*Zeitschrift für Anomalistik* Band 10 (2010), S. 58-115

# Konzeptuelle Replikationsstudie zu Experimenten zur Außersinnlichen Wahrnehmung[1]

Günter Daniel Rey[2], Katharina Berens, Elena Dietz, Manuela Hesser, Sandra Schäfer, Anne Schirmer[3]

**Zusammenfassung.** – Das Experiment untersucht mit Hilfe von Zenerkarten Fähigkeiten in der Außersinnlichen Wahrnehmung für die Bereiche Telepathie, Hellsehen und Präkognition. Angenommen wurde, dass a) auf Grund von paranormalen Fähigkeiten der Probanden eine überzufällig hohe Trefferzahl (Anzahl richtig erkannter Symbole) in den Teilbereichen auftritt und b) diese außerdem durch Motivation und die Einstellung gegenüber paranormalen Phänomenen moderiert wird. Es wurde ein dreifaktorieller Versuchsplan verwendet, der die 96 studentischen Probanden zunächst auf dem ersten, dreifachgestuften Faktor zufällig einem durch ein Feedback erzeugten motivationalen Zustand zuteilte. Zwei messwiederholte Faktoren schlossen sich an, nämlich die Testung in den drei Teilbereichen sowie eine 25-malige Wiederholung des Rateversuchs. Des Weiteren wurden vor dem Experiment Neurotizismus- und Extraversionswerte, sowie die paranormale Überzeugung der Probanden erhoben. Signifikante Effekte konnten weder für Telepathie (erwartete Trefferrate: 480;

1 *Redaktionelle Vorbemerkung*: Der nachfolgende Beitrag ist im Redaktionskollegium und unter den unabhängigen Gutachtern kontrovers diskutiert worden. Das liegt in erster Linie daran, dass diese Studie ein experimentelles Design benutzt, das in der parapsychologischen Forschungspraxis seit mehr als einem halben Jahrhundert keine Verwendung mehr findet und das – zu Zeiten, als es noch eingesetzt wurde – stets um aufwendige Vorkehrungen gegen Manipulationen und sensorische Lecks ergänzt war, die in dieser Studie jedoch fehlen. Die Resultate dieses Experiments stehen mithin – ganz gleich, wie sie ausgefallen wären – unter entsprechenden methodologischen Vorbehalten, was die Autoren selbst einräumen. Die Redaktion hat sich gleichwohl zur Publikation entschlossen, weil einerseits die Nichtveröffentlichung parapsychologischer Experimente dem Vorwurf der Datenselektion Vorschub leistet (Stichwort „File-Drawer-Problem") und weil es andererseits keineswegs an der akademischen Tagesordnung ist, dass im Rahmen der studentischen Ausbildung an einer deutschen Universität ein parapsychologisches Projekt durchgeführt wird. Dies hat entsprechende Aufmerksamkeit verdient.

2 Dipl.-Psych. Dr. Günter Daniel Rey studierte Soziale Verhaltenswissenschaften und Philosophie an der Fernuniversität Hagen sowie Psychologie an den Universitäten Frankfurt/M. und Trier. Er ist wissenschaftlicher Mitarbeiter am Lehrstuhl für Entwicklungspsychologie und Pädagogische Psychologie der Universität Würzburg.

3 Katharina Berens, Elena Dietz, Manuela Hesser, Sandra Schäfer und Anne Schirmer waren zum Zeitpunkt der Durchführung dieser Studie (2007) Psychologie-Studentinnen an der Universität Trier.

erreichte Trefferrate: 493; p = .26) und Hellsehen (480; 490; p = .31), noch für Präkognition (480; 456; p = .89) erzielt werden. Die unter b) postulierten moderierenden Einflüsse konnten ebenfalls nicht nachgewiesen werden.

*Schlüsselbegriffe*: Telepathie – Hellsehen – Präkognition – Zenerkarten – konzeptuelle Replikation

**Conceptual Replication Study of Experiments on Extrasensory Perception**

**Abstract.** – The experiment used Zener cards to investigate individual abilities in three categories of extrasensory perception: telepathy, clairvoyance and precognition. It was assumed that a) such abilities significantly increase the numbers of correct hits (correctly identified symbols) and that b) these abilities are moderated by motivational aspects and the general attitude towards paranormal phenomena. A three-factor model was used for the experiment. Ninety-six students were randomly assigned to three feedback-induced levels on the first between-subject motivational factor. Two within-subject factors were also used: firstly, testing in the three categories of extrasensory perception and, secondly, a 25-fold repetition of each trial. Results did not show any significant findings for either telepathy (expected correct hits: 480; achieved correct hits: 493; p = .26), clairvoyance (480; 490; p = .31) or precognition (480; 456; p = .89). The expected moderator effects according to b) also failed to reach significance.

*Keywords*: Telepathy – clairvoyance – precognition – Zener cards – conceptual replication

## *Einleitung*

Außersinnliche Wahrnehmung (ASW) beschreibt das Wissen um oder die Reaktion auf ein äußeres Ereignis, das bzw. die nicht über die bekannten Sinneswege vermittelt wird (Lucadou, 1997). Dabei werden in der parapsychologischen Literatur mitunter drei Basisphänomene der ASW unterschieden: Telepathie, Hellsehen und Präkognition (z.B. Steinkamp, 2005). Unter Telepathie versteht man die „direkte" psychische Informationsübertragung zwischen Personen ohne Beteiligung bekannter Kommunikationskanäle. Hellsehen bezeichnet die „direkte" Wahrnehmung von objektiven Vorgängen oder Sachverhalten, die niemandem bekannt sind. Das Vorauswissen über zukünftige Vorgänge, die rational nicht erschließbar sind und auch nicht als Folge dieser Voraussage auftreten, wird Präkognition genannt (Lucadou, 1997). Eine Erörterung der Frage, ob und inwiefern sich die aufgeführten drei Phänomene der ASW (Telepathie, Hellsehen und Präkognition) überhaupt voneinander unterscheiden lassen, soll an dieser Stelle nicht erfolgen (siehe hierzu stattdessen Irwin, 1999), da diese Frage nicht zentraler Gegenstand der vorliegenden Untersuchung ist.

Zur Erforschung der ASW wurden u.a. sogenannte „Forced-choice"- oder „Restricted-choice"-Experimente durchgeführt (Irwin, 1999). Dabei handelt es sich um Experimente, in denen die Probanden eine von mehreren vorgegebenen Antwortalternativen auswählen müssen. Beispielsweise werden den Versuchspersonen in zufälliger Reihenfolge sogenannte Zenerkarten dargeboten, die jeweils eines von fünf verschiedenen Symbolen zeigen (Kreis, Wellen, Kreuz, Quadrat oder Stern). Die Probanden sollen nun z.B. einem Empfänger das jeweils präsentierte Symbol Karte für Karte telepathisch übertragen. Derartige „Forced-choice"-Experimente gelten als älteste und einfachste Art der experimentellen Testung von Phänomenen zur ASW (Steinkamp, 2005) und gehen auf Experimentalserien von Joseph Banks Rhine zurück.

Eine ältere Meta-Analyse (Pratt *et al.*, 1940) zu Experimenten zur ASW umfasste 145 Berichte über experimentelle Studien zur ASW, duchgeführt im Zeitraum zwischen 1882 und 1940. Insgesamt schloss diese Analyse mindestens 77.796 Probanden und knapp fünf Millionen Einzelversuche ein. Die hochsignifikanten Ergebnisse der Auswertung deuteten nach Angaben der Autoren auf die Existenz von ASW hin.

Auch neuere Untersuchungen zur ASW greifen auf „Forced-choice"-Experimente zurück (Irwin, 1999). Im Gegensatz zu diesen Untersuchungen verwenden beispielsweise Ganzfeld-Experimente „Free-response"-Techniken zur ASW (Irwin, 1999), bei denen eine Versuchsperson einer vollständig homogenen, d.h. glatten, strukturlosen, gleichmäßigen und das ganze Sehfeld füllenden Fläche ausgesetzt wird. Zudem wird den Probanden mittels Kopfhörern ein sogenanntes „weißes Rauschen" (die Rauschenergie verteilt sich konstant über alle Frequenzen) dargeboten. Der im Ganzfeld befindliche Proband soll sodann beispielsweise Informationen von einer anderen Person, dem Sender, empfangen.

In der parapsychologischen Forschung kann die Befundlage zur ASW als uneindeutig dargestellt werden. Neben der bereits aufgeführten, die Existenz von ASW stützenden Meta-Analyse von Rhine und Kollegen (Pratt *et al.*, 1940) führten auch Bem & Honorton 1994 in ihrer Meta-Analyse zu Ganzfeld-Experimenten ein statistisch bedeutsames Ergebnis mit 30 Studien an ($p$ = 0.002, einseitig), welches die Existenz von ASW ebenfalls belegt. Fünf Jahre später, in Reaktion auf Bem & Honorton (1994), publizierten Milton & Wiseman (2001) eine Meta-Analyse über 30 Studien zu Ganzfeld-Experimenten zur ASW mit einer Trefferquote im Bereich der zufälligen Wahrscheinlichkeit ($p$ = 0.24, einseitig). Im Jahre 2001 bestätigten Storm & Ertel mit ihrer Metaanlyse, die 79 Ganzfeld-Studien aus dem Zeitraum 1974 bis 1996 einschloss, die Sichtweise von Bem & Honorton ($p = 7.78 \times 10^{-9}$). Milton & Wiseman (2001) wiederum zweifelten diese Belege für ASW aufgrund methodischer Probleme an. Wegen der uneinheitlichen Befundlage zur ASW bei der Verwendung von Ganzfeld-Studien soll im vorliegenden Experiment geprüft werden, ob Telepathie, Hellsehen und Präkognition bei der Auswahl von Zenerkarten (also in einem „Forced-choice"-Experiment) zu überzufällig hohen Trefferraten

führen. „Psi-missing" (z.B. Irwin, 1999) wurde dabei nicht in Betracht gezogen, um eine einseitige Testung zu gewährleisten und hierdurch die Teststärke der vorliegenden Untersuchung zu erhöhen (zum Vergleich der Teststärke bei ein- vs. zweiseitiger Testung siehe z.B. Bortz, 2005).

### *Moderierende Einflussfaktoren*

Die Forschung zur ASW berücksichtigt neben Persönlichkeitsdispositionen (z.B. Extraversion oder Neurotizismus) und Einstellungen der Personen gegenüber paranormalen Phänomenen (vgl. Kennedy, 2005; Lawrence, 1993), auch den Einfluss einer Rückmeldung über die Leistung der Versuchspersonen in Untersuchungen zur ASW (z.B. Braud, 2002) als potentielle Moderatoreffekte.

Bezüglich der Persönlichkeitsdimension Extraversion konnten Honorton, Ferrari & Bem (1998) in einer über 60 Studien umfassenden Meta-Analyse eine hochsignifikante, jedoch relativ geringe positive Korrelation ($r = .09$; $p = .000004$) zwischen der ASW-Leistung und der Persönlichkeitsdimension Extraversion nachweisen. Für „Forced-choice"-Experimente handelte es sich nach den Autoren der Meta-Analyse bei dieser Korrelation um einen artifiziellen Befund, da sich ein signifikanter Zusammenhang nur zeigte, wenn bereits im Vorfeld der Erfassung der Persönlichkeitsdimension die außersinnlichen Fähigkeiten getestet und die entsprechenden Ergebnisse den Versuchspersonen mitgeteilt wurden. Wurde Extraversion hingegen im Vorfeld gemessen, so konnte der aufgeführte Zusammenhang nicht nachgewiesen werden ($r = -.02$; n.s.). Palmer & Carpenter (1998) weisen die Behauptung eines artifiziellen Befundes jedoch zurück und argumentieren, dass in den „Forced-choice"-Untersuchungen, die in der Meta-Analyse Berücksichtigung fanden, Reihenfolge der Testung (Extraversion – ASW-Leistung vs. ASW-Leistung – Extraversion) und Testsetting (Individual- vs. Gruppentests) miteinander konfundiert seien. Eine eigene Analyse der Autoren kommt zu dem Schluss, dass sich unter Berücksichtigung des Testsettings die Reihenfolge der Testung nicht auf den nachweisbaren Zusammenhang zwischen den Extraversionswerten und den ASW-Leistungen auswirkt. In der vorliegenden Untersuchung wird Extraversion mittels NEO-FFI (Borkenau & Ostendorf, 1993) im Vorfeld der Überprüfung der ASW-Leistung gemessen, so dass Zusammenhänge zwischen beiden Variablen nicht auf die Testreihenfolge zurückgeführt und als artifiziell zurückgewiesen werden können.

Hinsichtlich der Persönlichkeitsdimension Neurotizismus können bei individueller Testung negative Zusammenhänge zur ASW-Leistung detektiert werden (Irwin, 1999). Personen mit niedrigen Neurotizismuswerten erzielen demnach bessere Ergebnisse in Untersuchungen zur ASW. In Gruppentestungen ist die Befundlage zu diesem korrelativen Zusammenhang hingegen deutlich inkonsistenter (ebd.).

