27.12.2016

Replikation: Überzeugungskraft psychologischer Erklärungen

Marc Dressler

inspective.

Inhaltsverzeichnis

1 Einleitung: dem Leser zum Gruße …… 4

2 Replikationen …… 7
2.1 Entstehung der Replikationskrise …….. 7
2.2 Der Begriff der Replikation …..……. 10
2.3 Das Konzept des Reproduzierbarkeitsprojektes: Psychologie …….. 13
2.4 Seismik systematischer Replikationen …….. 16

3 Originalstudie ……... 20
3.1 Neurologie und Psychologie …….. 20
3.2 Dekonstruktion der Verführungskraft neurologischer Erklärungen ……... 25

4 Die Replikation …… 29
4.1 Rekonstruktion …….. 30
4.1.1 Modell …… 30
4.1.2 Einbettung …….. 31
4.1.3 Formalisierung der Linearen Regression im Hierarchischen Modell …… 32
4.2 Reanalyse …….. 35
4.2.1 Effektgröße …….. 36
4.2.2 Stichprobenumfang ……. 46
4.2.2.1 Präzisionsansatz ……... 46
4.2.2.2 Teststärkenansatz ……. 49
4.3 Direkte Replikation …… 57
4.3.1 Ergebnisse …… 58
4.3.2 Diskussion …… 63
4.3.2.1 Zur Replizierbarkeit der Originalstudie …… 63
4.3.2.2 Zur Replikation der Originalstudie …….. 68

5 Kleine Methodologie der Replikation ……... 72
5.1 Der Replikationserfolg ……... 72
5.2 Der Erfolg des Replikationserfolges …….. 75
5.2.1 Induktive Bestätigung …….. 76
5.2.2 Statistische Bestätigung ……... 80
5.2.2.1 Signifikanztheorie ……. 85
5.2.2.2 Entscheidungstheorie …… 87
5.2.2.3 Subjektivismus……... 89
5.2.2.4 Likelihood ……. 91
5.2.3 Wissenschaftlicher Fortschritt durch Replikation …… 93
5.2.3.1 Kumulativer Fortschritt ……. 93
5.2.3.2 Organisches Wachstum ……. 95
5.2.3.3 Versuch und Irrtum ……... 98
5.3 Zweifel am Replikationserfolg ……... 100
5.3.1 Statistik-Recycling …….. 100
5.3.2 Effektgröße, Teststärke, Konfidenzintervall, Meta-Analyse …….…….. 107

6 Die Replikationsindustrie in der Wissensgesellschaft ……..….. 110
6.1 Monetarisierte Replikationswahrscheinlichkeit ……. 113
6.2 Die Kunst der Produktion und Reproduktion …….. 118
6.3 Herausforderungen im Computerzeitalter ……... 121

7 Literatur …….. 124

1 Einleitung: dem Leser zum Gruße

Wie die Verführungskraft neurologischer Erklärungen symptomatisch ist für ihre Replikation, und wie die Replikation symptomatisch ist für Replikationen überhaupt, so sind Replikationen symptomatisch für den gesamten Forschungsbetrieb. Dies zu zeigen hat sich die Arbeit vorgenommen, nicht mehr, aber auch nicht weniger.

Die Arbeit nimmt ihren Ausgang in der Verführungskraft einzelner Erklärungen und endet mit der Verführungskraft von Replikationen. Dazwischen liegen, hierarchisch eingebettet, die Erörterung historischer und methodologischer Voraussetzungen für den Ruf nach Replikationen, Beschreibungen ihrer wandelbaren Gestalt und den Gründen wissenschaftlicher Not; es wird die Rede sein von der Gralssuche nach dem Archimedischen Punkt im Universum, an dem der Replikationshebel angesetzt werden kann, um die Forschungslandschaft von falsch-positiven Ergebnissen zu befreien.

Bei diesem Trachten wird in mancher Aporie manche taube Nuss zu knacken sein, wenn reflexiv auf Schlüsse geschlossen wird, vor Kreisen in Kreisen fliehend, wenn die philosophische Hintertreppe zum Laufrad wird und die Induktion auf zureichendem Grunde wie die Statistik auf unzureichendem Grunde nach Worten der Verbindung ringen, während die Sprache leerläuft (Wittgenstein 1990, §132). Es wimmelt an verführerischen Formen und Formeln, die individuelle Beurteilungen wissenschaftlicher Erklärungen im selben Maße bewegen wie kollektive Entscheidungen zur Auflage von Forschungsprogrammen. Wo es vordergründig um Rationalität geht, wirkt hintergründig oft ein dumpfer Wille (Schopenhauer 1977, S.148), dessen Bahnen unvorhersehbar sind, auch nicht einzufangen in Wahrscheinlichkeiten.

Die Einbettung von Replikationen in eine umfassende Reform der Wissenschaft, die befeuert wird von einer anhaltenden Methodenkritik an der kanonischen Statistik, wird entlang der Replikationskrise, die im zweiten Kapitel beschrieben wird, herausgearbeitet. Das dritte Kapitel gilt der zu replizierenden Studie von Weisberg, Taylor und Hopkins (2015) zur Dekonstruktion der Verführungskraft neurologischer Erklärungen und der darin formulierten Hypothese, dass psychologische Erklärungen mit irrelevanter Neuroinformation besser beurteilt werden als dieselben psychologische Erklärung ohne Neuroinformation.

Dann erfolgt im vierten Kapitel auf der Grundlage des Originaltextes exemplarisch die Rekonstruktion, Reanalyse und Replikation der Studie, wobei ambivalente Eigenheiten einer Replikation im Hierarchischen Modell herausgearbeitet werden, die ihre abschliessende Interpretation erschweren.

Die hier ausgeführte direkte Replikation steht zunächst im Kontext der Originalstudie und rückt im fünften Kapitel in den Kontext des Ethos der Open Science Collaboration (OSC). Im Kontext der Originalstudie wurde erörtert, was eine Replikation zu einer erfolgreichen Replikation macht; nun wird der Erfolg erfolgreicher Replikationen erörtert in dem Sinne, wie Replikationen zum wissenschaftlichen Fortschritt beitragen. Die Bestätigungsfunktion vorausgegangener Studien führt unmittelbar auf das Induktionsproblem, dessen Lösungsversuch mit den Mitteln der kanonischen Statistik diskutiert wird.

Die Diskussion des fünften Kapitels schließt damit, dass Replikationen einen Originalbefund nicht bestätigen, weil sie keinen Zustrom an Wahrscheinlichkeit begründen: war die Wahrscheinlichkeit für ein signifikantes Ereignis im Original *P*, dann ist die Wahrscheinlichkeit dafür auch in allen folgenden Studien *P*, sofern die Studien unabhängig sind, d.h. ihr Ausgang nicht von der Versuchsreihenfolge abhängt. Jeder Datensatz kann zustande kommen unter ganz verschiedenen Zuständen der Welt, sodass man aus den Daten nicht herauslesen kann, welcher Zustand der Welt sie hervorgebracht hat.

Dass daran auch die Verwendung von Effektgrößen oder Konfidenzintervalle nichts ändert, liegt dem hier vorgeschlagenen Ansatz zufolge an der philosophischen Grammatik der an der statistischen Problemstellung beteiligten Begriffe. Sie zementieren eine Dichotomie, die logisch nicht durchbrochen werden kann, praktisch aber niemanden vor Schwierigkeiten stellt, sodass die Lösung des Induktionsproblems nur eine politische sein kann, die die Grammatik der Begriffe nachhaltig verändert. Methodologische Erwägungen können aus sich heraus eine philosophische Grammatik nicht sinnvoll ändern. Diese Analyse steht im Einklang mit dem Ethos der OSC.

Im abschließenden sechsten Kapitel evaluiert der Autor jüngere Ansätze der Statistik und stellt sie in den Kontext einer industrialisierten Wissensgesellschaft. Besondere Aufmerksamkeit gilt der Verdrängung von wissenschaftlichen Werten der Objektivität durch den ökonomischen Wert der Effizienz. Vor dem Hintergrund der Produktion bekommt die Replikation als Reproduktion den kulturaffirmativen Charakter einer tech-

nischen Dienstleistung, der mindestens so plausibel ist wie die Bestätigungsfunktion einer Replikation, aber ablenkt von sozialwissenschaftlichen Tendenzen, die eine Replikation völlig obsolet machen.

Eine weitere unheimliche Tendenz in der Wissenschaft drängt sich auf bei der Beschäftigung mit psychologischer Methodologie: ist jeder fünfte Artikel in Psychologiefachzeitschriften fehlerhaft, unvollständig oder schlampig (Bakker & Wicherts 2011), stimmt so gut wie keine Statistik, weil die berichteten Werte nicht zueinander passen, falsch gerundet wird, die Freiheitsgrade nicht stimmen oder die *p*-Werte von *F*-Statistiken halbiert werden (Ioannidis 2005), ganz zu schweigen von den desaströsen Ergebnissen bei der Interpretation von Statistiken (Tversky & Kahneman 1971), möchte man meinen, dass Psychologen nicht wissen, was sie tun: οὐ γὰρ οἴδασιν τί ποιοῦσιν (Luk 23, 34).

2 Replikationen

Replikationen treten in ihrem historischen Kontext plastisch hervor. Schon in der Entstehung der Replikationskrise wird deutlich werden, dass es in ihr um eine groß angelegte Wissenschaftsreform geht, die ihren Auslöser hat in einer sich radikalisierenden Methodenkritik und einer damit einhergehenden Stagnation der Psychologie. Mit einem Bündel von Instrumenten und Maßnahmen, zu denen die Replikation zählt, soll die Psychologie einem kumulativen Fortschritt zugeführt werden. Weil die in die Replikation gesteckten Erwartungen nicht sämtliche von einer Replikationsform erfüllt werden können, werden mit Blick auf die Fortschrittsermöglichung sechs Replikationsformen entwickelt, von denen die Reanalyse sowie die direkte und die konzeptuelle Replikation weiterverfolgt werden. Schließlich wird der Ethos des treibenden Reformmotors hinter dem Reproduktionsprojekt: Psychologie vorgestellt und Replikationen als Grundwert im Sinne von erstrebenswerten Zielen herausgearbeitet. Die Resonanz auf das Projekt rundet das Kapitel ab.

2.1 Entstehung der Replikationskrise

Die Replikationskrise kann gesehen werden als Symptom einer umfassenderen Krise, die im Zuge einer reflexiven Bestandsaufnahme in der Wissenschaft länger schon schwelt: es fehlt ihr an einer einheitlichen Methodik und etablierten Gesetzmäßigkeiten (Mittelstaedt & Zorn 1984), stattdessen herrscht eine Orientierungslosigkeit vor, und ein kumulativer Fortschritt ist, insbesondere in der Psychologie, nicht auszumachen (Staats 1983, S.11). In diesem Krisenszenario einer stagnierenden und fragmentierten Wissenschaft finden sich Faktoren, die eine Replikation unmöglich, überflüssig, zufällig, zweifelhaft und am Ende doch notwendig machen (Rosenthal 1989).

Diese Faktoren resultieren aus einer Kritik an den verwendeten Tests und am Erkenntnisinteresse, das die Verwendung der Tests motiviert. Der Kritik zufolge haben Tests eine zu geringe Stärke, keine Aussagekraft oder sie werden fehlerhaft verwendet, interpretiert oder missbraucht; das Erkenntnisinteresse wiederum fokussiere ausschließlich auf den positiven Nachweis von Effekten mittels Signifikanztests.

Die Replikation eines Effekts ist unmöglich, wenn der Effekt gar nicht existiert. Nicht existierende Effekte schleichen sich ein in den Bestand wissenschaftlicher Erkenntnisse,

wenn die Daten zuvor so lange getrimmt werden, bis sie die Kriterien eines Nachweises erfüllen. Das Spektrum des Datentrimmens reicht bei Signifikanztests vom selektiven Sammeln und Analysieren der Daten bis hin zu ihrer verfälschenden Manipulation (Simmons, Nelson & Simonsohn 2011).

Die Replikation eines Effektes ist ebenfalls unmöglich, wenn der Effekt mit einem Test nicht nachgewiesen werden kann. Das ist bei Signifikanztests insbesondere dann der Fall, wenn Effekt und Stichprobe klein sind. Dann reicht die Teststärke nicht aus, um zwischen Signalausschlägen des Effekts und nur zufälligen Schwankungen in der Population zu differenzieren (Cohen 1977). Für mittlere Effektgrößen liegt die Teststärke in der Psychologie durchschnittlich bei nur 45 Prozent (Gigerenzer 1989).

Die Replikation eines Effekts ist überflüssig, wenn der Test ein Gütekriterium vorhält, das seine Reliabilität verbürgt. Als solches Gütekriterium gilt der *p*-Wert. Er steht für die Wahrscheinlichkeit, eine Statistik zu erhalten, die mindestens so extrem ist wie die erhobene – vorausgesetzt, die Nullhypothese trifft zu, d.h. vorbehaltlich der Annahme, dass kein Effekt existiert. Je kleiner der *p*-Wert ist, umso erdrückender ist die Beweislast gegen die Nullhypothese. Ein kleiner *p*-Wert wird häufig interpretiert als Wahrscheinlichkeit, den Effekt erfolgreich zu replizieren (Gorroochurn, Hodge et al. 2007).

Würde der *p*-Wert den Replikationserfolg vorwegnehmen, dürfte der *p*-Wert einer Hypothese, die eine andere Hypothese enthält, für denselben Datensatz nicht größer ausfallen als der *p*-Wert der anderen Hypothese. Das kommt aber vor (Barber & Ogle 2014). Tatsächlich variiert der *p*-Wert bei gleichgroßen Stichproben derselben Population zwischen 0.001 und 0.760 (Cumming 2008). Mithin können bei variierenden Stichproben verschieden große Effekte gleiche *p*-Werte besitzen und gleichgroße Effekte verschiedene *p*-Werte. Ist die Effektgröße null, ist jeder *p*-Wert gleichwahrscheinlich, d.h. er ist bei 5 Prozent der Stichproben kleiner als 5 Prozent und bei 50 Prozent der Stichproben kleiner als 50 Prozent usw. Selbst bei einer Teststärke von 90 Prozent variieren die *p*-Werte noch beträchtlich (Halsey 2015). Ein derart unzuverlässiges Maß eignet sich nicht für Angaben zur Replizierbarkeit (Kelly 2006; Head 2015).

Die Replikation ist zufällig, wenn ein Effekt für zu groß gehalten wird. Denn ein größenadäquater Test ist zu schwach für den Nachweis eines Effekts, wenn der Effekt in Wirklichkeit kleiner ist. Die Effektgröße wird überschätzt, weil sich die Forschung nur für das Eintreten eines Effektes interessiert und nicht für sein Ausbleiben. Sofern aber

nur die positiven Resultate eines Experiments veröffentlicht werden, kann die Größe eines Effekts statistisch nicht zur Mitte tendieren, weil die negativen Resultate zur ausgleichenden Korrektur fehlen (Rosenthal 1979).

Die Replikation ist zweifelhaft, wenn nur ein Instrument verwendet wird und dieses Instrument fundamentale Mängel aufweist. Das nahezu einzige Instrument der experimentellen Psychologie und Medizin, dem ungeachtet seiner Mängel doppelblind vertraut wird, ist der Signifikanztest (Hill 1965). Die Mängel des Signifikanztests manifestieren sich Kritikern zufolge in seiner Realitätsferne und Inkonsistenz. So kann man beispielsweise für jede Effektgröße einen Stichprobenumfang und eine Irrtumswahrscheinlichkeit α so angeben, dass die Nullhypothese verworfen werden muss, der bedingte Rückschluss (Mises 1951, S.140) aber nachträglich die Wahrscheinlichkeit etabliert, dass die Nullhypothese zu $1-\alpha$ Prozent falsch ist (Lindlay 1957).

Schwerer als solch exotische Paradoxien dürfte allerdings der Zweifel wiegen an der Eignung des Signifikanztests für den Nachweis realer Effekte. Dieser Zweifel setzt an am Zustandekommen des *p*-Wertes: In der Berechnung des *p*-Wertes machen die empirischen Daten, die erhoben wurden, nur einen geringen Anteil aus. Den weitaus größeren Anteil stellen fiktive Daten, die nicht erhoben wurden. Die fiktiven Daten sind die Daten, die, sofern die Nullhypothese zutrifft, vorliegen müssten, aber eben nicht vorliegen. Das impliziert, dass unter Umständen die Nullhypothese fälschlicherweise beibehalten wird, weil die Alternativhypothese keine Daten vorhersagt, die nicht erhoben wurden. Dass Fiktionen ausschlaggebender sind als Beobachtungen, unterminiert in solchen Fällen das empirische Fundament einer Erfahrungswissenschaft (Barber & Ogle 2014; Burnham 2014).

Replikationen sind notwendig, wenn sie die Voraussetzungen für ihre eigene Anwendung schaffen, es aber kaum welche gibt. Replikationen müssen die stereotype Testpraxis durchbrechen, die Rate falsch-positiver Resultate reduzieren, Effektgrößen zurückstutzen und auf dem Weg einer umfassenden Datenbereinigung (Kruskal 1981) die Entwicklung einer kumulativen Wissenschaft ermöglichen oder gar beschleunigen (Schmidt 1996). Ohne Replikationen wäre wissenschaftlicher Fortschritt nicht möglich (Hubbard, Vetter & Little 1998). Als integraler Bestandteil von Wissenschaft schaffen Replikationen somit die Voraussetzungen von etwas, dem sie – in veränderter Form – angehören wird, weil sonst der Fortschritt zum Stillstand käme, aber – deshalb in ver-

änderter Form – noch nicht angehören, sonst wäre die Wissenschaft längst fortgeschritten.

Wie viele Replikationen es gibt, ist umstritten. Da weder eine erfolgreiche noch eine gescheiterte Replikation veröffentlicht wird – die eine bringt nichts Neues, die andere nichts Signifikantes (Evanschitzky, Baumgarth, Hubbard & Armstrong 2007) –, ist von einer geringen Replikationsdichte auszugehen. Makel, Plucker und Hegarty (2012) schätzen den Anteil der Replikationen an veröffentlichten Studien auf 1 Prozent, wobei die veröffentlichten Replikationen hauptsächlich gelungene Replikationen seien. Dagegen enthalten laut Neuliep und Crandall (1993) Dreiviertel der veröffentlichten Studien Replikationen. Der behauptete Anteil variiert mal mehr, mal weniger, je nachdem, was die Forscher gerade unter Replikation verstehen.

2.2 Der Begriff der Replikation

Replikationen sind nur verstehbar als Bestandteil einer Forschungspraxis, die innerhalb einer Wissenschaftskultur besteht (Travis 1981; Hendrick 1990). Dieser krisengebeutelten Kultur entwachsen Desiderata, die in funktionale Eigenschaften eines methodischen Verfahrens übersetzt und anhand zweckbezogener Dimensionen in verschiedene Formen der Replikation ausdifferenziert werden. Jeder Replikationsform lässt sich dann eine Funktion in einem geordneten Wissenschaftssystem zuweisen.

Wenn eine Studie eine andere Studie repliziert, dann geschieht das auf dem Boden ihrer Vergleichbarkeit. Nun gibt es aber zahllose Vergleichsmöglichkeiten zur Bestimmung der Äquivalenz von Studien. Auch unter der Einschränkung der Messbarkeit verbleibt eine Unzahl möglicher Aspekte, aus der die maßgeblichen Aspekte ausgesondert werden müssen in dem Sinne, dass sie eine formale und materiale Äquivalenz von Studien gestatten, also Auskunft geben über Gemeinsamkeiten und Unterschiede im Gegenstand und seiner Handhabung im Experiment (Schwarz & Strack 2014). Erst die Nennung dieser Aspekte zusammen mit einer Begründung, warum die Aspekte maßgeblich sind, macht die Beurteilung von Äquivalenzen möglich (Klein et al. 2014).

So wesentlich die Begründung für die Möglichkeit von Replikationen ist, so schwierig ist sie auch, weshalb in veröffentlichter Experimentalstudien der eigentliche Inhalt im Diskussionsteil neben dem methodischen Apparat sich häufig diffus oder kärglich aus-

nimmt. Da der Ausgang von Replikationen schwerwiegende Folgen zeitigt, ist der Wunsch nach einer Standardisierung verständlich (King 1995; Clemens 2015; Shavit 2016) und ihre Begründung unumgänglich. Um bei der Begründung den Fallstricken einer transzendentalen Deduktion zu entgehen, erfolgt sie im folgenden nicht vom ersten Anfang aus, sondern vom letzten Ende her, also zweckbezogen statt ersturspünglich. Dieses Vorgehen rechtfertigt sich damit, dass eine Wissenschaftsreform eine Sozialreform ist, die – mit Replikationen – bestimmte Zwecke verfolgt (Campbell 1969).

Der übergeordnete Zweck eine Replikation liegt in der Selektion des Wahren gegenüber dem Falschen (Schlosberg 1951). Bescheidener: die Replizierbarkeit einer Studie macht ihr Resultat zu einem ernstzunehmenden Kandidaten für Wahrheit. Denn durch Wiederholung und Wiederwiederholung nähert man sich in induktiven Schritten der Wahrheit (Redi 1664) – nicht etwa in der Physik, sondern in der Medizin, wo der Replikationsbegriff historisch seine Wurzeln hat. Ging es Redi noch um die Wiederholung der eigenen Experimente, so nutzte sein Landsmann Fontana (1787) Replikationen bereits, um Fehler aufzudecken, die andere gemacht haben.

Weil der Abstand zur Wahrheit sich nicht angeben lässt, solange das Muster fehlt, das allein den Anspruch auf Annäherung einlösen könnte, behilft man sich zur Orientierung der induktiven Schritte mit Indikatoren, die Eigenschaften der Wahrheit verkörpern – zumindest soweit, dass wir damit epistemologisch befriedigt, d.h. hinreichend konfident sind (Dennis & Valacich 2015). Am Ende der Konfidenz steht die Anwendbarkeit des Experimentalbefunds, die sich auf eine Vorhersagekraft stützt, die wiederum getragen wird von der Genauigkeit und Präzision des Befunds[1], wobei die Genauigkeit der Anhäufung äquivalenter Befunde geschuldet ist, die sowohl zufällige Schwankungen (nach dem Gesetz der großen Zahlen) glättet als auch die Entpersonalisierung und Dekontextualisierung des Befundes leistet, auf der die Genauigkeit ruht – und die Anhäufung der Befunde verdanken wir endlich den Replikationen.

Replikationen versorgen die Wissenschaft mit einem stabilen, robusten Fundament, wie es die Wahrheit auch täte, wenn wir sie nur wüssten. Replikationen sind der Grundstein wissenschaftlicher Erkenntnisansprüche (Lindsay & Ehrenberg 1993), auf dem Wissen-

1 Ein Befund ist umso genauer, je weniger verzerrt seine Messung ist, und er ist umso präziser, je weniger sein Messwert schwankt, d.h. je schmaler sein Konfidenzintervall ausfällt (Hunter 2001; Thompson 2012).

schaftler ihr Gebäude errichten, indem sie Stein auf Stein schichten, weil es ihnen nicht um bloße Wiederholungen geht, sondern um kumulative Wiederholungen (Danziger 1988). Replikationen sind somit architektonischer Entwurf und Abrissbirne in einem: sie restringieren den Bau auf seine Substanz und erlauben zugleich Generalisierungen auf weitere Formen und Materialien (Allen & Preiss 1993). Mit anderen Worten: sie sind restriktiv und expansiv, selektiv und integrativ als auch instruktiv und interogativ.

Solch widersprüchliche Funktionen kann eine Replikationsform alleine unmöglich erfüllen. Es sind daher, dem scholastischen Willen zum System nachgebend, Dimensionen des wissenschaftlichen Experimentalraumes zu fixieren, in deren Koordinaten Replikationsformen verortet werden können, die mit den entsprechenden funktionalen Eigenschaften ausgestattet sind. Die Auswahl an Dimensionen ist beachtlich, da es an Klassifizierungen von Replikationen nicht mangelt (Lykken 1968; Neuliep & Crandall 1993; Tsang & Kwan 1999; Gómez & Jurista 2010; Peng 2011; Camfield & Palmer-Jones 2013; Clemens 2015).

Für die Begründung der Psychologie als Wissenschaft, dergemäß Replikationen die Funktion haben, die konstante Annäherung an die Wahrheit sicherzustellen mittels stabiler und genau umrissener Effekte, genügen drei Dimensionen: das Maß für die Messung, das Messdesign und das Gemessene (Datensatz). Da ein identischer Datensatz nur in Kombination mit einem identischen Design sinnvoll ist, ergeben sich sechs Replikationsformen, die in Tabelle 1 dargestellt sind.

	Maß	Design	Datensatz
Reanalyse	identisch	identisch	identisch
Reinterpretation	verschieden	identisch	identisch
Replikation, konzeptuell	identisch	verschieden	verschieden
Exploration, frei	verschieden	verschieden	verschieden
Replikation, direkt	identisch	identisch	verschieden
Exploration, instrumentell	verschieden	identisch	verschieden

Tabelle 1: Formen der Replikation in drei ausgewählten Dimensionen.

Die Freiheitsgrade der Replikationsformen nehmen von oben bis zur freien Exploration monoton zu und danach wieder ab. Die direkte Replikation erzeugt mit ihrem Freiheitsgrad einen neuen Datensatz und übernimmt die abhängigen und unabhängigen Variablen der Originalstudie; sie dient der Reliabilität und hat dafür zu sorgen, dass alle Messungen eine akzeptable Konstanz vorweisen. Die konzeptuelle Replikation übernimmt

nur die abhängige Variable der Originalstudie; sie dient der Validität und hat dafür zu sorgen, dass im Zuge einer Generalisierung Brücken geschlagen werden zu entfernteren Wissensgebieten und so die Resultate eine Kreuzvalidierung erfahren (Amir & Sharon 1990; Hones 1999; Berthon, Ewing & Carr 2002).

Erwähnenswert ist unter den Replikationsbegriffen noch die Pseudo-Replikation, da sie fast die Hälfte aller Replikationen in der Ökologie ausmacht (Hurlbert 1984). Eine der Replikationsformen wird zur Pseudo-Replikation, wenn die Daten mit statistischen Methoden analysiert werden, ohne dass die Voraussetzungen dafür erfüllt sind. Ihr schließt sich die Pseudo-Evaluation einer Replikationsform an, die verkennt, wozu die statistischen Methoden eingesetzt wurden. So macht es für die Beurteilung des Replikationserfolges einen großen Unterschied, ob es in einer Studie um die Schätzung eines Populationsparameters in den Grenzen eines Konfidenzintervalls geht oder um das Verwerfen einer Nullhypothese. Denn es kann durchaus vorkommen, dass sich aus demselben Datensatz die Beibehaltung der Nullhypothese ergibt, obwohl sich die Konfidenzintervalle von Originalstudie und Replikation hinreichend überlappen. Bezogen auf den ersten Zweck wäre die Replikation des Originaleffektes erfolgreich, bezogen auf den zweiten Zweck nicht (Greenwald, Gonzalez, Harris & Guthry 1996).

2.3 Das Konzept des Reproduzierbarkeitsprojektes: Psychologie

Das Reproduzierbarkeitsprojekt: Psychologie sieht in Grundwerten den Grund für eine Wissenschaftsreform, deren Transmissionsriemen Replikationen sind. Für Replikationen ist ein standardisiertes Protokoll verbindlich, das dabei helfen soll, die Reproduzierbarkeitsrate von psychologischen Effekten zu ermitteln und die Selbstkorrekturmechanismen der Wissenschaft anzuwerfen (Ioannidis 2014). Die Übersetzung der wissenschaftlichen Werte in die wissenschaftliche Praxis birgt jedoch unaufgelöste Spannungen.

Neben der Universalität, Neutralität und Skepsis ist im Ethos der Wissenschaft der Wert des Kommunismus zu nennen (Merton 1975, S.273), der bei der Open Science Collaboration (OSC) für den offenen Zugang zu Methoden, Analysen, Design, Daten und Material steht. Wissenschaftliche Erzeugnisse sind ein öffentliches Gut, das unter kollektiver Anstrengung produziert wurde und somit allen gehören. Damit alle Zugang

zu diesen öffentlichen Gütern haben, muss transparent sein, was wie produziert wurde und wo es zu finden ist. Dass qua Zitierung auch transparent sein soll, wer am Ende der kollektiven Anstrengung die Forschungsfrüchte geerntet hat, evoziert einen Individualismus, der in einem, dem ersten, Spannungsverhältnis zum Kommunismus steht.

Die Skepsis des Reproduzierbarkeitsprojektes drückt sich aus in einer Zurückhaltung gegenüber psychologischen Befunden und den eigenen Resultaten: Eines abschließenden Urteils enthält man sich. Statt zu behaupten, dass kein Effekt existiert, prüfen die Projektteilnehmer, ob eine Studie in der Lage gewesen wäre, den gesuchten Effekt nachzuweisen (Simonsohn 2015). Statt für ihre Resultate den Status eines Nachweises zu reklamieren, werten die Wissenschaftler sie als Denkanstoß und Instrument zur Hypothesenfortbildung in der Erforschung der Faktoren eines Effektes (Nosek 2016). Diese Haltung der Ataraxie steht in einem – zweiten – Spannungsverhältnis zur reformerischen Haltung, die auf eine Reformation wissenschaftlicher Standardverfahren und Anreizsysteme drängt. Die reformerischen Haltung ist alles andere als eine Enthaltung.

In diesem Zusammenhang steht auch das dritte Spannungsverhältnis, das im Wissenschaftssystem besteht zwischen systeminternen, stabilisierenden und systemexternen, verändernden Werten, die die Selbstkorrekturmechanismen des Systems am Laufen halten sollen. Die Motivation zur systematischen Veränderung war bisher nur durch einen Wiedereintritt möglich (Luhmann 1990, S.84 u. 546), dessen Form im System einer autonomen Wissenschaft erst noch mit einer inhaltlichen Rechtfertigung auszufüllen wäre, inwieweit Replikationen beitragen zum wissenschaftlichen Fortschritt. An dieser Stelle sei nur hingewiesen auf die Spannung, die zwischen reformerischem Imperativ und anarchischer Selbstbestimmung besteht.

Die Replikationen schließlich stehen etwas quer im Wertekanon der Open Science Collaboration. Lassen sich sieben der acht Module (OSC 2012) der Transparenz zu- und somit dem Kommunismus unterordnen, so ist die Replikation noch am ehesten der Skepsis zugehörig, soll sie doch – die Selbstkritik überspringend – Gelegenheiten schaffen zur wissenschaftlichen Selbstkorrektur, um fruchtbare Forschungsrichtungen effizient zu identifizieren und das Verständnis von psychologischen Effekten zu verbessern sowie, jetzt wieder anti-skeptisch, unsere Zuversicht bzw. Konfidenz in die Forschungsergebnisse zu steigern.

Das Protokoll der OSC sieht direkte Replikationen vor. Eine möglichst exakte Kopie ist nur machbar, wenn der Methodenteil der Originalstudie hinreichend detailliert ist, und Materialien sowie maschinenlesbare Auswertungsskripte hinterlegt sind. Eine direkte Replikation ist formal ein reiner Reliabilitätstest, der eigentlich keine Generalisierung zulässt (Smith 1970); doch reicht nach Nosek und Lakens (2014) die jeder Studie inhärente Einzigartigkeit für Verallgemeinerungen aus. Im Vergleich mit direkten Replikationen fehlt konzeptuellen Replikationen fehlt die skeptische Schubkraft, die bei Primärforschern Zweifel am Effekt auszulösen vermag. Scheitert die konzeptuelle Replikation eines Sekundärforschers, attribuieren sie das Scheitern einfach den verschiedenen Designs.

Scheitert dagegen eine direkte Replikation, hilft das Scheitern bei der Identifikation von Randbedingungen, die den Effekt moderieren. Ist alles bis auf den Datensatz identisch, kann man sich auf die Suche nach Faktoren machen, die bei der Originalstudie vorgelegen haben, nicht aber bei der Replikation. Des Weiteren kann das Scheitern ein ganzes Forschungsgebiet fragwürdig erscheinen lassen, auf Defekte zentraler Komponenten einer etablierten Theorie hinweisen oder einfach nur verdeutlichen, dass ein Effekt weniger robust ist als gedacht. Das Gelingen dagegen präzisiert die Effektgröße, was das Vertrauen in die Theorie stärkt (Brandt et al. 2014).

Um diese Wirkung auf die Forschergemeinde entfalten zu können, müssen die Replikationen eine Mindestteststärke von 80 Prozent aufweisen. Zudem sieht das Protokoll eine Kontaktaufnahme vor zu den Autoren der Originalstudie, um von ihnen das Ursprungsmaterial sowie die Zustimmung zum Replikationsvorhaben einzuholen. Der im Idealfall autorisierte Replikationsplan wird auf der Internetplattform des Open Science Frameworks vorab registriert, um das Trimmen von Daten zu erschweren, und das Vorhaben von Externen begutachten zu lassen.

Im Open Science Framework werden die Resultate der Replikationsversuche gesammelt, um die Replizierbarkeit von Psychologiestudien bzw. die Reproduzierbarkeit psychologischer Effekte zu schätzen. Erst durch die konzertierte Bündelung standardisierter Replikationsversuche in Replikationsbatterien (Rosenthal 1990) wird eine präzise Schätzung der Reproduktionsrate möglich. Als Prädiktoren gehen die Signifikanz (*p*-Werte), Effektgrößen, Meta-Analysen der Effektgröße und die Expertenansicht zur Replikationswahrscheinlichkeit eines Effekts ein in das Reproduktions-

modell. Dass die Datenanalyse vorgenommen wird mittels derselben Methoden, deren Kritik zur Entstehung der Replikationskrise beigetragen hat, markiert das letzte Spannungsverhältnis.

2.4 Seismik systematischer Replikationen

Huldigten die Forscher bislang in aller Ruhe dem Kult der isolierten Studie (Duke-Elder 1964), indem sie ihre statistischen Analysen stets auf nur einen Datensatz beschränkten, und zwar so, dass sie beiläufig oder versteckt (Fisher 1935; Tukey 1962) andere Studien immer wieder replizierten, ohne die Originaldaten zu berücksichtigen, und ohne der Öffentlichkeit von den gelungenen Replikationen zu berichten (Mulkay 1986; Easley, Madden & Dunn 2000; Jones 2010), so wurde die Wissenschaftsgemeinde jüngst erschüttert von den auf mehreren Datensätzen beruhenden Veröffentlichungen der ersten groß angelegten, systematischen Replikationen von Simons, Alogna, Zwaan et al. (2014) sowie von Klein, Ratliff, Vianello et al. (ManyLabs 2014) und der Open Science Collaboration (2015). Die Erschütterung samt Nachbeben markiert medial den Beginn der Replikationskrise (Tucker 2016).

Während die Forscherteams des ManyLabs-Projektes wenige Studien mehrfach replizierten und eine Erfolgsquote von 77 Prozent verzeichnen konnten, replizierten die Teams des Reproduzierbarkeitsprojektes: Psychologie viele Studien nur einmal und kamen dabei statt der erwarteten 77 Prozent auf nur 36 Prozent erfolgreiche Replikationen. Zudem lagen die Effektgrößen deutlich unter denen der Originalstudien – im Durchschnitt waren sie gerade mal halb so groß. Die davon ausgelösten Schockwelle verlief auf persönlicher und inhaltlicher Ebene; auf der inhaltlichen Ebene spaltete sie sich auf in die Front derjenigen, die aus den Ergebnissen ihre Schlüsse zogen, und derjenigen, die bezweifelten, dass aus den Ergebnissen Schlüsse gezogen werden können.

Die einen schlossen aus den Ergebnissen, dass viele Studien nicht reproduzierbar seien (Bishop 2016) und dass die ganzen wissenschaftlichen Bemühungen letztlich nur heiße Luft (Yong 2016) und von falschen Lehrsätzen durchsetzte Lehrbücher hervorgebracht hätten (Stewart-Wilson 2016). Andere betrachteten die Ergebnisse als integrales Moment wissenschaftlicher Selbstkorrektur, das allzu kühne Hypothesen aus dem Datenpool aussondere; das Aufdecken und Verwerfen falsch-positiver Resultate bedeute

keine Krise, twitterte Pinker am 30. August 2015. Wieder andere hielten die Erfolgsquote wahlweise für zu gering, weil die Replikationsversuche von verzerrten Effektgrößen ausgingen (Etz 2016), oder für zu hoch, weil die Replikationsversuche keine Validitätsprüfung der Originalbefunde beinhalteten (Mayo 2016).

Wieder andere sahen sich in ihrer Einschätzung bestätigt, dass zu viel Wissenschaftsmüll die Forschungslandschaft verschmutze (Gelman 2016) und die phantastischen Behauptungen der Pop-Psychologie (Gelman & Geurts 2016) endlich als solche entlarvt würden. So wuchs die Replikationskrise aus zu einer Glaubwürdigkeitskrise der Psychologie (Horgan 2016) und eskalierte trotz Warnungen davor, Replikationen als persönliche Angriffe zu verstehen (Hamermesh 2007), auf der persönlichen Ebene und nahm teilweise inquisitorische Züge an (Lynch 2015).

Als die Initiatoren der Open Science Collaboration, Nosek und Lakens, 2014 die Aufnahme einer – argumentativ dürftigen – Erwiderung auf eine gescheiterte Replikation in eine Sonderausgabe der Fachzeitschrift *Social Psychology* zu Replikationen ablehnten (Schnall 2014), brach in den sozialen Netzwerken ein Replikationskrieg (Meyer 2014) aus, in dem die Forscher der Replikationsprojekte als geistlose Wissenschaftler ohne eigene Ideen gebrandmarkt oder als Kettenhunde einer Replikationspolizei diffamiert wurden; auf der anderen Seite fehlte es nicht an Häme über Autoren, deren Studien nicht repliziert werden konnten (Donnellan 2013). Weil Psychologen davon überzeugt sind, dass ihre Resultate mit hoher Wahrscheinlichkeit repliziert werden, gehen sie mit gescheiterten Replikationen hart ins Gericht (Kahneman & Tversky 1973).

Wo das Ringen um Reputation in den Vordergrund rückt, steigern erfolgreiche Replikationen zwar das Ansehen der Autoren einer Originalstudie, aber eine gescheiterte Replikation stellt nicht nur die Resultate der Originalstudie infrage, sondern gleich auch sämtliche Resultate all ihrer Veröffentlichungen (Brown 2014), sodass Replikationen auf persönlicher Ebene für Forscher eher bedrohlich wirken. Das führt dazu, dass Forscher Replikationsversuchen gegenüber misstrauisch sind und den Sekundärforschern schädigenden Vorsatz unterstellen (Spellman 2015).

Die Schärfe, mit der die Krisendebatte geführt wird, macht deutlich, dass es sich bei Replikationen nicht nur um eine Frage der Methodik handelt, sondern auch um eine Frage der Dignität der Forscher. Insofern in die Antwort auf die Replikationsfrage Werturteile einfließen, ist keine rein argumentative Auflösung zu erwarten; im Gegenteil, es

können sogar mit demselben Argument konträre Positionen bezogen werden. So verurteilt Fiske (2016) Replikationen aufgrund des von ihnen ausgehenden methodologischen Terrorismus, den Mischel (2005) an derselben Stelle in denselben Worten begrüßt hat.

Dass persönliche Wertvorstellungen wissenschaftlicher Integrität in die Replikationskrise involviert sind und involviert sein müssen, ergibt sich schon daraus, dass die Rolle von Replikationen neben der Rolle von Publikationsprocedere und Verfügbarkeit von Materialien, Daten und Code Bestandteil sind einer umfassenden Reform des Wissenschaftssystems (Greiffenhagen & Reeves 2013). Denn Reformen werden von Forschern angestrengt, nicht vom Erforschten. Und die sind nicht immer einer Meinung. Die Norm einer institutionalisierten Skepsis (Merton 1975, S.277) und die Norm gegenseitigen Vertrauens (Williams 2015) lassen sich zwar unterschiedlichen Ebenen, einer inhaltlichen und einer persönlichen, zuordnen, doch konsistent handlungsleitend kann nur eine von ihnen sein – in der Forschungspraxis sind λόγος und ἔθος untrennbar verbunden.

Was darf man folgern, was muss man fordern? Zwei Fragen, die, aus verschiedenen begrifflichen Sphären kommend, im Verhalten der Akteure konvergieren und angesichts weitreichender Implikationen für Wissenschaft und Gesellschaft zur Zurückhaltung bei ihrer Beantwortung mahnen. Diese – skeptische – Zurückhaltung ist daher angebracht bei den Schlussfolgerungen aus Replikationsprojekten. Dass auf der Grundlage der OSC-Studie nicht endgültig entschieden werden kann, wie hoch die Erfolgsquote von Replikationen insgesamt ist (Gilbert, King, Pettigrew & Wilson 2016), räumt Nosek (2016) freimütig ein. Dort aber hört der Konsens schon auf.

Gilbert et al. (2016) begründen ihre Zurückhaltung mit Schwächen in der Replikationstreue und in der Replikationsmetrik. Die Replikationen der OSC-Studie seien weder hinreichend repräsentativ noch originalgetreu, um als direkte Replikationen durchgehen zu können. Außerdem fehle es ihnen an einem robusten Maß für die Evaluierung der eingetretenen Erfolgsquote. Letztere müsse bereinigt werden um die Basisrate der Replikationsversuche, die rein zufällig gelingen, um stichhaltige Aussagen treffen zu können zur Replizierbarkeit psychologischer Studien.

Der berechtigte Ruf nach einer metrischen Kalibrierung der empirischen Ergebnisse von Replikationsstudien geht etwas unter, weil Gilbert et al. (2016) sich in ihrer Argumenta-

tion eines Maßes bedienen, dessen Schwächen sie nicht nur übersahen, sondern das sie auch noch falsch interpretierten: Zur Bestimmung der Basisrate maßen sie den Erfolg einer Replikation daran, ob ihr Ergebnis in das Konfidenzintervall der Originalstudie fällt oder nicht. Dann aber wären die Replikationsversuche von den Studien am erfolgreichsten, die die geringste Teststärke haben. Denn je kleiner die Stichprobe ist, umso größer fällt das Konfidenzintervall ausfällt; und je größer das Konfidenzintervall ausfällt, umso größer ist die Wahrscheinlichkeit, dass ein replizierter Effekt in ihm liegen wird (Srivastava 2016). Darüber hinaus ist bei einem 95 Prozent-Konfidenzintervall die Erfolgsquote von Replikationen nur dann 95 Prozent, wenn eine Studie unendlich oft und mit demselben Stichprobenumfang aus derselben Population repliziert wird (Nosek, Anderson, Zuni et al. 2016).

Das ist bedauerlich, denn zum Versuch einer metrischen Kalibrierung gehörte beispielsweise eine breitere Erörterung der Rolle von Kontextfaktoren. Bavel, Mende-Siedlecki, Brady und Reinero (2016) vermuten, dass der Replikationserfolg abhängt von der Kontextsensitivität des Forschungsthemas, in das eine Studie eingebettet ist, weil die Replikation kognitionspsychologischer Studien häufiger gelang als die Replikation sozialpsychologischer Studien. Diese augenscheinliche Abhängigkeit ist allerdings zweifelhaft, weil innerhalb der Disziplinen der Replikationserfolg nicht von der Kontextsensitivität abhängt, d.h. stark kontextsensitive Studien der Kognitionspsychologie wurden nicht seltener erfolgreich repliziert als schwach kontextsensitive Studien der Kognitionspsychologie; Gleiches gilt für die Sozialpsychologie (Inbar 2016).

In diesem Zusammenhang ist allerdings zu bedenken, dass die Verwerfung der Nullhypothese umso mehr vom Zufall abhängt, je feiner die Population in Subpopulationen zerlegt wird; wenn man also die Kontextsensitivität separat betrachtet für beispielsweise die Identitätsbildung und Konfliktforschung in der Sozialpsychologie. Mit wachsender Zerlegung verringert sich der Umfang der Teilstichproben und vergrößert sich der Standardfehler, sodass die Wahrscheinlichkeit einer Verwerfung der Nullhypothese gegen Null strebt (Eid, Gollwitzer & Schmitt 2015, S.238). Wie man die Perspektive auch wenden mag, die Krise scheint sich von von jedem Winkelzug nähren zu können, sodass die Replicate-Initiative jeden Moment zur Repligate-Affäre auszuwachsen droht (Mayo 2016).

3 Originalstudie

Bevor Methoden, Design und Resultate der Studie3 aus Weisberg, Taylor und Hopkins 'Deconstructing the seductive Allure of Neuroscience Explanations' aus dem Jahr 2015 dargestellt werden, wird die Hypothese der Studie, dass psychologische Erklärungen besser beurteilt werden, wenn sie irrelevante Neuroinformationen enthalten, eingeführt über Art und Ausbreitung neurologischer Erklärungen und Abbildungen im Rahmen einer thematischen Auseinandersetzung des Verhältnisses von Psychologie und Neurologie.

3.1 Neurologie und Psychologie

War in der Antike das Gehirn dem Herzen noch untergeordnet und für die Klugheit des Menschen nur indirekt mitverantwortlich (Aristoteles II 6.744 b 12), so wurde in der Aufklärung der Zweck des Gehirns einzig im Denken gesehen, wie das Verdauen als Zweck des Magens, die Blutbewegung als Zweck des Herzens oder die Wahrnehmung als Zweck der Sinnesorgane galt (Gall 1791, S.175). In dieser Zeit wurde wie die Welt so auch das Gehirn kartographiert, und bestimmte Fähigkeiten des Menschen eingegrenzt in Hirnarealen, deren Fläche für das Ausmaß dieser Fähigkeiten stand. Von hier nahm der neuro-manische Imperialismus (Tallis 2011, S.73) seinen Ausgang.

Wahrnehmen, Denken, Fühlen gerieten zu Erscheinungsformen des materiellen Gehirns, auf das psychische Phänomene zurückführbar sein müssen. Selbst die Psychoanalyse suchte ihre Wissenschaftlichkeit (Freud 1973, S.19) in Nervenzellen, die psychische Energie leiten (Freud 1987, S.391), diagnostizierte später aber zwischen erregten Nervenzellen und seelischen Vorgängen eine Lücke, deren Schließen nicht Aufgabe der Psychologie sei (Freud 2016, S.13). Auch die Experimentelle Psychologie kam zu dem Schluss, dass selbst ein vollkommenes Verständnis der molekularen Vorgänge des Nervensystems nichts zur Erklärung beizutragen vermag, weshalb ein neuraler Erregungszustand begleitet wird von einer bestimmten Erfahrung (Münsterberg 1891, S.26); vielmehr hänge die Interpretation neurophysiologischer Befunde von psychologischen Erkenntnissen ab und nicht umgekehrt (Wundt 1914, S.197).

Die gegenläufigen Strömungen einer reduktionistischen Einheit der Wissenschaft (Skinner 1975, S.59) und einer pluralistischen Wissenschaftsdifferenzierung fanden schließlich in der potentiellen Reduzierbarkeit auf materielle Substrate ein Auffangbecken, das die fortschreitende Spezialisierung in immer neue Disziplinen öffentlich legitimiert, obwohl es keine belastbaren Hinweise gibt auf Zusammenhänge zwischen psychischen und neuralen Zuständen (Fodor 1974); vielmehr entfernen sich Psychologie und Neurologie immer weiter voneinander (LeMoal & Swendsen 2015). Die eine stürzt ab in die Krise, die andere schreitet fort von Triumph zu Triumph.

Für Gehirnforscher, deren Forschungsgegenstand als wichtigstes Organ des Menschen gehandelt wird, ist die Erntezeit angebrochen (Spitzer 2004, S.230). Das Etikett 'Neuro' zeichnet jetzt ausgereifte Wissenschaft aus. Als ausgereift gilt eine Wissenschaft, wenn sie eintritt in das Stadium ihrer technischen Verwertbarkeit (Böhme, Daele & Krohn 1973). Und verwertbar ist Technik, weil sie teleologisch einen Nutzen bietet. Den wiederum liefere die Neurologie derart mit, dass sich aus ihren Befunden ableiten lasse, was man unterlassen muss, um das Gehirn, das sich durch seinen Gebrauch verändert, nicht zu schädigen, oder was man tun muss, um die zerebrale Leistung zu maximieren (Hüther 2001). Zur Begründung dafür werden somatische Marker (Adophs, Tranel, Bechera & Damasio 1996) als Spuren vorgebracht, die Gedanken im Gehirn hinterlassen, Spuren, die man sichtbar machen kann (Spitzer 2012, S.18).

So werden wir zu Augenzeugen gedachter Gedanken und gemachter Erfahrungen. Bildgebende Verfahren wie das der funktionellen Magnetresonanztomographie erzeugen digitale Bilder, die aus der magnetischen Flussdichte des Hämoglobins berechnet werden und wie eine (Falschfarben-)Photographie des Gehirns aussehen. Sichtbar sind in diesen Bildern nicht Gedanken, sondern relative Unterschiede der Durchblutung in einem Areal von der Größe eines Pixels. Je nachdem, wie signifikant sich die Durchblutung der Areale unterscheidet, werden die Pixel eingefärbt. Die Analysemethoden der Tomographie sind also dieselben wie die der Psychologie, mit denselben Folgen, nämlich einer hohen Rate falsch-positiver Hirnaktivitäten (Eklund, Nichols & Knutsson 2016), aber dem Unterschied, dass sich die Auswertung des Tomographen zu einem bunten Relief des Gehirns materialisiert.

Die Bilder sind ein materielles Vehikel, das unter Experten als zeitweiliger Bedeutungsträger in der diskursiven Meinungsbildung zirkuliert, die Fachkreise aber auch verlassen

und in der Öffentlichkeit meinungsbildend wirken kann (Latour 1990). Tomographische Aufnahmen des Gehirns besitzen eine hohe Mobilität in der Gesellschaft, wo sie ihre symbolische Wirkung eine zeitlang unverändert entfalten. Diesen Symbobilen verdankt die Neurologie weit mehr ihren rasanten Fortschritt als neuen Erkenntnissen: ihr Fortschritt ist im wesentlichen ein technologischer (LeMoal & Swendsen 2015). So hat sich zwischen 2002 und 2012 die Anzahl von Veröffentlichungen auf der Grundlage von tomographischen Aufnahmen des Gehirns verdreifacht auf über 9 000 Publikationen im Jahr (Ioannidis 2014), ohne dass die Datenanalysen reliabel (Bennett & Miller 2010) wären oder dazu geeignet, die Funktionsweise eines 30 Jahre alten Mikroprozessors zu rekonstruieren, obwohl der Chip über eine einfache Architektur verfügt, Transistoren sich leicht manipulieren lassen und Übergänge von aktiven und inaktiven Elementen oder lokalen Feldern analog zur Oxygenierung des Blutes gemessen werden können (Jonas & Körding 2016).

Versuche, psychologische Artikel mit Aufnahmen vom Gehirn wissenschaftlich aufzuwerten, verzeichneten erste Erfolge (McCabe & Castel 2008), die jedoch nicht repliziert werden konnten (Michael et al. 2013). Keehner, Mayberry und Fischer (2011) fanden den Effekt moderiert von der Räumlichkeit: je räumlicher das Gehirn dargestellt wird, desto größer sei dessen Überzeugungskraft. Dagegen räumten Farah und Hook (2013) ein, dass der Informationsgehalt zwar von der Darstellung des Gehirns abhänge, sie beeinflusse aber die Bewertung eines wissenschaftlichen Artikels nicht. Dies bestätigten weitere Versuche von Hook und Farah (2013), die zudem reduktionistische Überzeugungen, im Gegensatz zu Hopkins, Weisberg & Taylor (2016), als Moderator ausschließen.

Somit scheint die Wirkung der Symbobile aus aus den Neurologie-Laboratorien geringer als gedacht, unabhängig von der Art der Darstellung des Gehirns (Gruber & Dickerson 2012). Die Radiation der Hirnforschung vermögen sie nicht zu erklären. Statt visueller könnte die Radiation aber terminologische Ursachen haben, die in der Erklärung selbst liegen. Das wäre dann bedenklich, wenn schlechte Erklärungen durch einen konstruierten Bezug zum Gehirn plötzlich überzeugend würden. Diese Bedenken muss man in der Neurologie, wo der Bezug zum Gehirn unvermeidlich ist, nicht haben. Hier kann man ohne Einbußen in der Überzeugungskraft argumentieren, dass ein

visualisiertes Gehirn deshalb überzeugend ist, weil das Gehirn wesentlich mit der Verarbeitung visueller Reize betraut ist (Umiltà 2008).

Weisberg, Keil, Goodstein, Rawson und Gray (2008) halten das für eine schlechte Erklärung, weil sie zirkulär ist. Denn die Bedeutsamkeit von Hirnscans wird hier mit Hirnscans begründet: Anhand visueller Eindrücke vom Gehirn konstruieren Neurologen ein Modell zerebraler Funktionalität, zu der auch das Sehen gehört. Durch die modellierte Lokalisierung im Gehirn wird das Sehen selbst potenziell sichtbar. Vor dem Hintergrund dieses Modells werden die Gehirnareale technisch visualisiert. Die Visualisierung zeigt, dass das Gehirn hauptsächlich mit der Verarbeitung der Gesichtswahrnehmung beschäftigt ist. Die hohe Hirnaktivität beim Sehen wird schließlich gleichgesetzt mit einem hohen Anteil des Sehens an Überzeugungen. Mit anderen Worten: Weil die Hirnscans überzeugend sind, gelangen wir – sehenden Auges – zu der Überzeugung, dass Hirnscans überzeugend sind. Visuelles (*Cerebrum videns*) und visualisiertes Gehirn (*Cerebrum visibilis*) stehen damit in einem engen Begründungszusammenhang.

Für den Umstand, dass immer wieder eine zertifizierende Überlegenheit von Hirnscans festgestellt wird gegenüber psychologischen Erklärungen, wie beispielsweise beim Test auf Demenz (Munro & Munro 2014), ist eine Vielzahl von Hypothesen vorgeschlagen worden, die mit sich verjüngendem Radius vom kulturellen Ganzen kommend im Anschluss an soziale Zusammenhänge psychologische Gründe einkreisen. Von der gesellschaftlichen Totalen verengt sich der Fokus zum Individuum und seinen Kognitionen; eine weitere Verengung auf neurologische Vorgänge gibt die Auflösung der Hypothesen nicht her, das wäre ja zirkulär.

