Cornelis Menke

Zum Ideal der Quantifizierung

Abstract: Quantification in sciences and humanities may be considered to serve two different forms of objectivity: first, it may serve the object of a truthful representation of reality (absolute objectivity); second, it may serve to ensure trust among researchers (mechanical objectivity). Taking the ›crisis of confidence‹ within contemporary psychology as an example, I consider the question which conditions must be fulfilled for quantification to serve these ideals. I shall argue that big data, while permitting the pursuit of new research questions and programmes, at the same time may undermine the value of quantitative measures by allowing for new research practices like a mechanical search for patterns in data, thus compromising both ideals of objectivity.

1 Das Ideal der Quantifizierung

Auf dem Balkon des Social Science Research Building der University of Chicago, 1126 59th Street, findet sich eine Variante eines berühmtes Ausspruchs des britischen Physikers Sir William Thomson, des späteren Lord Kelvin, in den Stein gehauen:

> »When you cannot measure * your knowledge is * meager * and * unsatisfactory *«
> Lord Kelvin

Die Inschrift auf dem Gebäude ›Eleven Twenty-Six‹ ist amerikanisiert (»meager« anstelle von »meagre«); neben den Anführungszeichen finden sich dort fünf kleine Blumen, hier durch * wiedergegeben, welche Auslassungen markieren.[1]

1 Die Deutung der Blumen als Auslassungszeichen ist schlagend; die beiden Blumen um ›and‹ gehen auf den Soziologen William F. Ogburn zurück, der das Diktum ausgewählt und dort Auslassungszeichen gesetzt hatte. (Vgl. Robert K. Merton, David L. Sills und Stephen M. Stigler: »The Kelvin Dictum and Social Science: An Excursion into the History of an Idea«, in: *Journal for the History of the Behavioral Sciences* 20 (1984), S. 319–331.) Thomas Kuhn hat die Blumen beim Zitieren ausgelassen und so eine weitere Version des Diktums geprägt (Thomas S. Kuhn: »The Function of Measurement in Modern Physical Science«, in: *Isis* 52.2 (1961), S. 161–193, hier S. 161; Nachdruck in Thomas S. Kuhn: *The Essential Tension*. Chicago, London 1977, S. 178).

https://doi.org/10.1515/9783110523300-010

Robert K. Merton, David L. Sills und Stephen M. Stigler haben gemeinsam die Geschichte des Diktums rekonstruiert; die Ursprungsvariante findet sich am Anfang der Vorlesung »Electrical Units of Measurement« von 1883:

> I often say that when you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind: it may be the beginning of knowledge, but you have scarcely, in your thoughts, advanced to the stage of *science*, whatever the matter may be.[2]

Das Diktum Kelvins ist in mehreren abweichenden Varianten tradiert, teils mit sprachlichen Abweichungen (»If you cannot measure …«), teils mit inhaltlichen. Eine ähnliche Aussage aus dem Jahr 1692 findet sich bei John Arbuthnot: »There are very few things which we know; which are not capable of being reduc'd to a Mathematical Reasoning; and when they cannot, it's a sign our Knowledge of them is very small and confus'd.«[3] Weitere Varianten der Aussage finden sich bei Roger Bacon, Francis Bacon, Leonardo da Vinci und Immanuel Kant.[4] Es ist nicht verwunderlich, dass das Diktum Kelvins, Abwandlungen davon und ähnliche Aussagen gerade in der Statistik häufig zitiert werden, der Wissenschaft von der Messung und deren Fehlern (neben anderem). Der Epidemiologe und Statistiker Austin Bradford Hill etwa, ein Vorreiter randomisierter klinischer Studien, hat sich auf einen ähnlichen Ausspruch Hermann von Helmholtz' bezogen, dem er die pointiertere Aussage »all science is measurement« zuschreibt.[5]

Die Wahl des Ausspruchs Thompsons für die Fassade des Social Science Research Buildings geht auf den Soziologen William Fielding Ogburn zurück, der von 1927 bis zu seiner Emeritierung an der University of Chicago lehrte; sie zeugt von dem Prestige des Ideals der Quantifizierung in den Wissenschaften.

---

2 William Thomson [Lord Kelvin]: *Popular Lectures and Addresses 1. Constitution of Matter*. London 1889, S. 73f.; vgl. Merton, Sills und Stigler: »The Kelvin Dictum and Social Science«, S. 326.

