Zeitschrift für Anomalistik Band 4 (2004), S. 52-101

Astrologie und Psi
Eine Fallstudie verstärkt die Zusammenhangshypothese

SUITBERT ERTEL[1]

Zusammenfassung – MG, ein junger Akademiker, der die Grundlagen indischer Astrologie erlernt hatte, ließ sich auf eigenen Wunsch vom Verfasser auf seine Fähigkeit untersuchen, Geburtshoroskope richtig zu deuten. Er absolvierte zwei Zuordnungstests, bei denen Horoskope von Politikern und Malern den beiden Berufskategorien zuzuordnen waren sowie einen gleichartigen Test, bei dem Horoskope von berühmten Schriftstellern und unbedeutenden Menschen verwendet wurden. Er erreichte abwechselnd mal signifikant wenig und mal fast signifikant viele Treffer. MG wurde mittels eines neuen Testverfahrens (Ballzieh-Test) auch auf außersinnliche Wahrnehmungsfähigkeit geprüft. Aus dem Balltest-Ergebnis, in dem einige Variablen hochsignifikant aus dem Rahmen der Zufallsvariation heraus fallen, ließ sich vermuten, dass MG wahrscheinlich Psi-Fähigkeit besitzt, wenn auch nur mäßig ausgeprägt. Da zudem die signifikante Streuung der Ergebnisse des Probanden in den astrologischen Tests typischen Psi-Effekten ähnlich ist, wird die weiter zu prüfende Hypothese nahe gelegt, dass herausragende Fälle von astrologisch richtigen Deutungen, wie sie gelegentlich glaubwürdig berichtet werden, möglicherweise auf hellseherischem oder telepathischem Wege zustande kommen – eine Hypothese, die schon anderweitig aufgestellt wurde und als untersuchenswert erachtet wird.

Schlüsselbegriffe: Astrologie – Zuordnungstest – Parapsychologie – Psi

Astrology and Psi.
A case study supports the hypothesis of a connection

Abstract – MG, a university graduate, who had taken courses in Indian astrology, asked the author to test his ability in interpreting natal charts. He took two tests in which he was given horoscopes of painters and politicians, his task being to say which was which. Another similar test used horoscopes of writers and ordinary people. MG's results deviated significantly or almost significantly from chance in the wrong and the right direction respectively. Deviations from chance of various variables were also observed when MG took a test of extrasensory perception (ball drawing test). MG´s astrological results appear to share certain characteristics of typical psi effects (psi-hitting and psi-missing) suggesting that psi (telepathy, clairvoyance) might give rise to extraordinary hits purportedly associated, at least in rare cases, with chart interpretations by astrologers. Pertinent hypotheses are amenable to empirical tests.

Keywords: astrology – matching test – parapsychology – psi

[1] Prof. Dr. Suitbert Ertel ist emeritierter Professor für Psychologie an der Universität Göttingen. Anschrift: Institut für Psychologie, Gosslerstr. 14, D-37073 Göttingen. E-Mail: sertel@gwdg.de

Einleitung

Michael G., nach einem gerade bestandenen Universitäts-Diplom in Meteorologie noch nicht berufstätig, kontaktierte mich telefonisch am Göttinger Institut für Psychologie im April 2000. In seiner vielen Freizeit, die ihm als Arbeitsuchender ungewollt zufiel, widmete er sich ausgiebiger als während seines Studiums seinen Nebeninteressen, u. a. der Transzendentalen Meditation und der indischen Astrologie, die er in einem Grundkurs erlernt hatte. Er hatte erfahren, dass ich astrologische Fragestellungen mit wissenschaftlichen Methoden bearbeitete und berichtete mir, seine Deutungen von Geburtshoroskopen würden oft erstaunlich treffend sein. Er bat mich um Überprüfung seiner Fähigkeit, Horoskope richtig zu deuten. Drei astrologische Tests führte ich mit ihm durch. Während eines Besuchs in Göttingen erfuhr MG auch von meiner parapsychologischen Forschungsarbeit, woraufhin er sich zusätzlich einem Test auf außersinnliche Wahrnehmung (Ballzieh-Test) unterzog. Die Ergebnisse der beiden Testserien mit MG legen die Hypothese eines Zusammenhangs zwischen Astrologie und Psi nahe, die zur Diskussion gestellt wird und die man prüfen sollte.[2] Sie wurde schon von Timm & Köberl (1986) zur Deutung positiver Astrologie-Effekte aufgestellt, blieb aber unbeachtet. Die folgende Darstellung stützt sich u.a. auch auf einen schriftlichen Bericht, den MG im Rückblick auf seine drei Astrologietests verfasste. Sie wurde von MG freundlicherweise in den Teilen, die sein Verhalten und seine Sicht darstellen, auf Korrektheit überprüft, seine Korrektur- und Ergänzungsvorschläge wurden berücksichtigt.[3]

Drei Astrologische Zuordnungstests

Erster Test: Politiker und Maler I

Methode

Geburtsdaten von 20 Politikern und 20 Malern (Tag, Zeit, Geburtsorte) wurden einem von Colin Miles stammenden Datensatz eminenter schottischer Persönlichkeiten entnommen. Die Daten der beiden Berufsrepräsentanten waren chronologisch über den gewählten Zeitraum ähnlich verteilt. Sie wurden in eine Zufallsfolge gebracht und MG zugesandt mit der Aufgabe, den Geburtsdaten die richtigen Berufsbezeichnungen zuzuordnen. Dasselbe Verfahren und Material war schon früher für ein „Astro-Quiz" im Internet verwendet worden (Ertel 1998).[4] Die mittlere Zufallserwartung für richtige Zuordnungen beträgt 20 Treffer.

---

[2] Der Terminus Psi soll auf hypothetische Faktoren verweisen, welche z.B. in einem Psi-Rate-Test überzufällige Trefferhäufigkeiten hervorrufen, für deren Vorkommen die derzeitige Naturwissenschaft keine allgemein anerkannten Erklärungsbegriffe zur Verfügung stellt. Der Ausdruck „hervorrufen" impliziert hier nicht notwendig einen kausalen Zusammenhang.

[3] Auch habe ich für ausführliche, kritische und produktive Anmerkungen der drei Gutachter meines Manuskripts zu danken.

[4] Das verwendete Material ist bei Ertel (1998) im Anhang abgedruckt, was MG erst nachträglich mitgeteilt wurde, dem offensichtlich der Artikel sowie auch die Zeitschrift, in der er erschienen war,

Nachdem MG die Horoskope errechnet und inspiziert hatte, erachtete er 16 der 40 Fälle für eine Berufszuordnung als zu unsicher (z.B. erschienen ihm die Aszendenten uneindeutig oder die Deutungsmöglichkeiten widersprüchlich etc.). Mit meiner Zustimmung beschränkte er die Berufszuordnungen auf 24 Fälle, die Treffer-Zufallserwartung wurde entsprechend auf 12 herabgesetzt.

Ergebnisse

Nur 6 der 24 Zuordnungen, die MG mir zusandte, waren richtig, die Abweichung vom Zufall hatte die falsche Richtung. Nach dem Signifikanztest von Rosenthal & Rubin (1989), den ich zuerst anwandte, erschien die Abweichung sehr signifikant: PI = 0.25 (PI = Proportion Index), Z = -2.828, p = .0024, einseitig).[5] Auch nach einem Binomialtest, der hier vorzuziehen ist, war die Abweichung vom Zufall sehr signifikant (p = .0113, einseitig).[6]

Diskussion

MG´s Kommentar war: „*Dieses Ergebnis überraschte mich außerordentlich, denn ich meinte, mir viel Mühe bei der Analyse gegeben zu haben. Ich sah mir die Horoskope sofort noch einmal an … verglich mit dem, was ich im Kurs gelernt hatte, überprüfte noch einmal, ob die Aszendenten richtig ermittelt worden waren, kam aber zu keinem anderen Ergebnis.*“ MG war also verunsichert. Die signifikante Trefferabweichung mit falscher Richtung schien ihm nur Sinn zu machen, wenn er davon ausgehen könne, dass sich hervorragende Berufserfolge nicht vorzugsweise bei Personen einstel-

---

weder bekannt noch zugänglich waren. Die 11 am Test von 1998 teilnehmenden Astrologen waren bei ihren Zuordnungen „not better than chance“ (Ertel 1998, S. 3).

[5] Nach dem zweiten der drei Kriterien von Kimmel (1957) zur Anwendung einseitiger statistischer Prüfungen ist hier eine einseitige Prüfung angezeigt (“Use the one-tailed test when results in the unpredicted direction will, under no conditions, be used to determine a course of behavior different in any way from that determined by no difference at all“, p. 353). Denn die signifikante Abweichung von MG’s Horoskopdeutungen in falscher Richtung kann „unter keinen Umständen“ (ebenso wenig wie ein insignifikanter Unterschied) seine Horoskop-Deutungsfähigkeit bestätigen. So brachte das Testergebnis MG auch dazu, auf eine eventuelle zukünftige Horoskop-Deutungspraxis zu verzichten. Mit einseitigem p-Wert testeten auch z.B. Timm & Köberl (1986) die Deutungsfähigkeit ihrer Astrologen.

[6] Die Anwendung des Rosenthal & Rubin-Signifikanz-Tests (R&R) kann aus zwei Gründen kritisiert werden: (1) Die untere Grenze für N trials, von R&R mit 25 festgelegt, wird hier um 1 unterschritten, das N der vorliegenden Untersuchung beträgt 24. (2) Bedeutsamer ist, dass meine nachträgliche Monte-Carlo-Überprüfung des R&R-Verfahrens, veranlasst durch Vermutungen eines Gutachters des Manuskripts, überraschenderweise zeigte, dass das R&R-Verfahren fast durchweg signifikantere p-Werte liefert als das Binomialverfahren. Ich habe daraufhin einen der beiden Autoren des R&R-Verfahrens gebeten, die Diskrepanz zu klären. Das R&R-Verfahren sollte bis zur Klärung, wenn überhaupt, mit Vorsicht verwendet und zumindest durch nicht-kontroverse Verfahren ergänzt werden. Indessen ist die statistische Evidenz der Gesamtergebnisse der vorliegenden Untersuchung unabhängig von der Klärung der noch offenen Frage zum R&R-Verfahren, da nicht die Treffer individueller Tests, sondern die Heterogenität der Testergebnisse insgesamt weiter unten zum Thema werden.

len, die von ihrer Anlage her mit entsprechender Berufseignung hinreichend ausgestattet sind (diese wären astrologisch richtig zu diagnostizieren), sondern bei denen, die aufgrund eines gewissen Mangels an Berufstalent zu besonderen Anstrengungen herausgefordert werden, was sich dem Geburtshoroskop nicht so leicht entnehmen ließe (analog der These Alfred Adlers von der „Überkompensation" bei Minderwertigkeitskomplexen). Bei dieser überraschenden Befundlage erschien dem Probanden und dem Versuchsleiter ein Ergänzungstest wünschenswert.

Zweiter Test: Politiker und Maler II

Methode

Der zweite Test war mit dem ersten nahezu methodisch identisch, nur wurden Geburtsdaten von Politikern und Malern französischer Nationalität verwendet, die dem Gauquelin-Datensatz entnommen wurden (Gauquelin & Gauquelin 1970). Diesmal schied MG 21 Horoskope von den 40 vorgegebenen aus, die ihm zu unsicher erschienen. Für die verbleibenden 19 Fälle beträgt die mittlere Zufallserwartung für richtige Berufszuordnungen 9,5. Erwartet wurde eine Replikation der ersten Beobachtung, denn sollte beim ersten Test ein Effekt vorgelegen haben, dann ist hier wie bei jeder anderen Effektforschung davon auszugehen, dass er sich unter gleichen Bedingungen wiederholt.[7]

Ergebnisse

MG gelangen diesmal vorwiegend richtige Berufszuordnungen, die Trefferzahl (13 richtige von 19 Zuordnungen) ist nach dem Binomialtest marginal signifikant ($p = .08$, einseitig).[8] Bemerkenswerter aber als das Einzelergebnis für sich genommen erscheint mir der Kontrast zwischen den Trefferquoten der ersten und zweiten Untersuchung, der post hoc, nach dem Chi^2-Test geprüft, auffällig ist ($Chi^2 = 8.107$, $df=1$, $p = .004$). Dieser Sprung im Ergebnis weckt den Verdacht, dass hier ein Effekt vorliegt, allerdings ein zunächst verwirrender.[9]

---

[7] Zur Gerichtetheit der Hypothese im vorliegenden Fall: Die Hypothese eines *konstanten* Effekts wird aufgestellt, weil unter gleichen Bedingungen Ursache-Wirkungs-Zusammenhänge in der Regel reproduziert werden. Kimmels Kriterium ist erfüllt, weil die Hypothese eines konstanten Effekts bei signifikant *positiver* Trefferabweichung (also in nicht erwarteter Richtung) ebenso geschwächt würde wie bei einem Zufallsergebnis. Nur aus einer post-hoc-Perspektive lässt sich argumentieren, dass man von vornherein neben einem Wiederauftreten der negativen Trefferabweichung unter der gleichen Bedingung alternativ auch eine positive Abweichung hätte erwarten sollen, so dass hier zweiseitig zu testen wäre. Doch dabei wird fälschlicherweise vorausgesetzt, dass der Experimentator schon vor Beginn des zweiten Astrologietests von Erkenntnissen hätte Gebrauch machen können, die er tatsächlich erst im weiteren Verlauf der Versuchsserie gewonnen hat.

[8] Der R&R-Test kann hier schon wegen deutlicher Unterschreitung des minimalen Trial-N nicht verwendet werden.

[9] Verwendete Daten: Treffer (T) und Fehler (F) für Test 1: 6 T, 18 F; für Test 2: 13 T, 6 F.

Diskussion

Mit diesem Ergebnis war MG wiederum unzufrieden: "…*was nützt eine Methode, die einmal funktioniert, aber ein anderes Mal gegensätzliche Resultate liefert?"* Vielleicht aber könne man die Befundlage mit einer abgewandelten Hypothese vereinbaren. Es könne sein, so meinte MG, dass ein Geburtshoroskop zu wenig spezifische Indikatoren für den beruflichen Lebenslauf des Horoskopeigners enthalte. Vielleicht enthalte es nur Indikatoren für das Ausmaß des späteren Lebenserfolges, nicht für die besondere berufliche Richtung, die zum Erfolg führt. MG bat mich deshalb, ihm Geburtsdaten einer Stichprobe sehr berühmter Persönlichkeiten zusammen mit Daten unbedeutender Personen zu schicken, noch einmal wollte er mittels seiner Astrologie eine Zuordnung versuchen.

Dritter Test: Berühmte Schriftsteller und nicht-berühmte Personen

Methode

Verwendet wurden die Geburtsdaten von 20 berühmten französischen Schriftstellern aus Gauquelins Professions-Datenpool und Daten von 20 „ordinary people" aus Gauquelins heredity-Datenpool (Gauquelin & Gauquelin 1970). Die Geburtsdaten streuten wieder ungefähr gleich über den betreffenden Zeitraum. Aus dieser Stichprobe schied MG 17 ihm unsicher erscheinende Fälle aus, 23 verblieben zur Zuordnung. Da im ersten Test eine negative, im zweiten eine positive Trefferzahl erreicht wurde, hatte die Erwartung für den dritten Test keine bestimmte Richtung.

Ergebnisse

Von den 23 verbleibenden Geburtshoroskopen ordnete MG nur 12 richtig zu (p: n.s., Binomialtest).

Diskussion

MG´s Trefferquote beim dritten Test, die nicht vom Zufall abweicht, entspricht dem, was kontrollierte astrologische Zuordnungstests in der Regel ergeben (vgl. Müller & Ertel 1992; Ertel 1998; siehe aber auch Timm & Köberl 1986). MG war enttäuscht, aber lernbereit. Er schreibt: "*[Doch hat das Experiment] mir sehr wichtige Erkenntnisse gebracht. Die Frage, ob ich mich eines Tages auch Astrologe nennen möchte, die ich noch von diesem Experiment abhängig gemacht habe, habe ich nun – vermutlich endgültig – verneint. Ich glaube nicht mehr daran, dass man den Lebensweg eines Menschen auch nur in groben Zügen aus dem Horoskop ablesen kann*". Auch glaube er nicht, dass andere Astrologen dazu fähig wären. An weiteren Tests war MG nicht interessiert. Er hatte nur vorgehabt, seine astrologischen Deutungen, die ihm selbst ziemlich treffsicher erschienen, mittels objektiver Methoden auf Haltbarkeit überprüfen zu lassen. Das Ziel war erreicht.

Das Testergebnis war für MG indessen kein Anlass, die Möglichkeit von Zusammenhängen zwischen Menschen und Planeten überhaupt auszuschließen. *„Ich bin davon überzeugt, dass es [solche] Wechselwirkungen … gibt, so wie es aus den Untersuchungen von Gauquelin und seiner Nachfolger*

hervorgeht". Doch diese seien selbst für große Stichproben von Personen so geringfügig, dass aus ihnen keine Schlussfolgerungen für den Einzelfall gezogen werden dürften.[10] Sein Fazit fasst er so zusammen: *„Die bisherigen Untersuchungen haben … gezeigt, dass beide Extrempositionen gegenüber der Astrologie (alles „Aberglaube" oder „alles Wahrheit") unzutreffend sind. Eine differenziertere Betrachtung des Themas Astrologie ist daher erforderlich."*

MG´s Schlussfolgerungen erschienen mir konsequent und zudem generell vorbildhaft zu sein für Astrologen, denen man die Fähigkeit zusprechen möchte, die Freiheit ihres Sternenglaubens, der sich durch keine soziale oder institutionelle Autorität begrenzt sehen muss, durch die weniger anfechtbare Autorität der menschlichen Vernunft begrenzt zu sehen.

Vernunft und Vorsicht ist indessen auch aufseiten des Wissenschaftlers geboten, der angesichts der drei astrologischen Testergebnisse nicht schon zur Tagesordnung übergehen kann. Denn die sprunghafte Verteilung von richtigen und falschen Zuordnungen über die drei Tests insgesamt (vgl. Tabelle 1) ist signifikant und also zu beachten ($Chi^2 = 8.41$, $df=2$, $p = .015$).[11]

Tabelle 1: Zusammenfassung der Astro-Test-Ergebnisse mit MG. Treffer- und Fehlerhäufigkeiten, P- und Z-Werte.

	Treffer	Fehler	Summen	P (Binom.)	Z
Test 1: Schottische PO und MA	6	18	24	.011	2.29
Test 2: Französische PO und MA	13	6	19	.084	1.38
Test 3: Französische SCHR und GL	12	11	23	.500	0.00
Summen	31	35	66		

Abkürzungen: PO = Politiker, MA = Maler, SCHR= Schriftsteller, GL = gewöhnliche Personen

Bei Fortsetzung der Testserie könnte sich die Treffer-Streuung zwar auf den Zufall einpendeln, aber das darf man nicht für sicher halten. Das paradoxe Schwanken zwischen extremen Trefferquoten in den aufeinander folgenden Tests könnte sich als ein realer Effekt erwei-

[10] Zur Frage nach der Realität des Gauquelin-Planeteneffekts vgl. Ertel (1988) und Ertel & Irving (1996). Die Autoren referieren auch ausführlich die einschlägigen Arbeitsergebnisse der belgischen, französischen, holländischen und amerikanischen Skeptikerorganisationen.

[11] Ein Gutachter des vorliegenden Artikels wollte hier das Summe-Z^2-Verfahren angewandt sehen, ein Signifikanztest, bei dem die Z-Werte der Trefferabweichungen aus mehreren unabhängigen Tests zusammengefasst und in einen Chi^2-Wert überführt werden (beschrieben von Timm 1983), wobei $Chi^2 = \Sigma(Z^2)$. Doch entspricht eine Anwendung dieses Verfahren, welches die Zufallsabweichung der Trefferhäufigkeit unabhängig von ihrer Richtung auf Signifikanz prüft, nicht dem vorliegenden Prüfungsziel. Hier steht die Frage an, ob die beobachtete *Schwankung* der Treffer von einem Versuch zum anderen signifikant vom Zufall abweicht (Frage nach der Heterogenität der Trefferhäufigkeiten). Unter den von Timm (1983) beschriebenen Formeln zur Kombination von unabhängigen Z-Werten halte ich für den vorliegenden Prüffall allenfalls $Chi^2 = (-2) \times \Sigma (\log(p))$ als Alternative für geeignet. Dieser Test ergibt ein vergleichbares $p = .018$ ($Chi^2 = 15.37$, $df = 6$).

sen.[12] Zwar ließe sich dieser mit Begriffen des wissenschaftlichen Mainstreams kaum angehen. Doch würde hier möglicherweise die Parapsychologie weiterhelfen können, jene Wissenschaftsdisziplin, die mit paradoxen Phänomenen ähnlicher Art seit langem vertraut ist.
Aus der parapsychologischen Forschung ist das Phänomen des „psi-missing" bekannt (Rhine 1952). Unter Psi-Missing versteht man Psi-Effekte in einer vom Probanden nicht gewünschten Richtung, z.B. überzufällig wenige Treffer, obgleich er viele Treffer erzielen wollte. (Die Bezeichnung ist irreführend, weil man ohne Erläuterung meint, „Psi-Missing" bezeichne das Nicht-Auftreten eines Psi-Effekts). Proband MG wollte in den Astrologietests möglichst viele Treffer erzielen, doch lag beim ersten Test die Trefferquote weit unter der Zufallserwartung, was einem Psi-Missing-Effekt ähnlich ist. War es Psi-Missing?
Dies ist nicht unwahrscheinlich. Denn im zweiten Test kam es zu einem signifikanten Umschlag der Abweichungsrichtung, d.h. zu einem erwünschten Treffer-Überhang. Sprunghafte Veränderungen in Psi-Test-Ergebnissen von einer Abweichungsrichtung in die andere, meist von „Psi-Hitting" zu „Psi-Missing", kamen bei den ca. 10% Psi-Begabten unter den ca. 300 insgesamt getesteten Probanden meiner bisherigen *parapsychologischen* Versuche (mehr dazu weiter unten) nicht selten vor.
Ich erlaube mir, diesen spekulativen Gedanken auszubauen und es für möglich zu halten, dass der Anteil psi-begabter Personen unter eminenten Astrologen vielleicht generell überdurchschnittlich ist. Vielleicht trägt telepathisch-hellseherische Fähigkeit mit dazu bei, dass man von astrologischen Lehren angezogen wird, sie gern erlernt und praktisch anwendet. Horoskop-Deutungen psi-begabter Astrologen könnten somit durch Psi mitgeprägt werden und gelegentlich zu hervorragenden Trefferquoten (Psi-Hitting) führen. Unter Psi-Einfluss wären dann allerdings in Ausnahmefällen auch gravierende Fehldeutungen (Psi-Missing) zu erwarten.
Die Spekulation geht weiter: Überdurchschnittliche Treffer- und Fehlertendenzen bei Horoskop-Deutungen psi-begabter Astrologen könnten mehr oder weniger kurzfristig wechseln. Bei kurzfristigem Wechsel zwischen Psi-Hitting und Psi-Missing hätte man mit Standard-Überprüfungen von der Art, wie ich sie mit MG durchführte, als Durchschnitt individueller Zeitreihen eher Zufalls-Trefferquoten zu erwarten. Eine Einebnung der Psi-Effekte würde auch bei aggregierten Daten vorkommen, die von größeren Stichproben von Horoskopdeutern stammen, Psi-Effekte einiger Teilnehmer in negativer Richtung würden die positiven Ausschläge anderer Teilnehmer insgesamt vermindern oder ganz zum Verschwinden bringen.
Die vorliegende Untersuchung galt dem Einzelfall MG, dessen Ergebnisse wurden nicht mit denen anderer Teilnehmer vermischt. Auch wirkte sich günstig aus, dass sich die Untersuchung über einen längeren Zeitraum (sieben Monate) hinzog, was dem offenbar langsame-

[12] Vielleicht sollte man bei Fortsetzung der astrologischen Zuordnungsversuche mit MG, falls es dazu käme, oder bei Versuchen mit anderen Astrologen auf Horoskope gewöhnlicher Menschen lieber verzichten, wenn möglich, denn das Zufallsergebnis des dritten Tests bei MG könnte durch die „Nicht-Eminenz" dieser Personengruppe mit bedingt sein, was analog wäre den negativen Befunden, die Gauquelin und andere mit nicht-eminenten Personen erhielten.

ren Oszillieren von Fehler- und Treffertendenzen bei MG genügend Möglichkeiten zur Entfaltung bot.

