Kommentare zu Stefan Schmidt, Peter Conrad und Harald Walach: Ein ASW-Experiment mit einem Hohlspiegel

Wolfgang Ambach[1]

Kommt Psi vom Fass oder aus der Dose?

Wer um Physik schon immer einen großen Bogen gemacht hat, wird bei der Beschreibung dieses Experiments vor Respekt erblassen. So viele Gesetze der Optik (vergessen seit achter Klasse), der Mechanik (geschwänzt in siebter Klasse), der Elektrik (verhasst seit neunter Klasse) und der magnetischen Induktion (vermutlich auch mal gehört) schlummern hier in gebündelter Form in einem komplexen experimentellen Aufbau, mit dem das Ziel verfolgt wird, Außersinnliches nachzuweisen. Vom Außersinnlichen weiß man ja ohnehin so wenig, dass die Anwendung mehrerer physikalischer Teildisziplinen durchaus vielversprechend, zumindest jedenfalls nicht unangemessen erscheint. Allerdings hat man von einer zusätzlichen Vakuumierung und Erhitzung des experimentellen Aufbaus, einer monochromatischen Beleuchtung und einer Bestrahlung mit Mikrowellen- oder Polizeifunkfrequenz abgesehen.

Ein echtes Experiment. Und ein Hauch von Tinguely

Wer um Physik dagegen keinen Bogen gemacht hat, fragt sich vielleicht, welche Vorstellung von einem möglichen Wirkmechanismus wohl die Einbringung eines mit Aluminium ausgekleideten Hohlspiegels in den experimentellen Aufbau motiviert hat. Wir wissen um die optisch und thermisch reflektierenden, elektrisch leitenden und nichtmagnetischen Eigenschaften von Aluminium und von der optisch vergrößernden, strahlenbündelnden Wirkung einer konkaven Struktur. Aber warum sollte, warum könnte das helfen, Hellsehen und dessen Nachweis zu fördern?

Die Autoren berufen sich auf eine historische Vermutung, die gleichfalls einer Begründung entbehrt. Bei weiterer Recherche muss man feststellen, dass offenbar auch gar nicht unterschieden wird, ob es die Versuchsperson ist, die sich selbst in einem Hohlspiegel (ähnlich einer Tonne oder einem großen Fass) befindet und das Hellsehen versucht, wie in Vorexperimenten,

1 Dr. med. Wolfgang Ambach leitet die Forschungsgruppe Klinische und Physiologische Psychologie am Institut für Grenzgebiete der Psychologie und Psychohygiene (IGPP), Freiburg i.Br., und ist Herausgeber der Anthologie *Experimentelle Psychophysiologie in Grenzgebieten* (Würzburg: Ergon, 2012).

oder ob das experimentelle Detail, das die Versuchsperson durch Hellsehen erkennen soll, innerhalb einer nur wenige Zentimeter großen Hohlspiegel-Pappröhre von der Art einer Dose angeordnet ist. Es scheint also gleichwertig zu sein, ob Psi aus dem Fass oder aus der Dose kommt. Hauptsache Hohlspiegel.

Weiter fällt die elektrisch leitende Verbindung der Elektroden in der Hand der Versuchsperson mit einem (oder keinem) der Hohlspiegel auf, genauer gesagt, mit den darin befindlichen Spulen. Bifilarspulen wurden verwendet, Doppeldrahtspulen mit gegenläufiger Anordnung; falls also – woher auch immer – ein Strom im Draht fließen sollte, wird das entstehende Magnetfeld durch diese Art der Spule so gering wie möglich gehalten. Es muss also irgend ein Konzept von elektrischem Strom und von magnetischer Wirkung dieses elektrischen Stroms den Versuchsaufbau mitbestimmt haben, mutmaßlich die Vorstellung eines Stroms zwischen Versuchsperson und Hohlspiegel (sonst hätte man ja eine nichtleitende Paketschnur als Verbindung nehmen können). Es muss auch die Vorstellung eines Magnetfeldes im Hohlspiegel gegeben haben (sonst machen Spulen keinen Sinn), genauer gesagt, eines Magnetfeldes, das es nicht geben soll. Und es musste ein Konzept davon geben, was diese physikalischen Phänomene mit Hellsehen zu tun haben könnten. Wir erfahren nichts über all das. Ich werde den Verdacht nicht los, dass die Autoren einfach all das verbaut haben, was beim letzten Löten übriggeblieben war.

Es bringt für das Verständnis dieses Experiments offensichtlich nichts, über irgendwelche physikalischen Wirkzusammenhänge nachzudenken oder irgendwelche Hypothesen über solche Wirkzusammenhänge nachvollziehen zu wollen. Physikalische Grundkenntnisse erlauben immerhin, den Versuchsaufbau in erster Näherung als wahllose, aber augenfällige Anordnung von potentiell optisch, elektrisch und magnetisch wirksamen Einzelelementen zu entlarven. Diese Einschätzung würde ich bereitwillig revidieren, würden entsprechende Zusammenhangshypothesen nachgeliefert und plausibel gemacht. Ohne eine logische Begründung für den Aufbau fühlt man sich eher an eines der bizarren und spektakulären Experimente des Fernsehcharakters „Professor Denzil Dexter“ aus der früheren britischen Fernsehserie *The Fast Show* erinnert: „Professor Dexter's experiments are in turn pointless and dangerous. They seem to stem from a combination of optimism and an uncertain grasp of reality. He has tried to pass his hand through a pane of glass by changing its molecular structure, made a special hat to harness gamma rays, and announced that he is investigating spacebats.“ (Britisch Broadcasting Corporation, 2003).

Bei der vorliegenden Studie scheint wirklich die Freude am ungezügelten Basteln und Experimentieren die treibende Kraft gewesen zu sein. Wir sind an dem Punkt, an dem zielgerichtetes Experimentieren in sinnfreie Kunst übergeht. Jean Tinguely hätte sicher seine wahre Freude an den Spulen und Spiegeln gehabt. Der kannte die inneren Zusammenhänge seiner Aufbauten allerdings recht genau.

Röhre, Spiegel und Spule im Kampf mit dem Zufall

Nachdem die Beschreibung des experimentellen Aufbaus und die Auswahl seiner einzelnen Elemente eine plausible Herleitung vermissen lässt, erscheinen die formulierten Hypothesen ebenfalls ohne Ableitung und somit beliebig. Wieder einmal folgt ein Psi-Experiment der Devise, wenn fehlerfrei gearbeitet wird, bedeutet jede Abweichung von der Zufallserwartung einen Bruch mit der konventionellen Vorstellung und ist somit spektakulär. Nach dieser Devise lassen sich auch ohne Theorie beliebige Experimente konzipieren, die allesamt folgende Gemeinsamkeiten haben:

1. Eine nachweisbar verzerrungsfreie Methodik, sowohl die Versuchsdurchführung als auch die Auswertung betreffend, ist Grundvoraussetzung.
2. Aus einem negativen Ausgang folgt nichts.
3. Ein positives Ergebnis sprengt (in Form einer probablistischen Aussage) den konventionellen Erklärungsrahmen, ist aber mangels Theorie kaum inhaltlich zu deuten.