Es sind jedenfalls nicht nur Laien, die neurologische Symbobile unkritisch hochhalten (Weisberg et al. 2008), auch kritische Forschungsgesellschaften haben ihre Fördertöpfe für die Nervenwissenschaften erheblich vergrößert (Hasler 2013, S.30). Die Förderprogramme wirken kulturverstärkend, indem sie auf die gegenwärtige Nachfrage nach Wissenschaft reagieren und das künftige Angebot an Wissenschaft steuern. Und sie verstärken eine Kultur, in der Gesundheit als speicherbare Ressource abgerufen werden kann zur Steigerung der Produktivität (Thornton 2011, S.17). Unter dem Dach der Supervenienz, die den Widerspruch zwischen Reduktionismus und ontologischem Pluralismus aufhebt (Davidson 1980, S.214), verkörpert die Plastizität des Gehirns das

Versprechen, Defizite einem natürlichen Mechanismus zuschreiben zu können und gleichzeitig die Kapazität vorzuhalten zur Optimierung eben dieses Mechanismus. Das Individuum wird begrenzt durch etwas, für dessen Grenzziehung es die Verantwortung trägt: es erfährt und generiert seine Gesundheit – Illustrierte konsultierend, die zu gesundheitsrelevanten Themen unkritisch über das Gehirn berichten (Ramani 2009).

So gerne man auf die gelenkte Meinungsbildung durch Massenmedien rekurriert, auch bezüglich der medialen Aufmerksamkeit für das Gehirn (Weisberg 2008), so leicht wird sie überschätzt. Die Wirkung der Medien auf die Meinung ihrer Rezipienten ist äußerst gering (Früh 1991, S.220); einen viel stärkeren Einfluss haben Menschen aus dem persönlichen Umkreis und sogenannte Meinungsführer (Schenk 2002, S.341). Massenmedium und Rezipient stehen in einer gegenseitig gekoppelten Wirkungsbeziehung, die als Transaktion bezeichnet wird (Früh 1991, S. 16). In anderen Worten: die Medien berichten größtenteils, was die Rezipienten rezipieren möchten, und was die Rezipienten rezipieren möchten, das entnehmen sie größtenteils den Medien. Wer also im Rückgriff auf die Medien die Popularität der Nervenwissenschaften begründen möchte, wird diesem Zirkel kaum entrinnen.

Als Meinungsführer gelten Wissenschaftler, die paradigmatisch mit Messungen und Berechnungen universale Mechanismen erforschen und diese in hieroglyphischen Formeln ausdrücken (Sperber 2010; Eriksson 2012). In diesem Paradigma der Mechanik wird das Gehirn metaphorisch vorgestellt als Motor der Seele (Fernandez-Duque, Evans, Colton & Hodges 2015), dessen Wartung technisch so komplex und teuer ist, wie nur eine Premiumwissenschaft komplex und teuer sein kann, und dennoch auf einfachen elektrochemischen Kausalzusammenhängen basiert (Legrenzi & Umiltà 2009, S.104), die einen direkten Zugriff auf psychische Zusammenhänge suggerieren (Eysenck & Keane 2015, S.617). In seiner Materialität verbürgt der zerebrale Motor einen hohen Realitätsbezug (McCabe 2008), und er verrät mit der Antriebstechnik des Seelenlebens seine reduktionistischen Wurzeln, die, wie tief auch immer, bis in die öffentliche Meinung reichen (Legrenzi & Umiltà 2009, S.12).

Begreift man die öffentliche Meinung als Resultierende aus individuellen Überzeugungen, könnte die Überzeugungskraft der Neurologie mediiert werden von der Einstellung des Einzelnen. Menschen mit einer gefestigten Einstellung lassen sich schwerer überzeugen: widerstrebt der Einstellung einer Person ein neurologisch eingefärbtes Argu-

ment, wirkt es weniger überzeugend als dasselbe Argument bei entgegengesetzter Einstellung (Scurich & Shniderman 2014), weil dann das Argument keine kognitive Dissonanz (Festinger 2001, S.247) mehr hervorrufen kann.

Beim Einzelnen verbleibt nach der Konfrontation mit einem Gehirn unter Umständen ein Gefühl des mehr oder weniger Überzeugtseins, das den Grad seines Verständnisses anzeigt (Trout 2002). Oder aber es kommt beim Einzelnen erst gar nicht zum Verständnis, weil er – wie alle Menschen – Gefühle schneller verarbeitet, als er denken kann (Zajonc 1980). In beiden Fällen wird gefühlsmäßig entweder der neurologischen oder der psychologischen Evidenz der Vorzug gegeben. Bei solchen Bauchentscheidungen (Gigerenzer 2008, S.12) spielt es fast keine Rolle, ob die Evidenz explanativer oder narrativer Art ist, weil sowohl Erzählungen (Dekker, Lee & Jolles 2014) als auch Erklärungen täuschen und den falschen Anschein wissenschaftlicher Beschlagenheit erwecken können. Trout (2008) nennt letztere placebische Erklärungen, weil sie nur die leere Hülle eines guten Gefühls vermitteln.

Eine letzte Hypothese sieht den Grund, weshalb es bei der Beurteilung möglicherweise gar nicht erst zu einem Verständnis kommt, in vom inhaltlichen Kern ablenkenden neuronalen Details, die so verführerisch sind, dass man ihnen nachgeht, statt dem eigentlichen Argument zu folgen. Die Ablenkung durch Details (Garner, Gillingham, Kulikovich & White 1989) ist meta-analytisch bestätigt (Rey 2012) und prima facie die robusteste der vorgetragenen Hypothesen; die Ablenkung durch Details führt denn auch die hier zu replizierende Originalstudie im Titel, so zwar, dass deren Dekonstruktion angekündigt wird.

3.2 Dekonstruktion der Verführungskraft neurologischer Erklärungen

Der Titel lässt, wenn nicht eine Rehabilitierung der Psychologie gegenüber der Neurologie, so doch eine Renormierung der Überzeugungskraft neurologischer Argumente erwarten. Ohne gleich in den antiken Modus des Gehirns als einem Kühlaggregat zurückzufallen, scheinen Weisberg, Taylor und Hopkins (2015) entschlossen, aus dem Schatten der Neurologie herauszutreten und verlorenes Vertrauen in die Psychologie zurückzugewinnen – genau wie das Reproduzierbarkeitsprojekt: Psychologie. Idealerweise gleich mit eigenem Finanzindex wie dem Nasdaq Neuro Insights Neuro Tech Index

NERV. Dann könnte man auch an eine Fusion avisieren. Denn die Psychologie wird laut Yurevich (2008) große Fortschritte machen, aber eben nur als Neuropsychologie.

In die von ihnen katalysierte Diskussion um die Wirkung neurologischer Komponenten in einer psychologischen Erklärung brachten Weisberg et al. (2008) ein verführerisches Detail ein, das es ihnen nicht nur erlaubte, gute Erklärungen von schlechten zu unterscheiden, sie konnten zudem die Güte von Erklärungen als unabhängige Variable manipulieren, um den Beitrag irrelevanter Neuroinformation zur Beurteilung einer Erklärung experimentell zu testen: Irrelevant ist eine Information dann, wenn sie die der Erklärung zugrundeliegende Logik nicht berührt; und die Erklärung gilt als schlecht, wenn sie einen logischen Zirkel beinhaltet. In dem so konzipierten Design bekamen Erklärungen ohne Neuroinformation schlechtere Noten als Erklärungen mit irrelevanter Neuroinformation – irrelevante Neuroinformation kompensierte die mangelhafte Qualität einer Erklärung.

Auf der Suche nach Gründen für diese Verzerrung veröffentlichten Weisberg et al. (2015) drei weitere Studien, die die Länge, Qualität und neurologischen Fachjargon als Moderatoren zum Gegenstand hatten und allesamt die Hypothese bestätigten, dass psychologische Erklärungen mit irrelevanter Neuroinformation für überzeugender gehalten werden als solche ohne, wobei Länge und Qualität einer Erklärung je für sich die Beurteilung der Erklärung signifikant verbesserten, nicht aber Fachtermini, die den technologischen Kontext der Hirnforschung zum Ausdruck brachten, wie beispielsweise fMRT-Scans der Cortex anstelle von bloßen Aufnahmen der Hirnrinde.

In der hier zu replizierenden dritten Studie bekamen die Teilnehmer zur Beurteilung auf einer Likert-Skala von −3 bis +3 online vier Phänomene der Psychologie samt Erklärung vorgelegt: zur Rechenkompetenz von Säuglingen, zum Aufmerksamkeitsblinzeln, zum räumlichen Denken und zum Sehen und Vorstellen. In der Mitte erfolgte ein Aufmerksamkeitstest getarnt in Form einer Beschreibung des Phänomens der sozialen Attribution eigener Kenntnisse, zu der die Teilnehmer statt einer Erklärung die Aufforderung zur Abgabe der Höchstbewertung erhielten. Auch zu diesem Phänomen konnten sie ihre Entscheidung im Freitext begründen, sodass bei abweichenden Beurteilungen ermittelt werden konnte, ob jemand Text und Erklärung aufmerksam gelesen hatte.

Die Teilnehmer wurden zufällig der Versuchsgruppe mit Neuroinformation und der Kontrollgruppe ohne Neuroinformation zugeteilt. Der Zufall entschied auch für jedes Phänomen, ob die zugehörige Erklärung – mit oder ohne Neuroinformation – gut oder schlecht ausfiel. Die zirkulär titulierten Erklärungen bestanden in einer bloßen Paraphrase der Explananda und erklären somit nichts. Beispielsweise soll wegen des späteren Eintreffens eines Ereignisses die zeitliche Beziehung zwischen beiden Ereignissen das Aufmerksamkeitsblinzeln 'erklären'. In der Versuchsgruppe rückte, damit alle Antworten dieselbe Länge einhielten, an die (Text-)Stelle der Autorität von Forschern ein neurologisches Testat: Dann sollen etwa Aufnahmen der Gehirnregion, die am räumlichen Denken beteiligt ist, zeigen, dass der Geschlechterunterschied durch das schwache Abschneiden der Frauen 'erklärt' wird.

Die Stichprobe entstammte Studierenden ohne ersten Abschluss und Crowdworkern (Mechanical Turks), die gegen Entgelt Aufgaben erledigen. Daraus ergab sich ein 2 (Sample: Studierende, MTurks) x 3 (Neuroinformation: ohne, mit, Jargon) x 2 (Qualität: gut, schlecht) Design. Demographisch wurden von den Teilnehmern Geschlecht, Alter und höchster Bildungsabschluss erhoben.

Weisberg et al. (2015) werteten die Daten aus mittels Regressionsanalyse im Gemischten Modell mit Zufallsachsenabschnitt und -steigung der Qualität von Erklärungen bezogen auf jeden einzelnen Teilnehmer, wobei für die Stufen der unabhängigen Variablen 'Neuroinformation', 'Sample' und 'Qualität' Dummy-Variablen verwendet wurden: der Achsenabschnitt stehe für die studentische Beurteilung schlechter Erklärungen ohne Neuroinformation. Die Phänomene bzw. Items dagegen waren effektcodiert mit der Rechenkompetenz von Säuglingen als Referenzkategorie. Wie beides zusammen konsistent zu interpretieren ist, wird uns im nächsten Abschnitt beschäftigen. Auf die Regressionsanalyse folgte schließlich eine qualitative Textanalyse der schriftlichen Begründungen für die jeweilige Beurteilung der Probanden, die varianzanalytisch abgeschlossen wurde.

Unterschiede in den Beurteilungen der Items betreffend kamen die Autoren zu folgenden Ergebnissen: kein Effekt des Geschlechts; Haupteffekte für Sample, Qualität und Neuroinformation; kein über die Neuroinformation hinausgehender Effekt für Jargon; Interaktionseffekte zwischen Items und Neuroinformation bzw. Qualität mit zwei Ausnahmen: keine Interaktion zwischen Item 2 (Aufmerksamkeitsblinzeln) und

Neuroinformation sowie zwischen Item 4 (Sehen und Vorstellen) und Qualität. Von den abgegebenen Begründungen bezogen sich 24 Prozent auf das Gehirn, 58 Prozent davon in einem positiven Sinn; Studierende und Crowdworker unterschieden sich nicht in der relativen Häufigkeit positiver Bewertungen in ihren Begründungen.

Mit 687 Zitierungen binnen eines Jahres zählt die Dekonstruktion der Verführungskraft neurologischer Erklärungen von Weisberg et al. (2015) zu den einflussreichen Veröffentlichungen der Psychologie. Die Replikation dieser bedeutsamen Arbeit verfolgt zwei Ziele: erstens die Bereinigung der Effektgröße um die Veröffentlichungsverzerrung, und zweitens soll exemplarisch die Bedeutung von Replikationen für die Geltung wissenschaftlicher Veröffentlichungen herausgearbeitet werden.

4 Die Replikation

Zuerst wird im Folgenden das Neuro-Effekt-Modell auf der Grundlage des Textes von Weisberg et al. (2015) rekonstruiert und dann reanalysiert. Die ausführliche Darstellung der Wege zur Berechnung des für die Replikation erforderlichen Stichprobenumfangs aus Effektgröße, Teststärke und Präzisionsgrad illustriert die zahllosen Freiheitsgrade und Fehlerquellen, die eine Replikation bietet, und die zu sehr verschiedenen Ergebnissen führen. Schließlich werden Ablauf und Ergebnisse einer nach gängigen Maßstäben erfolgreichen direkten Replikation geschildert. Dass die Replikation dennoch kein Erfolg ist, zeigt sich an der unreflektierten Operationalisierung hochwertiger Erklärungen. Weil sowohl das Konstrukt der Qualität als auch das Konstrukt der irrelevanten Neuroinformation ambivalent bleibt, ist bei der Interpretation der Replikationsergebnisse Zurückhaltung geboten.

4.1 Rekonstruktion

Zur Rekonstruktion findet sich im Text nur der Verweis auf ein Gemischtes Modell. Gemischte Modelle zeichnen sich aus durch eine Mehrebenenstruktur mit gemischten Effekten, die so heißen, weil sie einen festen und einen zufälligen Anteil besitzen. Während der Festanteil den Effekt auf der Beobachtungsebene verkörpert, sammelt der Zufallsanteil auf allen Ebenen deren Beiträge zum Effekt ein. Es wird im Gemischten Modell also die Abhängigkeit eines Effektes von verschiedenen Ebenen modelliert. Treffender wäre es daher, von unabhängigen und abhängigen Anteilen an einem Effekt zu sprechen, statt von festen und zufälligen Anteilen.

Gemischte Effekte kommen dadurch zustande, dass die Ebenen der Einflussfaktoren auf einen Effekt ineinander eingebettet sind, wie beispielsweise Individuen eingebettet sind in Berufsgruppen. Auch Messwiederholungen an derselben Versuchsperson können als Einbettung interpretiert werden, dergestalt, dass die Items eingebettet sind in jeweils einen Studienteilnehmer. Faktoren, die bei den Teilnehmern in Abhängigkeit vom Messzeitpunkt untersucht werden, im Unterschied zu Faktoren, die zeitunabhängig die Teilnehmer in Gruppen einteilen, werden Innersubjektfaktoren genannt. Durch ihre Randomisierung zu jedem Zeitpunkt kann nur die Qualität, im Unterschied zu Neuroinformation und Sample, Innersubjektfaktor sein.

4.1.1 Modell

In die Ebenen des Gemischten Modells kann man durch Gruppen eine monotone und durch Cluster eine relative Hierarchie interpretieren. Im Sinne einer monoton wachsenden Mengenfolge (Elstrodt 2009, S.23) erfolgt bei der Gruppenbildung eine intensionale Erweiterung einer bereits vollständig zerlegten Menge anhand der Merkmalsstufen eines Faktors. Somit konstituiert jeder Faktor eine neue Ebene, auf der mindestens doppelt so viele Gruppen angesiedelt sind wie auf der Ebene darunter. Bei der Clusterbildung dagegen wird anhand der Merkmalsstufen eines jeden Faktors die Grundmenge neu zerlegt – die Merkmale anderer Faktoren bleiben unberücksichtigt.

Im folgenden wird das Hierarchische Modell mangels Hinweisen im Artikel nach dem Gruppenkonzept entwickelt. Dass Weisberg et al. (2015) gar kein Modell entwickelt haben, wird sich erst in der Reanalyse herausstellen. Für die formale Entwicklung des Modells spielt die Interpretation eine untergeordnete Rolle. Bei der Anwendung des formalisierten Modells zum Nachweis des Effekts wird dagegen wegen der verschiedenen Freiheitsgrade des Modells unter einer Interpretation auf das Clusterkonzept zurückzukommen sein.

Aus dem Faktor 'Item' und den Kovariaten 'Qualität', 'Neuroinformation' und 'Sample' lassen sich aus Itemausprägungen und Samplekategorie vier hierarchisch eingebettete Gruppen bilden.

Es ergeben sich vier Ebenen:

Ebene 1: Individuen mit vier Items
Ebene 2: Gruppen mit Items in derselben Qualität
Ebene 3: Gruppen mit identischen Items (in derselben Qualität und Neuroausprägung)
Ebene 4: Gruppen der Studierenden bzw. Crowdworker mit identischen Items.

Eine multiple Regression im Linearen Modell bildet bei einer Versuchsperson *i* den Zusammenhang zwischen *t=1,…,4* Items und den Faktoren seiner Beurteilung folgendermaßen ab:

$$Y_i^t = \beta_0 + \beta_1 \cdot Item_i^t + \beta_2 \cdot Neuro_i + \beta_3 \cdot Quality_i + \beta_4 \cdot Sample_i + \beta_5 \cdot Item_i^t \cdot Neuro_i + \beta_6 \cdot Item_i^t \cdot Quality_i + \varepsilon_i$$

Im Linearen Modell wird vorausgesetzt, dass die Werte, die in die Regression eingehen, voneinander unabhängig sind.

Bei einer multiplen Regression im Hierarchischen Modell werden zufällige Schwankungen mit berücksichtigt sowohl auf dem Achsenabschnitt als auch in der Steigung auf allen Ebenen, in die verschiedene Gruppen eingebettet sind. Dadurch gehen Abhängigkeiten, die möglicherweise zwischen den Ebenen bestehen, in die Berechnung des Zusammenhangs zwischen abhängiger Variable und Prädiktoren mit ein. So hängt die Beurteilung einer Erklärung möglicherweise nicht nur ab von ihrer Qualität, sondern auch vom Item, in das die Qualität eingebettet ist, mit der Tendenz, dass die Versuchspersonen im Durchschnitt ein bestimmtes Item besser beurteilen als die übrigen drei Items. Somit entzerrt die separate Anpassung der Prädiktorenbeiträge auf jeder Ebene die Schätzung der Mittelwertsunterschiede, die dann nicht repliziert werden können, wenn deren Abhängigkeit zu übergeordneten Ebenen nicht berücksichtigt ist.

Weil im Versuchsdesign eine Person jeweils vier Items bewertet, können die Itemwerte selbst nicht als unabhängig voneinander betrachtet werden – sie stehen in Bezug zur jeweiligen Versuchsperson und sind an sie auf Ebene 1 in Form eines Quadrupels (Item1, Item2, Item3, Item4) gebunden.

4.1.2 Einbettung

Ebene 1: Die Anzahl der Individuen i=1,…,239 in der gesamten Stichprobe, die t=1, …,4 Items beurteilen.

Ebene 2: Die Anzahl $j=1,\ldots,2^4$ von Gruppen von n_j Versuchspersonen mit jeweils denselben Itemausprägungen bezüglich der Qualität:

$n_1 = 0$	$n_2 = 16$	$n_3 = 14$	$n_4 = 14$
$n_5 = 19$	$n_6 = 16$	$n_7 = 25$	$n_8 = 23$
$n_9 = 12$	$n_{10} = 18$	$n_{11} = 13$	$n_{12} = 20$
$n_{13} = 22$	$n_{14} = 12$	$n_{15} = 15$	$n_{16} = 0$

Ebene 3: Die Anzahl $k = 1,\ldots,2 \cdot j$ von Gruppen von n_k Versuchspersonen mit denselben Itemausprägungen bezüglich Qualität und Neuroinformation:

$n_1 = 0$ …

…

$n_{29} = 5$ $n_{30} = 10$ $n_{31} = 0$ $n_{32} = 0$

Ebene 4: Anzahl $l = 1,\ldots,2 \cdot k$ von Gruppen von n_l Versuchspersonen mit denselben Itemausprägungen bezüglich der Qualität und Neuroinformation im selben Sample:

$n_1 = 0$ …

…

$n_{61} =$ $n_{62} =$ $n_{63} = 0$ $n_{64} = 0$

4.1.3 Formalisierung der Linearen Regression im Hierarchischen Modell

Ebene 1: Die Items *t* werden von den Versuchspersonen *i* verschieden bewertet:

$$Y^t_{ijkl} = \beta_{0\,jkl} + \beta^t_{1\,jkl} \cdot Item^t_{ijkl} + \epsilon_{ijkl}$$

Der Achsenabschnitt $\beta_{0\,jkl}$ gibt den Durchschnittswert an, der über alle Versuchspersonen und Items hinweg fest, also unabhängig ist und sich zusammensetzt aus der Qualität, der Neuroinformation und dem Sample (Populationsmittelwert über alle Gruppen). Der Koeffizient $\beta_{1\,jkl}$ gibt die Änderung der Bewertung nach Qualität, Neuroinformation und Sample bei den Items über alle Versuchspersonen hinweg an. Schließlich steht ε_{ijkl} für individuelle itembezogene Zufallsschwankungen unter den verschiedenen Bedingungen.

Ebene 2: Die Items werden von den Versuchspersonen je nach Qualität ihrer Erklärung verschieden bewertet. Der zusätzliche Effekt der Qualität auf die Beurteilung, d.h. von Versuchsperson *i* mit Qualitätsausprägung *j*, ist modelliert als Einfluss, der mit zufälligen, also von den Ebenen abhängigen Schwankungen sowohl in den Achsenabschnitt als auch in den Steigungskoeffizienten der Grundgleichung eingeht und mit den Items auf der ersten Ebene interagiert.

$\beta_{0\,jkl} = \beta_{00\,kl} + \beta_{01} \cdot \mathit{Qualität}_{0\,jkl} + \upsilon_{0\,jkl}$ und

$\beta^{t}_{1\,jkl} = \beta^{t}_{10\,k} + \beta^{t}_{11} \cdot \mathit{Qualität}_{0\,jkl} + \upsilon_{1\,jkl}$.

Dabei gibt β_{00kl} den Beitrag der Qualität zum Gesamtdurchschnitt der Itembewertungen wieder, der sich nach Neuroinformation und Sample differenzieren lässt. Dagegen symbolisieren β_{01} und β^{t}_{11} , wie stark die unterschiedliche Bewertung einzelner Items von der Qualität ihrer Erklärung abhängt. Der Koeffizient $\beta^{t}_{10\,k}$ des Interaktionsterms wiederum steht für den festen Anteil der Qualität an den Schwankungen der Itembewertungen, unabhängig von Neuroinformation und Sample.

Der zufällige Einfluss der Qualität auf die individuelle Beurteilung (Innersubjektfaktor), d.h. wie hoch die Personen ansetzen bei der Beurteilung der einzelnen Items unter Berücksichtigung der hierarchischen Konstellation, wird durch $\upsilon_{0\,jkl}$ wiedergegeben. Der Einfluss schwankt zufällig oder ebenenbedingt mit $\upsilon_{1\,jkl}$ je Gruppe der Versuchspersonen, deren Items dieselbe Qualität haben.

Die Kovariate 'Qualität' dient der Erklärung der Variationen zwischen den individuellen Itembewertungen. Die Varianz für die Residuen misst die Variation auf der ersten Ebene, also das Ausmaß individueller Unterschiede innerhalb der 16 Gruppen mit Items derselben Qualität bezüglich der Bewertung der Items. Die Varianz für den Achsenabschnitt misst die Variation auf der zweiten Ebene, also das Ausmaß des festen Anteils unterschiedlicher Bewertungen zwischen den Gruppen mit Items derselben Qualität. Bei der Interaktion von Item und Qualität mit *Item* = −0.39 und *Qualität* = 0.57 verbessert sich beispielsweise die Beurteilung nur um 0.18, wenn man die Qualität erhöht, also die schlechte Erklärung durch die gute ersetzt.

Ebene 3: Die Items mit bestimmter Qualität werden von den Versuchspersonen je nach Neuroinformation verschieden bewertet. Der zusätzlichen Effekt von Neuroinformation auf die Beurteilung; d.h. von Person *i* mit Qualitätsausprägung *j* und Neuroausprägung *k*, ist modelliert als Einfluss, der mit den Items auf der ersten Ebene interagiert und auf die Größe der Qualitätsänderung mit einem festen β^{t}_{1000} eingeht:

$\beta_{00\,kl} = \beta_{000\,l} + \beta_{001} \cdot \mathit{Neuro}_{00\,kl}$ und

$\beta^{t}_{10\,k} = \beta^{t}_{1000} + \beta^{t}_{111} \cdot \mathit{Neuro}_{00\,k}$.

Die Neuroinformation soll die Variation in der Bewertung von Items derselben Qualität erklären, also die Variation zwischen den Gruppen. Die feste Abweichung von der mittleren Bewertung aller Gruppen derselben Itemqualität aufgrund der Neuroinformation wird mit β_{000l} angegeben. Die Koeffizienten β_{001} und β^{t}_{111} wiederum sind die Gewichte, die angeben, wie sehr sich die mittlere Bewertung der Gruppen derselben Itemqualität ändert bei Änderung der Neuroinformation bzw. ihrer Interaktion mit den einzelnen Items, vorausgesetzt alle übrigen Variablen werden konstant gehalten.

Der Effekt der Neuroinformation wird als fest angenommen. Daher unterliegt der Einfluss der Neuroinformation weder auf dem Achsenabschnitt noch im Steigungskoeffizienten zufälligen Schwankungen.

Die Varianz für die Residuen misst die Variation auf der ersten Ebene, also das Ausmaß individueller Unterschiede innerhalb der 32 Gruppen mit identischen Items bei der Bewertung der Items. Die Varianz für den Achsenabschnitt misst die Variation auf der dritten Ebene, also das Ausmaß des festen Anteils unterschiedlicher Bewertungen zwischen den Gruppen mit identischen Items.

Ebene 4: Schließlich werden auch identische Items von den Versuchspersonen verschieden bewertet, je nachdem, ob Studierende oder Crowdworker die Bewertung vornehmen. Der zusätzliche Effekt der Stichprobenzugehörigkeit auf die Beurteilung, d.h. von Person *i* mit Qualitätsausprägung *j* und Neuroausprägung *k* aus Sample *l* wird ebenfalls als fest und eigenständig angenommen, ohne mit anderen Faktoren zu interagieren:

$$\beta_{000l} = \beta_{0000} + \beta_{0001} \cdot Sample_{000l} \quad .$$

Das Sample soll die Variation in der Bewertung identischer Items erklären, also die Variation zwischen den Gruppen mit identischen Items. Der Wert für die konstante Abweichung von der mittleren Bewertung aller Gruppen aufgrund der Samplezugehörigkeit wird mit β_{0000} angegeben und die Gewichtung in der Änderung der Neuroinformation und deren Interaktion mit einzelnen Items wird mit dem Koeffizienten β_{0001} modelliert.

Die Varianz für die Residuen misst die Variation auf der ersten Ebene, also das Ausmaß individueller Unterschiede innerhalb der 64 nach Sample differenzierten

Gruppen mit identischen Items bei der Bewertung der Items. Die Varianz für die Konstante misst die Variation auf der vierten Ebene, also das Ausmaß des festen Anteils unterschiedlicher Bewertungen zwischen den nach Sample differenzierten Gruppen mit identischen Items.

Die sich ergebende Regressionsgleichung kann schließlich aufgeteilt werden in einen festen Teil und in einen zufällig variierenden Teil:

$Y^t_{ijkl} = C^t_{ijkl} + D^t_{ijkl}$, wobei

$$C^t_{ijkl} = \beta_{0000} + \beta^t_{1000} \cdot Item^t_{ijkl} + \beta_{01} \cdot Qualität_{0\,jkl} + \beta_{001} \cdot Neuro_{00\,kl} + \beta_{0001} \cdot Sample_{000\,l} + \beta^t_{11} \cdot Item^t_{ijkl} \cdot Qualität_{0\,jkl} + \beta^t_{111} \cdot Neuro_{00\,kl} \cdot Item^t_{ijkl}$$

$D^t_{ijkl} = \upsilon_{1\,jkl} \cdot Item^t_{ijkl} + \upsilon_{0\,jkl} + \epsilon_{ijkl}$.

In oben stehender Regressionsgleichung zählt man zehn Koeffizienten. Allerdings ist in der Gleichung der Koeffizient für nur eines der vier Items aufgeführt. Im festen Teil kommen wegen der Effektkodierung der Items dreimal zwei Koeffizienten hinzu; im zufälligen Teil ändert sich nichts, weil die Zufallskomponente $\upsilon_{0\,jkl}$ für alle Items gleichermaßen gilt; wäre sie von Item zu Item verschieden, sodass jedes Item seine eigene, zuordenbare Zufallskomponente besäße, wäre die Zufallskomponente keine Zufallskomponente. Insgesamt müssen also im Modell 16 Koeffizienten geschätzt werden.

4.2 Reanalyse

Das Modell wollte in allen erdenklichen Gruppierungsvarianten nicht zu dem zu Weisberg et al. (2015) hinterlegten Datensatz passen. Zwei Dinge an diesem Datensatz sind allerdings merkwürdig. Erstens enthält er weniger Versuchspersonen als im Text angegeben. Und zweitens enthielt er weder eine Versuchsperson, die nur gute Erklärungen vorgelegt bekommen hatte, noch eine Versuchsperson mit nur schlechten Erklärungen, was bei einer echt randomisierten Qualität extrem unwahrscheinlich ist: $P = \left(1 - \frac{1}{8}\right)^{239} < 2 \cdot 10^{-14}$, was aus der Wahrscheinlichkeit vier gleicher Items

$P=\left(\frac{1}{2}\right)^4+\left(\frac{1}{2}\right)^4=\frac{1}{8}$ bei einer Versuchspersonen komplementär folgt für 239 Versuchspersonen.

Damit konfrontiert stellte Hopkins einen revidierten Datensatz zur Verfügung mitsamt dem Auswertungscode. Anhand des Codes ließ sich dann entschlüsseln, dass die Autoren ihrer Analyse formal ein 2-Ebenen-Cluster-Modell zugrunde legten, und dass sie entgegen ihrer Behauptunge im Text sämtliche Variablen effektcodiert hatten. Außerdem stellte sich heraus, dass Hopkins und Kollegen bei der Effektkodierung der Items ein Fehler unterlaufen ist: die Codierung wird vor der Regressionsanalyse wieder aufgehoben. Dadurch unterscheiden sich die veröffentlichten Werte für die Koeffizienten numerisch von den tatsächlichen, unbeschadet der Signifikanzen.

Die nachfolgenden Berechnungen zur Bestimmung von Effektgröße und Stichprobenumfang nehmen ihren Ausgang im ursprünglichen Datensatz, sind aber im Ergebnis angepasst an den revidierten Datensatz.

4.2.1 Effektgröße

Der Weg zur Effektgröße ist nicht vorgezeichnet, schon gleich gar nicht im Hierarchischen Modell. Vor dem ersten Schritt sollte man sich daher bewusst machen, dass die Effektgröße ihre Bedeutung erst erhält, wenn klar ist, ein wie gearteter Effekt gesucht wird, und vor allem wo er gesucht wird. Suchen kann man den 'Neuroeffekt' beispielsweise in den Individuen, den Clustern oder den Gruppen. Die Suche in Gruppen böte sich hier an, um Unterschieden nachzugehen innerhalb von Gruppen mit identischen Items bzw. zwischen Gruppen, die sich nur in einer Merkmalsausprägung unterscheiden.

Weisberg et al. (2015) interessieren sich aber für Cluster, genauer: für eine Struktur aus Item-Clustern, die eingebettet sind in Versuchspersonen, die wiederum eingebettet sind in Kontroll-und Versuchsgruppe (Neuro-Cluster). Hat man sich für diese Struktur entschieden, ist zur Bestimmung der Effektgröße noch eine Entscheidung zu treffen zwischen dem Cluster mit Neuroinformation und dem Cluster ohne Neuroinformation, zumal die Cluster verschieden groß sind. Anzahl und Größe der Cluster oder Gruppen

sind nach Lipsey (1990, S.141) und Snijders (2005) für die Bestimmung der Effektgröße bedeutsamer als der Umfang der gesamten Stichprobe.

Die Wahl des Populationsausschnitts, in dem der Effekt verortet sein soll, wirkt sich aus auf die Größe der Standardabweichung, die zur Berechnung der Effektgröße benötigt wird. So kann man beispielsweise durch geschicktes Gruppieren die Varianz der gesamten Stichprobe zerlegen in die Varianzen der einzelnen Gruppen, sodass für die Untersuchung dann nur die Varianz bzw. der Standardfehler der Gruppe relevant wird, für die man sich letztlich interessiert. Zu beachten ist allerdings, dass durch das Gruppieren Freiheitsgrade in statistischen Tests verlorengehen.

Maßgeblich kommt es also bei der Bestimmung der Effektgröße darauf an, die Standardabweichung nicht nur präzise, sondern auch valide, d.h. für den richtigen Populationsausschnitt zu schätzen. Die Originalstudie weist eine Spannweite auf von Standardabweichungen, die von 0.0602 bis 1.9552 reicht. Die geringste Standardabweichung erhält man, wenn man die Varianzen der nach Items mit gleicher Qualität eingeteilten Gruppen separat betrachtet. Deutlich größere Standardabweichungen erhält man bei der Betrachtung von Clustern.

Entscheidet man sich für die Gruppen auf der Ebene der Neuroinformation (Ebene 3), ist man nach Hedges (2007) immer noch mit drei Standardabweichungen konfrontiert, die im Grunde gleichwertig sind und allein vom Forschungsinteresse abhängen: (1) die Standardabweichung $\sigma_{zwischen}$ der Beurteilungen der Items mit Neuroinformation gegenüber denjenigen der Items ohne Neuroinformation, (2) die Standardabweichung σ_{Fehler} der Beurteilungen der jeweiligen Gruppen gegenüber ihrem Gruppenmittelwert und (3) die Standardabweichung σ sämtlicher Items gegenüber dem Gesamtmittelwert, mit

$$\hat{\sigma}_{zwischen}=\sqrt{\frac{\sum_i^m (\hat{\bar{Y}}_{mit}-\hat{Y}_{mit})^2+\sum_i^m (\hat{\bar{Y}}_{ohne}-\hat{Y}_{ohne})^2}{n_{mit}+n_{ohne}-2}}=1.0472 \quad ,$$

$$\hat{\sigma}_{Fehler}=\sqrt{\frac{\sum_i^m \sum_j^n (Y_{mit}-\hat{Y}_{mit})^2+\sum_i^m (Y_{ohne}-\hat{Y}_{ohne})^2}{N-M}}=1.5898 \quad \text{und}$$

$$\hat{\sigma}=\sqrt{\frac{\sum_i^m \sum_j^n (Y_{mit}-\bar{Y})^2+\sum_i^m (Y_{ohne}-\bar{Y})^2}{N-2}}=1.9037 \quad ,$$

wobei $\hat{\hat{Y}}$ für den jeweiligen Gruppenmittelwert und $\overline{\hat{\hat{Y}}}$ für den Mittelwert der Gruppenmittelwerte steht, und *M* die Anzahl der Gruppen oder Cluster angibt.

Die Standardabweichung zwischen den Clustern kommt nur dann infrage, wenn man sich auf Clusterebene für aggregierte Effekte interessiert. Die Wahl der Gesamtstandardabweichung für die Analyse ist nur scheinbar naheliegend; denn es ist wenig plausibel, dass die Versuchseinheiten aus einem bestehenden Populationscluster gezogen wurden. Vielmehr wurden in der Studie die Versuchspersonen bzw. Items erst durch die Versuchsanordnung geclustert, sodass der Maßstab für die Standardisierung der Effektgröße innerhalb der Cluster erst erzeugt wird. Diese Clusterbildung soll im Modell ja gerade abgebildet sein.

Für das vorliegende Design empfiehlt sich daher die residuale Standardabweichung. Lässt man allerdings die Gruppierungen außen vor und konzentriert sich wie die Autoren der Originalstudie auf das Cluster der Versuchsgruppe, in der der Effekt vermutet wird, dann nimmt die Standardabweichung für die Residuen zu: $\hat{\sigma}_{Fehler} = 1.8939$.

Auskunft über die Genauigkeit bzw. die Verzerrung der geschätzten Standardabweichung durch den Stichprobenfehler gibt die für den Effekt relevante Intraklassenkorrelation: je kleiner der Intraklassenkorrelationskoeffizient ist, desto reliabler sind die Schätzungen aus der Stichprobe. Bei drei Ebenen gibt es nach Hox (2010, S.34) auf der dritten Ebene mehrere Intraklassenkorrelationen. Weil es im 2-Ebenen-Cluster-Modell um die Korrelation zweier beliebiger Einheiten aus einem der beiden Neuro-Cluster geht und darin zur Qualität keine Zufallskomponente hinzukommt, bleibt nur eine Intraklassenkorrelation übrig mit dem Koeffizienten

$$ICC_3 = \frac{\sigma^2_{\upsilon_{00kl}}}{\sigma^2_{\upsilon_{00kl}} + \sigma^2_{\varepsilon_{00kl}}} = \frac{0.2089}{0.2089 + 2.6887} = 0.0721 \quad .$$

Damit stimmen zwei beliebige Einheiten aus einem der beiden Neurocluster überein in ihrer Beurteilung der Items mit einer Wahrscheinlichkeit, die 7.2 Prozent größer ist als bei zwei Einheiten, die völlig zufällig aus der Population gezogen würden.

Aus den vorhandenen Daten lässt sich dann die Effektgröße schätzen anhand des Bestimmtheitsmaßes R^2. Dem durch das Bestimmtheitsmaß ausgedrückten Verhältnis von den Abweichungen zwischen den Versuchseinheiten zu den Abweichungen insgesamt entspricht im Hierarchischen Modell das Verhältnis von der Differenz der Varianz

der Residuen im Grundmodell, das nur aus dem Zufallsachsenabschnitt besteht, und der Varianz der Residuen im Modell mit sämtlichen Prädiktoren zur Varianz der Residuen im Grundmodell (Hox 2010, S.71). Somit gilt für das Bestimmtheitsmaß auf der Ebene der Items im Grunde:

$$R_1^2 = \frac{\left(\sigma^2_{\varepsilon_{ijkl}}\right)_{nurAbschnitt} - \left(\sigma^2_{\varepsilon_{ijkl}}\right)_{Ebene\,2}}{\left(\sigma^2_{\varepsilon_{ijkl}}\right)_{nurAbschnitt}}$$ und analog auf der Ebene ihrer Qualität:

$$R_2^2 = \frac{\left(\sigma^2_{\upsilon_{0\,jkl}}\right)_{nurAbschnitt} - \left(\sigma^2_{\upsilon_{0\,jkl}}\right)_{Ebene\,2}}{\left(\sigma^2_{\upsilon_{0\,jkl}}\right)_{nurAbschnitt}} \quad .$$

Allerdings führen diese Formeln, wie beim Datensatz der Originalstudie, begünstigt durch die unbalancierte Stichprobe, zu negativen Bestimmtheitsmaßen und damit zu einem formalen Widerspruch (Lehmann & Casella 1998, S.191). Den kann man nach Snijders und Bosker (1994) bis zur zweiten Ebene beheben, indem man die Varianzen von Zufallssteigungen berücksichtigt und bezogen auf das jeweilige Modell gewichtet. Für den Fall, dass die Prädiktoren im Modell ausgerichtet sind am Gesamtmittelwert und nur eine einzige Zufallssteigung vorgesehen ist, erfolgt die Korrektur auf Ebene1 dadurch, dass man die Residuenvarianzen $\sigma^2_{\varepsilon_{ijkl}}$ jeweils ersetzt durch:

$$\sigma^2_{\upsilon_{0\,jkl}} + \sigma^2_{\upsilon_{1\,jkl}} \cdot \left(\sigma_{zwischen} + \sigma_{Fehler}\right) + \sigma^2_{\varepsilon_{ijkl}} \quad ,$$

wobei $\sigma_{Zwischen}$ die Standardabweichung der Beurteilungen bedeutet von Items mit guten Erklärungen gegenüber Items mit schlechten Erklärungen, und σ_{Fehler} die Standardabweichung der Residuen, die innerhalb der Gruppen mit guten und schlechten Erklärungen verbleiben. Auf der zweiten Ebene muss zusätzlich die Größe der Gruppen berücksichtigt werden, sodass sich hier folgende Ersetzungen für $\sigma_{\upsilon_{0\,jkl}}$ ergeben:

$$\sigma^2_{\upsilon_{0\,jkl}} + \sigma^2_{\upsilon_{1\,jkl}} \cdot \left(\sigma_{zwischen} + \frac{1}{n} \cdot \sigma_{Fehler}\right) + \frac{1}{n} \cdot \sigma^2_{\varepsilon_{ijkl}} \quad .$$

Dabei steht *n* für die Gruppengröße einer balancierten Studie. Sind die Gruppen unterschiedlich groß, kann *n* nach Muthén (1994) angenähert werden durch:

$$\hat{n} = \frac{N^2 - \sum_j n_j^2}{N \cdot (J-1)} \quad .$$

Auf der Ebene der Neuroinformation kommen nur feste Komponenten zum Modell hinzu, sodass mit Blick auf das Bestimmtheitsmaß nur marginale Verzerrungen zu erwarten sind, wenn man die Neuroinformation wie einen zweiten Prädiktor auf der Ebene der Qualität behandelt, sofern man die kleineren Gruppengrößen auf dieser Ebene berücksichtigt, d.h. anstelle von *J* mit *K* rechnet, sodass $\hat{n}=31.12$ und:

$$R_3^2=\frac{\left(\sigma^2_{\upsilon_{0\,jkl}}+\sigma^2_{\upsilon_{1\,jkl}}\cdot\left(\sigma_{Zwischen}+\frac{1}{\hat{n}}\cdot\sigma_{Fehler}\right)+\frac{1}{\hat{n}}\cdot\sigma^2_{\varepsilon_{ijkl}}\right)_{nurAbschnitt}-\left(\sigma^2_{\upsilon_{0\,jkl}}+\sigma^2_{\upsilon_{1\,jkl}}\cdot\left(\sigma_{Zwischen}+\frac{1}{\hat{n}}\cdot\sigma_{Fehler}\right)+\frac{1}{\hat{n}}\cdot\sigma^2_{\varepsilon_{ijkl}}\right)_{Neuro}}{\left(\sigma^2_{\upsilon_{0\,jkl}}+\sigma^2_{\upsilon_{1\,jkl}}\cdot\left(\sigma_{Zwischen}+\frac{1}{\hat{n}}\cdot\sigma_{Fehler}\right)+\frac{1}{\hat{n}}\cdot\sigma^2_{\varepsilon_{ijkl}}\right)_{nurAbschnitt}}$$

$$=\frac{\left(0.021+0.21\left(1.42+\frac{1}{31.12}1.78\right)+\frac{1}{31.12}3.638\right)-\left(0.085+0.21\left(0.97+\frac{1}{31.12}1.93\right)+\frac{1}{31.12}2.663\right)}{\left(0.021+0.21\left(1.42+\frac{1}{31.12}1.78\right)+\frac{1}{31.12}3.638\right)}$$

$$=0.137\quad.$$

Weil Weisberg et al. (2015) dagegen keine Gruppierung vorgenommen haben und sich auf die beiden Neuro-Cluster beschränkten, ergibt sich ein $\hat{n}=470.77$ und ein $R_3^2=0.2615$. Somit klärt im gesamten Modell die Varianz der Koeffizienten auf der dritten Ebene 26.2 Prozent der Residuenvarianz im Grundmodell auf. Bezogen auf die zufällige Varianz $\left(\sigma^2_{\upsilon_{0\,jkl}}\right)_{nurAbschnitt}$ des Grundmodells entspricht das einer Varianz $\left(\sigma^2_{\upsilon_{0\,jkl}}\right)_{erklärt}$ von $0.0208\cdot0.265=0.0055$, die vollständig zurückgeht auf die Varianz des Modells auf der Ebene der Neuroinformation.

Der Anteil, den die Varianz der Neuroinformation an der Varianzaufklärung hat, lässt sich daran bemessen, wie stark die Varianz der Neuroinformation und die 26.2 Prozent der aufgeklärten Varianz zusammenhängen. Anders ausgedrückt: die Größe des Effektes, den Neuroinformation auf die Beurteilung der Items ausübt, liegt darin, in welchem Maße der Prädiktor 'Neuroinformation' und die aufgeklärte Varianz des Gesamtmodells miteinander korrelieren. Die Korrelation gewinnt man analog zu Hox (2010, S.240) am einfachsten, wenn man die multiple Regression reduziert auf eine einfache Regression der Form:

$Neuro_{00k}=\beta_{000}+\beta_{001}\cdot\upsilon_{00k}+\varepsilon_{00k}$ mit den Varianzen $\sigma^2_{Neuro_{00k}}$, $\left(\sigma^2_{\upsilon_{00k}}\right)_{erklärt}$ und $\sigma^2_{\varepsilon_{00k}}$.

Dabei regrediert die Neuroinformation auf die verbliebenen, unerklärten Zufallsschwankungen des Grundmodells, und nicht umgekehrt, weil es für die Varianz der Neuroinformation an der erklärten Varianz schlicht nichts mehr zu erklären gibt. Die obige Darstellung der Regression rechtfertigt sich allein durch den Zweck, die Stärke des Zusammenhangs zwischen den Faktoren zu bestimmen, denn im Ergebnis ist es für die Korrelation egal, welcher von beiden Regressor ist und welcher Regressand.

Bei der einfachen linearen Regression mit nur einem Prädiktor gilt, dass das Bestimmtheitsmaß R^2 dem Quadrat des Pearsonschen Korrelationskoeffizienten r entspricht. Somit gilt:

$$r^2 = \frac{\left(\sigma^2_{\upsilon_{00k}}\right)_{erklärt}}{\sigma^2_{Neuro_{00k}}} = \frac{0.0055}{0.906} = 0.0061 \quad ,$$

d.h., erklärte Beuteilungsschwankungen und Neuroinformation korrelieren mit $r = \sqrt{0.0061} = 0.078$. Daraus lässt sich nun leicht eine Effektgröße bestimmen. Nach Cohen (1977, S.24) gilt nämlich:

$$d_r = \frac{r}{\sqrt{1-r^2}} \cdot \frac{n_{mit}+n_{ohne}}{\sqrt{n_{mit} \cdot n_{ohne}}} = \frac{0.079}{\sqrt{1-0.006}} \cdot \frac{1040}{\sqrt{680 \cdot 360}} = 0.164 \quad .$$

Die Übertragung der standardisierten Effektgröße in ein Hierarchisches Modell schafft dort allerdings keinen Vergleichsstandard wegen der vielen Varianzkomponenten, die als Fehlervarianzen betrachtet werden können (Judd, Westfall & Kenny 2012). Zudem ist der Schätzwert für den Effekt im Hierarchischen Modell in der Regel kleiner als im Standardmodell, weil dort der Standardfehler bei verschachtelten Strukturen unterschätzt wird (Maas & Hox 2005). So kommt Lipsey (1990, S.78) dem iterativ gewonnenen $\hat{\sigma}_{\varepsilon_{ijkl}}$ sehr nahe, bleibt aber darunter mit der Formel für die gepoolte Standardabweichung:

$$\hat{\sigma} = \sqrt{\frac{(n_{mit}-1)s^2_{mit} + (n_{ohne}-1)s^2_{ohne}}{(n_{mit}-1)+(n_{ohne}-1)}} = \sqrt{\frac{679 \cdot 1.948^2 + 359 \cdot 1.790^2}{1038}} = 1.8939 < 1.8946 = \hat{\sigma}_{\varepsilon_{ijkl}} \quad .$$

Bei einem kanonischen Mittelwertsvergleich mit $\overline{Y}_{mit} = 0.288$ und $\overline{Y}_{ohne} = -0.033$ lässt sich die Effektgröße nach Rosnow und Rosenthal (1989) anhand des t-Wertes schätzen:

$$d_t = t \cdot \frac{\sqrt{n_{mit}+n_{ohne}}}{\sqrt{n_{mit} \cdot n_{ohne}}} = 2{,}675 \cdot \frac{\sqrt{1040}}{\sqrt{680 \cdot 360}} = 0.174 \quad .$$

Verwendet man die Dummy-Kodierung für die Neuroinformations-Cluster, kommt man zu einer vergleichbaren Effektgröße über die punktbiseriale Korrelation:

$$r_{pb}=\frac{(\hat{\mu}_{mit}-\hat{\mu}_{ohne})}{\sigma}\cdot\sqrt{q}=\frac{(0.321)}{1.8944}\cdot\sqrt{0.2263}=0.081 \quad \text{mit} \quad q=\frac{n_{mit}}{n_{mit}+n_{ohne}}\cdot\frac{n_{ohne}}{n_{mit}+n_{ohne}} \quad ,$$

denn dann ist $d_{r_{pb}}=\frac{r_{pb}}{\sqrt{q(1-r_{pb}^2)}}=\frac{0.081}{\sqrt{0.2263\cdot(1-0.081^2)}}=0.169$.

Will man im Falle einer multiplen Regression den Effekt aus dem Grundrauschen herauspräparieren, empfiehlt Lipsey (1990, S.85), die multiple Korrelation r_m zu bilden zwischen der Beurteilung und der Linearkombination der Items, um letztere auszupartialisieren:

$$\hat{\delta}_{r_{pb}}=\frac{d_{r_{pb}}}{\sqrt{1-r_m^2}}=\frac{0.169}{\sqrt{1-0.354^2}}=0.181.$$

Diese Korrektur behandelt alle Prädiktoren gleichermaßen auf einer Ebene. Im Hinblick auf mehrere Ebenen ist aber die Intraklassenkorrelation geeigneter als die multiple Korrelation. Korrigiert man die Effektgröße mittels Intraklassenkorrelation, erhält man für den Effekt auf der Neuroinformationsebene:

$$\hat{\delta}_{r_{pb}}=\frac{d_{r_{pb}}}{\sqrt{1-ICC_3}}=\frac{0.169}{\sqrt{1-0.0721}}=0.175 \quad \text{bzw.} \quad \hat{\delta}_r=\frac{d_r}{\sqrt{1-ICC_3}}=\frac{0.164}{\sqrt{1-0.0721}}=0.170 \quad .$$

Der Unterschied zwischen korrigierter und unkorrigierter Effektgröße beträgt nur 0.006 $\hat{\sigma}$, vergleicht man aber die Varianzen der beiden Effektgrößen, so ist im Standardmodell

$$\sigma^2_{\hat{\delta}_{r_{pb}}}=\frac{n_{mit}+n_{ohne}}{n_{mit}\cdot n_{ohne}}+\frac{\hat{\delta}_r^2}{2(n_{mit}+n_{ohne}-2)}=\frac{1040}{680\cdot 360}+\frac{0.175^2}{2\cdot 1038}=0.0043 \quad ,$$

wohingegen nach Hedges (2007) im Hierarchischen Modell mit *M* Clustern gilt:

$$\sigma^2_{\hat{\delta}_r}=\left(\frac{n_{mit}+n_{ohne}}{n_{mit}\cdot n_{ohne}}\right)\cdot\left(\frac{1+(\hat{n}-1)\cdot ICC_3}{1-ICC_3}\right)+\frac{\hat{\delta}_r^2}{2(N-M)}=0.0043\cdot\frac{1+469.77\cdot 0.0721}{0.9279}+\frac{0.170^2}{2(1040-2)}=0.158 \quad .$$

Somit ist für den Neuroinformations-Effekt die Varianz der Effektgröße im Hierarchischen Modell mit zwei Clustern mehr als das Dreißigfache größer als im Standardmodell, was ein ungleich größeres Konfidenzintervall für jenes zur Folge hat. Im Hier-

archischen Modell ist somit die Schätzung der Effektgröße deutlich unpräziser; die Validität der Schätzung erkauft man sich mit Einbußen in ihrer Präzision.

Dass die Effektgröße starken Schwankungen ausgesetzt ist, zeigt sich auch daran, dass Levy (1967) folgend die Beurteilungen von über einem Drittel der Items mit Neuroinformation mehr als die Hälfte vom Clustermittelwert nach unten abweicht. Entsprechend groß ist die Schnittmenge beider Cluster mit identischen Beurteilungen eines Items; mit $\alpha=0.57$ aus

$$z_\alpha=\frac{1}{2}\frac{r_{pb}}{\sqrt{1-r_{pb}^2}}\sqrt{\frac{(N-m-1)(n_{mit}+n_{ohne})}{(n_{mit}\cdot n_{ohne})}}=0.171 \quad .$$

Letztendlich liegen sämtliche Schätzungen der Effektgröße, sofern $\hat{\delta}_r$ die tatsächliche Effektgröße ist, bei einer Konfidenz von 90 Prozent mit einer Wahrscheinlichkeit von 75.8 Prozent innerhalb der Intervallgrenzen $0.170-1.28\cdot0.397$ und $0.170+1.28\cdot0.397$. Denn bei normalverteilten Mittelwerten mit Varianz σ_n^2 sind nach Estes (1997) aufgrund der Additivität der Varianzen die durchschnittlichen Mittelwertdifferenzen mit Erwartungswert *0* und Varianz $2\sigma_n^2$ verteilt. Die Fläche unter dieser Verteilung entspricht der Wahrscheinlichkeit, dass die Mittelwertdifferenz einer Replikation in das Konfidenzintervall fällt. Für die Transformation der Verteilung in eine Standardnormalverteilung ist $z=\frac{\bar{X}-\bar{\mu}}{\sqrt{2}\sigma_n}$, sodass gegenüber Standardtransformationen $z=\frac{z_{(1-\alpha/2)}}{\sqrt{2}}=1.17$ gilt. Das entspricht einseitig einer Wahrscheinlichkeit von 87.9 Prozent, von der man für die untere Grenze nach Cumming, Williams und Fidler (2004) des geschlossenen Intervalls noch einmal 12.1 Prozent abziehen muss.

Somit ist $\hat{\delta}_r=0.170\pm0.508$, was bedeutet, dass bei 100 Wiederholungen des Experimentes in 76 Fällen die Items mit Neuroinformation besser beurteilt werden als 36.7 Prozent bis 71.2 Prozent der Items ohne Neuroinformation.