Neben den Persönlichkeitsdimensionen Extraversion und Neurotizismus greift die vorliegende Untersuchung auch auf das Konzept der Selbstwirksamkeit von Bandura (1986) zurück. Nach diesem Konzept beeinflusst die Überzeugung, dass man in einer bestimmten Situation die angemessene Leistung erbringen und gezielt Einfluss auf die Dinge und die Welt nehmen kann, die Wahrnehmung, Motivation und Leistung der Person. Die Selbstwirksamkeit kann unter anderem durch die tatsächlich erbrachte Leistung, sowie durch die Bekräftigung durch Andere erhöht werden. Für das vorliegende Experiment wird folglich angenommen, dass ein von der Versuchsleiterin gegebenes fingiertes Feedback in einem kurzen Probedurchgang mit Zenerkarten die ASW-Leistung im nachfolgenden Hauptdurchgang moderiert. Eine negative Rückmeldung zu den Ergebnissen im Probedurchgang soll Versuchspersonen dabei demotivieren und signifikant höheren Trefferquoten in den Bereichen Telepathie, Hellsehen und Präkognition während des Hauptdurchganges entgegenwirken. Für die fingierte positive Feedback-Bedingung wird eine Motivationssteigerung und in der Folge eine Trefferzunahme postuliert. Versuchspersonen, welche kein Feedback erhalten, fungieren als Kontrollgruppe. Näheres zur Operationalisierung ist dem Methodenteil zu entnehmen.

Eine weitere potentielle Moderatorvariable hinsichtlich der erzielten Ergebnisse in Studien zur ASW stellt die Einstellungen der Person gegenüber paranormalen Phänomenen dar. Nach Lawrence (1993) weisen zahlreiche „Forced-choice"-Experimente zur ASW nach, dass Personen, die an ASW glauben (aus als „Schafe" bezeichnet, siehe z.B. Schmeidler, 1952; Schmeidler & McConnell, 1958) durchschnittlich höhere Trefferraten erzielen als sogenannte „Böcke", Personen, die nicht an derartige Phänomene glauben. Dieser Befund stelle zudem einen der konsistentesten Effekte in der gesamten parapsychologischen Forschung dar (siehe auch Irwin, 1999).

In seiner Metaanalyse (Lawrence, 1993) zu diesem Befund wurden 73 Einzelstudien aus den Jahren von 1947 bis 1993 berücksichtigt, in denen insgesamt über 4500 Probanden geteste wurden. Dabei zeigte sich ein hochsignifikanter und robuster, wenn auch von der praktischen Bedeutsamkeit her geringer positiver Zusammenhang ($r = .03$) zwischen dem Glauben an ASW und den gezeigten Leistungen in ASW-Experimenten. Aufgrund der vorliegenden Befundlage wird in der Untersuchung vermutet, dass der Einfluss von Telepathie, Hellsehen und Präkognition auf die Auswahl von Zenerkarten auch durch die Einstellung gegenüber paranormalen Phänomenen moderiert wird.

Auf Grundlage der aufgeführten Forschungsbefunde werden zusammenfassend folgende Hypothesen formuliert:

*Hypothese 1*: Telepathie, Hellsehen und Präkognition führen bei der Auswahl von Zenerkarten zu überzufällig hohen Trefferraten.

*Hypothese 2*: Der Einfluss von Telepathie, Hellsehen und Präkognition auf die Auswahl

von Zenerkarten wird durch (fingierte) Rückmeldungen der Fähigkeiten zur ASW sowie der Einstellung gegenüber paranormalen Phänomenen moderiert.

## *Methode*

### *Versuchsdesign*

Das Zenerkartenexperiment zur Überprüfung der Existenz von ASW bestand aus einem Probedurchgang, dem sich ein Hauptdurchgang mit der Untersuchung von Telepathie, Hellsehen und Präkognition anschloss. In dem Probedurchgang absolvierte jede Versuchsperson fünf Einzel-Trials, bei denen das jeweilige Kartensymbol anzugeben war. Für die Durchführung der fünf Trials hatten die Probanden drei Minuten Zeit. Der Probedurchgang sollte die Versuchspersonen mit den Zenerkarten vertraut machen; gleichzeitig diente dieser zur Realisierung der ersten unabhängigen Variable. Dem Experiment lag ein dreifaktorieller Versuchsplan mit Messwiederholung auf dem zweiten und dritten Faktor zugrunde. Dessen erste unabhängige Variable (UV) betraf die vorgegebene Rückmeldung seitens des Versuchsleiters bezüglich der Leistung der Versuchsperson im Probedurchgang. In der Kontrollgruppe wurde kein Feedback, in der ersten Experimentalgruppe ein positives („Bei deinem Probedurchgang war überzufällig viel richtig"), in der zweiten ein negatives („Bei deinem Probedurchgang war nichts richtig") gegeben. Statt symmetrischer Formulierungen wurden asymmetrische Formulierungen („überzufällig viel richtig" vs. „nichts richtig") eingesetzt, da sich auch die Trefferwahrscheinlichkeiten asymmetrisch verteilen.

Das negative Feedback mit der Behauptung, dass in den fünf Einzel-Trials des Probedurchganges kein Treffer erzielt wurde, tritt dabei mit einer Wahrscheinlichkeit von 32,77% auf. Eine mögliche Alternativformulierung („überzufällig viel falsch") analog zur ersten Gruppe mit positivem Feedback wäre an dieser Stelle nicht sinnvoll gewesen, da bei fünf Durchgängen gar keine überzufällig seltene Trefferausbeute in Erscheinung treten konnte (bei Verwendung der herkömmlichen 5%-Signifikanzgrenze). Zudem sollte die Versuchperson in dieser Versuchsbedingung demotiviert und nicht durch den Verweis auf eine besonders selten auftretende Trefferquote ermuntert werden. Umgekehrt verhielt es sich bei der ersten Experimentalgruppe, die ein positives Feedback erhielt. Hier hätte man dem Probanden auch mitteilen können, dass dieser sämtliche Einzel-Trials des Probedurchganges richtig gelöst hätte. Die Einzelwahrscheinlichkeit für diese Trefferausbeute beträgt jedoch nur 0.03%. Auch wenn davon auszugehen ist, dass den Probanden dieser Prozentwert nicht während des Versuchs zur Verfügung stand, so hätte dieses höchst unwahrscheinliche Ergebnis einzelne Probanden doch möglicherweise misstrauisch bezüglich des Feedbacks machen können. Um dies zu vermeiden, wurde auf

eine solche Formulierung des Feedbacks verzichtet und in dieser Versuchsbedingung lediglich auf die überzufällig hohe Trefferzahl verwiesen.

Die zweite UV fasste die ASW mit den drei Stufen Telepathie, Hellsehen und Präkognition zusammen. Als dritter Faktor wurde die 25-malige Wiederholung des Rateversuchs herangezogen, welche in allen drei außersinnlichen Wahrnehmungsbedingungen stattfand. Als abhängige Variable (AV) dienten die Übereinstimmungen zwischen dem von der Versuchsperson angegebenen und dem tatsächlich vorliegenden Symbol. Zu berücksichtigen ist hier, dass in der zu zweit durchgeführten Telepathie-Bedingung jeweils die Werte des „Empfängers" in die Auswertung eingingen.

### *Versuchsmaterial*

Das für das Experiment erforderliche Material umfasste 150 Zenerkarten, auf denen jeweils eines von fünf Symbolen (Kreis, Wellen, Kreuz, Quadrat oder Stern) abgebildet war. Sämtliche Zenerkarten wurden von einer Versuchsleiterin sorgfältig selbst erstellt. Dabei wurden zunächst die benötigten Symbole am PC generiert und ausgedruckt. Diese wurden im zweiten Schritt auf blauen, gleichgroßen und nicht transparenten Pappkarten geklebt. Die blauen Rückseiten dieser Karten sahen dem Anschein nach identisch aus, wobei nicht ausgeschlossen werden kann, dass geringfügige Individualmerkmale eine Diskriminierung der einzelnen Karten ermöglichten und somit im Verlauf der Untersuchung zu Lerneffekten führen könnten (siehe z.B. Irwin, 1999). Zu beachten ist jedoch, dass die Probanden für jede der drei Stufen (Telepathie, Hellsehen und Präkognition) der zweiten UV nur einen einzigen Durchgang mit 25 Karten vornahmen, so dass dieser potentielle Lerneffekt begrenzt gewesen sein dürfte.

Der selbsterstellte Fragebogen zur Einstellung gegenüber paranormalen Phänomene (siehe Anhang) enthielt 19 Aussagen, wobei die dazugehörigen fünfstufigen Likert-Skalen an beiden Enden beschriftet wurden. Die niedrigste Ausprägung „1" wurde mit „diese Aussage trifft auf mich absolut nicht zu", die höchste Ausprägung „5" mit „diese Aussage trifft auf mich absolut zu" gekennzeichnet. Auf die Frage „Wie stark treffen die folgenden Aussagen auf dich zu?" sollten die Probanden zu allen Aussagen eine subjektive Einschätzung abgeben, wobei der Fragebogen sechs ablehnende Einschätzungen gegenüber der Existenz paranormaler Phänomene enthielt. Die Ergebnisse dieser Aussagen wurden für die Datenauswertung umkodiert (z.B. statt des Wertes 1 eine 5). Insgesamt wurde versucht, den Fragebogen auf die studentische Stichprobe abzustimmen. So wurde beispielsweise auf Glücksbringer bei Klausuren verwiesen oder auf Mysteryserien wie z.B. *X-Faktor* und *Ghostwhisperer*, die sich bei Personen in diesem Altersbereich einer gewissen Beliebtheit erfreuen.

Des Weiteren wurden Subskalen des NEO-FFI (Borkenau & Ostendorf, 1993) zur Erfassung

der Persönlichkeitsdimensionen Extraversion und Neurotizismus sowie ein selbstkonstruierter Fragebogen zur Einstellung gegenüber paranormalen Phänomenen eingesetzt. Zur Anwendung kam auch ein Informationstext zum Thema Parapsychologie und eine Instruktion zu Entspannungsübungen (Krampen, 2004).

### *Versuchsablauf*

Der Versuch dauerte je nach Versuchsperson zwischen 40 und 60 Minuten. Zu jedem Termin wurden zwei Versuchspersonen rekrutiert. Ihre erste Aufgabe bestand darin, den Persönlichkeitsfragebogen sowie den Fragebogen zur Einstellung gegenüber paranormalen Phänomenen am Computer zu bearbeiten. Darauf folgten der einleitende Text zur Parapsychologie, welcher eine ernsthafte Beschäftigung mit dem Thema induzieren sollte, sowie ein schriftlicher Hinweis zum generellen Aufbau und Ablauf des Versuchs. Es schloss sich eine fünfminütige Entspannungsübung (Krampen, 2004) an, die auf dem Prinzip der progressiven Relaxation beruht und eine optimale Konzentration resp. Entfaltung der inneren Fähigkeiten bewirken sollte (vgl. Braud, 2002; Gatling & Rhine, 1946).

Zur Realisierung des ersten Faktors fand dann der dreiminütige Probedurchgang mit fünf Einzel-Trials statt, bei denen die Probanden das jeweilige Kartensymbol der Zenerkarten angeben sollten. Die Versuchsteilnehmer erhielten ein nach einer Zufallsliste variiertes Feedback, welches von der wahren Trefferquote unabhängig war.

Das Kartenexperiment erfolgte an drei Stationen. Die erste Station stellte die Telepathie-Bedingung dar. Bei diesem Vorgang der Gedankenübertragung fungierte eine Person zunächst als „Sender", der Partner im abgetrennten Nachbarraum als „Empfänger". Die Aufgabe des „Senders" bestand darin, nach jedem computergenerierten akustischen Signal, welches in Intervallen von je 15 Sekunden mit Hilfe zweier Lautsprecherboxen in den beiden Räumen dargeboten wurde, eine der 25 zuvor gemischten Zenerkarten zu „senden". Nach diesem Durchgang mit insgesamt 25 Karten wurden die Rollen gewechselt.

Danach gingen die Probanden zur zweiten Station, der Hellseh-Bedingung, über. Vor jedem Versuch wurden die 25 Karten neu gemischt und mit den Symbolen nach unten platziert. Die Aufgabe bestand darin herauszufinden, welches Zeichen sich unter welcher Karte verbarg.

Die dritte Station beschäftigte sich mit der Fähigkeit zur Präkognition. Hierfür lagen die 25 Zenerkarten gestapelt und etwas abseits von jedem Teilnehmer auf dem Tisch, anders als bei der Hellseh-Bedingung, bei der sich die Karten ausgebreitet und verdeckt vor der Versuchsperson befanden. Es galt zunächst, die Reihenfolge der Symbole der 25 Zenerkarten vorherzusagen. Erst nach einer solchen Vorhersage, die auf einem Zettel notiert wurde, kamen die Zenerkarten

ins Spiel. Die Versuchsleiterin mischte diese solange, bis der Proband „Stopp" sagte.