Bezogen auf die vorliegende Studie frage ich mich vor allem, welche Schlussfolgerungen hätten gezogen werden können, hätte man ein signifikantes Ergebnis in allen oder auch in einzelnen experimentellen Bedingungen beobachtet. Hätte man eine signifikante Abweichung von der Zufallserwartung gefunden – hätte das dann geheißen, Psi kommt aus der Röhre? Oder aus dem Spiegel? Oder aus der Spule? Die Tatsache, dass das hier beobachtete Ergebnis am Ende von solchen Fragen weitgehend befreit, macht die Sache einfacher, aber nicht nachvollziehbarer.

Ein Ausflug in die Kombinatorik

Die Durchsicht der Datenanalyse fördert einen Irrtum im Detail zutage, dessen Folgen ich quantitativ abzuschätzen versucht habe. Bei der Auswertung der Trefferquoten gehen die Autoren davon aus, dass die Trefferwahrscheinlichkeit unter Zufallsbedingungen bei $p = 0.1$ liegt und dass die erwarteten Trefferhäufigkeiten binomialverteilt sind. Während ersteres zutrifft, unterliegt der Annahme einer Binomialverteilung ein Trugschluss. Die einzelnen Rateereignisse sind nämlich – entgegen der Annahme der Autoren – nicht voneinander unabhängig.

Die Analyse, die die Autoren vornehmen, wäre für voneinander unabhängige Ereignisse korrekt, wie sie bei einem *Ziehen mit Zurücklegen* auftreten. Tatsächlich handelt es sich bei der hier beschriebenen Versuchsdurchführung jedoch um ein *Ziehen ohne Zurücklegen*: Es wird nicht jede Zahl aus einem Pool der Zahlen „1“ bis „10“ gezogen, sondern es werden jeweils vier Zahlen aus einem Pool von 40 Zahlen gezogen, in denen jede der Zahlen 1-10 vier mal

vorkommt. Dadurch hängen die Trefferwahrscheinlichkeiten innerhalb eines 4er-Durchgangs[2] voneinander ab. Komplizierend kommt hinzu, dass das Ausmaß dieser inneren Abhängigkeit der vier Trefferwahrscheinlichkeiten vom Rateverhalten jeder einzelnen Versuchsperson bei jedem einzelnen 4er-Durchgang abhängt. Rät eine Versuchsperson beispielsweise einmal vier gleiche Zahlen, etwa „1111", so ist die Wahrscheinlichkeit zwar für jede einzelne Zahl 0.1, aber für „genau 4 Treffer" nicht etwa 0.0001, wie nach der Binomialverteilung zu erwarten wäre, sondern 4/40 x 3/39 x 2/38 x 1/37 = 0.00001094, also nur gut ein Zehntel davon. Die Wahrscheinlichkeit für „genau 1 Treffer" liegt bei diesem Tip nach den Regeln der Kombinatorik bei 0.3125, also höher als man anhand einer Binomialverteilung erwarten würde (nämlich 0.2916). Rät eine Versuchsperson dagegen lauter verschiedene Zahlen, etwa „1234", so weicht die Verteilung der Trefferhäufigkeiten in der anderen Richtung von der Binomialverteilung ab, nämlich tritt „genau 1 Treffer" seltener auf, „genau 0 Treffer" oder „genau 4 Treffer" dafür häufiger.

Tabelle 1 gibt die nach der Binomialverteilung erwarteten Werte und die tatsächlichen Wahrscheinlichkeiten der Trefferverteilung bei einem 4er-Durchgang wieder; die letzteren werden nach den fünf möglichen Typen eines einzelnen Ratetyps aufgegliedert (lauter verschiedene Zahlen, ein Zwilling, zwei Zwillinge, ein Drilling, vier gleiche Zahlen).

Tabelle 1

Ratetyp		**p(genau 0 Treffer)**	**p(genau 1 Treffer)**	**p(genau 2 Treffer)**	**p(genau 3 Treffer)**	**p(genau 4 Treffer)**
„1234"	(lauter verschiedene)	0.657337	0.289412	0.049283	0.003852	0.000117
„1123"	(ein Zwilling)	0.655283	0.293037	0.048164	0.003429	0.000088
„1122"	(zwei Zwillinge)	0.653237	0.296633	0.047088	0.002976	0.000066
„1112"	(ein Drilling)	0.651056	0.300635	0.045607	0.002659	0.000044
„1111"	(vier gleiche)	0.644545	0.312507	0.041361	0.001576	0.000011
Binomialverteilung		0.656100	0.291600	0.048600	0.003600	0.000100

2 Im Artikel wird das Wort „Durchgang" inkonsistent benutzt: Es steht manchmal für eine einzelne zu erratende Zahl, manchmal auch für ein 4er-Set.

Mit der Abweichung der Trefferwartungen von der Binomialverteilung innerhalb eines 4er-Durchgangs weicht auch die Häufigkeitserwartung für die gesamte Studie von der Binomialverteilung ab, und dies in Abhängigkeit vom Rateverhalten jeder einzelnen Versuchsperson. Im Folgenden werden nur noch die beiden Ratetypen „1111“ und „1234“ betrachtet, die als Extreme eine Eingrenzung ermöglichen.

Noch komplizierter wird die Sache dadurch, dass in einem 4er-Durchgang nun die beiden Bedingungen „mit Spiegel“ und „ohne Spiegel“ enthalten sind. Wenn über diese Daten nicht gepoolt wird, etwa um diese beiden Bedingungen miteinander zu vergleichen, dann folgt die Häufigkeitsverteilung einem noch komplexeren Muster, da die vier abhängigen Ziehungen auf je zwei für die beiden Bedingungen aufgeteilt werden; die Resultate der beiden Bedingungen sind damit auch voneinander abhängig.

Aufgrund der zunehmenden Komplexität des Problems bin ich dem nicht weiter analytisch nachgegangen, sondern habe den Ausgang der Studie in einer Computersimulation auf seine Streuungsmaße hin untersucht. Die Darstellung will ich hier abkürzen.

Die Daten der gesamten Studie wurden unter Zufallsbedingungen 1000 Millionen mal simuliert. Die relativen Häufigkeiten, bei 220 geratenen Zahlen (wie in jeder einzelnen Bedingungskombination) „genau *k* Treffer“ zu erhalten, wurden bestimmt, und zwar einmal nach der Binomialverteilung und je einmal für jedes der 5 möglichen Rateverhalten. Letzteres wird der Komplexität bei variiertem Rateverhalten natürlich nicht gerecht, erfasst aber mit den Varianten „1111“ und „1234“ die beiden Extreme.