Die Hypothese eines Zusammenhangs zwischen Horoskopdeutung und Psi wäre durch Einzelfall-Untersuchungen an namhaften Astrologen weiter zu prüfen. Leider sind bisher fast immer nur Gruppen von Astrologen untersucht worden. Als Ausnahme bemerkenswert ist van Rossems (1933) Überprüfung der Horoskop-Deutungen von Leo Knegt, eines holländischen Astrologen von zu seiner Zeit hohem Rang. Van Rossems Methode war eine qualitative. Knegt hatte im Blindversuch Geburtshoroskope von zehn ausgeprägten Persönlichkeiten, die der Versuchsleiter ausgewählt hatte, charakterologisch und biographisch nach vorgegebenen kategorialen Richtungen auszudeuten.

Die Trefferleistung Knegts erscheint unvoreingenommenen Lesern des van Rossem-Berichts erstaunlich treffend [13], so dass selbst ein entschiedener Kritiker der Astrologie wie Rudolf Smit beeindruckt wurde, der über die van Rossem-Ergebnisse ausführlich berichtet (Smit 1997). Für Smit ist Leo Knegt eine „white crow beyond our wildest dreams".[14] Zwar lässt Smit die Frage unerörtert, ob es sich bei Knegt möglicherweise um eine Psi-Begabung gehandelt hatte. Doch hält er dies für möglich (persönliche Mitteilung auf Befragung). Auch weist schon van Rossem bei seinem Versuch einer Erklärung der außergewöhnlichen astrologischen Leistung Knegts vorsichtig in die Richtung des „Okkulten".

Dies ist der Hintergrund, vor dem die folgende Untersuchung verstanden werden sollte. Sie galt der Frage, ob MG, der mit seinen astrologischen Deutungen eine psi-ähnliche Inkonstanz der Abweichungen vom Zufall zeigte, möglicherweise Psi-Begabung besitzt.

Der parapsychologische Balltest: Das Zahlen-Ziehen

Vorbemerkung

Der Balltest, ein Novum in der parapsychologischen Methodenlandschaft, wurde von mir 1999 auf einer Tagung der *Society for Psychical Research* in Durham (UK) und 2000 in Freiburg auf einer Tagung der *Parapsychological Association* vorgestellt (Ertel 2000), vgl. Ertel (2003). Die Tätigkeit der Probanden in diesem Test ist eine motorisch-praktische und alltagsnäher als die bloße Kopfarbeit und Tastendrück-Tätigkeit, die den Probanden der konventionelleren Psi-Testverfahren abverlangt wird. Diesem Umstand vermutlich verdankt der Balltest höhere Effektstärken (zumindest bei etwa 10% unausgelesener Teilnehmer), welche die in der Literatur bisher berichteten Effektstärken oft um ein Vielfaches übertreffen.

---

[13] Beispiel: Knegts Blinddeutung des Horoskops von Person 6: „Perhaps best job for her is in the travel industry…, she will succeed best in the hotel world or in some position on a passenger ship." Die Dame hatte tatsächlich "the position of stewardess on a passenger ship" (Smit 1997, S. 8).

[14] Rudolf Smit wollte mittels eines Zuordnungstests erfahren, ob sich ein astrologisches Treffer-Phänomen wie das von Knegt wiederholt. Seine Ergebnisse waren negativ (Smit 1998). Doch war sein Versuch methodisch weniger günstig angelegt (was Smit in einer persönlichen Mitteilung bestätigte). Smits *Zuordnungs*methodik weicht von der *Gutachten*methodik van Rossems erheblich ab. Zum Unterschied der beiden Untersuchungstypen siehe Timm & Köberl (1986).

Von den zwei Standard-Grundformen des Balltests (Einbeutel- und Zweibeutel-Technik) wurde die Zweibeutel-Technik gewählt. Obgleich für die Einbeutel-Technik mehr Vergleichsdaten vorliegen (N = 360 vs. N = 28), erschien die Zweibeutel-Technik für den vorliegenden Einzelfall geeigneter, weil ihre größere Bedingungskomplexität den hypothetischen Psi-Dispositionen, die versteckt und gehemmt sein können, eine messbare Auswirkung auf alternativen Kanälen ermöglicht. Vom Zufall abweichende Häufigkeiten der Treffer, die die Probanden bewusst registrieren und auf die konventionell forschende Parapsychologen meist allein ihr Augenmerk richten, sind nicht die einzigen Psi-Indikatoren. Zum genaueren Verständnis des Ball-Tests und zur Akzeptanz seiner diagnostischen Leistung verweise ich auf Ertel (2003).

Methode

Dem Probanden werden zwei nicht-transparente Beutel ausgehändigt, in jedem befinden sich 50 Tischtennisbälle, auf denen ähnlich wie auf Lotto-Bällen Zahlen geschrieben stehen, hier jedoch nur die Zahlen von 1 bis 5, jede Zahl ist auf 10 Bällen vertreten. Der Proband hat die Bälle durch Wenden der beiden geschlossenen Beutel (dies wie beim Umlegen von Pfannekuchen) durcheinander zu bringen, dann in die Beutel zu greifen und blind aus ihnen je einen Ball zu ziehen. Die beiden Zahlen, die dann gezogen werden, werden in ein vorbereitetes Protokollblatt eingetragen. Anschließend werden die Bälle in die Beutel zurück gelegt, worauf das nächste Trial folgt.
Der Proband soll bei jedem Zug eine bestimmte Zahl zwischen 1 und 5 ziehen, die ihm auf dem Protokollblatt vorgegeben wird. Die Vorgabe der Zielzahl erfolgt indirekt, z.B. wird vorgegeben: „3+2“ (der Proband soll die 5 ziehen), oder „4-1“ (der Proband soll die 3 ziehen. Die Aufgaben wechseln („2+1“, „5-2“ usw.). Immer soll das Ergebnis der Additions- bzw. Subtraktionsaufgabe gezogen werden. Aufgaben von der Art „2+2“ (Wiederholung von Zahlen innerhalb der Aufgabe) und „4-2“ (die Ergebniszahl 2 wäre mit einer Aufgabenzahl identisch) kommen nicht vor.
Als *wünschbare* Ziehergebnisse werden dem Probanden vier Möglichkeiten vorgestellt, nach Güte absteigend gestuft. Sie seien hier verdeutlicht mit der Beispielaufgabe „**2+3“.** Wird „2+3“ als Aufgabe vorgegeben und werden dann gezogen …

- links und rechts die **5**, also die Ergebniszahl der Aufgabe „2+3“aus beiden Beuteln, dann ist das ein *Trefferpasch* (= *Pasch 1*), das ist das begehrteste Ergebnis;
- links die **5** und rechts z.B. die **4** oder umgekehrt links z.B. die **4** und rechts die **5**, m.a.W. die Ergebniszahl und eine beliebige andere Zahl, dann ist das ein *Eintreffer*-Fall;
- links und rechts z.B. die **1** aus beiden Beuteln, also die gleiche Zahl, aber nicht die Ergebniszahl **5**, dann ist das ein „*Pasch 2*“;
- links die **2** und rechts die **3** oder links die **3** und rechts die **2**, also die beiden Aufgabenzahlen, dann ist das eine „*Repetition*“ (der beiden vorgegebenen Aufgabenzahlen). Das Ziehen nur einer Aufgabenzahl wird nicht auch als wünschbar mit aufgeführt.

Der Proband kodiert die Güte seines Ballzieh-Erfolges auf seinem Protokollblatt sofort nach jedem erfolgreichen Zug. Ohne wünschbares Ziehergebnis der Gütestufen 1 bis 4 erfolgt

nur der Zahleneintrag, das völlige Verpassen eines der vier Erfolgsmöglichkeiten schlägt sich also nicht in einer Kodierung nieder.
MG absolvierte eine zuvor festgelegte Serie von 16 Runs mit je 60 Ziehungen. Da MG mit der TM-Meditationstechnik vertraut war, wurde er gebeten, die Hälfte der Runs nach vorheriger Meditation (M+), die andere Hälfte ohne Meditation (M–) durchzuführen, diese Bedingung sollte er nach jedem zweiten Run wechseln (M+, M+, M–, M–, M+, M+ usw.).[15] Erwartet wurde eine höhere Trefferzahl nach M+. Die Versuche fanden bei MG zuhause ohne Anwesenheit anderer Personen statt.[16]

Ergebnisse

Zunächst eine kurze Vorbemerkung zu den Signifikanztest-Ergebnissen: Die Vielzahl der Variablen (rund 20), die bei Zweibeutel-Daten berücksichtigt werden, erfordert eine Korrektur, die nach Bonferroni vorgenommen wurde: alle originär signifikanten p-Werte wurden mit 20 multipliziert (angenäherte einfache Berechnung, Notation p_k.). Auf p-Korrekturen darf verzichtet werden bei Hypothesen, die in der vorliegenden Untersuchung eingeführt wurden und zur Entscheidung anstanden (z.B. beim Test auf stärkere Abweichungseffekte bei der M+ Bedingung im Vergleich mit der M– Bedingung) und bei Replikationen signifikanter Effekte innerhalb der Testdaten, sofern diese voneinander unabhängig sind. Letzteres ist einem split-half-Korrelationstest analog, bei dem man nach Vorliegen eines signifikanten Effekts bei der einen Testhälfte die andere gezielt auf Effektkonstanz prüft.

Pasch 1 (Treffer-Paschs)

Die Trefferpaschs bei M+ (22) und M– (21) weichen vom Erwartungswert 19.2 nicht signifikant ab.

Einzeltreffer

Auch liegen die Einzeltreffersummen 159 (M+) und 146 (M–) in der Nähe der Zufallserwartung (153.6).

---

[15] Psi-Effekt-steigernde Auswirkungen verschiedener Meditationspraktiken sind vielfach berichtet worden (z.B. eindrucksvoll bei Rao & Rao 1982). Ein Übersichtskapitel über „Experimental studies of psi and meditation" von Honorton (1977, S. 442) schließt mit der Aussage: „The combined results for all of the studies involving psi tasks during or following meditation are highly significant ($P=6 \times 10^{-12}$).

[16] An dieser Stelle kann nicht auch noch die Frage behandelt werden, ob die Probanden bei Psi-Versuchen ohne Labor-Kontrolle *primäre* Psi-Effekte (Treffer, Paschs etc.) vielleicht auf betrügerischem Wege ins Protokollblatt einbringen. Bei einigen Probanden mit hohen Heim-Trefferquoten habe ich selbst Laborversuche unter meiner Kontrolle durchgeführt. Unter den Labor-Ergebnissen kommen z.T. so starke Abweichungen vom Zufall vor, dass man zur Evidenz nicht einmal Signifikanztests benötigt (Ertel 2003). Auch konnte gezeigt werden, dass bei Heimtest-Daten *sekundäre* Psi-Indikatoren zu finden sind, die sich einer Einwirkung durch Betrug entziehen, so dass an der Echtheit auch der er primären Psi-Effekte kaum begründet zu zweifeln ist. Bei MG, der im Balltest keine *primären* Psi-Effekte zeigte, wurde diese Frage ohnehin nicht akut.

Pasch 2 (Nicht-Treffer-Paschs)

Auch liegen die Nicht-Treffer-Paschs weder bei M+ (80) noch bei M– (81) signifikant über dem Erwartungswert 76.8.

Repetitionen

Auch liegen die Häufigkeiten beim Ziehen der beiden Aufgabenzahlen im Zufallsbereich.

Sondereffekt mit den Aufgabenzahlen

Bei den Häufigkeiten gezogener Aufgabenzahlen zeigt sich eine bemerkenswert große Abweichung vom Zufall, die mit der in der Instruktion mitgeteilten Wünschbarkeitsliste nichts zu tun hat.[17] Sie ist so zu verdeutlichen: Zieht der Proband die Bälle z. B. bei der Aufgabe „2+3", dann haben die 2 (die erste Aufgabenzahl) und die 3 (die zweite Aufgabenzahl) die gleiche Chance, gezogen zu werden, das gilt für den linken Beutel so wie für den rechten Beutel.
Berücksichtigt man nun alle Trials, bei denen links oder rechts oder aus beiden Beuteln Aufgabenzahlen gezogen wurden – MG zog unter M+ und M– insgesamt 455 mal Aufgabenzahlen – dann sollten die Ziehungshäufigkeiten der ersten und der zweiten Aufgabenzahl vom Erwartungswert (455/2 = 227.5) nicht weit abweichen. Tatsächlich aber liegt eine große Abweichung vor, MG zog die zweite Aufgabenzahl viel häufiger (264 mal) als die erste Aufgabenzahl (191 mal). Der Unterschied ist hochsignifikant: Chi^2 (goodness of fit) = 11.7, df=1, p=.0006, p_k=.012). Der Überhang der gezogenen zweiten Aufgabenzahl gegenüber dem Erwartungswert beträgt (264 – 227.5)/227.5 = 16 %.
Der Überhang der zweiten Aufgabenzahl kommt unter der M+ Bedingung (129 vs. 99, p = .047) wie unter der M– Bedingung vor (135 vs. 92, p = .0043).
Legt man die Verteilung der Abweichungen zugrunde, die bei den 27 Probanden vorkamen, die den Zweibeuteltest bisher absolvierten (a.M. der 27 absoluten Abweichungswerte = .041, sd = .034), dann ergibt sich für MG, dem 28. Probanden, mit seiner Abweichung von .16, ein Z-Wert von 3.46, der ebenfalls sehr signifikant ist (p = .0003).[18] MG weicht mit seiner

[17] Die hier darzustellende Auffälligkeit hat nichts mehr mit den marginal erwünschten *Repetitionen* zu tun, bei denen gezogene Aufgabenzahlen nur interessierten, wenn sie *beide* beim gleichen Trial links und rechts gezogen wurden.

[18] Ein Proband, ein Ex-Astrologe, zeigte im Zweibeutel-Test eine Präferenz für das Ziehen der *ersten* Aufgabenzahl. Dabei kam noch eine Wechselwirkung hinzu (p_k = .02), d.h. er präferierte die *ersten* Aufgabenzahlen vor allem beim Ballziehen aus dem *linken* Beutel (starke Präferenz), die *zweiten* Aufgabenzahlen präferierte er beim *rechten* Beutel (schwächere Präferenz). Da von drei Astrologen bzw. Ex-Astrologen, die bisher mit dem Balltest getestet wurden, alle drei signifikante Abweichungen bei den sekundären Variablen zeigten, sehe ich die vorliegende Argumentation gestützt. Die Frage, ob solche Abweichungen vom Zufall einen Sinn machen, hat kaum Priorität gegenüber der Aufgabe, zunächst psi-bedingten Ordnungstendenzen nachzugehen, wo immer sie sich zeigen, um sie bei hinreichender Stärke als Basis für die weitere Forschung zu betrachten, von deren Ergebnissen am Ende vielleicht auch auf die Sinnfrage ein Licht fallen wird.

Bevorzugung der zweiten Aufgabenzahl also auch hochsignifikant von derjenigen Erwartung ab, die sich durch die Testergebnisse der übrigen Probanden definiert.

Meditationsbedingung

Tabelle 2 gibt die Treffersummen der acht Runs mit vorbereitender Meditation (M+) und ohne Meditation (M–) wieder. In der M+ Bedingung erzielt der Proband mehr Treffer als in der M– Bedingung. Nur in einem der acht aufeinander folgenden M+ und M– Runpaare sind Treffer bei M– häufiger als bei M+, was nach dem Vorzeichentest marginal signifikant ist (p = .062).[19]

Tabelle 2: Trefferzahlen, linker und rechter Beutel, unter der Bedingung M+ (mit Meditationsvorbereitung) und M– (ohne Meditation) über acht Paare von M+/M– Runs. Treffer ohne Trefferpaschs im Kleindruck.

		Aufeinander folgende acht M+ / M– Run-Paare								
		1	2	3	4	5	6	7	8	Summe
M+	Links	9 $_{7}$	13 $_{7}$	14 $_{12}$	22 $_{18}$	19 $_{15}$	8 $_{6}$	13 $_{12}$	6 $_{5}$	104
	Rechts	13 $_{11}$	20 $_{14}$	8 $_{6}$	9 $_{5}$	7 $_{3}$	15 $_{13}$	10 $_{9}$	17 $_{16}$	99
M–	Links	11 $_{8}$	10 $_{9}$	6 $_{4}$	8 $_{7}$	13 $_{10}$	15 $_{12}$	16 $_{12}$	14 $_{10}$	93
	Rechts	9 $_{6}$	14 $_{13}$	11 $_{9}$	19 $_{18}$	13 $_{10}$	7 $_{4}$	14 $_{10}$	8 $_{4}$	95
M+	Summe	22	33	22	31	26	23	23	23	203
M–	Summe	20	24	17	27	26	22	30	22	188
	Differenz	2	9	5	4	0	1	-7	1	15

Treffererwartung pro Durchgang: 12 für links oder rechts, 24 insgesamt.

Ein weiterer Unterschied zwischen M+ und M– fällt auf, wenn man die Treffer herausgreift, die keine Pasch-Treffer sind, also nur diejenigen Treffer eines Beutels, bei denen gleichzeitig ein Nicht-Treffer im jeweils anderen Beutel vorlag. Wenn diese Einbeutel-Treffer Zufallstreffer sein sollten, dann würde ihre Anzahl beim linken Beutel mit der Anzahl beim rechten Beutel nicht zusammenhängen. Doch findet sich unter der M+ Bedingung eine signifikant negative links-rechts Korrelation der Treffersummen (Pitman-r = **-.**92, p=.0015, p_k=.03. Zu Pitmans exaktem Korrelationstest siehe z.B. Lienert (1973, S. 662). Unter der M– Bedingung hat die links-rechts-Korrelation die gleiche Richtung, ist aber schwächer (r = -.34, n.s.).
Die links-rechts Korrelation der Treffer über alle 16 Runs (8 M+, 8 M–) ist ebenfalls negativ, sie beträgt rho = -.67 (p = .003, p_k = .06; der rechenintensive Pitman-Test ist wegen der größeren Anzahl Beobachtungen hier zu zeitraubend). Das heißt, wenn die Anzahl der Treffer links und rechts nicht gleich ist (also keine Paschs vorliegen), dann konzentrieren sie sich,

[19] Die Bedeutsamkeit dieses ohnehin nur marginal signifikanten Ergebnisses ist weiter abzuschwächen, da die Methode der Unterschiedsprüfung (Wahl des Vorzeichentest) vor dem Experiment nicht festgelegt war. Der Wilcoxon-Test ergibt allerdings auch p = .07. Gepaart wurden Test 1 (M–) mit Test 3 (M+), Test 2 (M–) mit Test 4 (M+) usw.

vor allem unter der M+ Bedingung, entweder auf den linken oder auf den rechten Beutel und sind dann beim jeweils anderen Beutel geringer.[20]

Der Standard-Ergebnisausdruck einer statistischen Analyse von Balltest-Daten berücksichtigt neben den bereits genannten noch andere Sekundärvariablen, die hier aus Platzgründen nicht beschrieben werden können. Um den Effekt der Meditationsbedingung insgesamt zu prüfen, wurden paarweise alle Z-Werte der M+ und M– Bedingung (Abweichungen vom Zufall) ermittelt, die mindestens p=.05 erreichten, d.h. aus der Gesamtheit wurde ein Z-Werte-Paar dann herausgelöst, wenn Z mindestens bei einer Bedingung, bei M+ oder M–, die für Signifikanzberechnungen festgelegte Abweichung erreichte. So wurden mit einem zwar willkürlich gewählten, aber ansonsten bias-freien Kriterium für beide Bedingungen die stärksten Abweichungen vom mittleren Zufall abgeschöpft (eine Bonferroni-Korrektur ist für den Auslesezweck nicht erforderlich). Es ergaben sich 20 Paare von Z_{M+} vs. Z_{M-} Werten, die gepaarten absoluten Z-Werte wurden dem Mann-Whitney-U-Test zugeführt.[21]

Wenn in den Testsitzungen von MG kein Psi vorhanden war, dann sollte kein mittlerer Unterschied zwischen den Z_{M+} und Z_{M-} Werten vorliegen. Nach dem Ergebnis des Mann-Whitney-U-Tests scheint jedoch die Null-Hypothese nicht haltbar zu sein: U = 112.5, Z = 2.37, p = .009. Unter der Meditationsbedingung wurden demnach sehr signifikant größere Abweichungen vom Zufall beobachtet als unter der Nicht-Meditationsbedingung. Obgleich, wie oben beschrieben, über die Primärvariablen Treffer, Paschs usw. keine Anzeichen einer Psi-Wirkung erkannt wurden, erscheint mir mit den insgesamt größeren Abweichungen vom Zufall auf Nebenkanälen der Verdacht einer Psi-Wirkung bei MG bekräftigt.

Schlussdiskussion

MG ist kein Überflieger unter den 28 Probanden, die den Zweibeutel-Balltest bisher durchführten.[22] Dennoch produziert auch er Abweichungen von der Erwartung, die nicht mehr im Zufallsbereich liegen. Es ragen heraus die Unterschiede zwischen der Bedingung M+ (mit vorbereitender Meditation) und M– (ohne Meditation), sowie einige Sondereffekte, vor allem die mit den Aufgabenzahlen.

Interessanterweise kommen Abweichungen vom Zufall nicht bei den bewusst wünschbaren Zieh-Variablen vor (Treffer, Paschs). Psi-Tendenzen bei MG, die vermutlich trotzdem vor-

---

[20] *Negative* Links-Rechts-Korrelationen der gleichen Art kommen auch bei anderen psi-begabten Probanden häufiger vor als positive Korrelationen, was vielleicht ein Hinweis darauf ist, dass Psi-Energie, was immer dies auch sei, begrenzt ist und innerhalb einer Zeiteinheit auf verschiedenen Kanälen nur „verteilt" werden, nicht aber auf mehreren Kanälen gleichzeitig gesteigert werden kann.