3 Zitiert nach ebd., S. 325.

4 Vgl. Stephen M. Stigler: *The History of Statistics. The Measurement of Uncertainty before 1900*. Cambridge MA, London 1986, S. 1.

5 »The essence of an experiment in the treatment of a disease lies in comparison. To the dictum of Helmholtz that ›all science is measurement,‹ we should add, Sir Henry Dale has pointed out, a further clause, that ›all true measurement is essentially comparative.‹« (Austin Bradford Hill: *Bradford Hill's Principles of Medical Statistics*. London 1991 [1955], S. 5). Der Ursprung des Zitats ist dunkel.

2 Quantifizierung und mechanische Objektivität

Woher stammt und worauf gründet sich das Prestige quantitativer Methoden in der Wissenschaft? Die übliche Antwort – es gründe sich auf den Erfolg quantitativer Methoden in den Naturwissenschaften – hat der Wissenschaftshistoriker Theodore Porter für unbefriedigend gehalten, da nicht klar sei, warum Methoden, die sich beim Studium von Sternen und Molekülen bewährt hätten, ein Vorbild für die Humanwissenschaften darstellen sollten, und vorgeschlagen, die Blickrichtung umzukehren: Statt das Prestige der Quantifizierung außerhalb der Naturwissenschaften durch deren Erfolg innerhalb zu erklären, sei es fruchtbar, die Rolle von Quantifizierung in der Geschäftswelt und den Sozialwissenschaften zu untersuchen, um ihre Rolle in den Naturwissenschaften besser zu verstehen. Porter betrachtet Quantifizierung wesentlich als eine *Kommunikationsstrategie* – eine Form der standardisierten Darstellung und Übermittlung von Informationen in und durch Zahlen, deren Strenge und Einheitlichkeit dort zum Vertrauen beitragen könne, wo etwa eine persönliche Bekanntschaft mit anderen Forschern nicht gegeben sei: »Quantification«, so Porter, »is a technology of distance.«[6] Mit Quantifizierung sei ein bestimmtes Ideal der Objektivität – »mechanical objectivity« – verbunden, das auf ein Zurückdrängen des Urteilens und einen Kampf gegen subjektive Einflüsse abziele, und es sei diese Art der Objektivität, auf die sich die Autorität der Wissenschaft maßgeblich gründe.

Der Siegeszug der Statistik in den Wissenschaften müsse wenigstens teilweise als eine Antwort auf »conditions of mistrust and exposure to outliers« verstanden werden.[7] Zwar begriffen die Anwender der Statistik diese nicht ganz zu Unrecht als eine der Mathematik entstammende Disziplin; doch sei auffällig, dass statistische Methoden sich eben nicht ›abwärts‹ in der Hierarchie der Wissenschaften verbreitet hätten, von der Mathematik und Physik in die Lebens- und Sozialwissenschaften, sondern im Gegenteil von den ›weichen‹ Disziplinen, der Psychologie und der Medizin, am bereitwilligsten aufgenommen worden seien: eben solchen, in welchen es, innerhalb der Disziplin wie auch von außen, an Vertrauen mangele: »Lack of trust is [...] characteristic of new or weak disciplines. It might almost be taken as the defining feature of weak disciplines.«[8]

---

6 Theodore M. Porter: *Trust in Numbers. The Pursuit of Objectivity in Science and Public Life*. Princeton 1995, S. ix.

7 Ebd., S. 200.

8 Ebd., S. 200.

Ein Beispiel Porters ist die Fehlertheorie, eine der frühesten Routineanwendungen inferentieller Statistik: Während die ersten Anwendungen statistischer Methoden etwa zur Identifizierung von Ausreißern (groben Messfehlern) von Astronomen für eigene Zwecke und an eigenen Daten durchgeführt worden seien, habe sich im 19. Jahrhundert mit größeren Observatorien die Frage gestellt, inwieweit dem Urteil von Assistenten beim Ausschließen von Ausreißern vertraut werden könne – »The problem of recruiting, training, and supervising unprofessional labor was central to the early history of error theory.«[9] In die Psychologie wiederum habe die Statistik über standardisierte Intelligenztests Einzug gehalten, ausgehend von den Vereinigten Staaten, wo (anders als in Europa) der Zugang zu Bildungseinrichtungen Anfang des 20. Jahrhunderts zunehmend auf der Grundlage standardisierter Tests erfolgte.[10]