Eine Tabelle, die die relativen Häufigkeiten aller möglichen Versuchsergebnisse wiedergibt, wird aus Platzgründen den Autoren separat zur Verfügung gestellt. Beim Ratetyp „1111“, der im einzelnen 4er-Durchgang eine Rarifizierung von „genau 4 Treffer“ und „genau 0 Treffer“ zugunsten häufigeren Vorkommens von „genau 1 Treffer“ hatte, ist auch in der Gesamtverteilung der möglichen Studienausgänge eine größere zentrale Tendenz festzustellen als bei der Binomialverteilung. Beim Ratemuster „1234“ geht die notwendige Korrektur in die umgekehrte Richtung.

Bezogen auf den 50%igen Trefferüberhang (33 Treffer bei Erwartungswert 22) in der Bedingung „mit Spiegel, offen“ ergaben sich in der Simulation die in Tabelle 2 dargestellten Häufigkeiten am oberen Ende der Verteilung:

Tabelle 2

	p(= 33 Treffer)	p(≥ 33 Treffer)	p(> 33 Treffer)
binomial	0.005094	0.012298	0.007203
Ratetyp „1111“	0.004784	0.011308	0.006525
Ratetyp „1234“	0.005128	0.012407	0.007280

Die exakten Wahrscheinlichkeiten im konkreten Experiment würden sich nur aufwendig unter Berücksichtigung des tatsächlichen Rateverhaltens der einzelnen Teilnehmer bestimmen lassen. Im Falle vorherrschenden Rateverhaltens vom Typ „lauter verschiedene Zahlen“ („1234“) wäre ein Trefferüberhang als signifikanter, im Falle des vorherrschenden Typs „1111“ als weniger signifikant einzuschätzen, als es die Berechnung anhand der Binomialverteilung ergibt. Die Abschätzung der beiden Extremverhalten erlaubt immerhin den Schluss, dass die tatsächlichen Wahrscheinlichkeiten, die unter Berücksichtigung von „Ziehen ohne Zurücklegen“ gewonnen werden, in jedem Fall nur ganz marginal von denen abweichen, die anhand einer Binomialverteilung berechnet würden. Die Auswirkung des methodischen *faux pas* auf die Signifikanzgrenzen ist also praktisch zu vernachlässigen. Für den Fall einer Replikation wird allerdings empfohlen, jede einzelne Zahl unabhängig aus einem Pool von zehn zu ziehen.

Noch mehr Statistik: Die Logik der Interpretation

Entscheidender als mögliche Unterschiede in der vierten Kommastelle dieser durch Simulation bestimmten Wahrscheinlichkeiten dürften jedoch folgende Überlegungen sein:

(1) Welche Wahrscheinlichkeit interessiert, um sie mit dem Signifikanzniveau zu vergleichen: p(„genau 33 Treffer“), p(„mindestens 33 Treffer“) oder p(„mehr als 33 Treffer“)? Da eine diskrete Verteilung der Trefferquoten vorliegt, ist dies (im Gegensatz zur kontinuierlichen Normalverteilung) ein Problem. Die von den Autoren angegebene Online-Ressource zur Berechnung liefert übrigens komfortablerweise auch ein „mid-p“, das wohl anhand einer speziellen Interpolation gewonnen wurde. Freilich kann eine solche Interpolation generell diskutiert werden. In Ermangelung des entsprechenden Algorithmus habe ich zur überschlägigen Berechnung eines mid-p auf eine Spline-Interpolation zurückgegriffen. Für die hier empirisch gewonnene Verteilung liefert dies eine Irrtumswahrscheinlichkeit für zweiseitige Testung zwischen 1.78% (Ratetyp „1234“) und 1.97% (Ratetyp „1111“). Auch die Autoren ließen letztlich offen, ob sie sich (nach ihren Berechnungen) bei dem genannten Trefferüberhang in der einen Teilbedingung für eine Irrtumswahrscheinlichkeit von 1.3% oder 1.8% entscheiden.

(2) Wird einseitig oder zweiseitig getestet? Da in Psi-Experimenten regelmäßig auch eine signifikant verminderte Trefferquote von Interesse ist, entspricht das Vorgehen der Autoren, zweiseitig zu testen, guter wissenschaftlicher Praxis.

(3) Welche Korrektur für multiples Testen ist angebracht? Die von den Autoren angewandte Bonferroni-Korrektur erscheint angemessen, da die Beobachtungen für „offen" und „verdeckt" voneinander unabhängig und die Beobachtungen für „mit Spiegel" und „ohne Spiegel" trotz des hier vorgebrachten Einwands voneinander immerhin *weitgehend* unabhängig sind. Wir landen damit für das genannte Einzelereignis bei einer Irrtumswahrscheinlichkeit im Bereich von 7% bis 8%.

(4) Wichtiger als alle bisherigen Rechenübungen ist am Ende die Frage, ob es inhaltliche Hypothesen für eine Trefferhäufung in genau der Bedingungskombination gibt, in der diese beobachtet wurde. Sehr begrüßt wird, dass die Autoren dieses Teilergebnis als explorativ werten und es damit von der ursprünglichen Fragestellung absetzen. Ganz im Sinne von Cohens „The earth is round. $p < .05$" (Cohen, 1994) sollte bei der finalen Bewertung von Studienergebnissen nicht die Frage im Vordergrund stehen, ob das Signifikanzniveau, das wie ein Fallbeil zwischen 0.049 und 0.051 schwebt, gerade noch erreicht wird oder gerade nicht mehr. Gerade bei dem genannten Trefferüberhang sollten bei der Frage, ob dieses exploratorische Ergebnis in einer Folgestudie weiter abgeklärt werden sollte oder nicht, letztlich nicht minimale Zahlendifferenzen rund um 0.05 sondern vielmehr inhaltliche Zusammenhangshypothesen ausschlaggebend sein. Wenn es diese nicht gibt, wäre auch eine Folgestudie nur ein konzeptloses Suchen nach Signifikanz.

Experimentelle Ökonomie

An dem Experiment nahmen 22 Versuchspersonen teil, die jeweils 40 Zahlen erkennen sollten, nämlich zehn mal vier gleichzeitig präsente, „verdeckte" Zahlen. Aus der Beschreibung des experimentellen Ablauf lässt sich mit Wahrscheinlichkeit abschätzen, dass das einzelne Experiment kaum länger als eine halbe Stunde in Anspruch genommen haben dürfte. Macht 11 Stunden Datenerhebung, die freilich noch mit einem Zuschlag für die Rekrutierung, die Einholung des Einverständnisses und ähnliches zu versehen sind.