[21] Es wird nicht bestritten, dass zwischen einigen Variablen Abhängigkeiten vorliegen, der Wilcoxon-Test setzt unabhängige Variablen voraus. Doch halte ich die Abhängigkeiten für minimal, ein Gutachter hielt offenbar stärkere Abhängigkeiten für denkbar.

[22] Spitzenkandidaten im Zweibeutel-Test erreichen allein hinsichtlich der Treffer bis zu 50% Überhang, z.B. erzielte K.G., ein 16-jähriger Schüler, in 31 Runs bei einer Treffererwartung von 744 eine Treffersumme von 1130. Unter meiner Kontrolle blieben seine Trefferquoten auf ungefähr gleichem Niveau.

handen sind, scheinen den Hauptkanälen zu ihrer Manifestation auszuweichen, die durch die Versuchsanordnung angeboten wurden, um stattdessen periphere Kanäle zu suchen, die der bewussten Kontrolle entzogen sind. Für eine unbewusste Psi-Hemmung spricht u.a. auch ein starkes Schwanken zwischen erhöhten und erniedrigten Trefferzahlen (Effekt-Heterogenität) *innerhalb* der Testserie. Auch unter den Balltest-Ergebnissen bei anderen Probanden waren Psi-Effekte nicht selten, die offenbar durch eine Art Hemmung in ihr Gegenteil umgeschlagen waren und die unerwünschte Richtung genommen hatten.

Die hier dargestellten Ähnlichkeiten der Ergebnisse bei MG im parapsychologischen und im astrologischen Test wollen nicht *beweisen,* dass ein Zusammenhang zwischen Psi und Astrologie besteht, zumal die Effekte nicht stark waren. Doch scheint mir, dass nach den Testergebnissen mit MG Anlass genug vorhanden ist, der Frage weiter nachzugehen, ob astrologische Deutungen von Geburtshoroskopen, wenn sie gelegentlich überzufällig treffend ausfallen, mit außersinnlichem Informationszuwachs (Hellsehen oder Telepathie) zu erklären sind.

Ein Zusammenhang zwischen Astrologie und Psi ist schon früher für möglich gehalten worden (Brier 1974; Kelly & Locke 1981; Timm & Köberl 1986). Timm & Köberl (1986) diskutieren die These nach einer Re-Analyse von Daten, die aus einer Untersuchung des Freiburger Instituts für Grenzgebiete der Psychologie und Psychohygiene stammten. Astrologen (N=178) hatten die Güte ihrer Horoskop-Deutungen anhand von sechs verschiedenen Tests zu beweisen versucht. Dem Erstautor Ulrich Timm, bekannt für seine ausgeprägte methodische Vorsicht in der Grenzgebietsforschung, kam nach einer kritischen Analyse der Astro-Untersuchungsdaten zu dem Ergebnis, dass eine statistisch signifikante Deutungs-Gesamtleistung der Astrologen vorhanden sei (p=.002). Selbst nach übervorsichtigem Ausscheiden von zwei der sechs Testvarianten, die man bei extremer Skepsis für bias-beeinflusst halten könnte, blieb die astrologische Trefferleistung auf dem p=.01 -Niveau sehr signifikant.

Was die Erklärung des dieserart erhärteten Effekts betrifft, so hielten Timm & Köberl (1986) von vier denkbaren Möglichkeiten nur zwei für diskutabel: entweder war der Deutungserfolg durch Astrologie oder durch Psi vermittelt, wobei die Autoren Psi-Faktoren für die wahrscheinlicheren oder auch alleinigen Vermittler hielten: „*Die Astrologen erhalten ihre Informationen auf einem anderen [nicht-astrologischen] …Weg, z. B. durch ASW"* [ASW= außersinnliche Wahrnehmung] (Timm & Köberl 1986, S. 54).

Ein Zusammenhang zwischen Psi und Astrodeutung erschien den Autoren u.a. aus demselben Grund nahe liegend, der auch mich – bei der Analyse der MG-Daten und unabhängig von Timm & Köberl – dazu gebracht hatte, die Idee eines Astro-Psi-Zusammenhangs ins Spiel zu bringen. Timm & Köberl waren trotz der positiven Gesamtleistung der Astrologen des Freiburger Tests auf eine hohe Inkonstanz der individuellen Teilnehmerleistungen aufmerksam geworden. Die Deutungen und Zuordnungen einzelner Teilnehmer lagen oft in einem Teiltest richtig, im anderen völlig falsch. Daraus schlossen sie: „*Die…Inkonstanz der Astrologenleistung [ist] ideal mit der unberechenbaren ‚Elusivität von Psi' vereinbar, während sie dem Konzept einer auf präzisen, objektiven Deuteregeln und auf persönlichem Können basierenden Astrologie widerspricht"* (Timm & Köberl 1986, S. 54).

Offenbar ist die Vorstellung, dass außersinnliche Einflüsse bei der Deutungsarbeit der Astrologen einen Beitrag leisten könnten, selbst Astrologen nicht fremd: *„It's somewhat of a com-*

mon belief among astrologers" antwortete Ken Irving, Erst-Herausgeber von *American Astrology* und von *Planetos, An Online Journal,* auf meine Nachfrage (E-mail vom 24. Juni 2001). Dean & Kelly (2003, S. 177-180) zitieren verschiedene Astrologen, die einen Psi-Astrologie-Zusammenhang behaupten: „*British astrologer Charles Harvey (1994) … argues that there can be a psi component to astrology (a point most astrologers would agree with…)*". Alan Vaughan (1973): "*My own small experience with astrologers has given me the impression that their best hits are psychic rather than astrological*". Doris Chase Doane (1956)*: "it is almost impossible to read a birth-chart … without exercizing in some degree Extra-sensory Perception"*. Geoffrey Cornelius (1994): *"[Some unknown element] is involved in the astrological interpretation… [and] is broadly but consistently characterized … as either ESP or intuition"*. Dal Lee (1964): "*…concludes that astrological meanings are too broad to allow specific statements unless some ESP faculty is used, and that some astrologers have ESP at least some of the time….*"

Systematischere Untersuchungen zu dieser Hypothese lassen sich mit dem hier vorgestellten Balltest durchführen. Wenn Astrologie-Erfolg mit Psi zusammenhängt, dann

- sollten Star-Astrologen, die von ihrer Community als besonders erfolgreich anerkannt werden (Leo Knegt-Typen), auch im Balltest viele Erfolge bzw. Zufallsabweichungen zeigen, signifikant mehr als der Durchschnitt unausgelesener Probanden;
- sollten astrologische Deutungs- und Zuordnungserfolge mit dem natürlichen Schwanken der Psi-Bereitschaft kovariieren, das sich durch den Balltest erfassen ließe (bei psibegabten Probanden können die Trefferquoten schwanken, was vielleicht mit der Tagesverfassung oder mit anderen zeitabhängigen Faktoren zusammen hängt);
- sollten astrologische Deutungserfolge häufiger vorkommen, wenn sich die Versuchsteilnehmer vor dem Test in eine meditative oder andere psi-affine Verfassung bringen.

Von der Grundhypothese ausgehend sollten weitere Detail-Hypothesen ableitbar sein. Dem Einfallsreichtum forschungsbereiter Leser sind keine Grenzen gesetzt.

Postskriptum: Jüngste Untersuchungsergebnisse

MG bekundete nach langem Zögern zwischenzeitlich Interesse, sich mit einer anderen Variante des Balltests testen zu lassen, bei dem nur ein Beutel und eine Trefferwahrscheinlichkeit von .10 verwendet wird (die Zahlen von 1 bis 5 kommen gleich häufig mit grünen oder roten Punkten vor; Heimtest, ohne Meditation). In acht Runs erzielte er 6, 2, 7, 6, 13, 6, 1, 7 Treffer (beobachtet pro Run: 6.0, erwartet 6.0). Die Summe der acht binomialen Z^2 -Werte ergibt $Chi^2 = 19.6$, df = 8, p = .01. Wieder also tritt bei MG keine konstante Abweichung von der Erwartung, sondern ein signifikanter Wechsel zwischen Trefferüberschuss und Treffermangel auf. Zwei weitere signifikante Nicht-Treffer-Abweichungen kamen vor: Die Ziehhäufigkeit der fünf Zahlen (unabhängig vom Raten) ist unausgewogen ($Chi^2 = 15.6$, df = 4, p = .005, $p_k = .03$). Sodann clustert bei MG das Ziehen gleicher Farben (Runs-Test Z = -2.1, p = .018). Seine längste Serie (16 mal hintereinander wurde „grün" gezogen) ist nach einem Simulationstest überzufällig lang (p = .01), unter den 47 mit gleichem Test getesteten Studenten waren „nur" 14 Hintereinander-Ziehungen einer gleichen Farbe der Rekord. Dies

entspricht der vermuteten Tendenz bei MG, zum Austragen vorhandener Psi-Tendenzen sekundäre Kanäle zu verwenden.

Literatur

Brier, B. (1974): Precognition and the philosophy of science. Humanities Press, New York.

Dean, G.; Kelly, I.W. (2003): Is astrology relevant to consciousness and psi? *Journal of Consciousness Studies* 10 (6-7), 175-198.

Ertel, S. (1988): Gauquelins Planetenhypothese: Stein des Anstoßes oder Prüfstein der Vernunft? *Psychologische Rundschau* 39 (4), 179-190.

Ertel, S. (1998): Astro-Quiz: Can astrologers pick politicians from painters? *Correlation* 17 (1), 3-8.

Ertel, S. (2000): Psi experiments on untrodden ground. Old problems in view of new method. Paper presented at the 14th Annual Convention of the Parapsychological Association, Freiburg, Germany, August 17-20.

Ertel, S. (2003): The Ball Drawing Test. Psi from untrodden ground. In: Thalbourne, M.A. (Ed.): Parapsychology in the 21st Century: The Future of Psychical Research. MacFarland, Jefferson, NC.

Ertel, S.; Irving, K. (1996): The Tenacious Mars Effect. Urania Trust, London.

Gauquelin, M. ; Gauquelin, F. (1970): Professional Notabilities Series A, Volumes 1-6. Hereditary Experiment. Series B. Volumes 1-6. LERRCP, Paris.

Honorton, C. (1977): Psi and internal attention states. In: Wolman, B.B. (Ed.): Handbook of parapsychology. Van Nostrand, New York.

Kelly, E.F.; Locke, R.G. (1981): A note on scrying. *Journal of the American Society for Psychical Research* 75, 221-227.

Kimmel, H.D. (1957): Three criteria for the use of one-tailed tests. *Psychological Bulletin* 54, 351-353.

Lienert, G.A. (1973): Verteilungsfreie Methoden in der Biostatistik. Vol. I. Hain, Meisenheim.

Müller, A.; Ertel, S. (1992): Astrologisches Zuordnungsexperiment mit Ärzte-Horoskopen. *Zeitschrift für Parapsychologie und Grenzgebiete der Psychologie* 34(3/4), 217-221.

Rao, P.V.K.; Rao, K.R. (1982): Two studies of ESP and subliminal perception. *Journal of Parapsychology* 46, 185-207.

Rhine, J.B. (1952): The problem of psi-missing. *Journal of Parapsychology* 16, 90-129.

Rosenthal, R.; Rubin, D.B. (1989): Effect size estimation for one-sample multiple-choice-type data: Design, analysis, and meta-analysis. *Psychological Bulletin* 106 (2), 332-337.

Rossem, C.P. van (1933): Twee occulte problemen. De mensch buiten zijn lichaam, uittredingsverschijnselen.. De mensch in den horoscoop, experimenten op het gebied der astrologie. W P van Stockum & Zoon, The Hague.

Smit, R. (1997): Leo Knegt: A white crow beyond our wildest dreams. *Correlation* 16(1) , 3-18.

Smit, R. (1998): Results of the Knegt follow-up test. *Correlation* 17 (2), 72-75.

Timm, U. (1983): Statistische Selektionsfehler in der Parapsychologie und in anderen empirischen Wissenschaften. *Zeitschrift für Parapsychologie und Grenzgebiete der Psychologie* 25 (3/4), 195-229.

Timm, U.; Köberl, T. (1986): Re-Analyse einer Validitätsuntersuchung an 178 Astrologen. *Zeitschrift für Parapsychologie und Grenzgebiete der Psychologie* 28 (1/2), 33-55.

Kommentare zu Ertel: Astrologie und Psi

EMIL BOLLER[23]

Einwände gegen die Psi-Interpretation der Einzeltests in Ertels Fallstudie

Astrologie ist durch einen großen Widerspruch geprägt. Auf der einen Seite erfahren die meisten Astrologen im direkten Umgang mit Ratsuchenden immer wieder Bestätigungen für die Richtigkeit ihrer Horoskopdeutungen. Auf der anderen Seite scheitern Bemühungen die Behauptungen von Astrologen in kontrollierten Untersuchungen empirisch zu belegen regelmäßig.

Ertel demonstriert diese Diskrepanz sehr schön anhand eines Einzelfalles. Erfreulicherweise zeigte sich die beteiligte Person als lernfähig. Sie glaubt nun nicht mehr daran, „dass man den Lebensweg eines Menschen auch nur in groben Zügen aus dem Horoskop ablesen kann“. Eine behauptete Fähigkeit systematisch zu überprüfen hat sich auch in der Beratungstätigkeit bewährt. Oft verhilft sie Ratsuchenden zu einer realistischeren Einschätzung ihrer vermeintlichen paranormalen Fähigkeiten. Insgesamt erweisen sich Astrologen jedoch wenig bereit, die Ergebnisse von inzwischen mehr als 500 empirischer Studien, die zum größten Teil nach 1950 erstellt wurden, überhaupt zur Kenntnis zu nehmen (Dean & Kelly 2003).

Anhand von drei Zuordnungstests mit heterogenen Ergebnissen und eines neuartigen Psi-Tests versucht Ertel in einem Einzelfall einen Zusammenhang zwischen Astrologie und Psi herzustellen. Allerdings sind die gewählten Tests bzw. ihre Umsetzung zur Klärung der Frage ungeeignet.

---

[23] Emil Boller ist Diplom-Psychologe und war viele Jahre wissenschaftlicher Mitarbeiter am Institut für Grenzgebiete der Psychologie und Psychohygiene (IGPP) in Freiburg/Breisgau. Anschrift: Birkenweg 1, D-79100 Freiburg. E-Mail: boller@anomalistik.de

Einwände gegen die Handhabung und die Psi-Interpretation der Zuordnungstests

Ertel erwägt Psi als mögliche Erklärung für die gefundene Variabilität in den Ergebnissen der Zuordnungstests. Bevor man ihm darin folgt, sollte man sich vergegenwärtigen, was eigentlich in einem Zuordnungstest gemessen wird und wie und unter welchen Bedingungen sich Psi darin äußern könnte.

Üblicherweise wird für einen Zuordnungstest eine Stichprobe von Geburtsdaten von Menschen zusammengestellt, die sich meist in einem Merkmal voneinander unterscheiden. Es wird angenommen, dass Astrologen in der Lage sind, aufgrund der Deutung des zugehörigen Geburtshoroskops die einzelnen Fälle der richtigen Kategorie zuzuordnen. Um diese Aufgabe erfolgreich zu lösen, muss der Astrologe zutreffende Vorstellungen darüber haben, durch welche Eigenschaften sich z.B. Politiker von Malern oder erfolgreiche Schriftsteller von gewöhnlichen Menschen unterscheiden und wie sich diese Unterscheidungsmerkmale im Horoskop ausdrücken. Unabhängig davon, ob diese Vorstellungen zutreffen, ist mit der Auswahl der Geburtsdaten der Ausgang solch eines Tests eigentlich determiniert. Es stellt sich dann die Frage, was mit solch einem Test tatsächlich gemessen wird: Die Güte des astrologischen Systems, die Qualität der Annahmen bezüglich der die Gruppen unterscheidenden Merkmale oder die Qualität der Stichprobe von Personen und ihrer zugehörigen Geburtszeitpunkte?

Angenommen, Psi – verstanden als die Nutzung paranormal erworbener Zusatzinformationen – käme in einem Zuordnungsexperiment zum Tragen, dann würde sich das auf zweierlei Weise ausdrücken:

(1) Es lassen sich Zuordnungen finden, die entgegen den astrologischen Regeln getroffen wurden und die richtig (oder konsistent falsch) sind.

(2) Bei aus astrologischer Perspektive uneindeutigen Horoskopen wird überzufällig häufig die richtige (falsche) Zuordnung getroffen.

In dem von Ertel beschriebenen Einzelfall handelt es sich um einen Astrologieanfänger, der gerade einen Grundkurs in indischer Astrologie absolvierte und schulbuchmäßig vorging, wie aus seinem Kommentar zum Ergebnis des ersten Zuordnungsexperiments ersichtlich wird. Nach der Rückmeldung des unerwartet ungünstig ausgegangenen ersten Zuordnungstests überprüfte er seine Deutungen mit dem im Kurs Gelernten und fand keinen Fehler. Daher gibt es wenig Grund anzunehmen, dass Fall 1 eingetreten ist. Die Möglichkeit für den Fall 2 wurde durch Ertel leichtfertig verspielt, indem er dem Astrologieanfänger erlaubte, aus astrologischer Sicht uneindeutige und widersprüchliche Fälle (immerhin 45%) nicht zuzuordnen. Dabei gelten gerade uneindeutige Situationen als psi-induktiv. Aufgrund der Beschreibung der Zuordnungstests halte ich die Psi-Interpretation der Ergebnisse für wenig plausibel. Die gefundene erhöhte Variabilität der Einzelergebnisse ist wahrscheinlich auf methodische Artefakte zurückzuführen, wobei die Ausgrenzung widersprüchlicher Fälle und die kleinen Stichproben eine zentrale Rolle spielen dürften. Dazu bemerken Dean & Kelly (2003): „Claimed success in matching charts to case histories (Clark, 1961) was consistent with the use of tiny samples, typically ten birth charts, whose disproportionately huge sampling variations were mistaken for genuine effects (Eysenck and Nias, 1982, pp. 86-7), a point confirmed by later studies and meta-analysis (Dean, 1986)."

Einwände gegen den Balltest

Der Balltest in der von Ertel beschriebenen Form ist als Psi-Test mehr als fraglich. Sieht man davon ab, dass der Test ohne Kontrollen in der privaten Umgebung des Teilnehmers stattfand, gibt es noch weitere Einwände. Während beim Zahlenlotto das Volumen der Kugeln im Verhältnis zum Ziehungsgerät recht klein ist und diese sich idealisiert dargestellt wie Gasmoleküle in einem Raum verhalten können, sind die Tischtennisbälle in einem Beutel untergebracht, deren Hülle sich an die Bälle schmiegt. Im Gegensatz zu den Bällen in Ziehungsgeräten sind die Bälle in einem Beutel in einem viel geringerem Maße gegeneinander beweglich. Sie verhalten sich eher wie die Moleküle in einer sehr zähen Flüssigkeit. Verschärft wird dieses Manko noch dadurch, dass die Beutel zwischen den Ziehungen nur „gewendet" werden, während bei früheren Experimenten von Ertel noch gefordert wurde, dass der Beutel vor der jeweils nächsten Ziehung geschüttelt wird. Eine Durchmischung der Bälle und damit die Unabhängigkeit der einzelnen Ziehungen voneinander ist somit nicht gewährleistet. Der Art und Weise, wie einzelne Teilnehmer den Test durchführen, d.h. wie sie den Ball ziehen (rühren sie in der Tasche, bevor sie ziehen, ziehen sie gezielt nur an bestimmten Positionen oder nutzen sie das ganze Volumen der Beutel) und wie die gezogenen Bälle zurückgelegt werden (systematisch oder unsystematisch, wobei der Teilnehmer immer weiß, an welchen Positionen die zuletzt gezogenen Bälle zurück gelegt wurden) kommt daher eine überragende Bedeutung zu. Nur ist dieses Verhalten keiner Analyse zugänglich, solange die Tests unbeobachtet in häuslicher Umgebung stattfinden. In einem Selbstversuch mit von Ertel dem IGPP zur Verfügung gestellten Versuchsmaterial gelang es mir unter Einhaltung der Versuchsanleitung nur durch systematisches Zurücklegen der Bälle und trotz des Schüttelns des Beutels zwischen den einzelnen Ziehungen, eine erhöhte Trefferzahl zu produzieren. Besonders Bälle, die in die Ecken des Beutels zurückgelegt wurden, waren recht sichere Treffer.

Die von Ertel als sekundäre Psi-Indikatoren bezeichneten Auffälligkeiten in den Daten sollten erst dann als solche interpretiert werden, wenn definitiv ausgeschlossen ist, dass diese durch unbewusste oder gezielte Verhaltensmuster in Verbindung mit der unzureichenden Durchmischungsprozedur zustande gekommen sind. Schon vor Jahren machte nicht nur ich Ertel im Rahmen eines WGFP-Workshops, wo er den Balltest vorstellte, auf dieses Problem aufmerksam.

Auch beim Balltest wird ein Psi-Effekt also mehr postuliert, als dass er überzeugend nachgewiesen wird.

Fazit

Der von Ertel nahe gelegte Zusammenhang zwischen Astrologie und Psi ist in hohem Maße konstruiert. Gegenargumente werden nicht in Erwägung gezogen. Dies zeigt sich besonders deutlich darin, dass Dean & Kelly (2003), welche die gleiche Thematik sehr detailliert behandeln, nur als Quelle für Zitate von Astrologen, die einen Psi-Astrologie-Zusammenhang behaupten, herangezogen werden, während ihre profunde Auseinandersetzung mit der Thematik mit keinem Wort gewürdigt wird. Daher erlaube ich mir deren Bilanz im vollen Umfang zu zitieren (Dean & Kelly 2003, S. 195):

"Our concern in this article has been to measure the performance of astrology and astrologers. A large-scale test of time twins involving more than one hundred cognitive, behavioural, physical and other variables found no hint of support for the claims of astrology. Consequently, if astrologers could perform better than chance, this might support their claim that reading specifics from birth charts depends on psychic ability and a transcendent reality related to consciousness. But tests incomparably more powerful than those available to the ancients have failed to find effect sizes beyond those due to non-astrological factors such as statistical artifacts and inferential biases. The possibility that astrology might be relevant to consciousness and psi is not denied, but if psychic or spirit influences exist in astrology, they would seem to be very weak or very rare. Support for psychic claims seems unlikely."

Literatur

Clark, V. (1961): Experimental astrology. *In Search*, Spring 1961, 101-112.

Eysenck, H.J.; Nias, D.K.B. (1982): Astrology – Science or Superstition? St Martin's Press, New York.

Dean, G. (1986): Can astrology predict E and N? Part 3: Discussion and further research. *Correlation* 6 (2), 7-52.

Dean, G.; Kelly, I.W. (2003): Is astrology relevant to consciousness and psi? *Journal of Consciousness Studies* 10 (6-7), 175-198.