Das Resultat deckt sich mit den Effektgrößen vergleichbarer Studien, insofern als diese im Konfidenzintervall [−0.338; 0.678] liegen. Für ablenkende Mathematikdetails (Eriksson 2012) lässt sich eine Effektgröße $d=0.224$ bestimmen, für ablenkende Gehirnaufnahmen $d=0.260$ (McCabe & Castel 2008) und $d=0.352$ (Keehner et al. 2011) bzw.

d=0.081 (Michael et al. 2013), *d=0.136* (Gruber & Dickerson 2012) und *d=0.173* (Hook & Farah 2013). Auf *d=0.402* lässt sich die Größe des Neuro-Effektes bei seinem erstmaligen Nachweis (Weisberg et al. 2008) taxieren; Replikationen erbrachten *d=0.667* (Scurich & Shniderman 2014), *d=0.84* (Minahan & Siedlecki 2016) und *d=0.166* (Hopkins, Weisberg & Taylor 2016). Signifikant weichen ab Fernandez-Duque et al. (2015) mit bestätigendem *d=1.616* und Tabacchi & Cardaci (2016) mit ablehnendem Cramérs *V=0.520*, was nach Cohen (1977, S.79) *d=0.8* entspricht.

Interpretiert man das Resultat im Rahmen der binomialen Effektgrößendarstellung (BESD), die als Bezugsgröße für die Schwelle eines erfolgreichen Effekt-Nachweises den Gesamtmedian heranzieht (Rosenthal & Rubin 1982), sodass *(d+r/2)·100* Prozent der Versuchsgruppe und *(d−r/2)·100* Prozent der Kontrollgruppe über dem Median liegen. Folglich lässt sich der hier postulierte Effekt nachweisen in 20.3 Prozent der Items mit Neuroinformation und in 12.5 Prozent der Items ohne Neuroinformation.

Bei der Bestimmung des Konfidenzintervalls wie auch beim Poolen der Standardabweichung wird vorausgesetzt, dass die Cluster dieselben Varianzen besitzen. Das ist bei randomisierten Studien zwar zu erwarten, dennoch können nach Snijders und Bosker (1999, S.126) im Verlauf einer Studie Heteroskedastizitäten auftreten. Das Verhältnis der Varianzen ist in der vorliegenden Studie mit *VR=1.186* und einer Kurtosis von *−1.182* zwar moderat, doch statistisch ist die Verteilung der Residuen weder homogen (White *p=0.003*) noch normal (Kolmogorov-Smirnov *p<0.001*), wodurch die Größenangaben des Effekts verzerrt werden. Angesichts der in Konsequenz variierenden Effektgrößen (Grissom & Kim 2001) soll daher der Neuro-Effekt kurz im Lichte alternativer Charakterisierungen betrachtet werden.

Für die Effektgröße erhält man einen gegen Ausreißer robusteren Schätzer, wenn man statt der Mittelwerte die Mediane der Cluster vergleicht und deren Differenz bezieht auf ein Variabilitätsmaß wie den Median der absoluten Abweichung vom Median (MAD). Ein solches Maß ist die Biweight-Standardabweichung, sodass nach Wilcox (2012, S.455) gilt:

$$d_M = \frac{Mdn_{mit} - Mdn_{ohne}}{\sigma_{bw}} = \frac{1-0}{1.855} = 0.539 \quad \text{mit} \quad \sigma_{bw} = \sqrt{N} \cdot \frac{\sqrt{\sum_i \mathbf{1}_{A_i} (A_i - Mdn)^2 (1 - A_i^2)}}{\left| \sum_i (\mathbf{1}_{A_i} - A_i^2)(1 - 5A_i^2) \right|} = 3.440 \quad ,$$

wobei

$A_i = \frac{Y_i - Mdn}{9 \cdot MAD}$ ist und $\mathbf{1}_{A_i} = \begin{cases} 1, \text{ falls } |A_i| < 1 \\ 0, \text{ falls } |A_i| > 1 \end{cases}$ die Indikatorfunktion.

Folglich liegt bei der Hälfte der Items mit Neuroinformation deren Beurteilung um eine Stufe höher als die der ebenfalls aufsteigend geordneten Items ohne Neuroinformation. Allerdings können bei Verwendung des Medians wertvolle Informationen bezüglich der Beurteilungen verloren gehen, weshalb sich nach Wilcox (1995) eine Darstellung d_Q in Form von Quantilen empfiehlt, deren Konfidenzintervalle unter Verwendung der Harrell-Davis-Schätzer für Quantile, hier zum Konfidenzniveau *0.9*, simultan bestimmt werden:

$d_{0.1} = 0.04 \pm 0.41$ $d_{0.2} = -0.01 \pm 0.04$ $d_{0.3} = -0.01 \pm 0.05$ $d_{0.4} = 0.69 \pm 0.44$ $d_{0.5} = 0.97 \pm 0.38$

$d_{0.6} = 0.01 \pm 0.04$ $d_{0.7} = 1.00 \pm 0.00$ $d_{0.8} = 0.29 \pm 0.46$ $d_{0.9} = 0.90 \pm 1.07$ $d_{1.0} = 0.00 \pm 1.31$

.

Daraus wird ersichtlich, dass sich die Items in den Spitzen, bei beträchtlichen Schwankungen, so gut wie nicht in ihrer Beurteilung unterscheiden. Demnach generiert der Effekt seine Größe im wesentlichen aus der Breite mittlerer Beurteilungen. Die Breite wiederum impliziert, dass die Verteilungen der Items mit und ohne Neuroinformation zu einem hohen Grad überlappen. Je größer die Überlappung, desto schwieriger ist die Zuordnung einer Item-Beurteilung zu dem Cluster, in dem der Effekt wirksam sein sollte. Als Maß für den Grad der Überlappung bietet sich also die Trefferquote der Zuordnung an.

Formal wird die Trefferquote wiedergegeben durch die Wahrscheinlichkeit, dass das *i*-te Item im Lichte seiner Beurteilung Y_i ein Item mit Neuroinformation ist:

$$P(Neuro_{mit}|Y_i) = \frac{p_{mit} \cdot e^{-\frac{1}{2}D^2_{mit,i}}}{p_{mit} \cdot e^{-\frac{1}{2}D^2_{mit,i}} + p_{ohne} \cdot e^{-\frac{1}{2}D^2_{ohne,i}}} = \frac{\frac{17}{26} \cdot e^{-\frac{1}{2} \cdot \frac{1486}{1040}}}{\frac{17}{26} \cdot e^{-\frac{1}{2} \cdot \frac{1486}{1040}} + \frac{9}{26} \cdot e^{-\frac{1}{2} \cdot \frac{1499}{1040}}} = 0.655 \quad ,$$

wobei D^2 für die quadrierten Mahalanobis-Distanzen steht, und *p* für die Wahrscheinlichkeiten, dass ein Item zufällig mit bzw. ohne Neuroinformation versehen ist. Setzt man diese Wahrscheinlichkeit ins Verhältnis zur erwartbaren Wahrscheinlichkeit zufälliger Treffer, erhält man nach Hess, Olejnik und Huberty (2001) einen Index, der den Anteil der Treffer angibt, die nicht zufällig zustande kommen:

$$d_I = \frac{P(Neuro_{mit}|Y_i) - \frac{p_{mit} \cdot n_{mit}}{N}}{1 - \frac{p_{mit} \cdot n_{mit}}{N}} = \frac{0.655 - 0.428}{1 - 0.428} = 0.397 \quad .$$

Demnach bringt das Experiment 1.5 mal mehr Zufallstreffer hervor als effektbedingte Treffer; d.h., zwei von drei Beurteilungen von Items mit Neuroinformation decken sich mit der Beurteilung von Items ohne Neuroinformation. Der hohe Überlappungsgrad indiziert eine geringe Teststärke, die zur Berechnung des Stichprobenumfangs benötigt wird.

4.2.2 Stichprobenumfang

In Mehrebenenmodellen hat jede Ebene ihren eigenen optimalen Stichprobenumfang, der sich nach den Versuchseinheiten bemisst, die auf der Ebene untersucht werden. Diese Versuchseinheiten sind in der vorliegenden Studie Cluster. Die Durchschnittsgröße der Cluster ist relativ unbedeutend im Vergleich zu ihrer Anzahl (Snijders 2005).

Zur Bestimmung des optimalen Stichprobenumfangs lassen sich zwei Ansätze unterscheiden. Der eine Ansatz führt über die Teststärke zum Stichprobenumfang, der andere über die Präzision der Effektschätzung. Zum Zwecke einer Triangulation werden beide Ansätze verfolgt und dann der größere Stichprobenumfang ausgewählt, um den Ansprüchen beider Ansätze zu genügen. Denn Tests, die stark genug sind, um einen bestimmten Effekt nachzuweisen, können dennoch nicht stark genug sein, um dessen Größe in der Population hinreichend präzise anzugeben. Und umgekehrt: Tests, die im gewünschten Maße präzise sind, können dennoch für den Nachweis eines Effektes zu schwach sein.

4.2.2.1 Präzisionsansatz

Beim Präzisionsansatz geht es darum, Werte zu schätzen, die möglichst genau den Werten in der Population entsprechen, und weniger darum, statistisch signifikante Schätzwerte zu generieren. Die Präzision eines Schätzwertes wird angegeben durch dessen Konfidenzintervall. Legt man fest, in welchen Grenzen der Effekt liegen soll und zu welchem Konfidenzniveau der Effekt innerhalb dieser Grenzen liegen soll, wird zur

Berechnung des Stichprobenumfangs nur noch die richtige Standardabweichung des den Effekt auslösenden Prädiktors in der Population benötigt; die Effektgröße selbst geht nicht in die Berechnung ein. Die Ober- und Untergrenzen des Intervalls ergeben sich aus der Weite *w,* die zur Effektgröße addiert oder von ihr subtrahiert wird, sodass *2w* die volle Breite des Konfidenzintervalls markieren.

Die Weite *w* ohne Berücksichtigung der verschiedenen Ebenen wird nach Kelley und Rausch (2003) angegeben mit $w = t_{(1-\alpha/2;\, N-2)} \cdot \sigma \cdot \sqrt{\frac{n_{mit} + n_{ohne}}{n_{mit} \cdot n_{ohne}}}$. Daraus folgt für den Umfang: $n_{ohne} = \frac{26 \cdot t^2_{(1-\alpha/2;N-2)} \sigma^2}{17 \cdot w^2}$. Soll die Weite genau dem 90%-Konfidenzintervall der Effektgröße aus der Originalstudie entsprechen, dann ist im Standardmodell $w = 2 \cdot 0.084$, weil die Standardabweichung der Effektgröße kleiner ist als im Hierarchischen Modell: $\sigma_{\hat{\delta}_{rpb}} = 0.066$.

Die zugehörige gepoolte Standardabweichung ist dagegen die des Residuums auf der Neuroinformationsebene, die sich mittels Formel auf $\hat{\sigma}_{Fehler} = 1.6397$ schätzen lässt und in iterativen Durchläufen nach dem Maximum Likelihood-Prinzip zu $\hat{\sigma}_{L3} = 1.6496$ konvergiert. Eingesetzt erhält man

$$n_{ohne} = \frac{26 \cdot 1.65^2 \cdot 1.650^2}{17 \cdot 0.168^2} = 401.6 \quad .$$

Für den Nachweis des Effektes wären folglich 402 Items ohne Neuroinformation erforderlich, wofür 101 Versuchspersonen benötigt würden. Wegen $N_R = \frac{17}{9} \cdot n_{ohne} + n_{ohne}$ käme man insgesamt auf 292 Versuchspersonen.

Weil im Standardmodell die Varianz, und damit der Standardfehler, systematisch unterschätzt wird, wird dort allerdings auch der Stichprobenumfang zu klein angesetzt. Das gilt in gleicher Weise, wenn man den Stichprobenumfang analog zur Effektgröße mittels Bestimmtheitsmaß ermittelt. Nach Kelley und Maxwell (2003) gehen dann in die Bestimmung der Weite des Konfidenzintervalls die Bestimmtheitsmaße sämtlicher Prädiktoren des Modells ein:

$N_R = \left(\frac{z_{1-\alpha/2}}{w}\right)^2 \cdot \frac{1-R^2}{1-\frac{1}{r_{jj}}} + n_{Prädiktoren} + 1$, wobei r_{jj} das j-te Hauptdiagonalenelement der invertierten Korrelationsmatrix aller Prädiktoren ist. In der Originalstudie besitzt die 7×7 -Korrelationsmatrix die Form:

$$\boldsymbol{A} = \begin{pmatrix} 1 & 0.03 & 0.09 & -0.06 & 0.31 & 0.15 & 0.02 \\ 0.03 & 1 & 0.01 & 0.03 & 0.37 & -0.06 & 0.16 \\ 0.09 & 0.09 & 1 & 0.15 & 0.10 & 0.12 & 0.10 \\ -0.06 & 0.03 & 0.15 & 1 & 0.02 & 0.14 & 0.05 \\ 0.31 & 0.37 & 0.10 & 0.02 & 1 & 0.17 & 0.10 \\ 0.15 & -0.06 & 0.12 & 0.14 & 0.017 & 1 & -0.02 \\ 0.02 & 0.16 & 0.10 & 0.05 & 0.10 & -0.02 & 1 \end{pmatrix} \begin{matrix} Item1 \\ Item2 \\ Item3 \\ Item4 \\ Qualität \\ Neuroinformation \\ Sample \end{matrix}$$

und damit lautet die Inverse:

$$\boldsymbol{A^{-1}} = \begin{pmatrix} 1.14 & 0.1 & -0.08 & 0.11 & -0.36 & -0.11 & 0 \\ 0.10 & 1.22 & -0.06 & -0.02 & -0.49 & 0.15 & -0.14 \\ -0.08 & -0.06 & 1.06 & -0.14 & -0.03 & -0.10 & -0.09 \\ 0.11 & -0.02 & -0.14 & 1.05 & -0.01 & -0.14 & -0.03 \\ -0.36 & -0.49 & -0.03 & -0.01 & 1.33 & -0.20 & -0.04 \\ -0.11 & 0.15 & -0.10 & -0.14 & -0.20 & 1.09 & 0.03 \\ 0 & -0.14 & -0.09 & -0.03 & -0.04 & 0.03 & 1 \end{pmatrix} .$$

Für $j=6$ (Neuroinformation) ist die Toleranz $r_{66}=1.09;$ Einsetzen ergibt:

$$N_R = \left(\frac{1.65}{0.168}\right)^2 \cdot \frac{1-0.262}{1-\frac{1}{1.09}} + 7 + 1 = 870.2 \quad .$$

Das entspräche 219 Versuchspersonen, mit 73 Personen im Cluster ohne Neuroinformation und 146 Personen im Cluster mit Neuroinformation.

Solange die Varianz, die durch die Ebenen oder Cluster bedingt ist, bei der Varianzaufklärung unberücksichtigt bleibt, wird der Umfang unterschätzt. Würde der Beitrag der Ebenen oder Cluster fehlen, würden die Varianzen der Variablen formal mindestens gleich bleiben, in der Regel aber zunehmen, um für die fehlende Varianzaufklärung aufzukommen: Ebenen oder Cluster kaschieren das tatsächliche Ausmaß der Varianz der Variablen. Um dieses Ausmaß herauszupräparieren, müssten für jede Variable die Intraklassenkorrelationen in die obigen Formeln eingebaut werden. Deren Berechnung aber ist sehr aufwändig, weil die Intraklassenkorrelation von Variable zu Variable verschieden ist.

Um den Stichprobenumfang für eine präzise Replikation im Hierarchischen Modell zu schätzen, ist daher eine Strukturierung erforderlich, die über das Schema von Versuchs- und Kontrollgruppe hinausgeht. Im vorliegenden Modell setzen sich die beiden Neuro-Cluster zusammen aus jeweils 16 Gruppen, deren Items dieselbe Qualität haben (k=16). Für die Weite des Intervalls in der so strukturierten Population ist dann die Standardabweichung $\sigma_{\hat{\delta}_r}$ des Hierarchischen Modells maßgeblich, sodass *w=2·0.508*.

Unter der Annahme, dass die Gruppenmittelwerte normalverteilt sind, ist das Verhältnis der Varianzen von Stichprobe und Population χ^2-verteilt und es gilt nach Pornprasertmanit und Schneider (2014) mit Konfidenz *1-α*:

$$w = 2 \cdot t_{(1-\alpha/2;k-2)} \cdot \sqrt{\frac{\chi^2_{(1-\alpha;k-2)}\left(\sigma^2_{\varepsilon_{ijkl}} + n \cdot \sigma^2_{\upsilon_{0jkl}}\right)}{(k-2) \cdot k \cdot n \cdot p \cdot (1-p)}} \quad \text{, mit} \quad p = \frac{n_{mit}}{n_{mit}+n_{ohne}} \cdot \left(1 - \frac{n_{mit}}{n_{mit}+n_{ohne}}\right)$$

aufgelöst nach der Clustergröße

$$n = \frac{4 \cdot t^2_{(1-\alpha/2;k-2)} \cdot \chi^2_{(1-\alpha;k-2)} \cdot \sigma^2_{\varepsilon_{0jkl}}}{w^2 \cdot \left((k-2)k \cdot p - 4 \cdot t^2_{(1-\alpha/2;k-2)} \cdot \chi^2_{(1-\alpha;k-2)} \cdot \sigma^2_{\upsilon_{0jkl}}\right)}$$

$$= \frac{4 \cdot 1.76^2 \cdot 21{,}10 \cdot 2.72}{1.016^2 \cdot \left(14 \cdot 16 \cdot 0.22 - 4 \cdot 1.76^2 \cdot 21.10 \cdot 0.18\right)} = 310.3 \quad .$$

Das sind balanciert je Neuro-Cluster 311 Versuchspersonen. Insgesamt entspricht das 623 Versuchspersonen, wovon 407 Personen im Cluster mit Neuroinformation und 216 Personen im Cluster ohne Neuroinformation benötigt werden.

4.2.2.2 Teststärkenansatz

Bestimmt man den Stichprobenumfang über die Teststärke, muss zuerst die Größe der Teststärke festgelegt werden, mit der ein Effekt nachgewiesen werden kann, sofern er existiert. Üblicherweise toleriert man hier eine Irrtumswahrscheinlichkeit von 20 Prozent, begnügt sich also mit einer Teststärke von 80 Prozent. Ob es aber folgenreicher ist, die Nullhypothese fälschlicherweise zu verwerfen, als sie fälschlicherweise beizubehalten (Neyman 1942), ist vom Einzelfall abhängig. Sofern es keine ausschlaggebenden Gründe dagegen gibt, ist der Fehler erster Art daher gleichzusetzen mit dem Fehler zweiter Art (Lipsey 1990, S.142).

Angesichts der beschränkten Ressourcen lässt sich ein einheitliches *0.05*-Niveau bei den Irrtumswahrscheinlichkeiten kaum halten, weshalb hier der Mittelweg mit $\alpha=\beta=0.1$ beschritten werden soll, sofern dieser gangbar ist. Will man nun bei einer Teststärke $1-\beta$ von 90 Prozent und einer Irrtumswahrscheinlichkeit α von 10 Prozent einen einseitigen Test aufsetzen, der eine Effektgröße von $\delta=0.170$ nachweist, ist wiederum die Diskrepanz zwischen Hierarchischem und Standardmodell zu beachten.

Im Standardmodell nimmt man an, dass die Mittelwertdifferenz ΔY der beiden Neuro-Cluster für den Fall, dass die Nullhypothese H_0 wahr ist, normalverteilt ist mit $\mu_0=0$ und der Varianz σ^2; und für den Fall, dass die Alternativhypothese H_1 wahr ist, ΔY ebenfalls normalverteilt ist mit $\mu_1=\overline{Y}_{\text{mit}}-\overline{Y}_{\text{ohne}}=0.321$ und derselben Varianz σ^2. Die Irrtumswahrscheinlichkeiten α und β sind verbunden über den kritischen Wert der Beurteilungsdifferenz, weshalb beide – bei gegebener Effektgröße und gleichem Umfang – nicht zugleich minimiert werden können. Der kritischen Wert für den Effekt, der in der Replikation nachgewiesen werden soll, erfüllt die Verteilungsfunktionen $F(\Delta Y_{krit}|H_0)=1-\alpha$ und $F(\Delta Y_{krit}|H_1)=\beta$, sodass z-transformiert gilt:

$$\mu_0+z_{(1-\alpha)}\cdot\sigma_n=\mu_1+z_{(\beta)}\cdot\sigma_n \quad \text{mit} \quad \sigma_n=\sqrt{\frac{\hat{\sigma}^2}{n_{mit}}+\frac{\hat{\sigma}^2}{n_{ohne}}}=\sqrt{\frac{26\cdot\hat{\sigma}^2}{17\cdot n_{ohne}}}$$

$$\Leftrightarrow\ n_{ohne}=\frac{26}{17}\cdot\frac{\left(z_{(0.95)}-z_{(0.1)}\right)^2\cdot\sigma_{L3}^2}{\hat{\mu}_1^2}$$

$$\Leftrightarrow\ n_{ohne}=\frac{26}{17}\cdot\frac{\left(1.28-(-1.28)\right)^2\cdot 1.650^2}{0.321^2}=264.8\ .$$

Demnach müsste die Stichprobe für die Replikation insgesamt mindestens 193 Personen umfassen, 67 Personen im Cluster ohne Neuroinformation und 126 Personen im Cluster mit. G*Power 3.1.9.2 errechnet unter den angeführten Bedingungen (t-Test für Mittelwertdifferenzen zweier unabhängiger Mittelwerte, einseitig, bei Stichprobenverhältnis 17:9) den Stichprobenumfang $N_R=1008$, mit $n_{ohne}=348$ und $n_{mit}=660$, was insgesamt 252 Versuchspersonen entspräche.

Das Standardmodell führt erwartungsgemäß auch im Teststärke-Ansatz zu einem zu geringen Stichprobenumfang, nicht nur weil es eine unterschätzte Varianz der Effektgröße impliziert, die das Modell aus der Unterschätzung des Standardfehlers vererbt, und die sich fortpflanzt auf die Teststärke infolge der lokalen Asymmetrie der Teststärke

um die Effektgröße; d.h., weil die Teststärke zu größeren Effekten hin stärker zu- oder abnimmt, als sie zu kleineren Effekten hin zu- oder abnimmt. Das Standardmodell setzt zudem voraus, dass die Standardabweichung eines jeden Prädiktors konstant ist. Diese Voraussetzung ist nicht erfüllt in einem Mehrebenenmodell wie dem vorliegenden, in dem der Prädiktor der Qualität zufälligen Schwankungen unterliegt.

Zur Bestimmung des Stichprobenumfangs im Hierarchischen Modell anhand der Teststärke nutzt man nach Snijders (2005) ebenfalls die Beziehung zwischen Effektgröße (bzw. Mittelwertdifferenz) und Standardfehler:

$$\frac{d}{\sigma_n} \approx z_{(1-\alpha)} + z_{(1-\beta)} \quad ,$$

die sich ergibt aus dem *t*-Test und dem Umstand, dass *t*-verteilte Werte in großen Umfängen annähernd normalverteilt sind. Allerdings wird hier der – vom Umfang abhängige – Standardfehler der zu erhebenden Stichprobe ermittelt, statt den Standardfehler der erhobenen Stichprobe zugrunde zu legen. Für α=β=0.1 ist dann ein Standardfehler erforderlich in Höhe von:

$$\sigma_n = \frac{\hat{\delta}}{z_{(1-\alpha)} + z_{(1-\beta)}} = \frac{0.170}{1.285 + 1.285} = 0.0638 \quad .$$

Für diesen Stichprobenstandardfehler lässt sich mit der geschätzten Populationsstandardabweichung der Stichprobenumfang bemessen. Für die Populationstandardabweichung wurde im vorliegenden Modell mit Neuro-Clustern die Standardabweichung der Residuen angesetzt, die auf der Ebene der Neuroinformation numerisch gegen $\hat{\sigma}_{L3} = 1.6496$ konvergiert.

Der Zusammenhang zwischen Standardfehler, Standardabweichung und Stichprobenumfang lautet u.a. nach Howell (2010, S.234):

$$\sigma_n = \sqrt{\frac{\sigma_{L3}^2}{n_{mit}} + \frac{\sigma_{L3}^2}{n_{ohne}}} = \sqrt{\frac{26 \cdot \sigma_{L3}^2}{17 \cdot n_{ohne}}} \quad , \text{ umgeformt ergibt sich:}$$

$$n_{ohne} = \frac{26 \cdot \sigma_{L3}^2}{17 \cdot \sigma_n^2} = \frac{26}{17} \cdot \frac{1.650^2}{0.064^2} = 1022.4 \quad .$$

Für den Nachweis des Effektes sind folglich 1023 Items ohne Neuroinformation erforderlich, wofür 256 Versuchspersonen benötigt werden. Wegen $N_R = \frac{17}{9} \cdot n_{ohne} + n_{ohne}$

kommt man insgesamt auf 740 Versuchspersonen. Das sind fast dreimal so viele Versuchspersonen, wie an der Originalstudie teilgenommen haben. Grund dafür ist vor allem die geringe Teststärke von 62 Prozent bei α=0.10 ($1-\beta$=0.47 bei α=0.05) der Originalstudie.

Bestimmt man den Stichprobenumfang anhand der Teststärke, so ist zu bedenken, dass Teststärke und Nullhypothese über α untrennbar miteinander verbunden sind. Somit beruhen die Umfangsberechnungen auf der Annahme, dass die Nullhypothese zutrifft, dass also kein Effekt existiert. Nimmt man dagegen an, dass ein Effekt existiert, dann sind die Mittelwertdifferenzen nicht mehr normalverteilt, sondern folgen der *t*-Verteilung mit einem Nonzentralitätsparameter λ_t, der umso größer ausfällt, je größer der Effekt ist (Steiger 2004), was zur Folge hat, dass die Teststärke zunimmt (Tiku 1971; Gupta & Perlman 1974) und ein Effekt früher signifikant wird (David & Johnson 1951), weil die Standardabweichung ihre Eigenschaft als Maß für die Variation verliert (Pearson 1931; Pearson & Hartley 1951). Mit

$$\lambda_t = \frac{\beta_{001}}{\sqrt{\frac{4\left(\sigma_\varepsilon^2 + n_{L_1} \cdot n_{L_2} \cdot \sigma_\upsilon^2\right)}{n_{L_1} \cdot n_{L_2} \cdot n_{L_3}}}} = \frac{0.16}{\sqrt{\frac{4\left(1.65^2 + 470.77 \cdot 2 \cdot 0.21^2\right)}{470.77 \cdot 2 \cdot 2}}} = 1.70$$

nach Moerbeek und Teerenstra (2016, S.184) erhält man im 3-Ebenen-Modell für die Originalstudie – unter Berücksichtigung dessen, dass dort Zufallseffekte nur auf der Merkmalsebene der Items (Qualität der Erklärungen) vorgesehen sind – einen nur unwesentlich größeren Wert (66 Prozent bei α=0.10 und 52 Prozent bei α=0.05) für die Teststärke, an der letztendlich noch interessiert, wie präzise sie geschätzt ist (Smithson 2001).

Ein Maß für die Präzision der Schätzung ist ihr Konfidenzintervall. In dessen Berechnungen gehen die Residuenvarianzen ein, die vor dem Hintergrund, dass die Nullhypothese nicht zutrifft, *F*-verteilt sind mit einem Nonzentralitätsparameter λ_F. Geht man von einer nonzentralen Verteilung der Residuen aus, lässt sich für die geschätzte Teststärke das Konfidenzintervall zu einem 90 Prozent-Niveau angeben, indem man zuerst mit *α=0.05* und *0.95* das Konfidenzintervall für den Nonzentralitätsparameter berechnet, und dann für die Teststärke genauso jeweils den kritischen Wert bestimmt, dessen Wahrscheinlichkeitsdichte integriert die Fehlerwahrscheinlichkeit des Tests für Falsch-

Negative angibt – und somit seine Stärke, jeweils für den unteren und den oberen Nonzentralitätsparameter.

Bei der Interpretation eines nonzentralen Konfidenzintervalls ist deren Zusammenhang mit der Größe eines Effektes zu bedenken, insbesondere in der Hinsicht, dass das asymmetrische Intervall sich zu kleineren Effektgrößen hin verengt und ganz verschwindet, wenn kein Effekt vorliegt, obwohl man von einem Konfidenzintervall erwarten sollte, dass seine Weite unabhängig ist vom Wert der eingeschlossenen Punktschätzung, wie hier für die Teststärke (Lecoutre & Poitevineau 2014, S.84).

Der Nonzentralitätsparameter der *F*-Verteilung mit den Freiheitsgraden ν_1 und ν_2 sollte nach Thomas (1997) zudem nicht aus dem Originaldatensatz bestimmt werden; sonst läuft die Berechnung hinaus auf eine bloße Reformulierung von dessen statistischer Signifikanz. Angesichts der Effektgrößen, die in der Literatur zum Neuro-Effekt genannt werden, liegt $d_{priori}=0.2$ ($r_{priori}=0.1$) nahe zur Schätzung des Nonzentralitätsparameters

$$\lambda_F=\left(\frac{r^2_{priori}}{1-r^2_{priori}}\right)\cdot\nu_2\approx\frac{\sum_i\left(Y_i-\overline{Y}\right)^2}{\hat{\sigma}^2}=\frac{657.61}{2.99}=219.64 \quad .$$

Für λ_F lässt sich ein 90 Prozent-Konfidenzintervall berechnen aus dessen zentraler χ^2-Verteilung mit ν Freiheitsgraden, die sich nach Taylor und Muller (1995) überführen lässt in eine *F*-Verteilung mit $\nu_1=1040-2\times4$ (Anzahl der verschiedenen Items) und $\nu_2=2\times4-1$ Freiheitsgraden, sodass sich für den Nonzentralitätsparameter eine untere Schranke $\hat{\lambda}_u=167.90$ und eine obere Schranke $\hat{\lambda}_o=264.73$ ergeben.

Aufgrund der Monotonie nonzentraler *F*-Verteilungen lässt sich aus dem Konfidenzintervall für den Nonzentralitätsparameter eine Konfidenzintervall für die Teststärke angeben unter Verwendung der Wahrscheinlichkeiten $1-F\left(F_{(1-\alpha;252;7)_{krit}} \mid \hat{\lambda}_u\right)$ und $1-F\left(F_{(1-\alpha;252;7)_{krit}} \mid \hat{\lambda}_o\right)$. Somit liegt für $\alpha=0.1$ die Teststärke bei direkten Replikationen mit ebenfalls 260 Versuchspersonen langfristig in neun von zehn Fällen zwischen *0.574* und *0.701*, was den weit höheren Stichprobenumfang bei einer angestrebten Teststärke von *0.90* erklärt.

Als Erklärung für den großen Stichprobenumfangkommt die ungleiche Größe der Neuro-Cluster hinzu. Gegenüber einem balancierten Design ist das Cluster ohne Neuroinformation um das 18/26-fache kleiner. Bei einer solchen Verkleinerung der Kontrollgruppe muss nach Lipsey (1990, S.140) die Versuchsgruppe $u=\left(2-\frac{1}{\sqrt{18/26}}\right)^{-2}=1.570$ mal größer ausfallen, damit die Teststärke unverändert bleibt.

Da kein zwingender Grund ersichtlich ist, weshalb in der Replikation ebenfalls ein Neuro-Cluster-Größenverhältnis von 17:9 angestrebt werden soll, um nachzuweisen, dass die Items mit Neuroinformation besser beurteilt werden als die Items ohne Neuroinformation, kann unter der Annahme eines hypothetischen Verhältnisses von 1:1 die ermittelte Anzahl der Versuchseinheiten im Cluster mit Neuroinformation um das 1,57-fache reduziert und die Studie mit gleichgroßen Clustern repliziert werden.

In einer Versuchsanordnung, in der das Versuchscluster die Hälfte der Stichprobe ausmacht, werden demnach $\frac{1935}{1.57}$ Items, also 308 Versuchspersonen benötigt. Die Replikation ist folglich mit gleichgroßen Gruppen vorzunehmen, von denen jede aus 308 Versuchspersonen besteht, sodass insgesamt 616 Personen an der Studie teilnehmen müssen, um bei einem Signifikanzniveau von 10 Prozent und einer Teststärke von 90 Prozent den von Weisberg et al. (2015) beschriebenen Effekt nachweisen zu können. Wegen

$$t_{(1-\alpha/2;k-2)}=\frac{w}{2}\cdot\sqrt{\frac{(k-2)\cdot n\cdot k\cdot p\cdot(1-p)}{\chi^2_{(1-\alpha;k-2)}\cdot\left(\sigma^2_{\varepsilon_{ijkl}}+n\cdot\sigma^2_{\upsilon_{0\,jkl}}\right)}}=\frac{1.016}{2}\cdot\sqrt{\frac{14\cdot308\cdot16\cdot0.5\cdot0.5}{21.10\cdot\left(2.72+308\cdot0.18\right)}}=1.91$$

entspricht dieser Stichprobenumfang dem des Präzisionsansatzes bei einem 1:1-Design, sodass bei 616 Teilnehmern die erwartete Effektgröße bei unendlich vielen Wiederholungen mit einer Konfidenz von 92.5 Prozent im Intervall ihrer doppelten Standardabweichung liegt.

Da für die Akquise von 616 Versuchspersonen weder genügend Zeit noch genügend Mittel zur Verfügung stehen, muss die Teststärke von 90 auf 80 Prozent abgesenkt werden. Dann ist ein Standardfehler von $\sigma_n=0.120$ anzusetzen, welcher 380 Items je Neurocluster erforderlich macht. Somit werden für die Replikation im Cluster mit

Neuroinformation 96 Versuchspersonen benötigt und im Cluster ohne Neuroinformation ebenfalls 96 Versuchspersonen, insgesamt also 192 Versuchspersonen.

Die Reduktion des geplanten Stichprobenumfangs geht auf Kosten der Präzision: die Schätzwerte für den Effekt aus unendlich vielen Replikationen wird nun mit einer Konfidenz von 71.5 Prozent überdeckt vom Intervall, das doppelt so breit ist wie der Schätzwert für den Effekt aus der Originalstudie. Würde man für die Replikation eine Sicherheits-Teststärke (Perugini, Gallucci & Constantini 2014) anstreben, der das zehnte Percentil eines 80 Prozent-Konfidenzintervalls der Effektgröße zugrunde liegt, um abzusichern, dass für 90 Prozent aller direkten Replikationen der Stichprobenumfang ausreicht für eine Replikationsteststärke von 80 Prozent, wären sagenhafte 39 276 Versuchspersonen erforderlich.

Die Erfolgsaussichten für einen Nachweis des Effektes werden zudem getrübt durch die geringe Wahrscheinlichkeit dafür, dass die Mittelwertdifferenz der Replikation auch nur dasselbe Vorzeichen hat wie die Mittelwertdifferenz der Originalstudie. Nach Killeen (2005) wird diese Wahrscheinlichkeit angegeben durch:

$$P_{rep}=\frac{\delta}{\sigma_{rep}} \quad \text{, mit} \quad \sigma_{rep}=\sqrt{2\sigma_{\delta}^2+\sigma_{\delta}^2} \quad ,$$

wobei σ_δ dem Umstand Rechnung trägt, dass die Größe des Effektes in der Population variiert, was hierarchisch als Zufallsschwankung υ_{0ikl} modelliert ist, sodass $\sigma_\delta^2=(\sigma_{\upsilon_{0jkl}}^2)_{nurAbschnitt}$ ist und unter identischen Bedingungen

$$P_{rep}=\frac{0.170}{\sqrt{2\cdot 0.158+0.021}}=0.530 \quad .$$

Die höchste Wahrscheinlichkeit einer Replikation des Vorzeichens, die sich mit der vorliegenden Effektgröße und -varianz erreichen lässt, ist 66.6 Prozent, wofür

$$n=\frac{z_{(p_{rep})}^2}{\delta^2-2\sigma_\delta^2 z_{(p_{rep})}^2}=2138$$ Items, also 536 Versuchspersonen nötig wären.

Dass die Replikationswahrscheinlichkeit *p=0.53* kleiner ist als die Wahrscheinlichkeit *1−β=0.8*, einen Effekt zu finden, sofern er in der Population vorhanden ist, liegt an der Größe des Standardfehlers des Effektes, der so groß ist wie der Effekt selbst. Das hat bei einem Konfidenzniveau von 0.9 zur Folge, dass fast ein Drittel der konfidenten Schätzungen ein negatives Vorzeichen besitzt. Zieht man dieses Drittel vom Konfidenz-

niveau ab, so kommt man grob auf eine Replikationskonfidenz von 0.61. Die Replikationswahrscheinlichkeit ist also kleiner als die Teststärke, weil bei gleicher Varianz das Intervall günstiger Ereignisse schmaler ist.

Sieht man von Varianz und Konfidenzintervall der Effektgröße ab und beschränkt sich auf das Konfidenziniveau des Nicht-Verwerfens der Nullhypothese bei der Überprüfung des Effekts, wird die Teststärke der geplanten Replikation restituiert. Dann gilt nach Greenwald, Gonzalez und Harris (1996) für die Wahrscheinlichkeit der Replikation zur in der Replikation erwünschten Konfidenz von 90 Prozent:

$$p_{rep} = 1 - P\left(z_\alpha \leqslant \frac{t_{krit} - t}{\sqrt{1 + \frac{t_{krit}^2}{2 \cdot df}}}\right) = 1 - P\left(z_\alpha \leqslant \frac{1.645 - 2.675}{\sqrt{1 + \frac{1.645^2}{2 \cdot 1037}}}\right) = 0.849 \quad .$$

Betrachtet man die Wahrscheinlichkeit P_S dafür, dass die Beurteilung eines zufällig gezogenen Items mit Neuroinformation besser ausfällt als die Beurteilung eines zufällig gezogenen Items ohne Neuroinformation, kommt man zu konsistenten Werten, auch wenn sie sich im Betrag etwas unterscheiden. Folgt man dem Ansatz des allgemeinen Sprachgebrauchs von McGraw und Wong (1992), so ist diese Wahrscheinlichkeit *0.54* aufgrund

$$z_{P_s} = \frac{\overline{Y}_{mit} - \overline{Y}_{ohne}}{\sqrt{s_{mit}^2 + s_{ohne}^2}} = \frac{0.288 + 0.033}{\sqrt{3.796 + 3.202}} = 0.121 \quad .$$

Greift man dagegen zurück auf die Statistik *U* von Mann-Whitney, die hier angibt, wie häufig Beurteilungen der Items mit Neuroinformation in einer geordneten Rangfolge vor denen der Items ohne Neuroinformation stehen – wobei eine bessere Beurteilung einen höheren Rang bedeutet –, erhält man nach Grissom (1994) für κ mögliche Rangunterschiede die Wahrscheinlichkeit

$$P_{S'} = \frac{U}{\kappa} = \frac{U}{n_{mit} \cdot n_{ohne}} = \frac{108645}{680 \cdot 360} = 0.44 \quad \text{, sodass } P_S = 1 - P_{S'} = 0.56.$$

Die geringe Wahrscheinlichkeit, bei Items mit Neuroinformation eine bessere Beurteilung anzutreffen, zusammen mit einer Replikationswahrscheinlichkeit, die ebenfalls kaum größer als ein Münzwurf ist, kontrastieren die Zuversicht der Autoren, die auf Nachfrage davon ausgehen, dass der Nachweis des Effektes – bei einer Teststärke von *0.8* – mit einer Wahrscheinlichkeit von *0.75* gelingen wird.

Aufgrund des geringen *p*-Wertes von *0.008* für den Regressionskoeffizienten der Neuroinformation stehen die Chancen noch schlechter, mit der Replikation ein ebenso signifikantes Ergebnis zu erzielen. Bei einem *p*-Intervall für Replikationen mit 90 Prozent Konfidenz *[0; 0.085]* ist nach Cumming (2008) die Wahrscheinlichkeit, in der Replikation einen *p*-Wert zu erhalten, der höchstens so groß ist wie der *p*-Wert der Originalstudie, gegeben durch

$$P_p = 1-\Phi\left(d\sqrt{\frac{N}{2}}+\Phi^{-1}(1-\frac{p}{2})\right)+\Phi\left(d\sqrt{\frac{N}{2}}-\Phi^{-1}(1-\frac{p}{2})\right)$$

$$=1-\Phi\left(0.17\cdot\sqrt{\frac{192}{2}}+2.652\right)+\Phi\left(0.17\cdot\sqrt{\frac{192}{2}}-2.652\right)=0.383 \quad .$$

4.3 Direkte Replikation

Die zu Weisberg et al. (2015) hinterlegten Items wurden aus dem Englischen ins Deutsche übertragen, nach dem von Hopkins bereitgestellten Muster graphisch umgesetzt, mit Effektgröße und geplantem Stichprobenumfang aus der Reanalyse in einem Replikationsvorhaben zusammengefasst und auf der Plattform des OSF vorabregistriert. Der vorgesehene Erhebungszeitraum von vier Wochen sollte verkürzt werden können, sobald absehbar sein sollte, dass die angestrebte Anzahl von 192 Versuchspersonen erreicht wurde.

Teilnehmen konnte an der online-Studie jedwede Person mit Internetzugang – nur einmal unter einer IP. Die Teilnehmer wurden eingestimmt auf interessante Befunde aus der Psychologie und darauf hingewiesen, dass mit der Studie die Qualität der zugehörigen Erklärungen erforscht werden solle. Nach Abschluss ihrer Eingaben wurden die Teilnehmer darüber aufgeklärt, dass sie entweder Erklärungen mit oder ohne Neuroinformation zur Beurteilung erhalten hatten. Studierende der Psychologie bekamen für die Teilnahme 0.5 Versuchspersonenstunden gutgeschrieben, Nicht-Studierende konnten an der Verlosung von drei Einkaufsgutscheinen in Höhe von zehn Euro teilnehmen.

Im Unterschied zu Studie3 von Weisberg et al. (2015) wurde das Merkmal 'Neurologischer Jargon' ausgespart, sodass insgesamt ein 2 (Neuroinformation: ohne, mit) x 2 (Sample: Studierende, Nicht-Studierende) x 2 (Qualität der Erklärung: schlecht, gut)

Design vorliegt für vier Items, deren Beurteilung (−3 bis +3) als abhängige Variable firmiert. Weil im Sample kulturbedingt keine Crowdworker zu erwarten waren, wurde die Kategorie erweitert auf Nicht-Studierende. Im Studienverlauf des Originals wurde zudem die jeweilige Erklärung erst 10 Sekunden nach der Phänomenbeschreibung eingeblendet; in der Replikation dagegen bekamen die Teilnehmer Phänomen samt Erklärung gleichzeitig eingeblendet, sie konnten aber ihre Beurteilung erst nach 10 Sekunden abschicken.

4.3.1 Ergebnisse

Die Stichprobe der Replikation umfasst 244 Teilnehmer. Nach drei Wochen waren 346 Abrufe der Studie registriert. 102 Einträge mussten gelöscht werden: 49 Artefakte ohne Beurteilungen; 43 Teilnehmer hielten nicht bis zum Ende durch. 75 haben beim Aufmerksamkeitstest nicht die dort verlangte Beurteilung abgegeben, bei 66 von ihnen ist allerdings aus den Erklärungen ersichtlich, dass sie den Testbefund gelesen hatten („Ich sehe es nicht ein, eine vorgegebene Antwort zu wählen. Ich bin aufmerksam und entscheide selbst.“). Eine Person hat den richtigen Wert angegeben, allerdings zu keinem Befund eine Erklärung abgegeben und auch keine demographischen Angaben gemacht.

Von den 244 in der Stichprobe verbliebenen Versuchspersonen sind 159 Studierende (114 weiblich, 45 männlich; Durchschnittsalter 19.3 Jahre, Spannweite 19-58 Jahre) und 85 Nicht-Studierende (48 weiblich, 36 männlich 1 sonstiges; Durchschnittsalter 48,2 Jahre, Spannweite 19-77 Jahre). Insofern sorgte im wesentlichen die größere Lebensreife von Fernuniversitätsstudierenden dafür, dass die Teilnehmer durchschnittlich 11 Jahre älter sind als in der Originalstudie.

Im 4-Ebenen-Modell mit Neuroinformation und Qualität auf derselben Ebene, beide eingebettet in die Items, die eingebettet sind in die ins Sample eingebetteten Versuchspersonen, wurde für die Gemischte Lineare Regression wie im Original ein 2-Ebenen-Cluster-Modell aufbereitet mit Zufallsachsenabschnitt und -steigung auf der Qualitätsebene, bezogen auf die Ebene der Versuchspersonen. Im Unterschied zum Original wurden alle Variablen effektcodiert mit Item1, ohne Neuroinformation, Studierenden und schlechten Erklärungen als Referenzkategorien, sodass der Achsenabschnitt den

Gesamtmittelwert und die Regressionskoeffizienten jeweils die Abweichung vom Gesamtmittelwert angeben.

Für einen Effekt des Geschlechts fanden sich varianzanalytisch wie im Original keine Hinweise. Die Regressionsanalyse erbrachte signifikante Effekte für Neuroinformation (*t (974)=1.86, p=0.06*), Qualität (*t=4.76, p<0.01*) und Sample (*t=3.51, p=0.01*). Interaktionseffekte fanden sich zwischen Items und Qualität, nicht aber zwischen Items und Neuroinformation. Eine Gegenüberstellung von Original, Reanalyse und Replikation findet sich in Tabelle 2.

	Original		Reanalyse		Replikation	
Prädiktor	β	[95% CI]	β	[95% CI]	β	[95% CI]
Konstante	0.02	[–0.19, 0.24]	0.12*	[0.00,0.24]	0.01*	[–0.73;0.09]
Item2	0.33*	[0.06, 0.62]	0.23	[0.05,0.41]	0.41*	[0.23;0.60]
Item3	–0.91*	[–1.24,-0.62]	–1.01*	[–1.20,–0.82]	–1.57*	[–1.76;–1.38]
Item4	0.98*	[0.69, 1.25]	0.88*	[0.71, 1.05]	1.17*	[1.00;1.35]
Neuroscience	0.27*	[0.05, 0.51]	0.16*	[0.04, 0.28]	0.11*	[-0.01;0.23]
Group	0.15*	[0.05, 0.27]	0.15*	[0.04, 0.26]	0.23*	[0.10;0.35]
Qualität	0.57*	[0.35, 0.78]	0.38*	[0.27, 0.48]	0.27*	[0.16;0.39]
Neuro*Item2	–0.37*	[–0.68,-0.07]	–0.26*	[–0.44,-0.08]	–0.08	[–0.27;0.10]
Neuro*Item3	–0.04	[–0.36, 0.27]	0.08	[–0.11, 0.26]	0.09	[–0.09;0.28]
Neuro*Item4	–0.05	[–0.38, 0.28]	0.06	[–0.11, 0.23]	0.09	[–0.08;0.27]
Quality*Item2	0.12	[–0.18, 0.44]	0.33*	[0.15, 0.51]	0.33*	[0.15;0.52]
Quality*Item3	–0.39*	[–0.67,-0.07]	–0.19*	[-0.38,-0.01]	–0.28*	[–0.48;-0.10]
Quality*Item4	–0.57*	[–0.85,-0.25]	–0.35*	[–0.51,-0.18]	–0.19*	[–0.37;-0.17]

*Tabelle 2: Gemischtes Modell für Lineare Regression: Koeffizienten geben Abweichung vom Gesamtmittelwert an. Nota bene die verschiedenen Alpha-Niveaus (Original * p<0.05, Replikation * p<0.1)!*

Gute Erklärungen (*M=0.25, SD=2.02*) wurden besser beurteilt als schlechte (*M=−0.35, SD=2.02*), ebenso wie Erklärungen mit Neuroinformation (*M=0.04, SD=2.06*) besser beurteilt wurden als Erklärungen ohne Neuroinformation (*M=−0.15, SD=2.04*), vgl. Abbildung 1; Studierende (*M=0.24, SD=2.08*) gaben insgesamt höhere Beurteilungen ab als Nicht-Studierende (*M=−0.18, SD=2.02*).

Ergänzt man die erhobenen Beurteilungen um normalverteilte Zufallsdaten mit dem Erwartungswert und der Standardabweichung der Stichprobe, jeweils entsprechend der Kombination aus Neuroinformation und Qualität, so als hätten die Versuchspersonen dasselbe Item einmal mit und einmal ohne Neuroinformation zur Beurteilung vorgelegt bekommen – wodurch der Datensatz sich verdoppelt, kann inferenzstatistisch geprüft werden, ob der Übergang von schlechten Erklärungen ohne Neuroinformation zu schlechten Erklärungen mit Neuroinformation signifikant verschieden ist zum Übergang von guten Erklärungen ohne Neuroinformation zu guten Erklärungen mit Neuro-information. Das ist bei $F(1, 1950)=2.33$ und $p=0.13$ nicht der Fall.

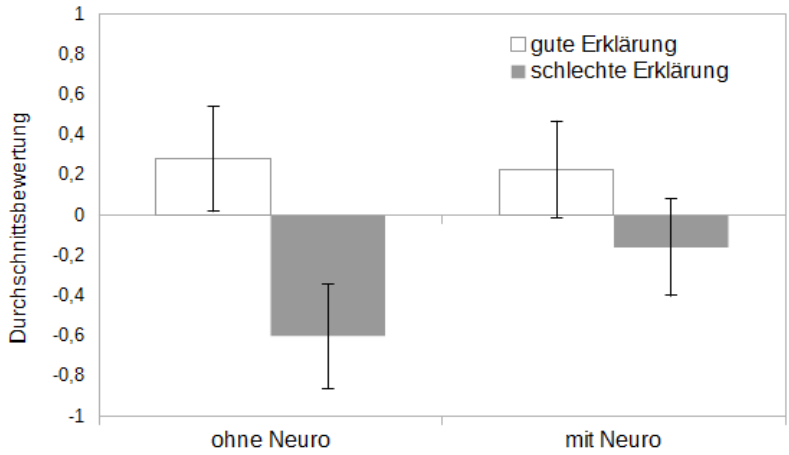

Abbildung 1: Durchschnittliche Beurteilung nach Neuroinformation und Qualität mit 95 Prozent-Konfidenzintervall.

Separate Regressionsanalysen für jedes Item im Standardmodell führten nur bei Item3 zu vergleichbaren Koeffizienten (Tabelle 3); Neuroinformation trug nicht signifikant bei zur Beurteilung der Erklärung von Item1, dafür aber das Sample zur Beurteilung der Erklärung von Item4; letzteres hat bei Item2 ein umgekehrtes Vorzeichen.

Die Resultate verzerren könnte der Ersteffekt, wenn die Qualität der ersten Erklärung die Beurteilung der folgenden Erklärung determinierte. Die Beurteilungen der zweiten Erklärung von Personen, deren erste Erklärung gut war (*M=0.37, SD=1.83*), sind allerdings nicht signifikant verschieden zu denen von Personen, deren erste Erklärung schlecht war (*M=0.32, SD=1.93*). Auch ein Decken- oder Bodeneffekt ist nicht feststellbar: aus den 976 Items lassen sich für aufeinanderfolgende Items 732 Paare bilden, von denen jeweils 163 Paare in der Qualitätskombination der Erklärungen übereinstimmen – von den 163 Paaren nun, in denen eine schlechte Erklärung auf eine gute folgte, folgte nur fünfmal (3.1 Prozent) eine −3-Beurteilung auf eine −3-Beurteilung, und von den 163 Paaren, in denen eine gute Erklärung auf eine schlechte folgte, folgte nur zweimal (1.2 Prozent) eine +3-Beurteilung auf eine +3-Beurteilung, sodass weder die Niederstbeurteilung einer guten Erklärung einen Boden, noch die Höchstbeurteilung einer schlechten Erklärung eine Decke markiert, die die durchschnittliche Beurteilung spürbar verzerrt hätte.

	Original				Reanalyse				Replikation			
	Item1	Item2	Item3	Item4	Item1	Item2	Item3	Item4	Item1	Item2	Item3	Item4
Neuro	0.27*	-0.09	0.24*	0.22*	0.28*	-0.10	0.24*	0.22*	0.00	0.02	0.21*	0.21*
Qualität	0.55*	0.69*	0.17	-0.01	0.57*	0.71*	0.18	0.03	0.42*	0.60*	-0.00	0.10
Sample	0.04	0.32*	0.18	0.08	0.03	0.32*	0.18	0.08	0.25	-0.07	0.33	0.55*

Tabelle 3: Ein-Item-Modell für lineare Regression: Koeffizienten geben Abweichung an vom Gesamtmittelwert (p<0.05).*

Im Versuchscluster mit Neuroinformation bezogen sich 38.4 Prozent der abgegebenen Beurteilungen auf die Neuroinformation; am häufigsten bei Item1 (26x), am seltensten bei Item3 (15x). 71.0 Prozent der Bezüge auf die Neuroinformation waren positiv konnotiert. Das Sample unterschied sich signifikant (*t (202) = 4.19, p<0.01*) hinsichtlich der positiven Bezugnahme auf Neuroinformation, nicht aber hinsichtlich der Häufigkeit von neurologischen Begriffen in den Begründungen insgesamt: Studierende haben die Erklärungen nicht nur als besser beurteilt, sie haben in ihren Begründungen auch häufiger positiv auf neurologische Begrifflichkeiten Bezug genommen als Nicht-Studierende, ohne insgesamt häufiger von ihnen Gebrauch zu machen. Da in der Mehrebenenanalyse nur bei Studierenden der Effekt der Neuroinformation signifikant (*t (627)=1.78, p=0.076*) ist, geht der Effekt in der Replikation zurück auf die Studierenden, die auf das Vokabular der Neurologie besser ansprachen.

Der in der Reanalyse beschriebene Weg zur Effektgröße kann in der Replikation nicht beschritten werden, weil die für den Einfluss der Ebenen maßgebliche Zufallsvarianz im Modell, das nur aus dem Achsenabschnitt besteht, $\left(\sigma^2_{\upsilon_{0\,jkl}}\right)_{nurAbschnitt} = 5 \cdot 10^{-15}$ ist, woraus sich für den Faktor der Neuroinformation keine – vom Nulleffekt verschiedene – Effektgröße ableiten lässt.

Aus dem Mittelwertvergleich folgt für *t (974)=1.434* die Effektgröße *d=0.09;* weil aber Mittelwertvergleiche in Gemischten Modellen die Größe überschätzen, ist sie zutreffender aus der punktbiserialen Korrelation $r_{pb}=0.03$ zu schätzen, die, korrigiert um die Intraklassenkorrelation ein $\hat{\delta}_r = 0.06$ ergibt mit der Varianz

$$\sigma^2_{\hat{\delta}_r} = \left(\frac{N}{n_{mit} \cdot n_{ohne}}\right) \cdot \left(\frac{1 + (\hat{n} - 1) \cdot ICC_3}{1 - ICC_3}\right) + \frac{\hat{\delta}^2_r}{2(N - M)} = 0.0042 \cdot \frac{1 + 476.7 \cdot 0.0924}{0.9076} + \frac{0.06^2}{2(976 - 2)} = 0.158 \quad .$$

Das zugehörige 90 Prozent-Konfidenzintervall lautet folglich:

$0.06 \pm 1.28 \cdot 0.46 = [-0.53; 0.66]$.