3 Krisen der Quantifizierung

Beiden mit der Quantifizierung von Wissenschaften verbundenen Idealen der Objektivität ist die Statistik nur teilweise gerecht geworden: Weder verbürgt sie in den einzelnen Disziplinen eine »objektive« Abbildung der Wirklichkeit, noch hat sie das Problem des Vertrauens gelöst. Beide Aspekte finden sich in den Benennungen der Krisen statistischer Quantifizierung wieder, die teils als Replikationskrisen (*replication crises*), teils als Vertrauenskrisen (*crises of confidence*) bezeichnet werden. Das jüngste und gegenwärtig prominenteste Beispiel ist die Vertrauenskrise in der Psychologie. Bedenkt man, dass statistische Tests ihren Einzug in die Psychologie nicht zuletzt im Kontext von Experimenten zum Nachweis parapsychologischer Phänomene gehalten haben, ist es nicht ohne Ironie, dass einer der Anlässe für die Vertrauenskrise eine Veröffentlichung des Psychologen Daryl J. Bem im Jahr 2011 war, in welchem dieser über »Experimental Evidence for Anomalous Retroactive Influences on Cognition and Affect« berichtete, die mit Anlass zur gegenwärtigen Krise gab.[11]

9 Ebd.

10 Ebd., S. 209–211.

11 Vgl. Daryl J. Bem: »Feeling the Future. Experimental Evidence for Anomalous Retroactive Influences on Cognition and Affect«, in: *Journal of Personality and Social Psychology* 100.3 (2011), S. 407–425; zur Bedeutung parapsychologischer Experimente vgl. Ian Hacking: »Telepathy. Origins of Randomization in Experimental Design«, in: *Isis* 79.3 (1988), S. 427–451.

Die Komplexität der Krise und mögliche Ursachen können hier nur angedeutet werden: Eigentümlich für die Psychologie ist die Dominanz einer Form statistischen Testens (des sogenannten *Null Hypothesis Significance Testing*, kurz NHST), welche den Schwerpunkt auf die Angabe eines bestimmten numerischen Werts, des sogenannten *p*-Werts, legt. Dieser gibt auf einer Skala von 0 bis 1 an, wie häufig ein experimenteller Befund oder ein noch ›extremerer‹ ist, *wenn es keinen Effekt gibt*, man es also nur mit statistischem Rauschen zu tun hat. Eine Konvention legt in der Psychologie eine Grenze bei einem *p*-Wert von 0,05 fest, unterhalb der ein Testergebnis als »statistisch signifikant« gilt und für eine Veröffentlichung in Betracht kommt. NHST verbindet, nicht ganz konsistent, Ansätze verschiedener statischer Schulen.[12]

Die Diskussion der letzten Jahre innerhalb der Psychologie hat verschiedene Schwächen dieser Praxis herausgearbeitet: So stehen Forschern viele Handlungsoptionen offen, um die Wahrscheinlichkeit für ein statistisch signifikantes Ergebnis deutlich zu erhöhen (›researcher degrees of freedom‹). Mehrere davon zählt man zu den ›fragwürdigen Forschungspraktiken‹ oder ›Questionable Research Practices‹, kurz QRP, – Praktiken in einem Graubereich zwischen methodisch tadellosem Vorgehen und wissenschaftlichem Fehlverhalten. Spätestens 2015 wurde aus einer Vertrauenskrise eine Replikationskrise, als die *Open Science Collaboration* die Ergebnisse eines großangelegten Replikationsversuchs von 100 Studien aus drei renommierten psychologischen Fachzeitschriften publizierte, deren Erfolg (in der Sicht der Autoren) deutlich hinter dem Wünschbaren zurückblieb.[13]

4 Quantifizierung und Forschungspraktiken

Mit den Geisteswissenschaften, auch den ›digitalen‹, hat all dies auf den ersten Blick wenig zu tun – dass Geisteswissenschaftler das Diktum Kelvins als Wahlspruch überhaupt in Erwägung ziehen würden, ist schwer vorstellbar. Dennoch

12 Vgl. Gerd Gigerenzer, Zeno Swijtink, Theodore Porter, John Beatty und Lorenz Krüger: *The Empire of Chance. How Probability Changed Science and Everyday Life*. Cambridge 1989, Kap. 3.