VOLKER GUIARD[24]

Statistik mangelhaft

Als Gutachter dieses Artikels hatte ich – nach einer äußerst schwierigen und langwierigen Diskussion mit Herrn Ertel – die Wahl, entweder die Publikation wegen statistischer Mängel abzulehnen oder einer Veröffentlichung zuzustimmen und dann die Statistik zu kommentieren, wobei ich mitunter auch Informationen aus der Gutachtertätigkeit einfließen lasse. Ich habe mich für den zweiten Weg entschieden, wobei es bei meinen Bemerkungen nicht immer nur um statistische Mängel geht, sondern mitunter auch um Klarstellung von Dingen, die im Artikel nicht deutlich werden.
In meinem Kommentar gehe ich nur auf methodisch-statistische Aspekte ein. Es gäbe sicherlich viele weitere Dinge zu diskutieren, ich hoffe jedoch, dass dies von anderen Kommentatoren übernommen wird.

1. Konfirmatorische und exploratorische Tests und das multiple Testproblem

Als *konfirmatorisch* bezeichnet man einen Test, wenn er eine vor dem Versuch (bzw. vor Ansehen der Daten) aufgestellte Hypothese prüfen soll. Die Ergebnisse solcher Tests gelten als

[24] PD Dr. Volker Guiard arbeitet am Forschungsbereich Genetik und Biometrie des Forschungsinstituts für die Biologie landwirtschaftlicher Nutztiere in Dummerstorf bei Rostock. Anschrift: Zum Laakkanal 14, D-18109 Rostock, E-Mail: guiard@anomalistik.de

die eigentliche Aussage eines Versuches. Werden mehrere solche Tests durchgeführt (multiple Testprozedur) so ist unter der Annahme, dass bei allen Tests die entsprechende Nullhypothese wahr ist, die Wahrscheinlichkeit, dass mindestens einer der Tests fälschlich Signifikanz anzeigt, größer als das vorgegebene Signifikanzniveau (hier 0,05). Aus diesem Grunde werden die Test korrigiert. Bei der Auswertung seines parapsychologischen Ballzieh-Tests verwendet Ertel zu diesem Zweck die Bonferroni-Korrektur, wobei die Anzahl der konfirmatorischen Tests bekannt sein muss. Die Anzahl der Tests gibt Ertel mit „rund 20" an.
Entschließt man sich nach Betrachtung der Daten dazu, die Signifikanz einiger interessant erscheinender Datenauffälligkeiten zu testen, so nennt man solche Tests *exploratorisch*. Dabei ist das Fehlerrisiko bedeutend größer, da nur dort getestet wird, wo sich bereits gewisse Auffälligkeiten zeigten. Damit können diese Testergebnisse nicht als die eigentlichen Ergebnisse des Versuches gelten. Sie dienen nur dazu, auf eventuelle Phänomene vage hinzuweisen, deren genauere (konfirmatorische) Prüfung jedoch zukünftigen Versuchen vorbehalten bleibt. Auch bei solchen Tests könnte man die Bonferroni-Korrektur oder analoge Korrekturen verlangen (vgl. Timm 1983, S. 215). Ertel verzichtet aber hierbei auf eine solche Korrektur, was ich akzeptierte, sofern sich bei eventuellen positiven Ergebnissen die Euphorie in Grenzen hält. Bei diesem weniger strengen Vorgehen muss man bei den entsprechenden Bestätigungstests in Wiederholungsversuchen eher mit Enttäuschungen rechnen.
Ertel verwendet die Begriffe „konfirmatorisch" und „exploratorisch" nicht explizit. Bei seinem Balltest gilt als Faustregel, dass ein statistischer Test als konfirmatorisch zu betrachten ist, falls die Bonferroni-Korrektur angewendet wurde, anderenfalls als exploratorisch.
In dem einleitenden Text zu den Ergebnissen des Ballzieh-Tests meint Ertel, „bei Hypothesen, die in der vorliegenden Untersuchung eingeführt wurden [V.G.: gemeint ist natürlich: *vor* dem Versuch] und zur Entscheidung anstanden", auf die Bonferroni-Korrektur verzichten zu dürfen. Diese Tests gehören aber zu den konfirmatorischen Tests, ein Verzicht auf die Bonferroni-Korrektur ist daher hier nicht gerechtfertigt. Als Beispiel nennt Ertel einen Test, den ich nachfolgend unter Punkt 9 noch kommentieren werde. Dieser Test wurde jedoch nicht vor, sondern nach Sichtung der Daten eingeführt. Er ist also nicht konfirmatorisch, sondern exploratorisch, womit auch der eigentliche Grund des Korrekturverzichts gegeben ist. Er ist kein Beispiel für eine vor dem Versuch eingeführte Hypothese.

2. Analyse aufeinanderfolgender Versuche und die „split-half"-Methode

In der Einleitung zu den Ergebnissen des Ballzieh-Tests schreibt Ertel, dass „bei Replikation signifikanter Effekte innerhalb der Testdaten, sofern diese voneinander unabhängig sind", auf die Bonferroni-Korrektur verzichtet werden könne. Dieses sei „einem split-half Korrelationstest analog, bei dem man nach Vorliegen eines signifikanten Effekts bei der einen Testhälfte die andere gezielt auf Effektresonanz prüft". Diese Formulierung bezieht sich aber nicht auf das multiple Testproblem, sondern auf einen anderen Sachverhalt, der zur Bonferroni-Korrektur ohnehin keine Beziehung hat. Zur Erläuterung betrachten wir zunächst den Fall zweier analoger Versuche, die nacheinander durchgeführt werden. Hierbei könnte man den ersten Versuch explorativ auswerten, wobei eine signifikante Auffälligkeit in den Daten als Hypothese formuliert werden könnte, welche dann mit dem unabhängig durchzuführenden zweiten Versuch konfirmatorisch getestet wird. Hier dient also der erste Versuch ledig-

lich zur Hypothesengenerierung. Der Test dieser Hypothese erfolgt dann nur mit dem zweiten Versuch, ist also vom ersten Versuch unabhängig. Analog kann man auch vorgehen, falls ein (umfangreicher) Versuch durchgeführt wurde, sofern man dessen Daten (unabhängig von den Werten) zufällig in zwei Gruppen teilt, wobei dann eine Gruppe die Rolle des ersten Versuches übernimmt und die zweite Gruppe die des zweiten Versuchs. Dieses Vorgehen ist mit „split-half" gemeint. Beide Vorgehensweisen sind bei dem Ballzieh-Test jedoch gar nicht verwendet worden, statt dessen aber öfter bei den astrologischen Zuordnungstests, und zwar fehlerhaft. In den weiteren Punkten werde ich darauf zurückkommen.

3. Der Test von Rosenthal & Rubin (1989) und der Binomialtest

Sowohl Ertel (durch Simulation) als auch ich (theoretisch) stellten fest, dass der Test von Rosenthal & Rubin (1989) fehlerhaft ist.[25] Hier soll deswegen weiterhin nur der Binomialtest betrachtet werden. Aber auch dieser Test hat seine Tücken, er setzt nämlich voraus, dass die beiden Gruppen (z.B. Politiker und Maler) den gleichen Umfang haben. MG hat jedoch einige dieser Personen als nicht zuordnungsfähig ausgeschlossen. Damit hätten aber die Anzahlen der verbleibenden Politiker und Maler zufällig recht unterschiedlich ausfallen können. Zum Glück war dieses hier aber nicht der Fall. Durch Simulation zeigte ich, dass durch die leicht unterschiedlichen Gruppengrößen das Risiko α nicht vergrößert wurde. Andernfalls hätte der Test wie bei einer 2×2-Tafel (Berufe×Zuordnungen) durchgeführt werden müssen (Chi-Quadrat-Test oder exakter Test von Fisher).

4. Einseitig oder zweiseitig?

Für das erste astrologische Zuordnungsexperiment verwendet Ertel einen einseitigen Test. In der Fußnote 5 begründet er dieses unter Verwendung des hierfür üblichen Motivs, so wie es (z.B.) bei Kimmel (1957) zu finden ist. Zur Verdeutlichung dieses Motivs stelle man sich vor, jemand hätte ein neues Verfahren entwickelt, von dem er eine bessere Leistung erhofft, als von dem bisher üblichen Verfahren. In einem Experiment soll das neue Verfahren seine bessere Leistung nun unter Beweis stellt. Sollte es sich tatsächlich als signifikant besser erweisen, so wird man dieses Verfahren in die Praxis einführen. Bei Nichtsignifikanz bleibt man jedoch bei dem alten Verfahren. Aber auch in dem unerwarteten Fall, dass das neue Verfahren signifikant schlechter sein sollte, ist die praktische Konsequenz (nämlich die Nichtverwendung in der Praxis) dieselbe, wie im Fall der Nichtsignifikanz. Daher ist die Unterscheidung zwischen „Nichtsignifikanz" und „signifikant schlechter" uninteressant, weswegen hier nur ein einseitiger Test (genauer: ein rechtsseitiger Test in Richtung „besser") durchgeführt wird. Diese nur von dem Zweck des Versuchs und nicht von den Daten abhängige Entscheidung ist natürlich vor dem Versuch zu treffen.
Bei dem ersten Zuordnungsexperiment fiel nun die Trefferanzahl zu gering aus. Ertel schreibt: „Die Abweichung vom Zufall hatte die falsche Richtung." Es war also eine eher zu große Trefferanzahl erwartet worden, d.h., es war ein rechtsseitiger Test geplant. Nach Ansehen der Daten führt Ertel nun jedoch einen linksseitigen Test durch. Diese unzulässige

[25] Ich beabsichtige, den theoretischen Fehler dieses Tests in einem zukünftigen Artikel darzulegen.

Richtungsänderung steht im krassen Widerspruch zu der Formulierung von Kimmel. Der richtige, also rechtsseitige p-Wert, ist ungefähr[26] durch p = 1 - p = 1 - 0,0113 = 0,9887 gegeben. Wegen der vorherigen Festlegung auf den rechtsseitigen Test interessiert der linksseitige p-Wert nicht bzw. höchstens nur exploratorisch. Bei dem zweiten astrologischen Zuordnungsexperiment schreibt Ertel: „Erwartet wurde eine Replikation der ersten Beobachtung." Daraus und aus der Fußnote 7 folgt, dass hier ein linksseitiger Test geplant wurde. Hier wurde also – wie oben unter Punkt 2 beschrieben – ein erster Versuch zur Hypothesengenerierung verwendet und ein zweiter Versuch zur Hypothesenprüfung. Aber auch im zweiten Versuch wurde dann nach Ansehen der Daten die Testrichtung geändert. Der richtige p-Wert ist also ungefähr durch 1 - p = 1 - 0,08 = 0,92 gegeben.
In der Fußnote 7 betont Ertel in korrekter Weise, dass die post-hoc-Entscheidung für einen zweiseitigen Test hier unzulässig wäre. Eine prä-hoc-Entscheidung für einen zweiseitigen Test wäre nur sinnvoll, wenn keine konkrete Abweichungsrichtung erwartet worden wäre, was für Ertel aber nicht zutraf. Vermutlich dachte Ertel während der Auswertung dieser astrologischen Zuordnungstests noch nicht an die in der Einleitung zitierte Arbeit von Timm und Köberl (1986), welche eine Beziehung zwischen Astrologie- und Psi-Effekten annahm, womit die für Psi typische Streuung auch für die Astrologie-Effekte denkbar wird.

5. Die Heterogenitätstests

Die drei Trefferquoten der astrologischen Zuordnungstests kann man jeweils darstellen als Summe der mittleren Trefferquote plus der Abweichung je Test. Diese Abweichungen fielen bei diesen drei Versuchen recht unterschiedlich, d.h. heterogen aus. Als Test dieser Heterogenität gibt Ertel (im Anschluss an den dritten astrologischen Versuch) den hierfür korrekten Chi^2-Test an (Chi^2-Test für eine 2×3-Kontingenztafel; $Chi^2 = 8{,}41$; $df = 2$; $p = 0{,}015$). In Fußnote 9 begründet Ertel, warum er den Summe-Z^2-Test (Timm 1983, S. 222, Formel 5), welcher hier $p = 0{,}069$ liefern würde, für unzulässig hält. Im Gegensatz zu der Annahme von Ertel stellt die Quadrierung der Z-Werte und damit die Unabhängigkeit von ihrem Vorzeichen (Richtung) hier aber nicht das eigentliche Problem dar. Auch bei dem von Ertel verwendeten Chi^2-Test spielt die Richtung der individuellen Abweichungen keine Rolle. Das Problem liegt jedoch darin, dass bei dem Summe-Z^2-Test nicht nur die individuellen Abweichungen, sondern auch die mittlere Trefferquote das Ergebnis beeinflusst. Dieser Test prüft also nicht nur die Heterogenität, sondern den gemeinsamen Effekt von Heterogenität und mittlerer Trefferquote. Im Extremfall würde dieser Test auch dann Signifikanz anzeigen, falls alle drei Trefferquoten identisch wären und sehr stark von 0,5 abwichen. Auch bei Nichtheterogenität könnte dieser Test also Signifikanz ergeben. Andererseits hält Ertel den Test $Chi^2 = -2 \cdot \Sigma \ln(p)$ (Formel 6 von Timm 1983) für geeignet. Wollte man diesen Test jedoch zur Heterogenitätsprüfung verwenden, so müsste man die zweiseitigen p-Werte verwenden (Timm 1983), also die mit 2 multiplizierten p-Werte aus der Tabelle 1. Damit erhält

[26] Bei kontinuierlichen Variablen gilt obige Gleichung exakt. Da die Trefferzahl keine kontinuierliche Variable ist, kann es einige Unregelmäßigkeiten geben, weswegen die Gleichung hier nur ungefähr gilt.

man $Chi^2 = 11{,}15$ ($df = 6$) und $p = 0{,}084$, also keine Signifikanz. Außerdem hat dieser Test denselben Nachteil, wie der Summe-Z^2-Test; auch er reagiert nicht nur auf Heterogenität, sondern auch auf die mittlere Trefferquote.
Zu dem von Ertel korrekt verwendeten Chi^2-Heterogenitätstest ist nun noch zu bemerken, dass er nur als exploratorisch gelten kann und nicht als konfirmatorisch, da die Entscheidung, die Heterogenitätshypothese zu testen, post hoc, also nach Feststellung der Heterogenität gefällt wurde. Auch das oben im Punkt 2 genannte „split-half"-Prinzip kann hier nicht geltend gemacht werden, da die zur Hypothesenprüfung verwendeten Daten nicht von den Daten unabhängig sind, welche zur Hypothesengenerierung verwendet wurden. Der Chi^2-Test verwendet nämlich alle drei Versuche, auch diejenigen, welche auf die Heterogenität erstmalig hinwiesen. Das Entsprechende gilt natürlich erst recht für den mit den ersten zwei Versuchen durchgeführten Chi^2-Test. Hier betont Ertel aber selbst, dass es sich um einen post-hoc-Test handelt.
Wollte man, nachdem die ersten zwei Versuche auf die Heterogenität aufmerksam machten, einen konfirmatorischen Heterogenitätstest durchführen, so müsste man mindestens zwei weitere Versuche durchführen und nur diese neuen Versuche für den Chi^2-Test verwenden.

6. Simultantest aller astrologischen Effekte

Um bei multiplen Tests das in Punkt 1 erwähnte Anwachsen des Risikos erster Art zu vermeiden, verwendet Ertel bei dem Ballzieh-Test die Bonferroni-Korrektur. Aus dem gleichen Grund ist es auch üblich, zunächst einen Simultantest für alle möglichen Effekte durchzuführen und nur dann, wenn dieser Test signifikant ausfällt, die einzelnen Effekte zu testen. Dieses Verfahren bietet sich für die astrologischen Zuordnungsversuche an. Die Gesamtheit aller Effekte wird durch die jeweiligen Effekte der drei Versuche repräsentiert, und zwar der jeweilige Gesamtbetrag dieser Effekte und nicht nur die Effektkomponenten, die den Unterschied der drei Effekte ausmachen. Also auch die gemeinsame Effektkomponente der mittleren Trefferquote ist hierbei zu berücksichtigen. Ihr galt sogar das primäre Interesse. Der Simultantest prüft also den gemeinsamen Effekt von Heterogenität *und* der mittleren Trefferquote. Dieser Simultantest kann nun mit dem in Punkt 4 genannten Summe-Z^2-Test oder mit $Chi^2 = -2 \Sigma \ln (2p)$ durchgeführt werden (Timm 1983 empfiehlt hierfür den ersteren Test). Wie oben erwähnt, zeigen beide Tests aber keine Signifikanz.

7. Sondereffekt mit Aufgabenzahlen

Unter Punkt 5 der Ergebnisse des Ballzieh-Tests (Sondereffekt mit Aufgabenzahlen) wird zunächst festgestellt, dass unter allen 455 Zügen, bei denen links oder rechts Aufgabenzahlen gezogen wurden, die zweite Aufgabenzahl signifikant häufiger auftritt, also die erste. Hierbei verwundert zunächst, dass lediglich bei 455 Zügen Aufgabenzahlen auftraten, obwohl der Erwartungswert der Anzahl solcher Züge 614,4 beträgt. Eine Nachfrage beim Autor ergab jedoch, dass bei der Formulierung „links oder rechts" das „oder" exklusiv gemeint war. Züge, bei denen beide Aufgabenzahlen gezogen wurden, blieben also unberücksichtigt.

In dem nachfolgenden Test wird der von MG erzielte Überhang (0,16) mit dem mittleren Überhang von 27 weiteren Probanden (0,041) verglichen. Hierbei entsteht die Frage, warum der zuerst durchgeführte exakte Test der Nullhypothese „Überhang (MG) = 0“ nun noch mal mit der approximativen Nullhypothese „Überhang (MG) = 0,041“ wiederholt wird, wobei der Wert 0 durch seine empirische Schätzung 0,041 ersetzt wird. Der Sinn dieses Tests ergibt sich jedoch aus der Tatsache, dass der Wert 0,041 seinerseits ebenfalls signifikant von 0 abweicht, der bei MG beobachtete Effekt war also im Mittel auch bei den bisherigen Probanden vorhanden. Insofern zeigt dieser zweite Test, dass der Überhang von MG auch weit über dem bisher beobachteten Durchschnitt liegt. Es ist dabei nur verwunderlich, dass der p-Wert des Tests „0,16 gegen 0,041“ kleiner ausfällt, als der p-Wert des Tests „0,16 gegen 0“. Dieses liegt zum einen daran, dass nun die aus den bisherigen 27 Ergebnissen errechnete Standardabweichung (sd = 0,034) verwendet wurde, die hier kleiner ausfiel als der (durch Simulation geschätzte) theoretische Wert (sd = 0,0467). Weiterhin war der erste Test zweiseitig und der zweite Test einseitig.
Da bei diesem zweiten Test eine Schätzung der Standardabweichung (mit 26 Freiheitsgraden) verwendet wurde, wäre die Verwendung des t-Tests angebracht. Damit erhält man (einseitig) p = 0,001 (pk = 0,02).

8. Der Pitman-Korrelationstest

Unter Tabelle 2 berechnet Ertel die Korrelation ρ der Wertepaare, die sich aus den Trefferanzahlen des linken und des rechten Beutels je Durchgang ergeben. Hierbei sollte die Hypothese $H_0 : \rho = 0$ gegen die linksseitige Hypothese $H_A : \rho < 0$ geprüft werden. Für die jeweils 8 Durchgänge von M+ bzw. M- wurde der nichtparametrische Pitmann-Test (ein Permutationstest) verwendet. Die Korrelation für alle 16 Durchgänge (M+ und M- gemeinsam) wurde jedoch mit dem für Normalverteilung üblichen Test geprüft, da hier der Permutationstest zu aufwendig war. Die Festlegung auf einen linksseitigen Test ist vermutlich aus der Beobachtung, dass auch bei anderen Probanden häufiger negative Korrelationen auftraten, motiviert. Aus unerfindlichen Gründen wurden von Ertel die beidseitigen Treffer (Pasch-Treffer) nicht als Treffer gezählt. Man kann theoretisch zeigen, dass die Korrelation unter der Nullhypothese „kein Psi“ durch $\rho = -0{,}19$ gegeben ist, sofern man die Paschs ignoriert. Damit ist es also nicht verwunderlich, wenn negative Korrelationsschätzungen häufiger auftreten als positive. Die hier verwendeten Tests sind damit unzulässig. Korrekterweise muss hier die Nullhypothese $H_0 : \rho = -0{,}19$ getestet werden. Das war nur mit einem Test möglich, bei dem die gesamte Prozedur der Datenentstehung (ohne Psi) und der Auswertung (ohne Paschs) 100000 mal simuliert wurde. Der Anteil der so ermittelten Korrelationen, die kleiner sind als die Korrelation aus den Originaldaten, entspricht dem p-Wert. Diese p-Werte habe ich in der folgenden Tabelle angegeben. Werden die Paschs als Treffer gewertet, so entspricht dieses Simulationsverfahren dem Test der Nullhypothese „ $\rho = 0$ “.
Nach Bonferroni-Korrektur ist auch hier lediglich bei M+ und ohne Paschs festzustellen, dass die Korrelation -0,919 signifikant kleiner ist als der entsprechende Nullhypothesenwert -0,19.

Pasch:	Ohne		mit	
H_0:	$\rho = 0{,}19$		$\rho = 0$	
	r	p	r	P
M +	-0,919	0,002	-0,64703	0,0413
M –	-0,349	0,358	-0,38653	0,17474
M+, M–	-0,665	0,015	-053267	0,01722

9. Globaler Test von Sekundärvariablen

Gegen Ende seines Manuskripts beschreibt der Autor einen sehr speziellen Vergleich zwischen M+ und M-. Dabei wurden eine Reihe von Sekundärvariablen in nicht näher beschriebener Weise jeweils aus den Originaldaten berechnet und zwar getrennt jeweils für M+ und M-. Aus all diesen Wertepaaren wurden dann diejenigen ausgewählt, bei denen mindestens bei einem Wert der Signifikanztest mit $p \le 0{,}05$ ausfiel. Diese Bedingung wurde von 20 Variablen erfüllt. Die entsprechenden Absolutwerte $|z_{M+}|$ und $|z_{M-}|$ wurden dem Mann-Whithney-U-Test zugeführt, wobei für M+ im Mittel größere Abweichungen vom Zufall (unabhängig von der Abweichungsrichtung) beobachtet wurden als bei M-. Dieser Test lieferte $z = 2{,}37$ und $p = 0{,}009$. Es wurde nicht erwähnt, ob p einseitig oder zweiseitig berechnet wurde, aber aus der Kombination dieser beiden Ergebniswerte folgt die Einseitigkeit des Tests.

Dieses Testvorgehen ist zulässig, sofern die Voraussetzung unabhängiger Daten nicht verletzt ist und die (im Prinzip willkürliche) Auswahlregel vorher festgelegt wurde. (Ob dieses Testvorgehen aber auch effizient ist, soll hier nicht diskutiert werden.)