Schließlich lässt sich mittels *t*-Test prüfen, ob die Mittelwerte sich signifikant unterscheiden, weil die Varianzen der Originalstudie und der Replikation ganz offensichtlich homoskedastisch sind. Das ist mit *p=0.74* klar nicht der Fall:

$$t(2014) = \frac{\Delta\overline{Y}_{Orig} - \Delta\overline{Y}_{Repl}}{\hat{\sigma} \cdot \sqrt{\frac{1}{N_{Orig}} + \frac{1}{N_{Repl}}}} = \frac{0.32 - 0.19}{2.66 \cdot \sqrt{\frac{1}{1040} + \frac{1}{976}}} = 0.54, \text{ mit } \hat{\sigma} = \sqrt{\frac{(N_{Orig}-1)\hat{\sigma}^2_{Orig} + (N_{Repl}-1)\hat{\sigma}^2_{Repl}}{N_{Orig} + N_{Repl}}} = 2.665 \quad .$$

Auf demselben Weg kann allerdings nicht geprüft werden, ob die Effektgrößen aus Original und Replikation identisch oder auch nur äquivalent sind, weil Effektgrößen, entgegen ihrem Namen, keine Größen vorstellen. Vergleichbare Größen sind Effektgrößen nur dann, wenn ihre Standardfehler einen konstant proportionalen Bestandteil der Größe ausmachen, d.h. wenn sie in den Studien im gleichen Maße in den gemessenen Effekt eingehen. Das ist so gut wie nie der Fall, weil die Faktoren, die den Effekt beeinflussen, nicht dieselben sind wie die Faktoren, die den Standardfehler beeinflussen. Effektgrößen sind daher eher zu verstehen als Maß für die Nachweisbarkeit eines Effekts (Baguley 2012, S.240); so können kleine Effekte ein großes δ haben, wenn sie so leicht nachweisbar sind wie kleine Steine in klarem Wasser.

	Original		Reanalyse		Replikation	
	N=270	[95% CI]	N=260	[90% CI]	N=244	[90% CI]
$\overline{Y}_{mit} - \overline{Y}_{ohne}$	0.32	[0.21; 0.43]	0.32	[0.22; 0.42]	0.19	[0.08; 0.30]
$\sigma^2_{\varepsilon_{ijk}}$	---		2.6625	[2.40; 2.92]	2.6936	[2.43; 2.96]
ICC_1	---		0.0044		0.0000	
ICC_2	---		0.0744		0.0840	
ICC_3	---		0.0272		0.0924	
R^2	---		0.2615		0.2863	
p	<0.05		0.01		0.06	
Effektgröße	---		0.17	[-0.33; 0.67]	0.06	[-0.53; 0.66]
Teststärke	---		0.62	[0.57; 0.70]	0.38	[0.35; 0.42]

Tabelle 4: Kenngrößen der Stichproben, α=0.1.

4.3.2 Diskussion

Zur Beurteilung des Replikationserfolges ist in einem ersten Schritt auf die Determinanten einer direkten Replikation einzugehen und der Grad ihrer Einhaltung im Zusammenhang mit möglichen Moderatoren im Modell der Originalstudie zu diskutieren. Sodann kann die Replikation in Beziehung gesetzt werden zur Originalstudie hinsichtlich der Reproduzierbarkeitsprädiktoren. Die konkrete Beziehung wird im letzten Abschnitt in den Kontext gestellt von Replikationen im allgemeinen.

4.3.2.1 Zur Replizierbarkeit der Originalstudie

Dem Anspruch, dass die einzigen Unterschiede zum Original unvermeidbare Unterschiede sein sollten (Brandt et al. 2012), konnte strenggenommen nicht Folge geleistet werden. Die simultane Einblendung von Phänomen und Erklärung wäre technisch vermeidbar gewesen, sie dürfte aber den Erfolg der Replikation ebenso wenig beeinflusst haben, wie unterschiedliche Rechnerleistung, Bildschirmauflösung, Betriebssysteme oder Browser (Plant 2016, Steenbergen & Bocanegra 2016). Das entscheidende Erfolgsmoment setzt früher an: an der Effektgröße.

Die standardisierte Effektgröße suggeriert im Namen eine stabile Vergleichbarkeit von Effekten; eine Stabilität, die ausgerechnet aus einem Maß für die Variation erwachsen soll: dem Standardfehler. Der Standardfehler bemisst die Variation in einer Stichprobe, unterliegt aber, wie andere Streuparameter auch, Schwankungen, die nicht zuletzt von der Individualität einer Stichprobe herrühren. Alles, was den Standardfehler beeinflusst, aber nicht die Standardabweichung, verzerrt die Schätzung der Effektgröße (Lecoutre & Poitevineau 2014, S.71). Und davon gibt es eine ganze Menge (Youden 1969). Jeder Versuch, Effekte als feste Größen zu normieren oder auch nur zu interpretieren, führt daher nur in eine weitere metrische Verirrung (Thompson 2001).

Die Vielfalt der Standardabweichungen, welche sich in der Reanalyse dargeboten hatte, verdeutlicht gerade an den abweichenden Beträgen, deren Schätzung stärker vom zugrundegelegten Modell abhängt als von der Kontingenz der Stichprobe, wie wichtig neben der Angabe der Größe die Angabe ihres Zustandekommens ist. Eine Angabe, die umso dringlicher ist, je weiter verzweigt die Ursachenketten eines Effektes sind – wie im Hierarchischen Modell. Wo neben den Versuchspersonen nicht nur der Änderungs-

anteil der Items variiert, sondern auch deren fester Anteil am Gesamtmittelwert und der zufällige Fehler selbst (Judd et al. 2012), müssen sowohl die Ebenen als auch ihre Einbettung äußerst genau beschrieben sein (Courgeau 2003, S.209).

Das ist nicht erfolgt. Nicht einmal die Anzahl der Ebenen wird von Weisberg et al. (2015) genannt, geschweige denn vorgenommene Gruppierungen entlang bzw. quer zu den Ebenen. Mithin fehlen Angaben zu Intraklassenkorrelationen und Varianzaufklärung. Auch die Residuenvarianz des nur aus dem Achsenabschnitt bestehenden Modells wäre hilfreich gewesen. Schließlich hätte eine Differenzierung der Alternativhypothese zum Neuro-Effekt nach Ebenen erhellend sein können; doch wird weder modelliert, wie die Merkmale der Neuroinformation auf ihrer Ebene wirken, noch was davon auf der Individualebene ankommt – welche intrinsische Plausibilität die verschiedenen Erklärungen haben könnten, wird nicht thematisiert. Damit mangelt es der Studie schon an elementaren Komponenten zur Rekonstruktion eines Gemischten Modells (Langer 2003).

Gelingt die Rekonstruktion des Modells dennoch, kann die Reanalyse nur noch an der Codierung scheitern. Und sie muss scheitern, wenn im Gemischten Modell auch noch Dummy- und Effektkodierung gemischt werden. Was der Rechner anstandslos verarbeitet, ist im Ergebnis keiner sinnvollen Interpretation mehr zugänglich. Die Koeffizienten können nicht mal die Abweichung vom Kategorienmittelwert, mal die Abweichung vom Mittelwert der Referenzkategorien verkörpern. Sind die Items effektcodiert, dann müssen es auch die übrigen unabhängigen Variablen sein, sonst verliert der Achsenabschnitt seinen Sinn (Eid et al. 2015, S.682).

Nur bei einheitlicher Effektcodierung repräsentiert der Achsenabschnitt den Gesamtmittelwert. Wie aus Tabelle 2 hervorgeht, liegen im Unterschied zur Originalstudie (*M=0.18; SD=1.90*) Reanalyse und Replikation (*M=−0.04; SD=2.05*) sehr nahe am jeweiligen Gesamtmittelwert. Die Werte stimmen nicht ganz überein, weil in der Mehrebenenanalyse die Koeffizienten der Ebenen-Merkmale in wechselseitiger Abhängigkeit bestimmt werden, wodurch ein 'abhängiger' Gesamtmittelwert zustandekommen kann: der Koeffizient auf der dritten Ebene drückt den Effekt der Neuroinformation auf den Gesamtmittelwert aus, insofern beide Neurocluster sich nicht unterscheiden in den Ausprägungen der übrigen Prädiktoren – also nicht unter der allgemeineren Voraussetzung,

alle anderen Prädikatorausprägungen einfach nur unverändert zu lassen (Agresti 2015, S.10).

Bei der Beurteilung der Erklärungen mit oder ohne Neuroinformation könnte, wie eingangs beschrieben, ein Verstehensgefühl (Trout 2002) leitend sein, das metakognitiv deren Richtigkeit signalisiert (Thompson, Turner & Pennycook 2011). Das Verstehensgefühl klingt ab bei kognitiver Dissonanz (Klaczynski 2000), und das langsame Denken (Kahneman 2011, S.21) dringt durch. Damit ist jedoch nur wiedergegeben, wie die Versuchspersonen zu ihrem Urteil gelangen, nicht aber, warum sie gerade ein bestimmtes Urteil fällen, und bspw. Erklärungen mit Neuroinformation besser beurteilen als dieselben Erklärungen ohne Neuroinformation. Die Gründe dafür suchen Psychologen in den Individuen und Epistemologen in den Erklärungen.

Rhodes, Rodriguez und Shah (2014) koppelten Neuroinformation an Erklärungen zu strittigen Themata, woraufhin der Neuro-Effekt verschwand, sobald die Erklärung der Einstellung der Versuchsperson zuwiderlief. Insofern könnte das Wünschenswerte via Einstellung den Effekt moderieren, gestützt von hartnäckigen Vorurteilen (Lord, Ross & Lepper 1979) und dem erworbenen Wissenshorizont (Klahr, Dunbar & Fay 1990, S.384). Die Neuroinformation könnte als Primer ein Vorwissen aktivieren, das zur Beurteilung der Erklärungen herangezogen wird (Harp & Mayer 1998) und so den Eindruck vermitteln, ein Phänomen besser verstanden zu haben (Rhodes et al. 2014). Demnach mangelt es in der Psychologie nicht an Theorievorschlägen, es fehlt lediglich am Willen zum Modell, an der Disziplin, die Vorschläge zu einem nomologischen Netz zu verweben (Cronbach 1984, S.149), statt sie in einer Regression zu einer beziehungslosen Kolonne von Koeffizienten nacheinander aufzufädeln.

Die Erklärungen ihrerseits könnten vom Schein der Wissenschaftlichkeit zehren oder durch ihre bloße Form oder innere Logik bestechen. Würde jegliche wissenschaftliche Information psychologische Erklärungen aufwerten, könnte man nicht von einem Neuro-Effekt sprechen. Fernandez-Duque et al. (2015) maßen allerdings unter den Wissenschaften für die Neurologie den größten Effekt. An Weisberg et al. (2015) kritisieren sie die Form der Erklärungen. Obwohl die Länge der Erklärungen als eigenständiger Effekt erwiesen und effektiv kontrolliert worden sei, so würde doch durch den Einbau der Neuroinformation in die Textmitte der Lesefluss so stark unterbrochen, dass die Zirkularität der schlechten Erklärungen verblasse.

Zirkuläre Erklärungen sind nicht notwendig schlecht. Es gibt virtuose Zirkel und vitiöse Zirkel und beliebige Abstufungen dazwischen. Virtuose Zirkel verlaufen über die Prämissen zur Konklusion und – rückversichernd – zurück zu den Prämissen, wie beispielsweise in den Rückkopplungsschleifen der Lernprozesse einer Künstlichen Intelligenz. Vitiös ist ein Zirkel dagegen, wenn in einer Argumentation die Gültigkeit der Konklusion der Gültigkeit der Prämissen vorausgeht und zur Rechtfertigung der Prämissen vorausgesetzt wird: Elektronen existieren, weil sie in einer Blasenskammer sichtbar gemacht werden können (Hahn 2011).

Eine Tautologie ist, anders als Rips (2002) vermutet, kein Zirkel. Die Argumentationsschritte einer Tautologie bestehen in Äquivalenzumformungen und nicht in Erklärungen. Ebensowenig ist der Satz „Die Blume blüht blau, weil sie blaue Blüten trägt“ zirkulär. Wie die schlechten Erklärungen von Weisberg et al. (2015) hat er die die Form einer Erklärung, ist aber keine Erklärung, sondern ein Pleonasmus. Mit Pleonasmen sind die Wenigsten vertraut, weil sie außerhalb der Poetik so gut wie nicht vorkommen, auch nicht in journalistischen Erzeugnissen, was die Konstruktion der 'zirkulären' Erklärungen in der Studie artifiziell macht und so die Wahrnehmung ihrer Qualität verzerrt im Vergleich zu gewohnten Argumentationsfiguren, guten wie schlechten (Girotto 2009).

Bleibt als letzte Ingredienz noch die Irrelevanz der neurologischen Termini. Diese haben sich Weisberg et al. (2015) von Experten attestieren lassen. In der Tat wird zu den Phänomenen eine Hirnregion als Erklärung angekündigt, im Nachsatz aber eine psychologische Erklärung geliefert: so sorgte bei Item1 das Gehirn dafür, dass die Säuglinge überrascht waren, nur eine Puppe zu sehen, wo sie zwei vermuteten. Für den erklärenden Zusammenhang zwischen Beobachtung und Überraschung ist das Gehirn sicher irrelevant. Man könnte allerdings schon den Bezug zum Gehirn als Erklärung auffassen; als diejenige nämlich, die psychologische Phänomene im Körper lokalisiert (Legrenzi & Umiltà 2009, S.104).

Je mehr es an einer integrativen Theorie fehlt, desto wichtiger wäre die theoretische Absicherung der Versatzstücke, die im statistischen Modell einen erratischen Eindruck hinterlassen (Schmidt & Hunter 2015, S.556). Solange aber offenbleibt, was für gute Erklärungen relevant ist und was nicht, bleibt viel Spielraum für die Interpretation dessen, was die Versuchspersonen beurteilt haben. Das Testergebnis gleicht dann einer

jungfräulichen Empfängnis, die die ganze Replikation samt Kontext infragestellt. Wie soll man falsch-positive Befunde entlarven, wenn das Positive nicht identifizierbar ist? Die vermeintliche Identifikation eines Effektes gerät zur Interpretation statistischer Artefakte. Sind nicht nur das Modell und sein Verwendungszweck vage (Claeskens 2016), sondern sind es auch dessen Grundbegriffe, wird eine Replikation gegenstands- und sinnlos, weil die Fehlertoleranz für falsch-positive Befunde sich beliebig weiten lässt.

Doch nicht nur potentielle Fehler im Modell, auch ganz reale Fehler in Datenanalyse und -vortrag erschweren eine Replikation: in Weisberg et al. (2015) stimmen die Angaben zur Stichprobenumfang nicht nur mit dem hinterlegten Datensatz nicht überein, sondern auch nicht mit dem revidierten Datensatz (80 statt 90 Versuchspersonen in Kontrollgruppe). Dazu passt, dass im Artikel in Studie3 plötzlich die Länge der Erklärungen wieder als Faktor auftaucht, obwohl alle Erklärungen dieselbe Länge haben und ergo die Länge nicht ins statistische Modell eingeht. Kommen fehlcodierte Daten hinzu, wird immer unklarer, was eigentlich repliziert werden soll. Im Grunde handelt es sich bei der vorliegenden Replikation um eine Replikation der Reanalyse – und bei der Reanalyse um eine Sekundäranalyse (Glass 1976)

4.3.2.2 Zur Replikation der Originalstudie

Um die Replikation der Reanalyse des Originals im Kontext seiner Reproduzierbarkeit besser interpretieren zu können, werden kurz Rubriken für den Ausgang von Replikationen samt ihrer Eigenschaften vorgestellt, den Eigenschaften der Replikation gegenübergestellt und anschließend eine Einordnung vorgenommen.

Unterschieden werden durchgängig erfolgreiche von gescheiterten Replikationen. Die gescheiterten Replikationen unterteilen Brandt et al. (2014) hinsichtlich der Signifikanz von Nullhypothese und Effektgrößendifferenz: eine gescheiterte Replikation ist informativ, wenn die Nullhypothese nicht verworfen werden kann oder der Effekt der Replikation signifikant abweicht vom Originaleffekt; eine gescheiterte Replikation ist praktisch, wenn die Nullhypothese verworfen werden kann und der Effekt der Replikation signifikant abweicht vom Originaleffekt; und die gescheiterte Replikation ist

unstimmig, wenn weder die Nullhypothese verworfen werden kann noch der Effekt der Replikation signifikant abweicht vom Originaleffekt.

Eine erfolgreiche Replikation sensu lato liegt vor, wenn der Effekt der Replikation in der Richtung übereinstimmt mit dem Originaleffekt; sensu stricto muss der Effekt der Replikation darüber hinaus signifikant sein (OSC 2012). Für Brandt et al. (2014) muss der Effekt der Replikation auch noch mindestens so groß sein wie der Originaleffekt. Asendorpf et al. (2013) vindizieren dafür das Konfidenzintervall: bei einer erfolgreichen Replikation überlappen die Konfidenzintervalle sich substantiell und das Konfidenzintervall der Originalstudie schließt den Effekt der Replikation ein (Gilbert et al. 2016) bzw. das Konfidenzintervall der Replikation schließt den Originaleffekt ein (Srivastava 2016). Camerer et al. (2015) zählen schließlich noch die Reproduktionserwartung der Beteiligten hinzu.

Ein letztes Replikationskriterium ist die Teststärke. Und zwar die der Originalstudie: fällt der Effekt der Replikation größer aus als die Mindestgröße, die ein Effekt hätte haben müssen, um in der Originalstudie nachweisbar zu sein, gilt sie als erfolgreich. In diesem Fall hätte die Teststärke der Originalstudie 33 Prozent betragen (Simonsohn 2015). Fördert die Replikation einen Effekt von dieser Mindestgröße zutage, wäre er in mindestens zwei von drei Replikationen nachgewiesen worden. Hätte Studie3 von Weisberg et al. (2015) eine Teststärke von 33 Prozent gehabt, wäre die Nachweisschwelle bei einem Alpha-Niveau von 90 Prozent d=0.04 gewesen und damit kleiner als der Effekt der Replikation.

Der Effekt der Replikation ist sowohl gleichgerichtet als auch signifikant, die Replikation somit sensu stricto gelungen. Dass der Effekt des Originals und der Effekt der Replikation sich signifikant ($t(2014)=15.4,\ p<0.01$) unterscheiden, sei hier trotz der großen Interpretationsbandbreite von Effektgrößen nur deshalb berichtet, weil es angesichts der relativ großen Konfidenzintervalle unerwartet ist. Noch unerwarteter wäre der signifikante Unterschied allerdings, wenn in der Population gar kein Effekt existierte. Daher sprächen die unterschiedlichen Effektgrößen nicht gegen den Erfolg der Replikation. Für eine erfolgreiche Replikation spricht jedenfalls, dass die Effektgrößen im jeweils anderen Konfidenzintervall enthalten sind und sich die Intervalle zu 82.5 Prozent überlappen.

Die Reproduktionserwartung von Hopkins ist mit 75 Prozent angesichts der Werte für die Replikationsprädiktoren aus dem vorigen Abschnitt optimistisch, hat sich aber nicht als falsch erwiesen. Der Mangel an theoretischer Fundierung der Studie macht eine fachpsychologische Expertise entbehrlich. Auch Design und Umsetzung sind wenig anspruchsvoll und leicht zu wiederholen. So kamen die Replikatoren denn auch zu der Selbsteinschätzung, dass sie das Design realisieren können, ohne zu wissen, was sie tun.

Nach OSC (2015) liefert eine Meta-Analyse ein weiteres Maß für den Erfolg einer Replikation bzw. für die Existenz des Effekts. Eine Meta-Analyse in Form einer integrativen Datenanalyse (Curran & Hussong 2009) resultiert aus der Zusammenführung beider Datensätze. Poolt man Original und Replikation, erhält man 1240 Items mit Neuroinformation (*M=0.18, SD=1.99*) und 776 Items ohne Neuroinformation (*M=−0.10, SD=1.94*) bei einem Gesamtmittelwert von 0.07 (*SD=1.98*). Daraus resultieren eine Teststärke zwischen 73.6 und 78.4 Prozent und eine Effektgröße $d = 0.07 \pm 0.24$, die wiederum signifikant (*t(2014)=3.14, p=0.02*) ist. Das Konfidenzintervall der Effektgröße umfasst nicht nur den Nulleffekt, seine Weite entspricht zudem dem Siebenfachen der Effektgröße.

Das ist ernüchternd, aber nicht gescheitert. Den einzigen Hinweis für ein Scheitern gibt nach vorstehender Rubrizierung der Umstand, dass der replizierte Effekt kleiner ist als der Originaleffekt. Das ist zu erwarten in einem Wissenschaftssystem mit Veröffentlichungsverzerrung, das den Vortrag zensiert von kleinen oder nicht-existenten Effekten, die die vorgetragenen Effektgrößen auf das Populationsmaß zurückstutzen würden (Lynch 2015). Außerdem besagt der Umstand nicht, dass kein Effekt besteht (Simonsohn 2015).

Bedenkt man das unerschöpfliche Angebot an Variationsmaßen, die große Spannweite des erforderlichen Stichprobenumfangs, der von 192 bis 740 Versuchspersonen reicht und vor allem das breite Konfidenzintervall für die Effektgröße, wird verständlich, wie eine erfolgreiche Replikation derart ernüchternd ausfallen kann. Außer zusätzlichen Daten scheint nichts gewonnen. Wie auch, wenn die Mehrheit (57 Prozent) in der Beurteilung der Erklärungen übereinstimmen, sodass anhand eines Skalenwertes nicht entschieden werden kann, ob die Beurteilung zu einem Item mit oder ohne Neuroinformation gehört? Oder wenn von zufällig gezogenen Beurteilungen diejenigen mit Neuroinformation kaum besser ausfallen (54 Prozent) als diejenigen ohne Neuro-

information bzw. nur 56 Prozent der Beurteilungen mit Neuroinformation auf einer Rangordnung der Beurteilungen vor den Beurteilungen der Kontrollgruppe liegt?

Als zutreffender Prädiktor für den Ausgang der Replikation erwies sich auch die hohe Wahrscheinlichkeit (85 Prozent) für das Verwerfen der Nullhypothese bei einer doch recht geringen Wahrscheinlichkeit (38 Prozent) für einen höchstens gleichgroßen *p*-Wert und bei einer ausgewogenen Wahrscheinlichkeit (53 Prozent) für die Vorzeichengleichheit des Effektes. Letztere ist etwas in Verruf geraten wegen überschätzter Effektgrößen (Iversen, Lee & Wagenmakers 2009): die Wahrscheinlichkeit für Vorzeichengleichheit deckt sich nur mit relativen Häufigkeiten, wenn Effektgröße und Population außergewöhnlich groß sind (Trafimow et al. 2010). Verwendet man zur Berechnung der Replikationswahrscheinlichkeit statt der Effektgröße Konfidenzintervalle, dann haben die eine so große Spannweite, dass sie quasi nichts mehr aussagen (Froman & Shneyderman 2004).

Die Grenzen der vorliegenden Replikation stecken im wesentlichen die großen und uneinheitlichen Varianzen ab, die mangels Modell, das die Variation erklären würde, den Replikationsausgang als bloß statistisches Artefakt qualifizieren. Aussagen zu Existenz und Charakter des Effekts sind mit sehr großen Unsicherheiten verbunden, die noch einmal vergrößert werden durch die Einführung eines statistischen Modells ohne empirisches Vorbild (Claeskens 2016). Demnach mag der Neuro-Effekt existieren; sofern der Effekt existiert, variiert er beträchtlich bis hin zur Umkehr des Vorzeichens (Tabacchi 2016). Die Identifikation und die funktionale Verknüpfung der Faktoren, die Grund sind für diese Variation, brächte die Psychologie einer Theorie näher als die fortgesetzte Korrelation freischwebender Konstrukte.

5 Kleine Methodologie der Replikation

Forschung lässt sich nach Ioannidis (2015) unterteilen in belassene und replizierte. Replikationen fordern Forschung heraus. Auch die Psychologie ist auf herausfordernde Replikationen angewiesen (Cohen 1994). Methodologisch bestätigen sie zuvörderst ein erfolgreich repliziertes Experiment (Finifter 1972) und sind insofern ein wichtiges Kriterium für die Bestätigung einer Theorie (Polio & Gass 1997), sodass in der Forschung nichts unberührt bleiben und ein experimenteller Befund erst nach mehreren Replikationen in den Wissensbestand der Wissenschaft aufgenommen werden darf (Bridgman 1928, Allen 1993; Mackey 2012).

Erfolgreiche Replikationen lassen sich unterteilen in zweifellose und zweifelhafte. Der Erfolg einer zweifelhaften Replikation wird doppelt infrage gestellt. Einmal wird bezweifelt, ob die Replikation tatsächlich erfolgreich war; das andere Mal wird bezweifelt, ob die erfolgreiche Replikation tatsächlich als Bestätigung gelten kann. Daran schließt sich die Frage an, was eine bestätigende, erfolgreiche Replikation bestätigt: Einen Effekt? Eine Hypothese? Oder eine Theorie? Und schließlich: Warum bestätigt eine erfolgreiche Replikation eine Hypothese oder Theorie? – Was nach Simon (1977, S.40) die wichtigere Frage ist.

Der Status von Replikationen in der Theorieentwicklung führt uns auf die Erkenntnistheorie. Im folgenden werden daher in einem epistemologischen Rahmen die methodologischen Anforderungen an Replikationen entwickelt, denen Replikationen genügen müssen, um als Moment einer Selbstkorrektur der Wissenschaft zum Fortschritt derselben beitragen und experimentelle Befunde bestätigen zu können. Bevor im letzten Abschnitt ein Ausblick gegeben wird auf die Bedeutung und Verwendung von Replikationen, erfährt ihr Methodenarsenal eine kritische Durchleuchtung im Hinblick auf die methodologischen Anforderungen im nächsten Abschnitt.

5.1 Der Replikationserfolg

Solange ein experimenteller Befund belassen bleibt oder einem Forschungsbereich entstammt, der sich im Stadium der Finalisierung befindet, kann keine Krise entstehen: Replikationen gelingen nach demselben Rezept, scheitern können nur unfähige Experi-

mentatoren. Anders verhält es sich, wenn ein Experiment fernab der Routine einen noch nicht etablierten Befund hervorbringt. Dann kann das Scheitern auch an einem falsch-positiven Befund liegen, weil es noch keine Kriterien gibt für eine erfolgreiche Replikation (Collins 1984). Welche Eigenschaften und Unterschiede zwischen Original und Replikation bestehen, lässt sich nicht abgrenzen gegen die Fähigkeiten der Experimentatoren. Die Fähigkeit eines Experimentators bemisst sich allein am korrekten Befund des Experiments, den korrekten Befund kann aber nur ein fähiger Experimentator erzeugen (Collins 1992, S.130).

Der Grund für Erfolg oder Misserfolg ist bei zweifelhaften Experimenten noch nicht freigelegt. Was bestätigt wird und wodurch es seine Bestätigung erfährt, ist vorerst unentschieden. Hier am Scheidepunkt kennt keiner den Weg: das Fortschreiten birgt Risiken, für die keine Theorie bürgt. Theorie und Erfahrung, Methoden und Fähigkeiten sind ineinander verwoben (Danziger 1985) und Kriterien der Bestätigung müssen erst erarbeitet werden, was häufig übergangen oder vernachlässigt wird (Oakes 1985, S.163). Wo eine Bestätigung möglich und die Entscheidung für oder gegen den Befund rational ist, wird vorausgesetzt, dass die Bestätigung einen Grund hat und die Entscheidung nach einem Grundsatz getroffen wird. Das besagt der Satz vom zureichenden Grunde: Nihil est sine ratione cur potius sit, quam non sit (Wolff 1736, §70).

Bestätigen Replikationen einen Befund in der Weise, dass die Bestätigung begründet ist, dann ist, wenn die Vernachlässigung von Replikationen sträflich ist, die Vernachlässigung der Philosophie – im Sinne einer πρώτη φιλοσοφία, ob nun Ontologie, Epistemologie oder Metatheorie (Meehl 1992) – ebenfalls sträflich. Schließlich bereitet der Satz vom zureichenden Grunde aller Wissenschaft den Boden (Schopenhauer 1977, S.16). Dass der Satz vom zureichenden Grunde selbst ein Grundsatz ist, der mit der Rechtfertigung von Grundsätzen sich selbst rechtfertigt, ist charakteristisch für die reflektierte Reflexivität der Philosophie. Hier biegt sich der Spaten zurück (Wittgenstein 1990, §217). Sind die letzten Gründe freigelegt, muss der letzte Spatenstich sich aus sich selbst begründen, d.h. selbst Grund sein. Die Bestätigung eines Befundes hat einen Grund ebenso wie die Bestätigung selbst einen Grund hat, man also über den Grund zur Bestätigung kommt, die, begründet, einen Befund bestätigt, indem sie einen Grund für den Befund anführt (Davidson 1980, S.11).

Für eine erfolgreiche Replikation ist ein Grund zum Zwecke der Bestätigung eines Befundes notwendig, weil sie ohne einen solchen unmöglich wäre; der sei hier geschenkt für den Einzelfall, nicht aber für den allgemeinen Begründungszusammenhang, in dem die erfolgreiche Replikation selbst, als Grund für die Bestätigung, steht. Der reflexiv-bestätigende Aspekt einer Replikation kommt methodisch zum Tragen in den Selbstkorrekturmechanismen der Wissenschaft (Campbell 1985; Ioannidis 2012), in denen Replikationen maßgeblich sind (Nosek et al. 2015): Insofern Replikationen selbst Bestandteil sind der Wissenschaft und experimentelle Befunde liefern, haben sie entweder dieselbe Rate an falsch-positiven Befunden wie die Originalexperimente, dann aber können sie die Wissenschaft nicht um falsch-positive Befunde bereinigen (Tsang & Kwan 1999) und müssten selbst repliziert werden usw.; oder aber sie kontrollieren ihre Fehlerrate selbst, dann aber bringen Replikationen in die Wissenschaft etwas ein, was Replikationen von allen anderen Experimenten unterscheidet (Darmant & Matalon 1986).

Dass die philosophischen Aussagen zur Tragweite von Replikationen ganz genauso die Tragweite philosophischer Aussagen autonom begründen müssen, ist ein so populärer wie irreführender Winkelzug (Mulkay 1984; Radder 1992). Die Reflexivität fundamentaler Argumente ist in der Philosophie ein geläufiger Topos (Nelson 1973, S.92), wohingegen die reflexive Seite von Replikationen weniger ins Auge springt: Replikationen können ihre Verlässlichkeit nicht aus Experimenten schöpfen, deren Unverlässlichkeit Grund dafür ist, dass Replikationen zur Korrektur der Experimentalforschung erforderlich sind. Die Etablierung von Replikationen als Korrekturfaktoren im Wissenschaftsbetrieb wirft fundamentale Fragen mit epistemologischem Widerhall auf, der in Begründungszusammenhängen die unvermeidliche Zirkularität der Reflexivität anklingen lässt.

Folgert man daraus, dass viele Replikationen denselben Effekt erzeugten und es diesen Effekt wirklich gibt – wobei die Gründe für die Existenz des Effektes seine Replikation nicht ausschließen –, dass Replikationen verlässlich sind, begeht man einen verallgemeinerten Übertragungsfehler, weil die Gründe für die Wahrheit der Prämissen abhängig sind von davon unabhängigen Gründen für die Wahrheit der Konklusion: die Existenz des Effektes ist nicht nur abhängig von den Replikationen, sondern auch von der Verlässlichkeit der Replikation. Die Gründe für einen Effekt lassen sich nicht übertragen auf die Verlässlichkeit von Replikationen. Vielmehr muss man, um berechtigter-

weise dem Grund zu folgen, den eine Replikation für einen Effekt anführt, berechtigte Gründe haben für die Verlässlichkeit einer Replikation. Sind Grund und Gründe voneinander unabhängig, gelangt man in einen Replikationsregress; sind sie voneinander abhängig, tut sich ein vitiöser Zirkel auf; verzichtet man auf eine Begründung, können Replikationen gegenüber anderen Forschungspraktiken nicht rechtfertigt werden – sie verhalten sich indifferent zueinander (ισοσθένεια).

Verzichtet man auf eine Begründung, kann man in Analogie zu Grundsätzen Replikationen anführen als Grundwert, der neben Universalität, Kommunismus, Skepsis und Neutralität normiert, wie die Forschungspraxis aussehen soll – was sich mit dem Wertekanon der OSC (2012) deckt. Als eine dem epistemologischen Begründungszusammenhang entzogenen Norm geben Replikationen einen Weg vor zur Erlangung wissenschaftlicher Erkenntnis und wissenschaftlichen Fortschritts (Nosek et al. 2015). Wie zielführend dieser Weg ist, wird eine Erörterung des möglichen Erfolgs von Replikationen weisen.

5.2 Der Erfolg des Replikationserfolges

Ein Erfolg von Replikationen liegt vor, wenn zwei direkte Replikationen eine Hypothese besser bestätigen als zwei verschiedene Experimente (Franklin & Howson 1984). Logisch gewendet müsste man aus den Befunden der Replikationen mehr schließen können als aus den Befunden verschiedener Experimente. Der gültige Schluss wiederum bezieht seine Gültigkeit aus den Prämissen. Die Gültigkeit kann man herleiten oder stipulieren. Daher kann ein Schluss bereits berechtigt sein, wenn ihn die Prämissen bestätigen (Boghossian 2014). Im Begriff der Bestätigung steckt ja der Verzicht auf einen Beweis und damit auf deduktive Schlüssigkeit. Verzichtet man nicht völlig auf die Schlüssigkeit (Earp & Trafimow 2015), verbleiben für die Replikationen induktive Schlüsse zur Bestätigung der replizierten Befunde. Die Relation zwischen Befunden und dem daraus Geschlussfolgerten, der Replikationen ihren Zugewinn verdanken gegenüber Primärexperimenten, heißt Wahrscheinlichkeit (Jeffreys 1973 S.23). Das induktive Schließen mit Wahrscheinlichkeiten führt nachfolgend in die Statistik, die nach epistemologischen Gesichtspunkten in der Wissenschaft verortet wird, bevor die daraus resultierenden Schwierigkeiten zur Sprache kommen.

5.2.1 Induktive Bestätigung

Der induktive Erkenntnisgewinn von Replikationen ist ein doppelter: außer der sukzessiven Bestätigung eines Befundes durch direkte Replikationen zeichnen Generalisierungen durch konzeptuelle Replikationen Induktionsschlüsse aus (Sun & Pan 2011). Hat man genügend Instanzen eines Befundes repliziert, gilt der Befund als Tatsache, die auch in belassenen Weltwinkeln Bestand hat. Der virtuose Zirkel von der Hypothese zum Experiment und von dort über die Replikation zur Revision der Hypothese, die erneut anhand eines Experimentes und seiner Replikation geprüft wird, verleiht dem induktiven Schließen einen kumulativen Charakter (Platt 1967).

Zu einer Hypothese gelangt man einmal auf dem Wege wiederholter Beobachtungen, das andere Mal, indem man die Hypothese aus einer Theorie ableitet. Die letztere, hypothetico-deduktive Methode (Popper & Eccles 1982, S.505) ist äußerst umstritten. Mill (1974, S.193) sprach gar Deduktionen überhaupt eine Sonderstellung ab und subsumierte sie unter enumerative Induktionen. Folgert man aus der Fehlbarkeit aller Replikationen und aus dem Umstand, dass es sich bei der vorgelegten Studie um eine Replikation handelt, dass die vorgelegte Studie fehlbar ist, dann ist die Konklusion wahr unabhängig von der Universalprämisse, die falsch wäre, wäre die Konklusion nicht wahr.

Damit Bestätigung und Generalisierung induktiv Gültigkeit erlangen können, muss vorausgesetzt werden, dass gemachte und (noch) nicht gemachte Beobachtungen einander gleich sind, der Kosmos also seine Ordnung nicht ändern darf. Diese Voraussetzung ist der Grundsatz der Induktion. Der gilt aber nicht a priori, weil die Annahme eines sich verändernden Kosmos nicht ausgeschlossen werden kann; er gilt aber auch nicht empirisch, weil gerade der Erfolg eines Induktionsschlusses infrage steht, der ohne Induktionsgrundsatz nicht gewährt ist (Hume 1854, S.32). So wäre es zirkulär, zu argumentieren, dass induktive Schlüsse in der Zukunft verlässlich sind, weil sie in der Vergangenheit verlässlich waren (Lipton 1991, S.10).

Humes Skepsis räumt den Induktionsgrundsatz ab mit einer Wucht, die Induktionsschlüsse auf nicht gemachte Beobachtungen ein für alle Mal zu verunmöglichen scheint (Whewell 1967, S.2; Ramsey 1978, S. 99; Gigerenzer 2004). Da ein Grundsatz aber gilt oder nicht gilt, könnte eine Abstufung zwischen wahr und falsch eine Brücke bauen zu

Bestätigung und Generalisierung. Von der diskreten Wahrheit zum Wahrscheinlichkeitskontinuum scheint es nur ein kleiner Schritt. Statt also entweder verlässlich oder unverlässlich zu sein, könnte die Verlässlichkeit von Induktionsschlüssen mehr oder weniger wahrscheinlich sein (Keynes 1973, S.244). Eine Verbindung zwischen Induktion und Wahrscheinlichkeitstheorie, deren Festigkeit entscheidend ist (Hogben 1957, S.14), stellt die Sukzessionsregel her (Laplace 1814, S.27).

Die Sukzessionsregel beruht auf dem Satz vom unzureichenden Grunde, der besagt, dass – als Ausdruck der Isosthenie – alle Ereignisse gleich wahrscheinlich sind, wenn es keine Gründe dafür gibt, dass die Ereignisse verschieden wahrscheinlich sind (Raman 1994). Insofern Laplace (1814, S.6) weniger apodiktisch Gleichwahrscheinlichkeit postuliert, wenn es keinen Grund gibt, das Gegenteil zu denken, rückt er induktive Wahrscheinlichkeiten (Cohen 1991, S.356) in die Nähe individuell-subjektiver Urteile (Fisher 1930). Weiß man folglich von einer Replikation, dass sie scheitern kann oder gelingen, und erzielt man *s* Erfolge bei *n* Replikationen, dann ist die Wahrscheinlichkeit, in der *n+1*ten Replikation wieder erfolgreich zu sein: $P\left(X_{n+1}=1|\sum_{i=1}^{n} X_i=s\right)=\frac{s+1}{n+2}$, wobei das Wissen um die Möglichkeit des Scheiterns oder Gelingens berücksichtigt ist in der hinzugezählten Beobachtung des Scheiterns und des Gelingens. Die Annahme der Gleichwahrscheinlichkeit schließt die Annahme ein, dass vorausgegangene Ereignisse nachfolgende Ereignisse nicht beeinflussen. Das ist unbegründet (Pearson 1920), kann auch gar nicht anders sein: ein Grundsatz begründet und wird nicht begründet.

Mit der Sukzessionsregel hielten nach und nach die formalen Grundsätze der Wahrscheinlichkeitstheorie Einzug in das induktive Schließen (Broad 1928). Damit ging die Induktion auf in einem formalen Kalkül, in dem die Wahrheitsfunktion zu einer Bestätigungsfunktion mutierte (Carnap 1950, S.2 u. 193; Steinfeld 1979, S.7): Wahrscheinlichkeit wird definiert als Grad der Bestätigung einer Hypothese (Hacking 1965, S.136). Sie hat ihr empirisches Pendant im Grad der Überzeugung eines Subjekts. So kann zwischen tatsächlicher und rationaler Überzeugung unterschieden werden (Carnap 1966): der Grad der Bestätigung einer Hypothese ist ein Maß für den Grad der Überzeugung, den eine rationale Person annehmen würde angesichts eines Befundes, der die Person mehr oder weniger überzeugt (Oppenheim 1952).

In der Parallelkonstruktion manifestiert sich der hybride Charakter der induktiven Logik, die logischen wie auch empirischen Ansprüchen genügen muss. Einerseits existiert die Wahrscheinlichkeit in Form von relativer Häufigkeit, andererseits muss eine Wahrerscheinlichkeit einspringen in Form von nach Überzeugungskraft gewichteter Bestätigungen (Carnap 1950, S.163), weil die bloße Sukzession von Ereignissen das Induktionsproblem nicht löst; einerseits gilt die logische Implikation, andererseits kann man dem Antezedens beipflichten und das Sukzedens dennoch ablehnen, sprich eine sehr wahrscheinlich richtige Hypothese verwerfen, ohne sich in formale Widersprüche zu verwickeln.

Einerseits können Induktionsschlüsse nicht bedingungslos wahr sein, sondern nur mehr oder weniger wahrscheinlich (Wright 1941), andererseits kann es ein wahrscheinliches Wissen ohne Gewissheit nicht geben (Wright 1957, S.152), die außerhalb des induktiven Schließens nahezu unerreichbar ist (Neta 2006), und innerhalb des induktiven Schließens definitiv unerreichbar ist (Jeffreys 1961, S.43). Unter Verwendung der bedingten Wahrscheinlichkeit (Bayes 1763) gilt für eine Hypothese *H*, den experimentellen Befund *b* und das Vorwissen *V*:

$P(H|b \wedge V)=\frac{P(H|V)\cdot P(b|H \wedge V)}{P(b|V)}$, d.h. die Wahrscheinlichkeit einer Hypothese verhält sich zur Wahrscheinlichkeit des Befundes wie die Wahrscheinlichkeit der Hypothese unter der Bedingung des Befundes zur Wahrscheinlichkeit des Befundes unter der Hypothese. Folgt nun ein bestätigender Befund b_n auf den andern b_{n-1}, dann nähert sich die Wahrscheinlichkeit einer erneuten Bestätigung der Gewissheit:

$\lim_{n \to \infty} P(b_n|b_1 \wedge b_2 \wedge ... \wedge b_{n-1} \wedge V)=1$ (Jeffreys 1961, S.44).

Damit scheint das Induktionsproblem gelöst. Doch die Kopplung der Bestätigung an die Überzeugung erwirkt die Lösung nur unter Isomorphiebedingungen, d.h. wenn jeder bestätigende Befund dieselbe Überzeugungskraft hat bzw. eine Person davon überzeugt werden kann, einem Befund jeweils dieselbe Überzeugungskraft zuzugestehen. Ohne weitere Begründung wird daraus eine petitio principii, in der vorausgesetzt wird, was eigentlich zu zeigen wäre, nämlich die Bestätigung einer Hypothese: jeder positive Befund gilt automatisch als Bestätigung der Hypothese (Lamal 1990). Es bleibt also offen, wie aus der formalen Konjunktion bisheriger Befunde auf das Eintreten künftiger Befunde geschlossen werden kann. Die Existenz einer Formel für induktives Schließen

verbürgt nur, dass induktives Schließen mathematisch abgebildet werden kann, sofern induktives Schließen begründet ist. Die induktive Bestätigung selbst ist keine Funktion der mathematischen Wahrscheinlichkeit (Cohen 1991, S.167)

Dasselbe gilt für Reichenbachs (1932, S.5) Versuch einer kumulativen Bestätigung einer Hypothese aus einer Sukzession von experimentellen Befunden:

$P(H(b))=\lim_{n\to\infty}\frac{1}{n}\mathbf{1}_i(H(b_i))$ mit $\mathbf{1}_i(H(b_i))=1$, falls b_i die Hypothese bestätigt. Das Grundproblem bleibt bestehen, in verifikationistischer wie in probabilistischer Perspektive. Hume bezweifelt, dass das Eintreten eines Ereignisses die Wiederkehr des Ereignisses rechtfertigt, egal ob die Wiederkehr gewiss oder nur wahrscheinlich sein soll. Ein experimenteller Befund bestätigt weder die Wahrheit noch die Wahrscheinlichkeit einer Hypothese. Und selbst wenn man einer Sukzession positiver Befunde eine bestätigende Funktion zugesteht, kann wegen der Unterbestimmtheit von Hypothesen (Quine 1964, S.79; 1996 S.16) insofern nicht von der Bestätigung einer Hypothese gesprochen werden, als dieselben Befunde eine Unzahl unterschiedlicher Hypothesen ebenfalls bestätigen (Goodman 1973, S.73).

Die Gültigkeit des Induktionsschlusses kann nicht einfach festgemacht werden an der Überzeugung (Wright 1957, S.21). Im Grunde müsste die Überzeugung bei einem Befund umso stärker zunehmen, je geringer das in der Hypothese zum Ausdruck gebrachte Vorwissen ist; mit anderen Worten: die Rückschlusswahrscheinlichkeit[2] müsste umso größer sein, je geringer die Ausgangswahrscheinlichkeit war (Mises 1951, S.140). Das trifft empirisch nicht zu: für außergewöhnliche Befunde braucht es eine außergewöhnliche Überzeugungskraft (Laws 2016). Soweit die Grade der Überzeugung nicht isomorph sind zu den Graden der Bestätigung, beschreibt der Kalkül der induktiven Logik propositionale Einstellungen gegenüber einer Hypothese, sagt aber nichts aus zur Wahrscheinlichkeit der Hypothese.

Ein solcher Isomorphismus wäre eine überstarre Verbindung (Wittgenstein 1990, §197), die unfehlbar einen infinitesimalen Grad der Bestätigung an die passende Hypothese koppeln würde. Eine solche Verbindung gibt es nicht (Carnap 1950, S.193). Und selbst wenn es sie gäbe, könnten Naturgesetze in der Gestalt universaler Generalisierungen

2 Die Rückschlusswahrscheinlichkeit entspricht dem Produkt aus Ausgangswahrscheinlichkeit und Likelihood (Jeffreys 1961, S.57)

keine Bestätigung erfahren, weil die Bestätigung im Sinne der Sukzession relativ zu den vorausgegangenen Befunden konzipiert ist (Hempel 1977, S.86). Daraus folgt formal, dass der Bestätigungsgrad sämtlicher Naturgesetze null ist, so als wären sie nie bestätigt worden (Carnap 1962, S.571).

Es hilft wenig, die Anzahl der erforderlichen Replikationen zu begrenzen, indem man sie abhängig macht von der Solidität vorausgegangener Experimente (Lamal 1990), weil die Solidität in einer logischen Abhängigkeit zur Induktion steht: ein solides Experiment ist ein gut bestätigtes Experiment. Egal wie häufig ein Befund repliziert wird, die Häufigkeit allein reicht nicht hin zur Bestätigung des Befundes. Positive Befunde bestätigen eine Hypothese genauso wenig, wie negative Befunde eine Hypothese zweifelhafter machen (Cohen 1991, S.175).

Direkte wie konzeptuelle Replikationen scheinen ein Pseudoproblem (Hacking 1983, S.231) darzustellen hinsichtlich der Bestätigung von (generalisierenden) Hypothesen. Replikationen sind nach Hacking nur sinnvoll, wenn die Replikation besser ausgeführt werden soll als das Originalexperiment, um zu genaueren Ergebnissen zu kommen. Verfechter bestätigender Replikationen müssen also den Induktionsgrundsatz als Grundsatz hinnehmen (Ramsey 1978, S.88) oder induktives Schließen als irreführend ablehnen (Neyman 1957, Medawar 1963). Für die einen sind Replikationen elementar, sodass sie keiner weiteren Begründung mehr bedürfen, für die anderen stellt sich die Aufgabe, den methodologischen Rahmen so zu modifizieren, dass erfolgreiche Replikationen vorausgegangene Experimente bestätigen. Einen solchen Rahmen bietet die Statistik.

5.2.2 Statistische Bestätigung

Die Statistik bringt etwas mit, was der Induktion schmerzlich fehlt. Ist die Induktion der Versuch, ausgehend vom kargen Boden spärlicher Beobachtungen auf künftige Beobachtungen zu schließen, so düngt die Statistik den Boden mit einer Überfülle von möglichen Beobachtungen. Um unbeobachtete Ereignisse reicher als die Induktion, die allein auf beobachteten Ereignissen fußt, geht die Statistik den Weg der indirekten Bestätigung von Hypothesen. Auf diesem Weg sorgt sie für eine begriffliche Klarheit, die die Schwierigkeit der Bestätigung von Hypothesen deutlich macht. Je nach Bewältigungsstrategie werden in der Statistik drei Richtungen verfolgt: die der Signifikanztheorie, die

der Entscheidungstheorie und die des Subjektivismus. Die Signifikanztheorie inkorporiert den Induktionsgrundsatz, die Entscheidungstheorie lehnt ihn ab und der Subjektivismus benötigt ihn nicht.

Die möglichen Beobachtungen gelangen über Verteilungen in die Inferenzbasis statistischen Schließens. Aus einzelnen Beobachtungen werden Annahmen getroffen, auf deren Grundlage Beobachtungen intrapoliert werden, die man machen würde, wenn man ein Experiment unendlich oft wiederholen würde (Guttman 1985). Dann schätzt man ab, wie wahrscheinlich die einzelnen Beobachtungen unter den getroffenen Annahmen sind. Nimmt man nach dem Satz vom unzureichenden Grunde hypothetisch an, dass Kopf und Zahl beim Münzwurf gleichwahrscheinlich sind, Kopf und Zahl also gleichverteilt sind, und beobachtet man zehnmal hintereinander Kopf, dann wird unter der Annahme der Gleichverteilung abgeschätzt, wie wahrscheinlich die Gleichverteilung ist angesichts der Beobachtung von zehn Kopfwürfen und (!) angesichts einer unendlichen Menge möglicher Würfe. Die Beobachtungen bestätigen die Hypothese also nicht direkt, sondern zusammen mit möglichen Beobachtungen, d.h. indirekt (Nagel 1949, S.52). Ausschlaggebend für die Bestätigung der Hypothese ist letztlich ihr Zusammenhang mit der Schätzung (Serlin 1985).

Während also in der Induktion mühselig von einer Replikation zur nächsten geschritten wird, sind in der Statistik bereits zahllose virtuelle Replikationen enthalten. Dafür ist die Induktion bzw. induktives Wissen von einem konkreten Ereignis robuster gegenüber zusätzlichen Erkenntnissen, weil man in der Statistik definitionsgemäß nichts weiß von einem konkreten Ereignis und nur Aussagen treffen kann zu Klassen von Ereignissen; eine Beschränkung, die die Wahrscheinlichkeit statistischer Aussagen nicht berührt (Keynes 1973, S.450). Allerdings muss man sich darüber im klaren sein, dass induktive Wahrscheinlichkeiten eine Relation zwischen Aussagen sind, während statistische Wahrscheinlichkeiten eine Relation zwischen Ereignisklassen darstellen (Hempel 1977, S.61).

Der Statistik zum Durchbruch verhalf Fishers (1922; 1928, S.1) Unterscheidung von Population und Stichprobe, die der Statistik in der Psychologie zum Status einer monolithischen Logik (Gigerenzer et al. 1989, S.107) verhalf. Population und Stichprobe verdeutlichen die in der Induktion angelegte Dichotomie von Wahrscheinlichkeit und Wahrerscheinlichkeit, von relativer Häufigkeit und subjektiver Überzeugung. Einmal ist

etwas der Fall, das andere Mal erscheint es einem so, d.h. einmal ist etwas wahrscheinlich, das andere Mal erscheint es einem wahrscheinlich. Genauso ist die Population, wie sie ist, und wie sie erscheint, ist die Stichprobe. Die Stichprobe ist die Erscheinung der Population, oder: dem Statistiker erscheint die Population als Stichprobe. Population und Stichprobe führen demnach zurück auf die Unterscheidung von νούμενον und φαινόμενον (Kant 1974, B308), die sich eignet zu fruchtbaren Gegenüberstellungen von Eigenschaften der Population und Eigenschaften der Stichprobe, die in Abbildung 2 das dichotome Grundmodell statistischen Schließens veranschaulichen.

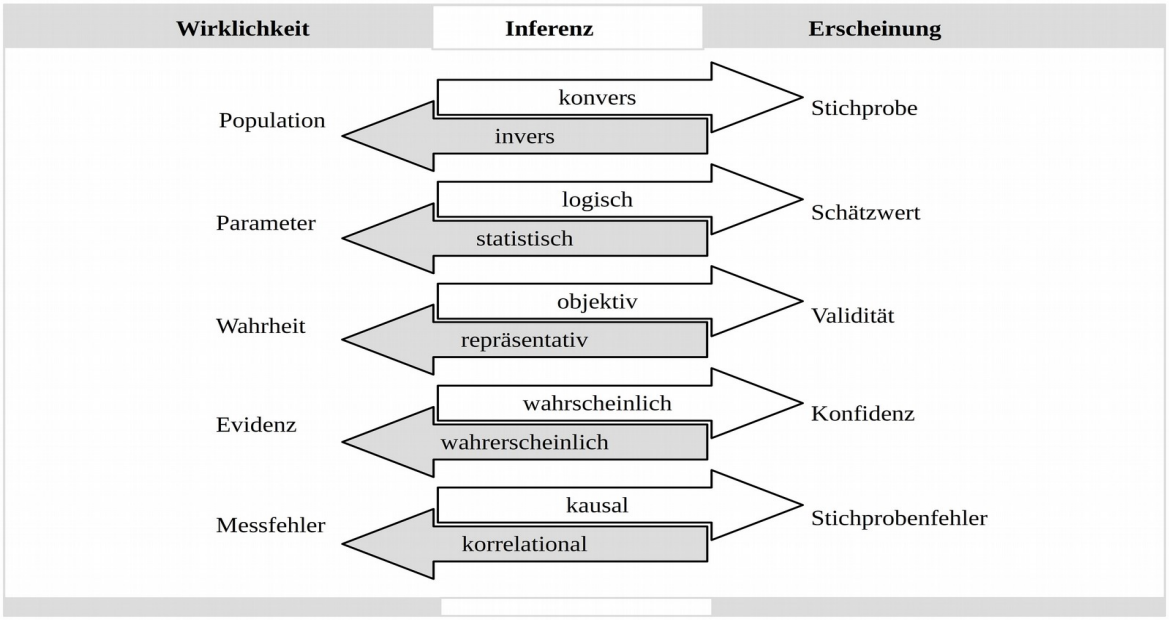

Abbildung 2: Gegenüberstellung von Eigenschaften der Population und der Stichprobe mit inferenziellen Übergängen.

