering the already established order of nature, but rather as affording primary data for rational study; and any world that we may contemplate is no longer an independent existence whose nature demands or determines them, but rather a logical construct, formed and shaped and modified so as to afford a true picture of the relations which the observations exhibit.

„Im wissenschaftlichen Alltag ist eine Tendenz zum ‚lokalen‘ Konstruktivismus oder Operationalismus vorhanden, wenn der Forschungskontext experimentell geprägt ist. Wenn größere theoretische Zusammenhänge nicht wichtig oder gar nicht vorhanden sind, wird oft so getan, also ob eine Funktion operational definiert sei. Man akzeptiert im lokalen Kontext operationale Definitionen in dem Sinn, daß die mit gegebenen Meßmethoden ermittelten Werte fraglos als Werte für die Größen angesehen werden, die man messen möchte“ sagt Balzer (1997: 183). Bridgman (1927: 5) selbst schreibt:

[...] the concept is synonymous with the corresponding set of operations.³⁸

³⁸ Koch (1992) führt jedoch aus, dass es Bridgman vornehmlich um die Präzisierung bereits vorhandener Begriffe ging und definitiv nicht um die Festlegung eines Begriffs mithilfe eines gewissen Mess-Verfahrens. Er schreibt (ibd., S. 265, Hervorhebungen im Original):

Balzer führt sofort ein starkes Argument gegen eine solche „abhängige“ Art der Messung an. Es kommt zu einer Inflation von (theoretischen) Größen und Funktionen: „Dieser Ansatz hat jedoch weitreichende Konsequenzen für den globalen Aufbau der Wissenschaft, insofern für jede Größe oder Funktion jeweils genau eine Meßmethode ausgezeichnet ist, die die Größe *operational definiert*. Die anfängliche Einfachheit führt dadurch später zu größerer Komplexität. Es ist nämlich nicht möglich, verschiedene Meßmethoden zur Messung der gleichen Funktion zu benutzen.“

Wäre man nun konsequent, müsste man an dieser Stelle explizit Äquivalenzbedingungen formulieren, also definieren, wann eine Reihe von Messmethoden doch ein und dieselbe Größe erfassen. Dem ist jedoch nicht so, denn er fährt fort: „In mehr theoretischem Kontext will man dann aber von einer solch strengen Festlegung dann nichts mehr wissen.“ Andererseits ist man sich des Problems bewusst, denn „In der theoretischen Literatur wird sehr wohl die Problematik des einen oder anderen Begriffs, seiner Bedeutung, und seiner Meßmethoden erörtert [...] es wird versucht zu klären, was unter der zu messenden Funktion, der Bedeutung des ‚Begriffs‘, zu verstehen ist und man versucht zu begründen, wieso die gemessenen Werte dieser Bedeutung gerecht werden“ (Balzer 1997: 183).

Vollmer (1993), widmet ein ganzes Kapitel (Nr. 9) der Diskussion instrumentalistischer Interpretationen von Theorien, wobei insbesondere Invarianzargumente eine Rolle spielen: „In mathematischer Sprechweise sind alle die Wechselwirkungen, die schließlich zu irgendwelchen Sinneseindrücken führen, *Projektionen*, und jeder Wechsel der jeweiligen Perspektive ist eine *Transformation*. Diese Transformationen bilden Gruppen³⁹, und reale Objekte sind nichts weiter als die *Invarianten* solcher Gruppen“ (ibid., S. 173, Hervorhebungen im Original.) Balzer (1997: 183) schließt sich dieser Meinung an: „Demgegenüber beschreibt die vorher skizzierte Lösung die tatsächlichen Verhältnisse, wo *eine* Funktion, etwa die klassische Masse, durch *eine* Theorie implizit definiert, aber durch *viele* verschiedene Meßmethoden gemessen wird, weit besser.⁴⁰ Bei ihr wird die Identität von Werten, die nach verschiedenen Methoden gemessen wurden, zu einer empirischen Hypothese über verschiedene Meßmethoden.“

Zurückhaltender ist Hand (1996: 454ff), der einige prominente Beispiele aus den Sozialwissenschaften anführt, wo sich eben Messinstrument und Ge-

“Quite early on, for instance, the force of the notion of operation seems to shift from a set of activities *constitutive* of the ‘meaning’ of a concept to ones which help us *find* the meaning. Also, from an early point on, Bridgman repeatedly indicated that operations are necessary conditions for meanings, not *sufficient* ones. Moreover, he *stressed* - constantly and emphatically - that he was not advocating a theory or method of *definition* (or, still less, of concept formation), but only a method for analyzing, perhaps sharpening, the meanings of concepts *already in place*. He did not advocate ‘operational *definition*’ but rather ‘operational analysis’, or ‘operational method’.”

³⁹ Im mathematischen Sinne. Anm. des Autors dieses Buches.

⁴⁰ Anmerkung von mir: Bis auf den Bezug zu einer inhaltlichen Theorie, ist dies genau das, was die Messtheorie tut, wenn sie darangeht, äquivalente Skalen zu definieren.

messen es nicht so einfach trennen lassen. Wir hatten schon das Beispiel von Intelligenztests (siehe S. 17) erwähnt. Es dürfte klar sein, dass ohne weitere Maßnahmen Messwerte zunächst einmal sehr stark und substanziell von der speziellen Art der Messung abhängen (siehe das grundlegende Diagramm (1.1), S. 10). Es ist ja gerade die klassische - repräsentative und realistisch interpretierte - Messtheorie, welche versucht, hier mehr Klarheit zu schaffen. Sie wurzelt nicht zuletzt in dem ernsthaften Bemühen, Instrument und Objekt, Messvorgang und empirische Fakten, empirische Strukturen und Relationen zwischen Zahlen auseinander zu halten bzw. zu trennen.

Der Operationalismus erscheint genau dann gerechtfertigt, wenn man naiv bzw. schlecht misst. Ohne weitere Überlegungen oder Maßnahmen ist zunächst einmal anzunehmen, dass die Daten aus einer nicht überschaubaren Wechselwirkung von Objekt und Messung zustande gekommen sind. Anders gesagt sind erhebliche Anstrengungen experimenteller wie auch theoretischer Art notwendig, um über operationale Begriffe und Aussagen hinaus zu gelangen. Leider war es, wie Michell (1997) in seinem Buch ausführlich und überzeugend darlegt, in den Sozialwissenschaften nun so, dass der Operationalismus, auch von Stevens, ziemlich kritiklos akzeptiert wurde, letztlich wohl auch, um eigene Ergebnisse wissenschaftlich reputierlich präsentieren zu können. Wir stützen uns im folgenden wesentlich auf Kapitel 7 des vorgenannten Buches, welches den doppeldeutigen Titel „A definition made to measure“ trägt.

Der Operationalismus der Psychophysik

Die Geschichte spielte sich in den Grundzügen wie folgt ab: Harvard um 1930. Campbells (1920) klassische Theorie des Messens wird seit 1927, als das Mitglied der Physik-Fakultät, Bridgman (Nobelpreis 1946), seinen Operationalismus vorstellt, kontrovers diskutiert. Aus Europa kommt mit Feigl 1930 das erste Mitglied des Wiener Kreises nach Massachusetts. Als Positivist ist er von Bridgmans Auffassung begeistert.⁴¹ Die Psychologie, vom „oberflächlichen“ Behaviorismus geprägt, greift die neue Philosophie bereitwillig auf, und der Psychophysiker Stevens (1935) präsentiert eine an die Psychologie adaptierte Version. Stevens (1936a: 95) schreibt völlig unzweideutig: „Scientific psychology is operational and as such can have nothing to do with any private or inner experience for the simple reason that an operation for penetrating privacy is self-contradictory. Therefore, we need no longer think of immediate experience as the subject-matter of psychology.“

Damit ist für ihn der Weg frei, z. B. bei der Tonwahrnehmung, wie folgt zu argumentieren. “[...] the response of the observer who says ‘this is half as loud as that’ is one which, for the purpose of erecting a subjective scale, can

⁴¹ Etwa schreibt Schlick (1986: 171): „Der Sinn eines Satzes [ist] die Methode seiner Verifikation.“

be taken at its *face value*“ (Stevens (1936b), meine Hervorhebung.) Michell (1997: 169f) schreibt dazu:

This conclusion followed from his operationalism: one tone's being half as loud as another may be operationally defined by the operation used to determine it, i.e., by the subject judging it to be so [...] The relation of magnitude between sensations was seen as being defined by the judgement operation itself [...] The operationalist concludes that a ratio scale is obtained because of the character of the assignment rule, viz., the subject is instructed to judge ratios.

Im Ferguson-Komitee (1940) trafen die fundamental verschiedenen Standpunkte pointiert aufeinander, und Stevens psychophysikalische Skalen waren ein Kristallisationspunkt der Auseinandersetzung. Wie nicht anders zu erwarten, konnte keine Einigung erzielt werden, da für die Partei um Campbell nach wie vor galt: “not every assignment of numerals is measurement.” Bezüglich der Tonskalen bringt Michell (1997: 174) die Situation wie folgt auf den Punkt:

“For such a procedure to count as ratio scale measurement, according to the non-operationalist, subjects' responses must manifest a special sort of empirical structure [...] Thus, to take numerical judgements as constituting measurements at face value, in according with operationalism, is to resort to postulating what one wants instead of looking to see what is the case.” Dies wertet Michell (1997: 172) wie folgt: “This is a way of thinking which non-operationalists would adjudge viciously circular, but that is because they presume the existence of independent attributes and regard as trivial those which can *only* be defined via rules for making numerical presumptions.” (Hervorhebung im Original.)

In einer Reihe von Artikeln zeigt Michell (2000, 2002, 2003a, 2004, 2005), dass der Operationalismus im Allgemeinen und Stevens' Erbe im besonderen - zumindest in der Psychologie - noch lange nicht überwunden ist. Im Abstract von Michell (2002) heißt es: “Stevens's theory of measurement has been an important methodological resource within psychology for half a century. It advanced the representational theory of measurement and promised to open up to scientific investigation the issue of the structure of psychological attributes [...] However, neither Stevens nor the psychometricians who adopted his theory used it in that way. They used it to foreclose scientific investigation of that issue.”