Auch wenn die Sekundärvariablen ursprünglich alle voneinander unabhängig sein sollten, so wird trotzdem durch das hier verwendete Auswahlverfahren bei den ausgewählten Paaren z_{M+} und z_{M-} eine Abhängigkeit erzeugt. Zur Erläuterung nehmen wir an, dass die für die Auswahl berechneten p-Werte zweiseitig definiert sind. Dann ist $p \le 0{,}05$ gleichbedeutend mit $|z| \ge 1{,}96$. Gilt nun bei einem ausgewählten Wertepaar $|z_{M-}| \ge 1{,}96$, so kann für $|z_{M+}|$ jeder beliebige positive Wert auftreten. Gilt jedoch $|z_{M-}| < 1{,}96$, so muss $|z_{M+}| \ge 1{,}96$ gelten, da anderenfalls dieses Variablenpaar nicht ausgewählt worden wäre. Das bedeutet aber, dass bei ausgewählten Wertepaaren die statistische Verteilung von $|z_{M+}|$ von dem konkreten Wert von $|z_{M-}|$ abhängt. Insbesondere muss bei kleinem $|z_{M-}|$ ($|z_{M-}| < 1{,}96$) der Wert von $|z_{M+}|$ recht groß sein ($|z_{M+}| \ge 1{,}96$), womit also $|z_{M-}|$ und $|z_{M+}|$ negativ korreliert sind. Damit ist also die Voraussetzung des Mann-Whithney-Tests verletzt, womit der Wert p = 0,009 nur als ein vorläufiger Wert gelten kann. Um einen adjustierten p-Wert p_a zu ermitteln, erzeugte ich in einer Simulationsstudie mit 100000 Läufen jeweils 200 standardisiert

normalverteilte Variablenpaar z_{M-} und z_{M+} und wendete obige Auswahlregel an. Die zufällige Anzahl ausgewählter Paare hat dabei den Erwartungswert 19,5. Für 3,1 % der (roh-) p-Werte des jeweils angewendeten Mann-Whithney-Tests galt $p \leq 0,009$. Damit beträgt also der adjustierte p-Wert $p_a = 0,031$. Eine Bonferroni-Korrektur entfällt hier, da es sich nicht um einen konfirmatorischen Test handelt. Die Idee zur Durchführung eines solchen Tests kam dem Autor erst in späteren Versionen des Artikels, nachdem ihm entsprechende Besonderheiten bei den z-Werten der Sekundärvariablen auffielen. Damit kann dieser Test nur als exploratorisch betrachtet werden.

Oben zeigte ich, dass trotz der durch die Auswahlregel erzeugten Abhängigkeit das Ergebnis mit Hilfe einer Simulationsstudie korrigierbar ist, sofern die Ausgangsvariablen untereinander unabhängig sind. Wenn man jedoch bedenkt, dass hier eine Fülle von Sekundärvariablen als unterschiedliche Funktionen aus ein und demselben Datenmaterial berechnet wurden, so muss zunächst mit Abhängigkeiten gerechnet werden bzw. die Unabhängigkeit zwischen den Variablen wäre nachzuweisen. Mir lagen von einigen der Sekundärvariablen die Bezeichnungen vor, welche einen ungefähren Eindruck von der genauen Definition dieser Variablen vermitteln. Dabei konnte man zwischen einigen Variablen Abhängigkeiten zumindest erahnen und mitunter sogar zeigen. Besonders problematisch sind multiple Abhängigkeiten (zwischen mehr als zwei Variablen). Ihre Widerlegung ist sehr schwierig, da die Fülle der potenziell möglichen Variablenkombinationen kaum überschaubar ist. Diese Abhängigkeiten könnten aber das Ergebnis des Mann-Whithney-Tests eventuell stark verzerren.

Aber auch bei Vorhandensein dieser Abhängigkeiten wäre eine Korrektur des Mann-Whitney-Tests noch möglich, indem man die oben beschriebene Simulationsstudie durchführt, dabei aber anstelle der standardisiert normalverteilten Variablen die Sekundärvariablen simuliert. Genauer gesagt, es werden in jedem Simulationslauf die Originaldaten des gesamten Experiments – unter Annahme der Nullhypothese „kein Psi“ – simuliert und aus ihnen alle Sekundärvariablen berechnet.

In der Fußnote 21 betont Ertel, dass er diese Abhängigkeiten zwischen den Variablen für minimal hält. Er berief sich mir gegenüber auf seine Erfahrungen mit dem Ballzieh-Test und auf sein „Weltwissen“. Diese Informationsquellen sind hierfür aber zu ungenau, insbesondere im Hinblick auf multiple Abhängigkeiten. Außerdem, falls eine Psi-Wirkung vorliegen sollte, wie Ertel vermutet, so könnte sich das auch auf die Abhängigkeitsstruktur auswirken. Eine verlässliche Aussage über die Abhängigkeiten unter H_0 wäre dann anhand praktischer Erfahrung nicht möglich.

Auch die Auswahlregel wurde von Ertel nicht immer so verwendet, wie er sie im Artikel beschrieb. Es gibt Sekundärvariablen, welche für den linken und den rechten Beutel jeweils gesondert berechnet wurden (l und r). Hier kann man die Summe $l + r$ beider Variablen als eine weitere Variable definieren. Hierbei beschloss Ertel sinnvoller Weise, nur die Einzelvariablen r und l zu verwenden und nicht ihre Summe, da letztere von den ersteren abhängt. Von diesem Beschluss wich er aber ab, sofern r und $l + r$ signifikant waren, l jedoch zwar das gleiche Vorzeichen, aber keine Signifikanz aufwies. In diesem Fall führte er die Variablen r und $l + r$ dem Mann-Whitney-Test zu, womit er bewusst Abhängigkeiten zwischen Variablen (Korrelation = 0,5) in Kauf nahm.

Um zu prüfen, ob eine Sekundärvariable x signifikant ist, wurde der Wert $z = (x - \mu)/\sigma$ berechnet. Hierbei ist also die Kenntnis des Erwartungswertes μ und der Standardabweichung σ von x erforderlich. Bei den oft recht kompliziert definierten Sekundärvariablen ist die Bestimmung dieser Parameter durchaus nicht trivial. σ könnte auch durch die Stichprobenstandardabweichung s geschätzt werden, sofern man die Variable x für jeden Durchgang gesondert berechnet und aus diesen Einzelwerten dann s bestimmt. Für μ ist dieser Weg jedoch nicht möglich, da der Schätzwert von μ dann mit x identisch wäre, woraus stets $z = 0$ folgte. Es wäre daher interessant gewesen, wenn die nicht ganz einfache Bestimmung von μ zumindest an einem Beispiel demonstriert worden wäre.

10. Zum Postskriptum

Die Basis des Postskriptums, welches mir bei der Begutachtung des Artikels noch nicht vorlag, war der glückliche Umstand, dass MG sich nun doch noch zu weiteren Tests bereit erklärte. Somit hätten die Hypothesen, die sich bisher im exploratorischem Sinne als auffällig erwiesen, eventuell konfirmatorisch bestätigt werden können. Doch leider wurde eine andere Versuchsvariante durchgeführt – und neue, bisher nicht betrachtete Hypothesen getestet. Leider ist unbekannt, welche Hypothesen erst nach Ansehen der Daten gewählt wurden, so dass nicht klar ist, inwieweit die Ergebnisse nur als exploratorisch oder sogar als konfirmatorisch zu betrachten sind.

Interessant ist, dass die Heterogenitätshypothese mit jenem Test geprüft wurde, welcher von Ertel vorher in Fußnote 11 abgelehnt wurde. Das soll nicht heißen, dass dieser Test deswegen sinnlos sei, aber es ist zumindest eine gewisse Inkonsequenz. Wie ich unten den Punkten 5 und 6 ausführte, prüft dieser Test sowohl die Heterogenität als auch die mittlere Trefferquote. Aber für die konkrete Ausführung dieses Tests gibt es mehrere Varianten. Ertel schreibt leider nicht, welche Variante er wählte. Nach Auswertung mit vielen Varianten scheint mir, dass Ertel die Z-Werte als Normalverteilungsordinaten der anhand des aus der Binomialverteilung ermittelten p-Wertes berechnete. Zumindest war bei dieser Testvariante mein Ergebnis ($Chi^2 = 19.96$, $p=0.01$) dem Ergebnis von Ertel ($Chi^2 = 19.6$, $p=0.01$) am ähnlichsten. Bei diesem Test ist der p-Wert der einzelnen Trefferanzahlen die Wahrscheinlichkeit, dass man unter der Nullhypothese diese oder eine noch größere Trefferanzahl erhält. Wegen der Zweiseitigkeit bei dem Heterogenitätstest hätte man dann aber konsequenter Weise für Trefferanzahlen, die unter dem Erwartungswert (=6) liegen, den linksseitigen p-Wert berechnen sollen, d.h. die Wahrscheinlichkeit, dass man unter der Nullhypothese diese oder eine noch *kleinere* Trefferanzahl erhält. Dann wäre $Chi^2 = 14.096$ und $p = 0.079$. Diese Testvariante ist aber sehr konservativ. Ein Kompromiss besteht darin, die Z-Werte aus $Z = (\text{Treffer} - p_0 * n)/\sqrt{n \cdot p_0 \cdot (1 - p_0)}$ zu berechnen ($p_0=0.1$=Trefferwahrscheinlichkeit unter H_0, $n=60$). Man erhält damit $Chi^2 = 17.039$, $p=0.0297$.

Nehmen wir aber Ertels Absicht, nur die Heterogenität testen zu wollen, ernst, so müssten wir den üblichen Chi^2-Test anwenden (siehe Punkt 5). Dieser entspricht dem vorherigen Test, sofern wir für p_0 die aus den Daten ermittelte mittlere Trefferquote $\hat{p}_0$ verwenden. Da

$\hat{p}_0$ aber in diesem Falle mit der theoretischen Trefferquote p_0 übereinstimmte, erhalten wir den selben Wert Chi² = 17.039. Für die Freiheitsgrade ist aber nun der Wert 8-1=7 zu verwenden, womit wir p=0.017 erhalten.
Unter den vielen möglichen Charakteristika der Daten werden sich sicherlich auch welche finden, die sich als signifikant erweisen. So geschehen mit der Unausgewogenheit der Ziehhäufigkeit der einzelnen Ziffern und der Länge der Folgen mit gleichen Farben (Run-Test). Bei dem p-Wert des Tests für die unausgewogene Ziehhäufigkeit gab Ertel sogar einen Bonferroni-korrigierten p-Wert (p=0.005 und p_k=0.03) an. Bisher hatte Ertel zu konfirmatorisch gedachten Tests den p_k-Wert angegeben, so dass man indirekt erkennen konnte, welche Tests als konfirmatorisch gemeint waren. Ist also dieser Test hier nun als einziger konfirmatorisch gedacht? Aus dem Verhältnis von p und p_k folgt, dass dann 30 Tests geplant waren, von denen also nur einer signifikant war. Oder wurde bei den anderen Tests im Postskriptum nur die Angabe von p_k vergessen? Hätten die anderen p-Werte auch noch jeweils mit 30 multipliziert werden sollen?
Ob nun die unausgewogene Ziehhäufigkeit oder die langen Farben-Runs irgendetwas mit Psi-Fähigkeiten zu tun haben, ist eine andere Frage. Aber diese Frage möchte ich, wie zu Beginn geschrieben, den anderen Kommentatoren überlassen.

ROB NANNINGA & JAN WILLEM NIENHUYS[27]

Statistischer Irrsinn

An der 1998 von Ertel in *Correlation* veröffentlichten Untersuchung nahmen 11 Astrologen teil. Sie bekamen die Geburtsdaten von 40 Schotten, die je zur Hälfte entweder Politiker oder Maler waren. Die Astrologen hatten die Aufgabe, beide Teilgruppen zu unterscheiden, aber wie bei derartigen Studien mit angemessener Verblindung üblich, scheiterten sie.
Zwei Jahre später wurde Ertel dann von dem jungen Meteorologen MG kontaktiert, der Kurse in indischer Astrologie besucht hatte und zu der Überzeugung gelangt war, dass seine Horoskopdeutungen oft richtig lägen. Er bat Ertel um einen Test, woraufhin ihm dieser die Daten jener Schotten sandte, im Vertrauen darauf, dass MG nicht in *Correlation* nachsehe, wo die richtigen Antworten bereits veröffentlicht waren.
Nur bei 24 Horoskopen traf MG die geforderten Zuordnungen. Er scheiterte kläglich, weil er nur in 6 der 24 Fälle richtig lag. Ertel bemerkte nun das für ihn merkwürdige statistische Phänomen, dass MG „sehr signifikant" *schlechter* als die Zufallserwartung abgeschnitten hatte. Er erhielt hierfür seinen Wert p=0.011, indem er den p-Wert *einseitig* berechnete.[28] Ein ein-

[27] Rob Nanninga ist Redaktionsleiter der niederländischen Zeitschrift *Skepter*. Anschrift: Westerkade 20, NL-9718 AS Groningen. E-Mail: r.nanninga@wxs.nl. Dr. Jan Willem Nienhuys ist Mathematiker und Redakteur von *Skepter*. Anschrift: Dommelseweg 1A, NL-5581 VA Waalre. E-Mail: j.w.nienhuys@tue.nl.

[28] Dabei unterstellte er als Nullhypothese, dass MGs Entscheidungsprozess dem Modell Münzwurf folgte, also immer eine Chance von 50 % für die Auswahl z.B. des Berufs „Maler" gegeben war. Dies ist aber keineswegs selbstverständlich, weil MG nur einen Teil der Horoskope für sich zur Be-

seitiges Testen ist hier aber unsinnig. Man kann dies nur tun, wenn man eine klare Hypothese prüft. MG wollte wissen, ob seine astrologischen Überzeugungen richtig sind. Dann kann man nicht einfach hinterher vorgeben, es sei erwartet worden, dass er schlechter als der Zufall abschneide. Deshalb muss dieser p-Wert mit dem Faktor 2 multipliziert werden.

Um sein Handeln zu begründen, zitiert Ertel Kimmel (1957, S. 353): "Use the one-tailed test when results in the unpredicted direction will, under no conditions, be used to determine a course of behavior different in any way from that determined by no difference at all." Es ist kaum möglich, dies unzweideutiger zu formulieren. In diesem Fall bedeutet es: Prüfe nur dann einseitig auf signifikante Fehlzuordnungen, wenn selbst die genauesten Zuordnungen von Horoskopen zu Berufen zu keiner anderen Schlussfolgerung Anlass geben als jener: „Ein weiterer Beleg, dass Astrologie Unsinn ist; ich bedauere, MG, aber Ihre Astrologie funktioniert einfach nicht." Wir überlassen es dem Leser, zu beurteilen, wie Ertel sein einseitiges Testen zu rechtfertigen versuchen wird, vermutlich wird er unseren Einwand einfach ignorieren.

Für sich allein genommen mag es etwas Auffallendes sein, dass MG nur 6 Treffer erzielt hat. Aber dass von 12 geprüften Astrologen einer dabei ist, der einen Ausreißer nach oben oder unten produziert, ist alles andere als bemerkenswert. In diesem Fall trat der Ausreißer auf, nachdem die offizielle Untersuchung bereits abgeschlossen war. Hätte MG nur wenige Treffer mehr erzielt, hätten wir niemals etwas von ihm gehört.

Ertel lud MG dann zu einem zweiten Test ein, diesmal um 20 französische Politiker von 20 französischen Malern zu unterscheiden. Von 19 zu beurteilenden Horoskopen ordnete er 13 korrekt zu. Wieder führte Ertel einen einseitigen Test durch, seiner Philosophie verpflichtet, dass man einseitig testen sollte, wenn es anders ausgeht als erwartet. Er erhielt $p=0.08$ (statt zweiseitig $p=0.17$) und nannte dies „marginal signifikant".

Er wurde dann noch weiter angespornt, als er seine Aufmerksamkeit auf die Differenz zwischen diesen beiden Zufallsergebnissen richtete. Der Chi-Quadrat-Test ergab eine signifikante Differenz zwischen beiden Resultaten ($p=0.004$). Spätestens an dieser Stelle sollte klar geworden sein, dass Ertel statistische Instrumente nach seinem Belieben auswählt, nachdem er sich Ergebnisse bereits angesehen hat. Ursprünglich gab es nämlich keinen Grund für die Annahme einer Differenz zwischen beiden Resultaten, weil beide Experimente fast identisch waren. Es wäre vernünftiger gewesen, die Daten der beiden Experimemte einfach zu addieren, und wir vermuten, dass Ertel genau dies getan hätte, wenn MG im zweiten Experiment erneut eine niedrige Trefferquote erzielt hätte.

Wir müssen an dieser Stelle darauf hinweisen, dass die hinter der Berechung von p-Werten stehende offizielle Philosophie lautet, dass die Wahl des statistischen Tests ein Teil der Versuchsplanung zu sein hat. Wenn die Daten bereits vorliegen, ist es bedeutungslos, p-Werte durch statistische Verfahren zu berechnen, die erst ausgewählt wurden, nachdem man sich die Daten bereits angesehen hat.

MG kam zu der Schlussfolgerung, dass er Berufe nicht anhand von Horoskopen bestimmen könne, aber er vermutete stattdessen, dass er berühmte Personen von „gewöhnlichen" Leu-

---

arbeitung selektiert hatte und er versucht haben könnte, die Berufe „Maler" und „Politiker" mit gleicher Häufigkeit zu vergeben.

ten zu unterscheiden in der Lage sein könnte. Ertel führte einen dritten Test durch, bei dem MG 20 berühmte französische Schriftsteller von gewöhnlichen Franzosen zu unterscheiden hatte. Bei 12 von 23 Horoskopen tippte er richtig, genau entsprechend der Zufallserwartung. Für einen gewöhnlichen Menschen wie MG war klar, was diese Resultate bedeuten: Horoskope sind wertlos, um etwas über das Leben von Menschen herauszufinden. Aber Ertels Gedanken waren andere: Die Trefferquote war zunächst niedrig, dann hoch, dann durchschnittlich entsprechend der Zufallserwartung. Es konnte also eine hohe Varianz vermutet werden, und ein von ihm daraufhin durchgeführter Chi-Quadrat-Test sagte das auch: $p=0.015$. Nun vermutete Ertel etwas parapsychologisches und schlug vor, dass Astrologen aufgrund von parapsychischen Fähigkeiten manchmal erfolgreich sein könnten. Vielleicht sei MG ja paranormal begabt, weil auch Hellsichtige abwechselnd Psi-Missing- und Psi-Hittimg-Phasen durchlebten.

Um MGs außersinnliche Fähigkeiten zu prüfen, unterzog der ihn einem „Balltest". Wir übergehen hier die Einzelheiten dieses Tests und beschränken uns auf die Feststellung, dass MG zwei Bälle aus zwei Beuteln zu ziehen hatte, die mehr oder minder auf vier verschiedene Weisen einfachen Daten entsprechen konnten, die auf einem Blatt Papier notiert waren. MG führte diese Tests alleine bei sich zuhause durch, ohne jede Aufsicht, und er notierte die Resultate auch selbst. Dies erforderte einige Konzentrationskraft, weil man dabei leicht Fehler machen kann. In einigen Fällen führte er zunächst eine Transzendentale Meditation durch.

Die Ergebnisse waren eindeutig: keine parapsychischen Fähigkeiten. Für jede der vier Zuordnungsmöglichkeiten zu den vorher festgelegten Daten entsprach die Trefferquote in etwa der Zufallserwartung, mit oder ohne Meditation. Aber erneut unternahm es Ertel, wieder irgend etwas Signifikantes aus den Daten zu entnehmen. Sein forschender Geist bemerkte eine seltsame Art von Entsprechung, die er bis dahin nicht bedacht hatte, und in der Tat fand er einen „hochsignifikanten Effekt". Ertel gab zwar zu, 20 verschiedene Methoden der statistischen Analyse ausprobiert zu haben, und er korrigierte dies, indem er p mit dem Faktor 20 multiplizierte. Allerdings bezieht sich der Faktor 20 nur auf eine der grundlegenden Zuordnungsarten; wenn Ertel wirklich fair gewesen wäre, hätte er mit der Zahl der Analysen multiplizieren sollen, die er nach der Inspektion aller denkbaren Ergebnisse bereit gewesen wäre durchzuführen.

Leider sind wir nach unseren früheren Erfahrungen mit dem Autor zu der Einschätzung gelangt, dass es wenig sinnvoll zu sein scheint, mit Ertel zu diskutieren. Die Diskussionen sind endlos, Ertel scheint nie einzusehen, warum er kritisiert wird, und er gräbt nur immer weitere neue, weit hergeholte Spekulationen hervor. Der Kritiker wird sich früher oder später entscheiden, dass er wichtigere Dinge zu tun hat, woraufhin Ertel dann triumphierend schlussfolgert, er habe die Diskussion gewonnen. Zusammenfassend zeigt unseres Erachtens der zu kommentierende Artikel lediglich, wie man mit Statistiken alles beweisen kann, und dass Ertels Methoden denen des sog. „mexikanischen Scharfschützen" (Beck-Bornholdt & Dubben 1997, S. 38) entsprechen: Feuere deine Schüsse zuerst auf ein Scheunentor, um danach die Zielscheiben rund um die Einschlagslöcher aufzumalen.

Literatur

Beck-Bornholdt, H.-P., Dubben, H.-H. (1997): Der Hund der Eier legt. Rowohlt, Hamburg.

ULRIKE VOLTMER[29]

„Dem Einfallsreichtum forschungsbereiter Leser sind keine Grenzen gesetzt“

Der Autor betrachtet seine Testergebnisse als einen Beleg für die „Zusammenhangshypothese“ zwischen Astrologie und Psi, wie schon aus der Überschrift hervorgeht. Im Artikel werden zwei unterschiedliche Testreihen vorstellt, zum einen in Form eines astrologischen Zuordnungstests, zum anderen in Form eines „parapsychologischen Balltests“, den der Autor selbst kreiert hat. Ertels Intention zielt darauf ab zu zeigen, dass „astrologische Leistungen“ mit Psi-Begabung zu tun haben. Einmal abgesehen von der Frage, ob der „Balltest“ und die für ihn verwendete Auswertungsmethode der Kritik der Kollegen standhält oder ob der „junge Akademiker, der die Grundlagen indischer Astrologie erlernt hat“ und mit der „TM-Meditationstechnik“ vertraut war“, überhaupt ein auffälliges Ergebnis erzielt hat, richtet sich mein Augenmerk auf Ertels Schlusssatz, den ich als Titel meines Diskussionsbeitrags gewählt habe. Der Autor möchte offenbar dazu auffordern, dass sich Leser Tests unterziehen oder Tests kreieren, um herauszubekommen, ob bei dem einen oder anderen Astrologen Psi-Fähigkeiten vorliegen und sie zudem „auffälliger“ bei Horoskopzuordnungstests abschneiden. Der Aufforderung, über diese Hypothese nachzudenken und darauf kreativ und nicht ohne Humor zu reagieren, will ich gerne folgen.

Ertel selbst zeigt in seinem Artikel an keiner Stelle, wie die zwei von ihm vermuteten Leistungen oder besser „Auffälligkeiten“ aufeinander bezogen werden können. Inwiefern haben die Ergebnisse der Testreihen miteinander zu tun? Dazu hätte Ertel einen astrologischen Zuordnungstest ohne Deutungsbemühung des Probanden durchführen können. Zum Beispiel wäre es ein Leichtes gewesen, den Probanden aufzufordern, Geburtsdaten und Berufs-Kärtchen einander zuzuordnen.