Eine Statistik setzt sich zusammen aus Parametern, die die Verteilung aller möglichen Ereignisse regieren, die wiederum konzipiert sind als (stetige) Zufallsvariablen, sodass eine Hypothese noumenal einen Wert zum Ausdruck bringt, der anhand der beobachteten Ereignisse, also anhand realisierter Zufallsvariablen, phänomenal geschätzt wird. Nimmt man in das Universum der Statistik Individuen als Elemente der Ereignisklassen hinzu, dann lassen sich acht Schlussformen klassifizieren, zu denen in der kritischen Diskussion eine neunte hinzugefügt werden muss; bei den ersten beiden Schlussformen fällt die Konklusion mit den Prämissen zusammen, weil die Prämisse der Konklusion Konklusion der Prämisse ist; die drei darauffolgenden Schlussformen sind der Induktion zuzurechnen (Carnap 1950, S.207), die übrigen betreffen die Statistik:

- beim kasuistischen Schließen schließt man von einem Individuum aufs andere Individuum und umgekehrt mittels Deskription;
- beim hermetischen Schließen schließt man vom Individuum aufs Universum und umgekehrt mittels Analogie;
- beim deklarativen Schließen schließt man vom Individuum auf die Stichprobe mittels vollständiger Enumeration;
- beim induktiven Schließen s.s. schließt man vom Individuum auf die Population mittels Induktion;
- beim universalen Schließen schließt man von der Stichprobe aufs Universum, d.h. auf kosmische Invarianten oder Naturgesetze, mittels Extrapolation;
- beim Trend-Schließen schließt man von einer Stichprobe auf die andere Stichprobe mittels Nivellierung von Unterschieden auf dem Weg der Durchschnittsbildung;
- beim konversen Schließen schließt man von der Population auf die Stichprobe mittels Deduktion.

Im Zentrum des wissenschaftlichen Interesses steht das inverse Schließen von der Stichprobe auf die Population. Hierin assistiert die Statistik den Wissenschaften (Cox 1958), je nach Ausrichtung mit eigenen Mitteln, die sich formal kaum unterscheiden, inhaltlich dafür umso mehr. Weil die drei Richtungen der Statistik dieselben Formeln verwenden und häufig numerisch zum selben Ergebnis kommen, werden ihre Unterschiede gerne übersehen. So zum Beispiel der Unterschied zwischen inversem Schließen und projektivem Schließen: beim projektiven Schließen schließt man von fiktiven Stichproben auf die Population, d.h. auf eine Überfülle unbeobachteter Ereignisse werden die beobachteten Ereignisse projeziert, sodass die Prämissen eines projektiven Schlusses im wesentlichen fingierte Stichproben sind.

Signifikanz- und Entscheidungstheoretiker sowie Subjektivisten sind sich darin einig, dass das Ergebnis eines Experimentes Evidenzen beibringt für die Bestätigung einer Hypothese. So ist es eine Aufgabe der Statistik, diese Evidenz zu quantifizieren und zu interpretieren in einem indeterminierten, probabilistischenen Maß. Die Suche nach der Bestätigung in der Evidenz führt somit auf die Konfidenz. Diese statistische Evidenz sui generis gilt als Interpretation der experimentellen Evidenz (Birnbaum 1962). Der Konfidenz kriterial zugeordnet ist das Konfidenzintervall, das einen Populationsparameter

schätzt. Ein Intervall zum Konfidenzniveau von 95 Prozent beschreibt die Überzeugung, dass das Intervall ein Repräsentant ist einer unendlichen Sukzession von Konfidenzintervallen, von denen auf lange Sicht 95 Prozent den Populationsparameter einschließen; oder aber es verkörpert die Überzeugung, dass durchschnittlich 83 Prozent der Replikationen einen Schätzwert generieren, der innerhalb des Intervalls liegen würde (Cumming 2012, S.133).

Nur wenn man unter Evidenz Gewissheit versteht (Oakes 1985, S.15) in dem Sinne, dass die Konfidenz zwingend gegensätzliche Evidenzen ausschließt (Kyburg 1974, S.153), kann es keine statistische Evidenz geben – eine empirische Evidenz im übrigen auch nicht (Unger 1975, S.39; Wittgenstein 1990, §§ 324-326). Als Evidenz in der schwächeren Form von Hinweis oder Indiz kann die Konfidenz durchaus ein Maß statistischer Bestätigung abgeben. Schließlich trägt sie bei zur statistischen Validität.

Statistische Validität bezieht sich auf das Ausmaß der Vermeidung von Schätzfehlern durch hohe Teststärke oder Unabhängigkeit der Stichproben, präzise Hypothesen und unverzerrte Interpretation der Testergebnisse (Westermann 2000, S.321). Der statistischen Validität kommt nur eine untergeordnete Bedeutung zu, weil sie abhängt von Design, Stichprobe und Operationalisierung, für die Modellkonstrukte erforderlich sind. Die Konstrukte beziehen ihre Validität aus umfassenden Theorien, welche im Validierungszusammenhang einer Metatheorie bedürfen, die erklärt, warum konkurrierende Theorien ausgeschlossen werden können (Fiedler, Kutzner & Krueger 2012). Neben der bereits genannten Konstruktvalidität adäquater Modellierung sind noch zu nennen die interne Validität durch Kontrolle der abhängigen Variablen, die eine suffiziente Statistik verbürgt, und die externe Validität einer erfolgreichen Generalisierung der Testergebnisse, zu der konzeptuelle Replikationen beitragen sollen (Cook & Campbell 1979, S.85-91). Bei den Validitätskriterien handelt es sich insofern ebenfalls um eine Abschwächung von Wahrheitskriterien, als sie sich speziell auf statistische Tests beziehen: die Validität ist – wie die Reliabilität – eine Eigenschaft von Tests und keine Eigenschaft der Wirklichkeit (Kane 1992).

5.2.2.1 Signifikanztheorie

Die Tests, mit der Signifikanztheoretiker die Bestätigung einer Hypothese anstreben, sind Signifikanztests. Seine Signifikanz bezieht der Test aus der Grenzwahrscheinlichkeit, zu der man bereit ist, die Nullhypothese, dass Versuchs- und Kontrollgruppe sich nicht unterscheiden, fälschlicherweise zu verwerfen. Dieses Fehlerniveau ordnet der Nullhypothese keine Wahrscheinlichkeit zu: wenn eine Nullhypothese zur Fehlerwahrscheinlichkeit von 5 Prozent nicht verworfen werden kann, heißt das nicht, dass sie zu 95 Prozent wahr ist. Das Fehlerniveau leistet man sich in der Population, wohingegen die im *p*-Wert ausgedrückte Wahrscheinlichkeit der beobachteten Stichprobe zuzurechnen ist. Von der Stichprobe zur Population kann nur invers geschlossen werden, sodass die Wahrerscheinlichkeit einer Hypothese angesichts der beobachteten Ereignisse sich nur angeben lässt mittels Likelihood oder Fiduzialwahrscheinlichkeit, die Grade des Vertrauens in eine Hypothese ausdrücken sollen. Die Bestätigung der Hypothese resultiert letztlich aus Wiederholungen (Fisher 1935, S.28f; Tukey 1969) – nicht in Form realer Replikationen, sondern in Form virtueller Replikationen: Der *p*-Wert steht für die Wahrscheinlichkeit einer Stichprobe angesichts unendlich vieler virtueller Stichproben desselben Umfangs.

In der Logik des Signifikanztests folgt aus der Beschaffenheit der Population, wie häufig eine Stichprobe mit bestimmten Ereignissen vorkommt. Die relative Häufigkeit einer Stichprobe im Verhältnis zu gleichgroßen Stichproben mit allen denkbaren Beobachtungen entspricht dann der Wahrscheinlichkeit der Stichprobe. Justiert man die Parameter der Verteilung der Stichproben gemäß der Nullhypothese, dann bedeutet eine geringe Wahrscheinlichkeit bzw. ein kleiner *p*-Wert, dass die Fläche unter der durch die Nullhypothese parametrisierte Wahrscheinlichkeitsdichte bis zur gezogenen Stichprobe sehr klein ist, was darauf hindeutet, dass Versuchs- und Kontrollgruppe sich doch unterscheiden. Nun hat man es meist mit nur einer Stichprobe zu tun und kennt somit nur die Realisationen der Zufallsvariablen beim einmaligen Ereignen einer Stichprobe. Um die relative Häufigkeit der Stichprobe dennoch spezifizieren zu können, ergänzt Fisher die Stichprobe um sämtliche Stichproben, die in derselben Weise hätten gezogen werden können – unter der Annahme, dass die Nullhypothese zutrifft. Als Funktionswerte der Stichprobenverteilung sind sämtliche Stichproben gleichwertig, was Mayo und Cox (2010, S.303) als Verankerung in der Wirklichkeit preisen und Krueger (1999) als ohne

Bezug zur Erfahrung kritisiert. Für Oakes (1985, S.4) sind Ereignisse, die den Ausgang eines Experimentes hätten beeinflussen können, es aber nicht getan haben, schlichtweg irrelevant.

Durch die Erzeugung einer Ereignisklasse, die aus der wirklichen und allen möglichen Stichproben erzeugt wird, besteht für Signifikanztheoretiker kein epistemologischer Grund zur Replikation eines Experimentes. Replikationen dienen in erster Linie dazu, den Standardfehler zu verringern (Fisher 1935, S.66). Bestätigung erfährt eine Hypothese im wesentlichen aus dem Experiment selbst. Der Signifikanztest trägt zur Einschätzung der Evidenz nur exhaurierend bei (Dingler 1907, S.29); er begnügt sich mit dem Verwerfen von Nullhypothesen. Dass ein Signifikanztest aus logischen Gründen überhaupt keine bestätigende Wirkung entfalten kann (Bredenkamp 1980, S.18), ist konstruiert aus der Implikation 'Wenn die Nullhypothese zutrifft, dann sind die Stichproben mit μ und σ verteilt'. So als könne man durch die Negation des Sukzedens das Antezedens endgültig falsifizieren (Popper 1989, S.207). Aber aus dem Eintreffen des unter der Nullhypothese sehr seltenen Ereignisses, dass der Stichprobenschätzer sehr weit von μ bzw. von Null abweicht, kann eben nicht per Kontraposition die Nullhypothese für falsch erklärt werden: wenn das Ereignis in allen nachfolgenden Experimenten kaum noch auftritt, ist es sehr selten und damit von der Nullhypothese korrekt vorhergesagt.

Mit der Deduktion verbindet die Signifikanztheorie weit weniger als mit der Induktion. Der wesentliche Unterschied zur Induktion besteht darin, dass ein hinzugekommene Befund nicht direkt den vorausgegangenen Befunden zugerechnet wird und unverändert in die Bewertung einer Hypothese eingeht; kommt in der Statistik ein Befund zur Stichprobe hinzu, dann nimmt die Stichprobe eine neue Gestalt an, die insgesamt ihre Position in der Stichprobenverteilung verändert und so zu einer Neubewertung der Hypothese führt. Daher geht in der Statistik ein zusätzlicher Befund nur indirekt ein in die Bewertung einer Hypothese (Bakan 1966). Wie die Bewertung im Einzelfall ausfällt, kann signifikanztheoretisch nicht vorhergesagt werden, sondern hängt im wesentlichen ab von der Erfahrung des Experimentators (Fisher 1935, S.217).

Es ist in der totalen Evidenz (Carnap 1947) die Summe der Gründe, die einen Forscher zum Verwerfen oder Annehmen einer Hypothese veranlasst. Diese Gründe sind vielschichtig und kaum quantifizierbar (Neyman & Pearson 1928). Insofern gibt es auf der

phänomenalen Stichprobenseite kein referenzielles Gegenstück zum Grad individueller Überzeugung (Neyman 1957). Individuelle Überzeugungen sind von Psychologen zu modellieren und nicht von Statistikern. Ohne ein solches Modell psychischer Phänomene ist ein Induktionsschluss sinnlos (Neyman 1955). Forscher verhalten sich, wie sie sich verhalten; und nach angemessener Wiederholung eines Experiments und allem Abwägen empirischer und statistischer Evidenz, entscheiden die Forscher, ob sie eine Hypothese verwerfen oder nicht (Pearson 1955). Das ist mit induktivem Verhalten (Neyman 1957) gemeint. Mehr nicht.

5.2.2.2 Entscheidungstheorie

Buchstäblich entscheidend ist, was die Forscher als relevant für ihre Entscheidung betrachten (Raiffa & Schlaiffer 1961, S.32; Salmon 1984, S.32). Das kann nicht alles in einem Schätzwert repräsentiert sein. Doch kommt es in der Entscheidungstheorie weniger an auf die Bedeutsamkeit eines Schätzwertes als vielmehr auf dessen funktionale Erzeugung (Neyman 1936). Wichtig ist, dass jeder Versuch, einen Parameter zu schätzen, eine Entscheidungssituation darstellt. Der Test einer Hypothese gilt entsprechend als ein Spezialfall des allgemeineren Entscheidungsproblems, für welche Hypothese unter mehreren man sich entscheiden sollte (Wald 1971, S.18). Dazu wird die Nullhypothese ersetzt durch eine Klasse von Alternativhypothesen, die daran bemessen werden, wie häufig sie sich als falsch erweisen. Wie in der Signifikanztheorie geht es auch hier um die Minimierung theoretischer Fehler.

Im einfachsten Fall hat man es mit einem Test von zwei Hypothesen zu tun unter der Annahme, dass eine von beiden wahr ist. Unter dieser Annahme wird ein Entscheidungskonfidenzintervall derart festgelegt, dass die für wahr gehaltene Hypothese verworfen wird, wenn der Testwert innerhalb des Intervalls liegt bzw. dass die Alternativhypothese verworfen wird, wenn der Testwert außerhalb des Intervalls liegt; die jeweils andere Hypothese gilt als angenommen. Die Wahrscheinlichkeit dafür, dass der Testwert innerhalb resp. außerhalb des Intervalls zum Liegen kommt, heißt in der Entscheidungstheorie Testgröße resp. Teststärke.

Über die Teststärke wird der Fehler kontrolliert, die für wahr gehaltene Hypothese fälschlicherweise anzunehmen. Kontrolliert man wie im Signifikanztest nur den Fehler,

die für wahr gehaltene Hypothese fälschlicherweise zu verwerfen, lassen sich immer Gründe finden, das Signifikanzniveau hochzuschrauben, um an der für wahr gehaltenen Hypothese festhalten zu können; weil dadurch aber die Wahrscheinlichkeit eines Fehlers zweiter Art zunimmt, rückt bei einer Fehlervermeidungsstrategie die Bedeutsamkeit von Alternativhypothese in den Blick (Neyman & Pearson 1928).

Die Bedeutsamkeit von Alternativhypothesen in der Statistik ist die wesentliche Neuerung der Entscheidungstheorie. Sie ermöglicht den konsekutiven Vergleich zweier Hypothesen, was einer Induktion durch Elimination gleichkommt (Gigerenzer et tal. S.102). Auch hier verläuft die Bestätigung nur indirekt. Anhand der Teststärke kann zur effektiven Eliminierung von Hypothesen der erforderliche Stichprobenumfang berechnet werden, wozu Signifikanztheoretiker den Fertigkeiten von Experimentatoren vertrauen mussten.

Wer experimentiert und regelmäßig Befunde produziert, will erstens Hypothesen stützen auf seine Befunde und zweitens angeben, wie wahrscheinlich die Hypothesen sind (Edwards, Lindman & Savage 1963). Mit anderen Worten: der Experimentator würde als echter Empiriker seine Schlüsse gerne invers aus einer realen Stichprobe ziehen, statt aus einer fiktiven Stichprobenklasse; und er würde gerne in absoluten Werten wissen, inwieweit die Hypothesen zutreffen, statt sein Forschungsverhalten auszurichten an relativen Werten, insoweit eine Hypothese wahrscheinlicher ist als andere. Genauer: es geht ihm um die Wahrerscheinlichkeit von Hypothesen. Die Wahrscheinlichkeit qua relative Häufigkeit greift dafür zu kurz.

Signifikanz- und Entscheidungstheoretiker bekommen Schwierigkeiten, wenn sie für einmalige Ereignisse Wahrscheinlichkeiten angeben sollen. Für Leben auf dem Mars bspw. lässt sich mit einem auf Häufigkeiten restringierten Probabilismus keine Wahrscheinlichkeit angeben (Neyman 1942; 1957). Wohl aber lässt sich angeben, wie wahrscheinlich einem Leben auf dem Mars erscheint. Die (invers gewonnene) Wahrerscheinlichkeit ist genauso zu reklamieren wie die (konvers gewonnene) Rückschlusswahrscheinlichkeit, wo sinnlos viele Experimente sinnvoll sein sollen, d.h. wo experimentelle Befunde Hypothesen bekräftigen oder abschwächen können (McGuigan 1956).

5.2.2.3 Subjektivismus

Subjektivisten können Parametern, und damit den sich auf sie beziehenden Hypothesen, direkt Wahrscheinlichkeiten zuordnen (Steinfeld 1979, S.3). Sie stützen sich dabei auf die Formel von Bayes (1763) für die bedingte Wahrscheinlichkeit, also die Wahrscheinlichkeit für das Zutreffen einer Hypothese unter der Bedingung, dass der Befund einer Stichprobe vorliegt:

$$P(H|b)=\frac{P(H)\cdot P(b|H)}{P(b)} \quad .$$

Sagt die Hypothese H den Befund b voraus (H="Alle Items mit Neuroinformation werden besser beurteilt" und b=besser beurteiltes Item mit Neuroinformation), dann kann man mit $P(b|H)=\frac{P(b\cap H)}{P(H)}=1$ wegen $b\subset H$ herleiten, dass bestätigende Befunde eine Hypothese bestätigen, indem sie die Hypothese wahrscheinlicher machen: $P(H|b)\geq P(H)$. Wie im übrigen auch etwas ohne Neuroinformation, das kein Item ist, aber schlechter beurteilt wurde, die Hypothese, dass Items mit Neuroinformation besser beurteilt werden, ebenfalls bestätigt, weil die Hypothese H^*="Alles ohne Neuroinformation oder was kein Item ist, wird schlechter beurteilt" zu H logisch äquivalent ist (Hempel 1945).

Allerdings erkauft man sich die Herleitung einer Hypothesenbestätigung durch erfolgreiche Replikationen mit einer Beschränkung der Bestätigungsextension: was bestätigt wird, ist weniger eine Hypothese im eigentlichen Sinne, sondern die subjektive Überzeugung, dass die Hypothese zutrifft. Die Überzeugungen formen ein System, das dem Wahrscheinlichkeitskalkül nachgebildet ist. Die Subjektivierung radikalisiert die Wahrerscheinlichkeit und immunisiert sie zugleich gegenüber Einflüssen aus der Wirklichkeit (Finetti 1972, S.21). Positiv gewendet machen noumenale Ereignisse phänomenale Erfahrungen nicht überflüssig. Denn in Erfahrungen findet sich die Evidenz, mit der Hypothesen bewertet werden.

In einer allgemein akzeptierten Wahrscheinlichkeit für die Bewertung einer Evidenz nimmt der Subjektivismus, dem Satz vom unzureichenden Grunde folgend, seinen Ausgang, die daher den Namen Ausgangswahrscheinlichkeit trägt. Jede neue Evidenz führt in induktiven Schritten zu einer Neubewertung der aktuellen Ausgangswahrscheinlichkeit, die deren Wert je nach Evidenz mal mehr, mal weniger übertrifft. Das Über-

zeugungssystem verfügt wie das Wissenschaftssystem über die Anlagen zur Selbstkorrektur: wenn Individuen mit Widersprüchen in ihren Überzeugungssystem konfrontiert werden, wenden sie sich aus freien Stücken den rationalen Argumenten zu (Savage 1972, S.58), um kognitive Dissonanzen zu verhindern bzw. Konsistenz im Denken und Handeln herzustellen (Finetti 1931).

Aufgabe der Wahrerscheinlichkeitstheorie ist es, auf Fehlschlüsse hinzuweisen und probabilistische Zusammenhänge so verständlich darzustellen, dass es den Individuen möglich ist, Inkonsistenzen in ihrem Verhalten aufzudecken (Savage 1972, S.57). Diese Inkonsistenzen sind realiter allerdings sehr beharrlich und weit entfernt vom Ideal des kalkülen Kopfes. Die Versuchspersonen von Kahneman und Tversky (1973) bewerteten die Wahrscheinlichkeit eines Ereignisses nach der Repräsentativität des Ereignisses für die Ereignisklasse, statt nach dessen Häufigkeit in der Population. Die Personen ließen zudem bei ihren Bewertungen durchgehend die Basisrate eines Ereignisses außer Acht und verstießen damit erheblich gegen den Wahrscheinlichkeitskalkül. Und sie verstießen gegen den Kalkül auch noch, nachdem sie auf ihre fehlerhaften Angaben aufmerksam gemacht wurden. Konfrontiert damit, dass positiv und negativ formulierte Wahrscheinlichkeiten zum selben Ergebnis führen, ließen sich nur wenige Versuchspersonen in ihrem inkonsistenten Urteil umstimmen. (Slovic & Tversky 1974).

Dem Subjektivismus fehlt ein äußerer Maßstab zur Bewertung eines Befundes. Es gelingt ihm nicht, von den Befunden invers auf einen Populationsparameter zu schließen. Die Ausgangswahrscheinlichkeit hält ihn gewissermaßen im Reich der Erscheinungen zurück, denn die Wahl einer Ausgangswahrscheinlichkeit ist nicht weniger fiktiv als die Stichprobenverteilungen der Signifikanz- und Entscheidungstheorie. Aus Gründen der Konsistenz müssen die Befunde als Evidenzen interpretiert werden, sodass sich die Fiktion transitiv überträgt von Befunden auf Überzeugungen. Ohne Ausgangswahrscheinlichkeit aber gibt es im Subjektivismus keine Rückschlusswahrscheinlichkeit. Mit der inversen Wahrscheinlichkeit bzw. Wahrerscheinlichkeit verkettet ist die Subjektivität über die Likelihood, die den Missing Link verkörpern könnte zwischen Evidenz und Ereignis.

5.2.2.4 Likelihood

Die Likelihood ist eine Funktion $\mathfrak{L}$ eines Populationsparameters θ angesichts eines realisierten Experimentes *B*, wobei davon ausgegangen wird, dass alle aus dem Experiment zur Befundlage *b* aggregierten Befunde b_i dieselbe Wahrscheinlichkeitsdichte besitzen. Diese wird dem Parameter zugeordnet, sodass die Likelihood angibt, wie wahrscheinlich es ist, einen Befund zu erhalten, wenn in der Population der Parameter gelten würde: $\mathfrak{L}(\theta|b)=P(b|\theta)$. Weil θ unbekannt ist, wird θ vollständig zerlegt in θ_i und jedem θ_i eine Ausgangsverteilung $P(\theta_i)$ zugewiesen. Kombiniert man die Ausgangsverteilung mit der Likelihood, erhält man eine Rückschlussverteilung mit einer Klasse von Wahrscheinlichkeiten, die sich nach Winkler und Hays (1975, S.474) in der Formel zur bedingten Wahrscheinlichkeit zusammenfassen lassen zu

$P(\theta_i|b)=\frac{P(b|\theta_i)\cdot P(\theta_i)}{P(b)}$, wobei $P(b)=\sum_{j=1}^{n} P(b|\theta_j)\cdot P(\theta_j)$ (Satz der totalen Wahrscheinlichkeit).

Mithilfe der Likelihood kann berechnet werden, wie wahrscheinlich die Ausprägung eines Parameters ist bei kontingenter Befundlage. Damit ist ein Bezug hergestellt zur Bestätigung einer Hypothese. Denn in einer Hypothese ist die Ausprägung eines Parameters konkretisiert. Beim Neuroeffekt ist der infrage stehende Parameter eine Mittelwertdifferenz ($\theta=\mu_{mit}-\mu_{ohne}$) in der Population, sodass die einseitige Hypothese $\theta>0$ lautet und $b=\overline{Y}_{\text{mit}}-\overline{Y}_{\text{ohne}}$ die Itembeurteilungen aggregiert.

Eine Hypothese kann demnach verstanden werden als Funktion eines Parameters, und der Bestätigungsgrad einer Hypothese als eine Funktion eines Parameters angesichts der Befundlage. Interpretiert man die Befunde als Evidenzen, wird die Likelihood Bestandteil einer epistemologischen Bestätigungsfunktion. Dann reicht selbst die geringste Evidenz für einen hochplausiblen Parameterwert aus, um eine Hypothese zu bestätigen, die den Wert impliziert (Goodman 2001). Insofern die Hypothese in die Likelihood integriert ist und die Befunde als Evidenzen für oder gegen die Hypothese gelten, artikuliert die Likelihood die Überzeugungskraft einer Evidenz angesichts einer Hypothese. Allerdings wächst einer solchen Likelihood eine bestätigende Bedeutung erst zu im Verhältnis zur Likelihood konkurrierender Hypothesen, die die Befundlage ebenfalls implizieren (Goodman 1999).

Der aus dem Verhältnis zweier Likelihoods bestehende Bayes-Faktor drückt aus, inwieweit ein Befund eine Hypothese stärker bestätigt als eine andere Hypothese. Sind vor einem Experiment Alternativ- und Nullhypothese gleich überzeugend und fördern die Ergebnisse des Experiments einen Bayes-Faktor von 3 zutage, dann hat im Rückschluss die Alternativhypothese eine Überzeugungskraft von 75 Prozent (Etz & Vandenkerckhove 2016). Somit kann unter Verwendung der Likelihood begründet die Nullhypothese verworfen und die Alternativhypothese angenommen werden (Perlman & Wu 1999). Der Grund dafür ist allerdings stets ein relativer.

Aufgrund der Relativität hängt die objektive Bestätigung einer Hypothese an der glücklichen Auswahl aus dem unendlichen Universum möglicher Hypothesen. Praktisch kann die Likelihood nur für eine begrenzte Zahl von Hypothesen bestimmt werden, und selbst diese können nicht nach jedem Experiment veröffentlicht werden, sodass in einer Replikation Hypothesen nicht bewertet werden können, die im Original nicht berücksichtigt waren (Hacking 1965, S.222). Der Bayes-Faktor stößt also an praktische Grenzen, zu denen auch seine Anfälligkeit gegenüber Ausreißern zählt. Räumt man ein, dass Evidenzen auch irreführend sein können, kommt man nicht an der Bestimmung einer Stichprobenverteilung vorbei, mit der der Fehler kontrolliert wird, eine wahre Hypothese fälschlicherweise zu verwerfen (Pearson & Neyman 1930). Sonst ist es immer möglich, für eine Hypothese einen günstigen Bayes-Faktor zu erzeugen, selbst wenn die Vergleichshypothese wahr sein sollte, indem man letzterer in Form ihrer Ausgangswahrscheinlichkeit eine extrem hohe Überzeugungskraft zuschreibt (Mayo 2004, S.95).

Diese Verzerrung ist möglich, weil auch die Likelihood keine überstarre Verbindung herstellen kann zwischen erscheinender Evidenz und wirklichen Ereignissen (Oakes 1985, S.13). Der Bayes-Faktor bleibt daher subjektiv (Johnson 2013). Die Verstrickung der Likelihood in die Dichotomien der Statistik ist an der Isolation der statistischen Logik von der mathematischen Logik ersichtlich: die statistische Logik gründet sich auf empirische Evidenz, während die mathematische Logik sich auf begriffliche Relationen gründet. Die durchgängige Dichotomie im statistischen Denken erfordert auch eine Dichotomie der Evidenz mit einer logisch-objektiven und einer subjektiv-erscheinenden Seite. Und die Likelihood adressiert nur die subjektiv-erscheinende Evidenz (Royall 1997, S.16).

Deshalb berühren logische Widersprüche die Konsistenz statistischer Wahrscheinlichkeitskalküle nicht. Unter den beschriebenen Voraussetzungen ist mit ihnen vereinbar, dass mächtige Hypothesen, die weniger mächtige einschließen, im Verhältnis zu diesen weniger mächtigen Hypothesen keine höhere Likelihood erzielen, wie man sich leicht verdeutlicht an zwei gleichwahrscheinlichen Hypothesen H_1 und H_2 und einer dritten Hypothese $H_3 = H_1 \vee H_2$, die keine größere Likelihood besitzt als H_1 oder H_2, obwohl für die Wahrheit von H_3 die Wahrheit einer der beiden Hypothesen ausreicht und H_3 eigentlich wahrscheinlicher sein müsste. Verwendet man den Bayes-Faktor als Grad der Bestätigung, muss man daher die Hypothesen sehr genau anschauen, weshalb der Faktor nur bei überschaubaren Modellen sinnvoll ist. (Mayo & Cox 2010, S.281). Für Hypothesen zu Hierarchischen Modellen ist er dagegen nicht geeignet (Lele 2004, S.191).

5.2.3 Wissenschaftlicher Fortschritt durch Replikation

Wissenschaftlicher Fortschritt kann etwas Geradliniges (Kant 1974, B VII) an sich haben oder seine innere Notwendigkeit aus einer organischen Abfolge von Stadien beziehen (Condorcet 1988, S.80). Weniger zielgerichtet ist das herumtappende Forschen in einer durch und durch falliblen Welt. Für alle drei Arten der Wissenschaftsentwicklung wird nacheinander der Beitrag erörtert, den Replikationen leisten können.

5.2.3.1 Kumulativer Fortschritt

Ohne klare Hinweise, wie eine Hypothese bestätigt werden kann, rückt das Ziel einer kumulativen Wissenschaft (Curran, Hussong, Cai & Huan 2009) in weite Ferne. Das Ziel ist getragen vom Gedanken, dass die Wissenschaft umso weiter fortschreitet zu einem vollständigen Verständnis des Universums, je mehr Entdeckungen gemacht und repliziert werden (Zeigler 2012). Freilich wirken Replikationen per se nicht kumulierend (Finifter 1975), es müssen schon Replikationen mit hoher Teststärke von wichtigen Entdeckungen sein (Brandt et al. 2013). Allerdings geht es auch ohne: Astronomie, Archäologie und Paläontologie sind kumulative Wissenschaften, die keine Replikation kennen (Cumming 2012).

Die Kumulation beschreiben Schmidt und Hunter (2015, S.10) so, dass die Theorien sich mit der Zeit wie konzentrische Kreise ausdehnen, was aber über die Metapher hinaus einen Nachweis schuldig bleibt (Hedges & Olkin 1985). Leichter nachweisbar ist das Ausmaß der Übereinstimmung von Replikationen. An ihnen soll der Zuwachs an Sicherheit und Stabilität der Theorien abgelesen werden können, trotz methodischer Unzulänglichkeiten. Macht man Replikationen so zum Maßstab der Bestätigungsakkretion, umschifft man elegant das Bermudadreieck aus Zirkel, Regress und Isosthenie, ohne jedoch diejenigen zu befriedigen, die wissen wollen, ob Replikationen einen Befund bestätigen können. Macht man umgekehrt die Bestätigung abhängig von Replikationen und bemisst wissenschaftliche Kumulation an der Konsistenz von Replikationen, ist die Psychologie nicht weniger kumulativ als die Physik (Hedges 1987). Sehr viele erfolgreiche Replikationen vermindern Inkonsistenzen, machen einen Befund zugleich aber verdächtig (Francis 2012).

Kumulation und Fortschritt sind keine Synonyme; während die Kumulation das konservative Treiben normaler Wissenschaft (Kuhn 1991, S.37f) kennzeichnet, charakterisieren den Fortschritt eher Innovationen. Weil direkte Replikationen keine Freiheitsgrade haben, können sie bestenfalls kumulativ wirken, während den konzeptuellen Replikationen Innovationen offenstehen. Wendet man den Bayes-Faktor an auf eine direkte und eine konzeptuelle Replikation desselben Originalexperiments, erfährt dessen Hypothese im Verhältnis zu einer festen Alternativhypothese – wenig überraschend – eine stärkere Bestätigung durch die konzeptuelle Replikation (Franklin & Howson 1984).

Fortschritt ist zumindest ohne Kumulation möglich. Versteht man unter kumulativer Wissenschaft das Lösen neuer Probleme mit Lösungen, die auch alle bisherigen Probleme lösen, gibt es entweder in der Wissenschaftsgeschichte kein Beispiel dafür oder keinen Fortschritt. Vielmehr gibt es Gegenbeispiele: geladene elektrische Teilchen konnten das Problem der Abstoßung gleichgeladener Leiter nicht lösen im Unterschied zu den Wirbeln der Wirbeltheorie, die von Elektronen und Protonen verdrängt wurden (Laudan 1977, S.148); die wiederum wurden verdrängt von Materiewellen, obwohl sie ohne theoretische Verluste die Umlaufbahnen der Elektronen im Atom nicht ersetzen konnten (Heisenberg 1973).

Dennoch halten Forscher wie Nosek (2016) daran fest, dass der wissenschaftliche Fortschritt ein kumulativer Fortschritt sei. Dieser Fortschritt kann höchstens ein subjektiver

Fortschritt der Reduktion von Unsicherheit sein oder ein Fortschritt bzgl. der sozialen Bedeutung der Probleme, die eine neue Theorie löst (Laudan 1977, S.150), im Sinne von Problemen, die aktuell im Zentrum der wissenschaftsöffentlichen Aufmerksamkeit stehen (Schmidt & Hunter 2015, S.16). Das ist ein Fortschritt, der sich im Kreise dreht, insofern das Stellglied sich selbst korrigierend die Stellgröße regelt: bestätigte Hypothesen definieren in der Sphäre der Subjektivität ein Problem, das aufgrund seiner subjektiven Bedeutsamkeit im Rahmen einer Theorie gelöst wird, deren Hypothesen bestätigt sind oder bestätigt werden müssen.

5.2.3.2 Organisches Wachstum

Ohne erkennbaren Anker, der Erscheinungen in der Wirklichkeit verankert, ist der Fortschritt ein organisches Ganzes, dessen Wesen sich in der Entwicklung vollendet (Hegel 1991, S.24). Der Maßstab, den die Wissenschaft an sich selbst anlegt, dient der Vergleichung von sich mit sich selbst (S.76), sodass die Wissenschaft in ihrer Gesamtheit ein Kreislauf in sich selbst ist (Hegel 1990, I S.70). Kennzeichnend dafür ist der Ansatz, mit den Methoden der Statistik statistische Methoden zu verbessern. Das gilt auch für die Feststellung, dass signifikante und nicht-signifikante Befunde zu einem Forschungsgegenstand im Verhältnis 1:1 stehen, was eine Pattsituation indiziert, in der Stagnation wahrscheinlicher ist als Fortschritt (Schmidt & Hunter 2015, S.18).

Ist der Fortschritt unsicher, wird ein Maß für die Unsicherheit wünschenswert. Als Maß für die Unsicherheit dient die Replikationsrate, die den Wissenschaftsreformern viel zu gering ist: schließlich wird der Mangel an Replikationen verantwortlich gemacht für die Stagnation (Valentine et al. 2011) auf einem Gebiet, auf dem andere enorme Fortschritte erkennen (Stern 1911, S.iii). Die dramatische Darstellung der Häufigkeit falsch-positiver Befunde, gegen die Replikationen in Stellung gebracht werden sollen, gehört zu den Schattengeschichten der Psychologie (Rorty 2000, S. 396), die mit der Überwindung historischer Irrtümer dort einen Fortschritt suggerieren, wo es um die bloße Selbstvergewisserung geht, dass fehlende Replikationen ein echtes Problem sind und die Begriffe der Problemformulierung nicht angezweifelt werden.

Eine Fortschrittsgarantie gibt auch die 'neue' Statistik nicht (Ioannidis 2012); es sind immer wieder Phasen in der Wissenschaftsgeschichte möglich, in denen die Über-

zeugungskraft von Hypothesen schwindet, ja es ist noch nicht einmal ausgeschlossen, dass die Überzeugungen von Wissenschaftlern im Zuge der Herdenbildung auf eine falsche Hypothese konvergieren (Park, Peacey & Munafó 2014). Hypothesen überzeugen nicht mit ihrem aktuellen Wahrheitsgehalt, sondern mit dem Versprechen, eines Tages verifiziert zu werden. Es sind die vielversprechenden Hypothesen, denen Wissenschaftler nachgehen (Heisenberg 1973). Und diese Versprechen auf Verifikation soll u.a. mit Replikationen eingelöst werden. In der abgeschwächten Form einer bestätigenden Verifikation (Neurath 1979, S.136; Schlick 1986, S.146) legen Replikationen den Grundstein dafür, dass ein universaler Konsens – als Surrogat für Wahrheit – überhaupt möglich wird (Campbell 1921, S.29).

Die universale Gültigkeit ihrer Theorien ist ein Grundpfeiler der Wissenschaft (Pearson 1900, S.24): Naturgesetze gelten ausnahmslos, immer und überall (Lewin 1967, S.8). Die Naturgesetze des psychischen Geschehens sind es, die der Psychologie Rechenschaft geben über ihre Wissenschaftlichkeit (Wundt 1898, S.385). Mit der Gesetzmäßigkeit der Natur eng verknüpft ist der Begriff der Replikation (Schmidt 2009). Was nicht weiter wundert, ist eine Bestätigung einer Hypothese nur möglich gewesen unter Voraussetzung des Induktionsgrundsatzes, dass die Natur unter Beobachtung invariant bleibt. Doch selbst unter dieser Voraussetzung ist es mit der verifizierenden Funktion der Replikation nicht weit her: die klassische Mechanik wurde mit zahllosen Experimenten x-mal repliziert und verifiziert – und zu Beginn des 20. Jahrhunderts doch verworfen (Bondi 1975, S.3).

Die Verletzlichkeit von Naturgesetzen hat das Vertrauen der Wissenschaftler in ihre eigenen Errungenschaften erschüttert. Vermeintlich universal gültige Theorien wichen einer neuen Betrachtung der Naturgesetzlichkeit (Kuhn 1976, S.123f) und für gewiss gehaltene Überzeugungen wurden aufgegeben. Die Lektion der Wissenschaftsgeschichte lag darin, dass mehr und mehr Unsicherheit an die Stelle einstiger Gewissheiten treten konnte (Tukey 1969). Mit der wachsenden Unsicherheit erlebte die Statistik einen Aufschwung als Wissenschaft vom Umgang mit Vagheiten und Schwankungen (Savage 1972, S.154). Im selben Maße, wie Gewissheiten schwanden, rückte die Wahrscheinlichkeit nach vorne – um, wissenschaftlich gezähmt, Gewissheit wiederherzustellen (Nicod 1924 S.10). Die Wahrscheinlichkeit galt als Durchgangsstadium der Gnoseogenese von der Hypothese zur Gewissheit, solange die die Wahrheit determinierenden

Faktoren noch unentdeckt sind (Bernard 1947, S.60). Die Geltung wissenschaftlicher Aussagen erfuhr einfach eine stetige Differenzierung auf einer Skala, die von unmöglich bis gewiss reicht (Keynes 1973, S.10), mit dazwischenliegenden Gradierungen der Wahrscheinlichkeit.

Mit dem Argument, dass es schwierig bis unmöglich sei, sämtliche Einflussfaktoren eines Ereignisses aufzudecken, wird Wissenschaftlichkeit kompatibel gemacht mit Wahrscheinlichkeit (Neyman 1955). Die auf ihr abgelastete Wissenschaft kann die Wahrscheinlichkeit jedoch nicht tragen, weil sie relativ konzipiert ist (entweder relativ zu Häufigkeiten oder relativ zu einem subjektiven Wissensbestand) und nicht absolut, sie also für ihre Tragfähigkeit etwas benötigt (relevante Ereignisse oder relevantes Wissen), das sie tragen (taxieren) soll – sie soll das Ei ausbrüten, aus dem sie schlüpft. Das führt in die groteske Situation, dass ausgerechnet im Zuge der Grundlegung wissenschaftlichen Fortschritts auf ein solides Fundament dieses derart ins Schwanken gerät, dass es sich aufschaukelt zur Resonanzkatastrophe, die in Gestalt des Skeptizismus jedes wissenschaftliche Gebäude zum Einsturz bringt.

Statt eines aufragenden Gebäudes entpuppt sich der Fortschritt als wachsender Organismus, der vom Larvenstadium bis hin zum adulten Zustand für jedes Entwicklungsstadium passende Verhaltensmuster entwickelt, indem er experimentelle Befunde entweder direkt durch erfolgreiche Replikationen assimiliert oder indirekt durch eine vorausgehende konzeptuelle Akkomodation der Hypothesen (Piaget 1951, S.39). Der Organismus entwickelt somit einen Teil der Verhaltensmuster reflexiv entlang von Hypothesen zur Entwicklung eines Organismus. Heute würde man sagen, dass wissenschaftliche Hypothesen einem evolutionären Selektionsdruck ausgesetzt sind und nur die tauglichsten bzw. wahrscheinlichsten Hypothesen reproduziert und tradiert werden (Kantorovich 1993, S.3 u. 257).

Die Evolution der Wissenschaft vollzieht sich für Asendorpf et al. (2013) zu langsam, weshalb sie mit Replikationen und anderen Reformen den 'Fortschritt' beschleunigen wollen. Das geht in dem hier nachgezeichneten Verständnis einer statistisch-probabilistischen Wissenschaftlichkeit, derzufolge Gewissheit preisgegeben und auf eine vollständige Beschreibung der Wirklichkeit verzichtet wird (Krüger 1990, S.5), nur, wenn man das Ende des Organismus kennt oder seinen Zielzustand, der antrainiert werden soll. Ein solches Ziel ist wegen ihrer wissenschaftlichen Bedeutsamkeit eine hohe Replikations-

rate; dessen Rechtfertigung auf induktiv-statistischem Boden aufgrund eben der grundsätzlichen Fallibilität wissenschaftlicher Hypothesen (Nagel 1949, S.2) sich als unfruchtbar erwiesen hat, obwohl es diesem entwachsen ist. Der Rückzug von der Wahrheit in die Wahrscheinlichkeit nimmt der Replikation die Mittel zur Verifikation. Replikationen sind genauso fehlbar wie Originalexperimente, sodass dessen Beurteilung unverändert hypothetisch bleibt: das Ersetzen einer Hypothese durch eine andere entspricht schwerlich einer Begründung.

5.2.3.3 Versuch und Irrtum

Hypothesen sind keine Gründe, Hypothesen sind das, was begründet sein will. In der klassischen Mechanik hatten Hypothesen nichts verloren. Solange die Ursachen eines Phänomens nicht klar bewiesen sind, sollte man weiter experimentieren, statt Hypothesen zu fingieren (Newton 1999, S.943). Das Verdikt hielt nicht lange. Hypothesen waren leichter zur Hand als Beweise und immer mehr glichen sich – entgegen ihrer separatistischen Rhetorik – die Experimentalwissenschaften der spekulativen Philosophie an. Erst vereinzelt in Form von partikularen Hypothesen, die, unter anderem durch Replikationen, noch verifiziert werden konnten (Hanson 1958); doch bald schon wurden sie erhoben zum universalen Standard wissenschaftlicher Methodik (Whewell 1967, S.438; Popper 1989, S.199).

Die Verfechter der klassischen Mechanik benötigten keine Replikationen und nutzten sie bestenfalls aus Ehrgeiz und zur Steigerung ihrer Popularität; die Verfechter universaler Hypothesen verlangen keine Replikationen, es sei denn, ein Effekt ist so schwach, dass zufällige Schwankungen ihn überdeckt haben könnten (Oksanen 2001). Die Replikation kann dann die Überzeugung einer Person steigern, nicht aber mit dem universalen Fallibilismus aufräumen. Der bleibt, auch wenn die Replikationen heißlaufen (Klayman & Ha 1987). Weil die Überzeugungen losgelöst sind von ihrem Gegenstand – oder gar überhaupt keinen Referenten besitzen –, können insbesondere direkte Replikationen leicht falsche Überzeugungen festigen. Denn direkte Replikationen replizieren auch die Fehler der Originalstudie und tragen so zu ihrer Verbreitung bei (Rosenbaum 2001; Ioannidis 2012; McElreath & Smaldino 2015).

Wie auch immer eine Replikation ausgeht, das statistisch-induktive Modell bietet keine Interpretation des Befundes, die sich begründet ausweisen könnte gegenüber konkurrierenden Interpretationen (Earp & Trafimow 2015). Das nomologische Netz einer Theorie, die bei einer solchen Begründung assistieren könnte, muss, da sie den – angestrebten – Abschluss einer Serie von Experimenten markiert, also am Ende steht, nach dem Argument des verallgemeinerten Übertragungsfehlers (Neta 2016) seine Strapazierfähigkeit oder Robustheit (Nowotny, Scott & Gibbons 2005, S.210) aus anderen Quellen beziehen als aus den Experimenten (Simon 1955). Selbst wenn man die Fähigkeit zur Unterscheidung erfolgreicher und gescheiterter Replikationen zugesteht, sind die Verfechter von Replikationen immer noch mit interpretatorischen Schwierigkeiten konfrontiert, weil die Replikation hochwertiger Experimente häufig gelingt, wenn ihre Rückschlüsse korrekt sind; die Replikation minderwertiger Experimente gelingt aber auch dann häufig, wenn ihre Rückschlüsse inkorrekt sind (Rosenbaum 2001).

In einer fallibilistischen Wissenschaft lauern nicht nur Fehler erster Art, sondern aller Art. Die bekanntesten sind der (zufällige) Stichprobenfehler auf der phänomenalen Seite und der (systematische) Messfehler auf der noumenalen Seite, der wiederum unterteilt ist in Fehler, die aus einem lückenhaften Modell resultieren, und Fehler, die technischen Grenzen geschuldet sind (Youden 1972). Messfehler könne nicht ausgeschlossen werden, weil es keine Messinstrumente mit überstarrer Mechanik gibt; und Stichprobenfehler können nicht ausgeschlossen werden, weil der Stichprobenumfang niemals unendlich ist (Schmidt & Hunter 2015, S.35). Replikationen minimieren zwar den Stichprobenfehler durch Vergrößerung des Stichprobenumfangs, der die Standardabweichung der Stichprobe verengt auf den Standardfehler der Population. Doch ob dabei auch der systematische Fehler hinreichend reduziert wird, lässt sich angesichts der kategorialen Dichotomie nicht entscheiden (Cochran, Mosteller & Tukey 1954). Der systematische Fehler kann nur aus sich selbst heraus ausgemerzt werden: Die Überzeugung signalisiert die Abwesenheit von Fehlern, der Zweifel deren Präsenz, und beide entspringen demselben Organ, in dem sie angewiesen sind auf gegenseitige Korrektur, um das Falsche nicht zu behaupten und das Wahre nicht zu leugnen.

Die Überzeugung von der Richtigkeit einer Hypothese steht auf einer Stufe mit dem Bewusstsein möglicher Fehlerquellen. Was letztlich überwiegt, lässt sich genauso wenig

weder durch Experimente noch durch Nachdenken entscheiden, wie die sich anschließende Frage, ob die Entscheidung ein Fehler war. Die Statistik kann Fehler nicht kompensieren, wie noch Borel (1933 S.x) dachte. Die Rede von der Kontrolle der Fehler, erster oder zweiter Art, täuscht eine Sicherheit vor (Loscalzo 2012), die methodisch unbegründet ist. Stanley und Spence (2014) simulierten ideale Replikationen, indem sie ausschließlich den zufälligen Fehler manipulierten; schon die geringste Manipulation des Fehlers erzeugte ein breites Spektrum variierender Befunde – eine weitere Warnung vor allzu großer Zuversicht bei Replikationen. Familienähnliche Begriffe wie Wahrscheinlichkeit, Fehlbarkeit, Fehler oder Hypothesen verdichten sich in der Statistik zu einem hermetisch abgeschlossenen Konglomerat, sodass der Eindruck entsteht, als würden sich die Begriffe gegenseitig bedingen.

5.3 Zweifel am Replikationserfolg

Der Zwiespalt zwischen Reformieren und Bewahren im Reproduzierbarkeitsprojekt: Psychologie wird deutlich am Verhältnis der Replikationen zur kanonischen Statistik. Doch auch die vermeintlich neue Statistik mit Effektgrößen, Konfidenzintervallen und Meta-Analysen hat, wie sich zeigen wird, kräftige Wurzeln in der kanonischen Statistik.

5.3.1 Statistik-Recycling

Dass der Geist der Statistik die Experimentalforschung beherrscht, wird gerne beklagt. Dabei ist es einerlei, wer wen gerufen hat, der Statistiker den Experimentator oder umgekehrt. Im Ergebnis hat man Wahrscheinlichkeiten statt Gesetzlichkeiten. Es geht darum, möglichst viele Evidenzen beizubringen und sie zu Schätzwerten zu aggregieren, statt jede Evidenz einzeln zu würdigen. Infolge der 'Okkupation' der Experimentalforschung durch die Statistik wird die Wiederholbarkeit zu einer Hauptforderung, die an ein Experiment zu stellen ist (Lewin 1967, S.10). Selbst Verfechter der kanonischen Statistik beginnen, Replikationen wissenschaftliche Bedeutung einzuräumen (Chow 1998). Die Statistik als universales Forschungsinstrument, das in wiederholter Anwendung experimentelle Ergebnisse vergleichbar macht, bereitet Kumulation und Fortschritt gleichermaßen den Boden (Mischel 2009).

Innerhalb der vorherrschenden kanonischen Statistik thronen Doppeltblindstudien, in denen auch der Experimentator nicht weiß, ob die von ihm untersuchte Person der Versuchs- oder Kontrollgruppe angehört (Hill 1971, S.155; Cochrane 1972, S.22), über allen anderen statistischen Designs, egal wie oft diese wiederholt werden (Worral 2010). Die Königsdisziplin besteht folglich in der Replikation von Doppeltblindstudien (Rawlins 2008). Das sklavische Treten der Datenkelter erweckt den Eindruck, dass mehr und mehr statistische Hypothesen verwechselt werden mit substantiellen Theorien (Meehl 1967). Die Theorien der Psychologie gleichen einer generischen Abfolge linearer Modelle (Loftus 1996), deren Strukturen sich auf die psychologischen Modelle übertragen. Derart methodomorphe Theorien (Danziger 1985) können nicht überwunden werden mit den Methoden, an denen sich die Psychologie infiziert hat. Solange die kanonische Statistik in den Experimentalwissenschaft vorherrscht, können keine Gegenevidenzen aufkommen, weil sie nicht der Methodologie entstammen, die das Monopol innehat auf Begründung der Validität einer Evidenz, und insofern eine Gegenevidenz gar nicht erst als evident anerkannt werden würde. Verpflichtet man sich einer Methodik, gibt es aus dem Zirkel der Selbstbestätigung kein Entrinnen mehr. Es erstaunt daher nicht, dass gegen die Methodenmonokultur in politischen Begrifflichkeiten wie der Anarchie (Feyerabend 1991, S.246) oder des Pluralismus (Asendorpf et al. 2012) angeschrieben wird.

Solche methodologische Überlegungen werden gerne abgetan als Wissenschaftsfolklore oder epistemologische Idiosynkrasien, die der Wissenschaft keinen Mehrwert verschaffen (Medawar 1968, S.29; Feynman 1985, S.345; Hawking 2015; Rusconi et al. 2016). Doch welcher Wissenschaft? Welchen Mehrwert? Genau das ist ja Gegenstand der philosophischen Wissenschaftstheorie. Die Philosophie hat nicht nur eine längere und globalere Geschichte als die Experimentalwissenschaften, sie versteht auch immer schon – im Unterschied zu ihrem jüngsten Spross – die Rechtfertigung ihres Daseins und Soseins als Pflicht. Insofern ist die Philosophie stets Wissenschaft reflektierter Wissenschaft. Das bedeutet keine Privilegierung der Reflexion gegenüber der Aktion, in dem Sinne, dass die Philosophie der Experimentalwissenschaft ihre Methodik vorschreiben könnte. Will die Philosophie den Spiegel im Spiegel vermeiden, um nicht in einen vitiösen Zirkel oder infiniten Regress einzutreten, kann die reflektierte Reflexion selbst nur kontingente Aktion sein. Als solche steht sie gleichberechtigt neben den Experimentalwissenschaften. Und wie bei ihnen genügt die Kontingenz dessen, dass

sich Menschen mit philosophischen Fragen beschäftigen, zur Rechtfertigung ihre Existenz als einer sozial bedeutsamen. Wer also wie Murtaugh (2014) effektive Studien inklusive Replikationen fordert und ein Ende der Diskussion über Methoden der Datenanalyse, fordert eine Methodik ohne Methodologie und lässt so einen vermeidbaren Mangel an Reflektiertheit durchblicken.

Replikationen gleichen dem Trojanischen Pferd, das die Instrumente (Signifikanz, Konfidenzniveaus, *p*-Werte, Standardfehler und dergleichen) in seinem Innern mit sich führt, deren Versagen erst die hölzerne Konstruktion von Replikation veranlasst hat: als wollte man mit einem Eimer Wasser einen Sumpf trockenlegen, um ihn zu überqueren. Doch Unsicherheit lässt sich statistisch nicht vermeiden, bestenfalls lässt sich Wahrerscheinlichkeit minimieren. Für die Stabilität eines gangbaren Fortschritts benötigt die Statistik extra-statistische Assistenz. Dann aber ist Signifikanz überflüssig, wenn ein Experiment replizierbar ist, und sie ist bedeutungslos, wenn ein Experiment nicht replizierbar ist (Meehl 1990).

Schwierigkeiten mit dem Begriff der Replikation tauchen immer dann auf, wenn Replikationen kriterial eingesetzt werden sollen für den Nachweis des Erfolges eines Experimentes bzw. für die Bestätigung einer Hypothese. Dann eröffnet sich das probabilistische Dilemma: der Experimentalforscher erwartet von der Statistik Auskunft darüber, wie wahrscheinlich seine Hypothese zutrifft und wie sich deren Wahrscheinlichkeit verändert infolge von Wiederholungen des Experiments (Berkson 1942). Das kann die kanonische Statistik mit dem Signifikanztest an der Spitze jedoch nicht leisten. Denn im Kern fragt der Experimentator den Statistiker, was er tun muss, um seine Behauptung zu verifizieren. Dazu kann der Statistiker in der Regel nichts sagen. Was er aber sagen kann, ist, wann ein experimenteller Befund signifikant ist.

Ein signifikanter Befund bedeutet bekanntlich keine Bestätigung der Alternativhypothese. Dieses Defizit meint Berkson (1970, S.287) durch die Entscheidung heilen zu können, im Falle einer verworfenen Nullhypothese die Alternativhypothese einfach als bestätigt zu betrachten. Das aber führt geradewegs ins nächste Dilemma. Prüfte Berkson die Voraussetzungen seiner Entscheidung, käme er auf einen statistischen Test, dessen Anwendung seinerseits an bestimmte Voraussetzungen wie Unabhängigkeit der Stichprobe oder Homoskedastizität geknüpft ist. Der Entscheidung für eine Hypothese geht somit die Entscheidung voraus, mit welchem Test der experimentelle Befund analysiert

werden soll, der wiederum die Prüfung vorausgeht, ob die Voraussetzungen für die Anwendung des Tests erfüllt sind. Im Entscheidungsdomino muss der Forscher sich ohne Hilfestellung der Statistik entscheiden, ob er die Voraussetzungen prüfen möchte oder nicht. Prüft er sie, kann der Forscher durch die sich anschließende Entscheidung für einen der anwendbaren Tests manipulativen Einfluss nehmen auf die Höhe der Signifikanz; prüft er sie nicht, dann stimmt möglicherweise die Höhe der Signifikanz nicht, weil ein Test zur Anwendung kam, dessen Voraussetzungen nicht erfüllt waren (Ng & Wilcox 2011).