Für die Durchführung der Präkognitions- und der Hellseh-Durchgänge waren keine zeitlichen Begrenzungen vorgesehen. Zum Abschluss des Experiments wurden die Versuchsteilnehmenden abermals an den Computer gebeten, um einige Fragen über ihre Motivation bezüglich ihrer Teilnahme zu beantworten und ihre außersinnliche Fähigkeit einzuschätzen.

Bei allen drei Stationen achteten die Versuchsleiterinnen zwar darauf, dass keinerlei Betrugsmöglichkeiten von Seiten der Probanden ergriffen wurden. Allerdings wurden keinerlei spezifische Maßnahmen unternommen, um die diversen Möglichkeiten eines Manipulationsversuchs auszuschließen, da von derartigen Absichten der Probanden im Vorfeld nicht ausgegangen wurde. Dieser Aspekt wird im Diskussionsteil kritisch erörtert.

## *Stichprobe*

### Versuchspersonen

Insgesamt nahmen 96 Studierende an den Versuchen teil, wobei immer zwei Personen gleichzeitig zu einem Termin rekrutiert wurden. Daraus resultierten Versuchspersonenpaare. Die Mehrheit der Versuchsteilnehmer war weiblich (75 Frauen, 21 Männer). Das Durchschnittsalter der Probanden betrug 21.6 Jahre (*SD* = 3.0). Tabelle 1 enthält detaillierte weitere Angaben über die Stichprobenmerkmale.

**Tabelle 1:** Aufschlüsselung der Versuchsteilnehmer nach Geschlecht, Alter und Studienfach auf die drei Stufen der UV1 (Art der Rückmeldung im Probedurchgang).

| | | Art der Rückmeldung (im Probedurchgang) | | | Gesamt |
|---|---|---|---|---|---|
| | | kein Feedback | positives Feedback | negatives Feedback | |
| Versuchspersonenzahl | | 32 | 32 | 32 | 96 |
| Geschlecht | weiblich | 27 | 24 | 24 | 75 |
| | männlich | 5 | 8 | 8 | 21 |
| Alter | M | 21.2 | 21.6 | 22.0 | 21.7 |
| | SD | 2.0 | 3.9 | 3.0 | 3.1 |
| Psychologiestudierende | | 26 | 25 | 18 | 69 |
| andere Studienfächer | | 7[1] | 7[2] | 13[3] | 27 |

[1] Soziologie, Kunstgeschichte/Volkswirtschaft, Ägyptologie/Kunstgeschichte, Französisch/ Politik auf Lehramt, Jura, Englisch/Französisch auf Lehramt
[2] Informatik, Wirtschaftsinformatik, Französisch/Deutsch/Geschichte auf Lehramt, Französisch/ Spanisch/Computerlinguistik, Wirtschaftsmathematik, Englisch/Deutsch auf Lehramt, Theologie
[3] Informatik, Geschichte/Deutsch auf Lehramt, Soziologie, Geschichte/Politikwissenschaft, Geschichte/Soziologie, BWL, Germanistik/Politik, Geschichte/Politik/Philosophie, Jura, Pädagogik

### Versuchsleiterinnen

Das Experiment wurde von fünf Versuchsleiterinnen durchgeführt. Das Durchschnittsalter der Versuchsleiterinnen betrug 20.6 Jahre ($SD$ = 0.5). Sie führten die Untersuchung vornehmlich aufgrund ihres Interesses an der aufgeführten Fragestellung durch. Die Versuchsleiterinnen standen paranormalen Phänomenen teils offen bzw. neutral, teils kritisch gegenüber. Sie waren während des experimentellen Vorgangs weder optisch noch akkustisch gegenüber den Ergebnissen der Versuchsperson verblindet. Auch hinsichtlich des vorgegebenen fiktiven Feedbacks und deren vermuteter Wirkung auf die Probanden waren die Versuchsleiterinnen informiert. Während des Experiments waren die Probanden den Versuchsleiterinnen rückwärts zugewandt, so dass kein Blickkontakt bestand. Des Weiteren unterließen die Versuchsleiterinnen während des Experiments jegliche Kommunikation (mit Ausnahme der standardisierten Anweisungen) und stellten sicher, dass den Probanden lediglich Schreibutensilien zur Verfügung standen. Die Versuchsleiterinnen mischten zudem die Zenerkarten verdeckt und ordneten diese auf dem Tisch mit dem Symbol nach unten an.

## *Datenanalyse, Datenauswertung, psychometrische Kennwerte und Teststärkenbestimmung*

### Datenaggregation

Die erhobenen Daten wurden aus den Protokollbögen zu den Einzelversuchen und den Textdateien, die durch den Computer aus den Daten des Persönlichkeitsfragebogens und Fragebogens zur Einstellung gegenüber paranormalen Phänomenen sowie weiteren Angaben der Versuchspersonen automatisch erstellt wurden, in eine Excel-Datei (Microsoft Excel 2003) für deskriptiv- und inferenzstatistische Auswertungen überführt. Von den Protokollbögen wurden die Versuchspersonennummer, das Versuchspersonenpaar (dies gab darüber Auskunft darüber, welche Versuchspersonen gemeinsam die Station Telepathie mit einem Sender und einem Empfänger absolvierten), die Versuchsbedingung (positives, negatives und kein Feedback) sowie ein Kennwort übertragen, welches die Probanden im Versuch angegeben hatten.

Das Kennwort diente zur korrekten Zuordnung der Daten aus den Protokollbögen und der Textdatei aus der Computerbefragung. Des Weiteren wurden in die Excel-Datei die Einzelwerte zu den drei abhängigen Variablen Telepathie, Hellsehen und Präkognition eingetragen, wobei für jede der drei Stationen jeweils 25 Werte pro Versuchsperson übertragen wurden. Für einen Treffer wurde eine Eins, für keinen Treffer eine Null vergeben.

Im Gegensatz zu den Daten der Protokollbögen, die per Hand eingegeben werden mussten, wurden die Daten zum Persönlichkeitsfragebogen und zum Fragebogen zur Einstellung gegenüber paranormalen Phänomenen sowie die weiteren Angaben der Versuchspersonen automatisch in zwei Textdateien übertragen. Diese mussten lediglich über die Import-Funktion von Microsoft Excel in zwei Reiter der Excel-Datei übermittelt werden. Beide Textdateien enthielten zu jeder Versuchsperson die Startzeit in Minuten sowie das Datum der Durchführung des Versuchs. Außerdem wurden IP-Adresse des Computers, verwendeter Browser inklusive verwendetes Betriebssystem, die Frage, ob der Fragebogen vorzeitig abgebrochen wurde, sowie ein Kennwort zur Zuordnung zu den Daten der Protokollbögen protokolliert. Die erste Textdatei enthielt darüber hinaus noch Daten zu den beiden Subskalen Extraversion und Neurotizismus des NEO-FFI (Borkenau & Ostendorf, 1993) sowie Antworten zur Einstellung gegenüber paranormalen Phänomenen. Die zweite Textdatei erfasste hingegen nachträgliche Fragen zur Motivation, Einstellung und Stimmungslage während der Versuche sowie persönliche Angaben der Versuchspersonen wie Alter, Studienfach, Geschlecht, E-Mail-Adresse (freiwillige Angabe), eine persönliche Einschätzung der eigenen außersinnlichen Fähigkeiten sowie Kommentare und Anmerkungen zu dem Versuch.

## Datenauswertung

Die Gesamttrefferzahlen für die einzelnen Versuchspersonen, aufgeschlüsselt für die drei abhängigen Variablen, sowie zahlreiche weitere deskriptiv- und inferenzstatistische Berechnungen wurden in Microsoft Excel 2003 und SPSS 13.0 berechnet. Die inferenzstatistische Auswertung des Datensatzes erfolgte mit Hilfe der Binomialverteilung, bei der gegen einen theoretischen Erwartungswert getestet wird. Die Auswertung mittels Binomialverteilung stellt für den vorliegenden Datensatz die exakte von mehreren möglichen Auswertungsmethoden dar (vgl. z.B. Burdick & Kelly, 1977).

## Psychometrische Kennwerte des selbsterstellten Fragebogens

Für die untersuchte Stichprobe von 96 Probanden ergab sich ein Gesamtmittelwert von 3.06 ($SD = 0.54$) auf der fünfstufigen Likert-Skala (zur Codierung in Zahlenwerte wurden die Werte 1 bis 5 verwendet). Während der Mittelwert (im Hinblick auf Boden- und Deckeneffekte) als nahezu optimal betrachtet werden kann, fällt die Standardabweichung eher gering aus. Durch-

führungs- und Auswertungsobjektivität können aufgrund der Computernutzung als gewährleistet gelten.

Zur Überprüfung der internen Konsistenz des neu entwickelten Fragebogens zur Einstellung gegenüber paranormalen Phänomenen lässt sich der Cronbachs-α-Wert berechnen. Da dieser Wert allerdings von der Stichprobengröße abhängig ist und sich für größere Stichproben trotz steigender Fehlermöglichkeiten automatisch eine höhere Reliabilitätsschätzung ergibt (z.B. Sponsel, 2004), sollte Cronbachs α nicht vorbehaltlos betrachtet werden. Dennoch hat sich der Cronbachs-α-Wert in der psychologischen Forschung etabliert und wird demnach auch hier verwendet. Der Cronbachs-α-Wert für die Skala zur Einstellung gegenüber paranormalen Phänomenen beträgt .818. Dieser Wert liegt im Vergleich zu den Skalen für Neurotizismus (.814) und Extraversion (.817) auf einem ähnlich hohen Niveau. Die Berechnung der Trennschärfen (mit „Part-whole-Korrektur") weist für 14 der 19 Items einen Wert von über 0.3 aus und bestätigt damit weitgehend die guten Ergebnisse des Cronbachs-α-Wertes. Belege zur Validität des selbsterstellten Fragebogens liegen bisher nicht vor. Lediglich die geringen Korrelationen zu den Extraversions- ($r$ = .05) und Neurotizismuswerten ($r$ = .22) deuten auf eine hinreichend diskriminante Validität zu diesen Konstrukten hin.

## Teststärkenbestimmung

Die Teststärkenberechnungen erfolgten mittels GPOWER 3.0 (Faul *et al.*, 2007). Bei der Festlegung der Effektstärke wurde auf die Konventionen von Cohen (1988) zurückgegriffen und ein „kleiner Effekt" ($h$ = 0.2) gewählt. Diese Vorgehensweise stellt sicher, dass bei ausreichender Teststärke auch noch kleine auftretende Effekte detektiert werden können. Das Alphafehlerniveau wurde gemäß gängigen Konventionen (z.B. Bortz, 2005) auf .05 gesetzt. Bei einer Stichprobe von 96 Studierenden und einer 25-maligen Wiederholung des Rateversuches ergeben sich für die drei ASW-Bedingungen Telepathie, Hellsehen und Präkognition jeweils 2400 Messungen. Die resultierenden Teststärken von jeweils $1-\beta > .99$ können als sehr hoch betrachtet werden und übertreffen gängige Konventionen (z.B. Bortz, 2005) von $1-\beta \geq .80$ überaus deutlich.

Bezüglich des postulierten Moderatoreffekts zwischen den Trefferraten für Telepathie, Hellsehen und Präkognition und den Fragebogenwerten zu Neurotizismus, Extraversion und Einstellung zu paranormalen Phänomenen wurde ebenfalls Bezug auf die Konventionen von Cohen (1988) genommen. Hier ergeben sich bei Annahme eines „mittleren Effektes" ($r$ = .3) und $\alpha$ = .05 Teststärken von jeweils $1-\beta$ = .92. Bei Annahme eines „kleinen Effektes" ($r$ = .1) reduziert sich die Teststärke allerdings drastisch ($1-\beta$ = .25).

## *Ergebnisse*

### *Hypothese 1*

Hypothese 1 besagt, dass die paranormalen Fähigkeiten Telepathie, Hellsehen und Präkognition überzufällig hohe Trefferraten im Kartenerkennen bewirken.

Bei insgesamt 2400 Versuchen in der Telepathie-Bedingung wurden 493 Treffer erzielt, beim Hellsehen hingegen 490 Treffer. Zu erwarten waren jeweils 480 Treffer. Im Präkognitions-Versuch wurden mit 456 weniger als die erwarteten 480 Treffer beobachtet.

Die inferenzstatistische Auswertung per Binomialverteilung ergibt in der Telepathie-Bedingung mit $p = .26$, in der Hellseh-Bedingung mit $p = .31$ und in der Präkognitions-Bedingung mit $p = .89$ ein auf dem Signifikanzniveau von $\alpha = .05$ statistisch nicht bedeutsames Ergebnis.

Die Hypothese, dass Telepathie, Hellsehen und Präkognition bei der Auswahl von Zenerkarten zu überzufällig hohen Trefferraten führen, kann unter der Annahme eines sehr kleinen Effekts auf Basis dieser Untersuchung zurückgewiesen werden.

Abbildung 1 stellt die beobachteten und erwarteten, binomialverteilten Häufigkeiten der Treffer über alle Faktorstufen der drei unabhängigen Variablen hinweg dar.