Im Folgenden unterbreite ich einen Vorschlag und vereinfache diesen dann so weit, dass die Astrologie ganz außen vor bleibt, jedoch die „astrologische Motivation“ der Probanden im Zentrum der Tests steht. Am Schluss bleibt dann nur noch ein „parapsychologischer Balltest“ übrig oder gar ein Test, durch den geklärt werden könnte, ob durch eine astrologische Motivation überzufällige Ergebnisse zu erzielen sind:

Dazu möge Herr Ertel die Geburtsdaten von ausgewählten Personen auf je einen Ball schreiben und gebe diese Bälle mit den Geburtsdaten in den Sack A. Denn nehme er etwa fünf Berufssorten, bezeichne sie mit den Ziffern 1, 2, 3, 4 und 5, schreibe diese Ziffern auf die gleiche Anzahl Bälle und stecke diese in den anderen Sack B. Nun fordere man den Pro-

---

[29] Ulrike Voltmer ist Diplom-Psychologin und Musikwissenschaftlerin sowie Schriftführerin der Gesellschaft für Anomalistik. Anschrift: Metzer Str. 65, D-66117 Saarbrücken.
E-Mail: voltmer@anomalistik.de

banden auf, jeweils aus dem einen und dem anderen Sack zusammengehörige Bälle zu ziehen.

Das lässt sich noch vereinfachen: Man suche prominente Vertreter einer Berufsgruppe. Wer den gleichen Beruf hat, bekommt eine Ziffer, sagen wir 1 bis 5. Die Ziffern werden auf Bälle geschrieben und kommen in Sack A. In Sack B befinden sich die Bälle mit den Ziffern, die bestimmte Berufe vertreten. Es ist ratsam, dass die gleiche Ziffer in beiden Säcken auch die gleichen Berufe vertritt, dann kann man das erreichte Ergebnis beim Ballziehen leichter vergleichen. Ähnliche Berufe könnten zudem benachbarte Ziffern tragen, dann kann man bei der statistischen Auswertung noch eine ausgeklügeltere Methode verwenden. .

Oder noch einfacher: Man gibt dem Probanden zu den Ertel-Bällen in beiden Säcken eine Erklärung, die wie folgt lautet: „In mühevoller Arbeit habe ich die Geburtsdaten von prominenten Angehörigen bestimmter Berufsgruppen herausgesucht, ich habe Gauquelin-Daten benutzt, habe ansonsten noch weitere bestgesicherte Quellen eruiert. Dann habe ich jedem Horoskop, deren Horoskopeigner einem Beruf angehört, eine Nummer gegeben. Gleiche Ziffern stehen für gleiche Berufe. Auch im anderen Sack bedeuten die gleichen Ziffern gleiche Berufe. Bitte ziehen Sie richtige zueinandergehörige Bälle."

Interessant wäre es nun festzustellen, ob die Ergebnisse mit und ohne astrologische Motivation verschieden ausfallen. Um sicherstellen, dass die Astro-Motivation wirklich erhalten bleibt, sollte man bei den Ziffern der Bälle in Sack A vielleicht doch besser Geburtsdaten verwenden (diese können willkürlich gewählt sein, denn der Test hat ja mit Astrologie nichts zu tun). So könnte man herausbekommen, ob die Astro-Motivation den Psi-Faktor beeinflusst.

Allerdings ließe es sich noch raffinierter anstellen. Man könnte auch gleich testen, ob gar „die Astrologie" oder der Glaube daran den „Zufall" moderiert: Dazu müsste man bei einem weiteren Test die Daten von *tatsächlich vorhandenen* Personen, die einer Berufsgruppe angehören, benutzen. Es wäre sicher spannend zu überprüfen, ob die Ergebnisse verschieden ausfallen, wenn bei einer Testreihe die Bälle in Sack A echte Daten von bestimmten Berufsangehörigen tragen und bei einer anderen Testreihe die Bälle in Sack B mit ausgedachten willkürlichen Daten versehen sind.

Doch was ist, wenn „zufällig" einmal ein Datum gewählt wird, an dem wirklich die Geburt einer Person stattgefunden hat, die tatsächlich der unterstellten Berufsgruppe angehört? Denn der Versuchsleiter selbst könnte gar durch eine astrologische Motivation angesteckt worden sein, wodurch seine eigene Psi-Begabung wirksam wird, die zufällig passende Daten hervorbringt.

Wenn nach Belegen für die Zusammenhangshypothese zwischen Astrologie und Psi gesucht wird, dann sollte dies auch in der Versuchsplanung zum Ausdruck kommen. Dabei darf sicherlich auch der „astrologischen Motivation" ein gewisser Stellenwert zukommen (eventuell anstelle von Meditationsbemühungen). Kann jene als „psi-generierende" Variable nachgewiesen werden?

Der Autor antwortet

SUITBERT ERTEL

Kritik sollte korrigieren, nicht demolieren

1. Zu Emil Boller: Sind die verwendeten Tests nicht valide? [30]

1.1. Einwände gegen den astrologischen Zuordnungstest

Boller hält nicht für „plausibel", dass die stark schwankenden Trefferquoten, die MG beim Zuordnen von Horoskopen zu Berufen zeigte, auf Psi zurückzuführen seien. Psi habe kaum wirksam werden können, da MG die ihm mehrdeutigen Horoskope aus dem Aufgabensample ausschließen durfte. Gerade „uneindeutige Situationen" seien „psi-induktiv".
Antwort: Es gibt Gegenevidenz. Man kann die Probanden eines Balltests, der Psi-Effekte prüft, abwechselnd so instruieren, dass sie beim Zahlenraten eindeutige oder uneindeutige Voraussagen machen. Eindeutig ist die Voraussage unter der Standardbedingung (*eine* Zahl wird vorausgesagt), uneindeutig im Sinne von „unentschieden" ist sie, wenn man bei jedem Zug zwei Zahlen voraussagen lässt (z. B. „jetzt ziehe ich die 3 oder die 5"). Psi-begabte Probanden sollten nach Boller bei uneindeutigen Voraussagen („oder"-Bedingung) besser abschneiden als bei eindeutigen. Fünf psi-begabte Probanden absolvierten den Test unter beiden Bedingungen (siehe Tabelle 1). Bei drei Probanden sanken die Effektstärken der Treffer (siehe Effektdifferenzen in der letzten Spalte, hier sind *Psi-Koeffizienten* als Effektstärken zu verwenden, nach Timm 1971[31]). Die Probanden reagierten sehr stark auf die Bedingungsvariation, aber uneinheitlich. Tendenziell größer ist der Psi-Koeffizient bei eindeutigen Voraussagen (unter der „Nur-A"-Bedingung) im Vergleich zu den uneindeutigen Voraussagen (unter der „A-oder-B"-Bedingung)[32]. Zu Bollers Annahme würde passen, wenn der Koeffizient kleiner wäre.
Boller hält den Psi-Faktor als Ursache für MGs Trefferschwankungen auch deshalb nicht für plausibel, weil die geringe Zahl der Trials pro Run (24, 19, 23 Trials) eine erhöhte Variabilität als Artefakt erzeugen soll. Schwer verständlich ist dieses Argument, denn das Signifikanzkonzept, basierend z. B. auf Z-Werten einer Normalverteilung, berücksichtigt doch N,

---

[30] Ich habe mit Emil Boller über den Vorabdruck seiner Kritik einen E-mail-Austausch gehabt, um Missverständnissen vorzubeugen. Ich habe mich bemüht, seine mir mitgeteilten Antworten zu berücksichtigen, doch auf seinen Wunsch von E-mail-Zitaten abgesehen.

[31] Der Psi-Koeffizient (Thouless 1935; Thouless 1970; Timm 1971) berücksichtigt den Umstand, dass die Wahrscheinlichkeiten von Zufallstreffern in den beiden Bedingungen verschieden sind (.2 und .4). Effektstärken lassen sich auch durch Rosenthal & Rubins PI-Werte vereinheitlichen, dann sind die Effektdifferenzen zwischen den beiden Bedingungen ausgeprägter. Doch scheint die traditionelle Berechnung der Psi-Leistung hier ein adäquateres Modell zu sein.

[32] Vanyas (8 Jahre) wurde unter der Kontrolle seiner Mutter Tanya getestet. Alle anderen Daten wurden unter Heimbedingungen ohne Experimentator gewonnen. Bei Vanya, Tanya und Galina K. wurde im Herbst 2003 der Balltest auch unter fremder Kontrolle am IGPP in Freiburg durchgeführt, wo alle drei Probanden hochsignifikante Psi-Effekte zeigten. Ein Bericht darüber wurde bei der *Zeitschrift für Parapsychologie und Grenzgebiete der Psychologie* eingereicht.

die Zahl der Messwiederholungen. Eine größere Variabilität bei geringer Trialzahl, die Boller wahrscheinlich meint, findet man bei *Effektgrößen*, die bei einer geringen Zahl von Effektmessungen weniger zuverlässig sind und weiter streuen. Irrtumswahrscheinlichkeiten (p-Werte) aber bleiben davon unbeeinflusst.

Tabelle 1: Ergebnisse von drei psi-hochbegabten Probanden im Standardversuch unter der „Nur-A"-Bedingung (eine Zahl raten, MCE = .2) und unter der „A-oder-B"-Bedingung (MCE = .4).

	Eine Zahl raten (Standardbedingungg) MCE = .20					Zwei Zahlen raten („A oder B") MCE = .40					Differenz Psi-Koeff
	Trials	**Treffer**	**Treffer Proport**	**P Binom**	**Psi-Koeff**	**Trials**	**Treffer**	**Treffer Proport**	**P Binom**	**Psi-Koeff**	
Yaroslava	960	315	.328	$>10^{-15}$	.160	1820	515	.283	$>10^{-15}$	-.293	**.453**
Vanya	960	440	.458	$>10^{-15}$	.323	960	504	.525	10^{-14}	.208	**.115**
Galina K.	960	370	.385	$>10^{-15}$	.232	960	459	.478	10^{-6}	.130	**.102**
Tanya	1200	405	.338	$>10^{-15}$	.172	960	499	.520	10^{-13}	.200	**-.28**
Galina B.	960	232	.242	10^{-5}	.057	960	475	.495	10^{-8}	.158	**-.215**
Summe	5040	1762	.350	$>10^{-15}$	.187	5660	2452	.433	10^{-6}	.055	**.132**

1.2. Einwände gegen den Ballzieh-Test.

Boller hält auch den Ballzieh-Test für „mehr als fraglich", die Trefferüberhänge seien nicht überzeugend als Psi-Effekte nachgewiesen. Warum nicht? Er erwähnt einen eigenen Ballversuch, den er vor Jahren durchführte, bei dem er die Möglichkeit einer mnestischen Unterstützung beim Zahlenziehen prüfte. Es gelang ihm, nach gezieltem Zurücklegen der Bälle in den Beutel, deren Nummern und Positionen er sich zu merken versuchte (z.B. „alle Fünfen in die linke Ecke unten, alle Einsen in die rechte Ecke oben" oder so ähnlich), und nach offensichtlich unzureichendem senkrechtem Auf- und Abschütteln der Bälle (sie hopsen nur ein bisschen von ihrer Ausgangslage hoch und wieder zurück) eine erhöhte Trefferzahl der zurück gelegten Nummern zu erreichen. Boller ist der Meinung, dass auch unter Normalbedingungen, ohne systematisches Zurücklegen, keine Zufallsverteilung der Bälle zustande komme und dass allein aus diesem Grund der Test „ungeeignet" sei.

Als ich vor Jahren mit Herrn Boller darüber eine ausführliche Korrespondenz führte, die er in seinem Kommentar leider nicht erwähnt, gab ich zu bedenken, dass es erstens nicht darauf ankommt, vor jedem Trial jedem der 50 *Bälle* im Beutel eine zufällige Position zuzuteilen. Es ist lediglich notwendig, dass bei jedem Trial die 5 *Zahlen* mit gleicher Wahrscheinlichkeit erreichbar sind. Zwar kann man dies nicht mit Sicherheit voraussetzen, man kann es aber – zweitens – empirisch prüfen. Wenn eine mnestische Unterstützung von der Art, wie

sie Boller willkürlich herstellte, auch ohne Absicht unter Standardbedingungen auftritt – dies ist der Kern des Bollerschen Einwands –, dann müssten die dadurch bedingten Zusatztreffer bedingungsabhängig variieren:

1. Wenn die Probanden mnestische Unterstützung erhalten sollen durch ein vielleicht unbewusstes Sicherinnern an zurückgelegte Bälle und Zahlen, die nicht gut durchgemischt wurden, dann kann sich ein Treffervorteil nicht schon beim ersten Zug ergeben. Die Trefferquoten müssten „im Laufe eines Tests", innerhalb eines Runs also, ansteigen (Voraussage im Sinne Bollers). Da sich die Probanden hinsichtlich durchschnittlicher Trefferquoten erheblich unterscheiden, müsste man vor allem bei den erfolgreicheren Ballziehern eine Lernkurve erwarten. Tabelle 2 zeigt für die ersten 15 Trials eines Runs, dass die Trefferüberhänge *vom ersten Zug an* auftreten, auch steigen die Treffer über die 15 Trials nicht an. Die Serien der Trefferquoten der erfolgreicheren Probanden und die der weniger erfolgreichen sind beide gleich stationär. Die Trefferhäufigkeiten korrelieren mit ihren Positionen 1-60 r = -.14 (n.s.) bei Probanden mit weniger Treffern und r = -.13 (n.s.) bei Probanden mit mehr Treffern. Die Korrelationen haben die einer Lernhypothese entgegen stehende negative Richtung.[33]

Tabelle 2: Trefferhäufigkeiten beim Balltest (Einbeutel-Version) für Trials 1 bis 15, ein Run umfasst 60 Trials. Getrennt für Probanden mit geringeren Tref ferraten (Durchschnitt pro Run <13 Treffer) und solchen mit höheren Trefferraten (Durchschnitt pro Run 13 und mehr Treffer). MCE = 12 Treffer. N Probanden = 231. Ein Proband absolvierte 4 oder 6 Runs.

Probd.-	Trial-Folge														
Erfolg	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
geringer	132	114	133	117	126	119	126	132	117	136	128	124	116	104	125
höher	140	125	138	152	124	129	122	142	121	136	150	140	155	130	145
Summe	272	239	271	269	250	248	248	274	238	272	278	264	271	234	270

2. Wenn man Probanden anweist, wiederholt gleiche Zahlen zu ziehen, anstatt sie, wie im Standardverfahren, beim jeweils nächsten Zug eine Zahl frei raten zu lassen, wenn man z.B. die Eins 12 mal hintereinander, dann die Zwei 12 mal hintereinander bis zur Fünf, alle 12 mal hintereinander, als Zielzahl vorgibt (zusammen 60 Trials, ein Run), dann sollte dies einer unbewussten mnestischen Unterstützungsoperation sehr entgegen kommen. Das Unbewusste des Probanden kann sich dann zeitweise z.B. allein auf die linke Ecke im Beutel konzentrieren und dorthin alle Einsen zurücklegen, nachdem sie gezogen wurden (die Position der anderen zurück gelegten Zahlen brauchen bis zum Wechsel der zu ziehenden Serie nicht gemerkt zu werden). Unter dieser Bedingung mit ihrer mnestischen Erleichterung müssten mehr Treffer erzielt werden als unter der Standardbedingung. Umgekehrt müssten weniger Treffer als unter der Standardbedingung dann erzielt werden, wenn man die Probanden anweist, die Zahlen 1 2 3 4 5 1 2 3 4 5 in dieser Folge als Zielzahlen zu nehmen (oder in der Folge 5 4 3 2 1), wobei die zu ziehenden Zahlen also bei jedem Trial wechseln und so ein mnestisches Rehearsal erschwert wird. Hier könnte der Proband die Zahl, deren Position

[33] Die Daten wurden unter Heimbedingungen ohne Experimentator gewonnen.

seinem Unbewussten momentan wieder einfällt und die er dann mit erhöhter Trefferchance gleich raten möchte, nicht mehr frei und verzögerungslos wählen. Tabelle 3 zeigt, dass die Trefferquoten bei den Probanden, die den Test unter der Standard- und unter den drei Zusatzbedingungen durchführten, die von Boller zu erwartenden Unterschiede nicht aufweisen. Die 11111-Bedingung ergibt *nicht mehr* Treffer, die 12345- und die 54321-Bedingungen ergeben *nicht weniger* Treffer als die Standardbedingung. Auch diese auf Bollers Boden eingepflanzte Hypothese schlägt also keine Wurzeln.

Tabelle 3: Trefferzahlen unter der Standardbedingung (freie Wahl beim Zahlenra ten) und unter den drei Bedingungen vorgegebener zu ziehender Zah lenfolgen (11111, 12345, 54321). Trialsumme in allen Fällen 960. Die MCE beträgt bei den 5 Alternativen 192 Treffer. Der Treffer-Überhang beträgt beim MCE-level (192) 0%, 100% würde er bei doppelt so vielen Treffern betragen usw.

	Standard Freie Wahl		11111-Bedingung		12345-Bedingung		54321-Bedingung	
	Treffer	% Überhang	Treffer	% Überhang	Treffer	% Überhang	Treffer	% Überhang
Galina K.	370	92.7	342	78.1	343	78.6	387	101.6
Tanya	298	55.2	303	57.8	281	46.4	307	59.9
Galina B.	232	20.8	209	8.9	234	21.9	246	28.1
Summe	900	56.3	854	48.3	858	49.0	940	63.2

Fazit: „Gegenargumente werden *nicht in Erwägung gezogen*“ (kursiv hinzugefügt) – so beklagt sich Boller, der empirische Evidenzen, die ich gegenüber seinen Gegenargumenten vorgebracht habe, seit Jahren kennt und in seinem Kommentar *nicht in Erwägung zieht*.

Bollers Bedenken sind beachtenswert. Doch hat sich der Kritiker zu sehr darauf versteift zu fordern, dass alle Bälle im Beutel vor jedem Trial neu zufallsverteilt sein müssen. Seine Forderung mag einer Idealvorstellung entspringen, die, für sich genommen, platonisch attraktiv erscheinen mag, seine Analogie mit den Gasmolekülen besticht. Doch fehlt der Übertragung dieser Analogie als Modell für den Balltest jede empirische Rechtfertigung.[34] Hätte ich in

[34] Für methodisch unverzichtbar halten es Parapsychologen generell, dass alle denkbaren Möglichkeiten einer Beeinflussung der Trefferquoten durch Bias und Täuschung ausgeschlossen werden. Diese Forderung hat nach meiner Auffassung schädliche Folgen. So wie Spinnen-Phobiker durch ständiges Vermeiden von Kontakten mit Spinnen der emotionalen Erfahrung ausweichen, dass diese Tiere eigentlich harmlos sind, so werden Psi-Forscher, solange sie die Möglichkeit von Bias und Täuschung für ihre Probanden rigoros ausschließen, nicht lernen können, dass Bias und Täuschung bei ihren Probanden vielleicht gar nicht oder nur in seltenen Ausnahmefällen zum Einsatz kommen. Zudem signalisieren die „precautions“ mit ihrem technischen Aufwand den Probanden einen be-

den Untersuchungsreihen der ersten Stunde, die lange vor Bollers Einwänden durchgeführt wurden, Anzeichen für Lerneffekte, für Treffervorteile beim wiederholten Ziehen gleicher Zahlen usw. gefunden – ich habe danach gesucht – ich hätte die Weiterentwicklung des Balltests sofort eingestellt. Boller lehnt den Balltest trotz aller Entkräftung seiner spekulativen Einwände weiterhin ab.[35] Er lobt MG, was dessen Lernfähigkeit angesichts erwartungswidriger empirischer Beobachtungen betrifft. Es wäre zu seinem Vorteil, wenn er sich MG auch noch zum Vorbild nähme[36].

2. Zu Volker Guiard: Ist die verwendete Statistik nicht ausreichend?

Nein, die Statistik ist „mangelhaft", schlussfolgert Volker Guiard im Titel seines Kommentars, in welchem er über neun Seiten statistische Details ausbreitet. Dass seine frühere Diskussion mit mir „schwierig und langwierig" war, wie er sagt, werden die Leser bei der Lektüre seines Kommentars ermessen können, die sicher nicht weniger oft als ich die meisten Absätze wieder und wieder lesen müssen, um zu verstehen, worauf der Autor hinaus will. Vor allem fehlen Begründungen für die von ihm vergebene Schulnote. Auf meine Bitte, zu jedem der zehn Punkte seines Beitrags eine klärende Schlussfolgerung mit Bewertung des

---

trächtlichen Mangel an Vertrauen, was auf Psi hemmend wirken dürfte. Diagnostische Methoden in der Mainstream-Psychologie (Fragebögen, Intelligenztests usw.) werden durchweg ohne phobischen Ausschluss von Fälschungsmöglichkeiten eingesetzt. Der zusätzliche denkbare Einfluss eines „faking" wird, wie es sich gehört, im Verlauf einer Testentwicklung gelegentlich als gesonderte Frage eigens untersucht.

[35] Boller wendet en passant noch ein, dass Kontrolle durch einen Experimentator fehlt. Wenn sein Bedenken schwerwiegend wäre, müsste man auch sein eigenes Ball-Experiment mit einem Fragezeichen versehen, das er ohne Kontrolle durchführte. Doch man vertraut ihm. Auch kann man zur Entwarnung die günstigen Ergebnisse heranziehen, die bei einem Vergleich von Balltest-Ergebnissen unter Heim- und Laborbedingungen (N = 29) erzielt wurden und die auf der 47. Convention der Parapsychological Association in Wien von mir vorgestellt wurden (Ertel 2004).