Es spricht nichts dagegen, in philosophischen Werken zur Messtheorie, z. B. Savage und Ehrlich (1992), die klassische, repräsentationale Auffassung des Messens neben eine „liberale“ Auffassung von Messen zu stellen. Die klassische Auffassung wurde ja zurecht als zu eng kritisiert, und es ist oftmals nicht gerechtfertigt, den Einfluss der Operationalisierung oder des Instruments einfach zu vernachlässigen. Selbst wenn die Operationalisierungen in den „weichen“ empirischen Wissenschaften nicht mit den Präzisionsmethoden der zeitgenössischen Naturwissenschaften konkurrieren können - es hieße alle quantitativen Bemühungen in den Sozialwissenschaften zu diskreditieren, dürften jene nicht von „Messungen“ sprechen. Jedoch: Die Maxime zu messen was messbar ist und messbar *zu machen*, was noch nicht messbar ist,

kann auch so interpretiert werden, dass man qualitativ hochwertige Messungen nicht geschenkt bekommt. Das heißt, man sollte nicht gleich bei jeder Zuordnung von Zahlen zu empirischen Sachverhalten von einer Messung zu sprechen. Vielmehr sollte man immer überprüfen, welche Eigenschaften das Messinstrument hat, welche Fehlerquellen vorhanden sind, wie die Übereinstimmung mit anderen Methoden ist und wie sich garantieren lässt, dass man das misst, was man messen will. Insofern ist Stevens Position des “anything goes” tatsächlich sehr liberal.

Was für einen Wissenschaftler aber auf keinen Fall zulässig sein darf, ist, aus Bequemlichkeit einer operationalistischen Philosophie zu folgen und infolge dessen darauf zu verzichten, empirische Hypothesen auch empirisch zu überprüfen. Tut man dies, so verlässt man den Rahmen neuzeitlicher Wissenschaft. Zwar ist es akzeptabel, zunächst einfach anzunehmen, ein z. B. psychologisches Attribut habe eine gewisse Struktur (abhängig oder unabhängig vom Beobachter). Der nächste Schritt muss dann jedoch sein, eine solche Hypothese systematisch und kritisch zu testen. Stellt die Messtheorie hierfür geeignete Werkzeuge zur Verfügung - und das tut sie - so ist es die Pflicht der Experimentatoren, jene auch zu nutzen! Michell (2002) bezeichnet es deshalb zurecht als einen Skandal, dass die Messtheorie vorwiegend zur *Rechtfertigung* von Messverfahren benutzt wird, anstatt sie als *Werkzeug zur Überprüfung* empirischer Hypothesen einzusetzen. Gerade auch bei den Skalenniveaus der Messtheorie handelt es sich um empirische Hypothesen, welche von Fall zu Fall zu überprüfen sind. Einfach nur anzunehmen, es läge gerade das Niveau vor, welches man gerade benötigt, heißt, via Zitat das eigene Tun zu rechtfertigen, anstatt empirische Evidenz vorzulegen. Michell (2000: 639) fasst schon ganz zu Beginn seines Beitrags wie folgt zusammen: “A pathology of science is defined as a two-level breakdown in the process of critical inquiry: first, a hypothesis is accepted without serious attempts being made to test it; and, second, this first-level failure is ignored” (siehe auch ibd., S. 641 und 648ff).

2.6 Forschungsstrategien I (Deduktion und Induktion)

[...] denn die Ueberzeugung beruht in allen Dingen entweder auf Schlüssen oder auf der Induktion. (Aristoteles 1877: 142)

Tukeys Kritik (S. 41) an Lucas Standpunkt (S. 35) besteht tatsächlich aus zwei Teilen. Sein zweites Argument lautet: “A lack of appreciation that all useful conclusions are not fundamental” (Tukey 1986b: 246). Diese Kritik ist tiefgründig, und Tukey widmet ihr im schon genannten Sammelband mehrere weitere Artikel. Schon deren Titel deuten an, worauf es ihm ankommt: “Conclusions vs. decisions” (1960b) und “Analyzing data: sanctification or

detective work” (1969).⁴² Letztlich geht es um prinzipiell verschiedene Herangehensweisen:

2.6.1 Deduktives Vorgehen

Tukey (1961: 153) beschreibt die Welt der Mathematik: “[Mathematicians] work with *abstract, uncertain* entities (like a continuous function defined on $[0, 1]$) just so long as the *assumptions* about them which are to be used are *unquestioned* (like continuity). Unquestioned assumptions lead to secure conclusions. It is only in mathematics that a deep intellectual challenge and high intellectual standard are combined with security of conclusion [...] to most mathematicians, a good problem is one that is carefully and crisply formulated, one that has a clear-cut answer [...] It is by insistence upon clear-cut formulations and by solving clear-cut problems, that mathematics progresses, and constructs its imposing and important structures.”

Wie wir schon sehr früh (S. 6) ausgeführt haben, ist diese Arbeit von großem Nutzen für die empirischen Wissenschaften. Tukey (1961: 154) erläutert: “Specific results [of mathematics] are indeed useful. But aid in formulating problems [...] is more important, as are both (i) assistance in crisper thinking, and (ii) the provision of an adequate supply of parts and schemes from which enough *different*, alternative, mathematical models can be constructed.”

Dabei darf man jedoch nicht vergessen, dass die mathematische Art zu denken und zu argumentieren ziemlich eingeschränkt und einschränkend ist. (Tukey 1961: 153) sagt dazu: “How does one make a clear-cut problem? By choosing very precise and clear-cut hypotheses. And how does one solve problems completely? By wringing everything possible out of these hypotheses. These are basic mathematical techniques, essential in their place. Every clear-cut problem is artificial, separated from the real world by idealization after idealization.”

Und genau dies wird laut Tukey (1961: 154f) zu einer Gefahr, falls man solche Methoden auf die empirischen Wissenschaften überträgt: “The danger of mathematics to the outside world in general, and to science in particular, is simple. Pure mathematics must take its assumptions most seriously, [...] questioning them not at all. Pure mathematics must value its results in its own terms, with far less attention to the relation of the assumptions to the real world than to the aesthetic nature of the results. Upon such actions hang its speedy progress. Yet these are just what science and technology must *not* do. Science and technology [...] must use the products, and avail itself of the aid of mathematics, yet dare not accept its attitudes.”

⁴² Das Thema wird auch in Tukey (1962, 1986b) ausführlicher behandelt. Alle Hervorhebungen in den nun folgenden ausführlichen Zitaten finden sich im jeweiligen Original.

Er plädiert deshalb für eine emanzipierte Einstellung der Mathematik gegenüber, welche der Erfahrung den höchsten Wert beimisst: "Science makes, as it must, ever-growing use of mathematical models, with whose power inevitably come the dangers of overacceptance, of transformation of assumptions into axioms, of placing mathematical models above experience (rather than beside or below it). The older sciences have learned to make powerful use of mathematical models [...] ⁴³ without letting these mathematical models use them. The newer branches of science and technology must learn to do the same." In einem präzisierten Sinne wiederholt Tukey damit Galilei (S. 3) bzw. das Programm der neuzeitlichen Wissenschaften.

Aus alledem lässt sich unschwer seine Kritik an der Messtheorie ableiten und schärfen: Weil die Messtheorie von ihren Protagonisten im mathematischen Stil betrieben wird, werden aus hypothetischen Annahmen Axiome und die aus ihnen hergeleiteten Ergebnisse stehen als normative Forderungen an die empirischen Wissenschaften im Raum.

Ein solches Vorgehen führt umgehend zu prinzipiellen Diskussionen, was richtig und bedeutsam oder falsch und unzulässig ist. Es führt jedoch nicht zu fruchtbaren empirischen Resultaten, letztlich weil es gegen Galileis *Maxime* (S. 42) verstößt, dass sich unser Geist und seine Werkzeuge der Natur anzupassen haben und nicht umgekehrt. Jede Theorie, die "mathematics rather than science" ist, die also ihre Annahmen kaum hinterfragt, wohl aber auf die Strenge der Herleitung ihrer Ergebnisse größten Wert legt, ergeht es ebenso wie der Messtheorie: In vielen Fällen ist sie schlicht nicht anwendbar - d.h. ihre Prämissen sind nicht adäquat - oder aber, wenn man sie doch anwendet, passen ihre Vorhersagen nicht zu den empirischen Befunden. Es ist die Überbetonung des (logisch-strengen, mathematisch-exakten) Weges von den Prämissen zu den Folgerungen, welche zu einer Unwucht führt. Die for-

⁴³ Hier erwähnt Tukey (1961: 154) ausdrücklich die Physik. Liest man herausragende Physiker wie z. B. Jaynes (2003), insbesondere den bereits erwähnten Anhang B oder Feynman (2007), so ist dieses Selbstbewusstsein - basierend auf exzellenten mathematischen Fertigkeiten - durchgängig zu spüren. Feynman widmet ein ganzes Kapitel der Beziehung zwischen Mathematik und Physik. Dabei unterscheidet er zwischen einer „Euklidischen Methode“, welche die Systematik, also die Axiome samt der Ableitung der Theoreme in den Mittelpunkt stellt, und einer „babylonischen Methode“ (S. 60ff), welche viele Resultate und ihre Beziehungen einfach nur nebeneinander stellt und bei praktischen Problemen einfach den Formalismus verwendet, der gerade nützlich ist. Feynman schreibt ausdrücklich: „Die Rolle der Mathematik in der Physik [...] ist gar nicht zu überschätzen (S. 49). Zugleich lässt er keinen Zweifel daran, wie mathematische Verfahren zu verwenden sind: „In der Physik brauchen wir die babylonische Methode und nicht die Euklidische [...] (S. 62). Er fährt fort (S. 74): „Mathematischer Rigorismus von großer Präzision ist in der Physik nicht unbedingt empfehlenswert.“ Denn: „Wer die Physik verstehen will, muß sich aufs Abwägen verstehen, [...] weil die Gesetze oft über den Bereich ihrer Ableitung hinaus gelten (S. 65).“ Und auf derselben Seite folgert er: „Wir haben also umfassende [physikalische] Prinzipien, die verschiedene Gesetze umschließen, und können die Beziehungen der verschiedenen Zweige der Physik untereinander nur verstehen, wenn wir ihre [mathematische] Ableitung nicht allzu ernst nehmen (Hervorhebung von mir).

male Stärke der Theorie wird dann durch ihre empirische Schwäche mehr als aufgewogen. Marc Twain (2006) könnte hierzu mit scharfer Zunge bemerken:

It ain't what you don't know that gets you into trouble. It's what you know for sure that just ain't so.

Ein lehrreiches Beispiel hierfür ist die Protophysik, ein konstruktivistischer Ansatz, welcher zum Ziel hat, die wichtigsten physikalischen Theorien aus Logik, Mathematik und sogenannten „idealen Normen“ herzuleiten. Sie ist von ihrer Grundeinstellung her noch extremer als die Messtheorie in ihrer strengsten Interpretation. Lorenzen (2000: 210f), einer ihrer Hauptvertreter, schreibt: „Die wahren Sätze der Protophysik sind solche Sätze, die auf der Basis von Logik, Arithmetik und Analysis, Definitionen *und* den idealen Normen, die Messen ermöglichen, verteidigbar sind. Diese idealen Normen sind deutlich unterschieden von den Konstruktionsregeln der Arithmetik, und sie sind keine Definitionen [...] Wir schreiben mit Normen vor, wie sich die Materie ‚verhalten‘ soll [...] Im Gegensatz zu material-analytischen Bestimmungen, wo wir Regeln für unsere Prädiktoren vorschreiben, um sie der Welt anzupassen, zwingen wir nun die Materie, sich unseren idealen Normen anzupassen.“ Es ist dann nur konsequent, dass „[...] diese protophysikalisch wahren Sätze, die von den idealen Normen abhängen, ‚mathematisch-synthetische‘ Wahrheiten genannt werden können.“ (Hervorhebung im Original.)