### *Hypothese 2*

Hypothese 2 postuliert, dass der Einfluss von Telepathie, Hellsehen und Präkognition auf die Auswahl von Zenerkarten durch die (fingierte) Rückmeldung der Fähigkeiten zur ASW sowie durch die Einstellung gegenüber paranormalen Phänomenen moderiert wird.

Die Wahrscheinlichkeiten für die zweite Hypothese der Untersuchung zeigen unter der Binomialverteilung, aufgeteilt nach Feedback und den drei Formen der ASW, keine *p*-Werte unter dem angenommenen Signifikanzniveau von $\alpha = .05$ (siehe Tabelle 2).

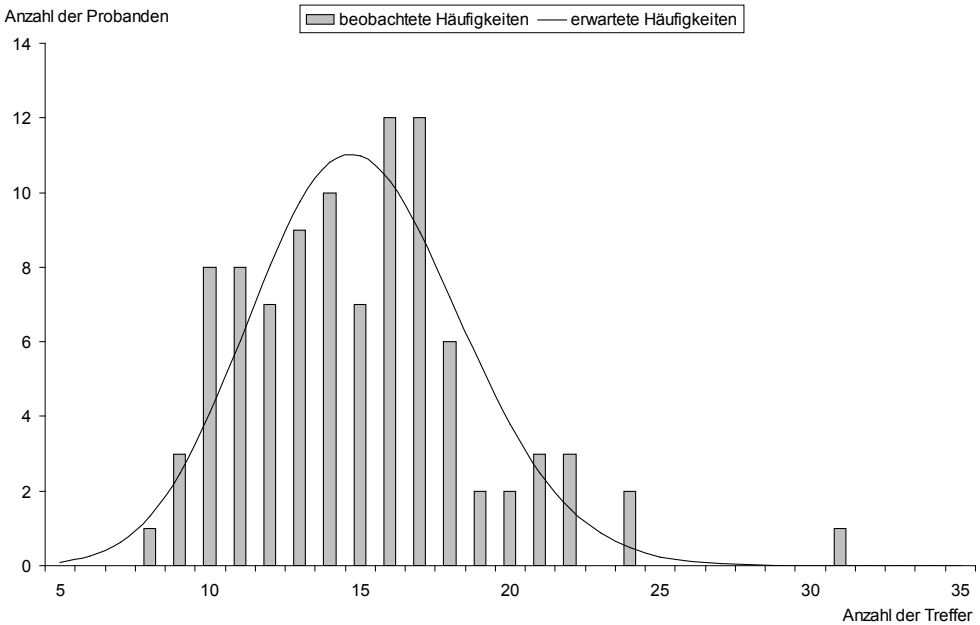


**Abbildung 1:** Beobachtete und erwartete, binomialverteilte (geglättete) Häufigkeiten der Treffer über alle Faktorstufen der drei unabhängigen Variablen hinweg. Der Erwartungswert für eine einzelne Versuchsperson beträgt 15 (Anzahl an Treffern).

| Feedback | Telepathie | Hellsehen | Präkognition | Gesamttreffer (ASW) |
|---|---|---|---|---|
| ohne Feedback | .34 | .41 | .90 | .67 |
| positives Feedback | .28 | .31 | .80 | .47 |
| negatives Feedback | .38 | .38 | .44 | .35 |

**Tabelle 2:** Darstellung der Wahrscheinlichkeiten (*p*-Werte) unter der Binomialverteilung aufgeteilt nach Feedback-Bedingung und ASW-Bedingung.

Punkt-Moment-Korrelationen ($r$ zwischen -.18 und .12). zwischen Treffern in den ASW-Bedingungen und den Fragebogenwerten zu Neurotizismus, Extraversion und Einstellung zu paranormalen Phänomenen demonstrieren ebenfalls keine signifikant von Null verschiedenen Korrelationen ($.08 \leq p \leq .99$). Tabelle 3 liefert einen Überblick über die Korrelationen der genannten Variablen.

| Fragebogen | Telepathie | Hellsehen | Präkognition | Gesamttreffer (ASW) |
|---|---|---|---|---|
| Neurotizismus | .00 (.99) | -.05 (.63) | .12 (.26) | .04 (.70) |
| Extraversion | -.16 (.11) | -.01 (.89) | .08 (.46) | -.05 (.63) |
| Einstellung gegenüber paranormalen Phänomenen | .01 (.93) | -.18 (.08) | .05 (.66) | -.07 (.51) |

**Tabelle 3:** Korrelationen und deren $p$-Werte (zweiseitige Testung) zwischen den Fragebogenwerten zu Neurotizismus, Extraversion sowie Einstellung gegenüber paranormalen Phänomenen und den ASW-Bedingungen.

Auf Basis der ausreichenden Teststärke für einen mittleren Effekt kann auch die zweite Hypothese, dass die (fingierte) Rückmeldung der Fähigkeiten zur ASW sowie die Einstellung gegenüber paranormalen Phänomenen den Einfluss von Telepathie, Hellsehen und Präkognition die Auswahl von Zenerkarten moderieren, verworfen werden.

### *Sonstige Befunde*

Eine Auswertung der Untersuchung, aufgeschlüsselt nach den einzelnen Versuchsleiterinnen, erbrachte keinerlei statistisch überzufällige Trefferraten. Sowohl bei gemeinsamer Berücksichtigung der drei paranormalen Modalitäten ($.10 \leq p \leq .88$) als auch bei getrennter Analyse von Telepathie ($.19 \leq p \leq .82$), Hellsehen ($.13 \leq p \leq .85$) und Präkognition ($.26 \leq p \leq .96$) zeigte sich bei keiner der fünf Versuchsleiterinnen ein statistisch bedeutsames Ergebnis. Aufgrund der fehlenden Signifikanzen erübrigt sich eine Adjustierung des Alphafehlers.

Unabhängig von dem globalen, nicht signifikanten Ergebnis lassen sich für eine Person außergewöhnlich hohe Trefferanzahlen feststellen (Telepathie: 12, Hellsehen: 9, Präkognition: 10 Treffer, über alle ASW-Bedingungen folglich 31 von 75 möglichen Treffern [41.3% statt der 20% zu erwartender Treffer]). Ein solches Ergebnis kommt wahrscheinlichkeitstheoretisch in nur etwa 2 von 100.000 Fällen zustande. Bei einer weiteren Testung der ASW-Fähigkeiten dieser Person ergab sich erneut eine überzufällige Trefferanzahl von 9 in der Präkognition-Bedingung. Jedoch nur noch 3 Treffer in der Hellseh-Bedingung festgestellt. Die gesamte Tref-

ferzahl dieser Person kommt somit nur in etwa 5 von 100.000 Fällen zustande, wobei der zuletzt gezeigte Abfall der Trefferquote sich im Sinne des Decline-Effekts (z.B. Lucadou, 1997) interpretieren ließe. Der Proband selbst führte in dem anschließenden Gespräch mit dem Versuchsleiter auftretende Ermüdungserscheinungen an.

Eine Betrachtung der Trefferverläufe über die jeweils 25 „Kartenvorhersagen" zeigt keine statistisch bedeutsame Veränderung der Auswahl von Zenerkarten über die Zeit (siehe Abbildung 2). Die Verläufe bewegen sich relativ zufällig um den erwarteten Mittelwert von 19.2 Treffern (1/5 mal 96 Teilnehmer) zu jedem der insgesamt 75 Messzeitpunkte. Es bestehen keine signifikanten Zusammenhänge zwischen der Trefferrate und dem Zeitpunkt, weder bei Telepathie ($r$ = .11; $p$ = .61) noch bei Hellsehen ($r$ = -.22; $p$ = .30) oder bei Präkognition ($r$ = .35; $p$ = .09). Wenngleich die abnehmenden Trefferhäufigkeiten aus Hypothese 1 für Telepathie, Hellsehen und Präkognition einen Decline-Effekt deskriptivstatistisch stützen, so zeigen die Vorzeichen der Korrelationen, dass lediglich unter der Hellseh-Bedingung ein schwacher, nichtsignifikanter Decline-Effekt in Erscheinung tritt. Demnach konnte ein Decline-Effekt in der vorliegenden Studie nicht nachgewiesen werden, der besagt, dass die Trefferleistungen im Verlauf der Untersuchung absinken.

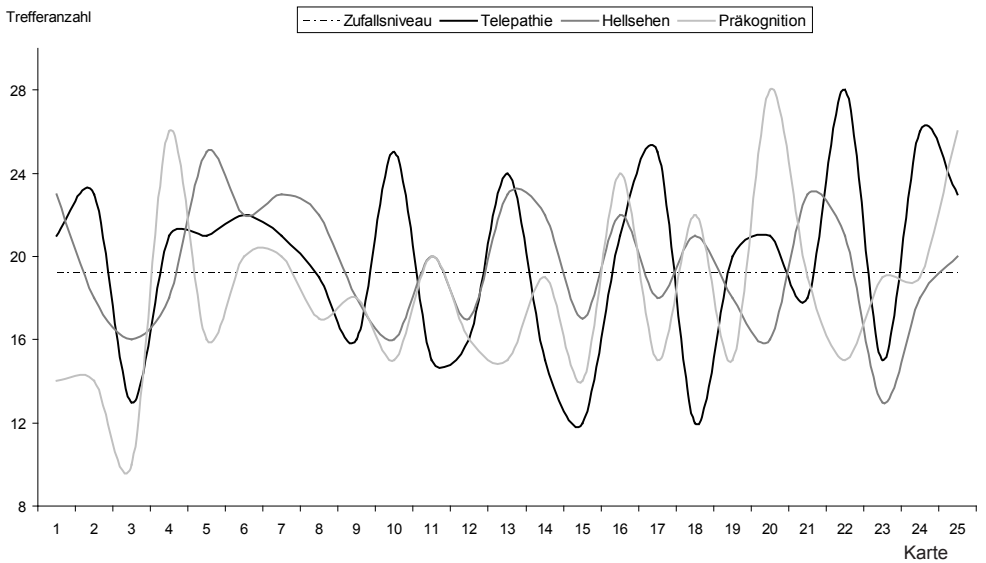


**Abbildung 2:** Zeitlicher Verlauf (geglättet) der Trefferhäufigkeiten, aufgeschlüsselt für die Bedingungen Telepathie, Hellsehen und Präkognition. Das Zufallsniveau für jeden liegt bei 19.2 Treffern (96 Versuchspersonen x 0.2 Trefferwahrscheinlichkeit).

### *Diskussion*

In der vorliegenden Untersuchung wurden keinerlei überzufällige Trefferraten bei der Auswahl von Zenerkarten verzeichnet und somit ASW nicht nachgewiesen. Auch moderierende Einflüsse durch (fingierte) Rückmeldungen der Fähigkeiten zur ASW sowie der Einstellung gegenüber paranormalen Phänomenen auf den Einfluss von Telepathie, Hellsehen und Präkognition bei der Auswahl von Zenerkarten zeigen sich nicht. Aufgrund der ausreichenden Power (≥ .92) kann die Nullhypothese – zumindest für die jeweils postulierten kleinen bzw. mittleren Effekte (siehe Methodenteil) – angenommen werden. Der ermittelte Befund widerspricht der uneinheitlichen Befundlage in der Literatur zur ASW nicht.

Interessant erscheint die außergewöhnlich hohe Trefferzahl einer einzelnen Versuchsperson, die die Vermutung nahe legt, dass – sofern ASW überhaupt existiert – diese nur bei wenigen Personen unter spezifischen Situationen in Erscheinung tritt. Träfe dies zu, so wäre es sinnvoll, sich auf diese zuvor ausgewählten Personen in Folgestudien zu konzentrieren. Zu beachten ist jedoch, dass das außergewöhnliche Ergebnis des Probanden auch durch einen Betrug oder auf andere Weise (z.B. „sensory leakage", d.h. einer Informationsübertragung über konventionelle, nicht paranormale Wege) zustande gekommen sein könnte, da keinerlei spezifische Kontroll- und Sicherheitsmaßnahmen gegenüber Manipulationsmöglichkeiten vorgenommen wurden.

Insgesamt weist das durchgeführte Experiment diverse (methodische) Schwächen auf. Neben den bereits erwähnten fehlenden Maßnahmen gegenüber Betrugsmöglichkeiten sind die selbsterstellten Zenerkarten zu nennen, die durch geringfügige Individualmerkmale eine Diskriminierung der einzelnen Karten ermöglicht haben könnten. Auch das Mischen per Hand kann an dieser Stelle moniert werden. Eine weitere Beanstandung bezieht sich auf die gerichtet formulierte erste Hypothese, die eine überzufällig unterdurchschnittliche Trefferrate der Probanden („psi-missing") unberücksichtigt lässt. Erst im Verlauf der Untersuchung hat sich herausgestellt, dass eine hinreichende Anzahl an Probanden für die Studie rekrutiert werden konnte, die auch eine zweiseitige Testung unter Berücksichtigung von „psi-missing" mit akzeptabler Teststärke erlaubt hätte. Außerdem kann bemängelt werden, dass die experimentell induzierte Motivation der ersten unabhängigen Variable nicht überprüft wurde. Es ist somit nicht sichergestellt, dass die fingierten Rückmeldungen tatsächlich zu drei unterschiedlichen motivationalen Zuständen geführt haben.