[36] Boller wird die Validität des Ballziehens vielleicht anerkennen, wenn er vom Ergebnis des folgenden Experiments erfährt. Bei der Durchführung dieses Experiments wurde die Möglichkeit mnestischer Unterstützung ausgeschlossen, da die Bälle nach dem Ziehen nicht wieder in den Beutel zurückgelegt wurden. Eine hoch psi-begabte Probandin (Katarina), die zuhause ohne Kontrolle den Zweibeutelversuch nach dem Standardverfahren, d.h. mit Zurücklegen der Bälle, durchgeführt hatte (960 Trials), führte auch 480 Trials ohne Zurücklegen durch. Mit Zurücklegen der Bälle betrugen die Trefferüberhänge beim Beutel links 92.7% und rechts 70.8%. Ohne Zurücklegen der Bälle waren die Trefferüberhänge insgesamt ähnlich hoch (118.7% links, 55.2% rechts). Einflüsse der „Closed-deck"-Anordnung auf die Trefferquoten (entspricht hier der Anordnung „ohne Zurücklegen"), die die statistische Auswertung modifizieren könnten, sind bei der hohen Targetzahl pro Run (60) zu vernachlässigen (Burdick & Kelly, 1977, S. 88-91). Da das von Boller vermutete systematische Zurücklegen und unzureichende Schütteln in Katarinas Versuch *ohne Zurücklegen* nicht auftreten konnte, hätten die Trefferquoten gegenüber der Standardbedingung sinken müssen – sofern trefferssteigernde Effekte im Standardversuch (mit Zurücklegen) tatsächlich existierten. Die Trefferquoten sind aber nicht gesunken – ein weiterer Beleg dafür, dass sich die Einwände Bollers empirisch nicht rechtfertigen lassen.

jeweils lokal Behandelten hinzu zu fügen, wenn auch nur in einem Satz, ging er nicht ein. So machte *ich* mich auf die Suche nach Guiards Bewertungsgründen und fand:

1. „Die Bonferroni-Korrektur wurde nicht immer verwendet"

... meint Guiard wohl unter Punkt 1, wenn ich ihn recht verstehe. Doch für den einzigen Fall, den er monieren zu müssen glaubt, schreibt er: „Diese Tests gehören aber zu den konfirmatorischen Tests, ein Verzicht auf die Bonferroni-Korrektur ist daher nicht gerechtfertigt." Ein Flüchtigkeitsfehler in der Formulierung? Wenn nein, gibt Guiard mir Recht, wenn ja, fehlt die Begründung.
Fazit: In Guiards Punkt 1 erkenne ich keinen Grund für eine Abwertung der verwendeten Statistik.

2. „Ertels Analogie mit der Split-half-Korrelation ist unzutreffend"

Hier unterlief Guiard offenbar ein zweiter Flüchtigkeitsfehler. Ich schrieb nicht, was Guiard zitiert: „Letzteres ist einem split-half Korrelationstest analog, bei dem man nach Vorliegen eines signifikanten Effekts bei der einen Testhälfte die andere gezielt auf Effekt*resonanz* [!] prüft". Ich schrieb „Effekt*konstanz*" und meinte, dass wenn man eine erste Hälfte der Daten exploratorisch (*mit* Bonferroni-Korrektur) geprüft hat, man die zweite Hälfte konfirmatorisch (*ohne* Korrektur) prüfen darf. Wenn es sich im ersten Teil trotz Bonferroni-korrigierter Signifikanz um einen *statistical fluke* handeln sollte, wird eine Prüfung auf Konstanz den Zufall, der schon bei den Daten des ersten Signifikanztests vorlag, reproduzieren. Zur Überprüfung der Konstanz finden keine multiplen Tests mehr statt, der Signifikanztest für die zweite Datenhälfte ist kein exploratorischer mehr.
Fazit: Da Guiard praktisch dasselbe mit anderen Worten sagt, finde ich in seinem Punkt 2 ebenfalls kein Argument für die Wahl seines Titels.

3. „Ertels Anwendung des Binomialtests bei der Analyse der astrologischen Zuordnungsdaten war o. k."

Durch gesonderte Prüfung bestätigt Guiard die Richtigkeit meiner intuitiv getroffenen Entscheidung.
Fazit: Auch Punkt 3 hat kein negatives Bewertungsvorzeichen.

4. „Ertels Anwendung eines einseitigen Tests hat zum Ergebnis 1-p, nicht p, wenn die Abweichungsrichtung erwartungswidrig ist"

Meine Aussage, dass die klägliche Trefferzahl 6 bei 24 astrologischen Zuordnungen mit $p = .0113$ nach dem Binomialtest signifikant sei, treffe nicht zu, behauptet Guiard. Die Zufallswahrscheinlichkeit sei $1\text{-}p = .9887$, und dieser Wert interessiere nicht.
Entgegnung (hier notgedrungen etwas ausführlicher): Wenn man eine Alternativhypothese prüft wie „MGs Astrologie verhilft ihm zu überzufällig vielen Treffern", dann ist ein einseitiger Test geboten. Er ist nicht nur geboten, wenn der Proband dann tatsächlich überzufällig viele Treffer zeigt (Bestätigung der Alternativhypothese) oder wenn seine Trefferzahl aus dem Zufallsbereich nicht heraus fällt (Akzeptieren der Nullhypothese). Der einseitige Test ist auch bei überzufälligem *Treffermangel* einzusetzen. Grund: Der Forscher sollte das Ergebnis mindestens für ebenso bedeutsam halten wie ein signifikantes Ergebnis in der erwarteten

Richtung. Mehr noch als bei einem signifikanten Test mit bestätigter Richtung sollten Konsequenzen gezogen werden. Nicht nur ist ja im vorliegenden Fall durch Signifikanz in der Gegenrichtung die Ausgangshypothese der astrologischen Fähigkeit bei MG nicht bestätigt worden. Vielmehr scheinen sich bei MGs Zuordnungstätigkeit ganz andere, unerwartete Faktoren durchgesetzt zu haben, die so stark sind, dass sie sich, sogar wenn man mit ihnen gerechnet und sie deshalb einseitig geprüft hätte, als sehr signifikant erwiesen hätten. Mit anderen Worten: *Signifikant* heißt *bedeutsam*, für den Forschungsprozess *sehr* bedeutsam sind auch und vor allem richtungskonträre signifikante Effekte. Sie veranlassen den Forscher nicht nur, z.B. nach ungünstigen Versuchsbedingungen zu suchen, die der zu prüfenden statistischen Ausgangshypothese abträglich gewesen sein können. Das tut man bei nichtsignifikanten Ergebnissen. Ein signifikantes Ergebnis in der Gegenrichtung dagegen führt zu tiefgreifenderen Überlegungen: Stimmen denn die inhaltlichen Voraussetzungen der Ausgangshypothese überhaupt, sind diese vielleicht umzuwerfen und durch ganz andere zu ersetzen? Man kann auf eine Fährte gestoßen sein, die vielleicht zu einer Entdeckung führt. Es handelt sich um das Gegenteil von dem, was Guiard als „uninteressant" (nicht bedeutsam) abtun will.

Eine einseitige Signifikanz in nicht erwarteter Richtung durch *zweiseitige Prüfung* abzuschwächen, wie das Nanninga & Nienhuys tun (siehe weiter unten), verbietet sich, denn weder hat man von Anfang an erwartet, dass astrologische Fähigkeit mal trefferfördernd, mal treffervermindernd wirken könnte (zweiseitige Alternativhypothese), noch rechnet man damit, nachdem das erste erwartungswidrige Ergebnis vorliegt. Die Ausgangsthese der astrologischen Deutungsfähigkeit ist durch das richtungsfalsche signifikante Ergebnis indiskutabel geworden.

Ein „krasser Widerspruch zu der Formulierung von Kimmel", wie Guiard meint (so auch Nanninga & Nienhuis später), liegt keineswegs vor, im Gegenteil. Kimmels Formulierung besagt, dass man bei Vorliegen einer großen Abweichung in der Gegenrichtung dann einseitig prüfen darf, wenn man „unter keinen Umständen" die Ausgangshypothese (im vorliegenden Fall die These einer astrologischen Fähigkeit bei MG), etwa durch Umdeutungen, zu retten versucht. MG hatte zunächst einen solchen Versuch gemacht mit der Zusatzhypothese einer Adlerianischen „Überkompensation", der eine „Minderwertigkeit" zugrunde liegt, welche durch MGs astrologische Deutungsfähigkeit ans Tageslicht gekommen sein könne. Natürlich habe ich solche Rettungsversuche nicht unternommen, der Kimmelschen Forderung entsprechend.

Fazit: Was unter Punkt 4 abgehandelt wurde, ist kein Fehler. Aufseiten meines Kritikers wäre ein Umlernen wünschenswert.

4. „Ertel hat die Heterogenität von MGs drei Zuordnungsergebnissen zwar richtig geprüft, dabei aber manches übersehen"

Gegen die primäre Überprüfung der großen Streuung der Trefferquoten mit Hilfe des Chi-Quadrat-Tests, die MG beim astrologischen Zuordnen zeigte, hat Guiard nichts einzuwenden ($p = .0015$). „Korrekt", sagt er. Seine weiteren Ausführungen betreffen zwei andere Signifikanztests, auf die ich in meinem Artikel nur deshalb auch noch zu sprechen komme, weil Guiard in der früheren Korrespondenz den einen zusätzlich angewandt sehen wollte

(Summe Z^2-Test, vgl. Timm 1983, Formel 5), den anderen (Timm 1983, Formel 6) brachte ich daraufhin als mögliche weitere Alternative ins Spiel.
Die Einschränkung, die Guiard bei dem von ihm empfohlenen Summe Z^2-Test vornimmt (wenn das Mittel der Treffer-Z-Werte nicht Null ist, dann ist das Summe Z^2-Verfahren kein reiner Heterogenitätstest) ist zwar richtig, im vorliegenden Fall aber irrelevant, denn das mittlere Treffer-Z der MG-Daten liegt bei Null. Doch Guiards Behandlung dieses Punktes veranlassen den Leser, der ohnehin die Details nicht so schnell nachvollziehen kann, sie als Hinweis auf einen Fehler in der mit „mangelhaft" bewerteten Statistik des Verfassers zu interpretieren.
Fazit: Was Guiard in Punkt 5 an Irrelevantem vorbringt, ist nicht als Minuspunkt zu werten.

6. „Ertel hätte die Gesamtheit der Effekte bei der astrologischen Zuordnung anders prüfen sollen"

Es handelt sich hier um eine Wiederaufnahme von Punkt 5, sein „korrekt"-Urteil war Guiard anscheinend zu viel des Guten. Er meint, ich hätte nicht nur die Heterogenität der Trefferquoten, sondern gleichzeitig auch die Richtung der mittleren Zufallsabweichung der Treffer prüfen sollen. „Ihr galt sogar das primäre Interesse". Der Summe-Z^2-Test also (Formel 5) oder der Test der Formel 6 seien anzuwenden, die resultierenden p-Werte sind $p=.069$ und $p=.084$.
Erwiderung: Mein Interesse beim ersten astrologischen Test galt möglichen Abweichungen der Trefferzahlen von der Zufallserwartung, man kann dies mein „primäres" Interesse nennen. Doch dieses Interesse war aufgrund der Zuordnungsergebnisse von MG aufzugeben. Stattdessen war ich, unterstützt durch den neuen Gedanken, dass bidirektionale Psi-Effekte im Spiele sein könnten, nur noch an der großen Streuung der Abweichungen von der MCE, an der „Heterogenität" interessiert. Es ist falsch, zu behaupten, ich sei beim Einsatz des Prüfverfahrens meinem „primären" Interesse gefolgt und hätte nicht das diesem Interesse entsprechende Verfahren gewählt. Mit dieser für nicht eingeweihte Leser unauffälligen Unterstellung erschleicht sich Guiard die scheinbare Legitimation zur Anwendung des Summe-Z^2-Tests, mit dem Heterogenität plus Abweichung der Trefferzahl vom MCE gleichzeitig geprüft werden. Den Zweck dieser Maßnahme von Guiard kann man am Ergebnis ablesen: Die von mir berichtete Signifikanz der Prüfung auf Heterogenität allein, die $p=.0015$ betrug, wird durch den Summe-Z^2-Test aus dem Signifikanzbereich hinausbefördert: $p=.069$ bzw. $p=.084$. Die Unwahrheit der Unterstellung, auf der diese tendenziöse Datenanalyse aufbaut, sollte dem Kommentator selbst spätestens bei Behandlung seines Punktes 10 aufgefallen sein. Denn dort sagt er auf einmal ganz richtig, zur Begründung einer anderen Kritik, dass es mir doch „hauptsächlich" „auf die Heterogenität ankomme".
Fazit: Tendenziös konstruierte Gründe zur Abwertung der verwendeten Statistik zählen nicht.

7. „Ertels Analyse des Sondereffekts bei den Aufgabenzahlen ist in Ordnung"

Guiard vollzieht mein Vorgehen im Detail nach, findet offenbar nichts Fehlerhaftes, wendet zusätzlich einen eigenen Signifikanztest an und bestätigt das Hauptergebnis.
Fazit: Ein Argument für eine schlechte Schulnote ist auch unter Punkt 7 nicht zu finden.

8. „Ertels Signifikanz für Korrelationen der Zweibeutel-Trefferquoten wird mit einem korrekteren Verfahren im wesentlichen reproduziert"

Guiard fand, dass die von mir berechneten negativen Korrelationen der Trefferquoten für die beiden Beutel mit ihrem gegensinnigen Auf und Ab von Run zu Run nicht ganz stimmen können, da ich die Pasch-Fälle (Treffer gleichzeitig links und rechts) ausgelassen hatte. Dies erschien mir zwar zunächst sinnvoll, denn wenn z. B. nur Paschs vorgekommen wären – das wäre ein Psi-Effekt einer anderen Kategorie – , dann würde sich eine Korrelation = 1 ergeben. Das wäre ein Artefakt, dessen Beitrag zur Korrelation der Trefferhäufigkeiten wollte ich ausschließen. Doch das Ausweichen vor einem Artefakt wurde von mir unbemerkt durch ein Artefakt mit Gegenrichtung erkauft, was Guiard durch Neuberechnung (mithilfe von Simulationen) korrigiert hat. Die Signifikanz der Korrelation für M+ (Meditationsbedingung) veränderte sich von p=.002 auf p=.04, die der Korrelation für beide Bedingungen zusammen (M+ und M–) veränderte sich von .015 auf .017.
Fazit: Guiard hat einen tückischen Denkfehler bei mir aufgedeckt, seine Korrektur führt indessen nur zu numerischen Abschwächungen der p-Werte und nicht etwa auch zu Veränderungen an den Schlussfolgerungen, die aus dem aus dem Ergebnis gezogen wurden.

9. „Ertel hat anzunehmende Abhängigkeiten zwischen Variablen nicht berücksichtigt und bei der Anwendung eines Mann-Whitney-Tests die notwendige Voraussetzung nicht durchweg eingehalten"

Den Effekt der Meditationsbedingung (M+) habe ich u. a. auch für eine Vielzahl von Variablen gleichzeitig, ohne auf diese einzeln eingehen zu müssen, durch einen Mann-Whitney-Test-Vergleich zwischen der M+ und M–Bedingung prüfen wollen. Variablen, die offensichtlich nicht unabhängig voneinander sind, habe ich dabei vorschriftsmäßig nicht verwendet. Damit gab sich Guiard nicht zufrieden. Er vermutete unerkannt gebliebene Abhängigkeiten, deren Einfluss, wenn er vorliegen sollte, ich selbst für geringfügig hielt (Fußnote 19 im Artikel). Wegen der Komplexität der Auswertungen beim Zweibeutel-Test hätte ich, um die Berechnungen für Leser voll nachvollziehbar zu machen, umfangreiche Erläuterungen geben müssen, was mir bei der untergeordneten Bedeutung dieser Zusatzauswertung nicht angemessen erschien. Zudem hoffte ich, dass man darauf vertrauen würde, dass bei der Anwendung des Mann-Whitney-Tests keine schwerwiegenden Fehler gemacht würden. Ich gebe zu, dass Guiard mit Akribie Gründe vorgebracht hat, die, wenn sie alle zutreffen, beim Mann-Whitney-Test zur Abschwächung des berichteten p-Wertes der M+-Effekte führen würden.
Wie schwer würden diese artifiziellen Beiträge im Kontext der Untersuchung wiegen? Die Schlussfolgerungen, die aus unabhängigen p-Werten anderer Signifikanztests gezogen wurden, werden dadurch nicht tangiert. Darüber hinaus lassen sich noch nicht berichtete Ergebnisse des Summe-Z^2-Tests hinzufügen, dessen Anwendung Guiard bevorzugt.[37] Er hätte den Summe-Z^2-Test auch auf die Daten der Tabelle 2 meines Artikels anwenden sollen und hätte die Ergebnisse der nachfolgenden Tabelle 4 erhalten (oberer Teil Zweibeutel-Ergebnisse). Dazu hätte er anmerken können: Nach Anwendung des von mir bevorzugten Signifikanztests sind die Trefferquoten unter der Bedingung mit Meditation signifikant vom

[37] Verwendet wurde die von Guiard angegebene Formel für Z-Werte.

Zufall verschieden (p=.04 für den linken, p=.004 für den rechten Beutel, p=.001 für beide zusammen). Unter der Bedingung ohne Meditation ist die Abweichung nicht signifikant."

Tabelle 4: Ergebnisse der Summe-Z^2-Signifikanzprüfungen für den Zweibeuteltest (Version #1) unter variierenden Bedingungen sowie für den Einbeuteltest (Version #2). df = Zahl der Runs mit je 60 Trials. [38]

Versions-Nummer	Testversion	Beutel	Bedingung	df	Summe Z^2	P	df	Summe-Z^2	p
1	Zweibeutel	Links	Mit Meditation	8	22.5	.004 **	16	38.4	.001 **
1	Zweibeutel	Rechts	Mit Meditation	8	15.9	.04 *			
1	Zweibeutel	Links	Ohne Meditation	8	9.0	.34	16	20.4	.202
1	Zweibeutel	Rechts	Ohne Meditation	8	11.4	.18			
1	Zweibeutel	Summiert über Bedingungen und Beutel					32	59.3	.0026 **
2	Einbeutel		Ohne Meditation	8	17.0	.029 *			
1+2	Beide Tests			40	75.9	.0005 **	* signifikant **sehr signifikant		

Fazit: Guiard hat von mir unterschätzte Abhängigkeiten unter den Variablen bei meiner Anwendung des Zusatztests teilweise aufgezeigt, teilweise für denkbar gehalten. Da eine Anwendung des Mann-Whitney-Tests, wenn sie gegen Kritik in allen Details abgesichert sein will, im vorliegenden Fall mit riesigem Arbeitsaufwand verbunden wäre (siehe Guiards Simulationen), da der Test im übrigen nur Zusatzinformation liefert, mit seiner unvollständigen Darstellung aber Angriffsflächen bietet, hätte ich auf ihn verzichten sollen. Die übrigen vorliegenden und aus den tabellierten Daten ohne Aufwand ableitbaren statistischen Bekräftigungen des Verdachts von Psi-Effekten in MGs Zweibeutel-Test, auf die es letztlich ankommt, genügen. Aber die waren für Guiard kein Thema.

10. „Ertel hat die in den Einbeutel-Daten von MG signifikante Heterogenität, die er dort bestätigt fand, überschätzt"

Die Überschätzung ist darauf zurückzuführen, dass beim Summe-Z^2-Verfahren dann, wenn man anschließend an wiederholte explorative Beobachtungen mit sowohl Plus- als auch Minus-Abweichungen vom Zufall rechnet (wie hier bei MG), die Z-Werte nicht zweiseitig festgelegt wurden. Die Konsequenzen einer Nichtbeachtung dieser Regel sind jedoch kaum erheblich. Anstatt p=.01 beim Einbeutel-Test ergibt sich p=.029 (siehe Tabelle 4). Bei den Ergebnissen mit dem *Zweibeutel-Test* in Tabelle 4 sind die p-Wert-Unterschiede noch gering-

[38] Auf die Ergebnisse des Einbeutel-Verfahrens wird unter (10) einzugehen sein.

fügiger, die Klassifizierung der Ergebnisse als signifikant bzw. sehr signifikant sind bei einseitiger und zweiseitiger Z-Wert-Bestimmung praktisch gleich.
Guiard meint nun aber, ich hätte so wie bei der Analyse der Daten der astrologischen Zuordnung das Chi²-Verfahren verwenden sollen, das die Heterogenität allein prüft, worauf es mir ja hauptsächlich ankäme (jetzt schätzt er meine Absicht auf einmal richtig ein!). Dass ich bei der Analyse der Daten des Einbeuteltests sein präferiertes Verfahren anwende, sei „inkonsequent".
Der Grund für den Verfahrenswechsel: Ich hielt wegen des Vorkommens zweier beobachteter Häufigkeiten unter 5 (die Häufigkeiten 1 und 2 kommen vor) das Chi²-Verfahren nicht für anwendbar. Tatsächlich aber ist es laut Lehrbuch, das ich jetzt konsultierte, bei dem hohen N und df der vorliegenden Untersuchung auch anwendbar.
Eine legitime Anwendung des Chi²-Verfahrens, das allein die Heterogenität prüft, ist mir natürlich auch für diesen Fall sehr recht, und Guiards Angaben lassen erkennen, dass das Ergebnis signifikanter wird. Die Prüfung auf Heterogenität allein bringt den p-Wert von .029 auf .017.[39] Doch Tabelle 4 enthält die von Guiard bevorzugten Summe-Z²-Ergebnisse. Eine Anwendung des Chi²-Verfahrens (Heterogenität allein) auf *alle* Daten der Tabelle 4 verbessert indessen die dort eingetragenen Signifikanz-Werte sonst nur wenig. Zum Beispiel erreicht man bei den Zweibeutel-Daten *links plus rechts* für Heterogenität allein p=.0024, zusammen mit der Effektrichtung (ermittelt durch Summe-Z²) ergibt sich, wie tabelliert, p=.0026.
Ansonsten stellt Guiard bei diesem Punkt nur Fragen, die er besser per Email mit mir hätte klären können. Der Leser wird durch durchschimmernde Verdachtsäußerungen vom hypothesenbestätigenden Ergebnis abgelenkt.
Fazit: An der Beobachtung, dass MG in einem neuen Ballversuch (Einbeutel-Test) die „Heterogenität" seines früheren Trefferverhaltens reproduzierte, die er zuerst beim astrologischen Zuordnen, dann beim Ballversuch mit zwei Beuteln zeigte, hat sich durch Guiards Beseitigen eines Haars in der Suppe nichts geändert.

Schlussfazit zu Guiard: Guiards fachmännisches Aufspüren von Unkorrektheiten bei meiner statistischen Arbeit ist mir willkommen. Immerhin hat er bei drei der zehn von ihm aufgeführten Punkte Unkorrektheiten ausfindig gemacht. Sie waren erfreulicherweise geringfügig und konsequenzenlos. Doch ist die bei ihm durchschlagende Tendenz abzulehnen, die sich in den griffigen zwei Worten seines Titels („Statistik mangelhaft") äußert und darin, dass er seinen langen Ausführungen keine Schlussfolgerungen beifügt. Das Schlussfolgern überlässt er den Lesern, sie werden und sollen offenbar seine zehn Punkte als zehn Gründe für seine vernichtende Benotung verstehen. Guiard drückt damit dem verantwortlichen Forscher von „Astrologie und Psi" und damit auch seiner Thematik in unverantwortlich tendenziöser Weise, mit einem vollen Paket fadenscheiniger Gründe, einen üblen Stempel auf.
Dabei könnte der Einsatz der statistischen Kompetenz von Volker Guiard produktiv sein, wenn er seine Aufgabe nur darin sähe, andere Forscher bei ihren Problemlösungsversuchen

[39] Der publizierte Wert p=.01, bei dem Heterogenität plus Effektrichtung berücksichtigt waren, der Wert aber überschätzt wurde, liegt in der Nähe von .017, der am Ende der Überarbeitung als beste Schätzung anzusehen ist.

zu unterstützen. Vielleicht sind es grundlegendere Weltbild-Commitments, die es Guiard schwer machen, seine Fähigkeiten auch bei der Erforschung anomalistischer Phänomene mit der Neutralität eines guten Wissenschaftlers einzusetzen.[40]

3. Zu Nanninga und Nienhuys: „Statistischer Irrsinn"?

Die Betitelung des Kommentars von Nanninga und Nienhuys („Statistischer Irrsinn") stellt den Titel Guiards, der sich mit einer schulmeisterlichen Benotung begnügte, in den Schatten: Die Autoren geraten mit einer solchen Etikettierung meiner Untersuchung ins Umfeld psychiatrischer Diagnosen.