Weil in der Welt alles unsicher, zweifelhaft und fehlbar geworden ist, muss man allem Sicheren, Zweifelhaften und Fehlbaren mit einer Statistik beikommen, um valide Resultate zu erzielen. Das ist ein Trugschluss. Ein statistischer Test gibt aus, womit man ihn füttert: Fallibilität. Statistisch signifikante Ergebnisse besagen nur, dass Hypothesen fallibel sind (Serlin 1987). Denn die Schlussfolgerungen aus einem Signifikanztest sind nur vorläufige (Fisher 1955). Und mit statistischen Tests verharrt man in der Dimension, die ihr zugedacht wurde: der phänomenalen Dimension der Stichprobe. In der noumenalen Dimension ist die Population eindeutig definiert und die Schlussfolgerungen sind logisch; in Stichproben, auf die die Tests angewendet werden, kann das Testergebnis dagegen einer Vielzahl von Populationen zugehören. Denn die Population, auf die geschlossen werden soll, ist eine unendliche Klasse von im wesentlichen fiktiven Ereignissen, die alle vergangenen, gegenwärtigen und zukünftigen Ereignisse umfasst unter einer spezifizierten Experimentalbedingung. Diese Schlussfolgerung gelingt mit einem Signifikanztest nur dann, wenn die Population selbst eine Zufallsstichprobe aus der unendlichen Klasse ist. Daher können Überzeugungen in der Population nicht einfach übertragen werden auf Testergebnisse von einer Stichprobe. Sollen Signifikanztests Überzeugungen verstärken, muss etwas aus der Population hinzukommen (Fisher 1955).

Das schwächste Bindeglied zwischen Stichprobe und Population ist der Standardfehler, der umso kleiner ausfällt, je größer die Stichprobe aus einer Population ist. Macht man die Stichprobe nur groß genug, wird jeder Test signifikant und jede Nullhypothese verworfen (Bakan 1966; Meehl 1978; 1990). Das gilt selbstverständlich für jede andere Hypothese auch (Jones, Derby & Schmidlin 2000). Ist die Auflösung eines Tests groß genug, werden noch so kleine Abweichungen sichtbar und Nullhypothesen zu absoluten Begriffen, die sich jeder Verifikation entziehen (Unger 1975, S.202). Insofern befördern

Replikationen die Verwerfung von nahezu allen Hypothesen. Was übrig bliebe, wäre ein löchriger Flickenteppich von aneinandergereihten Befunden aus zusammenhangslosen, freistehenden Tests (Ashcroft 2004). Replikationen knüpfen kein, nomologisches Netz, in dem die Knoten Theoriebestandteile repräsentieren, die über Kanten in einer funktionalen Beziehung zueinander stehen (Meehl 1978). Replikationen affirmieren den Ist-Zustand der Wissenschaft je nach Einsatz und untermauern das probabilistisch Weltbild. Das wirkt eher reaktionär als reformerisch.

Die Replikation kann epistemologisch nicht mehr leisten als der Signifikanztest, den sie hinter ihren Bohlen verbirgt; in ihrem Innern führt sie die Methoden der Studien mit, die sie repliziert. Das sind in der Regel Signifikanztests, die weder ein Passepartout noch ein Panaceum verkörpern für die Lösung wissenschaftlicher Probleme (Cantor 1932). Eine gedankenlose Anwendung des Signifikanztests, die sich am Schließautomatismus in der Psychologie zeigt (Bakan 1966), verbietet sich ganz offensichtlich in der Differenzielle Psychologie oder Psychotherapie (Toomela 2007). Der Geltungsbereich statistischer Schlüsse ist beschränkt auf Klassen und Teilklassen oder Aggregate; zu Individuen sagen sie nichts aus. Weniger offensichtlich ist die Beschränkung ihres Geltungsbereichs auf Stichproben, auf hypothetische Populationen; auf die Population an sich fällt nur ein Schlagschatten. Ist man sich dieser Schwächen bewusst, kann der Signifikanztest schadlos eingesetzt werden (Greenwald et al. 1996), aber eben nicht mehr ubiquitär. Wenn, wie bereits für die Replikationen festgestellt, ein gutes Experiment in der Regel keinen Signifikanztest benötigt, folgt umgekehrt, dass ein Experiment schlecht ist, sollte ein Signifikanztest nötig sein (Healy 1978).

Statistische Signifikanz deckt sich nicht mit wissenschaftlicher Signifikanz (Johnson 1999), weshalb kein isolierter Test ausreicht, wie signifikant auch immer, zur Etablierung einer objektiven Erkenntnis (Fisher 1935, S.14). Allerdings reicht auch das mehrmalige Verwerfen einer Nullhypothese dazu nicht aus. Wichtige Fortschritte erzielt man in der Wissenschaft häufig ohne Tests, sehr selten aufgrund von Tests und manchmal trotz der Tests (Morrison & Henkel 1970, S.311). Dass Signifikanztests den Fortschritt behindern (Cohen 1994), liegt weniger am Test selbst als vielmehr an seiner unreflektierten Anwendung. Wüssten Psychologen was sie tun, würden sie den *p*-Wert nicht als Evidenzkriterium interpretieren.

Der *p*-Wert ist die Wahrscheinlichkeit, eine Stichprobe zu erhalten, die mindestens so extrem ist wie die erhobene unter der Annahme, dass die Nullhypothese wahr ist. Das ist keine Evidenz einer Stichprobe für die Population, und das ist noch nicht einmal Evidenz einer Stichprobe für Stichproben, weil der *p*-Wert stark variiert (Cumming 2008). Infolgedessen überschätzt der *p*-Wert die Evidenz gegen die Nullhypothese (Goodman 2001), die gegenüber dem *p*-Wert jede beliebige Größenordnung annehmen kann (Berger 1987). Das liegt an den vielen Faktoren, die außer der Effektgröße den *p*-Wert beeinflussen, wie Standardabweichung, Stichprobenumfang, Nichtlinearität oder Heteroskedastizität (Berkson 1942). Daher gibt der *p*-Wert höchstens einen äußerst schwachen Hinweis auf das Vorzeichen der Effektgröße im Falle einer Replikation (Jones & Tukey 2000). Im Kontext der Reproduzierbarkeit eines Experimentes reicht es nicht aus, zu wissen, wie selten eine Stichprobe ist bei Gültigkeit der Nullhypothese, ohne zu wissen, wie selten die Stichprobe ist bei Gültigkeit alternativer Hypothesen (Barber & Ogle 2014).

Einen solchen Vergleich unternimmt der Bayes-Faktor. Beim subjektivistischen Ansatz wird gar nicht erst versucht, den *p*-Wert in Verbindung zu bringen mit der Population; hier steht der *p*-Wert für Eigenschaften der Stichprobe selbst (Burnham & Anderson 2014). Allerdings variieren die Schätzer des Subjektivismus ebenfalls stark, sodass die Kritik an der Signifikanz- und Entscheidungstheorie bezüglich Kriterien der Reproduzierbarkeit auch am Subjektivismus greift (Garcia-Pérez 2012). Das liegt daran, dass der Subjektivismus entweder eine Ausgangsverteilung der Parameter bestimmen muss unter Einbeziehung sämtlicher Befunde oder die Dichte einer Ausgangswahrscheinlichkeit, die neutral ist gegenüber allen Hypothesen über den Parameter (Efron 1978). Diese Festlegung ist nicht nur subjektiv, sie zieht auch die Schwierigkeiten fingierter Stichprobenverteilungen bei der Beurteilung von Hypothesen nach sich.

In der Signifikanz- und Entscheidungstheorie erlauben die Schlüsse keine Angaben zur Wahrscheinlichkeit von Hypothesen. Denn in aller Regel entsprechend sich $P(H_0|b)$ und $P(b|H_0)$ nicht. Ohne die Unterscheidung von Wirklichkeit und Erscheinung entgeht man dadurch einem vitiösen Begründungszirkel. Macht man aber die Wahrerscheinlichkeit zum Fundament der Wahrscheinlichkeit, schreibt also der Wahrscheinlichkeit eine subjektive und eine objektive Seite zu (Edwards 1963), ist die Konsistenz des Wahrscheinlichkeitskalküls nur auf einer Seite erreichbar. Der Bayes-Faktor bezieht

aus Likelihood und bedingter Wahrscheinlichkeit zwar die Form einer Verknüpfung von Objektivität mit Subjektivität, von Parameter mit Schätzwert/Testergebnis, doch das Verhältnis zweier Likelihoods lässt sich nur im Rahmen subjektiver Überzeugungen widerspruchsfrei interpretieren, denn bei einem gegebenen Befund können zwei Hypothesen im selben Verhältnis zueinander stehen, wie zwei andere Hypothesen zum selben Parameter, deren Größe sich aber ergibt aus einem völlig anderen Befund. Die Verhältnisgleichheit ist nur nachvollziehbar, wenn die Befunde interpretiert werden als individuelle Evidenz (Hacking 1972). In der subjektivistischen Wendung des Induktionsproblems könnte nur subjektive Gewissheit objektive Wahrheit garantieren. Doch auch die stärkste Überzeugung, die auf vergangenen Befunden basiert, ist nicht notwendig gewiss, ihre Validität ist stets bezweifelbar (Akaike 1998, S.428). Gewissheit müsste dagegen notwendig vorliegen bei jeder logisch wahren Aussage, bzw. müsste die Überzeugung null sein bei jeder Kontradiktion (Earman 1996, S.5), die alle unabhängig sind von empirischen Befunden – was nicht der Fall ist. Insofern eignet sich die Likelihood nicht zum statistischen Fundament (Oakes 1985, S. 13).

Der Subjektivismus ist nicht konzipiert zur Erkenntnis der Wirklichkeit. Seine statistischen Testergebnisse sind nur dort sinnvoll, wo man etwas über den Grad der Überzeugung einer Person erfahren möchte (Mayo 1983). Subjektivisten nutzen die Befunde von Replikationen, um die Wahrerscheinlichkeit, d.h. den Grad ihrer Überzeugung an die Befunde anzupassen. Das führt zur Modifikation von Hypothesen, aber weder zu ihrer Verwerfung noch zu ihrer Generalisierung noch zur Innovation alternativer Hypothesen (Gelman & Shalizi 2013). Vielmehr eröffnet der Subjektivismus Forschern zusätzliche Möglichkeiten, ihre Befunde so lange zu durchforsten, bis sie darin einen berichtenswerten Indikator finden und ein weiteres falsch-positives Resultat in die Welt setzen (Simmons et al. 2011). Die bedingte Wahrscheinlichkeit ist letztlich Ausdruck methodenübergreifender Klugheit in dem Sinne, dass voreilige Schlüsse vermeidbar sind unter Berücksichtigung relevanter Randbedingungen (Hogben 1957, S.25).

Angesichts des Missverhältnisses von Anspruch und Wirklichkeit der kanonischen Statistik wird verständlich, dass unter der Fahne der Replikationskrise methodische Neuerungen erkämpft werden sollen, die Replikationen nicht nur verlangen, sondern Replikationen auch effektiver machen im Auftrag wissenschaftlichen Fortschritts. Diese Neuerungen sind eher Neubewertungen bereits etablierter Instrumente oder haben

zumindest dieselben Wurzeln wie die Signifikanz- und Entscheidungstheorie oder der Bayes-Faktor. Die zögerliche Radikalität ist im Einklang mit der Reformorientierung ihrer Protagonisten; einen revolutionären Befreiungsschlag aus der Krise meiden sie trotz vereinzelter martialischer Rhetorik. Ihre Reformkandidaten heißen Effektgröße, Teststärke, Konfidenzintervall, Replikation und Meta-Analyse.

5.3.2 Effektgröße, Teststärke, Konfidenzintervall, Meta-Analyse

Inzwischen gibt es einen artenreichen Zoo an Effektgrößen, die alle in verwandtschaftlicher Beziehung stehen und letztlich abstammen von Cohens *d* – der standardisierten Effektgröße, die eine Mittelwertdifferenz ins Verhältnis setzt zur Standardabweichung (Cohen 1977, S.20). Die Effektgröße ist also eine Eigenschaft der Stichprobe, nicht der Population. Deshalb ist der Rückschluss auf die Population ein die Stichprobe transzendierender Schluss, und folglich inkohärent (Oakes 1983, S.93). Bedenkt man darüber hinaus, dass sich der Maßstab für die Effektgröße ändert mit dem Stichprobenumfang, dann zeigt die Effektgröße ähnliche Schwächen wie das Signifikanzniveau (Murray & Dosser 1987). Standardisierte Effektgrößen verlieren an Aussagekraft, wenn das Treatment die Variation beeinflusst (Cumming 2012), weil die unterschiedliche Variation sich nicht unterscheiden lässt von unterschiedlichen Effekten (Greenland, Schlesselman & Criqui 1986).

Die Effektgröße geht mit Stichprobenumfang, Signifikanzniveau und Standardabweichung direkt ein in die Berechnung der Teststärke. Dadurch infiziert sich die Teststärke mit der Fehlervariation und wird abhängig von den destabilisierenden Faktoren der kanonischen Statistik (Muller & Benignus 1992). Die Teststärke ist zentraler Bestandteil der Entscheidungstheorie; ohne die Absicht, eine Entscheidung zu konkurrierenden Hypothesen zu treffen, wird die Teststärke nicht benötigt (Gigerenzer et al. 1989). Die Teststärke soll sicherstellen, dass ein Effekt, sofern er in der Population existiert, mit dem Test nachgewiesen wird. Ein Test mit nicht-signifikantem Befund hat post hoc immer eine geringe Teststärke, unabhängig vom Stichprobenumfang (Nakagawa & Foster 2004). D.h., ein hoher *p*-Wert impliziert eine geringe Teststärke. Somit spräche eine geringe Teststärke nicht gegen eine Beibehaltung der Nullhypothese, – will man nicht die Evidenzlogik des Signifikanztestes ad absurdum führen. Vielmehr lässt sich mit

einer hohen Teststärke der *p*-Wert beliebig minimieren (Meehl 1978). Damit verschafft die Teststärke keine objektive Evidenz.

Statt der gewünschten Evidenz verbleibt einem unter Berücksichtigung der Teststärke auch nur die Konfidenz. Konfidenzintervalle werden gewöhnlich dann berichtet, wenn die p-Werte bescheiden sind (Ioannidis 2014). Sie bestehen aus der Menge aller Parameterwerte, die zum festgelegten Konfidenzniveau nicht verworfen werden können, d.h. aus den Werten, die bei einem Signifikanztest *p*-Werte hervorbrächten, die kleiner sind als das Alphaniveau. Konfidenzintervalle setzen sich wie die *p*-Werte zusammen aus der Parameter-Punktschätzung, dem Standardfehler und der Stichprobenverteilung. Liefern die *p*-Werte keine Evidenz, so tun das Konfidenzintervalle auch nicht (Murtaugh 2014). Ohne suffiziente Statistik sind Konfidenzintervalle nichtssagend (Oakes 1985, S.129). Allerdings kommt man zu suffizienten Statistiken nur, wenn man das Induktionsproblem im Messraum gelöst hat und alle Faktoren eines Effekts im Stichprobenraum bestätigt sind.

Replikationen sind allerdings nicht nur angedacht als Mittel zur Überwindung des Induktionsproblems, sondern auch als Mittel zur Beschreibung der Gegenwart in Wissenschaft und Forschung. So soll anhand von Replikationen die gesamte Replikationsrate gemessen werden gegenüber der durchschnittlichen Teststärke der replizierten Elemente (OSC 2012). Original und Replikationen sollen demnach meta-analytisch ausgewertet werden. Von Meta-Analysen verspricht man sich einen Mehrwert an Evidenz gegenüber isolierten Replikationen. Die Lebenslinie von replizierten Effekten verläuft deutlich kumulativer, wenn sie meta-analytisch zusammengefasst werden, als wenn man bloß ihre Abfolge nach dem Signifikanzkriterium betrachtet (Braver, Thoemmes & Rosenthal 2014).

Ziel einer Meta-Analyse ist das perfekte Experiment: Es wird geschätzt, wie die Befunde der Experimente ausgesehen hätten, wenn sie ohne Beschränkungen ausgeführt worden wären. Durch die Schätzung sollen Eigenschaften der Population auf der Stichprobenebene gewissermaßen kalibriert werden (Schmidt & Hunter 2015, S. 36 u. 556). Zur Ausmittelung des Stichprobenfehlers wird der gewichtete Durchschnitt und die Varianz der Parameterwerte berechnet. Daraus ergibt sich die Varianz, die allein aus dem Stichprobenfehler zu erwarten ist. Die erwartete Stichprobenvarianz wird abgezogen von der beobachteten Stichprobenvarianz. Ist das Ergebnis null, gehen die unter-

schiedlichen Parameterwerte aus den Studien zurück auf den Stichprobenfehler (Hunter & Schmidt 1996).

Die prima facie-Vorteile sind statistischer Natur: höhere Teststärke, Gewichtung nach Stichprobenumfang oder geringere Anfälligkeit gegenüber Ausreißern (Head et al. 2015). Verrechnet man ungeprüft die berichteten Parameterwerte, erhält man mit Meta-Analysen einfach zusätzliche Statistiken, die die Schwächen ihrer Zeitgenossinnen teilen (Guttman 1985). Differenziert man nach der Güte eines Experiments, fließen in die Meta-Analysen Beurteilungen ein, die entweder dem Ergebnis schon vorgreifen oder bloß subjektiv sind (Eysenck 1994); sie sind transzendent oder konfident, nicht aber evident. Diesfällig sind Meta-Analysen teilweise irreführend, teilweise suboptimal (Hedges & Olkin 1985), als Maß für Evidenz sind sie nicht geeignet (Oakes 1985, S.163). Zur Kalibrierung fehlt ihnen die Eichgröße. Die bisherigen Erörterungen resümierend könnte man Meta-Analysen schlicht vorhalten, induktiv zu sein, oder Teil des Induktionsproblems (Eysenck 1995).

Als Voraussetzung für Meta-Analysen können Replikationen epistemologisch nicht vorgebracht werden, weil Meta-Analysen ebenso wenig Evidenz kumulieren wie Replikationen. Dass es mit einer einfachen Replikationsbatterie nicht getan ist, wissen auch ihre Verfechter. Weil eine Replikation keine belastbaren Rückschlüsse zulässt, führt Rosenthal (1990) weitere Kriterien an, wie Homogenität der Befunde oder Unabhängigkeit der Studien, am besten noch versehen mit einem Replikationsindex, in dem die Noten zusammengefasst sind, die unabhängige Experten vergeben haben für die jeweilige Methodik von Untersuchungen eines Effektes. Hier schließt sich denn auch der Kreis wissenschaftlicher Selbstkorrektur: auch Replikationen sind etwas, das Wissenschaftler korrigieren können.

6 Die Replikationsindustrie in der Wissensgesellschaft

Wenn aber eine Replikation, ja schon die bloße Replizierbarkeit nicht von Bedeutung ist für die Validierung experimenteller Befunde, und mithin der Fähigkeiten eines Experimentalforschers (Jeffreys 1973, S.204), was sagt dann die Reproduktionsrate aus? Dass ein Erfolg der nächsten Replikation unwahrscheinlich ist? Dass das Induktionsproblem unlösbar ist? Dass die Psychologie keine Wissenschaft ist? Oder dass in der Psychologie einiges verbessert werden muss? Viermal Nein!

Der Erfolg der nächsten Replikation hängt durchaus ab von der Reproduktionsrate, insofern nämlich, als die Psychologen in irgendeiner Weise auf die veröffentlichte Reproduktionsrate reagieren und gegebenenfalls auf die Replizierbarkeit ihrer Experimente mehr achten werden – je nachdem, wie groß der innerwissenschaftliche Selektionsdruck wird, erfolgt auch die Adaptation der Wissenschaftler und ihrer Theorien (Karmiloff-Smith & Inhelder 1977). Außerdem sollte, insofern die Reproduktionsrate eine Reproduktionsrate bedeutsamer Effekte ist, einer unabhängigen Replikation gerade ihre Bedeutsamkeit entgegenstehen, weil bedeutsame Effekte die Öffentlichkeit erreichen, und damit die Replikationen mit völlig anders informierten Versuchspersonen erfolgen als im Original, dessen Versuchspersonen noch nicht von dem Effekt gehört haben konnten (Stroebe & Strack 2014).

Das Induktionsproblem ist nur konsistent unlösbar. Und konsistent unlösbar heißt, dass eine Lösung regelwidrig denkbar ist. Eine solche Regelwidrigkeit bestünde in der Änderung der Regeln oder der Hinzunahme neuer Regeln. Regeln, die den Übergang vom Einzelnen zum Allgemeinen erlauben. Mit der kategorialen Scheidung zwischen Wirklichkeit und Erscheinung, zwischen Population und Stichprobe ist das epistemologische Schicksal der Statistik bereits besiegelt. Es braucht eine Regel, um zwischen beiden Begriffswelten zu vermitteln; wodurch zwei weitere Vermittlungsregeln erforderlich würden, usw. Eine solche Regel ist der Induktionsgrundsatz, der die Formulierung des Induktionsproblems so normiert, dass es ohne ihn nicht lösbar ist. Relevant sind allerdings nur die Regeln, die auch befolgt werden (Wittgenstein 1990, §§185-242). Dass zwischen einer Regel und ihrem Befolgen keine innere Notwendigkeit besteht (Kripke 1989, S.56), dass man beispielsweise ab und zu Rechenregeln verletzt, oder dass zu einem Verhalten meist mehrere Regeln passen, ist nur Ausdruck der logonormen Regel für 'Regelfolgen'. Dieser Zirkel eines sich reflexiv selbst normierenden

Sprechverhaltens ist nicht vitiös, weil wir in der Sprache über die Sprache sprechen. Die Sprache nimmt daran keinen Schaden.

Daher ist es auch kein Schaden, wenn es in der Schließenden Statistik nichts (rück-)zuschließen gibt (Simonton 2014). Statistische Schlüsse genießen gegenüber anderen Schlüssen keine Privilegien; aus sich heraus begründet kein Schluss seine Norm. Die Gründe, die Norm eines Schlusses rechtfertigen, sind dieselben, die seine Gültigkeit rechtfertigen. Diese Gründe sind manchmal in unserem Verhalten zu finden, ihren Ausdruck finden sie darin immer. Replikationen werden zur Norm, wenn Forscher replizieren; warum sie das tun, ist damit noch nicht gesagt. Das gilt auch für die konfidenzsteigernde Funktion von Replikationen (Nosek et al. 2015): wenn eine erfolgreiche Replikation als Evidenz gelten soll, müssen Gründe dafür dargetan werden, die nicht auf der Evidenz einer Replikationsnorm beruhen, also Gründe, die außerhalb der philosophischen Grammatik (Wittgenstein 1990, §295) liegen, – unter der Voraussetzung, dass Wissenschaftler ihr Verhalten nicht beeinflussen lassen von vitiös-zirkulären Argumenten.

Die Psychologie ist schon deshalb eine Wissenschaft, weil sie, unter anderem, um den Begriff der Replikation ringt. Verpflichtet man sich in diesem Ringen auf die Norm elegant polierter Trivialitäten eines Induktionskalküls, dann sind Replikationen empirische Schnipsel, mehr nicht (Allport 1968, S.8). Sprachliche Engpässe, die andere (Smith & Smith 2014) Engstirnigkeit nennen, wirken wie ein methodischer Hemmschuh, demzufolge jedes Phänomen in dieselbe Form gepresst werden muss, um ihm das Siegel der Signifikanz einzuprägen, wie der kritische Rationalismus glaubte, der Wissenschaft das Siegel der Falsifizierbarkeit einprägen zu müssen. Popper und die Neopositivisten sind dem Methoden-Marketing einer sich emanzipierenden Naturphilosophie aufgesessen, dessen Exponenten von Hooke (1707, S.3-70) bis Bois-Reymond (1884) reichen, und das in der Auseinandersetzung zweier Kulturen (Snow 1959) gipfelte, die zur Jahrtausendwende neu aufflammte (Sokal & Bricmont 1997, S.14f).

Die wissenschaftliche Revolution entzauberte die Welt nicht nur, sie machte aus ihr einen sterilen Automaten, der fehler- und reibungslos funktioniert: Beherrschen durch Berechnen (Weber 1985, S.593) verspricht, Bürden und Risiken des Lebens in den Griff zu bekommen, vorausgesetzt, man besitzt vollkommene Kenntnisse der Weltmechanik (Laplace 1840, S.4). Die Mechanik löst dieses Versprechen nur insoweit ein, als wir

unser Handeln orientieren an der in der Grammatik der Mechanik normierten Kausalzusammenhänge und uns nicht daran stören, dass wir uns ihnen experimentell nur annähern können. Der Erfolg der Methodenmonokultur liegt darin, dass sie die Früchte erntet, die sie sät, ohne je den Gedanken an anderes Saatgut aufkommen zu lassen.

Das eschathologische Moment der wissenschaftlichen Moderne wirkt im Alltag der Wissenschaften, nicht aber am letzten Tag der Wissenschaften: am Tag der Philosophie. Dort wird über die Gründe Gericht gehalten. Ein Vertrösten auf die Zukunft ist dann nicht mehr möglich, dann münden Grammatik und Rhetorik in Sprechverhalten (1. Johannes 2, 3-6), wo nur noch Welt ist, ohne stützende Theorie – wie praktisch sie auch immer gewesen sein mag (Lewin 1951, S.169). In einer derart archaischen oder besser prosphaten Welt stoßen die Menschen andauernd auf Evidenzen, experimentelle und nicht-experimentelle, die sie zum Handeln motivieren. Wer in dieser Welt experimentiert, macht Entdeckungen, die ihn und andere zum Weiterexperimentieren motivieren können; genauso können die Entdeckungen auch zu Replikationen animieren. Mit anderen Worten: am Grunde stehen Experimente auf ihren eigenen Füßen, ohne den Schutz einer formalen Fassade von Berechnungen, die für die Evidenz irrelevant ist (Hogben 1957, S.21).

Die Menschheit ist am Anfang nicht stehen geblieben und am Ende noch nicht angekommen. Sie hat Lebensformen kultiviert, verändert oder ausgemerzt. Die Frage des Pilatus, quid veritas (Joh 18, 38), schrumpft angesichts der großen Transformationen der Weltgeschichte (Polanyi 1957, S.44) zu einer historischen Randnotiz. Sofern überhaupt argumentiert wurde, galten die Argumente nicht der Wahrheit, sondern der Überzeugung anderer von der eigenen Position (Mercier & Sperber 2011). Noch heute heute fällt es Menschen schwer, Argumente zu ersinnen, die ihre Position widerlegen würden (Poletiek 1996). Den Niederungen unhintergehbaren Verhaltens glaubt die moderne Wissenschaft entstiegen zu sein und hat doch nur die Metaphysik durch die Physik ersetzt. Die zugehörige epistemologische Theorienkinematik logischer Geschlossenheit ist kategorial verschieden (Austin 1946) von der Dynamik des Wissenschaftsbetriebs. Auf der Pferderennbahn galoppieren keine Kugeln im Vakuum; die Hengste und Stuten müssen sich ihren Weg bahnen über Grasnarben und Rossäpfel. Dort kommt man manchmal narrativ schneller vorwärts als argumentativ, eine zusammenhängende

Geschichte wirkt oft überzeugender als eine konsistente Argumentation (Bateson 1972, S.247; Leamer 1974). Dass das anders sein soll, führt auf das vierte Nein.

In der Psychologie muss nichts verbessert werden, es kann aber etwas verändert werden. Eine Veränderung lässt sich ethisch rechtfertigen, aber nicht epistemologisch. Veränderungen wissenschaftlicher Gepflogenheiten implizieren eine Modifikation der Normen und Bewertungsmaßstäbe, hinter denen ein Veränderungswille stehen muss (Schopenhauer 1977, I S.144). D.h., Replikationen müssen gewollt sein; sie haben nur einen Wert, wenn man ihnen einen Wert gibt. Und einen Wert gibt man Replikationen, indem man repliziert oder praktisch Replikationen fördert. Das Beschwören einer Krise, in der der Goldstandard (Liu et al. 2008) der Experimentalwissenschaft eine deflationäre Abwertung erfährt, wäre an Dramatisierung eines Grundwertes kaum zu überbieten, hätte man sich an die Dauerkrisen nicht schon längst gewöhnt. Der Erfolg des Manövers lässt sich aus der Medienresonanz ebenso wenig vorhersagen, wie der Erfolg einer Replikation aus statistischen Kennwerten.

6.1 Monetarisierte Replikationswahrscheinlichkeit

Erfolgreiche Vorhersagen sind so beeindruckend wie schwierig. „Det er svært at spå, især om fremtiden“, soll Bohr gesagt haben während eines Vortrages zur Quantenmechanik, die bekanntlich vor der Psychologie die Wahrscheinlichkeitswende vollzogen hat. Der Schwierigkeit stellt sich Meehl (2002) mit einer Theorienaktuarik, in der pragmatische Indikatoren versammelt werden für erfolgreiche Theorien, d.h. Theorien, deren Lebenszeit länger ist als 50 Jahre (Meehl 2004), um Prognosen abzugeben zur Lebensdauer von Theorien. Die Aktuarik kommt aus ohne freilaufende Verfahren zur Beibringung von Evidenzen, die das individuelle wie kollektive Forscherverhalten determinieren. Sie zollt einer Wissenschaft Tribut, die sich entwickelt wie das Leben – völlig unbegründet und unbegründbar (Mahoney 1980). Ohne einen archimedischen Grund außerhalb des Verhaltens, an dem ein Hebel für nachhaltige Veränderungen ansetzen könnte, geben Werte dem Verhalten eine Orientierung; auch der Philosophie, die sich mit Wert und Wahrheit auseinandersetzt (Steiner 1987, S.271). Zu einem bestimmten, als wertvoll erachteten Zweck sind auch philosophische Theorien für den Moment zusammengebastelt (Lévi-Strauss 1962, S.27) aus empirischen Schnipseln und nicht für

die Ewigkeit gebaut aus zeitlosem Stahlbeton. Zumindest im Ansatz ist die Theorienaktuarik eine solche Épistémologie bricolée.

Die Planken, die die Wissenschaft tragen sollen, werden in der Aktuarik ausgetauscht auf den offenen Weltmeeren mit Treibholz, das die Wellen anschwemmen (Neurath 1932). Die Tragfähigkeit bastelt Meehl (2002) aus Qualitätsindikatoren einer Theorie, wie die Sparsamkeit ihrer Grundsätze, der Innovationsgrad, ihre Präzision, ableitbare und nicht ableitbare Befundlagen sowie ihre Kompatibilität mit bestehenden Theorien oder ihre Reduzierbarkeit auf diese. In der statistischen Aufarbeitung der Indikatoren wird die Stichprobenverteilung ersetzt durch Algorithmen, die die Befunde einer Befundlage gewichten nach ihrer Häufigkeit und ihrer Bedeutsamkeit (Dawes, Faust & Meehl 1989). Die so gewonnenen Prognosen sind klinischen wie epistemologischen Prognosen überlegen, was die Wahrscheinlichkeit ihres Zutreffens anbelangt (Grove & Meehl 1996). Damit noch nicht zufrieden, wechselt Meehl (2002) vom Beschreibungsmodus wertender Indikatoren in den Begründungsmodus transzendierender Objektivitätsmaße, indem er (2004) die Zusammenstellung der Indikatoren unter Meta-Metagrundsätze stellt und mittels Faktoranalyse die Wahrheitsähnlichkeit der Faktoren feststellen möchte. Die Wahrheitsähnlichkeit rekurriert auf die Wahrerscheinlichkeit einer Konvergenz theoretischer, metatheoretischer und meta-metatheoretischer Evidenzen. Denn die Wahrerscheinlichkeit in der Konklusion wird vorausgesetzt in den statistischen Operationen der Metatheorie, die Bestandteil sind der Prämissen.

Damit teilt die Theorienaktuarik mit dem positiv-prädiktiven Wert (Ioannidis 2005) den epistemologischen Selbstwiderspruch, der sich ergibt, wenn man mit Methoden Ergebnisse korrigieren möchte, die dieselben Methoden hervorgebracht haben – noch dazu, wenn diese Methoden, wo nicht als fehlerhaft, dann doch als fehlbar betrachtet werden. Der positiv-prädiktive Wert, der angibt, wie wahrscheinlich es ist, dass ein positiv berichteter Effekt existiert, benötigt zu seiner Verifikation, also den Nachweis dass der positiv-prädiktive Wert die wahre Wahrscheinlichkeit angibt, eine Evidenz seiner Richtigkeit, die nicht der Wert selbst – oder ein Derivat des Wertes – ist. Gibt es eine solche Evidenz, ist der positiv-prädiktive Wert überflüssig, gibt es eine solche Evidenz nicht, ist er zweifelhaft. Berechnet wird der positiv-prädiktive Wert aus der Teststärke, dem Alphaniveau und der Ausgangswahrscheinlichkeit bzw. Basisrate:

$$PPV = \frac{(1-\beta)\cdot P(b)}{(1-\beta)\cdot P(b)+\alpha} \quad .$$

Bei einem Alphaniveau von 5 Prozent und einer Teststärke von 20 Prozent existiert unter Verwendung des Satzes vom unzureichenden Grunde der Effekt mit einer Wahrscheinlichkeit von zwei Dritteln; geht die Basisrate runter von 50 auf 10 Prozent, reduziert sich auch die Wahrscheinlichkeit auf ein Viertel. Bei größeren Teststärken erhöhen sich nach dieser Formel die Wahrscheinlichkeiten, dass ein Effekt existiert, entsprechend (Button, Ioannidis, Nosek et al 2013). Die Formel besticht durch ihre einfache Form. Dass sie klärungsbedürftige Resultate zeitigt, wie den Rückgang der Effektwahrscheinlichkeit unter die Basisrate bei geringen Teststärken und hoher Ausgangswahrscheinlichkeit, täuscht über das eigentliche Problem hinweg: die Basisrate. Diese der Population zugeordnete Eigenschaft ist nur in den seltensten Fällen verfügbar und wird daher aus einer Stichprobe geschätzt. So zieht man ohne Not den gesamten epistemologischen Begründungsapparat ins Boot. Einfacher wäre es, große Stichprobenumfänge zu fordern, standardisierte Designs, die Angabe von Effektgrößen oder eine größere Anzahl an Replikationen (Moonesinghe, Khoury & Janssens 2007), zumal am Grunde dann doch wieder die totale empirische Evidenz entscheidet (Ioannidis 2005).

Die Frage, ob eine eine Wahrscheinlichkeit wahr ist, spielt die gegenläufigen Grammatiken ähnlich klingender Begriffe gegeneinander aus. Ohne Normverstöße ist keine positive Antwort möglich – dass beide Begriffe keinen gemeinsamen Gebrauch haben, deutet allerdings darauf hin, dass eine Norm, die beide verknüpft, nicht gebraucht wird. Jeder klärende Normverstoß läuft hinaus auf eine Regeländerung. Die Regeländerung erbringt freilich keinen epistemologischen Zugewinn in Form von gesteigerter Konsistenz oder beseitigter Zweifel, wie die Novellierung des Strafrechts neue Tatbestände schafft, aber keine neuen Sachverhalte; erst in der Urteilsbegründung sind Tatbestand (und Rechtsfolge) verhaltenswirksam.

Wie in der Aufklärung der Gerichtshof der autonomen Vernunft (Kant 1974, B779) als letzte Instanz der Wahrheitsfindung galt, so übernehmen nach der neoliberalen Transformation sich selbst organisierende Märkte die Rolle der Vernunft, wohlwissend, dass Märkte pervertiert sind in Klassenkatalysatoren (Marx 1957, S.388), die die Kluft zwischen Arm und Reich immer weiter aufspreizen (Piketty 2013, S.5). In Märkten gelten offenbarte Überzeugungen als wahre Überzeugungen, d.h. von einem Konsumen-

ten, der sich für ein Produkt entschieden hat, wird angenommen, dass er sich wieder für dieses Produkt entscheiden werden wird, um konsistent zu handeln (Samuelson 1931).

Trotz ungleicher Allokation von Vermögen und Wissen (Hanson, Oprea & Porter 2006), wird die totale empirische Evidenz in Märkten aggregiert zur Nachfrage nach Wissen (Almenberg, Kittlitz & Pfeiffer 2009) und der Gleichgewichtspreis per Dekret erklärt zum Maß der Vernunft. Auf der Angebotsseite handeln Wissenschaftler mit ihren Hypothesen, deren vorhergesagte Ereignisse sie in der Zukunft einlösen müssen. Zur Bestimmung der Wahrscheinlichkeit, dass eine Hypothese wahr ist, wird der Gleichgewichtspreis, den eine Hypothese am Wissenschaftsmarkt erzielt, ins Verhältnis gesetzt zum Loswert, der den Gesamtumsatz widerspiegelt aller miteinander konkurrierender Hypothesen, auf deren Wahrheit andere Marktteilnehmer wetten (Hanson 1985). Lautet die Hypothese, dass ein Experiment replizierbar ist, dann ist – im Marktgleichgewicht bei einem Kaufpreis von zwei Euro und einer Gewinnsauszahlung von zehn Euro – die Replikationswahrscheinlichkeit 20 Prozent.

Der Replikationsindex erfährt als Aktienindex auf Vorhersagemärkten eine Demokratisierung dadurch, dass im Grunde jeder mit seinem Geld auf den Erfolg einer Hypothese wetten kann, wie auf den Ölpreis oder Sportergebnisse. In der Casino-Epistemologie aggregiert das Teilwissen der Marktteilnehmer zum Freihandelsmosaik der Wahrheit (Hayek 1945). Ausgehend von der richtigen Überlegung, dass jeder Marktteilnehmer für sich entscheiden muss, wie evident ihm ein experimenteller Befund erscheint, wird im Aggregat eine Entscheidungshilfe angeboten, für eine Entscheidung, die bereits getroffen und aggregiert wurde. Betrachtet man die Vorhersagemärkte in einer reversiblen Zeitreihe, wird im Aggregat der Wetten das Induktionsproblem sichtbar, mit dem Unterschied, dass von partikularen Befunden vorwärtsgeschlossen wird auf generelle Eigenschaften der Population. Solange aber die grammatische Polarität von Stichprobe/Population bzw. Partikularem/Generellem in Gebrauch bleibt, kann widerspruchsfrei weder ein Rückschluss noch ein Vorwärtsschluss zwischen beiden Polen vermitteln.

Schlüsse hin oder her, Vorhersagemärkte benötigen ein Pendant zur Börse als einer Clearingstelle, die – nach vereinbarten Zeitraum – darüber richtet, ob eine Hypothese sich bewahrheitet hat bzw., ob eine Replikation erfolgreich war. Wenn das verbindlich gelingt ohne Aggregierung von Marktteilnehmern, dann ist auch zur Vorhersage keine solche Aggregierung notwendig. Im Reproduktionsprojekt: Psychologie konnten die

Teilnehmer darauf wetten, ob ein Befund repliziert werden würde. Kriterium für eine erfolgreiche Replikation war ein $p<0.05$, sodass schließlich 29 von 41 Replikationen richtig vorhergesagt wurden. Den Marktpreis als Wahrscheinlichkeit interpretiert, konnten im Zuge des Projektes auch Studien, deren Replikation wahrscheinlicher war als 50 Prozent, nicht repliziert werden (Dreber, Nosek, Pfeiffer et al. 2015). Drei Punkte sollten einen daher skeptisch stimmen: Erstens ist das Knüpfen des Replikationserfolges an den *p*-Wert problematisch aufgrund von dessen unkontrollierbaren Schwankungen, und folglich wäre der Marktpreis anders ausgefallen, wenn den Teilnehmern diese Problematik vor der Wette mitgeteilt worden wäre. Zweitens wird der Marktpreis verschieden ausfallen, je nachdem, welche Erfolgskriterien für eine Replikation gewählt werden. Drittens schließlich addieren sich die Wetten nicht notwendig linear auf zu einer fallenden Geraden, sondern können jede beliebige Form annehmen (Sonnenschein 1972; Debreu 1974; Mantel 1976), sodass insgesamt die Vorhersagen instabil sind und wenig aufschlussreich.

Nehmen wir dennoch an, Hypothesen würden eines Tages gehandelt werden wie Wertpapiere, dann würde Effizienz zur Grundnorm wissenschaftlichen Handelns, wodurch wissenschaftsendemische Maßstäbe verschwänden und genau die Instrumente blockiert würden, die zusammen mit den Replikationen an Wert gewinnen sollen. So würde beispielsweise die Teststärke leiden, weil es für Wissenschaftler jetzt schon effizienter ist, Studien mit geringer Teststärke zu realisieren (Bakker, Dijk & Wicherts 2012). Und Wissenschaftler betreiben heute schon Werbung für sich und ihre Institution, bei der sie gar zu leicht der Versuchung erliegen, Forschungserfolge anzukündigen, die bei Investoren und Steuerzahlern überzogene Erwartungen auslösen. Die lassen zwar den Börsenkurs in die Höhe schießen, platzen aber früher oder später. Es ist nicht einsichtig, weshalb Vorhersagemärkte besser funktionieren sollten als Finanzmärkte. Wie deren Index alles abbildet, nur nicht die realökonomischen Verhältnisse (Keynes 1936, S.161f), so ist nicht zu erwarten, dass ausgerechnet der Börsenindex für Hypothesen deren Wahrheitsgehalt getreulich wiedergibt.

Werden Hypothesen wie Wertpapiere gehandelt, werden Wissenschaftler wie ihre Nachbarn aus Industrie und Wirtschaft ihre Wetten absichern, indem sie auf populäre Forschungsfelder setzen, die – wie die Neurowissenschaften – in Forschungsgeldern geradezu ertränkt werden. (Pedersen & Hendricksen 2014). Wenn es zu viele Förder-

gelder gibt für zu wenige Forschungsfelder, entsteht der ideale Nährboden für pluralistische Ignoranz, derzufolge Menschen in unübersichtlichen Situationen ihr Verhalten nicht an den eigenen Überzeugungen ausrichten, sondern daran, wovon sie überzeugt sind, dass andere davon überzeugt sind (Katz & Allport 1931, S.343). Kurz: eine Ökonomisierung der Wissenschaft würde aus Universitäten und Akademien einen Hort der Ignoranz und Unvernunft machen. Das dürfte schwerlich mit dem Selbstverständnis der Wissenschaft in Einklang zu bringen sein.

Wahrheit darf nicht käuflich sein, Wahrscheinlichkeit keinen Preis besitzen (Resnik 2007, S.35). Dennoch ist es vom Wahren zu Waren nicht weit. Die Experimentalwissenschaften haben längst die Enge des Versuchslabors gesprengt und sich gemausert zu administrierter Großforschung (Solla Price 1963, S.6) mit internationalen Forschungsprogrammen (Lakatos 1978), die von Wissenschaftsorganisationen aufgelegt werden, um ein übergreifendes Versprechen einzulösen: Beherrschen durch Berechnen. Es zählen Taten und Tatsachen.

6.2 Die Kunst der Produktion und Reproduktion

Tatsachen sind aus einer Tat hervorgegangen (Latour 2000). Dem Motto 'nullius in verba' der Royal Society getreu, geht es den Experimentalwissenschaften um Interaktion, nicht um Reflexion. Wirklichkeit wird geschaffen und öffentlich demonstriert im Modus des Vorführens, nicht im Modus des Beweisens. Factum et verum convertuntur – Gewissheit gibt es nur von Getanem (Vico 1957, S.72). Eifriges Tun geht den Tatsachen voraus, die zum Maßstab des Tuns werden, wie jedes künstlerische Bricolieren sich an seinen eigenen Maßstäben misst. Die Zuwendung zu den Artes mechanicae ging einher mit einer Abwendung von den Artes liberales (Kagan 2000, S.21). Die hohe Kunst der Wissenschaft bestand fortan darin, Dinge zum Laufen zu bringen (Knorr-Cetina 1977).

Das Neue an Experimenten in der Wissenschaft lag darin, dass die Experimente Phänomene produzieren, Phänomene, von denen die Experimentatoren bis heute überzeugt sind, dass sie nicht produziert werden, sondern unabhängig vom Experiment existieren. Doch ihren historischen Erfolg verdanken Experimente im wesentlichen ihrer produktiven Wiederholbarkeit (Heisenberg 1973). Grundsätzlich kann sich jeder vom Funktionieren einer Apparatur durch Nachmachen selbst überzeugen. Produktion und

Reproduktion sind das Kerngeschäft eines Experimentalforschers, mit offensichtlichen Schnittstellen zur Fließfertigung von Automobilen, Medikamenten oder Fachartikeln. Unterm Strich ist es seine Produktivität, die einen Forscher auszeichnet (Solla Price 1963, S.41). Zur Reproduktion werden Experimente herangezogen, die in ein Forschungsprogramm passen. Im Rahmen eines Forschungsprogramms versuchen die Wissenschaftler, ihre Phänomene instrumentell miteinander zu verknüpfen, um so ihren Herrschaftsbereich auszudehnen. Die Reproduktion fungiert als Produktion, nicht als Verifikation (Knorr-Cetina 1977). Konzeptuelle Replikationen produzieren aus Produkten neue Produkte. Die Funktion ökonomischer Wertschöpfung scheint die primäre Funktion von Replikationen zu sein; die epistemologische Bestätigungsfunktion dagegen eher ein Deckmantel.

Die Experimentalwissenschaften haben sich als Technik bewährt, und diese Bewährung ist auch Maß für die Gültigkeit von Wahrscheinlichkeitstheorie und Statistik (Gnedenko 1957, S.11). Gerade die Statistik machte ihre größten Fortschritte im Kontext des Brauwesens und der Düngemittelindustrie (Healy 1978). Industrielle Qualitätskontrolle und die Ertragssteigerung von Nutzpflanzen sind Musterexemplare praxis-geladener Tatsachen, für die Replikationen zur Problemformulierung mehr beitragen als zur Problemlösung. Weil Flasche für Flasche abgefüllt wird und der Kreislauf von Aussaat und Ernte sich stet wiederholt, formiert sich das ökonomische Ziel, mit möglichst wenig Aufwand einen bestimmten Ertrag zu erzielen, wie auch das technische Ziel der Homogenisierung von Gerstensaft oder der Schädlingsresistenz von Gerstepflanzen. Auch die Psychologie stellt sich der Forderung der Kosteneffizienz beim Unterfangen, mithilfe von Replikationen Herrschaft über menschliches Verhalten zu erlangen, also 'humane' Tatsachen zu schaffen (Tweney 2004).

Effekte sind das Produkt psychologischer Experimente und Replikationen die Vorstufe zur Serienproduktion. Diese Wissensprodukte kann man in Kühlhäusern lagern oder zwischen Buchdeckeln verwahren. Der kumulative Charakter wissenschaftlichen Fortschritts ist aus ökonomischer Perspektive geradezu greifbar. Forschungsprogramme befeuern die Konjunktur derjenigen, die Dinge zum Laufen bringen und am Laufen halten. Die Programme erlauben auch die Replikation teurer oder zeitaufwändiger Experimente. Dass Forschungsteams Monate bis Jahre brauchen, um ihre eigenen Experimente zu replizieren, die an im Labor gezüchteten Zellen oder Lebewesen vorgenom-

men wurden (Wiesel 2013), ist kein Ausschlusskriterium. Wenn sich die Reproduktion eines Produktes in irgendeiner Weise lohnt, wird sich jemand finden, der sie angeht.

Aus epistemologischen Gründen lohnen Replikationen sich nicht. Die Wissenschaften verdanken ihren Fortschritt nicht wasserdichten Induktionsschlüssen (Karmiloff-Smith 1988), sondern ihren Produkten und dem Versprechen ihrer Beherrschbarkeit. Wissenschaft schreitet zu Zielen hin fort und nicht von einem epistemologisch zertifizierten Grund weg (Franklin 1993). Fortschritt ist relativ zu den Zielen einer Wissenschaft kumulativ (Laudan 1977, S.7). Daher ist es wichtig, sich auf klar beschriebene Ziele zu einigen (Rappaport 1977, S.18). Wissenschaftlicher Spürsinn (Fiedler 2011) und die Verführungskraft replikationslogischer Erklärungen norden den Fortschritt ein. Vorausblickend, voraussagend schreitet die moderne Wissenschaft voran, den Zielen zu; nicht rückblickend, rückschließend auf letzte Gründe, um die Distanz dorthin zu messen. Warum? Weil die Ziele der Wissenschaft wertvoller scheinen als ihr Ursprung.

Wertvorstellungen sind höchstens als sehr allgemeine Werte im Sinne eines Weltethos (Küng 1990) universal. Will man konkret Werte der Wissenschaft umwerten, kann man das jederzeit tun – durch hartnäckige Überzeugungsarbeit; zum Einlenken zwingen kann man die Anderen nur physisch (Fichte 1801). Ob man für eine erfolgreiche Umwertung den schwergängigen Motor einer Methodenkrise anwerfen muss oder besser leicht und lustvoll archaische Instinkte anspricht (Nietzsche 1988, S.383), steht dahin. Beide Wege, der asketische und der hedonistische, stehen der Wissenschaft offen. Der reflektierte Replikator (Ioannidis 2009), der aus der übervollen Welt behutsam Überzeugung um Überzeugung abschichtet, schöpft anfänglich nur aus dem Vollen, weil eine Theorie nicht aus nichts bestehen kann, und geht von dort den Weg zum einsamen Cogito (Descartes 1992, II, 3), der auf das Oxymoron eines regressiven Progresses führt (Brunswik 1955). Der aggressive Entdecker (Ioannidis 2009), der wildert, wo er kann, jagt, was ihn interessiert und medial mit seiner Beute protzt, trachtet danach, die Fülle noch zu mehren, und geht den Weg zum massenhaften αὐξάνσομεν.

Gigerenzer (2009) sieht zwei Vorbereitungen, die getroffen sein müssen, bevor die Psychologie sich auf den Weg machen kann. Erstens sollten die isolierten Hypothesen, die Theorien noch nicht einmal parodieren, zumindest zu komplexeren Modellen ausgebaut werden; und zweitens sollten die Modelle zu Theorien integriert werden, die eine kollektive Anstrengung repräsentieren. In der Psychologie vertreten zu viele Psycho-

logen etwas, das sie für ihre eigenen Theorien halten, und an dem sie beharrlich festhalten, egal wie oft Replikationen ihrer Experimente scheitern (Gelman & Geurts 2017). Theorien sind nicht nur Voraussetzung der Replizierbarkeit eines Experiments (Cesario 2014; Vartanian 2014), insofern sie die Folie bieten, vor der ein Befund interpretiert und der Erfolg einer Replikation bewertet werden kann. Sie sind auch – aus demselben Grund – ausschlaggebend für die Beurteilung der Replikationswahrscheinlichkeit eines Experiments (Miller 2009). Theorien haben also größeres epistemisches Gewicht als der reale Befund und die fiktive Stichprobenverteilung zusammen. Ohne ein Netz wissenschaftlicher Theorien, aus dem Behauptungen ihre Plausibilität beziehen, kann es weder Selbstkorrektur noch Fortschritt der Wissenschaft geben (Campbell 1985). Ein Netz aus Theorien verstärkt Hypothesen dadurch, dass man anhand verschiedener und unabhängiger Methoden immer wieder zum selben Ergebnis, wie bspw. zur Avogadro-Zahl, kommt, wodurch die Hypothesen kreuzvalidiert werden (Cartwright 1991).

In jeder Replikation steckt ein Mindestmaß an Zweifel am Originalexperiment. Wer das Originalexperiment in Zweifel zieht, muss dessen Generalisierung in Zweifel ziehen. Wer die Generalisierung in Zweifel zieht, muss die bestätigende Evidenz in Zweifel ziehen. Wer die bestätigende Evidenz in Zweifel zieht, muss die Vorhersagen, die die Generalisierung beinhaltet, in Zweifel ziehen. Und wer in Zweifel zieht, dass die Vorhersagen zutreffen, muss selbst Wissenschaft betreiben (Fodor 1997). Als Wissenschaftler oder als Bürgerwissenschaftler (Coyne 2016). Wissenschaftlicher Fortschritt ist zu wichtig, als dass man ihn beamteten Wissenschaftlern überlassen könnte. Fortschritt heißt hier zur Tat schreiten.

6.3 Herausforderungen im Computerzeitalter

Zur Tat schreiten zusehends Rechner mit auf ihnen installierten Programmen. Ein laufendes Programm ist für Simon (1992) ein Moment der Wahrheit und des Fortschritts. Für King (2003) ist die Automatisierung des Auffindens, Teilens und Archivierens, des Konvertierens, Analysierens und Verbreitens von Daten längst in online-Reichweite, was Replikationen erheblich vereinfachen würde. Die Bedeutung des Einfachen haben Computer grundlegend verändert (Efron 1978), weshalb vereinzelt spekuliert

wird auf eine informationstechnologische Lösung der Replikationskrise (Spellman 2015).

In der Tat ist Big Data im Computerzeitalter (Efron & Hastie 2016, S.267) ein Steuerparadies für die von Statistikern entwickelten Werkzeuge. Sie splitten die Wetteinsätze aus Myriaden von Quellen auf der Suche nach charakteristischen Mustern (Markowetz et al. 2014) und verteilen die Einsätze in praktikablen Portionen auf die einzelnen Rechner im Netzwerk zur Ausführung von Monte Carlo-Simulationen anhand der Daten. Dadurch vergrößert sich der Datenumfang von Simulation zu Simulation. Die resultierenden – nicht-parametrischen – Simulationen werden dann verbunden zu einem einzigen Datensatz, aus dem eine Stichprobe gezogen wird, die den Konsens aller Rechner repräsentiert (Scott et al. 2016).

Die Daten der gesamten Population strömen in Echtzeit aus Smartphones und Mobiltelefonen aus online-Auktionen und sozialen Netzwerken, aus Spielplattformen, Suchmaschinen oder Blogs, was einer Vollerhebung gleichkommt. Stichproben – und Stichprobenverteilungen – erübrigen sich. Ohne Theorien im herkömmlichen Sinn sagen rechnergestützte Sozialwissenschaften das Verhalten sozialer Gruppen, Aggregate oder Kategorien erstaunlich präzise voraus (Chang, Kauffman & Kwon 2013), sodass Ethologie und Psychologie sich zu einer Wissenschaft vereinen (Gomez-Marin et al. 2014).