Der Einsatz eines selbstkonstruierten Fragebogens zur Einstellung gegenüber paranormalen Phänomenen ist für Kritik ebenfalls anfällig. Auch wenn es innerhalb der parapsychologischen Forschung üblich ist, selbstkonstruierte Tests einzusetzen (Goulding & Parker, 2001) und deren Fragen – wie im vorliegenden Fall – auf die verwendete Stichprobe abzustimmen, so hätte man für die vorliegende Untersuchung eher einen etablierten und bewährten Fragebogen wie beispielsweise die „Paranormal Belief Scale" (Tobacyk & Milford, 1983) oder die „Australi-

an Sheep-Goat Scale“ (Thalbourne & Haraldsson, 1980) heranziehen sollen.

Weitere Kritikpunkte betreffen die fehlende Verblindung der Versuchsleiterinnen, die Vertrautheit zwischen den Versuchspersonen sowie die Stichprobenrepräsentativität. Beispielsweise kann Telepathie möglicherweise deshalb nicht detektiert werden, weil fehlende Vertrautheit bzw. Fremdheit zwischen den Versuchspersonen die direkte psychische Informationsübertragung beeinträchtigt haben könnte. Zudem ist die verwendete Gelegenheitsstichprobe nicht repräsentativ, da es sich bei allen Probanden um Studierende der Universität Trier handelt, was die externe Validität der Studie deutlich einschränkt.

Zusammenfassend kann festgehalten werden, dass das vorliegende Experiment im Falle positiver Befunde in diverser Hinsicht hätte kritisiert und in Frage gestellt werden müssen. In diesem Falle wäre beispielsweise unklar geblieben, ob etwaige signifikante Ergebnisse aufgrund von gezielten Manipulationsversuchen der Probanden oder aber aufgrund der Existenz von ASW eingestellt haben. Da jedoch insgesamt keinerlei überzufällige Befunde im Sinne eines Nachweises von ASW aufgetreten sind, ist anzunehmen, dass die oben aufgeführten potentiellen Alternativerklärungen in der Studie keine bzw. in statistischer Hinsicht lediglich eine unbedeutende Rolle gespielt haben. Folgestudien sollten den aufgeführten diversen Kritikpunkten jedoch eine deutlich größere Beachtung schenken.

## Literatur

Bandura, A. (1986). *Social Foundations of Thought and Action: A Social Cognitive Theory*. Englewood Cliffs, NJ: Prentice-Hall.

Bem, D.J., & Honorton, C. (1994). Does psi exist? Replicable evidence for an anomalous process of information transfer. *Psychological Bulletin*, 115, 4-18.

Borkenau, P., & Ostendorf, F. (1993). *NEO-Fünf-Faktoren-Inventar (NEO- FFI) nach Costa und McCrae, Handanweisung*. Göttingen: Hogrefe.

Bortz, J. ([6]2005). *Statistik für Human- und Sozialwissenschaftler* (6. Aufl.). Berlin: Springer.

Braud, W. (2002). Psi-favorable conditions. In Rammohan, V.G. (Ed.), *New Frontiers of Human Science: A Festschrift for K. Ramakrishna Rao* (S. 95-118). Jefferson, NC: McFarland.

Burdick, D.S., & Kelly, E.F. (1977). Statistical methods in parapsychological research. In Wolman, B.B. (Ed.), *Handbook of Parapsychology* (S. 81-130). New York: Van Nostrand Reinhold.

Cohen, J. ([2]1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale, NJ: Laurence Erlbaum.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, Instruments*

*& Computers,* 39, 175-191.

Gatling, W., & Rhine, J.B. (1946). Two groups of PK subjects compared. *Journal of Parapsychology*, 10, 120-125.

Goulding, A., & Parker, A. (2001). Finding psi in the paranormal: Psychometric measures used in research on paranormal beliefs/experiences and in research on psi-ability. *European Journal of Parapsychology*, 16, 73-101.

Honorton, C., Ferrari, D.C., & Bem, D.J. (1998). Extraversion and ESP performance: A meta-analysis and a new confirmation. *Journal of Parapsychology*, 62, 255-276.

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

Kennedy, J.E. (2005). Personality and motivations to belief, misbelieve, and disbelief in paranormal phenomena. *Journal of Parapsychology*, 69, 263-292.

Krampen, G. (2004). Differentielle Indikation von autogenem Training und progressiver Relaxation. *Entspannungsverfahren*, 21, 6-27.

Lawrence, T.R. (1993). Gathering in the sheep and goats... A meta-analysis of forced-choice sheep-goat ESP studies, 1947-1993. In *The Parapsychological Association 36th Annual Convention. Proceedings of Presented Papers* (S. 75-86). Durham, NC: The Parapsychological Association.

Lucadou, W. von (1997). *Psi-Phänomene. Neue Ergebnisse der Psychokinese-Forschung*. Frankfurt/M.: Insel.

Milton, J., & Wiseman, R. (2001). Does psi exist? Reply to Storm and Ertel (2001). *Psychological Bulletin*, 127, 434-438.

Palmer, J., & Carpenter, J.C. (1998). Comments on the extraversion-ESP meta-analysis by Honorton, Ferrari, and Bem. *Journal of Parapsychology*, 62, 277-282.

Pratt, J.G, Rhine, J., Smith, B., Stuart, C., & Greenwood, J. (1940). *Extrasensory Perception After Sixty Years*. Boston, MA: Bruce Humphries.

Schmeidler, G.R. (1952). Personal values and ESP scores. Journal of Abnormal and Social Psychology, 47, 757-761.

Schmeidler, G.R., & McConnell, R.A. (1958). *ESP and Personality Patterns*. New Haven, CT: Yale University Press.

Sponsel, R. (2004). Cronbachs alpha. Von der numerologischen Kunst, eine „Reliabilität" aus dem Nichts zu zaubern. http://www.sgipt.de/wisms/mtt/tgk/calpha.htm [Zugriff: 16. Juli 2007].

Steinkamp, F. (2005). Forced-choice ESP experiments: Their past and their future. In Thalbourne, M.A., & Storm,L. (Eds.), *Parapsychology in the Twenty-first Century* (S. 124-163). Jefferson, NC: McFarland.

Storm, L., & Ertel, S. (2001). Does psi exist? Comments on Milton and Wiseman's (1999) meta-analysis of Ganzfeld research. *Psychological Bulletin*, 127, 424-433.

Thalbourne, M.A., & Haraldsson, E. (1980). Personality characteristics of sheep and goats. *Personality and Individual Differences*, 1, 180-185.

Tobacyk, J.J., & Milford, G. (1983). Belief in paranormal phenomena: Assessment instrument development and implications for personality functioning. *Journal of Personality and Social Psychology*, 44, 1029-1037.

## *Anhang*

**Fragebogen 1***: Verwendeter Fragebogen zur Erfassung der Einstellung gegenüber paranormalen Phänomenen*

Wie stark treffen die folgenden Aussagen auf dich zu?
1 = diese Aussage trifft auf mich absolut nicht zu.
5 = diese Aussage trifft auf mich absolut zu.

1. Ich glaube daran, dass es Dinge gibt, die nicht durch physikalische Gesetze erklärbar sind.
2. Mir fallen spontan Situationen ein, die ich mir nicht durch logisches Denken erklären kann.
3. Ich glaube an Horoskope und passe mein Verhalten daran an.
4. Ich besitze Tarotkarten oder ähnliches und benutze diese.
5. Ich schaue gerne Mysteryserien (z.B. x-Faktor, Ghostwhisperer) an.
6. Manchmal habe ich das Gefühl, beim Spiel den Würfelfall beeinflussen zu können.
7. Ich denke nicht, dass das Schicksal einen Einfluss auf mein Leben hat.
8. Ich habe gewisse Rituale oder Talismänner,[4] mit denen ich glaube[,] mein Glück beeinflussen zu können (z.B. Glücksbringer bei Klausuren).
9. Ich kann mir nicht vorstellen, dass manche Menschen außergewöhnliche Fähigkeiten haben (z.B. Telepathie).
10. Ich denke, dass Träume manchmal einen Hinweis auf Zukünftiges geben können.
11. Ich glaube an Gott.
12. Ich glaube an Wiedergeburt, Leben nach dem Tod oder ähnliches.
13. Ich glaube an den Sinn des Lebens und an das Gute in der Welt.
14. Ich kann mir nicht vorstellen, dass es Leben auf anderen Planeten gibt.
15. Ich glaube daran, dass es Menschen gibt, die eine Vorahnung haben.
16. Ich denke nicht, dass man zu Toten Kontakt aufnehmen kann.
17. Ich glaube nicht, dass eine optimistische Einstellung positiven Einfluss haben kann.
18. Der Gedanke daran, dass Übersinnliches existiert[,] macht mir Angst.
19. Ich hatte noch nie ein Déjà-vu-Erlebnis.

4 Der korrekte Plural von „Talisman“ wäre „Talismane“ gewesen. (Red.)

**Fragebogen 2**: *Fragen, die am Ende der Untersuchung zum Einsatz kamen*

1. Meine Motivation während des Experiments war (sehr gering – sehr hoch).
2. Ich fühle mich (sehr entspannt – sehr angespannt).
3. Ich habe während des Experiments versucht[,] positive Resultate zu erzwingen (trifft überhaupt nicht zu – trifft voll und ganz zu).
4. Meinen Stimmungszustand würde ich derzeit folgendermaßen bezeichnen (sehr gut – sehr schlecht).
5. Ich hatte dem Experiment gegenüber eine positive Einstellung (trifft überhaupt nicht zu – trifft voll und ganz zu).
6. Ich glaube daran[,] bei dem Versuch erfolgreich gewesen zu sein (trifft überhaupt nicht zu – trifft voll und ganz zu).
7. Ich habe Angst vor unheimlichen, nicht rational erklärbaren Dingen (trifft überhaupt nicht zu – trifft voll und ganz zu).
8. Meine Konzentration während der Untersuchung war (sehr gut – sehr schlecht).
9. Ich habe das Gelingen meiner Handlungen erwartet (trifft überhaupt nicht zu – trifft voll und ganz zu).
10. Ich glaube[,] parapsychologische Fähigkeiten zu besitzen (trifft überhaupt nicht zu – trifft voll und ganz zu).
11. Ich glaube, dass es Menschen gibt, die parapsychologische Fähigkeiten besitzen (trifft überhaupt nicht zu – trifft voll und ganz zu).
12. Ich glaube[,] mit Training meine parapsychologischen Fähigkeiten verbessern zu können (trifft überhaupt nicht zu – trifft voll und ganz zu).
13. Ich fand das Experiment langweilig (trifft überhaupt nicht zu – trifft voll und ganz zu).
14. Alter
15. Studienfach
16. Geschlecht
17. E-Mail (freiwillig)
18. Mit welcher Strategie bist Du bei unserem Experiment zu Deinen Antworten gelangt?
    a. Ich habe geraten.
    b. Ich habe versucht[,] logisch abzuwägen und zu zählen.
    c. Ich habe auf mein Gefühl gehört.
    d. Ich habe es gewusst.
    e. Ich habe folgende Strategie verwendet:

# *Kommentare zu Rey* et al.*:*

# *Konzeptuelle Replikationsstudie zu Experimenten zur außersinnlichen Wahrnehmung*

Wolfgang Ambach[5]

## Dieser Betrieb bildet aus

Der Beitrag von Rey *et al.* wird mit einer ungewöhnlichen redaktionellen Vorbemerkung eingeleitet: Die Studie entspreche nicht dem methodischen Standard, mit dem das historische Paradigma bei seiner letzten Anwendung vor über 50 Jahren verbunden war. Es werden zweierlei publikationspolitische Überlegungen dargelegt, aufgrund derer man den Artikel dann trotzdem zur Publikation angenommen habe.

### *Rechnen Sie mit dem Schrecklichsten*

Ein nettes Vorwort, und es spricht Bände. Der Leser erfährt, quasi bevor er zu Lesen beginnt, (a) dass hier ein offenbar veraltetes Forschungsparadigma zum Einsatz kam und noch dazu in dilettantischer Weise umgesetzt wurde. Weiterhin erfahren wir, (b) dass eine Nichtveröffentlichung – offenbar ungeachtet ihrer Gründe – Vorwürfen der Datenselektion Vorschub geleistet hätte, was man durch den Entschluss zur Veröffentlichung vermied. Schließlich nehmen wir zur Kenntnis, (c) dass der aufmerksamkeitswürdige Umstand, dass die Studie im Rahmen der studentischen Ausbildung an einer deutschen Universität durchgeführt wurde, weiterer Grund für die Annahme des Artikels war.