Nun frage ich die Autoren, wie viele Stunden mit der Sichtung der relevanten Literatur, der Planung des Experiments, der manuellen Herstellung des experimentellen Aufbaus, der Eingabe und statistischen Auswertung der Daten und vor allem mit dem Abfassen der schriftlichen Arbeit (Masterarbeit), evtl. diversen Referaten und schließlich mit dem Abfassen des vorliegenden Manuskripts und dem Lesen dieses Kommentars vergangen sind. Bis zum Beweis des Gegenteils schätze ich: 1000 Stunden. Unterstellt man diese (freilich völlig provisorische)

Zeitzuordnung, so machte die Datenerhebung nicht einmal zwei Prozent des insgesamt für das Experiment betriebenen Aufwands aus.

Dies ist unter ökonomischen Gesichtspunkten des Experimentierens nicht optimal. Freilich sind die Ressourcen begrenzt, die einem bestimmten Experiment zukommen können. Gleichzeitig soll das Experiment möglichst aussagekräftig sein, bezogen auf das Testen von Nullhypothesen also eine möglichst große Teststärke bei einer bestimmten, vorgegebenen Irrtumswahrscheinlichkeit aufweisen. Das Interessante ist nun, dass die Datenerhebung am augenfälligsten mit zeitlichem Aufwand verbunden ist, während die genannten übrigen Tätigkeiten in ihrem Zeitbedarf eher unterschätzt werden, dies sowohl von Laien als auch von den Experimentatoren selbst.

Dieses Phänomen betrifft nicht nur grenzwissenschaftliche Studien, sondern ist ubiquitär. In vielen Wissenschaftsbereichen kann ein zu kleiner Stichprobenumfang allerdings mit dem Argument beschönigt werden, man habe sich eben nur für Effekte mit ausreichend großer Stärke interessiert. Bei der vorliegenden Studie wird dieses Argument allerdings kaum angebracht sein, so dass die Autoren sich vermutlich den Einwand gefallen lassen müssen, sie hätten – rein nach der Versuchsökonomie – besser das Zehnfache an Versuchspersonen untersucht, den bereits gefassten Entschluss zur experimentellen Untersuchung genau dieser Fragestellung immer vorausgesetzt. Folgt man der Devise, lieber zehn Prozent mehr Aufwand für eine solche Studie betreiben, diesen zusätzlichen Aufwand ganz der zusätzlichen Datenerhebung widmen, und im Ausgleich dafür jede zehnte Studie sein lassen, so wird man letztlich mehr valide Aussagen aus seinem Werk ziehen können.

Freilich ist auch bei 220 Versuchspersonen bei einem Alpha-Fehler-Niveau von 5.0% mit einer Wahrscheinlichkeit von 5.0% ein falsch positives Ergebnis zu erwarten. Jedoch hätte man (a) dann ein strengeres Alpha-Niveau zugrunde legen können und (b) einen Effekt, wenn er denn tatsächlich besteht, mit wesentlich höherer Wahrscheinlichkeit gefunden. Das aktuelle Rätseln darum, was nun wohl dahintersteckt, dass in einer Teilbedingung ein 50%iger (!) Trefferüberhang beobachtet wurde, hätte es dann jedenfalls nicht gegeben.

Fazit: Ohne Sinn, aber mit Verstand

Die Studie wurde – mit Ausnahme eines kleinen, praktisch folgenlosen Irrtums, der die Kombinatorik betrifft – offensichtlich methodisch sauber durchgeführt und mit viel Umsicht interpretiert. Besonders gefällt die konsequent durchgehaltene Logik der Interpretation hypothesenbezogener und exploratorischer Ergebnisse. Die dargestellte Studie leidet allerdings schwer darunter, dass keine nachvollziehbaren Gründe für die experimentelle Anordnung genannt werden, so dass der Eindruck einer Beliebigkeit des Experimentierens entsteht. Insofern tritt

der wissenschaftliche Erkenntniszuwachs dem Eindruck nach hinter den erheblichen künstlerischen Wert des durchgeführten Experiments zurück. Unterm Strich also keine Sternstunde für das zielgerichtete Streben nach Wissenszuwachs durch Experimente, aber ein Glück für methodische Diskutierer, Bastler und Rechner, zugleich ein Lehrstück für Studierende, und vor allem ein längst fälliger Vorstoß in das zielfreie künstlerische Experimentieren.

Literatur

Britisch Broadcasting Corporation (2003). http://www.bbc.co.uk/comedy/fastshow/characters/denzil_dexter.shtml (letzter Zugriff: 6.12.2012).

Cohen, J. (1994). The earth is round ($p < 0.05$). *American Psychologist*, 49, 997-1003.

Suitbert Ertel[3]

Signifikanzkriterien allein können irreführen

Die Autoren Schmidt, Conrad und Walach (SCW) sind der Meinung, ihr Hellseh-„*Experiment zeigte keinen Hinweis auf ASW*“[4] und „*Die naheliegendste Interpretation ist hier […] das Vorliegen einer Zufallsschwankung*“.

In der experimentellen Parapsychologie ist vielen Forschern noch nicht bewusst, dass man ein statistisch nicht-signifikantes Ergebnis nicht ohne weiteres auf ein Defizit an Psi-Effekten zurückführen darf. Immer ist sowohl die Stärke von Psi-Effekten (sie können vorhanden, aber zu schwach sein) als auch die Häufigkeit ihres Auftretens zu berücksichtigen (die Zahl der Messwiederholungen kann für einen Nachweis zu gering sein). Tut man das nicht, fällt man leicht einem Beta-Irrtum zum Opfer: Man hält irrtümlicherweise die Null-Hypothese und damit das Nichtvorhandensein von Psi im Experiment für bestätigt.

Solche Urteile aber sind nicht erlaubt, solange man die Power eines statistischen Tests, das Zusammenspiel von Effektstärke (ES) und Auftretenshäufigkeit (N = Anzahl von Messwiederholungen) im Hinblick auf Signifikanz nicht beachtet. Speziell für Parapsychologen sind hier zwei wichtige Arbeiten von Jessica Utts (1988, 1991) wegweisend, die von SCW offenbar nicht bedacht wurden.

3 Prof. Dr. Suitbert Ertel ist emeritierter Professor für Psychologie an der Universität Göttingen.

4 ASW = Außersinnliche Wahrnehmung

Die methodologische Beziehung der drei entscheidenden Begriffe untereinander lässt sich mit der einprägsamen Abbildung 1 veranschaulichen.