1. Nanninga und Nienhuys halten es für „unsinnig", dass ich beim ersten astrologischen Zuordnungstest von MG, der eine geringe Trefferquote hatte, den p-Wert einseitig berechnete. Man könne „nicht einfach hinterher vorgeben, es sei erwartet worden, dass er schlechter als der Zufall abschneide". Ich hätte den p-Wert zweiseitig berechnen müssen. Antwort: Ich habe mit meiner Verwendung des einseitigen Tests meine anfängliche Absicht und Erwartung nicht geändert (siehe meine Replik zu Guiard, Punkt 4).

2. Einseitig werde nur getestet, sagen Nanninga und Nienhuys, „wenn man eine klare Hypothese prüft", man dürfe die Hypothese nach Inspektion der Daten nicht ändern. Antwort: Damit beschreiben sie *mein* Vorgehen recht gut, die Hypothese war klar und wurde nach Inspektion der Daten nicht geändert. Nanninga und Nienhuys leisten sich hier ein Eigentor: *Nach Inspektion der Daten* wollen sie die primäre Alternativhypothese („MGs Astrologie verhilft ihm zu Zuordnungstreffern") zurückziehen und zur Rechtfertigung eines zweiseitigen Testens eine nun wirklich „unsinnige" Erwartung einschmuggeln: Denn jetzt tun sie so, als besage die Alternativhypothese, dass MG aufgrund seiner astrologischen Kenntnisse mal einen auffälligen Überschuss, mal einen auffälligen Mangel an Zuordnungstreffern hervorrufen könnte.

3. Die Autoren tun sich wie Guiard schwer damit, ein nicht erwartetes signifikantes Ergebnis interessant zu finden, das man weiter erforschen sollte. Mit Ironie werten sie meine vorsichtigen Tastversuche ab, die ich nach Vorliegen von Ausreißer-Ergebnissen unternahm, um sie aufzuklären: „Sein forschender Geist bemerkte ..."; er unternahm es, „wieder irgend etwas Signifikantes aus den Daten zu entnehmen" usw. [41] Antwort: Die Strategie eines allseitigen Inspizierens von Daten, das über das hinausgeht, was durch geplante Nullhypothesen einge-

---

[40] Nach seiner Homepage-Information stellt sich Guiard als Mitglied der „Skeptiker"-Organisation GWUP dar, deren Auffassungen er wohl im wesentlichen teilt. Man findet bei Skeptikerorganisationen generell kaum eigene vorurteilsfreie Forschung, wohl aber eine offene Kampfansage an die Forschung derjenigen, die mehr für möglich halten als sie selbst. Guiards Kommentar über „Astrologie und Psi" kann bei der GWUP sicher mit Applaus rechnen, ebenso wie beim holländischen Gegenstück „Stichting SKEPSIS", eine Organisation, in der Nanninga und Nienhuys, die den anschließenden Kommentar schrieben, exponierte Positionen innehaben.

[41] Im schon veröffentlichten holländischen Original des vorliegenden Kommentars (Nanninga 2004) erspart sich der Autor sogar eine ironische Einkleidung seiner Absicht nicht und spricht geradewegs von „de dwaze professor" („der törichte/beschränkte/schwachköpfige Professor"). Seine mit ad hominems garnierte Kritik erschien im *Skepter*, den Nanninga selbst herausgibt.

grenzt wurde, ist Nanninga und Nienhuys offenbar fremd. Von der „Philosophie" der Explorativen Datenanalyse, EDA (Tukey), die einen unbeschwert offenen und lernbereiten Umgang mit gesammelten Daten propagiert, scheinen sie nichts zu halten, sie schreiben: „Wenn die Daten bereits vorliegen, ist es bedeutungslos, p-Werte durch statistische Verfahren zu berechnen, die erst ausgewählt wurden, nachdem man sich die Daten bereits angesehen hat". Kaum jemals aber sind bei der Versuchsplanung alle wichtigen Informationen vorauszusehen, die in den Daten stecken werden. In vielen Fällen lässt sich über einen Einsatz statistischer Verfahren erst nach gründlicher Dateninspektion entscheiden. Wer sich den Urteilsspruch von Nanninga und Nienhuys gegenüber Signifikanztests, die nach Dateninspektion gewählt werden, zu eigen macht, geht mit Scheuklappen an möglichen uneinkalkulierbaren Erkenntnissen vorbei. Das Gegenteil von Scheuklappen ist zu fordern: Eine Sensibilisierung für Unvorhersehbares.

4. Nanninga und Nienhuys betrachten meinen ersten Zuordnungsversuch mit MG nicht als den ersten Versuch einer Serie mit drei Folgen. Sie klassifizieren ihn als 12. Versuch einer Serie, die ich vor Jahren mit 11 unausgelesenen Astrologen unter anderen Bedingungen (Internetbefragung, westliche Astrologie usw.) – nur mit dem gleichen Versuchsmaterial – durchführte. Bei MG als einem der 12 Fälle habe man das Alpha-Niveau zu adjustieren, der p-Wert werde dann nichts mehr besagen. Antwort: Nehmen wir an, bei einem Patienten kommt es bei wiederholter Verabreichung eines Medikaments, das den Blutdruck senken soll, mal zu Senkungen, dann aber auch zu Steigerungen des Blutdrucks, und dies auch bei einer Wiederholung. Dann verhält man sich nicht so, wie man sollte, wenn man diese anomalen Kreislaufdaten mit den Behandlungsdaten von 11 anderen Personen in einen Topf wirft, bei denen das Präparat nicht gewirkt hat, wodurch die Daten des anomal reagierenden Patienten durch einebnende Mittelwertsbildung verschwinden. Vielmehr hat man die Anomalie bei diesem Patienten mit besonderer Aufmerksamkeit zu beachten und sich zu fragen, warum das Mittel bei ihm überhaupt gewirkt hat und warum auch mal in nicht erwünschter Richtung.

Generell ist hier anzumerken: Dass man bei Anwendung von Signifikanzberechnungen auf Merkmale, auf die man erst durch Daten-Inspektion aufmerksam wurde, sich des explorativen Charakters dieses Vorgehens bewusst sein soll und bei „Signifikanz"-Entscheidungen Alpha-Adjustierungen vorzunehmen hat, meine ich zu wissen und zu beachten. Doch liegen die dabei zu berücksichtigenden Bedingungen oft in einer Zone des Abwägens und Dafürhaltens. Nanninga und Nienhuys z. B. halten dafür, dass MG der verspätete Proband Nr. 12 einer Testserie sein soll, die vor sieben Jahren abgeschlossen wurde. Dieser Meinung dürfen sie sein. Doch wird diese Zone des freien Dafürhaltens von Nanninga und Nienhuys missbraucht, wenn sie mit freizügigen Rechtfertigungen den vom Untersucher eingebrachten p-Wert unter die Schwelle der Signifikanz manövrieren. [42]

---

[42] Einige Worte noch zum letzten Absatz in Nanninga & Nienhuys' Kommentar: „Nach unseren früheren Erfahrungen mit dem Autor", schreiben sie, hielten sie es „für wenig sinnvoll", mit mir zu diskutieren. „Die Diskussionen sind endlos, Ertel scheint nie einzusehen, warum er kritisiert wird, und er gräbt nur immer weitere neue, weit hergeholte Spekulationen hervor." Dann lassen sie durchblicken, dass sie die Kommunikation mit mir abgebrochen haben (sie hätten „wichtigere Din-

5. Abschließende Reflexion: Bei der nahezu alleinigen Kritik an den p-Werten durch Nanninga und Nienhuys (wie zuvor durch Guiard) kann bei methodentheoretisch weniger informierten Lesern der Eindruck entstehen, als sei mit der Entscheidung über Null-Hypothesen alles entschieden.[43] Mein Verständnis von der Strategie des NHST (*Null Hypothesis Significance Testing*) ist ein anderes. NHST ist lediglich ein Werkzeug, von dessen Verwendung im Rahmen der Forschung, der es dient, allenfalls Kriterien, aber keine Entscheidungen erwartet werden können. Nach Benutzung des Werkzeugs NHST mit einem Ergebnis, bei dem nach Regeln der Statistik über Nullhypothesen *entschieden* wurde, wird der Forscher auf die Ebene seiner *empirisch-inhaltlichen Hypothesen* (EIH) entlassen, von der er ausging und die ihn veranlasste, NHST zu benutzen.[44] Auf der Ebene EIH aber wird nichts mehr *entschieden*, hier wird nur *geurteilt*. So sind z. B. über die Güte des verwendeten statistischen Verfahrens, über das Gewicht des Analyse-Ergebnisses im Kontext der Untersuchung und des gesamten Forschungsprozesses nur Urteile zu treffen. Meine Kritiker übersehen dies, auch übersehen sie, dass ich angesichts der Ergebnisse der Untersuchung mit MG selbst nur hypothetische Urteile abgegeben habe, z. B. als ich schrieb, ich hielte die EIH über einen Zusammenhang zwischen auffälligen astrologischen Trefferleistungen und Psi für gestützt. Es sei also sinnvoll, sie weiter zu prüfen. Der letzte Balltest mit MG (Einbeutelverfahren, Postscriptum) lässt sich als eine erste weitere Überprüfung der Heterogenitätshypothese verstehen, die eine Form der allgemeineren Psi-Hypothese darstellt. Da das Ergebnis die Heterogenität signifikant repliziert hat (siehe Tabelle 4), darf man der Hypothese mehr abgewinnen als zuvor

---

ge zu tun"). Tatsächlich antworten sie nicht mehr auf E-mails, in denen ich an sie Fragen richte zur Aufklärung einer Anomalie in ihren Daten. Die Autoren nehmen mit ihrem Vorwurf Bezug auf eine Jahre zurückliegende Auseinandersetzung über ein anderes Thema („Marseffekt"), das mit dem vorliegenden Thema („Astrologie und Psi") nichts zu tun hat. Mit ihrer sachlich unmotivierten Abschweifung haben sie ihre aktuelle Aufgabe eines Kommentars aus dem Auge verloren, in dem allein die vorliegende Arbeit behandelt werden sollte. In einer Replik darauf, die verständlich sein sollte, hätte ich den von ihnen ausgesparten, ganz anderen und sehr umfänglichen Kontext mitzuteilen.

[43] Guiard selbst scheint so zu denken, denn er trug sich mit dem Gedanken, wie er schreibt, die Publikation meines Untersuchungsberichts insgesamt abzulehnen, und dies – wie er an anderer Stelle zugibt – obwohl er bei seiner Beurteilung die *inhaltlichen* Ausführungen des Verfassers ausklammerte!

[44] Empfehlung einer Übersichtslektüre: Mit beachtlicher wissenschaftstheoretischer Reflexion differenzieren in einem psychologischen Standard-Handbuch Hussy & Möller (1983) die Ebenen der *inhaltlichen* und *statistischen* Hypothesen, die engstens zusammen gehören, aber unter allen Umständen auseinander gehalten werden müssen. Basislektüre: Hager & Westermann (1983). Gründe zur Relativierung des NHST sind bei 14 Autoren in Harlow et al. (1997) zu finden. So werden etwa im darin zu findenden Aufsatz von Schmidt & Hunter die acht wichtigsten von 86 Argumenten analysiert, mit denen Autoren die traditionelle NHST-Strategie verteidigen. Guiard, Nanninga und Nienhuys werden zwar kaum bereit sein, sich z.B. ihrem provozierenden Urteil anzuschließen: "Statistical significance testing retards the growth of scientific knowledge; it never makes a positive contribution." (S. 37). Doch könnten sie ihre Argumente für eine ausschließlich (Guiard) oder fast ausschließlich (Nanninga und Nienhuys) an NHST-Begriffen orientierte Kritik im Lichte dieser Herausforderungen begründen und in dieser Zeitschrift zur Diskussion stellen. Als Kommentatoren sind sie den Autoren, die ihnen ausgeliefert werden, eine grundlegendere Rechtfertigung ihrer hinterfragbaren methodischen Grundposition schuldig.

und sich für weitere Forschung in der eingeschlagenen Richtung motivieren lassen – so habe ich *geurteilt*. [45]

Dass Nanninga und Nienhuys auf dieser Ebene des freien Urteilens zu anderen Einschätzungen kommen, dagegen ist grundsätzlich nichts einzuwenden. Wohl aber ist als unfein abzulehnen, dass sie ihre Urteile als *Entscheidungen* ausgeben mit der offensichtlichen Tendenz, unliebsame p-Werte abzuwürgen. Zudem entfalten sie dabei eine Autorität, die nur auf der NHST-Ebene berechtigt ist, wo mit zweiwertiger Logik letzte *Entscheidungen* getroffen werden wie z.B. die, dass p=.051 die Signifikanzschranke von .050 nicht erreicht. Eine solche Entscheidung ist im geschlossenen System der provinziellen Statistik unanfechtbar. Doch lässt sich die mathematische Autorität und Präzision von *Entscheidungen* über *statistische Hypothesen* nicht auf die Ebene des *Urteilens* über *wissenschaftliche Hypothesen* übertragen. Über Psi bei MG ist kein verbindliches Endurteil zu treffen. Auch hat die erneute Bestätigung der Hetereogenitätshypothese bei MG nur Wahrscheinlichkeit erhöht, aber nichts *entschieden*. Erst recht bleiben Weltbilder durch p-Werte in der Regel unbehelligt, sie werden durch Einzelereignisse, die sich auf der NHST-Ebene abspielen, kaum gefährdet, meine Kritiker haben hier nichts zu befürchten.

4. Zu Ulrike Voltmer: Wirkt sich „astrologische Motivation" auf den Balltest aus?

Ulrike Voltmers Beitrag ist der einzige problemlösungsorientierte unter den Kommentaren. Obgleich sich die Autorin nicht sicher ist, ob der Balltest den Ansprüchen eines validen Psi-Tests hinreichend genügt, lässt sie ihre experimentelle Fantasie spielen und schlägt für die Testdurchführung Bedingungsvariationen vor.

Nach meinem Verständnis ihrer Absicht ist sie allerdings bemüht, die Astrologie auszuklammern, woraufhin zu fragen wäre: Warum dies? Geht es in meinem Artikel doch eigentlich primär um die Grundfrage, ob astrologische Deutungserfolge mit Psi zusammenhängen, eine Frage, die die Autorin auch durchaus richtig paraphrasiert.

Voltmer möchte nur die mögliche Auswirkung „astrologischer Motivation" auf eine Psi-Variable überprüft sehen. Doch braucht man dabei die Astrologie nicht auf Eis zu legen. Man kann die Motivationsfrage mit der Grundfrage von „Astrologie und Psi" methodisch ohne weiteres verknüpfen, und zwar wie folgt:

Neues experimentelles Design, Grundidee U. Voltmer: Man selektiere für eine Untersuchung praktizierende Astrologen. Nur bei diesen kann „astrologische Motivation" beim Balltest überhaupt hervorgerufen werden. Die Frage ist nur, wie am besten? Frau Voltmer spielt fünf methodische Variationen durch, die sich auf drei Typen reduzieren lassen. Von diesen scheint mir der an erster Stelle genannte Typ der geeignetste zu sein, welcher mit dem an fünfter Stelle genannten weitgehend identisch ist.

Der rechte Beutel enthält 50 Bälle mit darauf geschriebenen Geburtsdaten von 50 bedeutenden Persönlichkeiten, welche fünf verschiedenen Berufen angehören. (Die Namen werden nicht auf die Bälle geschrieben). Die Berufe können mit 1, 2, 3, 4, 5 kodiert und auf die Codes können auf die Bälle des linken Beutels geschrieben werden (das entspräche Voltmers

---

[45] Empfohlener Merksatz: Auf der *Ebene der Statistik* zeigt Signifikanz an, was dort der Fall ist, auf der *Ebene der Problemlösung* zeigt sie an, dass es dort noch etwas Wichtiges zu tun gibt.

Vorschlag, z.B. 1 für Sportler, 2 für Maler usw.). Doch wird man auch z. B. Abkürzungen der Berufsbezeichnungen auf die Bälle schreiben können, anstatt sie von den Probanden umständlich aus Ziffern dekodieren zu lassen.
Die Aufgabe des Probanden würde darin bestehen, einen der fünf Berufe nach Belieben mental auszuwählen, d.h. ihn als nächstes Ziehergebnis vorauszusagen und aus dem linken und rechten Beutel gleichzeitig je einen Ball zu ziehen. Ein Treffer links ergäbe sich, wenn der angesagte Beruf auf dem gezogenen Ball zu lesen ist. Ein Treffer rechts ergäbe sich, wenn das dort gezogene Geburtsdatum das einer Person ist, die dem vom Probanden angesagten Beruf angehört.
Die Aufgabe wäre für praktizierende Astrologen fraglos interessanter (motivationsaktivierender) als Bälle zu ziehen, auf denen die Lösung langweiliger Additions- und Subtraktionsaufgaben vom ABC-Schützen-Niveau stehen (wie bei MGs Zweibeuteltest).
Das Design erfordert eigentlich einen Vergleich. Man könnte den hypothetisch angenommenen Effekt einer derart operationalisierten Motivationsstärke dadurch prüfen, dass man die astrologiepraktizierenden Probanden einmal unter den gerade beschriebenen Bedingungen testet, ein zweites Mal unter vergleichbaren Bedingungen, die keinen Bezug mehr zur Astrologie haben dürften.[46]
Doch würde ich es vorziehen, das Grund-Design beizubehalten, d.h. unter motivationsfreundlichen Bedingungen, wie soeben beschrieben, zwei Teilsamples von Astrologen zu testen, von denen ein Sample in ihrer Praxis auffallend gute astrologische Horoskopdeutungen zustande bringen (aufgrund von Urteilen aufmerksamer Beobachter der Szene), während das Vergleichssample nach dem Urteil derselben Beobachter keine auffälligen Erfolgsbilanzen haben. Bei Zutreffen der Grundhypothese würde man mit Anwendung des Balltests bei den erfolgreichen Astrologen mehr Treffer erwarten als bei den weniger erfolgreichen Astrologen. Im einzelnen ergäben sich nach den Voraussagen der Probanden, bei Zutreffen der Hypothese, 1. mehr Treffer im linken und/oder 2. im rechten Beutel und/oder 3. häufigeres gleichzeitiges Ziehen von Geburtsdatum und zugehörigem Beruf, auch wenn es sich dabei um nicht jeweils vorausgesagte Berufe handelt.
Bei MG, der in seiner astrologischen Praxis offenbar viele treffende Deutungen zustande brachte (wenn man seinem eigenen Eindruck probeweise einen Wahrheitskern zubilligt), würden wir im Balltest, nach Modifikation a la Voltmer, vielleicht überzeugendere Psi-Effekte erhalten als bei seinen früheren Zweibeuteltests, die im Artikel beschrieben wurden.
Wie dem auch sei, bei einem Balltest mit Astrologen sollten die intuitiv erfolgversprechenden methodischen Überlegungen von Ulrike Voltmer zur „astrologischen Motivation“ mit einbezogen werden.

Schlussfazit

Die Positionen der Autoren Boller, Guiard, Nanninga und Nienhuys, die meinen Artikel kommentierten, sind hinreichend transparent geworden. Man möchte ihnen raten, bei zu-

[46] Die Standardversion des Zweibeutel-Tests, die MG verwendete, wäre kein vergleichbarer Test, da dem Probanden bei Verwendung von Geburtsdaten im rechten Beutel diese Daten ja unbekannt sind und er über die Frage, ob das gezogene Geburtsdatum richtig oder falsch ist, nach den einzelnen Zügen im Ungewissen bleibt.

künftigen Kommentaren über Forschungsberichte ihre Aufgabe darin zu sehen, die jeweiligen Autoren bei ihren Problemlösungsbemühungen durch fruchtbare Kritik zu unterstützen, anstatt zu versuchen, ihnen durch vernichten wollende Kritik zu schaden. Dass im vorliegenden Fall die sachlichen Teile ihrer Ausführungen nicht überzeugen und sie auf dieser Ebene ihr Ziel gänzlich verfehlen, wird man nur nach zeitaufwendiger Lektüre meiner Replik erkennen. Dazu genügt kein Überfliegen der verwendeten Titel, mit denen zwei Kommentatoren auf der Schnellstraße fragwürdiger verbaler Etiketten, die das Urteil eiliger Leser prägen, eine Verunglimpfung des ganzen Forschungsprojekts erreichen wollen. Ulrike Voltmer geht auf die in meinem Artikel geäußerte Einladung ein, mit eigenen Ideen am hier gesponnenen Forschungsfaden anzuknüpfen, die vier anderen Kommentatoren versuchen, den Faden abzuschneiden. Der Kontrast zwischen wissenschaftsbelebender und wissenschaftslähmender Kommentierarbeit könnte kaum größer sein.

Literatur

Burdick, D.S.; Kelly, E.F. (1977): Statistical methods in parapsychological research. In: Wolman, B.B. (Ed.): Handbook of Parapsychology. Van Nostrand, New York.

Ertel, S. (2004): Psi test feats achieved alone at home: Do they disappear under lab control? In: Schmidt, S. (Ed.): Proceedings of the 47th Annual Convention of the Parapsychological Association.

Hager, W.; Westermann, R. (1983): Planung und Auswertung von Experimenten. In: Bredenkamp, J.; Feger, H. (Hrsg.) Hypothesenprüfung. Enzyklopädie der Psychologie. Forschungsmethoden der Psychologie, Bd. 5, 24-238. Hogrefe, Göttingen.

Harlow, L.L.; Mulaik, S.A.; Steiger, J.H. (1997, Eds.): What if there were no significance tests? Erlbaum, Mahwah.

Hussy, W.; Möller, H. (1983): Hypothesen. In: Herrmann, T.; Tack, W.H. (Hrsg): Methodologische Grundlagen der Psychologie. Serie I, Bd. 1. der Forschungsmethoden der Psychologie. Hogrefe, Göttingen.

Nanninga, R. (2004): Statistisch vuurwerk. De berekeningen van professor Ertel. *Skepter* 17, 30-31.

Thouless, R.H. (1935): Dr. Rhine's recent experiments on telepathy and clairvoyance and a reconsideration of J.E. Coover's conclusion on telepathy. *Proceedings of the Society for Psychical Research* 43, 24-37.

Thouless, R.H. (1970): The measurement of efficiency of ESP. *Journal of the Society for Psychical Research* 45, 323-325.

Timm, U. (1971): Die Messung von Psi-Leistungen. *Zeitschrift für Parapsychologie und Grenzgebiete der Psychologie* 13, 160-175.

Timm, U. (1983): Statistische Selektionsfehler in der Parapsychologie und in anderen empirischen Wissenschaften. *Zeitschrift für Parapsychologie und Grenzgebiete der Psychologie* 25 (3/4), 195-229.