Wir verzichten darauf, abermals auszuführen, dass eine solche Auffassung in diametralem Widerspruch zur Grundhaltung jeder empirischen Wissenschaft steht und zitieren lediglich Feynman (2007: 181): „Nicht hochtrabende Vorschriften sind unabdingbar ‚für die Existenz der Naturwissenschaften‘ oder für die Bestimmung der Eigenschaften der Natur. Diese werden stets durch unseren Werkstoff bestimmt, durch die Natur selbst.“ Lorenzen (2000: 200) räumt mit Blick auf die moderne Physik denn auch zurecht ein: „Die herrschende Meinung der Mathematiker ist seitdem, daß die euklidische Geometrie nur eine unter mehreren - mathematisch gleichberechtigten - ‚Geometrien‘ sei. Die Auszeichnung einer dieser ‚Geometrien‘ als *der* Geometrie des ‚wirklichen‘ Raumes wird der Physik überlassen.“

Interessant ist die Frage, wie die Protophysik versucht, ihrer allfälligen empirischen Widerlegung zu entgehen. Zum einen bezweifelt sie ganz direkt die entscheidende Bedeutung empirischer Methoden: „Einwände gegen diesen Anspruch der empirischen Physik, über die ‚Struktur‘ von Raum und Zeit entscheiden zu können, sind schon unmittelbar nach dem Bekanntwerden der Relativitätstheorie (um 1920)⁴⁴ vielfach erhoben worden - und von der konstruktiven Wissenschaftstheorie unter dem Titel ‚Protophysik‘ seit 1960.“ (Lorenzen 2000: 200)

Zum anderen lässt sich gerade bei einem Effekt wie der sogenannten Längenkontraktion der Relativitätstheorie trefflich argumentieren, dass es nicht der Maßstab, mit dem gemessen wird, ist, welcher sich verkürzt; sondern dass sich die Länge des gemessenen Objektes selbst ändert. Man ist gezwungen,

⁴⁴ Meine Anmerkung: Siehe insbesondere H. Dingler, der sich nicht zuletzt auf Kant beruft.

so zu argumentieren, wenn man zuvor - per Norm - bestimmt hat, welche Eigenschaften die Längenmessung besitzen muss.

Lassen die eigenen Normen schließlich einen realen physikalischen Effekt, wie z. B. die Krümmung des Raumes, nicht zu, so bleibt einem nur der schon bekannte Ausweg, die entsprechende theoretische Größe (einen gewissen Tensor in den Einsteinschen Feldgleichungen) lediglich mathematisch zu interpretieren. Daraus ergibt sich die skurrile Situation, dass man eine physikalische Theorie, welche man - anders als oftmals die Messtheorie - problemlos empirisch anwenden *kann*, nicht empirisch interpretiert, weil eben dies unseren eigenen Vorurteilen, also den a priori vorgenommenen Setzungen, widerspricht.

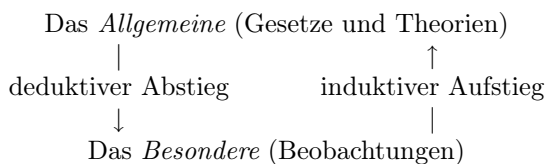
2.6.2 Deduktiv-induktives Schema

Die vorgestellten Beispiele sollten eines deutlich gemacht haben: Jede deduktiv-strenge Grundhaltung, die ihre Grundannahmen sehr ernst nimmt, aber zugleich wenig hinterfragt, tendiert dazu, Ergebnisse zu beweisen, welche als normative Forderungen formuliert werden. Wie nicht anders zu erwarten, kollidiert diese Haltung regelmäßig mit der Wirklichkeit, und dies umso heftiger, je weniger die Grundannahmen als zu überprüfende Hypothesen, sondern als nicht weiter zu hinterfragende Annahmen - Setzungen, Axiome - behandelt werden. Ein rigides System, welches seine Grundfesten nicht mit der Realität abgleicht, wird häufig unpassend sein, und zuweilen verfallen Theoretiker dann auf die (ironisch gemeinte) Lösung Herbert Marcuses: „Wenn die Tatsachen mit der Theorie nicht übereinstimmen – um so schlimmer für die Tatsachen.“ Ein derart isoliertes System ist dann zwar bestenfalls noch mathematisch-exakt, nicht aber mehr empirisch relevant.

Die axiomatischen Theoriengebäude der auch empirisch höchst erfolgreichen Physik werden deshalb zurecht bewundert. Sie vereinigen höchste formale Strenge mit nahezu perfekter empirischer Passung. Doch wie bei jedem axiomatischen System zahlt man einen Preis: Es handelt sich um eine geschlossene Welt, in die - via Grundannahmen - alles hineingesetzt werden muss, was später herauskommen kann (bzw. soll). Man findet also bildlich gesprochen nur dass wieder, was man zuvor als Struktur (wenn auch implizit, in den Axiomen) vorausgesetzt hat. Ist diese Struktur der realen Welt nicht angemessen - und das ist leider die Regel, nicht die Ausnahme - so hat man bestenfalls blutleere Mathematik betrieben, und darf darauf hoffen, dass die untersuchten formalen Zusammenhänge irgendwann später, in einer anderen Situation, anwendbar sein werden. Ist das formale System noch dazu eng, so entwickelt sich schnell eine mit Scheuklappen versehene Geisteshal-

tung, welche Shakespeare⁴⁵ mit den Worten “home-keeping youth have ever homely wits” verspottet.

Dass man hier ein Gegengewicht benötigt, ist offensichtlich und zumindest in den empirischen Wissenschaften unbestritten. Es gereicht der Philosophie zur Ehre, dass bereits Aristoteles ein passendes Modell entwickelt hat, das beiden Erkenntnisrichtungen gerecht wird und unter dem Namen deduktiv-induktives Schema auch heute noch jedem Wissenschaftstheoretiker geläufig ist.⁴⁶



Wie das Motto dieses Abschnitts zeigt, war Aristoteles von der Allgemeingültigkeit des Schemas überzeugt. Unzählige Philosophen haben sich seitdem dezidiert und vom Grundton her eher kritisch bis ablehnend mit dem rechten, induktiven Teil des Schemas auseinander gesetzt. Es hieße deshalb eine Geschichte der Philosophie bis in unsere Tage zu schreiben, ginge man auf die Einzelheiten der Diskussion um das „Induktionsproblem“ ein. Jeffreys (1973: 13) gibt eine ernüchternde Zusammenfassung:

This statement is at least as old as Plato, and was emphasized by Hume at considerable length. There is an element of uncertainty in all inferences of the kind considered. Most scientists have ignored Hume's difficulty. Some philosophers have overemphasized it by declaring all sensory information unreliable. Others have claimed to answer it, but all the answers depend on elementary mistakes [...]

Wir werden später, in Kapitel 4, insbesondere in Abschnitt 4.1, ausführlich darauf zurückkommen. Trotz aller prinzipieller Kritik gehen wir davon aus, dass Induktionsschlüsse möglich und sinnvoll sind.⁴⁷

⁴⁵ The Two Gentlemen of Verona (Die beiden Veroneser), 1. Akt, 1. Szene.

⁴⁶ Die nachstehende Abbildung wurde Schurz (2006: 49) entnommen.

⁴⁷ Als Evidenz hierfür sei an dieser Stelle nur angeführt, dass wir im Alltag ständig von ihr Gebrauch machen. Alle empirischen Wissenschaften stützen sich maßgeblich auf induktive Schlüsse. Ihre historische Entwicklung war ebenfalls induktiv: Sie begann bei der Beschreibung schritt weiter zur Systematisierung und erreichte ihren Höhepunkt in Erklärung, mithin war also eine fortschreitende Strukturierung - und damit zwangsläufig auch Mathematisierung - zu erkennen. Aus isoliertem bruchstückhaftem Wissen entstanden zusammenhängende Ansätze und schließlich umfassende Theorien. Erst recht zeigt sich die Bedeutung der induktiven Richtung, wenn man nicht nach Naturgesetzen sucht, sondern selber die Regeln vorgibt: Juristen sprechen von der normativen Kraft des Faktischen, Softwareentwickler setzen erst Standards, nachdem man Erfahrungen auf einem neuen Gebiet gesammelt hat, Kreativitätstrainer haben den Slogan “Storming, Norming, Performing” geprägt und schließlich ist der Volksmund davon überzeugt, dass probieren über studieren gehe.

2.6.3 Induktives Vorgehen

Historisch gesehen entwickelten sich die erfolgreichen empirischen Wissenschaften, gerade auch die Physik, nicht deduktiv sondern induktiv. Das heißt, sie sammelten zunächst möglichst zahlreiche Erfahrungen und gingen erst danach ans Systematisieren und Formalisieren. Auch der Messvorgang selbst, ihr wohl wichtigstes Hilfsmittel, macht hier keine Ausnahme. Wie nicht nur Duncan (1984) und Tukey (1969, 1986b) ausführen, wurden jahrhundertlang interessante Größen benannt und so gut als möglich gemessen. Erst im 19. Jahrhundert begann die Formalisierung des Messvorgangs. Große Namen sind mit dieser Anstrengung verbunden: Die Arbeiten von von Helmholtz (1887), Kelvin (1891) und Hölder (1901) gipfelten schließlich im Werk von Campbell (1920, 1928).⁴⁸ Tukey (1986b: 248) nennt das so erzielte Ergebnis „Monastic Measurement. Just as some have done for mathematics, measurement may be divided into ‘monastic’ and ‘secular’. The analogy of the ‘high church’ view, which we naturally call the ‘high monastery’ view is surely that provided by Norman R. Campbell [... His books] have been the source, proximate or remote, of many fears that assignment of numbers, many of which would have been perfectly useful, were not ‘measurements’ [...]. There is little doubt that measurement which fulfills Campbell’s requirements, exactly or approximately, is measurement which deserves the highest social status, the highest prestige that we can today imagine.“⁴⁹

Genau diese Reihenfolge, also gezielte empirische Erfahrung, gefolgt und begleitet von einer immer präziseren Fassung der zentralen Erkenntnisse, war zuverlässig erfolgreich.⁵⁰ Es ist genau diese Art von Mathematisierung, die als wichtige⁵¹ Tendenz in allen entwickelten Wissenschaften auszumachen ist. Im Prinzip⁵² spricht jedoch nichts dagegen, die Abfolge der Ereignisse umzukehren. Warum sollte man nicht bereits erzielte mathematische Kenntnisse nutzen, sie womöglich geeignet modifizieren oder ausbauen, und sie dann auf „jüngere“ Wissenschaften anwenden?