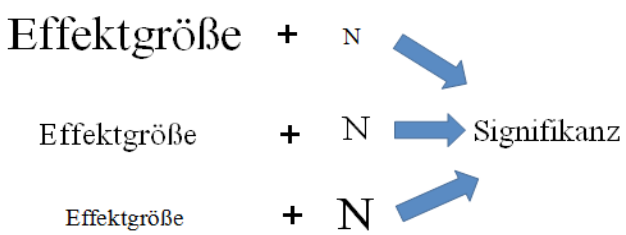

Abb. 1: Je kleiner die Effektgröße, umso mehr Trials (N) (Messwiederholungen) sind notwendig, um einen Effekt als zweifelsfrei oder fast zweifelsfrei vorhanden zu erweisen. Die variable Schriftgröße soll die relative Effektgröße bzw. die relative Häufigkeit von N andeuten. Der Signifikanzwert soll für dieses Modell als invarianter Wert aufgefasst werden.

Im Experiment der Autoren SCW wird ASW (speziell Hellsehen) geprüft. Die Versuchsteilnehmer raten das Vorkommen der Zahlen 0 bis 9 unter vier verschiedenen Bedingungen: Hohlspiegelbehälter verwendet (HSp-ja), Hohlspiegelbehälter nicht verwendet (HSp-nein), Proband ist bei den Trials informiert, ob ein Behälter HSp enthält (O = offen) oder nicht enthält (V = verdeckt). Die Bedingungen wurden so kombiniert, dass vier Kombinationen entstanden: HSp-ja mit O, HSp-ja mit V, HSp-nein mit O, HSp-nein mit V. Nur von einer Hypothese ist bei SCW die Rede: Hypothetisch erwartet wird eine durch Psi hervorgerufene höhere Trefferzahl bei den beiden HSp-ja-Bedingungen im Vergleich zu den beiden HSp-nein-Bedingungen. Von hypothetischen Erwartungen für die Bedingungen V und O und einer möglichen Wechselwirkung mit HSp ist in dem Artikel nicht die Rede.

Die Auswertung der Daten durch SCW stützt sich allein auf die Wahrscheinlichkeit einer Abweichung vom Zufall (Signifikanz p). Effektgrößen werden nicht berücksichtigt. Die Autoren beurteilen ihr eigenes Vorgehen so, als hätten sie die ASW-Hypothese und nicht den Unterschied zwischen HSp-ja und HSp-nein prüfen wollen: „*Das Experiment kann insgesamt von seinem methodischen Vorgehen als valider Test auf ASW gewertet werden*“.[5]

5 Abgesehen von der Hauptkritik in diesem Kommentar ist zu bemängeln, dass hier die Validität des verwendeten Rate-Tests als ASW-Test behauptet wird, ohne dass die Frage der Reliabilität des Tests auch nur angesprochen, geschweige denn untersucht wird. Die Validität eines Tests aber setzt seine Reliabilität voraus.

Tabelle 1
Enthält die Ergebnisse in übersichtlicherer Form:
HSp-ja = mit Hohlspiegel, HSp-nein = ohne Hohlspiegel, V = verdeckt O = offen

	Bedingung Hohlspiegel ja – nein	**Bedingung offen – verdeckt**	**N trials**	**Treffer-zahl**	**Z**	**ES**	**P**
1	HSp ja	O	220	33	2.36	.159	.009 *
2	HSp ja	V	220	22	.00	.000	n.s.
3	HSp nein	O	220	18	-.79	-.053	n.s.
4	HSp nein	V	220	27	1.01	.068	n.s.
5	HSp ja		440	55	1.67	.077	.05
6	HSp nein		440	45	.08	.004	n.s.
7	O		440	51	1.03	.049	n.s.
8	V		440	49	.072	.003	n.s.
9	Total		880	100	1.29	.043	n.s.
10	Balltest (N = 47 Vpn)		22560	2620	8.07	.054	$<10^{-10}$

ES = Z / sqrt(N)
* Für SCW ergibt sich hier p = .013 bzw. p = .018, den die Autoren als Zufallsschwankung abtun.

Die Hypothese der Autoren (in der HSp-ja-Bedingung sei ein ASW-Effekt stärker als in der HSp-nein-Bedingung) lässt sich auf Signifikanz prüfen. Man vergleiche das Ergebnis mit Hohlspiegel (Zeile 5) vs. Ergebnis ohne Hohlspiegel (Zeile 6). Der Unterschied der Trefferproportionen (55 vs. 45) ist nicht signifikant (errechnet mit Online-Hilfe durch vassarstats.net).

Allerdings wird bei diesem Vergleich die Bedingungsvariable offene vs. verdeckte Vorrichtung nicht berücksichtigt, über die keine Hypothese geäußert wird. Wenn man nun davon ausgeht, dass eine offene und damit für die Probanden etwas durchschaubarer gemachte Bedingung psi-förderlich sein müsste und eine verdeckte Bedingung nicht oder weniger psi-förderlich wäre und man deshalb für den anstehenden Vergleich die beiden Hohlspiegel-Bedingungen nur unter Offenheit verwendet, dann wird der Proportionsvergleich (Trefferzahlen 33 vs. 18, Zeile 1 vs. 3) sehr signifikant (Z = 2.33, p = .01).

Aber diese Hypothese habe nur ich und zudem post hoc aufgestellt. Beschränkt man sich auf die Hypothese, dass sich im vorliegenden Experiment nur unter vermutlich günstigen Bedingungen ASW manifestieren müsste und verzichtet man auf einen Vergleich mit den vermutlich weniger günstigen Bedingungen (verdeckte Darbietung), dann sind zwei signifikante Ergebnisse zu erkennen; einmal in Zeile 1 für die HSp-ja-Bedingung mit offener Darbietung (Binomial-Z = 2.36, p = .009) und in Zeile 5 für die HSp-ja-Bedingung, egal ob mit offener oder verdeckter Darbietung kombiniert (Binomial-Z = 1.67, p = .05). Meine eigene Post-hoc-Hypothese (vielleicht war diese auch bei SCW unterschwellig vorhanden) gewinnt somit empirisch etwas Plausibilität.

Die Hauptkritik meines Kommentars hat jedoch nicht die Vagheit der Hypothesenbildung der Autoren zum Gegenstand. Bedauerlicher ist, dass SCW den Einfluss der Effektgröße völlig vernachlässigen. Statistische Signifikanzwerte sind wesentlich mit bedingt durch Effektgrößen (s. Abb. 1). Hätten SCW die Zahl der Messwiederholungen im Experiment erhöht (diese sind manipulierbar), dann wären bei gleichbleibender Effektstärke (diese ist nicht oder nicht so leicht manipulierbar) höhere p-Werte erzielt worden.

Meine Voraussage lässt sich stützen, wenn man die Effektstärken ES der kritischen Bedingung des Experiments von SCW (Tabelle 1, Zeile 1, ES = .169, und Zeile 5, ES = .077) mit Effektstärken aus anderen Psi-Experimenten vergleicht. Tabelle 2 gibt einen Überblick über ES-Werte für 10 verschiedene Kategorien experimenteller Psi-Untersuchungen (Meta-Analysen). Die Effektgröße für „total" aus dem Experiment von SCW beträgt ES = .043 (Tabelle 1, Zeile 9). In Tabelle 2 sind für vier andere Experimente der Forced-Choice- Kategorie Effektstärken aufgeführt (Zeilen 5 bis 8), deren Meta-Analysen ES von .054, .048, .010 und .009 aufweisen. Mit einer so großen Zahl von Messwiederholungen, wie sie in Meta-Analysen eingehen, resultieren für diese Psi-Effekte sehr hohe Signifikanzwerte.