Jetzt, wo genügend unabhängige Daten vorhanden sind, die die kanonische Inferenzstatistik erübrigen, weil Algorithmen zur Vorhersage von individuellem Verhalten direkt aus der totalen informatischen Evidenz Muster extrahieren, mit neuen Daten abgleichen und entsprechend modifizieren und verfeinern zu einem Panoptikum gläserner Menschen, wie es sich bisher nur dem Laplaceschen Dämon dargeboten hat, jetzt, wo wir nie dagewesene Gewissheit und Herrschaft erlangen könnten, jetzt kommt Unsicherheit auf, ob es nicht besser wäre, die Wissenschaft würde ihr Versprechen nicht einlösen. Es sieht fast so aus, als fräße auch die wissenschaftliche Revolution ihre Kinder (Büchner 1987, S.22), allen Anfangserfolgen zum Trotz.

Die datenintensive Exploration fordert nicht nur die Methodik der Wissenschaften heraus (Kitchin 2014) dahingehend, dass der Grundsatz der Induktion außen vor bleibt, weil es für eine Echtzeitextrapolation unerheblich ist, ob die Welt sich gleich bleibt: die Muster im Datenmosaik sind immer relativ zum Bezugssystem, und ein sich änderndes Bezugssystem bringt nach dem Relativitätsprinzip (Einstein 1905) andere Muster hervor

als ein Inertialsystem. Spielt der Induktionsgrundsatz keine Rolle, dann sind Replikationen überflüssig. Die Frage nach dem Wert der Reproduzierbarkeit ist daher zu richten an künftige Forschung, nicht an vergangene Experimente. Psychologen sollten zumindest Stellung beziehen, ob sie produzieren und reproduzieren wollen oder die Effizienz von Algorithmen und Datenmatrizen überwachen, ob sie, auf nomothetischer Grundlage beherrschen oder auf ideographischer Grundlage verstehen wollen oder doch lieber etwas ganz anderes. Replikationen sind kein Muss. Allerdings muss man wissen, was man will, will man wissen, was man tut.

7 Literatur

Abelson, R.P. (1995). *Statistics as principled Argument.* Hillsdale: Lawrence Erlbaum.

Adolphs, R., Tranel, D., Bechara, H. & Damasio, A. (1996). Neuropsychological Approaches to Reasoning and Decision-Making. In: Damasio, A., Damasio, H. & Christen, Y. (Hg). *Neurobiology of Decision-Making.* (S.157-180). Heidelberg: Springer.

Aglioti, S. & Berlucchi, G. (2013). *Neurofobia.* New York: Cortina.

Agresti, A. (2015). *Foundations of Linear and Generalized Linear Models.* New Jersey: Wiley.

Akaike, H. (1981). Likelihood of a Model and Information Criteria. *Journal of Econometrics, 16,* 3-14.

Akaike, H. (1998). *Selected Papers.* New York: Springer.

Alexander, J. & Weinberg, J.M. (2007). Analytic Epistemology and Experimental Philosophy. *Philosophy Compass, 2,* 56-80.

Allen, M. & Preiss, R. (1993). Replication and Meta-Analysis: A necessary Connection. *Journal of Social Behavior and Personality, 8(6),* 9-20.

Allison, D.B. (2016). A Tragedy of Errors. *Nature, 530,* 27-29.

Allport, G.W. (1968). The historical Background of Social Psychology. In: Lindzey, G. (Hg). *The Handbook of Social Psychology. (*Bd 1, S.1-46). New York: Random House.

Almenberg, J., Kittlitz, K. & Pfeiffer, T. (2009). An Experiment on Prediction Markets in Science. *Public Library of Science One, 4(2),* e5800.

Altman, M. & McDonald, M.P. (2003). Replication with numerical Accuracy. *Political Analysis, 11(3),* 302-307.

Amir, Y. & Sharon, I. (1990). Replication Research: A 'Must' for the scientific Advancement of Psychology. *Journal of social Behavior and Personality, 5(4),* 51-69.

Anderson, J.R. (2007). *Kognitive Psychologie.* 6.Aufl. Heidelberg: Springer.

Angrist, J.D. & Pischke, J.S. (2010). The Credibility Revolution in Empirical Economics: How better Research Design is taking the Con out of Econometrics. *Journal of Economic Perspectives, 24(2),* 3-30.

Aristoteles. (1966). *De Generatione Animalium.* Brüssel: Brouwer.

Asendorpf, J.B., Conner, M., Fruyt, P. d., Houwer, J. d., Denissen, J.J., Fiedler, K. et al. (2013) Recommendations for Increasing Replicability in Psychology. *European Journal of Personality, 27,* 108-119.

Ashcroft, R.E. (2004). Current epistemological Problems in Evidence Based Medicine. *Journal of Medical Ethics, 30,* 131-135.

Austin, J.L. (1946). Other Minds. *Proceedings of the Aristotelian Society, 20,* 148-187.

Baayen, R.H., Davidson, D.J. & Bates, D.M. (2008). Mixed-Effects Modeling with crossed Random Effects for Subjects and Items. *Journal of Memory and Language, 59,* 390-412.

Baguley, T. (2012). *Serious Stats: A Guide to Advanced Statistics for the Behavioral Sciences.* New York: Palgrave Macmillan.

Bakan, D. (1966). The Test of Significance in psychological Research. *Psychological Bulletin, 66(6),* 423-437.

Baker, M. (2016). Is there a Reproducibility Crisis? *Nature, 533,* 452-454.

Bakker, M. & Wicherts, J.M. (2011). The (Mis)reporting of statistical Results in Psychology Journals. *Behavioral Research, 43,* 666-678.

Bakker, M., Dijk, A.v. & Wicherts, J.M. (2012). The Rules of the Game called Psychological Science. *Perspectives on Psychological Science, 7(6),* 543-554.

Barber, B. (1961). Resistance by Scientists to scientific Discovery, *Science, 134,* 596-602.

Barber, J.J. & Ogle, K. (2014). To *p* or not to *p*? *Ecology, 95(3),* 621-626.

Barnard, G.A. (1949). Statistical Inference. *Journal of the Royal Statistical Society, Series B, 11(2),* 115-149.

Barta, R. (2006). *Antropología del Cerebro: La Consciencia y los Sistemas simbólicos.* Valencia: Pre-Textos.

Bateson, G. (1972). *Steps to an Ecology of Mind: Collected Essays in Anthropology, Psychiatry, Evolution, and Epistemology.* Chicago: University Press.

Bavel, J.v., Mende-Siedlecki, P., Brady, W.J. & Reinero, D.A. (2016). Contextual Sensitivity in scientific Reproducibility. *Proceedings of the National Academy of Sciences, 113(23),* 6454-6459.

Bayes, T. (1763). An Essay towards Solving a Problem in the Doctrine of Chances. *Philosophical Transactions of the Royal Society, 53,* 376-408.

Begley, C.G. & Ioannidis, J.P. (2015). Reproducibility in Science: Improving the Standard for basic and preclinical Research. *Circulation Research, 116,* 116-126.

Begley, C.G., Buchan, A.M. & Dirnagl, U. (2015). Institutions must do their Part for Reproducibility. *Nature, 525,* 25-27.

Benjamini, Y. & Hochberg, Y. (1995). Controlling the False Discovery Rate: A practical and powerful Aproach to Multiple Testing. *Journal of the Royal Statistical Society, Series B, 57(1),* 289-300.

Bennett, R.M. & Hacker, P.M. (2010). *Die philosophischen Grundlagen der Neurowissenschaften.* Darmstadt: Wissenschaftliche Buchgesellschaft.

Bennett, C.M. & Miller, M.B. (2010). How reliable are the Results from functional Magnetic Resonance Imaging? *Annals of the New York Academy of Sciences, 1191,* 133-155.

Berger, J.O. & Sellke, T. (1987). Testing a Point Null Hypothesis: The Irreconcilability of p Values and Evidence. *Journal of the American Statistical Association, 82,* 112-122.

Berkson, J. (1942). Tests of Significance considered as Evidence. *Journal of the American Statistical Association, 37,* 325-335.

Bernard, C. (1947). *Principes de Médicine expérimentale.* Paris: Presses Universitaires de France.

Berthon, P., Pitt, L., Ewing, M. & Carr, C.L. (2002) Potential Research Space in Management Information Systems: A Framework for Envisioning and Evaluating Research Replication, Extension, and Generation. *Information Systems Research, 13(4),* 416-427.

Birnbaum, A. (1962). On the Foundations of Statistical Inference. *Journal of the American Statistical Association, 57(298),* 269-306.

Birnbaum, A. (1964). *The anomalous Concept of statistical Evidence: Axioms, Interpretations, and elementary Exposition.* New York: University Press.

Birnbaum, A. (1970). Statistical Methods in scientific Inference. *Nature, 225(5237),* 1033.

Birnbaum, A. (1972). More on Concepts of statistical Evidence. *Journal of the American Statistical Association, 67(340),* 858-861.

Bishop, D. (2016). There is a Reproducibility Crisis in Psychology and we need to act on it. *Bishopblog,* Blogeintrag vom 05.03.

Bissell, M. (2013). The Risks of the Replication Drive. *Nature, 503,* 333-334.

Bogen, J. (2001). 'Two as good as a Hundred': Poorly replicated Evidence in some Nineteenth-Century neuroscientific Research. *Studies in the History and Philosophy of Biological and Biomedical Sciences, 32(3),* 491-533.

Boghossian, P. (2006). *Fear of Knowledge: Against Relativism and Constructivism.* Oxford: Clarendon Press.

Boghossian, P. (2014). What is Inference? *Philosophical Studies, 169(1),* 1-18.

Bohannon, J. (2014). Replication Effort provokes Praise – and ‚Bullying' Charges. *Science, 344(6186),* 788-789.

Böhme, G., Daele, W.v.d. & Krohn, W. (1973). Die Finalisierung der Wissenschaft. *Zeitschrift für Soziologie, 2,* 128-144.

Bois-Reymond, E.d. (1887). *Über die Grenzen des Naturerkennens.* 6.Aufl. Leipzig: Veit & Comp.

Bondi, H. (1975). What is Progress in Science? In: Harré, R. (Hg). *Problems of scientific Revolution.* (S. 1-10). Oxford: Clarendon Press.

Bonnet, D.G. (2012). Replication-Extension Studies. *Current Directions in Psychological Science, 21(6),* 409-412.

Borel, É. (1933). *Traité du Calcul des Probabilités et de ses Applications.* Paris: Gauthiers-Villars.

Boring, E.G. (1919). Mathematical vs. scientific Significance. *The Psychological Bulletin, 16(10),* 335-338.

Brandt, M.J., IJzerman, H., Dijksterhuis, A., Farach, F.J., Geller, J., Giner-Sorolla, R., Grange, J.A., Perugini, M., Spies, J.R. & Veer, A.v. (2014). The Replication Recipe: What makes for a convincing Replication? *Journal of Experimental Social Psychology, 50,* 217-224.

Braver, S.L., Thoemmes, F.J. & Rosenthal, R. (2012). Continuously cumulating Meta-Analysis and Replicability. *Perspectives on Psychological Science, 7(6),* 333-342.

Bredenkamp, J. (1972). *Der Signifikanztest in der psychologischen Forschung.* Frankfurt/M.: Akademische Verlagsgesellschaft.

Breslow, N.E. & Clayton, D.G. (1993). Approximate Inference in Generalized Mixed Models. *Journal of the American Statistical Association, 88(421),* 9-26.

Bridgman, P.W. (1928). *The Logic of modern Physics.* New York: Macmillan.

Bridgman, P.W. (1959). *The Way Things are.* Harvard: University Press.

Broad, C.D. (1928). The Principles of problematic Induction. *Proceedings of the Aristotelian Society, 28,* 1-46.

Brown, A.N., Cameron, D.B. & Wood, B.D. (2014). Quality Evidence for Policymaking: I'll believe it when I see the Replication. *Journal of Development Effectiveness, 6(3),* 215-235.

Brunswik, E. (1938). The conceptual Framework of Psychology. *International Encyclopedia of Unified Science, 1(7),* 1-102.

Brunswik, E. (1955). Representative Design and probabilistic Theory in a functional Psychology. *Psychological Review, 62(3),* 193-217.

Büchner, G. (1987). *Werke und Briefe.* 8.Aufl. München: dtv.

Burnham, K.P. & Anderson, D.R. (2014). *p* Values are only an Index to Evidence. *Ecology, 95(3),* 627-630.

Button, K.S., Ioannidis, J.P., Mokrysz, C., Nosek, B.A. et al. (2013). Power Failure: Why small Sample Size undermines the Reliability of Neuroscience. *Nature Neuroscience, 14,* 365-376.
Camerer, C.F., Dreber, A., Forsell, E. et al. (2015). Evaluating Replicability of laboratory Experiments in Economics. *Science, 351(6280),* 1433-1436.
Camfield, L. & Palmer-Jones, R. (2013). Three 'Rs' of Econometrics: Repetition, Reproduction and Replication. *The Journal of Development Studies, 49(12),* 1607-1614.
Campbell, N. (1921). *What is Science?* London: Methuen.
Campbell, D.T. & Fiske, D.W. (1959). Convergent and discriminant Validation by the Multitrait-Multimethod Matrix. *Psychological Bulletin, 56(2),* 81-105.
Campbell, D.T. (1969). Reforms as Experiments. *American Psychologist, 24,* 409-429.
Campbell, D.T. (1973). The Social Scientist as methodological Servant of the experimenting Society. *Policy Studies Journal, 2(1),* 72-75.
Campbell, D.T. (1985). Toward an epistemologically-relevant Sociology of Science. *Science, Technologies & Human Values, 10(1),* 38-48.
Cantor, G. (1932). Historische Notizen über die Wahrscheinlichkeitsrechnung. In: *Gesammelte Abhandlungen mathematischen und philosophischen Inhalts.* Berlin: Springer.
Carnap, R. (1947). On the Application of Inductive Logic. *Philosophy and Phenomenological Research, 8,* 133-148.
Carnap, R. (1950). *Logical Foundations of Probability.* London: Routledge.
Carnap, R. (1966). The Aim of Inductive Logic. *Studies in Logic and the Foundations of Mathematics, 44,* 308-318.
Carnap, R. (1973). Notes on Probability and Induction. *Synthese, 25(3/4),* 269-298.
Carpenter, S. (2012). Psychology's bold Initiative. *Science, 335(6076),* 1558-1561.
Cartwright,N. (1991). Replicability, Reproducibility, and Robustness. *History of Political Economy, 23(1),* 143-155.
Cesario, J. (2014). Priming, Replication, and the hardest Science. *Perspectives on Psychological Science, 9(1),* 40-48.
Chamberlin, T.C. (1890). The Method of multiple Working Hypotheses. *Science, 15(366),* 92-96.
Chang, R.M., Kauffman, R.J. & Kwon, Y.O. (2014). Understanding the Paradigm Shift to Computational Social Science in the Presence of Big Data. *Decision Support Systems, 63,* 67-80.
Chow, S.L. (1998). Précis of statistical Significance: Rationale, Validity, and Utility. *Behavioral and Brain Sciences, 21(2),* 169-194.
Cialdini, R.B. (1984) *Einfluß: Wie und warum sich Menschen überzeugen lassen.* Landsberg: Moderne Verlags-Gesellschaft.
Claeskens, G. (2016). Statistical Model Choice. *Annual Review of Statistics and its Applications, 3,* 233-256.
Clemens, M.A. (2015). The Meaning of failed Replications: A Review and Proposal. *Institut für Zukunft der Arbeit Diskussionspapier Nr. 9000,* 1-24.
Cochran, W.G., Mosteller, F. & Tukey, J.W. (1954). Principles of Sampling. *Journal of the American Statistical Association, 49(265),* 13-35.
Cochrane, A.L. (1972). *Effectiveness and Efficiency: Random Reflections on Health Services.* Nuffield: Provincial Hospitals Trust.
Cohen, J. (1977). *Statistical Power Analysis for the Behavioral Sciences.* 2.Aufl. London: Academic Press.

Cohen, J. (1992). Statistical Power Analysis. *Current Directions in Psychological Science, 1(3),* 98-101.
Cohen, J. (1994). The Earth is round ($p < .05$). *American Psychologist, 49(12),* 997-1003.
Cohen, L.J. (1991). *The Probable and the Provable.* Aldershot: Gregg Revivals.
Collins, H. (1975). The seven Sexes: A Study in the Sociology of a Phenomenon, or the Replication of Experiments in Physics. *Sociology, 9,* 205-224.
Collins, H. (1981). Son of seven Sexes: The social Destruction of a physical Phenomenon. *Social Studies of Science, 11(1),* 33-62.
Collins, H. (1984). When do Scientists prefer to vary their Experiments? *Studies in History and Philosophy of Science, 15(2),* 169-174.
Collins, H. (1987). Misunderstanding Replication? *Sociology of Science, 26(2),* 451-459.
Collins, H. (1989). The Meaning of Experiment: Replication and Reasonableness. In: H. Lawson & L. Appignanesi (Hg). *Dismantling Truth.* (S.82-92). London: Weidenfeld & Nicholson.
Collins, H. (1991). The Meaning of Replication and the Science of Economics. *History of Political Economy, 23(1),* 123-142.
Collins, H. (1992). *Changing Order: Replication and Induction in scientific Practice.* Chicago: University Press.
Condorcet, J.A. (1988). *Esquisse d'un Tableau historique des Progrès de l'Esprit humain.* Paris: Flammarion.
Cook, T.D. & Campbell, D.T. (1979). *Quasi-Experimentation: Design and Analysis.* Boston: Houghton Mifflin.
Courgeau, D. (Hg). (2003). *Methodology and Epistemology of Multilevel Analysis.* Dordrecht: Kluwer.
Cox, D. (1958). Some Problems connected with Statistical Inference. *The Annals of Mathematical Statistics, 29(2),* 357-372.
Cox, D. & Mayo, D.G. (2010). Objectivity and Conditoinality in Frequentist Inference. In: Mayo, D.G. & Spanos, A. (Hg). *Error and Inference.* (S.276-304). Cambridge: University Press.
Coyne, J.C. (2016). Replication Initiatives will not salvage the Trustworthiness of Psychology. *Biomed Central Psychology, 4(28),* 1-11.
Cronbach, L.J. (1973). Beyond the two Disciplines of scientific Psychology. *American Psychologist, 2,* 116-127.
Cronbach, L.J. (1984). *Essentials of psychological Testing.* 4.Aufl. New York: Harper & Row.
Cumming, G., & Finch, S. (2001). A Primer on the Understanding, Use, and Calculation of Confidence Intervals that are based on central and noncentral Distributions. *Educational and Psychological Measurement, 61(4),* 532-574.
Cumming, G., Williams, F. & Fidler, F. (2004). Replication and Researchers' Understanding of Confidence Intervals and Standard Error Bars. *Understanding Statistics, 3(4),* 299-311.
Cumming, G. (2008). Replication and *p* Intervals. *Perspectives on Psychological Science, 3(4),* 286-300.
Cumming, G. (2012). *Understanding the new Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis.* London: Routledge.

Curran, P.J., Hussong, A.M., Cai, L., Huang, W. et al. (2008). Pooling Data from Multiple Longitudinal Studies: The Role of Item Response Theory in Integrative Data Analysis. *Developmental Psychology, 44(2),* 365-380.
Curran, P.J. (2009). The seemingly Quixotic Pursuit of a cumulative Psychological Science. *Psychological Methods, 14(2),* 77-80.
Curran, P.J. & Hussong, A.M. (2009). Integrative Data Analysis: The simultaneous Analysis of multiple Data Sets. *Psychological Methods, 14(2),* 81-100.
Damasio, A., Damasio, H. & Christen, Y. (Hg). (1996). *Neurobiology of Decision-Making.* Heidelberg: Springer.
Damasio, A. (1999). *The Feeling of what happens.* London: William Heinemann.
Danziger, K. (1985). The methodological Imperative in Psychology. *Philosophy of the Social Sciences, 15,* 1-13.
Danziger, K. (1988). On Theory and Method in Psychology. *Recent Trends in Theoretical Psychology, 3,* 87-94.
Darmant, G. & Matalon, B. (1986). Recherches sur les Pratiques de Vérification des Expériences scientifiques. *L'Année Sociologique, 36,* 209-238.
David, F.N. & Johnson, N.L. (1951). The Effect of Non-Normality on the Power Function of the *F*-Test in the Analysis of Variance. *Biometrika, 38(1/2),* 43-57.
Davidson, D. (1980). *Essays on Actions and Events.* Oxford: University Press.
Dawes, R.M., Faust, D. & Meehl, P.E. (1989). Clinical versus actuarial Judgment. *Science, 243(4899),* 1668-1674.
Deangelis, C.D. & Fontanarosa, P.B. (2010). The Importance of independent academic statistical Analysis. *Biostatistics, 11(3),* 383-384.
Debreu, G. (1974). Excess Demand Functions. *Journal of Mathematical Economics, 1,* 15-23.
Dekker, S., Lee, N.C. & Jolles, J. (2014). Over het Vóórkomen en Voorkómen van Neuromythen in het Onderwijs. *Neuropraxis, 18(2),* 62-66.
Dempster, A.P. (1966). New Methods for Reasoning towards posterior Distributions based on Sample Data. *The Annals of Mathematical Statistics, 37(2),* 355-374.
Dennis, A.R. & Valacich, J.S. (2015). A Replication Manifesto. *Association of Information Systems Transactions on Replication Research, 1,* 1.
Depaoli, S. & Schoot, R.v.d. (2015). Improving Transparency and Replication in Bayesian Statistics. *Psychological Methods, 1221,* 1-22.
Descartes, R. (1992). *Meditationes de Prima Philosophia.* Hamburg: Meiner.
Dickhaus, T. (2014). *Simultaneous Statistical Inference.* Berlin: Springer.
Diener, E. & Biswas-Diener, R. (2015). The Replication Crisis in Psychology. In: R. Biswas-Diener & E. Diener (Hg). *Psychology.* (7-14). Champaign: Noba Textbook Series.
Dietrich, C. (2010). Decision Making: Factors that influence Decision Making, Heuristics used, and Decision Outcomes. *Inquiries Journal, 2(2),* 1-10.
Dingler, H. (1907). *Grundlinien einer Kritik und exakten Theorie der Wissenschaften.* München: Theodor Ackermann.
Dingler, H. (1928). *Das Experiment: sein Wesen und seine Geschichte.* München: Ernst Reinhardt
Ditto, P.H. & Lopez, D.F. (1992). Motivated Skepticism: Use of differential Decision Criteria for preferred and nonpreferred Conclusions. *Journal of Personality and Social Psychology, 63(4),* 568-584.
Donnellan, B. (2013). Go big or go Home – A recent Replication Attempt. *The Trait-State Continuum,* Blogeintrag vom 11.12.

Donoho, D.L. (2010). An Invitation to reproducible Computational Research. *Biostatistics, 11(3),* 385-388.

Dreber, A., Pfeiffer, T., Almenberg, J., Isaksson, S., Nosek, B.A. et al. (2015). Using Prediction Markets to estimate the Reproducibility of scientific Research. *Proceedings of the National Academy of Science, 112(50),* 1543-1547.

Duke-Elder, S. (1964). The Saga of a Century. *Transactions of the American Ophthalmological Society, 63,* 192-203.

Duncan, G.J., Engel, M., Claessens, A. & Dowsett C.J. (2014). Replication and Robustness in Developmental Research. *Developmental Psychology, 50(11),* 2417-2425.

Durlak, J.A. (2009). How to select, calculate, and interpret Effect Sizes. *Journal of Pediatric Psychology, 34(9),* 917-928.

Dworsky, L.N. (2008). *Probably not: Future Prediction using Probability and Statistical Inference.* New Jersey: Wiley.

Earman, J. (1996). *Bayes or Bust?* 2.Aufl. Cambridge: The MIT Press.

Earp, B.D. & Trafimow, D. (2015). Replication, Falsification, and the Crisis of Confidence in Social Psychology. *Frontiers in Psychology, 6,* 1-11.

Easley, R.W., Madden, C.S. & Dunn, M.G. (2000). Conducting Marketing Science: The Role of Replication in the Research Process. *Journal of Business Research, 48,* 83-92.

Edwards, W., Lindman, H. & Savage, L.J. (1963). Bayesian Statistical Inference for psychological Research. *Psychological Review, 70(3),* 193-242.

Efron, B. (1978). Controversies in the Foundations of Statistics. *The American Mathematical Monthly, 85(4),* 231-246.

Efron, B. (1979). Computers and the Theory of Statistics: Thinking the Unthinkable. *Society of Industrial and Applied Mathematics Review, 21(4),* 460-480.

Efron, B. & Tibshirani, R. (1991). Statistical Data Analysis in the Computer Age. *Science, 253(5018),* 390-395.

Efron, B. & Hastie, T. (2016). *Computer Age Statistical Inference.* Cambridge: University Press.

Eid, M., Gollwitzer, M. & Schmitt, M. (2015). *Statistik und Forschungsmethoden.* 4.Aufl. Weinheim: Beltz.

Einstein, A. (1905). Zur Elektrodynamik bewegter Körper. *Annalen der Physik und Chemie, 17,* 891-921.

Eklund, A., Nichols, T.E. & Knutsson, H. (2016). Cluster-Failure: Why fMRI Inferences for spatial Extent have inflated false-positive Rates. *Proceedings of the National Academy of Science, 113(28),* 7900-7905.

Elstrodt, J. (2009). *Maß- und Integrationstheorie.* 6.Aufl. Heidelberg: Springer.

Eriksson, K. (2012). The Nonsense Math Effect. *Judgment and Decision Making, 7(6),* 746-749.

Estes, W.K. (1997). On the Communication of Information by Displays of Standard Errors and Confidence Intervals. *Psychonomic Bulletin & Review, 4(3),* 330-341.

Etz, A. & Vandenkerckhove, J. (2016). A Bayesian Perspective on the Reproducibility Project: Psychology: *Public Library of Science One, 11(2),* 1-12.

Evans, J.S. (2012). Questions and Challenges for the new Psychology of Reasoning. *Thinking and Reasoning, 18(1),* 5-31.

Evanschitzky, H., Baumgarth, C., Hubbard, R. & Armstrong, J.S. (2007). Replication Research's disturbing Trend. *Journal of Business Research, 60,* 411-415.

Evanschitzky, H. & Armstrong, J.S. (2013). Research with built-in Replication: Comment and further Suggestions for Replication Research. *Journal of Business Research, 66,* 1406-1408.

Eysenck, H.J. (1994). Meta-Analysis and its Problems. *British Medical Journal, 309(9),* 789-792.

Eysenck, H.J. (1995). Meta-Analysis squared – does it make Sense? *American Psychologist, 2,* 110-111.

Fahs, P.S., Morgan, L.L. & Kalman, M. (2003). A Call for Replication. *Journal of Nursing Scholarship, 35(1),* 67-72.

Farah, M.J. & Hook, C.J. (2013). The seductive Allure of ‚Seductive Allure'. *Perspectives on Psychological Science, 8(1),* 88-90.

Faucheux, C. (1976). Cross-cultural Research in experimental Social Psychology. *European Journal of Social Psychology, 6,* 269-322.

Faust, D. & Meehl, P.E. (1992). Using scientific Methods to resolve Questions in the History and Philosopy of Science: Some Illustrations. *Behavior Therapy, 23,* 195-211.

Fernandez-Duque, E. (1997). Comparing and Combining Data across Studies: Alternatives to Significance Testing. *Oikos, 79(3),* 616-618.

Fernandez-Duque, D., Evans, S., Colton, C. & Hodges S.D. (2015). Superfluous Neuroscience Information makes Explanations of psychological Phenomena more appealing. *Journal of cognitive Neuroscience, 27(5),* 926-944.

Fernández Vargas, M.A. (Hg). (2016). *Performance Epistemology: Foundations and Applications.* Oxford: University Press.

Festinger, L. (2001). *A Theory of Cognitive Dissonance.* Stanford: University Press.

Feyerabend, P. (1991). *Wider den Methodenzwang.* 3.Aufl. Frankfurt/M.: Suhrkamp.

Feynman, R. (1985). *Surely you're joking, Mr. Feynman!* New York: W.W. Norton & Co.

Fichte, J.G. (1801). *Sonnenklarer Bericht an das größere Publikum über das eigentliche Wesen der neuesten Philosophie.* Berlin: Realschulbuchhandlung.

Fiedler, K. (2011). Voodoo-Correlations are everywhere – not only in Neuroscience. *Perspectives on Psychological Science, 6(2),* 163-171.

Fiedler, K., Kutzner, F. & Krueger, J.I. (2012). The long Way from α-Error Control to Validity proper: Problems with a short-sighted false-positive Debate. *Perspectives on Psychological Science, 7(6),* 661-669.

Finetti, B.d. (1931). Sul Significato soggettivo della Probabilità. *Fundamenta Mathematicae, 17,* 298-323.

Finetti, B.d. (1972). *Probability, Induction and Statistics.* London: Wiley.

Finifter, B.M. (1972). The Generation of Confidence: Evaluating Research Findings by Random Subsample Replication. *Sociological Methodology, 4,* 112-175.

Finifter, B.M. (1975). Replication and Extension of social Research through secondary Analysis. *Social Science Information, 14(2),* 119-153.

Fisher, R.A. (1922). On the mathematical Foundations of theoretical Statistics. *Philosophical Transactions of the Royal Society, A, 222,* 309-368.

Fisher, R.A. (1928). *Statistical Methods for Research Workers.* 2.Aufl. London: Oliver & Boyd.

Fisher, R.A. (1930). Inverse Probability. *Mathematical Proceedings of the Cambridge Philsophical Society, 25,* 528-535.

Fisher, R.A. (1935). The Logic of inductive Inference. *Journal of the Royal Statistical Society, 98(1),* 39-54.

Fisher, R.A. (1935). *The Design of Experiments.* London: Oliver & Boyd.
Fisher, R.A. (1955). Statistical Methods and scientific Induction. *Journal of the Royal Statistical Society, Series B, 17(1),* 69-78.
Fiske, S. (2016). Mob Rule or Wisdom of Crowds? *Observer, 29(9),* 42-43.
Fodor, J. (1974). Special Sciences or: The Disunity of Science as Working Hypothesis. *Synthese, 28(2),* 97-115.
Fodor, J. (1997). Special Sciences: Still autonomous after all these Years. *Philosophical Perspectives, 11,* 149-163.
Fontana, F. (1787). Sopra la Forgente di molti Errori. In: Fontana, F. *Trattato del Veleno della Vipera de Veleni americani.* (S.165-176). Neapel: Nuova Società Letteraria.
Forcina, A. (2002). Probabilistic Modeling: An historical and philosophical Digression. In: J. Haitovsky, R. Lerche & Y. Ritov (Hg). *Foundations of statistical Inference.* (69-76). Heidelberg: Springer.
Francis, G. (2012). Too good to be true: Publication Bias in two prominent Studies from Experimental Psychology. *Psychonomic Bulletin & Review, 19(2),* 151-156.
Francis, G. (2012). The Psychology of Replication and Replication in Psychology. *Perspectives on Psychological Science, 7(6),* 585-594.
Franklin, A. & Howson, C. (1984). Why do Scientists prefer to vary their Experiments? *Studies in the History and Philosophy of Science, 15(1),* 51-64.
Franklin, A. (1994). How to avoid the Experimenters' Regress. *Studies in the History and Philosophy of Science, 25(3),* 463-491.
Freud, S. (1974). *Abriss der Psychoanalyse.* Frankfurt/M.: Fischer.
Freud, S. (1987). Entwurf einer Psychologie. In: *Gesammelte Werke.* Nachtragsband, (S.387-477). Frankfurt/M.: Fischer.
Freud, S. (2016). *Das Unbewusste.* Leipzig: Reclam.
Froman, T. & Shneyderman, A. (2004). Replicability reconsidered: An excessive Range of Possibilities. *Understanding Statistics, 3(4),* 365-373.
Früh, W. (1991). *Medienwirkungen: Das dynamisch-transaktionale Modell.* Opladen: Westdeutscher Verlag.
Fujioka, T. (2001). Asymptotic Approximations of the inverse Moment of the noncentral Chi-squared Variable. *Journal of the Japanese Statistical Society, 31(1),* 99-109.
Furchtgott, E. (1984). Replicate, again and again. *American Psychologist, 11,* 1315-1316.
Furman, J.L., Jensen, K. & Murray, F. (2012). Governing Knowledge in the Scientific Community: Exploring the Role of Retractions in Biomedicine. *Research Policy, 41(2),* 276-290.
Gall, F.J. (1791). *Philosophisch-medicinische Untersuchung ueber Natur und Kunst im kranken und gesunden Zustande des Menschen.* Wien: Rudolph Goesser & Co.
García-Pérez, M.A. (2012). Statistical Conclusion Validity: Some common Threats and simple Remedies. *Frontiers in Psychology, 3,* 1-11.
Garland, B. (Hg). (2004). *Neuroscience and the Law.* New York: Dana Press.
Garner, R., Gillingham, M.G., Kulikowich, J.M. & White, C.S. (1989). Effects of 'seductive Details' on Macroprocessing and Microprocessing in Adults and Children. *Cognition and Instruction, 6,* 41-57.
Gelman, A. & Shalizi, C.R. (2013). Philosophy and the Practice of Bayesian Statistics. *British Journal of Mathematical and Statistical Psychology, 66,* 8-38.

Gelman, A. (2016). Replication Crisis Crisis: Why I continue in my 'Pessimistic Conclusions about Reproducibility'. Blogeintrag vom 05.03.
Gelman, A. (2016). Why is the scientific Replication Crisis centered on Psychology? Blogeintrag vom 22.09.
Gelman, A. & Geurts, H.M. (2017). The statistical Crisis in Science: How is it relevant to Neuropsychology? *The Neuropsychologist,* (im Druck).
Gergen, K.J. (1973). Social Psychology as History. *Journal of Personality and Social Psychology, 26(2),* 309-320.
Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J. & Krüger, L. (1989). *The Empire of Chance: How Probability changed Science and Everyday Life.* Cambridge: University Press.
Gigerenzer, G. (1998). We need statistical Thinking, not statistical Rituals. *Behavioral and Brain Sciences, 21(2),* 199-200.
Gigerenzer, G. (2004). Mindless Statistics. *The Journal of Socio-Economics, 33,* 587-606.
Gigerenzer, G. (2008). *Bauchentscheidungen.* München: Goldmann.
Gigerenzer, G. (2009). Surrogates for Theory. *Observer, 22(2),* 21-23.
Gilbert, D., King, G., Pettigrew, S. & Wilson, T. (2016). Comment on 'Estimating the Reproducibility of Psychological Science'. *Science, 351(6277),* 1037-b.
Gilbert, D., King, G., Pettigrew, S. & Wilson, T. (2016). More on 'Estimating the Reproducibility of Psychological Science'. Webveröffentlichung vom 07.03.
Girotto, V. (2009). Un'Immagine del Cervello vale più di mille Parole. *Giornale Italiano di Psicologia, 36,* 325-328.
Glass, G.V. (1976). Primary, secondary, and Meta-Analysis of Research. *Educational Researcher, 5(10),* 3-8.
Gnedenko, B.W. (1957). *Lehrbuch der Wahrscheinlichkeitstheorie.* Berlin: Akademie-Verlag.
Gómez, O.S. & Juristo, N. (2010). Replication Types in experimental Disciplines. Tagungsbeitrag, September.
Gomez-Marin, A., Paton, J.J., Kampff, A.R., Costa, R.M. & Mainen, Z.F. (2014). Big behavioral Data: Psychology, Ethology and the Foundations of Neuroscience. *Nature Neuroscience, 17,* 1455-1462.
Goodman, N. (1973). *Fact, Fiction and Forecast.* 3.Aufl. Harvard: University Press.
Goodman, S.N. (1992). A Comment on Replication, *p*-Values and Evidence. *Statistics in Medicine, 11,* 875-879.
Goodman, S.N. (1999). Toward Evidence-Based Medical Statistics. *Annals of Internal Medicine, 130(12),* 1005-1013.
Goodman, S.N. (2001). Of *p*-Values and Bayes: A modest Proposal. *Epidemiology, 12(3),* 295-297.
Gopnik, A. & Meltzoff, A.N. (1997). *Words, Thoughts, and Theories.* Cambridge: MIT Press.
Gorroochurn, P., Hodge, S.E., Heiman, G.A., Durner, M. & Greenberg, D.A. (2007). Non-Replication of Association Studies: 'Pseudo-Failures' to replicate? *Genetics in Medicine, 9(6),* 335-341.
Greenfield, S. (2015). *Mind Change: How digital Technologies are leaving their Mark on our Brains.* New York: Random House.
Greenland, S., Schlesselman, J.J. & Criqui, M.H. (1986). The Fallacy of Employing standardized Regression Coefficients and Correlations as Measure of Effect. *American Journal of Epidemiology, 123(2),* 203-208.

Greenwald, A.G., Gonzalez, R, Harris, R.J. & Guthrie, D. (1996). Effect Sizes and p Values: What should be reported and what should be replicated? *Psychophysiology, 33,* 175-183.

Greiffenhagen, C. & Reeves, S. (2013). Is Replication important for Human Computer Interaction? *RepliCHI Proceedings*, 1-7.

Grissom, R.J. (1994). Probability of the superior Outcome of one Treatment over another. *Journal of Applied Psychology, 79,* 314-316.

Grissom, R.J. & Kim, J.J. (2001). Review of Assumptions and Problems in the appropriate Conceptualization of Effect Size. *Psychological Methods, 6(2),* 135-146.

Gross, M.T. (1997). The Need for Replication Studies – Is it really a done Deal? *Journal of Orthopaedic & Sports Physical Therapy, 25(3),* 161-162.

Grove, W.M. & Meehl, P.E. (1996). Comparative Efficiency of informal and formal Prediction Procedures: The clinical-statistical Controversary. *Psychology, Public Policy, and Law, 2,* 293-323.

Gruber, D. & Dickerson, J.A. (2012). Persuasive Images in popular Science: Testing Judgments of scientific Reasoning *and* Credibility. *Public Understanding of Science, 21(8),* 938-948.

Guest, O. (2016). Crisis in what exactly? *The Winnover,* 1-9, 14.06.

Gupta, S.D. & Perlman, M.D. (1974). Power of the noncentral *F*-Test. *Journal of the American Statistical Association, 69(345),* 174-180.

Guttman, L. (1985). The Illogic of Statistical Inference for cumulative Science. *Applied Stochastic Models and Data Analysis, 1,* 3-10.

Hacking, I. (1965). *Logic of Statistical Inference.* Cambridge: University Press.

Hacking, I. (1972). Likelihood. *The British Journal of the Philosophy of Science, 23(2),* 132-137.

Hacking, I. (1983). *Representing and Intervening.* Cambridge: University Press.

Hahn, U. (2011). The Problem of Circularity in Evidence, Argument, and Explanation. *Perspectives on Psychological Science, 6(2),* 172-182.

Halsey, L.G., Curran-Everett, D, Vowler, S.L. & Drummond, G.B. (2015). The fickle *p* Value generates irreproducible Results. *Nature Methods, 12(3),* 179-185.

Hamermesh, D.S. (2007). Replication in Economics. *The Canadian Journal of Economics, 40(3),* 715-733.

Hanson, R. (1958). The Logic of Discovery. *Journal of Philosophy, 55(25),* 1073-1089.

Hanson, R. (1985). Could Gambling save Science? Encouraging an honest Consensus. *Social Epistemology, 9,* 3-33.

Hanson, R., Oprea, R. & Porter, D. (2006). Information Aggregation and Manipulation in an experimental Market. *Journal of Economic Behavior & Organization, 60(4),* 449-459.

Harp, S.F. & Mayer, R.E. (1998). How seductive Details do their Damage: A Theory of cognitive Interest in Science Learning. *Journal of Educational Psychology, 90(3),* 414-434.

Harré, R. (1970). *The Principles of scientific Thinking.* London: Macmillan.

Harré, R. (Hg). (1975). *Problems of scientific Revolution: Progress and Obstacles to Progress in the Sciences.* Oxford: Clarendon Press.

Harris, R.J. (1985). *A Primer of Multivariate Statistics.* 2.Aufl. New York: Academic Press.

Hasler, F. (2013). *Neuromythologie.* 2.Aufl. Bielefeld: Transcript.

Hartshorne, J.K. & Schachner, A. (2012). Tracking Replicability as a Method of post-publication open Evaluation. *Frontiers in computational Neuroscience, 6(8),* 1-13.

Hawking, S. (2015). Zeitgeist Conference. In: Reisz, M. Is Philosophy dead? *Times Higher Education,* 22.02.

Head, M.L., Holman L., Lanfear, R., Kahn, A.T. & Jennions, M.D. (2015). The Extent and Consequences of *p*-Hacking in Science. *Public Library of Science Biology, 13(3),* 1-15.

Healy, M.J. (1978). Is Statistics a Science? *Journal of the Royal Statistical Society, Series A, 141(3),* 385-393.

Hedges, L. & Olkin, I. (1985). *Statistical Methods for Meta-Analysis.* New York: Academic Press.

Hedges, L. (1987) How hard is Hard Science, how soft is Soft Science? The empirical Cumulativeness of Research. *The American Psychologist, 42(2),* 443-455.

Hedges, L. &. Vevea, J. (1996). Estimating Effect Size under Publication Bias: Small Sample Properties and Robustness of a Random Effects Selection Model. *ournal of Educational and Behavioral Statistics, 21,* 299-332.

Hedges, L. (2007). Effect Sizes in cluster-randomized Designs. *Journal of Educational and Behavioral Statistics, 32,* 341-370.

Hegel, G.W.F. (1990). *Wissenschaft der Logik.* Frankfurt/M.: Suhrkamp.

Hegel, G.W.F. (1991). *Phänomenologie des Geistes.* 3.Aufl. Frankfurt/M.: Suhrkamp.

Heisenberg, W. (1973). Tradition in Science. *Bulletin of the Atomic Scientists, 29(10),* 4-10.

Hempel, C.G. (1945). Studies in the Logic of Confirmation. *Mind, 56,* 1-24.

Hempel, C.G. (1977). *Aspekte wissenschaftlicher Erklärung.* Berlin: de Gruyter.

Hendrick, C. (1990). Replications, Strict Replications, and Conceptual Replications: Are they important? *Journal of Social Behavior and Personality, 5(4),* 41-49.

Herrnson, P.S. (1995). Replication, Verification, secondary Analysis, and Data Collection in Political Science. *Political Science and Politics, 28(3),* 452-455.

Hess, B., Olejnik, S. & Huberty, C.J. (2001). The Efficacy of two Improvement-over-Chance Effect Sizes for two-Group univariate Comparisons under Variance Heterogeneity and Nonnormality. *Educational and Psychological Measurement, 61(6),* 909-936.

Hewitt, J.K. & Heath, A.C. (1987). A Note on Computing the Chi-Square Noncentrality Parameter for Power Analyses. *Behavior Genetics, 18(1),* 105-108.

Hill, A.B. (1965). Reflections on the controlled Trial. *Annals of the Rheumatic Diseases, 25,* 107-113.

Hill, A.B. (1971). *Principles of Medical Statistics.* 9.Aufl. London: Lancet,

Hochberg, Y. & Benjamini, Y. (1990). More powerful Procedures for Multiple Significance Testing. *Statistics in Medicine, 9,* 811-818.

Hogben, L. (1957). *Statistical Theory.* New York: W.W. Norton & Co.

Home, R.W. (Hg). (1983). *Science under Scrutiny.* Dordrecht: Reidel.

Hones, M.J. (1990). Reproducibility as a methodological Imperative in experimental Research. *Proceedings of the Biennial Meeting of the Philosophy of Science Association, 1,* 585-599.

Hook, C.J. & Farah, M.J. (2013). Look again: Effects of Brain Images and Mind-Brain Dualism on Lay Evaluations of Research. *Journal of cognitive Neuroscience, 25(9),* 1397-1405.

Hooke, R. (1707). *The posthumous Works.* London: Smith & Walford.

Hopf, E. (1934). On Causality, Statistics and Probability. *Journal of Mathematics and Physics, 13(1-4),* 51-105.
Hopkins, E.J., Weisberg, D.S. & Taylor, J.C. (2016). The seductive Allure is a reductive Allure: People prefer scientific Explanations that contain logically irrelevant reductive Information. *Cognition, 155,* 67-76.
Horgan, J. (1999) The undiscovered Mind: How the human Brain defies Replication, Medication, and Explanation. *Psychological Science, 10(6),* 470-474.
Horgan, J. (2016). Psychology's ongoing Credibility Crisis. *Scientific American Blog,* 07.03.
Hornbæk, K., Sander, S.S., Bargas-Avila, J.A. & Simonsens, G. (2014). Is once enough? On the Extent and Content or Replications in Human-Computer Interaction. *Proceedings of the Association for Computing Machinery,* 3523-3535.
Howell, D.C. (2010). *Statistical Methods for Psychology.* 7.Aufl. Belmont: Wordsworth.
Hox, J.J. (2010). *Multilevel Analysis: Techniques and Applications.* 2.Aufl. London: Routledge.
Hozo, S.P., Djulbegovic, B. & Hozo, I. (2005). Estimating the Mean and Variance from the Median, Range, and the Size of a Sample. *BioMed Central Medical Research Methodology, 5(13),* 1-10.
Hubbard, R., Vetter, D.E. & Little, E.L. (1998). Replication in Strategic Management: Scientific Testing for Validity, Generalizability, and Usefulness. *Strategic Management Journal, 19(3),* 243-254.
Hubbard, R. (2016). *Corrupt Research.* London: Sage.
Hume, D. (1854). *An Inquiry concerning the human Understanding.* Edinburgh: Adam & Charley Black.
Hume, D. (1854). *A Treatise of human Nature.* Edinburgh: Adam & Charley Black.
Hummel, H.J. (1972). *Probleme der Mehrebenenanalyse.* Stuttgart: Teubner.
Hunt, K. (1975). Do we really need more Replications? *Psychological Reports, 36,* 587-593.
Hunter, J.E. & Schmidt, F.L. (1996). Cumulative Research Knowledge and Social Policy Reform: The critical Role of Meta-Analysis. *Psychology, Public Policy, and Law, 2,* 324-347.
Hunter, J.E. (2001) The desperate Need for Replications. *Journal of Consumer Research, ,* 149-158.
Hurlbert, S.H. (1984). Pseudoreplication and the Design of ecological Field Experiments. *Ecological Monographs, 54(2),* 187-211.
Hüther, G. (2001). *Bedienungsanleitung für ein menschliches Gehirn.* Göttingen: Vandenhoeck & Ruprecht.
Inbar, Y. (2016). Association between contextual Dependence and Replicability in Psychology may be spurious. *Proceedings of the National Academy of Sciences, 113(34),* 4933-4934.
Ioannidis, J.P., Ntzani, E.E., Trikalinos, T.A. et al. (2001). Replication Validity of genetic Association Studies. *Nature Genetics, 29(3),* 306-309.
Ioannidis, J.P. (2005). Why most published Research Findings are false. *Public Library of Science Medicine, 2(8),* 696-701.
Ioannidis, J.P. (2009). Why most discovered true Associations are inflated. *Epidemiology, 19(5),* 640-648.
Ioannidis, J.P. (2012). Why Science is not necessarily self-correcting. *Perspectives on Psychological Science, 7(6),* 645-654.

Ioannidis, J.P. (2012). Scientific Inbreeding and Same-Team Replication. *Journal of Psychosomatic Research, 73(6),* 408-410.
Ioannidis, J.P. (2014). Why 'An Estimate of the Science-wise false Discovery Rate and Application to the top medical Literature' is false. *Biostatistics, 15(1),* 28-36.
Ioannidis, J.P. (2014). How to make more published Work true. *Public Library of Science Medicine, 11(10),* 1-8.
Ioannidis, J.P., Munafò, M.R., Fusar-Poli, P., Nosek, B.A. & David, S.P. (2014). Publication and other Reporting Biases in Cognitive Sciences: Detection, Prevalence, and Prevention. *Trends in Cognitive Sciences, 18(5),* 235-241.
Ioannidis, J.P. (2015). Failure to replicate: Sound the Alarm. *Cerebrum, 6,* 12-15.
Irwin, J.R. (2009). Equivalence of the Statistics for Replicability and Area under the ROC Curve. *British Journal of Mathematic and Statistical Psychology, 62,* 485-487.
Jager, L.R. & Leek, J.T. (2014). An Estimate of the Science-wise false Discovery Rate and Application to the top medical Literature. *Biostatistics, 15(1),* 1-12.
Iversen, G.I., Lee, M.D. & Wagenmakers, E.J. (2009). p_{rep} misestimates the Probability of Replication. *Psychonomic Bulletin & Review, 16(2),* 424-429.
Jeffreys, H. (1961). *Theory of Probability.* 3.Aufl. Oxford: Clarendon Press.
Jeffreys, H. (1973). *Scientific Inference.* 3.Aufl. Cambridge: University Press.
Jiang, J. (2007). *Linear and Generalized Linear Mixed Models and their Applications.* New York: Springer.
Johnson, D.H. (1999). The Insignificance of statistical Significance Testing. *Journal of Wildlife Management, 63(3),* 763-772.
Johnson, V.E. (2013). Revised Standards for statistical Evidence. *Proceedings of the National Academy of Sciences, 110(48),* 19313-19317.
Johnson-Laird, P.N. & Wason, P.C. (Hg). *Thinking: Readings in Cognitive Science.* (S. 307-314). Cambridge: University Press.
Jonas, E. & Körding, K. (2016). Could a Neuroscientist understand a Microprocessor? *bioRxiv Preprint,* 26. Mai, 1-15.
Jones, K.S., Derby, P.L. & Schmidlin, E.A. (2010). An Investigation of the Prevalence of Replication Research in Human Factors. *Human Factors, 52(5),* 586-595.
Jones, L.V. & Tukey, J.W. (2000) A sensible Formulation of the Significance Test. *Psychological Methods, 5(4),* 411-414.
Judd, C.M., Westfall, J. & Kenny, D.A. (2012). Treating Stimuli as Random Factor in Social Psychology. *Journal of Personality and Social Psychology, 103(1),* 54-69.
Kadane, J.B., Schervish, M.J. & Seidenfeld, T. (1999). *Rethinking the Foundations of Statistics.* Cambridge: University Press.
Kagan, J. (2000). *Die drei Grundirrtümer der Psychologie.* Weinheim: Beltz.
Kahneman, D. & Tversky, A. (1973). On the Psychology of Prediction. *Psychological Review, 80(4),* 237-251.
Kahneman, D. & Tversky, A. (1982). On the Study of statistical Intuitions. *Cognition, 11(2),* 123-141.
Kahneman, D., Slovic, P. & Tversky, A. (Hg). (1982). *Judgment under Uncertainty: Heuristics and Biases.* Cambridge: University Press.
Kahneman, D. (2011). *Thinking, fast and slow.* London: Random House.
Kahneman, D. (2014). A new Etiquette for Replication. *Social Psychology, 45(4),* 310-311.

Kane, E.J. (1984). Why Journal Editors should encourage the Replication of applied econometric Research. *Quarterly Journal of Business and Economics, 23(1),* 3-8.

Kane, M.T. (1992). An Argument-based Approach to Validity. *Psychological Bulletin, 112(3),* 527-535.

Kant, I. (1974). *Kritik der reinen Vernunft.* Frankfurt/M.: Suhrkamp.

Kantorovich, A. (1993). *Scientific Discovery: Logic and Tinkering.* New York: State University Press.

Karmiloff-Smith, A. & Inhelder, B. (1977). If you want to get ahead, get a Theory. In: Johnson-Laird, P.N. & Wason, P.C. (Hg). *Thinking: Readings in Cognitive Science.* (S. 293-306). Cambridge: University Press.

Karmiloff-Smith, A. (1988). The Child is a Theoretician, not an Inductivist. *Mind & Language, 3(3),* 183-195.

Katz, D. & Allport, F.H. (1931). *Students' Attitudes: A Report of the Syracuse University Reaction Study.* New York: Craftsman.

Keats, J. (2016). The Replication Crisis. *Discover Magazine, 8,* 14.

Keehner, M., Mayberry, L. & Fischer, M. H. (2011). Different Clues from different Views: The Role of Image Format in public Perceptions of neuroimaging Results. *Psychonomic Bulletin Review, 18,* 422-428.

Keiding, N. (2010). Reproducible Research and the substantive Context. *Biostatistics, 11(3),* 376-378.

Keil, F.C. (2006). Explanation and Understanding. *Annual Review of Psychology, 57,* 227-254.

Kelley, K., Maxwell, S.E. & Rausch, J.R. (2003). Obtaining Power or Obtaining Precision: Delineating Methods of Sample-Size Planning. *Evaluation and the Health Professions, 26(3),* 258-287.

Kelley, K. & Maxwell, S.E. (2003). Sample Size for Multiple Regression: Obtaining Regression Coefficients that are accurate, not simply significant. *Psychological Methods, 8(3),* 305-321.

Kelley, K. & Rausch, J.R. (2006). Sample Size Planning for standardized Mean Difference: Accuracy in Parameter Estimation via narrow Confidence Intervals. *Psychological Methods, 11(4),* 363-385.

Kelly, C., Chase, L.W. & Tucker, R.K. (1979). Replication in experimental Communication Research: An Analysis. *Human Communication Research, 5(4),* 338-342.

Kelly, C. (2006) Replicating empirical Research in Behavioral Ecology: How and why it should be done but rarely ever is. *The Quarterly Review of Biology, 81(3),* 221-236.

Kemeny, J.G. & Oppenheim, P. (1952). Degree of factual Support. *Philosophy of Science, 19(4),* 307-324.

Kennet, R.S. & Zacks, S. (2014). *Modern Industrial Statistics.* Sussex: Wiley.

Kennet, R.S. & Shmueli, G. (2015). Clarifying the Terminology that describes scientific Reproducibility. *Nature Methods, 12(8),* 699.

Keynes, J.M. (1936). *The General Theory of Employment, Interest, and Money.* London: Macmillan.

Keynes, J.M. (1973). *A Treatise on Probability.* London: Macmillan.

Killeen, P.R. (2005). An Alternative to Null-Hypothesis Significance Tests. *Psychological Science, 16(5),* 345-353.