13 Aus dem Abstract des genannten Artikels: »Replication effects were half the magnitude of original effects, representing a substantial decline. Ninety-seven percent of original studies had statistically significant results. Thirty-six percent of replications had statistically significant results.« (Open Science Collaboration: »Estimating the Reproducibility of Psychological Science«, in: *Science* 349 (2015), S. aac4716-1–aac4716-8, hier S. acc4716-1.)

kann eine Betrachtung der Probleme der Quantifizierung, die sich in anderen Feldern zeigen, lehrreich sein: Zwar sind diese oft in Teilen Probleme bestimmter Arten der Quantifizierung; zugleich machen sie aber auf einen weiteren Aspekt aufmerksam: dass nämlich Formen der Quantifizierung im Kontext von *Forschungspraktiken* beurteilt werden müssen.

Dies soll im Folgenden an einem Beispiel verdeutlicht werden: den sich wandelnden Praktiken im Umgang mit großen Datenmengen. Auf der einen Seite scheinen umfangreiche Daten vor allem Vorzüge zu haben: Je mehr Daten zur Verfügung stehen, umso sicherer sollten (*ceteris paribus*) statistische Beschreibungen der Daten sein; zugleich erlauben große Datensätze, mehr und neuartige Fragen zu stellen und zu untersuchen. Dennoch werfen große Datenmengen auch Probleme auf, denn sie ermöglichen neue Forschungspraktiken, etwa ein ungezieltes mechanisches Suchen nach Mustern – mit der Folge, dass quantitative Maßstäbe, die sich in zielgerichteter Forschung bewährt haben mögen, bei ungezielten Forschungspraktiken keinem der beiden angeführten Ziele der Quantifizierung, dem Ideal absoluter und dem mechanischer Objektivität, gerecht werden.

5 Fragwürdige Forschungspraktiken

In einer umfangreichen anonymen Befragung von über 2000 in der Forschung tätigen Psychologen haben Leslie K. John, George Loewenstein und Drazen Prelec versucht, die Verbreitung von Questionable Research Practices (QRPs) zu ermitteln. Die Befragung umfasste neben Auskünften darüber, in welchem Umfang man selbst in der Forschung auf QRPs zurückgegriffen habe, auch Einschätzungen darüber, wie verbreitet diese bei anderen Forschern seien. In der Studie wurden diejenigen, die angegeben hatten, bestimmte QRPs verwendet zu haben, auch nach einer Bewertung gefragt, in welchem Maße dies zu rechtfertigen gewesen sei (*defensibility rating*); mögliche Angaben waren 0 (nein), 1 (möglicherweise) oder 2 (ja). Die Studie hat nicht zuletzt aufgrund des Ergebnisses Aufmerksamkeit erregt, zu dem John, Loewenstein und Prelec kommen: dass nämlich der Anteil derjenigen, die fragwürdige Praktiken verwendet hätten, überraschend

hoch sei – »that some questionable practices may constitute the prevailing research norm.«[14]

Die Selbstauskünfte und Bewertungen erlauben eine vergleichende Betrachtung, welche Arten von QRPs unter Psychologen als (besonders) fragwürdig angesehen werden, und eine Spekulation über die Gründe. Datenfälschung (*Falsifying data*) wird nur von weniger als 1% der Befragten zugegeben und auch von diesen deutlich als Fehlverhalten bewertet (mit einem Durchschnitt von 0,16 bei der genannten Bewertungsskala). Seltener sind Falschangaben bei Veröffentlichungen hinsichtlich des Experimentdesigns und des Ergebnisses (»In a paper, reporting an unexpected finding as having been predicted from the start« bzw. »In a paper, ›rounding off‹ a *p* value (e.g., reporting that a *p* value of .054 is less than .05)«; die durchschnittliche Bewertung beträgt 1,50 bzw. 1,68). Nur 15,6% gaben an, die Datenerhebung bei einem Experiment schon einmal früher als geplant abgebrochen zu haben, nachdem das gesuchte Ergebnis schon gefunden wurde (»Stopping collecting data earlier than planned«), während 55,9% angaben, schon einmal weitere Daten erhoben zu haben, wenn das Ergebnis noch nicht gefunden wurde (»Deciding whether to collect more data after looking to see whether the results were significant«); die Bewertung der Legitimität weicht dabei zwischen beiden Praktiken allerdings kaum voneinander ab (Durchschnitt 1,76 bzw. 1,79).