Zu den Forced-Choice-Tests mit Effektstärken, die ähnlich hoch ausfallen wie im statistischen Test „total" von SCW (ES = .043, Tabelle 1, Zeile 9) gehört der Ball-Selektionstest (BST), den ich entwickelt habe (Ertel, 2005, 2007, 2008, 2010a, 2010b, 2010c). In Tabelle 1, Zeile 10, hatte ich die Ergebnisse einer Balltest-Durchführung mit Studienanfängern aus dem Jahr 2003 eingefügt. In diesem Experiment gab es ebenfalls 10 Target-Alternativen (Raten der Zahlen 1 bis 5 mit entweder roter oder grüner Farbe). Das Forced-Choice-Ergebnis von SCW (ES = 0.43) ist demnach hinsichtlich Effektstärke ähnlich hoch wie das des Forced-Choice-Tests BST von 2003 (ES = .054). Die Treffer-Proportion des Tests von SCW unterscheidet sich nicht signifikant von der Treffer-Proportion des BST-Tests von 2003, so dass man die Daten von SCW zu derselben Grundgesamtheit zählen darf, zu der die durch Psi-Effekte ausgewiesenen BST-Daten von 2003 gehören.

Tabelle 2
ES-Effektstärken bei 10 verschiedenen experimentell hervorgerufenen Psi-Phänomenen und deren Klassifizierung sowie Quellen der ES-Information aus Ertel (im Druck)

	Psi-Phänomen	Kategorie	Quelle der ES-Werte	ES
1	Telepathy dream studies	Free Response	RADIN (2007)	.256
2	DMILS	BIO-PK	SCHLITZ & BRAUD (1997)	.240
3	Remote viewing	Free Response	MILTON (1997)	.160
4	Ganzfeld	Free Response	STORM & ERTEL (2001)	.154
5	Ball Selection Test	Forced choice	ERTEL (2012)	.054
6	Card guessing	Forced choise	RHINE et al. (1940	.048
7	Precognition	Forced choice	STEINKAMP (1998)	.010
8	Clairvoyance	Forced choice	STEINKAMP (1998)	.009
9	Dice Tossing	Makro-PK	RADIN & FERRARI (1991)	.003
10	RNG	Mikro-PK	RADIN & NELSON (1998)	.0003

ES: Effect size
DMILS: Direct mental interaction with living systems
PK: Psychokinesis
RNG: Random number generator

Man könnte einwenden, dass das Balltest-Ergebnis von den Studenten unter ungewöhnlichen Heimbedingungen, nämlich ohne Aufsicht durch einen Experimentator, gewonnen wurde. Doch wurden von den 47 Balltest-Probanden 12 mit signifikanten Heimtest-Ergebnissen für eine anschließende Wiederholung des Versuchs unter meiner Aufsicht selektiert. Die derart ausgelesene Stichprobe von 12 Probanden erzielte unter Aufsicht eine Effektstärke von ES =

.118 (ein mehr als doppelt so hoher ES-Wert wie die 47 Probanden unter der Heimbedingung ohne Aufsicht gezeigt hatten). Damit darf man den Einwand, man könne das Ergebnis von SCW unter Aufsicht eines Experimentators nicht mit dem Ergebnis des Balltests vergleichen, weil das Balltest-Ergebnis ohne Aufsicht und somit durch Nicht-Psi-Faktoren entstanden sein könne, als entkräftet ansehen. Zudem steht zum Vergleich auch noch die Effektstärke von ES = .048 eines anderen Forced-Choice-Tests (Card guessing) zur Verfügung (Tabelle 2, Zeile 6), die mit der Effektstärke von SCW vergleichbar ist. Mit anderen Worten, das Ergebnis von SCW lässt sich in die Rangordnung früher gewonnener Psi-Effekte der Forced-Choice-Kategorie zwanglos einordnen.

Ich begrüße die Gelegenheit, das Experiment von SCW zu kommentieren, weil sich an ihm die bedauerlichen Konsequenzen eines Vorgehens aufweisen lassen, bei dem die Zusammenhänge zwischen Effektstärke ES, Beobachtungshäufigkeit N und Signifikanzlevel p nicht beachtet werden. Ein solches Vorgehen kennzeichnet die parapsychologische Forschung auch heute noch generell. Die oft beklagten statistisch unbefriedigenden Ergebnisse von Psi-Experimenten (wie das von SCW) wären mit Berücksichtigung dieser Zusammenhänge vermeidbar. Eine methodische Reform experimenteller parapsychologischer Forschung, die sich durch eine primäre Ausrichtung auf Testpower statt auf Signifikanz auszeichnen müsste (s. Utts, 1988, 1991), dürfte im Ergebnis auf Skeptiker einen stärkeren Eindruck machen als weniger effektive Strategien, mit denen die „deniers“ des Paranormalen zur Anerkennung von ASW-Effekten gebracht werden sollen.

Am Beispiel der Selektion von Probanden durch Vortests unter Heimbedingungen (meine Strategie) lässt sich auch zeigen, dass beliebige Experimente im Psi-Phänomenbereich von vorne herein mit psi-begabten Probanden durchgeführt werden können. Diese werden aller Voraussicht nach nicht nur höhere Psi-Effektstärken, sondern auch höhere Signifikanzen erzielen, als sie mit unausgelesenen Probanden-Stichprobem gewonnen werden.

Vorschlag für ein Dissertationsthema: Eine umfassende Meta-Analyse der bisherigen experimentellen Psi-Forschung in den verschiedenen Phänomen-Bereichen (siehe Tabelle 2) sollte durchgeführt werden mit dem Ziel, nur die jeweils eingesetzte Power der Tests nachträglich zu ermitteln. Das Ergebnis würde uns das Desaster in der bisherigen parapsychologischen Forschung vor Augen führen. Eine Reaktion darauf würde nach meiner Schätzung einen großen Sprung nach vorne zur Folge haben, wie man ihn auf andere Weise nicht auslösen könnte.

Literatur

Ertel, S. (2005). Psi test feats achieved alone at home: Do they disappear under lab control? *Australian Journal of Parapsychology*, 5, 149-164.

Ertel, S. (2007). Außersinnliche Wahrnehmung unter der Kontrolle organisierter Skeptiker. *Zeitschrift für Anomalistik*, 7, 236-269.

Ertel, S. (2008). Betrugsverdacht und sensorische Schlupflöcher. *Zeitschrift für Anomalistik*, 8, 143-153.