Wie hinreichend bekannt ist, ging gerade bei der erfolgreichsten Symbiose, jener zwischen Mathematik und Physik, zuweilen die mathematische Ent-

⁴⁸ Für eine ausführliche Darstellung der Geschichte des Messens, insbesondere in den Sozialwissenschaften, siehe z. B. Michell (1997, 1999, 2003a, 2005).

⁴⁹ Auch Duncan (1984) beschäftigt sich eingehend mit der Geschichte des Messens. Eine seiner wohlplatzierten Kritiken an Stevens lautet, dass auch ein zu *liberaler* Begriff des Messens, wie er in Stevens’ berühmter Definition zu Ausdruck kommt, schädlich ist. Man nimmt dem Begriff einen Gutteil seines Wertes, wenn man nahezu jeder Zuordnung von Zahlen zu Objekten als „Messung“ bezeichnet. Am Beispiel der Nominalskala führt er (ibid., S. 129ff) dies im Detail aus.

⁵⁰ Wir verweisen beiläufig auf Bacon (1620: Buch 1, Aphorismus I, erster Satz), dem auch das Motto von Kapitel 6 entnommen ist: „Man, being the servant and interpreter of Nature, can do and understand so much and so much only as he has observed in fact or in thought of the course of nature.“

⁵¹ - womöglich sogar wichtigste -

⁵² Siehe das auf Aristoteles zurückgehende Schema

wicklung der physikalischen Anwendung voraus. Von Neumann (1947: 191) nennt zwei konkrete Fälle: “Two strange examples are given by differential geometry and by group theory: they were certainly conceived as abstract, nonapplied disciplines and almost always cultivated in this spirit. After a decade in one case, and a century in the other, they turned out to be very useful in physics.” In vielen Fällen müssen sogar zuerst die mathematischen Werkzeuge vorliegen, um dann mit ihrer Hilfe bislang unerreichbare Probleme mit Aussicht auf Erfolg bearbeiten zu können.

Liest man die frühen Autoren der Messtheorie in den 1950er Jahren, so ist genau diese Orientierung an der Geschichte des Messens und den etablierten Naturwissenschaften zu spüren. Sie sind beeindruckt von den Erfolgen der Physik, nicht zuletzt von der von Einstein zu einem (weiteren) glänzenden Erfolg geführten Invarianzidee. Warum sollte diese so einfache wie durchschlagende Idee nicht auch in den anderen empirischen Wissenschaften von großer Bedeutung sein? Wären nicht genauso revolutionäre wie tiefe Erkenntnisse zu erwarten, adaptierte man die zentralen Überlegungen nur geeignet an die Verhältnisse in den weniger entwickelten Wissenschaften? Dieser grundlegende Optimismus ist historisch verständlich, und wäre dem Ansatz durchschlagender Erfolg beschieden gewesen, so wäre damit ein „großer Sprung“ direkt aus der „Steinzeit“ rudimentärer Theorien - basierend auf ziemlich zweifelhaften Messungen - hin zu Modellen “of the highest social status” (Tukey) gelungen. Tatsächlich muss Cliff (1992: 187) jedoch feststellen: “[...] regrettably, the promise of the axiomatic measurement movement has not been fulfilled,” und der Titel seines Aufsatzes bewertet die Messtheorie ganz klar als eine “revolution that never happened.”

Egal welchen Ansatz man wählt - den typischen, also ausgehend von empirischen Phänomenen zu formalisieren oder den selteneren, also ausgehend von mathematischen Theorien hin zu möglichen Anwendungsfelder - entscheidend ist, konsequent den Kontakt zur Empirie zu suchen und zu halten. Dies gilt selbstredend ganz besonders für jede praxisrelevante Theorie des Messens. Wie wir gesehen haben, ist die Messtheorie im Wesentlichen deswegen gescheitert, ihre Revolution im Ansatz verpufft, weil sie dem deduktiven Zweig der Erkenntnis zuviel Bedeutung beigemessen hat.⁵³ Die Frage drängt sich deshalb auf, welche Gestalt eine eher *induktiv motivierte* Messtheorie hätte.

2.7 Messtheorie (induktiv)

[Induction:] The glory of science and the scandal of philosophy.

C. D. Broad, zitiert nach Ladyman (2002: 52)

Der wesentliche Unterschied ist, dass aus Axiomen Hypothesen werden. Alle Annahmen, die in der mathematisch orientierten Messtheorie so selbst-

⁵³ Wir erinnern an Jaynes (2003: 674ff), siehe S. 45.

verständlich getroffen werden, sind bei jeder Anwendung auf ihre empirische Gültigkeit hin zu prüfen. Genau dies ist die nachdrücklich vorgebrachte Forderung Michells. Es genügt nicht, zu Beginn einer Untersuchung einfach zu behaupten, die Daten befänden sich auf einem gewissen Niveau oder die Messtechnik stelle sicher, dass eine gewisse Struktur vorliegt. Vielmehr sind gewichtige Gründe, also letztlich empirische Belege und stringente theoretische Argumente, vorzuzeigen, weshalb ein Messverfahren angemessen ist und eine Skala die Eigenschaften hat, welche man postuliert.

Oftmals wird dann die mathematisch motivierte, grobe Einteilung in wenige Skalenniveaus nicht ausreichen. Passt keines der von Stevens vorgeschlagenen Modelle, so helfen womöglich die von Krantz, Luce, Suppes, Tversky, Narens und anderen Messtheoretikern entwickelten axiomatischen Systeme weiter. Falls nicht, so ist ein neues Modell zu entwickeln, oder aber der deterministisch-enge Rahmen der Messtheorie ist ganz zu verlassen.

Es geht also viel weniger um eine theoretische Verfeinerung oder Präzisierung potenziell möglicher empirischer und numerischer Relative, als um eine angemessene Beschreibung realer Messungen. Dabei ist es definitiv nicht sinnvoll, ein Messmodell, das dem konkreten Messvorgang nicht gerecht wird, „trotz alledem“ zu verwenden. Statt sich subtil oder in aller Deutlichkeit von der Realität zurückzuziehen, sucht eine so verstandene Messtheorie intensiv den Kontakt zu ihrem Gegenstandsbereich, also den in der Praxis verwendeten Messverfahren.

Induktiv vorzugehen heißt, sich flexibel an wechselhafte Bedingungen anzupassen; Modelle zu entwickeln, die der Realität gerecht werden. Passt ein Modell nicht, so ist dieses geeignet zu modifizieren oder durch ein anderes zu ersetzen. Man muss offen sein für das Unerwartete, Unpassende, Überraschende. Es sind genau diese Herausforderungen des Alltags, der wissenschaftlichen Praxis, die eine Theorie wachsen (zuweilen aber auch sterben lassen) lassen. Entzieht sie sich ihnen, so degeneriert sie zu einer praktisch ziemlich irrelevanten Spielerei von Spezialisten. (Man denke an die Worte von Neumanns, S. 46, und Galileis, S. 42.)

Neben der (unmotivierten) Mathematisierung bzw. dem Rückzug auf die mathematische Theorie haben wir als zweite wesentliche „Distanzierungsstrategie“ den Operationalismus besprochen. Hier nun führt ein Wechsel der Blickrichtung zu völlig neuen Einsichten. D. h., statt den Operationalismus als (letzte bzw. bequeme) Zuflucht zu sehen, verstehen wir ihn im folgenden als ausbaufähigen Ausgangspunkt, der zunächst einmal von nicht-äquivalenten Verfahren ausgeht. Systematische Forschung sollte dann versuchen, diese wenig befriedigende Situation sukzessive zu verbessern. Denn warum scheitert die Idee äquivalenter Skalen, also vergleichbarer Messverfahren, häufig in der Praxis? Wohl nicht zuletzt aus dem einfachen Grund, dass verschiedene Verfahren *nicht* äquivalent sind. Hand (1994) diskutiert ein elementares Beispiel,⁵⁴ das wir hier zu Demonstrationszwecken unverändert übernehmen:

⁵⁴ Sein *Example 3*

Zwei Forscher (ein Engländer und ein Franzose) untersuchen eine Reihe von Autos A_i auf ihre „Effizienz“. Versteht man darunter die Wegstrecke, die mit einer bestimmten Menge Kraftstoff zurückgelegt werden kann, so ergibt sich die Auswertung in der ersten Zeile; interpretiert man „Effizienz“ hingegen als die Kraftstoffmenge, welche benötigt wird, um einen vorgegebene Strecke zu bewältigen, so kommt man zur Auswertung in der zweiten Zeile:

Effizienz	Gruppe A				Gruppe B			
Forscher:	A_1	A_2	Summe	$\varnothing$	A_3	A_4	Summe	$\varnothing$
English (miles per gallon)	1	4	5	2,5	2	2	4	2
French (gallons per mile)	1	0,25	1,25	0,625	0,5	0,5	1	0,5

Da die Autos der ersten Gruppe im Mittel weiter kommen als die Autos der zweiten Gruppe ($2,5 > 2$), wird der englische Forscher ersteren den Vorzug geben. Zum genau gegenteiligen Ergebnis kommt der französische Forscher, da die Autos der zweiten Gruppe im Mittel weniger Kraftstoff als die Autos der ersten Gruppe verbrauchen ($0,5 < 0,625$).

Dieses Ergebnis ist wenig befriedigend. Hand gibt eine Reihe von Interpretationen, eine davon (ibid., S. 324) ist der Messtheorie entlehnt: Die eigentlich interessierende Eigenschaft der Fahrzeuge ist „Effizienz“, welche auf die zwei genannten Arten gemessen wird. Nimmt man an, dass die gemessenen Größen (also „Englische Effizienz“ in miles per gallon bzw. „Französische Effizienz“ in gallons / mile) monoton von der (latenten) Eigenschaft „Effizienz“ abhängen, so sollte eine bedeutungsvolle Aussage invariant unter monotonen Transformationen sein. Das arithmetische Mittel ist nicht invariant unter monotonen Transformationen (siehe Ungleichung (2.4), S. 34), also ist es nicht angebracht, diese Statistik zu verwenden. Sie führt zu verfahrensabhängigen und damit letztlich bedeutungslosen Ergebnissen, wie die obige Auswertung nur allzu deutlich zeigt.

Bei dieser Vorgehensweise hat man - im Sinne der Messtheorie - die beiden Operationalisierungen als äquivalent betrachtet. Deshalb sind nur diejenigen Statistiken bedeutungstragend, die unter allen zugehörigen Transformationen invariant bleiben.