Weitere QRPs, nach denen gefragt wurde, umfassen unvollständige Berichte über Experimente (»In a paper, failing to report all of a study's conditions«, 27,7%) oder über Experimentserien (»In a paper, selectively reporting studies that ›worked‹«, 45,8%) sowie Datenkorrekturen (»Deciding whether to exclude data after looking at the impact of doing so on the results«, 38,2%). Unter allen QRPs, nach denen gefragt wurde, am weitesten verbreitet (63,4%) und als am wenigsten fragwürdig bewertet (Durchschnitt 1,84) ist allerdings, nur Teile dessen zu berichten, wonach in einem Versuch gesucht wurde (»In a paper, failing to report all of a study's dependent measures«).[15]

Die Selbstauskünfte und Bewertungen werfen die Frage auf, wie sich die Unterschiede bei den Angaben für die einzelnen Praktiken deuten lassen. Die Bewertungen zeigen, sieht man von Datenfälschung ab, kaum Schuldbewusstsein, wie die durchschnittliche Bewertung von meist über 1,50 anzeigt; bemerkenswert

14 Leslie K. John, George Loewenstein und Drazen Prelec: »Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling«, in: *Psychological Science* 23.5 (2012), S. 524–532, hier S. 524.

15 Ebd., S. 525.

aber ist die Spannbreite beim Eingeständnis der Praktiken. Vergleichsweise selten sind *ausdrückliche* Falschangaben in Veröffentlichungen (wie etwa das Abrunden von *p*-Werten). Dass die Datenerhebung selten bei einem positiven Ergebnis früher als geplant abgebrochen, aber oft über den geplanten Umfang hinaus fortgesetzt wird (bei vergleichbaren Bewertungen hinsichtlich der Rechtfertigung), könnte sich dadurch erklären lassen, dass der geplante Umfang vieler Experimente von vornherein an der unteren Grenze dessen liegt, was aussagekräftige Ergebnisse erwarten lässt. Häufig sind hingegen besonders Formen unvollständiger Berichte: fehlende Angaben über nicht-erfolgreiche Experimente, über Manipulationen im Experiment sowie, vor allem, über mitbetrachtete mögliche Wirkungen.

6 Probleme der Quantifizierung

Man kann an dieser Stelle die Frage aufwerfen, warum unvollständige Berichte überhaupt zu ›fragwürdigen‹ Forschungspraktiken gezählt werden. Die Verbreitung dieser Praxis, die John, Loewenstein und Prelec in der Psychologie gefunden haben, deutet jedenfalls darauf hin, dass unter Forschern ein Unrechtsbewusstsein wenig verbreitet ist.

Eine Antwort auf diese Frage ist eher technischer Natur und nimmt auf die Interpretation des *p*-Werts Bezug. Beträgt die Wahrscheinlichkeit eines statistisch signifikanten Resultats 5%, wenn tatsächlich kein Effekt, sondern nur die erwarteten statistischen Abweichungen innerhalb der Daten vorliegen – dies ist es ja, was ein *p*-Wert von 0,05 der Definition nach aussagt – so liegt umgekehrt die komplementäre Wahrscheinlichkeit dafür, unter dieser Voraussetzung *kein* statistisch signifikantes Resultat zu erhalten, bei 95%. Würde man das Experiment nun 20-mal wiederholen, betrüge die Wahrscheinlichkeit, unter den 20 Versuchen dennoch *wenigstens ein* statistisch signifikantes Resultat zu erhalten, mehr als 64% – denn die komplementäre Wahrscheinlichkeit, in den 20 unabhängigen Versuchen *nie* ein statistisch signifikantes Ergebnis zu erhalten, entspricht gerade $0.95^{20} \approx 36\%$.