Ertel, S. (2010a). On individual differences in extrasensory perception. In Rao, K.R. (Ed.), *Yoga and Parapsychology: Empirical Research and Theoretical Essays* (S. 331-350). Delhi: Motilal Banarasidass.

Ertel, S. (2010b). Psi in a skeptic's lab. *Journal of Scientific Exploration*, 24, 581-598.

Ertel, S. (2010c). Replikation von ASW-Heimtest-Ergebnissen im Labor. *Zeitschrift für Anomalistik*, 9, 108-139.

Ertel, S. (im Druck). Assessing psi ability: A challenge for psychometrics. In Broderick, D.G., & Goertzel, B. (Eds.), *The Science of Psi*.

Utts, J. (1988). Successful replication versus statistical significance. *Journal of Parapsychology*, 52, 305-320.

Utts, J. (1991). Replication and meta-analysis in parapsychology. *Statistical Science*, 6, 363-403.

Patrizio E. Tressoldi[6]

Die Falle der Nullhypothesen-Signifikanzprüfung
Ein Vorschlag für die Analyse der Resultate von Schmidt *et al.*

In ihrem ganz neuen Versuchsprotokoll für ASW-Tests mittels "Hohlspiegeln" berichten Schmidt *et al.*, ihre Resultate seien gemäß der klassischen Nullhypothesen-Signifikanzprüfung (NHST) fast ausnahmslos nichtsignifikant ausgefallen. Die Deutung ihrer Resultate lautet demnach folgerichtig: „Das Experiment zeigte keinerlei Hinweise auf ASW".

Dies ist ein weiteres, ja ein wiederkehrendes Beispiel dafür, dass eine mehr als unzulängliche statistische Praxis, die Gigerenzer *et al.* (2004) als „Nullritual" bezeichnet haben, zu einer fehlerhaften Interpretation experimenteller Resultate geführt hat.

Wie ich an anderer Stelle gezeigt habe (Tressoldi, 2012), führt mangelnde Kontrolle der statistischen Power zu einer irrtümlichen Nichtzurückweisung der Nullhypothese (also zu einem Fehler II. Art oder einem Beta-Fehler), insbesondere dann, wenn die untersuchten Phänomene

6 Dr. Patrizio E. Tressoldi ist in der psychologischen Forschung am Dipartimento di Psicologia Generale an der Universität Padua in Italien tätig. Er gilt als Experte für die Anwendung metaanalytischer Verfahren. (Email: patrizio.tressoldi@unipd.it).

eine geringe Effektstärke (ES) haben. Wie Schmidt *et al.* angesichts ihrer langjährigen Erfahrung auf diesem Gebiet sehr wohl wissen, bewegen sich die Effektstärken für ASW in Bereichen von weniger als 0.01 für Forced-Choice-Verfahren mit normalbewussten Versuchspersonen bis hin zu 0.28 für Experimente mit antizipatorischen psychophysiologischen Reaktionen.

Um mit ihrem Experiment eine hinreichende statistische Power (etwa von 0.90) für die Zurückweisung der Nullhypothese zu erreichen, wären für eine geschätzte Effektstärke von annähernd 0.20 mindestens 33 Versuchspersonen (für einen Binominaltest) bzw. 216 Versuchspersonen (für einen One-Sample t-Test) vonnöten gewesen.

In der folgenden Tabelle habe ich für die wichtigsten von Schmidt *et al.* erzielten Resultate die Effektstärken (ES) mit ihren 95%igen Konfidenzintervallen (CI) berechnet.

Effekt	**ES= z√n**	**95% CI**
Gesamt	0.28	-0.15-0.70
Hohlspiegel	0.37	-0.70-0.80
Kontrollbedingung	0.03	-0.39-0.45

Die Effektstärkenschätzungen für die Gesamt- und die Hohlspiegel-Ergebnisse bewegen sich im Bereich der besten Effektstärken, die wir bis heute überhaupt mit den unterschiedlichsten ASW-Versuchsprotokollen erzielt haben. Wenn wir die Konfidenzintervalle betrachten, ist ein wenig Vorsicht angebracht, weil deren Spannen negative Werte einschließen. Wiederum ist jedoch der Hinweis wichtig, dass diese Intervalle (bei geringer Präzision) wegen der niedrigen Zahl der Versuchspersonen sehr weit sind.

Warum verabschieden wir nicht das "Nullritual" und übernehmen stattdessen zuverlässigere statistische Verfahren wie diejenigen, die von der American Psychological Association (2010; Cumming, 2012) empfohlen werden?

(aus dem Englischen von Gerd H. Hövelmann)

Literatur

American Psychological Association. (⁶2010). *Publication Manual of the American Psychological Association*. 6th ed. Washington, DC: American Psychological Association.

Cumming, G. (2012).*Understanding the New Statistics: Effect Sizes, Confidence Intervals and Meta-Analysis.* New York: Routledge.

Gigerenzer, G., Krauss, S., & Vitouch, O. (2004). The null ritual: What you always wanted to know about significance testing but were afraid to ask. In Kaplan, D. (Ed.), *The Sage Handbook of Quantitative Methodology for the Social Sciences* (S. 391-408). Thousand Oaks, CA: Sage.

Tressoldi, P.E. (2012). Replication unreliability in psychology: Elusive phenomena or "elusive" statistical power? *Frontiers in Psychology*, 3, in advance of publication: doi: 10.3389/fpsyg.2012.00218.

Autorenantwort:

Stefan Schmidt, Peter Conrad, Harald Walach

Von der Spannung zwischen Ideal und Pragmatik

Wir möchten uns bei den drei Kommentatoren für ihre sorgfältige und detaillierte Begutachtung unserer Arbeit bedanken. Im Nachfolgenden wollen wir zu den Anregungen Stellung beziehen. Dabei fassen wir die Kommentare von Wolfgang Ambach und von Suitbert Ertel und Patrizio Tressoldi zur Experimentellen Ökonomie, Powerabschätzung bzw. Studiengröße in einem Punkt zusammen, da sie aus unserer Sicht auf denselben Sachverhalt abzielen.

Modell mit Zurücklegen

Wolfgang Ambach hat mit seiner Beobachtung, dass durch die Art und Weise der Zielzahlermittlung ein statistisches Modell ohne Zurücklegen verwendet wird, vollkommen recht. In der Tat ist damit die von uns verwendete Statistik in der Anwendung nicht ganz korrekt. Allerdings zeigt sich aber auch in der aufwändig durchgeführten Simulation, dass bei der hier vorgenommenen relativ kleinen Auswahl aus der großen Anzahl möglicher Ziele (4 Zahlen aus 40) die beiden Modelle nur minimale Abweichungen voneinander haben. In diesem Sinne könnte man auch jetzt, allerdings erst in Kenntnis der Ambach'schen Simulation, in einem pragmatischen Sinne die Binomialverteilung als einfach zu rechnende Approximation der komplexen Experimentalsituation annehmen. Diese Schlussfolgerung kann allerdings – wir denken, auch das hat Ambach klar gezeigt – nicht für andere Versuchsaufbauten verallgemeinert werden, da die Abweichungen von der Binomialverteilung jeweils von der Anzahl der gezogenen Ziele, von der Grundgesamtheit an Zielen und vom Rateverhalten der Versuchspersonen (das man grundsätzlich nicht als zufällig unterstellen kann) abhängt.