Eine andere Lösung, vorgetragen vom „Gentleman“ Hand, besteht darin, sich für eines der beiden Effizienzmaße zu entscheiden. Im konkreten Fall hält er die benötigte Kraftstoffmenge je Strecke für angemessener und gibt damit dem Ansatz des französischen Forschers den Vorzug. Eine solche Lösung setzt voraus, dass man die beiden Perspektiven als nicht gleichwertig betrachtet. Und genau diese Sichtweise scheint auch uns angemessener zu sein, was sehr deutlich wird, wenn man die ungebräuchlichen „Effizienzen“ des Beispiels einmal durch die weit anschaulichere Geschwindigkeit in m/s bzw. ihren Kehrwert (Einheit: s/m) ersetzt.

Mit den obigen Zahlenwerten besteht Gruppe A dann aus zwei unterschiedlich schnellen Fahrzeugen. Der Ansatz des englischen Forschers besteht darin, die Zeitspanne (z. B. eine Sekunde) festzuhalten und zu ermitteln, wie weit

die Fahrzeuge in dieser Zeit kommen. Er stellt fest, dass Gruppe A insgesamt 5 Meter zurücklegt, während Gruppe B lediglich vier Meter weit kommt. Hier gibt also das schnellste Fahrzeug A_2 den Ausschlag, und der englische Forscher entscheidet sich für Gruppe A. Die Perspektive des französischen Forschers ist es, den Zähler, nun also die Länge der Teststrecke (z. B. einen Meter) festzuhalten, und zu messen, wieviel Zeit die beiden Fahrzeuge zusammen (bzw. gemittelt) benötigen. Gruppe A benötigt für die Teststrecke insgesamt 1,25 Sekunden, was länger ist als Gruppe B (1 Sekunde). Das langsamste Fahrzeug A_1 gibt bei seiner Art der Messung den Ausschlag, und der französische Forscher entscheidet sich folgerichtig für Gruppe B.⁵⁵

Beide Verfahren sind also fundamental verschieden. Sie sind beide auf ihre Art *sinnvoll*, sollten jedoch *nicht* als *äquivalent* betrachtet werden. Anders gesagt: die verschiedenen Perspektiven sorgen für einen Unterschied - sie erfassen verschiedene Aspekte von „Effizienz“ bzw. „Schnelligkeit“. Mit beiden Verfahren gewinnen wir wichtige, wenn auch substantiell verschiedenartige Informationen über die Fahrzeuge. (Genau dies sollte man auch völlig unvoreingenommen erwarten: Andersartige Verfahren liefern substantiell verschiedene Daten.) Nur den Teil der Information auf die Fahrzeuge zurückzuführen, welcher sich bei allen Messverfahren zugleich zeigt, ist außerordentlich vorsichtig. Doch genau das ist die Haltung der Messtheorie, welche eine Eigenschaft nur dann der Empirie zuschreibt, wenn sie sich auf allen Skalen zeigt.⁵⁶

Typischerweise wird es in der Forschung so sein, dass sich ein spannender neuer Effekt nur mit einem einzigen Verfahren „gerade eben noch so“ - also am Rand der Messungenauigkeit - feststellen lässt. Bei einem einzigen Verfahren lässt sich zudem prinzipiell nicht unterscheiden, ob eine Information auf das Objekt der Messung oder das Verfahren zurückzuführen ist. Deshalb ist es zum einen - kritisch bzw. konservativ gedacht - sehr sinnvoll, zu fordern, dass das Resultat mit einer anderen Methode repliziert werde. Andererseits ist es im Sinne wissenschaftlichen Fortschritts aber auch nicht angebracht, in solchen Zweifelsfällen immer für einen möglichen Artefakt zu argumentieren und nie in Betracht zu ziehen, dass der Effekt nicht von der Perspektive, sondern der Realität herrührt.⁵⁷

⁵⁵ Die Unterschiede werden ganz deutlich, wenn man sehr heterogene Gruppen betrachtet. Besteht Gruppe A aus einer Schnecke und einem Rennwagen und Gruppe B aus menschlichen Zwillingen, so wird deutlich, dass Gruppe A beim englischen Verfahren haushoch überlegen ist: Je Zeiteinheit legt das Rennauto ein Vielfaches der Strecke der Zwillinge zurück, und die Entfernung, welche die Schnecke zurücklegt, ist völlig unerheblich. Beim französischen Verfahren ist hingegen Gruppe A chancenlos: Egal wie wenig Zeit das Auto für die Teststrecke benötigt, die Schnecke wird am allerlängsten unterwegs sein und damit den Ausschlag gegen Gruppe A geben.

⁵⁶ In der angewandten Mathematik spricht man in solchen Fällen von einem sehr *konservativen* Ansatz.

⁵⁷ Auch wenn er zumindest im Moment nur unter diesem einen Blickwinkel zu erkennen ist. So wird jeder Arzt vorsorglich zu weiteren Untersuchungen raten, wenn sich eine potenziell krankhafte Veränderung nur mit einer Methode nachweisen lässt.

2.8 Reale Messungen

Measurement: The real thing. (Duncan 1984: Kapitel 5)

Wie „stark“ wird das Gemessene vom Messvorgang bestimmt? Das Kontinuum der möglichen Antworten hat zwei Pole. Einer davon ist der Operationalismus, der alle Verfahren als fundamental verschieden, also nicht vergleichbar oder gar äquivalent auffasst:

Objektbereich 1 — Prozedur 1 — Daten 1
Objektbereich 2 — Prozedur 2 — Daten 2
Objektbereich 3 — Prozedur 3 — Daten 3

Konsequenterweise kann er keine Realität über das spezielle Verfahren hinaus erkennen - die Eigenschaften eines Phänomens sind nämlich untrennbar mit dem einzelnen Verfahren verwoben.⁵⁸ Weil die Realität nicht vom Verfahren getrennt werden kann, die tatsächlichen Eigenschaften mit anderen Worten also nicht isolier- oder identifizierbar sind, ist eine Variante hiervon die Objektbereiche gleich wegzulassen.⁵⁹ Kurz gesagt ist ein so verstandener Operationalismus gleichbedeutend mit einer *vollständigen* Skalenabhängigkeit der Ergebnisse.

Der entgegengesetzte Pol ist also, alle Messverfahren in „einen Topf“ zu werfen, d.h. als äquivalent zu betrachten:

Objektbereich	{	Prozedur 1 — Daten 1
		Prozedur 2 — Daten 2
		Prozedur 3 — Daten 3

Eine grobe, mathematisch orientierte Messtheorie, die nur wenige Skalenarten kennt, kommt diesem Pol ziemlich nahe. Weil man bei der Messung einer Größe eine (sehr große) Äquivalenzklasse von Verfahren betrachtet, wird das spezielle Verfahren nebensächlich, es ist lediglich ein Mittel zum Zweck, die Realität zu erkennen. Hierzu passt eine realistische Interpretation der Invarianten, also derjenigen Eigenschaften der Daten, welche unter allen Messverfahren gültig sind bzw. nicht von einer speziellen Skala abhängen. Skalenabhängige Aussagen sind empirisch bedeutungslos, während man Invarianten auf die Realität attribuieren sollte. Schematisch:

Objektbereich	{	Prozedur 1	}	Invarianten
		Prozedur 2		
		Prozedur 3		

⁵⁸ Bridgman (1927: 5): “[...] the concept is synonymous with the corresponding set of operations.”

⁵⁹ Hand (1996: 453): “Thus, an attribute is defined by its measuring procedure, no more and no less, and has no ‘real’ existence beyond that.”

So verschieden beide Vorstellungen auch erscheinen mögen, tatsächlich sind beide Spezialfälle des allgemeinen Modells (1.3). Das zeigt sich auch historisch. Die tiefschürfende Kritik an der traditionellen Messtheorie des 19. Jahrhunderts⁶⁰ führte nach Russell (1903: 158) zu einer strikten Trennung formaler und empirischer Sachverhalte: “The separation of number and quantity is thus complete: each is wholly independent of the other.”⁶¹ Aufgabe des 20. Jahrhunderts war es deshalb, den Zusammenhang von *number* und *quantity* zu modellieren, was in Form der Abbildung empirischer auf formale Strukturen denn auch geschah.

In der Praxis wird der häufigste Fall wohl jener sein, dass gewisse Messverfahren äquivalent sind, andere jedoch nicht. Das heißt, man hat schematisch:

Objektbereich 1	{	Prozedur 1 — Daten 1
		Prozedur 2 — Daten 2
Objektbereich 2	—	Prozedur 3 — Daten 3

Eine solche Situation lässt sich auf drei verschiedene Arten interpretieren:

- Betrachtet man die *Zeilen* bzw. die *letzte Spalte*, so erkennt man verschiedene Perspektiven, die alle zu unterschiedlichen Daten (1, 2, 3) führen, welche aber partiell vergleichbar sind (nämlich 1 und 2). Je nach Blickwinkel des Beobachters stellt sich die Situation verschieden dar.
- Betrachtet man die Verfahren, also die *mittlere Spalte*, so kann man von einem *partiellen* Operationalismus sprechen, denn das Ergebnis hängt von der Prozedur (1 & 2 versus 3) ab. Andererseits lässt sich aber auch von einem *partiellen* Realismus sprechen, denn aufgrund der Äquivalenz der Skalen 1 und 2 gibt es Invarianten, die als Eigenschaften der Realität angesehen werden können.
- Blickt man zuletzt auf die Objekte, so kann man entweder argumentieren, dass es sich um verschiedene Bereiche handelt (1 versus 2) oder aber um mehrere Facetten ein und desselben Gegenstands.⁶²

Diese differenziertere Betrachtungsweise hat viele Vorteile. Zum Ersten wird sie der Realität gerechter als die obigen extremen Modelle.⁶³ Zum Zweiten werden Objekte, Prozeduren und Daten auf eine Stufe gestellt. Jede der genannten Sichtweisen ist gerechtfertigt, und es kommt auf die konkrete Situation an, welche der möglichen Interpretationen am angemessensten ist. Zum Beispiel spricht man bei der Intelligenzmessung von *verschiedenen Facetten* der untersuchten Eigenschaft,⁶⁴ ähnlich wie beim Gleichnis von den

⁶⁰ Siehe z. B. Russell (1897) und Michell (2003a)

⁶¹ Michell (1997: 111) ergänzt: “and measurement must be understood anew.”

⁶² Etwa sprechen Cronbach et al. (1972) von einem “multifacet concept of observational procedures” in der Psychologie.

⁶³ Auch wenn sie zuweilen immer noch zu grob sein könnte. Wenn sich Prozeduren z. B. teilweise überlappen, so kann man sie nicht mehr in strikt getrennte Äquivalenzklassen einteilen.