Killeen, P.R. (2005). Replicability, Confidence, and Priors. *Psychological Science, 16(12),* 1009-1012.
Killeen, P.R. (2008). Replication Statistics. In: Osborne, J. (Hg). *Best Practices in quantitative Methods.* (S. 103-124). London: Sage.
King, G. (1995). Replication, Replication. *Political Science & Politics, 28,* 444-451.
King, G. (2003). The Future of Replication. *International Studies Perspective, 4(1),* 100-105.
Kirkwood, B.L. (1981). Bioequivalence Testing – A Need to rethink. *Biometrics, 37(3),* 589-594.
Kitchin, R. (2014). Big Data, new Epistemologies and Paradigm Shifts. *Big Data & Society, 2,* 1-12.
Klaczynski, P.A. (2000). Motivated scientific Reasoning Biases, epistemological Beliefs, and Theory Polarization. *Child Development, 71(5),* 1347-1366.
Klahr, D., Dunbar, K. & Fay, A.L. (1990). Designing good Experiments to test bad Hypotheses. In: J. Shrager & P. Langley, P. (Hg). *Computational Models of scientific Discovery and Theory Formation.* (S. 355-402). San Mateo: Morgan Kaufmann.
Klahr, D. & Simon, H.A. (1999). Studies of scientific Discovery: Complementary Approaches and convergent Findings. *Psychological Bulletin, 125(5),* 524-543.
Klayman, J. & Ha, Y.W. (1987). Confirmation, Disconfirmation, and Information in Hypothesis Testing. *Psychological Review, 94(2),* 211-228.
Klein, R.A., Ratliff, K.A., Vianello, M. et al. (2014). Investigating Variation in Replicability: A 'Many Labs' Replication Project. *Social Psychology, 45(3),* 142-152.
Klein, R.A., Ratliff, K. A., Vianello, M. et al. (2014). Theory Building through Replication. *Social Psychology, 45(4),* 307-310.
Kline, R.B. (2013). *Beyond Significance Testing: Statistics Reform in the Behavioral Sciences.* 2.Aufl. Washington: American Psychological Association.
Knorr-Cetina, K. (1977). Producing and Reproducing Knowledge: descriptive or constructive? *Sociology of Science, 16(6),* 669-696.
Kripke, S.A. (1989). *On Rules and Private Language.* Oxford: Basil Blackwell.
Krueger, J. (1999). Significance Testing does not solve the Problem of Induction. *Psycholoquy, 10(15),* 1-5.
Krüger, L. (1990). Method, Theory, and Statistics: The Lesson of Physics. In: Cooke, R. & Constantini (Hg). *Statistics in Science.* (S. 1-14). Dordrecht: Kluwer.
Kruskal, W. (1981). Statistics in Society: Problems unsolved and unformulated. *Journal of the American Statistical Association, 76(375),* 505-515.
Kubokawa, T., Robert, C.P. & Saleh, A.K. (1993). Estimation of Noncentrality Parameters. *The Canadian Journal of Statistics, 21(1),* 45-57.
Kuhn, T.S. (1991). *Die Struktur wissenschaftlicher Revolutionen.* 11.Aufl. Frankfurt/M.: Suhrkamp.
Kullback, S. & Leibler, R.A. (1951). On Information and Sufficiency. *Annals of Mathematical Statistics, 22,* 79-86.
Küng, H. (1990). *Projekt Weltethos.* München: Piper.
Kyburg, H.E. (1974). *The logical Foundations of statistical Inference.* Dordrecht: Reidel.
Kyburg, H.E. (1990). *Science and Reason.* Oxford: University Press.
Lakatos, I. (1978). *Methodologie der wissenschaftlichen Forschungsprogramme.* Braunschweig: Vieweg.

Lakens, D. (2016). The statistical Conclusions in Gilbert et al. are completely invalid. *The 20% Statistician,* Blogeintrag vom 06.03.
Lamal, P.A. (1990). On the Importance of Replication. *Journal of social Behavior and Personality, 5(4),* 31-35.
Langer, W. (2004). *Mehrebenenanalyse: Eine Einführung für Forschung und Praxis.* Wiesbaden: Verlag für Sozialwissenschaften.
Laplace, P.S. (1840). *Essai philosophique sur les Probabilités.* 6.Aufl. Paris: Bachelier.
La Sorte, M.A. (1972). Replication as a Verification Technique in Survey Research: a Paradigm. *The Sociological Quarterly, 13(2),* 218-227.
Latour, B. (1990). Visualization and Cognition: Thinking with Eyes. In: Lynch, M & Woolgar, S. (Hg). *Representation in scientific Practice.* (S. 19-68). Cambridge: MIT Press.
Latour, B. (2000). Factures/Fractures: De la Notion de Réseau à celle d'Attachement. In: Micoud, A. & Peroni, M. (Hg). *Ce qui nous relie.* (S. 189-208). Paris: Édition de l'Aube.
Laudan, L. (1977). *Progress and its Problems: Towards a Theory of scientific Growth.* London: Routledge & Kegan Paul.
Laws, K.R. (2016). Psychology, Replication and beyond. *BioMed Central Psychology, 4(30),* 1-8.
Laws, K.R. (2016). Is Psychology really in Crisis? *The Conversation,* 27.06.
Lawson, H. & Appignanesi, L. (Hg). (1989). *Dismantling Truth: Reality in the post-modern World.* London: Weidenfeld & Nicholson.
Leamer, E.E. (1974). False Models and post-Data Model Construction. *Journal of the American Statistical Association, 69(345),* 122-131.
Leamer, E.E. (1983). Let's take the Con out of Econometrics. *The American Economic Review, 73(1),* 31-43.
LeBel, E.P. (2015). A new Replication Norm for Psychology. *Collabra, 1(4),* 1-13.
Lecoutre, B. & Poitevineau, J. (2014). *The Significance Test Controversy revisited.* Heidelberg: Springer.
Legrenzi, P. & Umiltà, C. (2009). *Neuro-Mania.* Bologna: Il Mulino.
Lehmann, E.L. & Casella G. (1998). *Theory of Point Estimation.* 2.Aufl. New York: Springer.
Lele, S.R. (2004). Evidence Functions and the Optimality of the Law of Likelihood. In: Taper, M.L. & Lele, S.R. (Hg). *The Nature of scientific Evidence.* (S. 191-203). Chicago: University Press.
LeMoal, M. & Swendsen, J. (2015). Sciences of the Brain: The long Road to scientific Maturity and to present-Day Reductionism. *Comptes Rendus Biologies, 338(8-9),* 593-601.
Lenth, R.V. (2001). Some practical Guidelines for effective Sample Size Determination. *The American Statistician, 55(3),* 187-193.
Lévi-Strauss, C. (1962). *La Pensée sauvage.* Paris: Librairie Pion.
Levy, P. (1967). Substantive Significance of significant Differences between two Groups. *Psychological Bulletin, 67(1),* 37-40.
Lewin, K. (1951). *Field Theory in Social Science: selected theoretical Papers.* New York: Harper & Row.
Lewin, K. (1967). *Gesetz und Experiment in der Psychologie.* Darmstadt: Wissenschaftliche Buchgesellschaft.
Limentani, G.B., Ringo, M.C., Ye, F., Bergquist, M.L. & McSorley, E.O. (2005). Beyond the *t*-Test. *Analytical Chemistry, 6,* 221-226.

Lindlay, D.V. (1957). A statistical Paradox. *Biometrika, 44(1/2),* 187-192.
Lindlay, D.V. (2000). The Philosophy of Statistics. *Journal of the Royal Statistical Society, Series D, 49(3),* 293-319.
Lindsay, R.M. & Ehrenberg, A.S. (1993). The Design of replicated Studies. *The American Statistician, 47(3),* 217-228.
Lindsay, D.S. (2016). Replication and Effect Size in Psychological Science: Comments. Webveröffentlichung vom 06.03.
Lipsey, M.W. (1990). *Design Sensitivity: Statistical Power for experimental Research.* London: Sage.
Lipton, P. (1991). *Inference to the best Explanation.* London: Routledge.
Liu, Y.J., Papasian, C., Hamilton, J. & Deng, H.W. (2008). Is Replication the Gold Standard for Validating Genome-wide Association Findings? *Public Library of Science, 3(12),* e4037.
Loftus, G.R. (1996). Psychology will be a much better Science when we change the Way we analyze Data. *Current Directions in Psychological Science, 5(6),* 161-171.
Lombrozo, T. & Carey, S. (2004). Functional Explanation and the Function of Explanation. *Cognition, 99,* 167-204.
Lombrozo, T., Kelemen, D. & Zaitchik, D. (2007). Inferring Design: Evidence of a Preference for teleological Explanations in Patients with Alzheimer‘s Disease. *Psychological Science, 18(11),* 999-1006.
Lord, C.G., Ross, L. & Lepper, M. (1979). Biased Assimilation and Attitude Polarization: The Effects of prior Theories on subsequently considered Evidence. *Journal of Personality and Social Psychology, 37(11),* 2098-2109.
Loscalzo, J. (2012). Irreproducible experimental Results: Causes, (Mis)interpretations, and Consequences. *Circulation, 25(10),* 1211-1215.
Luh, W.M. & Guo, J.H. (2011) Developing Noncentrality Parameter for Calculating Group Sample Sizes in heterogeneous Analysis of Variance. *The Journal of Experimental Education, 79,* 53-63.
Luhmann, N. (1990). *Die Wissenschaft der Gesellschaft.* Frankfurt/M.: Suhrkamp.
Lykken, D.T. (1968). Statistical Significance in psychological Research. *The American Statistician, 47(3),* 217-228.
Lynch, M. & Woolgar, S. (Hg). (1990). *Representation in scientific Practice.* Cambridge: MIT Press.
Lynch, J.K., Bradlow, E.T., Huber, J.C. & Lehmann, D.R. (2015) Reflections on the Replication Corner: In Praise of conceptual Replication. *International Journal of Research in Marketing, 32(2),* 1-13.
Maas, C.J. & Hox, J.J. (2005). Sufficient Sample Sizes for Multilevel Modeling. *Methodology, 1(3),* 86-92.
Mackey, A. (2012). Why (or why not), when, and how to replicate Research. In: Porte, G. (Hg). *Replication Research in Applied Linguistics.* (S. 21-46). Cambridge: University Press.
Mahoney, M.J. (1980). Rationality and Authority: On the Confusion of Justification and Permission. *Social Studies of Science, 10(4),* 515-518.
Makel, M.C., Plucker, J.A. & Hegarty, B. (2012). Replication in Psychology Research: How often do they really occur? *Perspectives in Psychological Science, 7(6),* 537-542.
Makel, M.C. (2014). The empirical March: Making Science better at self-Correction. *Psychology of Aesthetics, Creativity, and the Arts, 8(1),* 2-7.

Mantel, R. (1976). Homothetic Preferencees and Community Excess Demand Functions. *Journal of Economic Theory, 12,* 197-202.
Markowetz, A., Balszkiewicz, K., Montag, C., Switala, C. & Schlaepfer, T.E. (2014). Psycho-Informatics: Big Data shaping modern Psychometrics. *Medical Hypotheses, 82(4),* 405-411.
Marx, K. (1957). *Das Kapital.* Stuttgart: Kröner.
Maxwell, S.E., Kelley, K. & Rausch, J.R. (2009). Sample Size Planning for statistical Power and Accuracy in Parameter Estimation. *The Annual Review of Psychology, 59,* 537-563.
Maxwell, S.E., Lau, M.Y. & Howard, G. (2015). Is Psychology suffering from a Replication Crisis? What does 'Failure to replicate' really mean? *American Psychologist, 70(6),* 487-498.
Mayo, D.G. (1983). An objective Theory of statistical Testing. *Synthese, 57,* 297-340.
Mayo, D.G. (2004). An Error-Statistical Philosophy of Evidence. In: Taper, M.L. & Lele, S.R. (Hg). *The Nature of scientific Evidence.* (S. 79-96). Chicago: University Press.
Mayo, D.G. (2016). Repligate returns (or, the non-Significance of non-significant Results are the new significant Results). Blogeintrag vom 04.03.
McCabe, D.P. & Castel, A.D. (2008). Seeing is Believing: The Effect of Brain Images on Judgments of scientific Reasoning. *Cognition, 107,* 334-352.
McElreath, R. & Smaldino, P.E. (2015). Replication, Communication, and the Population Dynamics of scientific Discovery. *Public Library of Science One, 8,* 1-16.
McGraw, K.O. & Wong, S.P. (1992). A Common Language Effect Size Statistic. *Psychological Bulletin, 111,* 361-365.
McGuigan, F.J. (1956). Confirmation of Theories in Psychology. *Psychological Review, 63(2),* 98-104.
McNutt, M. (2014). Reproducibility. *Science, 343(6168),* 229.
Medawar, P.B. (1963). Is the scientific Paper a Fraud? *Listener, 70,* 377-378.
Medawar, P.B. (1969). *Induction and Intuition in scientific Thought.* Philadelphia: American Philosophical Society.
Meehl, P.E. (1967). Theory-Testing in Psychology and Physics: A methodological Paradox. *Philosophy of Science, 34,* 103-115.
Meehl, P.E. (1978). Theoretical Risks and Tabular Asterisks. *Journal of Consulting and Clinical Psychology, 46,* 806-834.
Meehl, P.E. (1990). Appraising and Amending Theories. *Psychological Inquiry, 1(2),* 108-141.
Meehl, P.E. (1992). Cliometric Metatheory: The actuarial Approach to empirical, history-based Philosophy of Science. *Psychological Reports, 71,* 339-467.
Meehl, P.E. (2002). Cliometric Metatheory: II. Criteria Scientists use in Theory Appraisal and why it is rational to do so. *Psychological Reports, 91,* 339-404.
Meehl, P.E. (2004). Cliometric Metatheory: III. Peircean Consensus, Verisimilitude and Asymptotic Method. *British Journal of the Philosophy of Science, 55,* 615-643.
Meier, K.J. (1995). Replication: A View from the Streets. *Political Science & Politics, 28(3),* 456-459.
Mercier, H. & Sperber, D. (2011). Why do Humans reason? Arguments for an Argumentative Theory. *Behavioral and Brain Sciences, 34,* 57-73.
Merton, R.K. (1973). *The Sociology of Science.* Chicago: University Press.

Meyer, M.N. & Chabris, C. (2014). Why Psychologists' Food Fight matters: 'Important Findings' haven't been replicated, and Science may have to change its Ways. Webveröffentlichung vom 31.07.

Michael, R.B., Newman, E. J., Vuorre, M., Cumming, G. & Garry, M. (2013). On the (non)persuasive Power of a Brain Image. *Psychonomic Bulletin Review, 20,* 720-725.

Michie, S. & Johnston, M. (2012). Theories and Techniques of Behavior Change: Developing a cumulative Science of Behavior Change. *Health Psychology Review, 6(1),* 1-6.

Mill, J.S. (1974). *A System of Logic.* Toronto: University Press.

Miller, J. (2009). What is the Probability of Replicating a statistically significant Effect? *Psychonomic Bulletin and Review, 16(4),* 617-640.

Miller, M. (1986). Mehrebenen-Itemanalyse. In: M.v. Saldern (Hg). *Mehrebenenanalyse.* (S.82-98). Weinheim: Beltz.

Minahan, J. & Siedlecki, K.L. (2016). Individual Differences in Need for Cognition influence the Evaluation of circular scientific Explanations. *Personality and Individual Differences, 99,* 113-117.

Mischel, W. (2005). Alternative Futures of our Science. *Observer, 18(3),* 41-45.

Mischel, W. (2009). Becoming a cumulative Science. *Observer, 22(1),* 21-23.

Mises, R.v. (1951). *Wahrscheinlichkeit, Statistik und Wahrheit.* 3.Aufl. Wien: Springer.

Mittelstaedt, R.A. & Zorn, T.S. (1984). Econometric Replication: Lessons from the experimental Sciences. *Quarterly Journal of Business and Economics, 23(1),* 9-15.

Moerbeek, M. & Teerenstra, S. (2016). *Power Analysis of Trials with Multilevel Data.* Boca Raton: Chapman & Hall.

Moinester, M. & Gottfried, R. (2014). Sample Size Estimation for Correlations with pre-specified Confidence Interval. *The Quantitative Methods for Psychology, 10(2),* 124-130.

Monin, B. & Oppenheimer, D.M. (2014). The Limits of direct Replications and the Virtues of Stimulus Sampling. *Social Psychology, 45(4),* 299-300.

Moonesinghe, R., Khoury, M. J. & Janssens, A. C. J. (2007). Most published Research Findings are false – but a little Replication goes a long Way. *Public Library of Science Medicine, 4(2),* 218-221.

Morrison, D.E. & Henkel, R.E. (Hg). (1970). *The Significance Test Controversy.* Chicago: Aldane Publishing.

Morse, S.J. (2004). New Neuroscience, old Problems. In: B. Garland (Hg). *Neuroscience and the Law.* (S.157-200). New York: Dana Press.

Mulkay, M. (1984). The Scientist talks back: A One-Act Play, with a Moral, about Replication in Science and Reflexivity in Sociology. *Social Studies of Science, 14(2),* 265-283.

Mulkay, M. & Gilbert, G.N. (1986). Replication and mere Replication. *Philosophy of the Social Sciences, 16,* 21-37.

Muller, K.E., LaVange, L., Ramey, S.L. & Ramey, C.T. (1992). Power Calculations for General Multivariate Models including repeated Measures Applications. *Journal of the American Statistical Association, 87(420),* 1209-1226.

Muller, K.E. & Benignus, V.A. (1992). Increasing scientific Power with statistical Power. *Neurotoxicology and Teratology, 14,* 211-219.

Muller, K.E. & Pasour, V.B. (1997). Bias in Linear Model Power and Sample Size due to Estimating Variance. *Communications in Statistics: Theory and Methods, 26(4),* 839-851.

Muma, J.R. (1993). The Need for Replication. *Journal of Speech, Language, and Hearing Research, 36,* 927-930.

Munro, G.D. & Munro, C.A. (2014). 'Soft' versus 'hard' psychological Science: Biased Evaluations of sctientific Evidence that threatens or supports a strongly held political Identity. *Basic and Applied Social Psychology, 36,* 533-543.

Münsterberg, H. (1891). *Ueber Aufgaben und Methoden der Psychologie.* Leipzig: Verlag von Ambrosius Abel.

Murayama, K., Pekrun, R. & Fiedler, K. (2014). Research Practices that can prevent an Inflation of false-positive Rates. *Personality and Social Psychology Review, 18(2),* 107-118.

Murray, L.W. & Dosser, D.A. (1987). How significant is a significant Difference? Problems with the Measurement of Magnitude of Effect. *Journal of Counseling Psychology, 34(1),* 68-72.

Murtaugh, P.A. (2014) In Defense of *p* Values. *Ecology, 95(3),* 611-617.

Nagel, E. (1949). Principles of the Theory of Probability. *International Encyclopedia of Unified Science, 1(6),* 1-80.

Nakagawa, S. (2004). A Farewell to Bonferroni: the Problems of low statistical Power and Publication Bias. *Behavioral Ecology, 15(6),* 1044-1045.

Nakagawa, S. & Foster, T.M. (2004). The Case against retrospective statistical Power Analyses with an Introduction to Power Analysis. *Acta Ethologica, 7,* 103-108.

Nakagawa, S. & Parker, T.H. (2015). Replication Research in Ecology: Feasibility, Incentives, and the Cost-Benefit-Conundrum. *BioMed Central Biology, 13,* 88.

Nayfach, S. & Pollard, K.S. (2016). Toward accurate and quantitative comparative Metagenomics. *Cell, 166(5),* 1103-1116.

Nelder, J.A. (1986). Statistics, Science and Technology. *Journal of the Royal Statistical Society, Series A, 149(2),* 109-121.

Nelson, L. (1973). *Geschichte und Kritik der Erkenntnistheorie.* Hamburg: Meiner.

Neta, R. (2006). Epistemology factualized: New contractarian Foundations for Epistemology. *Synthese, 150(2),* 247-280.

Neta, R. (2013). What is an Inference? *Philosophical Issues, 23,* 388-407.

Neta, R. (2016). Epistemic Circularity and Virtuous Coherence. In: Fernández Vargas, M.A. (Hg). *Performance Epistemology.* (S. 224-248). Oxford: University Press.

Neuliep, J.W. (Hg). (1991). *Replication Research in the Social Sciences.* London: Sage.

Neuliep, J. W. & Crandall, R. (1993). Everyone was wrong: There are lots of Replications out there. *Journal of social Behavior and Personality, 8(6),* 1-8.

Neurath, O. (1932). Protokollsätze. *Erkenntnis, 3(1),* 204-214.

Neurath, O. (1979). *Wissenschaftliche Weltauffassung, Sozialismus und Logischer Empirismus.* Frankfurt/M.: Suhrkamp.

Newton, I. (1999). *Philosphia Naturalis Principia Mathematica.* 3.Aufl. Oakland: University of California Press.

Neyman, J. & Pearson, E. (1928). On the Use and Interpretation of certain Test Criteria for Purposes of Statistical Inference. *Biometrika, 20(1/2),* 175-240.

Neyman, J. & Pearson, E. (1933). On the Problem of the most efficient Tests of statistical Hypotheses. *Philosophical Transactions of the Royal Society, Series A, 231,* 289-337.

Neyman, J. (1936). Outline of a Theory of statistical Estimation based on the classical Theory of Probability. *Philosophical Transactions of the Royal Society, Series A, 234,* 333-380.
Neyman, J. (1941). Fiducial Argument and the Theory of Confidence Intervals. *Biometrika, 32(2),* 128-150.
Neyman, J. (1942). Basic Ideas and some recent Results of the Theory of Testing Hypotheses. *Journal of the Royal Statistical Society, 105(4),* 292-327.
Neyman, J. (1955). Statistics – Servant of all Sciences, *Science, 122(3166),* 401-406.
Neyman, J. (1955). The Problem of inductive Inference. *Communications on Pure and Applied Mathematics, 8,* 13-46.
Neyman, J. (1957). 'Inductive Bahvior' as a basic Concept of Philosophy of Science. *Review of the International Statistical Institute, 25(1/3),* 7-22.
Nezlek, J.B., Schröder-Abé, M. & Schütz, A. (2006). Mehrebenenanalysen in der psychologischen Forschung: Vorteile und Möglichkeiten der Mehrebenenmodellierung mit Zufallskoeffizienten. *Psychologische Rundschau, 57(4),* 213-223.
Ng, M. & Wilcox, R.R. (2011). A Comparison of Two-Stage Procedures testing Least-Squares Coefficients under Heteroscedasticity. *British Journal of Mathematical and Statistical Psychology, 64,* 244-258.
Nicod, J. (1924). *Le Problème logique de l'Induction.* Paris: Félix Alcan.
Niederman, F. & March, S. (2015). Reflections on Replication. *Association of Information Systems Transactions on Replication Research, 1,* 7.
Nietzsche, F. (1988). *Die fröhliche Wissenschaft.* Berlin: dtv/de Gruyter.
Nisbett, R. E., Krantz, D.H., Jepson, C. & Fong, G.T. (1982). Improving inductive Inference. In: Kahneman, D., Slovic, P. & Tversky, A. (Hg). *Judgment under Uncertainty: Heuristics and Biases.* (S. 445-461). Cambridge: University Press.
Nosek, B.A. & Lakens, D. (2014). Registered Reports: A Method to increase the Credibility of published Results. *Social Psychology, 45(3),* 137-141.
Nosek, B.A., Alter, G., Banks, G.C. et al. (2015). Promoting Open Research Culture. *Science, 348(6242),* 1422-1425.
Nosek, B.A., Anderson, C.J., Zuni, K. et al. (2016). Response to Comment on 'Estimating the Reproducibility of Psychological Science'. *Science, 351(6277),* 1037-c.
Nosek, B.A. (2016). Let's not mischaracterize Replication Studies. *Retraction Watch,* 09.03.
Nowotny, H., Scott, P. & Gibbons, M. (2005). *Wissenschaft neu denken.* 2.Aufl. Weilerswist: Velbrück.
Nuzzo, R. (2014). Statistical Errors. *Nature, 506,* 150-152.
Nuzzo, R. (2015). Fooling ourselves. *Nature, 526,* 182-185.
Oakes, M. (1986). *Statistical Inference: A Commentary for the Social and Behavioral Sciences.* New York: Wiley.
Oaksford, M. & Chater, N. (2007). *Bayesian Rationality: The probabilistic Approach to human Reasoning.* Oxford: University Press.
Oksanen, L. (2001). Logic of Experiments in Ecology: is Pseudoreplication a Pseudoissue? *Oikos, 94,* 27-38.
Open Science Collaboration. (2012). An open, large-scale, collaborative Effort to estimate the Reproducibility of Psychological Science. *Perspectives on Psychological Science, 7(6),* 657-660.

Open Science Collaboration. (2015). Estimating the Reproducibility of Psychological Science. *Science, 349(6251),* 943, aac4716.

Parisi, D. (2009). Non si può capire la Mente senza studiare il Cervello. *Giornale Italiano di Psicologia, 2,* 279-284.

Park, I.U., Peacey, M.W. & Munafò, M.R. (2014). Modelling the Effects of subjective and objective Decision Making in scientific Peer Review. *Nature, 506,* 93-95.

Parkhurst, D.F. (2001). Statistical Significance Tests: Equivalence and Reverse Tests should reduce Misinterpretation. *BioScience, 51(12),* 1051-1057.

Pashler, H. & Harris, C.R. (2012). Is the Replicability Crisis overblown? *Perspectives on Psychological Science, 7(6),* 531-536.

Pearson, E.S. & Neyman, J. (1930). On the Problem of two Samples. *Bulletin of the Polish Academy of Sciences,* 73-96.

Pearson, E.S. & Hartley, H.O. (1951). Charts of the Power Function for Analysis of Variance Tests, derived from the non-central F-Distribution. *Biometrika, 38(1/2),* 112-130.

Pearson, E.S. (1955). Statistical Concepts in the Relation to Reality. *Journal of the Royal Statistical Society, Series B, 17(2),* 204-207.

Pearson, K. (1900). *The Grammar of Science.* 2.Aufl. London: Black.

Pearson, K. (1920). The fundamental Problem of Practical Statistics. *Biometrika, 13(1),* 1-16.

Pedersen, D.B. & Hendricksen, V.F. (2014). Science Bubbles. *Philosophy & Technology, 27(4),* 503-518.

Peng, R. (2011). Reproducible Research in Computational Science. *Science, 334(6060),* 1226-1227.

Peng, R. (2016). A simple Explanation for the Replication Crisis in Science. *Simply Statistics,* Blogeintrag vom 26.08.

Perlman, M.D. & Wu, L. (1999). The Emperor's new Tests. *Statistical Science, 14(4),* 355-381.

Perugini, M., Gallucci, M. & Constantini, G. (2014). Safeguard Power as a Protection against imprecise Power Estimates. *Perspectives on Psychological Science, 9(3),* 319-332.

Pfeifer, M.P. & Snodgrass. (1990). The continued Use of the continued Use of retracted, invalid scientific Literature. *Journal of the American Medical Association, 263(10),* 1420-1423.

Piaget, J. (1951). *Introduction à l'Épistémologie Génétique.* Paris: Presses Universitaires de France.

Piketty, T. (2013). *Le Capitale au XXième Siècle.* Paris: Éditions du Seuil.

Plant, R.P. (2016). A Reminder on Millisecond Timing Accuracy and potential Replication Failure in Computer-based Psychology Experiments. *Behavior Research Methods, 48(1),* 408-411.

Platt, J.R. (1964). Strong Inference: Certain systematic Methods of scientific Thinking may produce much more rapid Progress than others. *Science, 146(3642),* 347-353.

Polanyi, K. (1957). *The Great Transformation.* Boston: Beacon Press.

Poldrack, R.A. & Poline, J.B. (2014). The Publication and Reproducibility Challenges of shared Data. *Trends in Cognitive Science, 19(2),* 59-61.

Poletiek, F.H. (1996). Paradoxes of Falsification. *Quarterly Journal of Experimental Psychology, Section A: Human Experimental Psychology, 49(2),* 447-462.

Polio, C. & Gass, S. (1997). Replication and Reporting: A Commentary. *Studies in Second Language Acquisition, 19,* 499-508.
Polkinghorne, D.E. (2007). Validity Issues in Narrative Research. *Qualitative Inquiry, 13(4),* 471-486.
Popper, K.R. (1989). *Logik der Forschung.* 9.Aufl. Tübingen: J.C.B Mohr.
Popper, K.R. & Eccles, J. (1982). *Das Ich und sein Gehirn.* München: Piper.
Pornprasertmanit, S. & Schneider, W.J. (2014). Accuracy in Parameter Estimation in clustered randomized Designs. *Psychological Methods, 19(3),* 356-379.
Porte, G. (2013). Who needs Replication? *Callico Journal, 30(1),* 10-15.
Quine, W.v.O. (1964) *Ontological Relativity and other Essays.* 2.Aufl. Harvard: University Press.
Quine, W.v.O. (1996). *Pursuit of Truth.* 3.Aufl. Harvard: University Press.
Radder, H. (1992). Experimental Reproducibility and the Experimenters' Regress. *Proceedings of the Biennal Meeting of the Philosophy of Science Association, 1,* 63-73.
Raiffa, H. & Schlaiffer, R. (1961). *Applied statistical Decision Theory.* Harvard: University Press.
Raman, K. (1994). Inductive Inference and Replications: A Bayesian Perspective. *Journal of Consumer Research, 20,* 633-643.
Ramani, D. (2009). The Brain Seduction: the public Perception of Neuroscience. *Journal of Science Communication, 8(4),* 1-8.
Ramón y Cajal, S. (1899). *Textura del Sistema nervioso del Hombre y de los Vertebrados.* Madrid: Moya.
Ramsey, F.P. (1978). *Foundations.* London: Routledge.
Rappaport, J. (1977). *Community Psychology.* New York: Holt.
Raudenbush, S.W. & Bryk, A.S. (2010). *Hierarchical Linear Models: Applications and Data Analysis Methods.* 2.Aufl. London: Sage.
Rawlins, M. (2008). De Testimonio: On the Evidence for Decisions about the Use of therapeutic Interventions. *Clinical Medicine, 8(6),* 579-588.
Redi, F. (1664). *Osservazioni intorno alle Vipere.* Florenz: Stella.
Reichenbach, H. (1932). *Wahrscheinlichkeitslogik.* Berlin: de Gruyter.
Resnik, D.B. (2007). *The Price of Truth: How Money affects the Norms of Science.* Oxford: University Press.
Rey, D.G. (2012). A Review of Research and Meta-Analysis of the Seductive Detail Effect. *Educational Research Review, 7,* 216-237.
Rhodes, R.E., Rodriguez, F. & Shah, P. (2014). Explaining the alluring Influence of Neuroscience Information on scientific Reasoning. *Journal of experimental Psychology: Learning, Memory, and Cognition, 40(5),* 1432-1440.
Rhodes, R.E. (2016). What Crisis – the Reproducibility Crisis. *The Psychologist, 29,* 508-512.
Richter, S.H., Garner, J.P. & Würbel, H. (2009). Environmental Standardization: Cure or Cause of poor Reproducibility in Animal Experiments? *Nature Methods, 6(4),* 257-261.
Rips, L.J. (2001). Two Kinds of Reasoning. *Psychological Science, 12(2),* 129-134.
Rips, L.J. (2002). Circular Reasoning. *Cognitive Science, 26,* 767-795.
Roediger, H.L. (2012). Psychology's Woes and a partial Cure: The Value of Replication. *Observer, 25(9),* 27-29.
Rorty, R. (2000). *Wahrheit und Fortschritt.* Frankfurt/M.: Suhrkamp.

Rosenbaum, P.R. (2001). Replicating Effects and Biases. *The American Statistician, 55(3),* 223-227.
Rosenthal, R. (1979). The 'File Drawer Problem' and Tolerance for Null Results. *Psychological Bulletin, 86(3),* 638-641.
Rosenthal, R. & Rubin, D.B. (1982). A simple general Purpose Display of Magnitude of experimental Effect. *Journal of Educational Psychology, 74,* 166-169.
Rosenthal, R. & Rubin, D.B. (1982). Comparing Effect Sizes of independent Studies. *Psychological Bulletin, 92(2),* 500-504.
Rosenthal, R. (1990). Replication in Behavioral Research. *Journal of Social Behavior and Personality, 5(4),* 1-30.
Rosenthal, R. (1990). How are we doing in Soft Psychology. *American Psychologist, 45(6),* 775-777.
Rosnow, R.L. (1981). *Paradigms in Transition: The Methodology of Social Inquiry.* Oxford: University Press.
Rosnow, R.L. & Rosenthal, R. (1989). Statistical Procedures and the Justification of Knowledge in psychological Science. *American Psychologist, 44,* 1276-1284.
Royall, R. (1997). *Statistical Evidence: A Likelihood Paradigm.* London: Chapman & Hall.
Royall, R. (2004). The Likelihood Paradigm for statistical Evidence. In: Taper, M.L. & Lele, S.R. (Hg). *The Nature of scientific Evidence.* (S. 119-137). Chicago: University Press.
Rukhin, A.L. (1993). Estimation of the Noncentrality Parameter of an *F*-Distribution. *Journal of Statistical Planning and Inference, 35,* 201-211.
Rusconi, E., Sedgmond, J., Bolgan, S. & Chambers, C.D. (2016). Brain matters… in Social Sciences. *Neuroscience, 3(3),* 253-263.
Ryle, G. (1937). Induction and Hypothesis. *Proceedings of the Aristotelian Society, Supplement, 16,* 36-62.
Saldern, M.v. (Hg). (1986). *Mehrebenenanalyse: Beiträge zur Erfassung hierarchisch strukturierter Realität.* Weinheim: Beltz.
Samuelson, P.A. (1938). A Note on the pure Theory of Consumer's Behaviour. *Economica, 5(17),* 61-71.
Sanabria, F. & Killeen, R.P. (2007). Better Statistics for better Decisions: Rejecting Null Hypothesis Tests in favor of Replication Statistics. *Psychology in the Schools, 44(5),* 471-481.
Salmon, W.C. (1984). *Scientific Explanation and the causal Structure of the World.* Princeton: University Press.
Savage, L.J. (1972). *The Foundations of Statistics.* 2.Aufl. New York: Dover.
Schaffer, S. (1986). Scientific Discoveries and the End of natural Philosophy. *Social Studies of Science, 16,* 387-420.
Schenk, M. (2002). *Medienwirkungsforschung.* 2.Aufl. Tübingen: Mohr Siebeck.
Scherbaum, C.A. & Ferreter, J.M. (2009). Estimating statistical Power and required Sample Sizes for organizational Research using Multilevel Modeling. *Organizational Research Methods, 12(2),* 347-367.
Schervish, M.J. (1996). *p* Values: What they are and what they are not. *The American Statistician, 50(3),* 203-206.
Schickore, J. (2011). What does History matter to Philosophy of Science? The Concept of Replication and the Methodology of Experiments. *Journal of the Philosophy of History, 5,* 513-532.
Schlick, M. (1986). *Philosophische Logik.* Frankfurt/M.: Suhrkamp.

Schlosberg, H. (1951). Repeating fundamental Experiments. *American Psychologist, 6,* 177.
Schmidt, F.L. (1996). Statistical Significance Testing and cumulative Knowledge in Psychology: Implications for Training of Researchers. *Psychological Methods, 1(2),* 115-129.
Schmidt, F.L. & Hunter, J.E. (2015). *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings.* 3.Aufl. London: Sage.
Schmidt, F.L. & Oh, I.S. (2016). The Crisis of Confidence in Research Findings in Psychology: Is Lack of Replication the real Problem? *Archives of Scientific Psychology, 4,* 32-37.
Schmidt, S. (2009). Shall we really do it again? The powerful Concept of Replication is neglected in the Social Sciences. *Review of General Psychology, 13(2),* 90-100.
Schnall, S. (2014). Clean Data: Statistical Artifacts wash out. *Social Psychology, 45(4),* 315-317.
Schnall, S. (2014). Speaking out on the ‚Replication Crisis'. *The Psychologist, 28(1),* 8.
Schooler, J.W. (2014). Metascience could rescue the 'Replication Crisis'. *Nature, 515,* 9.
Schopenhauer, A. (1977). *Über die vierfache Wurzel des Satzes vom zureichenden Grunde.* Zürich: Diogenes.
Schopenhauer, A. (1977). *Die Welt als Wille und Vorstellung.* Zürich: Diogenes.
Schwarz, N. & Strack, F. (2014). Does merely Going through the same Moves make for a 'direct' Replication? *Social Psychology, 45(4),* 305-306.
Scott, S.L., Blocker, A.W., Bonassi, F.V., Chipman, H.A. et al. (2016). Bayes and Big Data: the Consensus Monte Carlo Algorithm. *International Journal of Management Science and Engineering Management, 11(2),* 78-88.
Scurich, N. & Shniderman, A. (2014). The selective Allure of neuroscientific Explanations. *Public Library of Science ONE, 9,* 1-6.
Sedlmaier, P. & Gigerenzer, G. (1989). Do Studies of statistical Power have an Effect on the Power of Studies? *Psychological Bulletin, 105(2),* 309-316.
Serlin, R.C. & Lapsley, D.K. (1985). Rationality in psychological Research: The Good-Enough Principle. *American Psychologist, 40(1),* 73-83.
Serlin, R.C. (1987). Hypothesis Testing, Theory Building, and the Philosophy of Science. *Journal of Counselling Psychology, 34(4),* 365-371.
Shafir, E. (1993). Choosing versus Rejecting: Why some Options are both better and worse than others. *Memory and Cognition, 21(4),* 546-556.
Shavit, A. & Ellison, A. (2013). There and back again: Replication Standards in Long-Term Research. *Bulletin of the Ecological Society of America, 94,* 395-397.
Shavit, A. (2016). 'Location' Incommensurability and 'Replication' Indeterminacy: Clarifying an entrenched Conflation by Using an Involved Approach. *Perspectives on Science, 34(4),* 425-442.
Shrager, J. & Langley, P. (1990). *Computational Models of scientific Discovery and Theory Formation.* San Mateo: Morgan Kaufmann.
Simmons, J.P., Nelson, L.D. & Simonsohn, U. (2011). False-positive Psychology: Undisclosed Flexibility in Data Collection and Analysis allows presenting Anything as significant. *Psychological Science, 22(11),* 1359-1366.
Simon, H.A. (1955). Prediction and Hindsight as confirmatory Evidence. *Philosophy of Science, 22(3),* 227-230.
Simon, H.A. (1977). *Models of Discovery and other Topics in the Methods of Science.* Dordrecht: Reidel.

Simon, H.A. (1992). What is an 'Explanation' of Behavior? *Psychological Science, 3(3),* 150-161.
Simons, D.J. (2014). The Value of Direct Replication. *Perspectives in Psychological Science, 9(1),* 76-80.
Simons, D.J., Alogna, V.K., Zwaan, R.A. et al. (2014). Registered Replication Report: Schooler and Engstler-Schooler. *Perspectives in Psychological Science, 9(5),* 556-578.
Simonsohn, U., Nelson, L.D. & Simmons, J.P. (2014). *p*-Curve and Effect-Size: Correcting for Publication Bias using only significant Results. *Perspectives on Psychological Science, 9(6),* 666-681.
Simonsohn, U. (2015). Small Telescopes: Detectability and the Evaluation of Replication Results. *Psychological Science, 3,* 1-11.
Simonton, D.K. (2014). Significant Samples – not Significance Tests! The often overlooked Solution to the Replication Problem. *Psychology of Aesthetics, Creativity, and the Arts, 8(1), 11-12*.
Sing Chawla, D. (2016). How many Replication Studies are enough? *Nature, 531,* 11.
Singal, J. (2016). Here's a helpful Rundown of the current State of Psychology's Replication Crisis. *New York Magazine,* 19.09.
Singh, K., Ang, S.H. & Leong, S.M. (2003). Increasing Replication for Knowledge Accumulation in Strategy Research. *Journal of Management, 29(4),* 533-549.
Skinner, B.F. (1975). The steep and thorny Way to a Science of Behavior. In: Harré, R. (Hg). *Problems of scientific Revolution.* (S. 58-71). Oxford: Clarendon Press.
Slovic, P. & Tversky, A. (1974). Who accepts Savage's Axiom? *Systems Research and Behavioral Science, 19(6),* 368-373.
Smith, N.C. (1970). Replication Studies: A neglected Aspect of psychological Research. *American Psychologist, 25(10),* 970-975.
Smith, J.K. & Smith, L.F. (2014). Replicability? Not that again! *Psychology of Aesthetics, Creativity, and the Arts, 8(1),* 21-23.
Smithson, M. (2001). Correct Confidence Intervals for various Regression Effect Sizes and Parameters: The Importance of noncentral Distributions in Computing Intervals. *Educational and Psychological Measurement, 61(4),* 605-632.
Snijders, T.A. & Bosker, R.J. (1994). Modeled Variance in 2-Level Models. *Sociological Methods and Research, 22,* 343-363.
Snijders, T.A. & Bosker, R.J. (1999). *Multilevel Analysis: An Introduction to basic and advanced Multilevel Modeling.* London: Sage.
Snijders, T.A. (2005). Power and Sample Size in Multilevel Modeling. In: Everitt, B. S. & Howell, D. C. (Hg). *Encyclopedia of Statistics in Behavioral Science.* Bd 3. (S.1570-1573). Sussex: Wiley.
Sokal, A. & Bricmont, J. (1997). *Impostures intellectuelles.* Paris: Odile Jacob.
Solla Price, D.J.d. (1963). *Little Science, Big Science.* New York: University Press.
Sonnenschein, H. (1972). Market Excess Demand Functions. *Econometrica 40,* 549-563.
Sorić, B. (1989). Statistical 'Discoveries' and Effect-Size Estimation. *Journal of the American Statistical Association, 84(406),* 608-610.
Snow, C.P. (1959). *The two Cultures and the scientific Revolution.* Cambridge: University Press.
Spellman, B.A. (2015). A short (personal) future History of Revolution 2.0. *Perspectives on Psychological Science, 10(6),* 886-899.
Sperber, D. (2010). The Guru Effect. *Review of Philosophy and Psychology, 1,* 583-592.

Spitzer, M. (2004). *Nervensachen: Perspektiven zu Geist, Gehirn und Gesellschaft.* Stuttgart: Schattauer.
Spitzer, M. (2012). *Digitale Demenz: Wie wir unsere Kinder um den Verstand bringen.* München: Droemer.
Srivastava, S. (2016). Evaluating a new Critique of the Reproducibility Project. *The Hardest Science,* Blogeintrag vom 08.03.
Staats, A.W. (1983). *Psychology's Crisis of Disunity.* New York: Praeger.
Stanley, D.J. & Spence, J.R. (2014). Expectations for Replications: Are yours realistic? *Perspectives on Psychological Science, 9(3),* 305-318.
Starch, D. & Elliott, E. C. (1912). Reliability of Grading Work in Mathematics. *The School Review, 21(4),* 254-259.
Steenbergen, H.v. & Bocanegra, B.R. (2016). Promises and Pitfalls of Web-based Experimentation in the Advance of replicable Psychological Science. *Behavior Research Methods, 48,* 1713-1717.
Steegen, S., Tuerlinckx, F., Gelman, A. & Vanpaemel, W. (2016). Increasing Transparency through Multiverse Analysis. *Perspectives on Psychological Science, 11(5),* 702-712.
Steiger, J.H. (2004). Beyond the *F*-Test: Effect Size Confidence Intervals and Tests of close Fit in the Analysis of Variance and Contrast Analysis. *Psychological Methods, 9(2),* 164-182.
Steinfeld, T. (1979). *Philosphical Problems of Statistical Inference.* Dordrecht: Reidel.
Stern, W. (1911). *Die Differentielle Psychologie in ihren methodischen Grundlagen.* Leipzig: Johann Ambrosius Barth.
Stevens, S.S. (1939). Psychology and the Science of Science. *Psychological Bulletin, 36(4),* 221-263.
Stewart-Williams, S. (2015). A quick Guide to the Replication Crisis in Psychology. *Psychology Today,* Blogeintrag vom 06.09.
Stich, S. & Shaun, N. (1998). Theory Theory to the Max. *Mind & Language, 13,* 421-449.
Stodden, V. (2009). Enabling reproducible Research: Licensing scientific Innovation. *International Journal of Communications Law & Policy, 13,* 23-46.
Stodden, V. (2009). The legal Framework for reproducible scientific Research: Licensing and Copyright. *Computing in Science and Engineering, 11,* 35-40.
Stodden, V. (2015). Reproducing statistical Results. *Annual Review of Statistics and its Applications, 2,* 1-19.
Strathern, M. (2014). Innovation or Replication? Crossing and Criss-Crossing in Social Science. *Arts and Humanities in Higher Education, 13(1/2),* 62-76.
Stroebe, W. & Strack, F. (2014). The alleged Crisis and the Illusion of exact Replication. *Perspectives on Psychological Science, 9(1),* 59-71.
Sun, S. & Pan, W. (2011). The philosophical Foundations of prescriptive Statements and statistical Inference. *Educational and Psychological Review, 23,* 207-220.
Suppes, P. (1969). *Studies in the Methodology and Foundations of Science.* Dordrecht: Reidel.
Tabacchi, M.E. & Cardaci, M. (2016). Preferential Biases for Texts that include neuroscientific Jargon. *Psychological Reports, 0(0),* 1-11.
Tabachnik, B. G. & Fidell, L. S. (2014). *Using Multivariate Statistics*. 6.Aufl. Essex: Pearson.
Tallis, R. (2011). *Aping Mankind: Neuromania, Darwinitis and the Misrepresentation of Humanity.* Durham: Acumen.

Taper, M.L. & Lele, S.R. (Hg). (2004). *The Nature of scientific Evidence.* Chicago: University Press.

Taylor, D.J. & Muller, K.E. (1995). Computing Confidence Bounds for Power and Sample Size of the General Linear Univariate Model. *The American Statistician, 49(1),* 43-47.

Thomas, L. (1997). Retrospective Power Analysis. *Conservation Biology, 11(1),* 276-280.

Thompson, B. (1994). The pivotal Role of Replication in psychological Research: Empirically Evaluating the Replicability of Sample Results. *Journal of Personality, 62(2),* 157-176.

Thompson, B. (2001). Significance, Effect Sizes, stepwise Methods, and other Issues: Strong Arguments move the Field. *Journal of Experimental Education, 70,* 80-93.

Thompson, B. (2007). Effect Sizes, Confidence Intervals, and Confidence Intervals for Effect Sizes. *Psychology in the Schools, 44(5),* 423-432.

Thompson, V.A., Turner, J.A. & Pennycook, G. (2011) Intuition, Reason, and Metacognition. *Cognitive Psychology, 63,* 107-140.

Thompson, M. (2012). Precision in Chemical Analysis: A critical Survey of Uses and Abuses. *Analytical Methods, 4,* 1598-1611.

Thornton, D.J. (2011). *Brain Culture: Neuroscience and popular Media.* Rutgers: University Press.

Tiku, M.L. (1971). Power Function of the *F*-Test under non-normal Situations. *Journal of the American Statistical Association, 66(336),* 913-916.

Titchener, E.B. (1972). *Systematic Psychology.* Cornell: University Press.

Toomela, A. (2007). Culture of Science: Strange History of the methodological Thinking in Psychology. *Integrative Psychological & Behavioral Science, 41(6),* 6-20.

Trafimow, D., MacDonald, J.A., Rice, S. & Clason, D.L. (2010). How often is p_{rep} close to the true Replication Probability? *Psychological Methods, 15(3),* 300-307.

Travis, G.D. (1981). Replicating Replication? Aspects of the social Construction of Learning in Planarian Worms. *Social Studies of Science, 11(1),* 11-32.

Tressoldi, P.E. (2012). Replication Unreliability in Psychology: elusive Phenomena or 'elusive' statistical Power? *Frontiers in Psychology, 3(218),* 1-5.

Tricket, S.B. & Trafton, J.G. (2007). „What if...“ The Use of conceptual Simulations in scientific Reasoning. *Cognitive Science, 31,* 843-875.

Trout, J.D. (2002). Scientific Explanation and the Sense of Understanding. *Philosophy of Science, 69,* 212-233.

Trout, J.D. (2007). The Psychology of scientific Explanation. *Philosophy Compass, 2(3),* 564-591.

Trout, J.D. (2008). Seduction without Cause: Uncovering explanatory Neurophilia. *Trends in Cognitve Sciences, 12(8),* 281-282.

Tsang, E.W. & Kwan, K.-M. (1999). Replication and Theory Development in Organizational Science: A critical Realist Perspective. *The Academy of Management Review, 24(4),* 759-780.

Tucker, J. (2016). Does Social Science have a Replication Crisis? *The Washington Post,* 09.03.

Tukey, J.W. (1962). The Future of Data Analysis. *Annals of Mathematical Statistics, 33,* 1-67.

Tukey, J.W. (1969). Analyzing Data: Sanctification or Detective Work. *American Psychologist, 24,* 83-91.
Tversky, A. & Kahneman, D. (1971). Belief in the Law of Small Numbers. *Psychological Bulletin, 76(2),* 105-110.
Tweney, R.D. (2004). Replication and the experimental Ethnography of Science. *Journal of Cognition and Culture, 4(3),* 731-758.
Umiltà, C. (2008). La Neuropsicologia della Coscienza. *Sistemi Intelligenti, 3,* 395-404.
Unger, P. (1975). *Ignorance.* Oxford: Clarendon Press.
Urbach, P. (1981). On the Utility of Repeating the 'same' Experiment. *Australasian Journal of Philosophy, 59(2),* 151-162.
Valentine, J.C., Biglan, A., Boruch, R.F., Castro, F.G., Collins, L.M., Flay, B.R. et al. (2011). Replication in Prevention Science. *Prevention Science, 12,* 103-117.
Valpine, P.d. (2014). The Common Sense of *p* Values. *Ecology, 95(3),* 617-621.
Vartanian, O. (2014). Toward a cumulative Psychological Science of Aesthetics, Creativity, and the Arts. *Psychology of Aesthetics, Creativity, and the Arts, 8(1),* 15-17.
Venebles, W. (1975). Calculation of Confidence Intervals for Noncentrality Parameters. *Journal of the Royal Statistical Society. Series B, 37(3),* 406-412.
Vico, G. (1959). *La Scienza nuova.* Mailand: Rizzoli.
Wagenmakers, E.J. & Forstmann, B.U. (2014). Rewarding High-Power Replication Research. *Cortex, 51,* 105-106.
Wald, A. (1971). *Statistical Decision Functions.* New York: Chelsea Publishing.
Ward, A. (2004). How one Mistake leads to another: On the Importance of Verification/Replication. *Political Analysis, 12(2),* 199-200.
Wason, P.C. (1960). On the Failure to eliminate Hypotheses in a conceptual Task. *Quarterly Journal of Experimental Psychology, 12,* 129-140.
Wason, P.C. (1977). On the Failure to eliminate Hypotheses – a second Look. In: Johnson-Laird, P.N. & Wason, P.C. (Hg). *Thinking: Readings in Cognitive Science.* (S. 307-314). Cambridge: University Press.
Weber, M. (1985). *Wissenschaft als Beruf.* In: *Gesammelte Aufsätze zur Wissenschaftslehre.* (S. 581-613). Tübingen: Mohr Siebeck.
Wedgewood, R. (2006). The normative Force of Reasoning. *Noûs, 40(4),* 660-686.
Weisberg, D.K. (2008). Caveat Lector: The Presentation of Neuroscience Information in the popular Media. *The scientific Review of mental Health Practice, 6(1),* 51-56.
Weisberg, D.K., Keil, F.C., Goodstein, J. Rawson, E. & Gray, J.R. (2008). The seductive Allure of Neuroscience Explanations. *Journal of cognitive Neuroscience, 20(3),* 470-477.
Weisberg, D.K., Taylor, J.C. & Hopkins, E.J. (2015). Deconstructing the seductive Allure of Neuroscience Explanations. *Judgment and Decision Making, 10(5),* 429-441.
Westermann, R. (2000). *Wissenschaftstheorie und Experimentalmethodik.* Göttingen: Hogrefe.
Westlake, W.J. (1976). Symmetrical Confidence Intervals for Bioequivalence Trials. *Biometrics, 32(4),* 741-744.
Whewell, W. (1967). *The Philosophy of the Inductive Sciences.* 2.Aufl. London: Frank Cass & Co.
Wilcox, R. (1995). Comparing two independent Groups via multiple Quantiles. *Journal of the Royal Statistical Society, Series D, 44(1),* 91-99.

Wilcox, R. (2012). *Introduction to robust Estimation and Hypothesis Testing.* 3.Aufl. Amsterdam: Elsevier.
Wilkinson, L. (1999). Statistical Methods in Psychology Journals. *American Psychologist, 54(8),* 594-604.
Williams, R. (2015). Can't get no Reproduction: Leading Researchers discuss the Problem of irreproducible Results. *Circulation Research, 117,* 667-670.
Wineman, L. (2013). Interesting Results: Can they be replicated? *Monitor on Psychology, 44(2),* 38.
Winkler, R.L. & Hays, W.L. (1975). *Statistics: Probability, Inference, and Decision.* 2.Aufl. New York: Holt, Rinehart & Winston.
Wittgenstein, L. (1990). *Philosophische Untersuchungen. 7.*Aufl. Frankfurt/M.: Suhrkamp.
Wolff, C. (1736). *Philosophia prima sive Ontologia.* 2.Aufl. Frankfurt/M.: Libraria Rengeriana.
Worral, J. (2010). Evidence: Philosophy of Science meets Medicine. *Journal of Evaluation in Clinical Practice, 16,* 356-362.
Wright, G.H.v. (1941). Induktionsproblemet och Kunskapens Gränser. *Finsk Tidskrift, Logik, Vetenskapsfilosofi, 2,* 204-214.
Wright, G.H.v. (1957). *The logical Problem of Induction.* 2.Aufl. Oxford: Basil Blackwell.
Wundt, W. (1898). *Grundriss der Psychologie.* 3.Aufl. Leipzig: Wilhelm Engelmann.
Wundt, W. (1914). *Reden und Aufsätze.* Leipzig: Kröner.
Xu, R. (2003) Measuring explained Variation in Linear Mixed Models. *Statistics in Medicine, 22,* 3527-3541.
Yong, E. (2012). Replication Studies: Bad Copy. *Nature, 485,* 298-300.
Yong, E. (2016). Psychology's Replication Crisis can't be wished away. *The Atlantic,* 04.03.
Youden, W.J. (1960). The Sample, the Procedure, and the Laboratory. *Analytical Chemistry, 32(13),* 23A-37A.
Youden, W.J. (1972). Enduring Values. *Technometrics, 14(1),* 1-11.
Yurevich, A.V. (2008). Cognitive Frames in Psychology: Demarcations and Ruptures. *Integrative Psychological & Behavioral Science, 43(2),* 89-103.
Zajonc, R.B. (1980). Feeling and Thinking: Preferences need no Inferences. *American Psychologist, 35,* 151-175.
Zeigler, D. (2012). Evolution and the cumulative Nature of Science. *Evolution: Education and Outreach, 5,* 585-588.
Zwaan, R. (2016). Why continue to elicit false Confessions from the Data? *Zeitgeist: Psychological Experimentation, Cognition, Language, and Academia,* Blogeintrag vom 07.03.