Dieses Kleingedruckte als Fußnote zum Aufsatztitel erzählt nach meiner Auffassung allerdings weniger über den Artikel selbst als über die Bedingungen und das Spannungsfeld, denen die Forschung zu unkonventionellen Fragestellungen, aus der hier ein Beispiel präsentiert wird, generell ausgesetzt ist.

Ad (a): Während das Redaktionskollegium das Aufgreifen eines historischen Ansatzes an sich vielleicht noch mehrheitlich gutgeheißen hätte, wurde der Bruch mit früher zu diesem Ansatz etablierten Standards offenbar sehr bemängelt. Für die damals üblichen aufwendigen Vorkehrungen gegen Täuschung und sensorische Lecks gab und gibt es sicherlich gute Gründe. Dennoch möchte ich im Rückblick zu bedenken geben, dass auch die frühere Maximierung von

5 Dr. med. Wolfgang Ambach leitet die Forschungsgruppe Klinische und Physiologische Psychologie, Institut für Grenzgebiete der Psychologie und Psychohygiene (IGPP), Freiburg i.Br.

Abschirmung und Kontrolle zu keinem Zeitpunkt geeignet war, die Möglichkeit artifizieller Einflüsse gänzlich (und vor allem: in aller Augen!) auszuschließen. Gleichzeitig waren die Optimierungsversuche – so vermute ich – nicht nur durch forscherische Neugier motiviert, sondern auch durch das Bestreben, den Kampf gegen die Überzeugungsgegner, die Ungläubigen, eines Tages doch gewinnen zu können. Rey *et al.* haben mit dieser Tradition einfach gebrochen; sie waren einfach nur neugierig, ohne gleich wasserdichte Beweise liefern zu wollen. Die Autoren kämpften eindeutig nicht den gleichen Kampf wie die Forscher in der Rhine'schen Tradition, und sie hatten offenbar auch nicht das gleiche imaginierte Gegenüber. Ich frage mich, inwieweit die Kontroverse um die Publikationswürdigkeit des Artikels, soweit sie sich auf die Manipulationsvorkehrungen bezieht, tatsächlich wissenschaftlich motiviert ist, und in wie weit sie die verschiedenen Traditionszugehörigkeiten der Autoren und der einzelnen Reviewer widerspiegelt.

Ad (b): Selektive Publikation von Artikeln mit „Positivergebnissen" leistet nicht nur dem Vorwurf der Datenselektion Vorschub, sondern *ist* Datenselektion. An dieser Stelle wird deutlich, wie unfrei die Entscheidung über die Annahme oder Ablehnung eines Artikels wird, wenn ständig ein Überzeugungsgegner mit bereitgehaltenen Vorwürfen präsent ist, dem man keine Angriffsfläche bieten will. Das Problem des Selektionsvorwurfes, den man möglichst vermeiden will, vermischt sich dann mit der wissenschaftlichen Bewertung eines Artikels. Es wird vermutlich schwierig, vor lauter Vermeidung eines Bias in die eine Richtung nicht einen gegenteiligen entstehen zu lassen. Im vorliegenden Fall hat die Befürchtung von Vorwürfen offenbar überwogen und damit dem Artikel weitergeholfen.

Ad (c): Dass eine Studie zur außersinnlichen Wahrnehmung in der universitären Ausbildung Platz findet, ist zweifellos eher selten. Dies als Publikationskriterium zu nehmen, spiegelt allerdings deutlich wider, wie wenig zufrieden man mit der Nebenrolle ist, die den unkonventionellen Fragestellungen, etwa der hier behandelten Frage nach außersinnlicher Wahrnehmung, in Scientific Community und Lehre zukommt. Um dem Forschungsthema der außersinnlichen Wahrnehmung aus seinem Schattendasein herauszuhelfen, wurde hier offenbar ein Artikel gleichzeitig als zweitklassig eingestuft und dennoch durchgewunken.

Nimmt man (a), (b) und (c) zusammen, kann man zu der Schlussfolgerung gelangen, dass Forschung in Grenzgebieten mit mehr Selbstbewußtsein erfüllt sein sollte und sich mehr durch Wissen-Wollen als durch Vermeiden oder Besiegen von Gegenargumenten definieren sollte. Eine weitere Folgerung könnte sein, Forschung in Grenzgebieten nicht als abgesetzt („para") von der restlichen Wissenschaft zu begreifen, sondern als Teil *der* Wissenschaft. Dies würde allerdings bedeuten, dass zwar die Motivation für eine bestimmte (auch unkonventionelle) Fragestellung den eigenen Neigungen und Glaubenssätzen folgen darf, nicht aber die Interpretation von Studienergebnissen.

### *Wissenschaft und Lehre: Ein Loblied auf die Naivität*

Angesichts der vielen Jahrzehnte, in denen die Erforschung des Paranormalen sich zu einer Wissenschaft „neben" („para") der Wissenschaft entwickelt hat, klingen die eben angestellten Überlegungen, die Vorstellung *einer* Wissenschaft mit für alle gleichermaßen gültigen Axiomen, sicher visionär, vielleicht auch sehr naiv.

Der Artikel von Rey *et al.* gibt einer solchen naiven Sichtweise allerdings recht. Er zeigt exemplarisch und in weitgehend brillianter Weise auf, dass es sehr wohl möglich ist, ein umstrittenes Phänomen unvoreingenommen und ohne Überzeugungswillen wissenschaftlich zu untersuchen und darzustellen.

Es ist weder Nachteil noch Zufall zu nennen, dass die Autoren Uraltmaterial zur Grundlage ihrer Studie erkoren haben. Nach der redaktionellen Vorbemerkung hätte man vielleicht eine Art „J.B.-Rhine-Gedenkstudie" erwarten können, ganz auf Nachahmung des Äußeren bedacht. Aber auf den zweiten Blick wird deutlich, dass zwar die Versuchsidee und die Zenerkarten historisch sind, nicht aber die wissenschaftliche Aufbereitung der Versuchsergebnisse, die sich an aktuelle Standards anlehnt. Ich betrachte es auch nicht als Nachteil, dass keine Perfektion in den Details der Abschirmung betrieben wurde. Die Tatsache, dass, anders als zu J.B. Rhines Zeiten, auf langwierig optimierte Vorkehrungen gegen artifizielle Einflüsse, z.B. Täuschung oder Versuchsleiter-Einflüsse, verzichtet wurde, macht die Ergebnisse freilich, wie die Autoren ja auch diskutieren, eingeschränkt extrapolierbar. Insbesondere ein Sprengen des konventionellen Erklärungsrahmens würde in keinem Ergebnisfall zwingend sein. Ich vermute aber stark, dass es nie ernstgemeinte Absicht der Autoren war, einen gegen diverse Gegenargumente resistenten Nachweis außersinnlicher Wahrnehmung zu erbringen und „endlich und objektiv Psi nachzuweisen". Sonst hätten die Autoren die von ihnen selbst referierte Vorgeschichte und insbesondere Inhalt und Implikationen der von ihnen zitierten Meta-Analysen nicht verstanden. Vielmehr erkenne ich in dem Projekt den respektablen Ansatz, einen experimentellen Klassiker neu mit Leben zu füllen und ihn zur Grundlage für ein sauber durchgeführtes Lehrexperiment zu machen. Insofern könnte man die Darstellung der Studie prinzipiell auch als Lehrbeispiel für die allgemeine experimentelle Psychologie gelten lassen. Darüber hinaus sind es aus meiner Sicht – neben der exemplarischen Versuchsdarstellung – gerade die im Text diskutierten methodologischen Probleme, die die Studie auch für die Grenzgebietsforschung zu einem Lehrbeispiel und damit für ein einschlägiges Journal publikationswürdig machen.

Es ist ebenfalls weder Nachteil noch Zufall zu nennen, dass dieses Projekt (mutmaßlich) neben dem Erstautor von „Undergraduates" durchgeführt wurde, die das Kartenlegen nach J.B. Rhine ohne Ermüdung an den Fronten einer jahrzehntelangen Debatte wieder einmal angepackt haben. Die Autoren haben historisches Material der experimentellen Parapsychologie neu belebt, ja zum Teil sogar selbst hergestellt. Wie ausgeführt, wäre ohne die notwendige

Naivität, bitte immer als Ursprünglichkeit bzw. Unvoreingenommenheit verstanden, und auch ohne ein wenig Nostalgie, das Experiment vermutlich nicht zustandegekommen. Mit Blick auf die universitäre Einbindung des Projekts wäre es aufschlussreich gewesen und hätte auch der Wissenschaftlichkeit keinen Abbruch getan, wenn man in einem zusätzlichen Abschnitt des Artikels näheres über das Zustandekommen und die Rahmenbedingungen des Experiments (Experimentalpraktikum? Diplomarbeit? Praktikum? Auswahl des Themas? Kommentare der Betreuer? Bewertung?) erfahren hätte.

### *Detailanmerkungen*

Wie schon angeklungen ist, halte ich den Artikel für insgesamt gut gelungen und sehr verständlich zu lesen. Angenehm fällt auf, dass die Autoren durch den gesamten Artikel hindurch eine neutrale, wissenschaftliche Position beibehalten. Divergente Positionen werden zitiert und bleiben für den Leser als ungewertete Zitate erkennbar. Die verwendeten Methoden sind sehr ausführlich und in Anlehnung an Lehrbuchstandards dargestellt, was dem Ausbildungsrahmen entspricht, in dem die Studie durchgeführt wurde. Der Diskussionsteil ist freilich auffallend kurz, was vermutlich sowohl darauf zurückgeht, dass keine signifikanten Effekte oder Zusammenhänge gefunden wurden, als auch darauf, dass die Autoren auf vage Spekulationen verzichtet haben und sich an das Gebot des einfachsten möglichen Modellrahmens gehalten haben.

Im folgenden möchte ich aber auch einige kritische Anmerkungen zusammenfassen.

- Literaturreferenzen zu Originalia von J.B. Rhine hätten sich gut gemacht.
- Von den zitierten Meta-Analysen sind die eindrucksvollen *p*-Werte angegeben, leider aber nicht die Effektstärken, denen hier ebenfalls besondere Bedeutung zukommt.
- Man hätte ergänzend zur Vorgeschichte die von Studie zu Studie fluktuierende Natur der immer wieder gefundenen Signifikanzen erwähnen können, ebenso die Feststellung, dass (auch bei der beschriebenen uneinheitlichen Befundlage) bislang kein stabil reproduzierbares ASW-Phänomen gefunden wurde.
- Bedauerlich ist die Verwendung eines eigens erstellten Fragebogens zu paranormalen Einstellungen. Der Hinweis, dass der Einsatz selbstkonstruierter Tests üblich sei, hilft dem nicht ab, und bedauerlich ist nicht nur der unnötig betriebene Aufwand. Der Bezug zu den anderen erhobenen Persönlichkeitsdimensionen, und insbesondere die Vergleichbarkeit der Skala über diese Studie hinaus sind dadurch nicht einschätzbar. Neben den zitierten Skalen von Tobacyk & Milford (1983) und Thalbourne & Haraldsson (1980) wird an dieser Stelle ergänzend auf die deutschsprachige Skala von Schriever (1998) hingewiesen; jede dieser Skalen hätte Parallelen zu anderen Studien ziehen lassen.

- Die Erhebung der Skalen erfolgte teilweise nicht auf Papier, sondern am PC. Für die Vergleichbarkeit beider Varianten wird kein Beleg angeführt.
- Streng zu unterscheiden ist (soweit möglich) zwischen dem „Regressions-Effekt" (siehe Galton, 1886; Stigler, 1986) und dem vielfach postulierten „Decline-Effekt". Ersteres bezeichnet die Tendenz zur Effektverkleinerung bei Nachmessung eines Extremums, zweiteres bezeichnet (man korrigiere mich; ich verwende die Definition der Autoren) dass Trefferleistungen *im Verlauf einer Untersuchung* absinken. Die Autoren liefern mit Abb. 2 eine Darstellung des zeitlichen Verlaufs der Trefferhäufigkeiten, die keinen Rückgang innerhalb des Experiments erkennen lässt. Der beobachtete Trefferrückgang bei wiederholter Messung der einen, extremen Versuchsperson dagegen ließe sich (unter Minimierung des Modells) zunächst einmal als Regressions-Effekt deuten.

### *Effektstärken und Hypothesen*

Spezielle Beachtung findet der Umgang mit den Themen Effektstärke, Testpower und Hypothesenprüfung.

Den Zusammenhang zwischen Effektstärke, Stichprobenumfang und Testpower (womöglich vor der Durchführung eines Experiments) zu betrachten ist ebenso lehrbuchgemäß wie sinnvoll. Dem haben die Autoren auch entsprochen, indem sie aus der Erwartung eines „kleinen" oder „sehr kleinen" Effekts und ihrer Versuchspersonenzahl auf die Testpower geschlossen haben, die sie dann als ausreichend betrachten konnten. Konsequenterweise heißt es in der Diskussion dann, dass aufgrund der ausreichenden Testpower die Nullhypothese *angenommen* werden kann.