⁶⁴ Also sprachlicher, mathematischer, räumlicher Intelligenz usw.

sechs blinden Weisen, die ein und denselben Elefanten untersuchen. Im Beispiel von Hand wird man wohl eher auf die *Unterschiedlichkeit der Verfahren* abheben, ähnlich einem sportlichen Mehrkampf, der ganz verschiedene Fähigkeiten, wie Kraft, Geschicklichkeit, Ausdauer, Geschwindigkeit usw. testet. Schließlich kennt jeder Fotograf diverse Filter, entsprechend *verschiedenen Perspektiven*, welche zu unterschiedlichen Ergebnissen - jedes untrennbar mit dem speziellen Verfahren verwoben, und doch höchst relevant für die Praxis - führen. Im Infrarot erkennen Bienen Muster, die uns verborgen bleiben, eine Falschfarbenaufnahme kann Kontraste betonen und Schwarzweißbilder heben Konturen deutlich hervor.

Zum Dritten kann sich Fortschritt daran zeigen, dass skalenabhängige Merkmale nach und nach durch skalenunabhängige ersetzt werden. Dem entspricht, dass vergleichbare Messverfahren entwickelt werden (aus einer Prozedur werden viele) und aber enge operationale Definitionen durch Verallgemeinerungen ersetzt werden können, die von konkreten Messverfahren immer unabhängiger werden. Man startet pragmatisch-operationalistisch und kommt schließlich zu Theorien, die völlig zurecht den Anspruch auf eine realistische Interpretation erheben.⁶⁵

Fortschritt kann sich jedoch auch so darstellen, dass sich neue Perspektiven eröffnen, also mit qualitativ anderen Verfahren ganz neue Eigenschaften messen lassen. Blickt man zurück, so wurden gerade mit neuen Methoden überraschende Entdeckungen gemacht, die sich völlig außerhalb unseres bisherigen Erfahrungsschatzes befanden.

Schließlich animiert die Darstellung zu einer dezidiert *empirisch-induktiven* Sicht: Es ist durch Beobachtung und Experiment zu klären, welche Prozeduren tatsächlich (noch) als äquivalent angesehen werden sollten (siehe hierzu auch S. 75); es entscheidet sich in der Empirie, welche Aussagen „wirklich“ bedeutungsvoll sind; ebenso sollten für jede Gleichung die zu ihr gehörigen Symmetrien (siehe S. 20) ermittelt werden. Forschung ist ein offener und flexibler Prozess, der sich an der Realität orientiert. Messverfahren eröffnen Objektbereiche, welche nach und nach differenziert werden. Feste „A-Priori-Einstellungen“ (um nicht das abgenutzte Wort „Vorurteile“ zu verwenden) stören hierbei mehr als dass sie nützen. Und a priori zum Scheitern verurteilt ist der Versuch, aus prinzipiellen Überlegungen heraus der Realität vorschreiben zu wollen, wie sie sich zu verhalten habe.

Man erkennt nun auch, wann die Messtheorie erfolgreich ist: Bei der klassischen Temperaturmessung nutzt man im Prinzip immer denselben Effekt, nämlich die Wärmeausdehnung von Gegenständen (typischerweise einer Flüssigkeit, zuweilen aber auch eines Festkörpers). Die scheinbar unterschiedlichen Arten der Messung von Kälte bzw. Wärme bzw. Hitze (denn nur so ungenau ließ sich die Temperatur vor Einführung von Thermometern fassen) sind nicht wirklich verschieden, sondern durchaus vergleichbar. Da man zudem Messfehler hier vernachlässigen kann, kommt man schnell zu einer großen Anzahl

⁶⁵ Dazu mehr S. 164ff.

gleichwertiger Messmethoden bzw. Skalen, bei denen Nullpunkt und Einheit frei wählbar sind, also von Forscher zu Forscher variieren - und genau so war es auch historisch gesehen.

Das zur Messtheorie gehörige Schema zeigt, dass man in gewissem Sinne von einem Idealzustand ausgeht, der in der Praxis nicht allzu häufig der Fall ist. Die Physiker hatten sozusagen „Glück“, weil die Messung physikalischer Attribute - wie z. B. der Temperatur - problemlos möglich war. Bei vielen anderen Gegenstandsbereichen ist dem nicht so, und die optimistischen, weil sehr weitreichenden Annahmen der Messtheorie sind nicht erfüllt: Deshalb scheitert sie, wenn man es mit nicht-äquivalenten, also mit nicht gleichartigen Perspektiven zu tun hat oder aber viele Facetten eines Objekts eine Rolle spielen. Auch mit Messfehlern weiß sie nicht umzugehen, ebenso wenig mit Approximationen, die nicht exakt ihren engen Strukturen entsprechen. In allen diesen Fällen bleibt ihr nur der Rückzug in die Mathematik, d.h., ihre Schlüsse sind nicht direkt empirisch verwendbar.

Da dies oft der Fall ist, sollte man den Term „bedeutungsvoll“ besser durch den neutraleren Ausdruck „skalenabhängig“ bzw. „nicht von der speziellen Skala trennbar“ ersetzen. Skalenabhängigkeit kann je nach Situation bedeuten, dass eine Statistik oder Beziehung tatsächlich nicht bedeutungsvoll ist oder aber, dass sie nur relativ zur gewählten Perspektive bzw. bedingt unter dem verwendeten Verfahren sinnvoll ist. Perrins Resultat ist deshalb so überzeugend, weil er über ein Dutzend völlig verschiedenartiger Perspektiven wählte und trotzdem immer dasselbe Ergebnis erhielt. Und genau deshalb ist er auch der Kronzeuge jedes Realisten. Wir führen die Diskussion in Abschnitt 3.9 weiter.

2.9 Messtheorie und Statistik I

Statisticians have generally rejected the proscription of methods based on the limitations of permissible transformations.

Velleman und Wilkinson (1993: 67)

Fast so alt wie die Stevenssche Messtheorie ist die Frage, welche Auswirkungen ihre Resultate und Empfehlungen auf die Analyse von Daten haben. Inwieweit sollte die Messtheorie uns bei der Wahl von statistischen Verfahren leiten, könnte man auch fragen. Die Diskussion wird dadurch erschwert, dass die Messtheorie dezidiert mathematisch betrieben wurde und Mathematiker der empirischen Interpretation ihrer Resultate wenig Bedeutung zumessen. Falls Theoreme normative Implikationen haben, so sollten empirisch arbeitende Wissenschaftler jene ohne Widerworte beherzigen; ansonsten sind jene in Gefahr, „bedeutungslose“ Ergebnisse zu produzieren.

Das heißt, auf der einen Seite lässt sich mit Luce (S. 35) und Stevens (1946, 1951) argumentieren, dass die Messtheorie die Wahl des statistischen

Verfahren wenn nicht vorschreibt, so doch zumindest eine ganze Reihe von Möglichkeiten als „nicht zulässig“ ausschließt. So schreiben Backhaus et al. (2003: XVIII f.) in ihrer Einführung in multivariate Analysemethoden: „Das Skalenniveau bedingt sowohl den *Informationsgehalt der Daten* wie auch die *Anwendbarkeit von Rechenoperationen*. [Bei Nominalskalen] handelt es sich lediglich um eine Kodierung der Merkmalsausprägungen [...] Mit derartigen Zahlen sind daher keine arithmetischen Operationen (wie Addition, Subtraktion, Multiplikation oder Division) erlaubt [...] Die Rangwerte 1., 2., 3. etc. sagen nichts über die Abstände zwischen den Objekten aus [...] Daher dürfen auch ordinale Daten [...] nicht arithmetischen Operationen unterzogen werden.“ (Hervorhebungen im Original.)

Im Großen und Ganzen gefiel es Lehrbuchautoren sehr gut, mit der Skaleneinteilung der Messtheorie ein Ordnungsschema für statistische Verfahren zur Hand zu haben. Zuweilen ist dies sogar heute noch, insbesondere in den Sozialwissenschaften, der Fall, siehe z. B. Krauth (1988), Steyer und Eid (1993) und Bortz (2004). Insbesondere setzt man die Messtheorie ein, um die grundlegenden Unterscheidung von parametrischen und nichtparametrischen statistischer Verfahren zu motivieren. Damit ist folgendes gemeint: Statistische Analysemethoden nutzen oftmals nur einen gewissen Teil der in den Daten potenziell enthaltenen Informationen. Klassische parametrische Verfahren berücksichtigen die Distanz zwischen einzelnen Daten. Berechnet man z. B. die Varianz einer Stichprobe $x_1, \dots, x_n$, also

$$\sigma^2(X) = \sum_{i=1}^n (x_i - \bar{x})^2 / n$$

wobei die x_i beliebige reelle Zahlen sind und $\bar{x}$ ihr arithmetisches Mittel, so erkennt man sofort, dass die Abstände $x_i - \bar{x}$ eine wichtige Rolle spielen. Nichtparametrische Verfahren hingegen nutzen lediglich die Ranginformation, arbeiten also mit dem geordneten Datensatz $x_{(1)} \leq \dots \leq x_{(n)}$, ohne dass die Abstände zwischen den Zahlen in die Auswertung eingehen. Wie man unschwer erkennt, passt diese Differenzierung hervorragend zur Unterscheidung von Intervall- und Ordinalskalen in der Messtheorie.

Die im Großen und Ganzen kooperative Haltung der theoretischen Statistik ist nicht weiter verwunderlich, geht man in der mathematischen Statistik doch im wesentlichen von Zahlen - nicht: Daten - aus, die gewissen formalen Voraussetzungen, gerne „Modellannahmen“ genannt, genügen. Messtheoretische Ideen und statistische Verfahren ergänzen sich hier sehr gut. Schlimmstenfalls arbeitet man mit nicht kompatiblen Annahmen und jeder betrachtet innerhalb des von ihm gesetzten Rahmens, was sich daraus folgern lässt.

Nicht so in der angewandten Statistik. Aus den vielen bereits genannten Gründen waren angewandte Statistiker, die mit realen Daten arbeiten, erheblich kritischer, und die normativen Vorgaben der Messtheorie stießen nahezu unmittelbar auf entschiedenen Widerstand. Die am weitesten reichende Haltung ist, dass die Zahlen nicht wissen, wo sie herkommen (Lord 1953),

was nichts anderes heißt, als messtheoretische Überlegungen bei statistischen Auswertungen zu ignorieren. Im Kern will man sich nicht von anderer Stelle vorschreiben lassen, wie mit Daten umzugehen ist; dies umso mehr als restriktive Regeln ganz allgemein nur dann respektiert werden, wenn man bei deren Verletzung unsinnige Ergebnisse erzielt.

Genau das ist, wie wir oben gezeigt haben, zumeist nicht der Fall, weshalb in der Konsequenz die Vorgaben der Messtheorie in der Praxis - von Statistikern wie empirischen Forschern - in aller Regel beiseitegeschoben werden. Um zumindest auf dem Papier zu zeigen, dass man die Messtheorie kennt und ihr gerecht werden will, findet sich in Lehrbüchern bisweilen ein einschlägiger Vorspann, welcher sinngemäß den typischen und vielsagenden Satz enthält: „Oftmals werden - auch im vorliegenden Buch - Skalen benutzt, von denen man lediglich annimmt, sie seien intervallskaliert“ (Backhaus et al. 2003: XIX). Das heißt, weniger diplomatisch ausgedrückt, dass man bei der Datenanalyse genau die Annahmen voraussetzt, welche man gerade braucht (ohne dieselben empirisch zu prüfen), was genauso bequem ist wie ein sich selbst genügender Operationalismus im Labor.