Möchte man *p*-Werte als Aussagen über Häufigkeiten deuten, so kommt es daher, wenn man dieser Überlegung folgt, nicht allein auf die Ergebnisse ›erfolgreicher‹ Versuche an, sondern deren Wert vermindert sich, wenn viele Versuche unternommen wurden. – Die Überlegung lässt sich nun auch auf einen Versuch übertragen, in dem 20 verschiedene mögliche Wirkungen eines Eingriffs unter-

sucht werden: Wiederum beträgt die Wahrscheinlichkeit, dass eine dieser Wirkungen statistisch signifikant auftritt, bei 20 möglichen 64%, wenn jede einzelne eine Wahrscheinlichkeit von 5% besitzt.[16]

Diese Überlegung steht im Hintergrund der Auffassung, es stelle eine fragwürdige Forschungspraxis dar, bei einem Bericht über einen statistischen Befund untersuchte mögliche Wirkungen nicht zu erwähnen. Zugleich wird aber auch deutlich, wie problematisch die scheinbar objektiven statistischen Angaben sind: Damit die berechneten Angaben tatsächlich mit ihrer Interpretation als Häufigkeiten in Einklang stehen, müsste man ja nicht allein das gesamte Experiment darstellen, sondern auch gescheiterte Vorversuche und Versuche anderer Forscher mit in Betracht ziehen, was zunächst rein praktisch vielfach unmöglich ist.

Es ist aber zugleich kontraintuitiv: Spinnt man den Gedanken weiter, wären möglicherweise auch Versuche in einem ganz anderen Forschungsbereich für die Interpretation des eigenen Versuchs relevant. Betrachtet man diese Versuche zusammen, so würde die Wahrscheinlichkeit, bei wenigstens einem ein statistisch signifikantes Resultat zu erhalten, ja ebenfalls davon abhängen, wie viele Versuche insgesamt, zu ganz unterschiedlichen Fragestellungen, unternommen wurden. – Es ist nicht offenkundig, wo und mit welcher Begründung hier die Grenze zu ziehen wäre; die Überlegung soll an dieser Stelle einzig vor Augen führen, vor welche Probleme sich der Versuch der Quantifizierung von Geltungsfragen gestellt sieht, wenn man statistische Daten in einem größeren Umfang untersucht. Das fehlende Problembewusstsein vieler Forscher ist vor diesem Hintergrund vielleicht durchaus verständlich.

Die Suche nach Mustern in statistischen Daten wirft eine weitere Frage auf: in welchem Maße nämlich eine rein auf quantitative Maße gestützte heuristische Suche überhaupt erfolgversprechend ist. In einem Gedankenexperiment hat der Pragmatist Charles Sanders Peirce plausibel machen wollen, dass eine mechanische Mustersuche ohne Vorannahmen fruchtlos sei. Peirce schließt seine 1883 erschienene Abhandlung *A Theory of Probable Inference*, eine frühe Theorie der Wahrscheinlichkeitsschlüsse, mit einer kurzen Betrachtung, inwieweit sich diese Theorie auf die »Logic of Scientific Investigation«, die Methodologie, beziehen lasse. Die Betrachtung führt ein Problem vor Augen, das sich einer heuristischen Suche in großen Datenmengen stellt, insofern diese Suche sich rein auf die Betrachtung von statistischen Korrelationen und Abhängigkeiten stützt.

---

16 Eine frühe Betrachtung des Problems findet sich bei Ronald A. Fisher: *The Design of Experiments*. Edinburgh, London 1935, S. 66; voraussetzen muss man freilich, dass die einzelnen Folgen unabhängig sind.

Peirce imaginiert den Fall, ein außerirdisches Wesen suche nach Regelmäßigkeiten und Zusammenhängen im *United States Census*, der in den Vereinigten Staaten seit 1790 verfassungsgemäß alle 10 Jahre stattfindenden Volkszählung. Die Schwierigkeit, die sich stellt, ergibt sich nach Peirce daraus, dass viele mögliche Zusammenhänge, die man näher untersuchen könnte, von vornherein kaum als kausale Zusammenhänge zu deuten wären, was sich aber ohne inhaltliche Kenntnisse nicht erkennen lässt:

> Suppose a being from some remote part of the universe, where the conditions of existence are inconceivably different from ours, to be presented with a United States Census Report, – which is for us a mine of valuable inductions, so vast as almost to give that epithet a new signification. He begins, perhaps, by comparing the ratio of indebtedness to deaths by consumption in counties whose names begin with the different letters of the alphabet. It is safe to say that he would find the ratio everywhere the same, and thus his inquiry would lead to nothing. […] The stranger to this planet might go on for some time asking inductive questions that the Census would faithfully answer, without learning anything except that certain conditions were independent of others. At length, it might occur to him to compare the January rain-fall with the illiteracy. […]
>
> He would infer that in places that are drier in January there is, not always but generally, less illiteracy than in wetter places. A detailed comparison between Mr. Schott's map of the winter rain-fall with the map of illiteracy in the general census, would confirm the result that these two conditions have a partial connection. […] Now we, knowing as much as we do of the effects of winter rain-fall upon agriculture, upon wealth, etc., and of the causes of illiteracy, should come to such an inquiry furnished with a large number of appropriate conceptions; so that we should be able to ask intelligent questions not unlikely to furnish the desired key to the problem. But the strange being we have imagined could only make his inquiries haphazard, and could hardly hope ever to find the induction of which he was in search.[17]

Das Gedankenexperiment dient Peirce dazu, einen Punkt plausibel zu machen, der ihn über lange Zeit beschäftigt hat: Dass es nämlich verwunderlich und erklärungsbedürftig ist, dass man überhaupt je zu Theorien und Erklärungen gelangt, die sich bewähren. Die Erklärung, die Peirce für dieses Problem gibt, liegt darin, eine natürliche Anlage, richtig zu raten, anzunehmen. Den Mechanismus deutet Peirce in der angeführten Schrift nur an: Es seien angeborene, wenngleich rohe Ideen (der Kraft, des Raumes und der Zeit, der menschlichen Natur), auf

17 Charles S. Peirce: »A Theory of Probable Inference«, in: *Writings of Charles Sanders Peirce. A Chronological Edition, Volume 4: 1879–1884*, hg. v. Christian J. W. Kloesel. Bloomington, Indianapolis 1989 [1883], S. 408–450, hier S. 446–447.

welchen erfolgreiches Raten und damit auch die Naturwissenschaften und ›moral sciences‹ notwendig beruhten: »all human knowledge, up to the highest flights of science, is but the development of our inborn animal instincts.«[18]

Die von Peirce imaginierte Situation, ein mechanisches Durchsuchen von Daten ohne Vorannahmen und Fragestellungen, war zu seiner Zeit keine Forschungspraxis, die ernsthaft in Betracht gekommen wäre – das ›strange being‹ unterscheidet sich von menschlichen Forschern durch das Fehlen der forschungsleitenden Ideen, nicht in der Art der Forschung. Heute verfügen wir über Computerprogramme, die nach Korrelationen oder Hinweisen auf Kausalbeziehungen in Daten suchen.[19] Auch das Experimentieren hat sich gewandelt: In der Pharmaforschung werden heute in High-Throughput-Screenings oft abertausende Substanzen pro Tag auf mögliche Leitstrukturen zur Entwicklung neuer Arzneistoffe durchmustert; nicht zuletzt führt das stetige Wachstum der zur Verfügung stehenden Daten in vielen Forschungsfeldern dazu, dass blindes Suchen nach Mustern zu einer möglichen Heuristik der Forschung geworden ist.

7 Kelvins Diktum

Ob Peirce' Annahme, die blinde Suche nach Mustern und Korrelationen stelle eine aussichtslose Forschungspraxis dar, noch zutrifft, ist letztlich eine empirische Frage; die Schwierigkeit bei der Interpretation der *p*-Werte in statistischen Experimenten mit vielen untersuchten Wirkungen zeigen allerdings an, dass die Interpretation dieses quantitativen Maßes nicht unabhängig von den Forschungspraktiken ist, in welche es eingebettet ist. Die Verwendung von *p*-Wert-Verfahren in den Wissenschaften hat ihren Ausgang Anfang des 20. Jahrhunderts von agrarwissenschaftlichen Experimenten genommen; vor allem war es der Statistiker und Biologe Ronald A. Fisher, der viele der Verfahren während seiner Tätigkeit als Chief Statistician an der Rothamsted Experimental Station entwickelte. Eines seiner Ziele war es, exakte Verfahren zur Beurteilung gerade sehr kleiner Stichproben bzw. Versuche zu finden – der Grund dafür war eben, dass agrarwissenschaftliche Experimente aufwendig und teuer sind, so dass das Erheben gro-

18 Ebd., S. 450.

19 Siehe z. B. Clark Glymour, Richard Scheines, Peter Spirtes und Kevin Kelly: *Discovering Causal Structure. Artificial Intelligence, Philosophy of Science, and Statistical Modelin*g. Orlando 1987.