Die Autoren belegen, dass bei einer Effektstärke von Cohens $d$=0.2 eine große Chance bestanden hätte, den entsprechenden Effekt in dieser Studie zu bestätigen.

Hierzu ist zunächst anzumerken, dass aus den Literaturreferenzen, speziell den Meta-Analysen, im Text keine Effektstärken berichtet werden. Den Autoren sei zugestanden, dass sie sich eventuell nur für einen Effekt interessiert hätten, falls dieser größer als $d$=0.2 gewesen wäre; dies ist in der experimentellen Forschung zu anderen Fragestellungen durchaus üblich. Die einschlägige Literatur zu außersinnlicher Wahrnehmung, auch bei besonderer Beachtung der Meta-Analysen, liefert sehr uneinheitliche Ergebnisse bezüglich der gefundenen Effektstärken (siehe etwa Utts, 1991). Extrem hohe Signifikanzen sind selbstverständlich auch bei minimalen Effektstärken möglich, wenn eine sehr große Gesamtstichprobe in einer Meta-Analyse untersucht wird. Aus der grenzwissenschaftlichen Fragestellung ergibt sich, dass man sich hier durchaus auch für extrem kleine Effekte interessieren sollte, da auch diese von fundamentaler Bedeutung für unser Weltverständnis wären. Unterstellen wir beispielsweise eine Effektstärke

von $d$=0.05, so würde dieser Effekt bei der Stichprobengröße von N = 96 und einem Alpha-Niveau von 0.05 mit einer Wahrscheinlichkeit von nur etwa 10% als existent angenommen (Auch bei Abwesenheit des Effekts würde ein solcher Effekt gemäß Definition des Alpha-Niveaus mit einer Wahrscheinlichkeit von 5,0% angenommen). Die Autoren vertreten in der Diskussion ihre positive Annahme der Nullhypothese explizit nur für die berichteten, postulierten Effektstärken von $d$=0.2; allerdings wird diese Aussage bedeutungslos, wenn auch deutlich kleinere Effektstärken interessieren. Dann wäre auch die Ausführung der Autoren, dass auch eine zweiseitige Testung „mit akzeptabler Teststärke" möglich gewesen wäre, hinfällig. Vor dem Hintergrund unbekannter, womöglich minimaler Effektstärken wäre diese Studie noch deutlicher als ein kleines Puzzleteilchen im randlosen Puzzle der Wissenschaften erkennbar.

Ganz allgemein sei noch hinzugefügt, dass die (lehrbuchgemäße) Annahme der Nullhypothese, die sich aus dem Verwerfen der Alternativhypothese ergibt, nicht inhaltlich mit einer Aussage über die Nicht-Existenz eines Effekts gefüllt werden darf. Eine Feststellung von der Art, dass die Alternativhypothese verworfen wird, ist bei entsprechender Ergebnislage angemessen, nicht aber die Schlussfolgerung, dass das untersuchte Phänomen nicht existiert. Die letztere inhaltliche Schlussfolgerung läßt sich anhand einer Einzelstudie sicher niemals belegen, streng genommen auch nicht durch eine Vielzahl von Negativergebnissen; vielmehr wird hier die Frage nach der generellen Falsifizierbarkeit eines postulierten Phänomens aufgeworfen, die gar nicht Gegenstand der Studie von Rey *et al.* sein konnte.

## *Nabelschau nur wenn notwendig?*

Die Autoren führen einige wesentliche methodenkritische Aspekte im Diskussionsabschnitt auf, was den Extrapolationsbereich der Studie in angemessener Weise relativiert. Nicht vorbehaltlos zustimmen möchte ich der Ansicht, dass das vorliegende Experiment lediglich „im Falle positiver Befunde in diverser Hinsicht hätte kritisiert und in Frage gestellt werden müssen". Die Autoren schreiben, dass signifikante Ergebnisse dann nicht mehr eindeutig etwaigen Manipulationsversuchen oder aber außersinnlichen Wahrnehmungen zuzuordnen gewesen wären.

In Abweichung hiervon meine ich, dass auch eine „selektive Nabelschau", die nur bei positiven Ergebnissen durchgeführt wird, einen Bias erzeugt, nämlich in negativer (konservativerer) Richtung: Man stelle sich vor, die Ergebnisse (Effektstärken) von 1000 Experimenten seien normalverteilt um Null. Nun wird allen Studien mit negativen Effekten „geglaubt", bei den positiven aber herumgerechnet, herumgedeutet, korrigiert und relativiert. Das Ergebnis einer Meta-Analyse wäre folglich ein (womöglich signifikanter) Effekt mit negativem Vorzeichen und würde womöglich als „psi-missing" gedeutet. Wäre tatsächlich ein positiver Effekt wirksam, würde dieser durch das genannte Vorgehen unterschätzt bzw. schwerer gefunden. Die

Schlussfolgerung kann daher nur lauten, dass es einfach nicht erlaubt ist, die Wertigkeit und die Einschränkungen einer Studie (und damit die Tragweite möglicher Ergebnisse) zeitlich nach Kenntnisnahme der Ergebnisse festzulegen.

## *Fortsetzung folgt?*

Im Nachklang stellt sich zum einen die Frage, ob und wie die eine Versuchsperson, die die extremen Ergebnisse hatte (berichtete Signifikanz bei Erstmessung: $p$ = 2/100000), seit 2007 weiter untersucht wurde. Sofern dies nicht geschehen ist, wäre es aufschlussreich, im konkreten Fall zu erfahren, *warum* die weitere Untersuchung eingestellt wurde. Zum einen ist die Entwicklung beobachteter hochsignifikanter Effekte bei mehrfacher Versuchswiederholung Gegenstand diverser Debatten, zum anderen könnten sich Hinweise zum Umgang mit dem File-Drawer-Problem ergeben, und schließlich wäre der weitere Umgang mit dem beobachteten Extremum aus einer experimentellen Meta-Perspektive heraus interessant.

Dass grenzwissenschaftliche Projekte in der universitären Ausbildung relativ dünn gesät sind, ist zweifellos bedauerlich, gerade auch deshalb, weil kaum ein anderes Feld gleich gut geeignet ist, experimentelle Methodik zu lernen und zu lehren. Der Gewinn für die Ausbildung ist ebenso wenig zu unterschätzen wie das methodisch befruchtende Abfärben auf die experimentelle Forschung in anderen Gebieten.

Nachdem die Autoren mehrere Anregungen für Folgestudien aufzeigen und die vorliegende Studie im Jahr 2007 stattgefunden hat, stellt sich dem Leser weiterhin die Frage, ob inzwischen eine solche Folgestudie stattgefunden hat, ggf. mit welchen Modifikationen und welchen Ergebnissen. Für den Fall, dass eine solche Folgestudie bislang nicht stattgefunden hat, wage ich die These, dass die Autoren, sofern weiter forschungstätig und gesund, sich inzwischen Gebieten zugewendet haben, in denen die meisten Effektstärken über $d$=0.2 liegen.

## Literatur

Galton, F. (1886). Regression towards mediocrity in hereditary stature. *Journal of the Anthropological Institute*, 15, 246-263.

Schriever, F. (1998). Die Skala zur Erfassung paranormaler Überzeugungen (SEPÜ). *Zeitschrift für Parapsychologie und Grenzgebiete der Psychologie*, 40-41, 95-133.

Stigler, S.M. (1986). *The History of Statistics: The Measurement of Uncertainty before 1900*. Cambridge, MA: The Belknap Press of Harvard University Press.

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

## Roger Nelson[6]

### Hochwissenschaftliches Rauschen

Es ist wirklich wichtig, Studenten und junge Wissenschaftler dazu zu ermutigen, eigene Erfahrungen mit der Durchführung parapsychologischer Experimente zu sammeln. Unter dieser Prämisse wäre ich im Prinzip froh, einen Aufsatz wie den vorliegenden in einer guten wissenschaftlichen Zeitschrift publiziert zu sehen. Die Lektüre dieses Beitrags fördert tatsächlich jedoch zahlreiche fragwürdige Aspekte zutage, die mich letztendlich zu der Auffassung veranlassen, dass das Experiment nicht hinreichend gut durchgeführt und begründet ist, als dass seine Darstellung Teil derjenigen Literatur werden sollte, anhand deren Wissenschaftler den ASW-Forschungsstand beurteilen. Meine Bedenken sind dabei völlig unabhängig von der Frage, ob die zu überprüfende Hypothese selbst bestätigt oder verworfen werden muss.

Die Parapsychologie hat es verdient, dem gebietsfremden Leser anhand derjenigen hohen qualitativen Standards vorgestellt zu werden, die diese Forschung professionell auszeichnen. Der Beitrag von Rey und seinen Mitarbeiterinnen weist hingegen eine Fülle von Mängeln auf. Meine nachfolgenden Bemerkungen werden sich auf diese Unzulänglichkeiten konzentrieren. Ich bedauere, dass ich sehr viel weniger Positives zu sagen habe als ich gerne gesagt hätte, und ich hoffe, der Verfasser und seine Ko-Autorinnen werden mir einerseits meine Schonungslosigkeit und andererseits (und insbesondere) etwaige Missverständnisse nachsehen, die dem Umstand geschuldet sein mögen, dass Deutsch nicht meine Muttersprache ist.

#### *Forschungs-Design*

Meine Bedenken fallen in drei verschiedene Kategorien, die sich aber wechselseitig überschneiden. Erstens ist die methodische Anlage eines Forschungsvorhabens eine Kunst, die damit beginnt, eine angemessene, nutzbringende und klar formulierte Frage zu stellen. Eine eindeutige Fragestellung macht deutlich, um was es im Experiment geht, und sie erlaubt letztlich eine verständige Interpretation der Resultate sowie, so ist zu hoffen, neue Einsichten über den Gegenstand, von dem die Fragestellung handelt. Möglicherweise ist mir ja die eine oder andere Nuance entgangen, aber es scheint nicht klar, welche Motivation dieser Studie überhaupt zugrunde liegt.

Was genau wollen Rey *et al.* eigentlich zeigen? Was ist ihre grundlegende Frage? Falls sie die Arbeiten von J.B. Rhine mit Zener-Karten zur Telepathie, zum Hellsehen und zur Prä-

6 Dr. Roger Nelson war als Psychologe viele Jahre lang Mitarbeiter am Princeton Engineering Anomalies Research (PEAR) Laboratory der Princeton University und ist der Direktor und Koordinator des Global Consciousness Project.

kognition replizieren wollten, dann hätten sie eben das tun sollen, was Rhine getan hat – ein einfaches, direktes Experiment konzipieren und darauf gefasst sein, dieses Protokoll über einen längeren Zeitraum durchzuziehen.[7] Dies ist es jedoch nicht, was sie in ihrem Beitrag beschreiben. Stattdessen haben sie ein dreifaktorielles Design gewählt, dass augenscheinlich darauf bedacht ist, nicht nur Telepathie oder Hellsehen oder Präkognition, sondern alle drei ASW-Modalitäten zugleich zu untersuchen und sich gleichzeitig auch noch verschiedenartige mutmaßliche Moderatoren einschließlich Persönlichkeits- und Einstellungsvariablen sowie eines induzierten „Motivations"-Faktors vorzunehmen. Darüber hinaus wird das Experiment (seltsamer Weise aber nicht die statistischen Auswertungsverfahren) dadurch noch zusätzlich verkompliziert, dass verschiedene Experimentatoren, und zwar nicht weniger als fünf, mit unterschiedlichen selbstzugeschriebenen Grundeinstellungen zu den zu untersuchenden Phänomenen – von neutral bis skeptisch – als Versuchsleiter eingesetzt werden. Das ist allerhand.

Das daraus resultierende Projekt macht den Eindruck eines Lehrinstruments für einen Einführungskurs in die experimentelle Psychologie. Ich denke, dass das faktorielle Design und die statistische Auswertung im wesentlichen korrekt durchgeführt sind, sofern man denn die Grundvorstellungen der Autoren akzeptiert; allerdings stellt sich meines Erachtens noch eine ernstliche Frage hinsichtlich der Poweranalyse (dazu weiter unten).

Ich bin ganz entschieden der Auffassung, dass parapsychologische Forschungsfragen, so interessant sie auch sein mögen, nicht leicht (oder gar „nebenbei") zu behandeln sind. Insbesondere können Vorannahmen, wie sie für rein psychologische Fragestellungen etwa aus der Wahrnehmungs- oder Gedächtnispsychologie oder hinsichtlich klinischer Kategorisierungen angemessen sein mögen, nicht einfach auf ein Forschungsgebiet übertragen werden, in dem Effekte, sofern sich solche denn einstellen, im allgemeinen gering (ja überaus geringfügig) zu sein pflegen. Dies ist nicht der Ort, um ausführlicher darauf einzugehen, aber unzutreffende Vorannahmen dieser Art liegen vielen der strittigen Auseinandersetzung in der Parapsychologie – etwa hinsichtlich Metaanalysen zu Ganzfeld-, Forced-Choice- oder Mikro-PK-Experimenten – zugrunde. Einige dieser akademischen Streitfälle werden von Rey *et al.* zwar angesprochen, aber sie unternehmen keinen Versuch herauszufinden, weshalb die Metaanalysten miteinander im Streit liegen. Es scheint, als zitierten sie die betreffenden Studien lediglich zur Ausgestaltung des Hintergrunds und als täten sie dann so, als ob ihre eigene Untersuchung irgendetwas dazu beitrüge, diese (impliziten, aber gar nicht selbst gestellten) Fragen zu beantworten und den Problemen auf den Grund zu gehen.