Man sollte sich daran erinnern, dass auch ein entschiedener Gegner der Messtheorie wie Tukey weit davon entfernt ist, die Empfehlungen der Messtheorie einfach zu ignorieren. Tukey (1986b: 245) sagt: “Not every quantity measured on an ordinal scale is such as to deserve the calculation of a sample mean [...]” und auf der nächsten Seite heißt es: “If a finding that ‘variance increases as we move up the scale’ is only useful if it can be taken as a contribution to the fundamentals of psychology, then we must be very careful about making such statements.”⁶⁶

Deshalb begründen Mallows und Tukey (1982: 902f) explizit, warum sie trotzdem die Messtheorie im Regelfall nicht beachten: “To say we require fundamental measurement on an interval scale before using an arithmetic mean may perhaps be correct for that one-tenth of one percent of the time when we are concerned with fundamental measurement. Yet, who would deny the arithmetic mean to the climatologist - purely as a matter of principle - because most temperatures are measured on a practical scale or because interval-scale temperatures have been available only for about a century? Practical measurement is by far the most used and the most analyzed.” Tukey (1969: 729f) liest sich wie eine Folgerung hieraus:

Campbellian measurement has scared us far too long [...] if physics had had to grow by Campbellian measurement, it would never have reached the state in which Campbell found it. While giving Campbell the respect which his insightful wisdom deserves, it is time to strive off the chains with which response to his writings have bound us, whether through acceptance, through reaction, or through optimism.

⁶⁶ Er fährt fort: “But if it serves to guide us, perhaps in the design of an experiment, perhaps in the choice of a method of statistical analysis, perhaps in the directions in which we seek new or modified theories whose confirmation we realize must rest on approximate results, such a statement, although resting on a wholly approximate foundation, may be very useful.”

Measuring the right things on a communicable scale lets us stockpile information about amounts. Such information can be useful, whether or not the chosen scale is an interval scale. Before the second law of thermodynamics - and there were many decades of progress in physics and chemistry before it appeared - the scale of temperature was not, in any nontrivial sense, an interval scale. Yet these decades of progress would have been impossible had physicists and chemists refused either to record temperatures or to calculate with them. I hope it is clear that if I had a 1818 physicist as a client I would not let the then scale type of temperature stop me from recommending the calculation [...] of means of temperatures. With a 1968 psychologist I would take a similar view about many numbers not on interval scales.

Diese Meinung vertrat zeitgleich auch Stevens (1968: 856), wenn er in seinen späten Jahren einräumte “The widespread use on ordinal scales of statistics appropriate only to interval or ratio scales can be said to violate a technical canon, but in many instances the outcome has demonstrable utility.”

Kompromisse

Eine solche Lösung ist auch für Messtheoretiker möglich, welche wie Steyer, Eid oder Narens (siehe S. 48ff) die „bescheidene“ Auffassung vertreten, dass es sich beim Begriff der Bedeutsamkeit lediglich um einen ‚terminus technicus‘ handle. Bei dieser mathematisch-formalen Lesart fällt es leicht, nicht-bedeutungsvolle Aussagen und Statistiken zu akzeptieren, handelt es sich doch bei jenen lediglich um Aussagen, welche von einer speziellen Skala wesentlich abhängen.⁶⁷ Derartige Aussagen sind zwar nicht so gehaltvoll wie „unbedingte“, also Aussagen, die von der Skala unabhängig sind, aufgrund der zurückhaltenden Interpretation des Begriffs „bedeutungslos“ wird daraus jedoch *nicht* gefolgert, dass skalenabhängige Begriffe oder Aussagen keine empirische Bedeutung hätten.

Außerdem legt die genannte Haltung den weiteren Schritt nahe, Fortschritt daran festzumachen, dass sukzessive skalenabhängige Ergebnisse bzw. Aussagen durch skalenunabhängige ersetzt werden. Narens (2002: 408) schreibt: “As the fragment matures, the tendency of progression usually moves from non-meaningful to meaningful [...] formulations.” Eine etwas allgemeinere Lösung hatten wir auf anderem Wege in Abschnitt 2.8 hergeleitet.

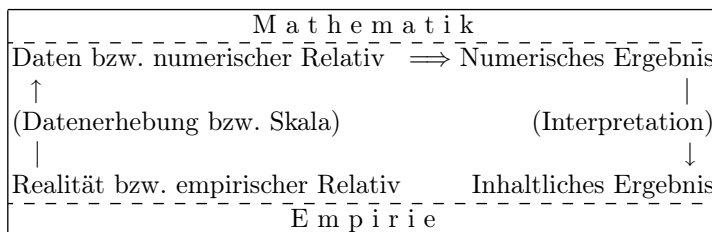
Auch bei einer strengen, normativen Lesart der Messtheorie sind tragfähige Kompromisse durchaus möglich. Das Hauptargument ist seit Burke (1953), dass es bei der Messtheorie um die Erhebung und anschließende Interpretation von Daten geht, während es die Statistik vornehmlich mit Verfahrensweisen zu tun hat, die auf mathematisch-formalen Annahmen basieren. Luce et al. (1990: 294) zitieren Michell (1986: 402): “The question never was one of permissible statistics or meaningfulness. It was always one of legitimate inference.” Dies kommentieren sie mit: “We give our view, which agrees with the

⁶⁷ So lässt sich bereits Pfanzagl (1968: 34) im Fall der Temperaturmessung mit °C und °F verstehen: “Thus contrary to the relation $x < y$, the relation $y = 2x$ [...] is meaningless without specification of the scale.”

latter but not the former of the two quoted sentences.” Aus dem folgenden geht dann hervor, dass Luce et al. (1990) nach wie vor eine strenge Version der Messtheorie vertreten, jedoch einen Unterschied zwischen den Annahmen statistischer Verfahren und der Messtheorie machen: “It seems resonable to require that the numerical relations defined by such funcitons be meaningful relative to the underlying empirical structure. But on the other hand, the logic of statistical inference seems to require only distributional assumptions, not assumptions about the way in which the measurements x_i are used to represent empirical relations [...]” (ibd., S. 294). Bortz und Döring (1995: 168) erklären den Unterschied präzise und deutlich:

In diesem Zusammenhang sei auf einen Irrtum aufmerksam gemacht, der seit der Einführung der vier Skalenarten [...] durch Stevens anscheinend nur schwer auszuräumen ist. Die Behauptung, parametrische Verfahren wie z. B. der t-Test oder die Varianzanalyse setzten intervallskalierte Daten voraus, ist in dieser Form nicht richtig. Die mathematischen Voraussetzungen dieser Verfahren sagen nichts über die Skaleneigenschaften der zu verrechnenden Daten aus. (Die Varianzanalyse setzt z. B. normalverteilte, unabhängige und homogene Fehlerkomponenten voraus.) Vor diesem Hintergrund wäre beispielsweise gegen die Anwendung varianzanalytischer Verfahren auf Daten wie z. B. Telefonnummern nichts einzuwenden, solange diese Zahlen die geforderten mathematischen Voraussetzungen erfüllen [...]. Gaito (1980) diskutiert die Hartnäckigkeit dieses Mißverständnisses anhand zahlreicher Literaturbeispiele⁶⁸ und fordert nachdrücklich, bei der Begründung der Angemessenheit eines statistischen Verfahrens zwischen *meßtheoretischen Interpretationsproblemen* und *mathematisch-statistischen Voraussetzungen* zu unterscheiden. Die Frage, ob verschiedene Zahlen tatsächlich unterschiedliche Ausprägungen des untersuchten Merkmales abbilden bzw. die Frage, ob - wie es die Intervallskala fordert - gleiche Zahlendifferenzen auch gleiche Merkmalsunterschiede repräsentieren, ist ein meßtheoretisches und kein statistisches Problem. Der statistische Test ‚wehrt‘ sich nicht gegen Zahlen minderer Skalenqualität, solange diese seine Voraussetzungen erfüllen. Die Skalenqualität der Zahlen wird erst bedeutsam, wenn man die Ergebnisse interpretieren will. Es sind dann meßtheoretische Erwägungen, die dazu veranlassen, die Ergebnisse einer Varianzanalyse über Nominalzahlen für nichtssagend zu erklären, weil die Mittelwerte derartiger Zahlen keine inhaltliche Bedeutung haben.“ (Hervorhebungen im Original.)

Wir fassen die Argumente im folgenden Schaubild zusammen, das strikt zwischen der mathematischen und der empirischen Ebene unterscheidet:



⁶⁸ Man vergleiche hierzu unsere Ausführungen am Anfang des Abschnitts, die womöglich verständlich machen, wie sich dieser Irrtum entwickeln konnte.

Das heißt, bei der Datenanalyse hat man es mit drei nur teilweise von einander abhängigen Bereichen zu tun. Zum ersten sind die Daten zu erheben, technisch gesprochen wird gemessen ($\uparrow$). Genau hierfür ist die Messtheorie zuständig. Beim Ergebnis der Erhebung handelt es sich in aller Regel um Zahlen.

Eben jene Zahlen werden im zweiten Schritt mithilfe statistischer Verfahren ausgewertet ($\Rightarrow$). Diese Verfahren basieren auf mathematisch-formalen Annahmen. Nur wenn jene im Wesentlichen erfüllt sind, sollte man die Verfahren benutzen. Der Grund hierfür ist, dass es sich bei den Verfahren letztlich um mathematische Umformungen handelt, die nur unter den gegebenen Voraussetzungen zulässig sind. Ansonsten kann man Gegenbeispiele angeben, die zeigen, dass das Resultat unsinnig oder beliebig ist. Man beachte, dass die Mathematik an dieser Stelle lediglich eine Hilfswissenschaft ist, die aus den Daten Informationen extrahiert bzw. in der Lage ist, viele Daten in wenigen aussagekräftigen Zahlen, Diagrammen und Bemerkungen zusammenzufassen.

Zuletzt interpretiert man die numerischen Ergebnisse, das heißt, man bezieht die Zahlen wieder auf die reale Welt ($\downarrow$). Und genau hier kommt erneut die Messtheorie oder allgemeiner gesprochen die Art und Weise der Datenerhebung ins Spiel. Gewisse empirische Schlussfolgerungen sind nämlich nur dann gerechtfertigt, wenn die Datenerhebung bereits gewissen Anforderungen genügt. Hat man nur die Medaillenränge notiert, so wird man wohl nicht fundiert behaupten dürfen, der Sieger eines sportlichen Wettbewerbs haben nur ganz knapp gewonnen. Natürlich lässt sich alles Mögliche behaupten, es kommt jedoch entscheidend auf die Art der Datensammlung (und Auswertung) an, ob die Behauptungen auch gerechtfertigt oder glaubhaft sind. Informatiker haben den drastischen Satz “Garbage in, garbage out” geprägt, und genau so ist es auch schlimmstenfalls bei der Analyse und Interpretation von Daten.