ßer Datenmengen nicht in Frage kam. Trotz der kleinen Stichproben, die der Forschung zur Verfügung standen, bereitete aber die Interpretation der *p*-Werte weniger Schwierigkeiten als heute. Dies rührt teils daher, dass die Wahl kleiner Stichproben die Kehrseite der hohen Kosten von Experimenten war – und die Kosten von Experimenten eine weitere Wirkung hatten: Sie schlossen nicht allein umfangreiches, sondern zugleich auch *ungezieltes* Experimentieren in größerem Umfang praktisch aus.

Den Wert von Quantifizierung an Forschungspraktiken zu binden, scheint theoretisch wie praktisch unbefriedigend: theoretisch, weil die Grenze zwischen ›explorativen‹ und ›testenden‹ Forschungspraktiken unscharf und willkürlich erscheint; praktisch, weil der Nutzen großer Datenmengen gerade auch darin zu liegen scheint, eine explorative Suche nach Zusammenhängen zu ermöglichen, nach denen man gezielt nicht gesucht hätte. Insofern aber die Quantifizierung nicht um ihrer selbst willen erstrebt wird, sondern eine Form der Objektivität wissenschaftlicher Geltungsansprüche begründen soll, ist der Bezug auf Forschungspraktiken dennoch sinnvoll: Kelvins Diktum ist völlig vereinbar damit, dass auch Wissen von etwas, das sich messen und in Zahlen ausdrücken lässt, oft dürftig und unbefriedigend ist.[20]

Bibliographie

Bem, Daryl J.: »Feeling the Future. Experimental Evidence for Anomalous Retroactive Influences on Cognition and Affect«, in: *Journal of Personality and Social Psychology* 100.3 (2011), S. 407–425.

Fisher, Ronald A.: *The Design of Experiments*. Edinburgh, London 1935.

Gigerenzer, Gerd, Zeno Swijtink, Theodore Porter, John Beatty und Lorenz Krüger: *The Empire of Chance. How Probability Changed Science and Everyday Life*. Cambridge 1989.

Glymour, Clark, Richard Scheines, Peter Spirtes und Kevin Kelly: *Discovering Causal Structure. Artificial Intelligence, Philosophy of Science, and Statistical Modeling*. Orlando 1987.

Hacking, Ian: »Telepathy. Origins of Randomization in Experimental Design«, in: *Isis* 79.3 (1988), S. 427–451.

Hill, Austin Bradford: *Bradford Hill's Principles of Medical Statistics*. London 1991 [1955].

John, Leslie K., George Loewenstein und Drazen Prelec: »Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling«, in: *Psychological Science* 23.5 (2012), S. 524–532.

Kuhn, Thomas S.: »The Function of Measurement in Modern Physical Science«, in: *Isis* 52.2 (1961), S. 161–193.

20 Ich danke der VolkswagenStiftung für die Förderung durch das Dilthey-Fellowship ›Wissenschaft und Werte‹, in dessen Rahmen diese Arbeit entstanden ist.

Kuhn, Thomas S.: *The Essential Tension*. Chicago, London 1977.
Merton, Robert K., David L. Sills und Stephen M. Stigler: »The Kelvin Dictum and Social Science. An Excursion into the History of an Idea«, in: *Journal for the History of the Behavioral Sciences* 20 (1984), S. 319–331.
Open Science Collaboration: »Estimating the Reproducibility of Psychological Science«, in: *Science* 349 (2015), S. aac4716-1–aac4716-8.
Peirce, Charles S.: »A Theory of Probable Inference«, in: *Writings of Charles Sanders Peirce. A Chronological Edition, Volume 4: 1879–1884*, hg. v. Christian J. W. Kloesel. Bloomington, Indianapolis 1989 [1883], S. 408–450.
Porter, Theodore M.: *Trust in Numbers. The Pursuit of Objectivity in Science and Public Life*. Princeton 1995.
Stigler, Stephen M.: *The History of Statistics. The Measurement of Uncertainty before 1900*. Cambridge MA, London 1986.
Thomson, William [Lord Kelvin]: *Popular Lectures and Addresses 1. Constitution of Matter*. London 1889.