Auswahl der geeigneten Wahrscheinlichkeit aus der Binominalverteilung

Ambach fragt, welche der möglichen Wahrscheinlichkeiten aus der Binomialverteilung die geeignete sei, um sie mit dem Signifikanzniveau zu vergleichen: mindestens 33 Treffer, genau 33 Treffer oder mehr als 33 Treffer? Legt man hier die Logik an, die auch zur Ermittlung eines p-Wertes bei kontinuierlichen parametrisch verteilten Daten (z.B. aus einer Normalverteilung) Anwendung findet, dann sollte die Wahl aus unserer Sicht auf „mindestens 33 Treffer" fallen. Denn auch bei kontinuierlichen Daten fragt man ja nie nach der Wahrscheinlichkeit, genau dieses Ergebnis erhalten zu haben, sondern nach der Wahrscheinlichkeit, dieses oder ein noch extremeres Ergebnis zu finden.

Fass oder Dose?

Ambach kritisiert, dass der von uns verwendete Versuchsaufbau durch keine klar dargestellte Theorie begründet sei, und damit hat er ebenfalls recht. Wir haben in der Publikation angedeutet, dass der Aufbau des Versuchsapparates durch einen der drei Autoren (PC) auf der physikalischen Theorie der Zeit des russischen Astronomen Kozyrev und vor allem auf der Interpretation dieser Theorie durch Kaznacheev und Trofimov beruht. Um zu verstehen, warum wir diesen Versuchsaufbau überprüfen, ohne die zugrunde liegende Theorie genauer zu erläutern, muss man den Kontext der Arbeit verstehen. Das Experiment entstand im Rahmen einer Masterarbeit, der Masterstudent ist der Zweitautor (PC). Dieser hat das Experiment auf der Basis der entsprechenden Theorien und Quellen (von denen einige auf Russisch verfasst sind) konzipiert. Diese theoretischen Arbeiten sind jedoch nicht in einer nachprüfbaren Fassung publiziert (i.e. nicht auf Englisch oder Deutsch, nicht in einer Zeitschrift mit Peer Review). Da somit zwei der Autoren (SS und HW) nicht in der Lage waren, die Operationalisierung zu überprüfen, haben wir uns darauf verständigt, den Versuchsaufbau unhinterfragt und per se in einer Pilotstudie empirisch zu testen.

Bei einem positiven Befund könnte man dann immer noch auf Basis des zugrunde liegenden Theoriegebäudes differenzierte Ableitungen vornehmen. Es würde aber auch die Möglichkeit bestehen, alternative theoretische Modelle zur Erklärung eines eventuellen Psi-Effekts zu bemühen. Aus der Perspektive eines normalen Psi-Experiments ist hier der Umstand gegeben, dass eine psi-förderliche Bedingung („psi-conducive", siehe z.B. Irwin, 1999) probeweise eingeführt wird und gegen eine Standard-Psi-Bedingung (hier Pappröhre ohne Spiegel und Elektrode) getestet wird. Damit beantwortet sich auch die Frage, was aus einem eventuell positiven Ergebnis gefolgert worden wäre. Es würden sich dann neben einer Replikation auch Folgeuntersuchungen empfehlen, die aufzuklären versuchen, welche Komponenten oder welche Kombination dieser Komponenten für den Effekt verantwortlich zeichnen könnten.

Experimentelle Ökonomie und statistische Power

Alle drei Kommentatoren sind sich einig in dem Punkt, dass zu einer guten Beurteilung dieses Experimentes eine größere Stichprobe notwendig gewesen wäre. Wolfgang Ambach bemerkt richtig, dass, wenn das Experiment doch schon einmal entwickelt wurde, man es auch erschöpfend durchführen sollte. Ertel und Tressoldi verweisen auf die geringe statistische Power der Untersuchung. Das Problem mit der statistischen Power ist, dass man diese a priori nur dann gut abschätzen kann, wenn man auch eine gute Schätzung der Effektstärke vorliegen hat. Dies war hier nicht der Fall, da ein Experiment mit diesem Versuchsaufbau nicht zuvor durchgeführt worden war.

In diesem Fall kann man zwei Dinge tun: 1) Man sucht ähnliche Experimente und versucht dann, die Effektstärke abzuschätzen. So gehen sowohl Ertel als auch Tressoldi vor. Dass dies nicht ganz trivial ist, sieht man schon daran, dass Tressoldi einen Schätzwert von ES(r)=.20 und Ertel einen Range von .009 bis .054 vorschlagen. Der Unterschied zwischen den beiden entsprechenden Fallzahlberechnungen ist immens! 2) Da bleibt die zweite Alternative: man macht eine Pilotstudie, um eine geeignete Effektstärke zu ermitteln. Diese beträgt in unserem Fall über alle 880 Trials (Fragestellung 1) ES(r)=.043, wie Ertel korrekt berechnet (Tressoldi ist mit ES(r)=.28 offensichtlich ein Rechenfehler unterlaufen). Auf Basis dieses Wertes muss man tatsächlich davon ausgehen, dass unsere Fallzahlen zu gering waren. Man muss aber auch berücksichtigen, dass die Effektstärke lediglich eine Schätzung der wahren Effekstärke und hier immer mit einen gewissen Stichprobenfehler behaftet ist.

Letztendlich war hier jedoch auch die Pragmatik eines lediglich 3-monatigen Zeitkorridors für eine Masterarbeit ausschlaggebend. Wir finden, dass dies im Rahmen einer Pilotstudie zu einer ersten Abschätzung und generellen Überprüfung des experimentellen Paradigmas im Sinne eines ‚Feasibility Tests' zu vertreten und so auch üblich ist. Ambach hat zwar recht, dass verglichen mit dem sonstigen Aufwand die Mühen der Rekrutierung von Personen und der Datenerhebung prozentual gering sind. Logistisch aber, weil Menschen eben nicht über Zeit als Meterware verfügen, sondern Wartezeiten und Verfügbarkeiten eingerechnet werden müssen, ist dies eben nicht trivial. Und so schluckte auch hier, wie so oft, die Pragmatik das Ideal.

Literatur

Irwin, H.J. ([3]1999). *An Introduction to Parapsychology* (Third Edition). Jefferson, NC: McFarland.