Alle drei Aspekte des statistischen Erkenntnisprozesses werden in der Definition von Efron und Tibshirani (siehe S. 4) angesprochen. Nur wenn man sich auf den rein mathematischen „Ast“ ganz oben beschränkt, lässt sich Lords Behauptung, dass die Daten nicht wissen, wo sie herkommen, vertreten. Nicht aber, wenn es darum geht, Daten inhaltlich zu interpretieren. Dann ist es sehr wohl relevant, wie die Daten zustande gekommen sind. Es ist gerade die orthodoxe Statistik in der Nachfolge R. A. Fishers, die diesen Gesichtspunkt betont und ihm allgemeine Anerkennung verschafft hat. (Nur eine Minderheit von Statistikern sind aus Gründen, die wir noch besprechen werden, etwas anderer Ansicht. Siehe insbesondere den Abschnitt 4.4.2.)

Die Interpretation ist gewissermaßen die Umkehrung der Erhebung der Daten (wobei man zwischendurch die Dienste der Mathematik in Anspruch genommen hat). Sie ist deshalb besonders schwierig, weil hier alle potenziell relevanten Aspekte zu berücksichtigen sind:⁶⁹ Erstens die Erhebung, zweitens die formale Analyse und drittens das ganze Umfeld. Genau deshalb ist

⁶⁹ “How accurate are my data summaries” (siehe S. 4) ist also nur ein kleiner Aspekt der Interpretation.

sie nie rein formal oder gar automatisiert zu bewältigen. Sie erfordert Fingerspitzengefühl bei der Gewichtung vieler Faktoren, Erfahrung bei der Bewertung von Zahlen und Fakten und insbesondere ein großes Verständnis des empirischen Bereichs, den man gerade untersucht. Ein numerisch höchst bedeutsam erscheinendes Resultat kann empirisch völlig unwichtig oder sogar irreführend sein, und andererseits kann ein statistisch unauffälliges Ergebnis inhaltlich außerordentlich bedeutsam sein. Eine originelle, weit über die formalen Aspekte hinausgehende Diskussion dieses Problems findet sich in Good (1983b), insbesondere S. 290.

Man erkennt, dass Messtheorie und mathematische Statistik (im Sinne mathematischer Analyseverfahren) im Wesentlichen zwei verschiedene Aspekte der Datenanalyse bearbeiten. Mathematische Voraussetzungen, für die formale Analyse der Daten höchst bedeutsam, gehören zur oberen, nicht-empirischen Ebene. Anders die Messtheorie: sie will gerade den Übergang von Empirie zum formalem Modell, von Fakten zu Zahlen, beschreiben. Schließlich fließen beide in die Interpretation der Resultate ein.

Eine integrative Position sollte damit die Argumente der Messtheorie und der Mathematik ernst nehmen. Es geht um "Scales and Statistics", und in seinem gleichnamigen Artikel hat P. Gardner (1975) eine sinnvolle Kompromisslösung formuliert. Zunächst unterscheidet er streng (wie wir gerade eben und viele andere Autoren auch) zwischen inhaltlich-empirischen und formal-mathematischen Aspekten. Um Platz für einen Kompromiss zu schaffen, argumentiert er sodann gegen die rigide Skaleneinteilung nach Stevens: "In short, the distinction between types of scales (and, by implication, appropriate statistics) is by no means clear-cut" (ibid., S. 47.) Insbesondere führt er eine "grey area" zwischen Ordinal- und Intervallskala ein, die er "ordinal/interval continuum" nennt und sehr praxisrelevant ist: "[This continuum is] occupied by a large number of instruments which can be conveniently labelled as *summated scales*." Dazu gehören z. B. "summated-rating attitude scales, in which the total score is the sum of a set of ordinal weightings [...]", also etwa der weiter oben besprochene APGAR-Test. Diese Instrumente sind von größter praktischer Bedeutung, denn "The summated scale category obviously includes a large proportion of all the instruments used in educational and psychological research." Soweit die Skalen bzw. die „Anhebung“ des Skalenniveaus über die Ordinalskala hinaus.

Auch von statistischer Seite aus kann man entgegenkommen. Wie die Physik und andere empirische Wissenschaften hat auch die Statistik erkannt, dass viele mathematische Voraussetzungen in der Praxis oft nicht exakt erfüllt oder sogar mehr oder minder stark verletzt werden. Anders als Mathematiker, die in solchen Fällen keine Aussagen machen bzw. nur eine Vermutung in den Raum stellen, haben Statistiker nun empirisch-systematisch getestet, wie sich ihre Verfahren verhalten, wenn die Annahmen eben nicht genau gelten oder sogar grob verletzt werden. Dabei hat sich im großen und ganzen erfreulicherweise herausgestellt, dass die Ergebnisse vieler Verfahren im We-

sentlichen gültig bleiben, falls man ihre Voraussetzungen nicht allzu sehr bzw. nicht allzu viele Voraussetzungen zugleich verletzt.

Hierfür hat sich der Terminus „robust“ eingebürgert, d.h. eine robuste Statistik bzw. ein robustes Verfahren ist ein(e) solche(s), die bzw. das bei Verletzungen der mathematischen Annahmen nicht in die Irre führt, sondern weiterhin sinnvolle Ergebnisse liefert.⁷⁰ Es gehört mittlerweile zur Routine, Verfahren auf ihre „Robustheit“ zu überprüfen und es ist immer ein Gütesiegel, wenn ein Verfahren nicht allzu empfindlich auf die Verletzung seiner Annahmen reagiert. (Siehe hierzu aber auch S. 336ff) Insbesondere sind auch viele parametrische Verfahren, die also Abstände zwischen Datenpunkten auswerten, ziemlich robust.

Insgesamt besteht der von P. Gardner (1975) ausgearbeitete Kompromiss also darin, das realistische Skalenniveau höher einzuschätzen als die traditionelle Messtheorie und zugleich robuste statistische Verfahren für derartige Daten zu verwenden. Dadurch geht man sowohl theoretisch fundiert als auch praxisnah vor.

Invarianz

Oftmals sind die einfachsten Strategien die besten. Schon auf den Seiten 61ff und 67 hatten wir diesbezüglich viele Vorzüge einer empirisch-induktiven Strategie aufgezählt. Das heißt, anstatt höchst theoretische und zuweilen wohl auch dogmatische Argumente auszutauschen um schließlich zu komplexen „diplomatischen“ Übereinkünften zu kommen, spricht viel dafür, die Theorie-Beladenheit der Diskussion hinter sich zu lassen und direkt-empirisch die Frage anzugehen: Welche Skalen und Auswertungen sind sinnvoll? Bei diesem Ansatz treten an die Stelle theoretischer Invarianzargumente empirische:

[This] argument is founded upon empirical studies of the effect of transformations of data upon [...] various statistics. The basis of this argument is that, if one alters the metric properties of scales and still reaches the same conclusions regardless of whether the data have been transformed or not, then, for the purposes of statistical inference, it cannot matter very much what the scale properties are. (P. Gardner 1975: 50)

Wie die Physiker des 19. Jahrhunderts misst man also zunächst, was noch sinnvoll messbar ist, und fragt sich danach, welche Klasse von Transformationen (immer noch) zu derselben statistischen Schlussfolgerung führen. Das heißt nicht, die Messtheorie zu ignorieren, sondern ihren Kerngedanken - Invarianz - empirisch umzusetzen. Dazu mehr in den Abschnitten 3.9 und 3.10.3.

⁷⁰ Siehe z. B. Gelman et al. (2004), Maronna et al. (2006) und die beiden Klassiker Huber (1981), Hampel et al. (1986).

Mathematik als gemeinsame Basis

Wir möchten dieses von Auseinandersetzung geprägte Kapitel mit einem Beispiel beenden, das die integrative Kraft der Mathematik beleuchtet. Es wäre nicht allzu überraschend, hätten Luce und Tukey, zwei exponierte Vertreter diametral verschiedener philosophischer Haltungen, mit vielen Worten gegeneinander gestritten. Tatsächlich beschränkt sich nach Wissen des Autors die Auseinandersetzung zwischen den beiden jedoch auf die zuvor erwähnten Zitate (siehe S. 35 und S. 41) aus den 1950er Jahren, und auch diese sind sachlicher, nicht persönlicher Natur.

Wie fruchtbar völlig verschiedenartige Denkansätze zusammenwirken können, bewiesen die beiden genannten Wissenschaftler mit ihrem gemeinsamen Artikel zu *Conjoint Measurement*. Dieser Artikel wird von *allen* Fachwissenschaftlern als ein großer Durchbruch der Messtheorie bezeichnet und er hat auch praktisch eine immense Bedeutung erlangt. Im Kern zeigt er, dass man mithilfe weniger Annahmen, allein durch systematische Vergleiche, zu Messungen auf Intervallskalenniveau gelangt: "The role usually played by the concatenation operation⁷¹ is replaced by the fact that the objects are ordered pairs. From the axioms we give, simultaneous measurement on interval scales is obtained [...]"⁷²

Es ist die Mathematik, welche den allgemeinverbindlichen Rahmen zur Verfügung stellt, innerhalb dessen selbst völlig verschiedene Ansichten fruchtbar zusammenwirken können. Dies ist keine geringe Leistung, verpuffen doch Energien nicht in fruchtlosen rhetorischen „Glaubenskämpfen“ - zu archivieren in umfangreichen Bibliotheken. Stattdessen wirkt der philosophische Dissens fachwissenschaftlich stimulierend und führt zu ebenso fundamentalen wie fundamental neuen Einsichten:

Not the least of the attractions of mathematics is that it allows us to temporarily ignore the question of interpretation [...] Mathematics [...] offers [...] a common language, so that we can dispute about ideas but not about grammar. (Moore 1990)

⁷¹ Diese Operation des Zusammenfügens war für Campbell grundlegend für eine Messung auf Intervallskalenniveau. Sie führt dazu, dass man Messwerte addieren „darf“. Das Paradebeispiel sind hintereinander gelegte Stäbe bei der Längenmessung.

⁷² Siehe Luce und Tukey (1964: 487), für eine prägnante aktuelle Darstellung siehe auch Luce (2000: 293f).

Statistik im Forschungsprozess

Eine Philosophie der Statistik als Baustein einer
integrativen Wissenschaftstheorie

Saint-Mont, U.

2011, XIV, 675 S. 4 Abb., Hardcover

ISBN: 978-3-7908-2722-4

A product of Physica-Verlag Heidelberg