7 Leser, die sich über die Rhineschen Forschungsstandards orientieren möchten, sind gut beraten, den Beitrag „A review of the Pearce-Pratt distance series of ESP tests" zu lesen, der bequem unter http://psychicinvestigator.com/demo/ESPdoc.htm zugreifbar ist. Gedruckte Fassung: Rhine & Pratt (1954).

### *Subtile Methoden*

Zweitens kommt der sauberen und sorgfältigen Umsetzung der experimentellen Verfahren besondere Bedeutung zu. Das gilt zwar für jede experimentelle Forschung, ist jedoch für die Psi-Forschung oder für jeden anderen Phänomenbereich, der es mit flüchtigen oder subtilen oder wahrlich kleinen Effekten zu tun hat, absolut ausschlaggebend. In solchen Forschungsbereichen können bereits geringfügigste Fehler irreführende Resultate zur Folge haben. Sie können einerseits faktisch nicht existente anomale Effekte vorspiegeln oder andererseits tatsächliche Abweichungen verbergen, aus denen wir etwas hätten lernen können. Skeptische Beobachter bestehen durchaus zu recht darauf, dass außergewöhnliche Behauptungen nach außergewöhnlichen Beweisen verlangen, nach Beweisen nämlich, die zuverlässig und valide und mithin glaubhaft sind. Die parapsychologische Forschung ist für den Zaghaften ebenso ungeeignet wie für denjenigen, der nicht willens oder in der Lage ist, uneingeschränkt und ehrlich sein Bestes zu geben. Nun präsentiert der Diskussionsteil des vorliegenden Aufsatzes uns eine bemerkenswerte Litanei von Problemen. Einige Leser mögen der Auffassung sein, die Autoren verdienten besondere Anerkennung wegen ihrer ehrlichen und ungeschminkten Zurschaustellung aller der Fehler, die aus ihrer Arbeit zu ersehen sind. Aber sollten wir uns nicht eher die Frage stellen, weshalb sie diese Fehler und Unzulänglichkeiten überhaupt zugelassen haben? Und als eingeladener Kommentator ist mir auch die Frage wichtig, weshalb ein Aufsatz mit so vielen Fehlern zur Publikation gelangt – mit Fehlern, die jede denkbare Schlussfolgerung dubios erscheinen lassen und Interpretationen letztlich unmöglich machen. Hier stellen sich offenkundig Probleme, die die Wissenschaftlichkeit der Literatur und die Einhaltung professioneller Standards grundsätzlich betreffen.

### *Poweranalyse*

Wir haben es mit einem Gebiet zu tun, in dem es auf Vorannahmen besonders ankommt. Diese sind für die Poweranalyse, die darüber Aufschluss gibt, mit welcher Chance bei einer Studie ein Effekt überhaupt nachgewiesen werden kann, von mehr als nur trivialer Bedeutung. Das ist hier von besonderer Wichtigkeit angesichts des Umstands, dass die Autoren nachdrücklich betonen, ihr Experiment besitze „ausreichende Power", um Schlussfolgerungen zu rechtfertigen – trotz einer offenkundig schmalen Datenbasis: nur einem Run von jedem der 96 Teilnehmer (die sich zudem über drei verschiedene experimentelle Fragen verteilen – ein weiterer Umstand, den wir hier der Einfachheit halber außer Betracht lassen). Unter der Nullhypothese, dass es Psi nicht gibt, sollte jeder der 96 x 25 = 2400 Rateversuche unabhängig sein (vorausgesetzt, dass die richtigen Antworten zufällig und voneinander unabhängig gewählt werden), so dass es den Anschein hat, als könne man von einer statistischen Binominalverteilung von n = 2400 ausgehen. Unter der Alternativ-Vermutung können die Antworten jedoch für jede Versuchsperson

unterschiedliche Trefferraten aufweisen (was ja nachweislich auch der Fall ist).

Folglich empfiehlt es sich, das Experiment als eine Folge von 96 Beobachtungen aufzufassen, jede binominal mit unterschiedlichem *p*. In diesem Fall wäre es angemessen, einen One-Sample t-test mit n = 96 durchzuführen und diese als 96 Beobachtungen aus einer Normalverteilung mit einer wahren Effektstärke von 0.2 zu behandeln. Es kommt entscheidend darauf an, welches Modell man wählt. Für n = 2400 ergibt eine binominale Poweranalyse, wie die Autoren feststellen, eine Zufallserwartung von 0.20, und eine postulierte Effektstärke von 0.2 würde eine Trefferrate von ca. 0.345 (mit Cohen's *h*) und statistische Power nahe 1 erreichen. Wenn wir andererseits jedoch das t-test-Modell mit n = 96 wählen, was ich angesichts der zu testenden Hypothese für den angemessenen Weg halte, dann ergibt sich als berechnete statistische Power für dieselbe Effektstärke ein weitaus weniger eindrucksvoller Wert von 62.4%.

### *Sinnvolle Interpretation*

Schließlich kommen wir zur Interpretation der Ergebnisse. Wären die ursprünglichen Forschungsfragen angemessen formuliert und das experimentelle Verfahren entsprechend ausgelegt und sorgfältig umgesetzt, dann sollte das Experiment Daten mit einem gewissen Informationswert zur Verfügung stellen. Die Resultate könnten beispielsweise die Alternativhypothese stützen oder aber vom erwarteten Zufallswert nicht zu unterscheiden sein. In jedem Fall erhoffen wir uns von einem Experiment, dass es uns – ganz gleich, wie es ausgeht – etwas Verwertbares über die betreffenden Phänomene oder über die jeweilige Forschungsfrage lehrt, die wir zur beantworten versucht haben. Erbringt das Experiment „positive Evidenz", dann ist die Aufgabe in gewissem Sinne einfacher, besonders bei einem Projekt wie dem vorliegenden, das eine Art Replikation sein will. In dem Maße, in dem es sich um eine valide Replikation handelt, ist die Deutungsarbeit dann praktisch schon getan, und wir könnten uns unmittelbar mit der Frage befassen, welches Licht die neuen Daten auf die Befunde vorhergehender Experimente werfen. Entsprechen die Ergebnisse hingegen der Zufallserwartung, dann gibt es eine Reihe möglicher Richtungen, die ihre Interpretation einschlagen könnte. Ein solches Resultat könnte beispielsweise zu der (hier von den Autoren vertretenen) Behauptung verleiten, dass die neuen Resultate in gewisser Weise die früheren bestreiten und so zur Heterogenität des historischen Forschungsstands beitragen.

Angesichts des Berichts der Autoren denke ich jedoch nicht, dass wir diese – oder überhaupt irgendeine sonstige nützliche – Schlussfolgerung ziehen können. Dafür gibt es mindestens zwei Gründe: Zum einen hat das Experiment keinerlei Resultate erbracht, die von einem statistischen Rauschen unterschieden werden könnten. Es gibt schlechterdings nichts zu interpretieren – wenigstens dann nicht, wenn man die meines Erachtens fadenscheinigen Argumen-

te der Autoren hinsichtlich (a) der Powerberechnung und (b) der Bedeutung von Zufallsbefunden nicht akzeptiert. Zum zweiten sind die Anlage und die Durchführung des Experiments so hochproblematisch, dass ich es aus den schon früher genannten Gründen für unmöglich halte, ihm irgendetwas Repräsentatives oder überhaupt Interpretierbares abzugewinnen.

## *Zahlreiche Fragen*

Bei der Vorbereitung dieser Stellungnahmehatte ich mich auf die Kommentierung eines interessanten neuen Beitrags eingerichtet, in dem Rey *et al.*, so erwartete ich, einen modernen Blick auf eine klassische Forschungsperspektive werfen. Bei der Lektüre stellte sich dann rasch Enttäuschung ein, da zahllose klärungsbedürftige Fragen auftauchten und der Eindruck immer unabweisbarer wurde, hier gehe es nicht so sehr um den Versuch, unsere Kenntnisse zu erweitern und zu vertiefen, als vielmehr um die Einübung einer willkürlich gewählten Allerweltsmethode, die für bestimmte Segmente der experimentellen Psychologie ihre Berechtigung haben mag, im Bereich der Erforschung eines möglicherweise erweiterten menschlichen Bewusstseins aber zwangsläufig versagen muss. Um verständlicher zu machen, weshalb dieser Kommentar, wie ich fürchte, zu einer ungnädig kritischen Stellungnahme geraten ist, füge ich hier in einem Anhang eine Übersicht über die Fragen, Probleme und Einwände an, die die Lektüre des Beitrags von Rey *et al.* aufgeworfen hat. Obwohl (oder weil) mein Kommentar durchweg kritisch gestimmt ist, hoffe ich, dass er zumindest in einem weiteren Zusammenhang insofern einen gewissen Nutzen haben wird, als er aus der Perspektive professioneller Parapsychologie ein Licht auf die außerordentliche Sorgfalt wirft, die die parapsychologische Forschung erfordert.

## *Anhang*

1. Weshalb Zener-Karten? Oder entscheidender: weshalb selbstgebastelte Zener-Karten? Die einschlägige Literatur lässt keinen Zweifel an den Schwierigkeiten, einschließlich der Gefahr möglicher irreführender oder verfehlter Schlussfolgerungen, die solches Testmaterial mit sich bringen kann. Es gibt vorzügliche moderne Alternativen, die problemlos zur Verfügung stehen, beispielweise Online-Eperimente, die das Boundary Institute (Richard Shoup) und das Institute of Noetic Sciences (Dean Radin) anbieten.

2. Weshalb wird ein bekanntermaßen kleiner Effekt dadurch weiter verwässert, dass man das Design eines Experiments, das vorgeblich eine Replikation darstellen soll, in multiple Forschungsfragen aufspaltet? Hierin kann ich nur die unbedachte Übertragung experimenteller Taktiken aus anderen Gebieten wie der Psychologie, die über vergleichsweise große und stabile Effekte verfügen, auf ein Gebiet mit notorisch kleinen,

subtilen Effekten sehen. Schon das Abstract des Beitrags liest sich eher wie ein Lehrkurs für ANOVA oder faktorenanalytische Verfahren. Dieses Experiment ist keine Replikation in irgendeinem nachvollziehbaren Sinn. Das Experiment tut weder das, was seine Vorbilder taten, noch das, was man aufgrund der umfangreichen Literatur ansonsten für sinnvoll hätte halten können.

3. Wenn ich das recht verstehe, gibt diese Studie vor, durch „inkonsistente Befunde der Ganzfeld-Metaanalysen" angeregt worden zu sein. Dort geht es jedoch um ein vollkommen anderes experimentelles Design. Zudem sind „inkonsistente Befunde" von Metaanalysen in der Wissenschaft an der Tagesordnung, zumal dann, wenn solche Analysen miteinander schwer verträglichen Absichten entspringen. Die von Rey *et al.* zur Rechtfertigung herangezogenen Arbeiten unterscheiden sich weniger phänomenologisch als hinsichtlich der von ihren Autoren jeweils in Ansatz gebrachten Selektionskriterien. Eine angemessene Aufarbeitung „inkonsistenter Befunde" in Metaanalysen würde mithin eher eine soziologische Einschätzung erfordern als die Replikation eines mehr als 50 Jahre alten Paradigmas (das selbst „inkonsistente Befunde" gezeitigt hat, die wiederum nicht in erster Linie phänomenologisch bedingt sind).

4. Meiner Auffasssung nach sollte ein Wissenschaftler, wenn er denn schon Moderatorvariablen untersuchen möchte, sein experimentelles Design so anlegen, dass diese Variablen in Abhängigkeit von anderen manipulierbaren Größen (und nicht mittels einer gekünstelten Klassifikation im Rahmen eines faktoriellen Designs) miteinander verglichen werden können. Ich gebe zu, dass dies ein Vorurteil meinerseits sein mag, aber die Evidenz für die Wirksamkeit solcher Modifikatoren scheint mir bestenfalls dürftig zu sein. Stellt man zudem die generelle Schwäche des mutmaßlichen anomalen Effekts in Rechnung, dann ist praktisch sichergestellt, dass sich das, was die ursprüngliche Forschungsfrage noch hätte zutage bringen können, bei einer Verzettelung über schlecht gerechtfertigte sekundäre Fragestellungen unwiderbringlich verflüchtigt. Und wenn ferner, wie im vorliegenden Fall, der zu betrachtende Faktor überhaupt nur dadurch gerechtfertigt ist, dass er in einem ganz anderen experimentellen (free-response) Forschungs-Design (Ganzfeld) eine gewisse Rolle spielt, dann ist seine Übertragung auf ein Forced-Choice-Design mindestens fragwürdig und mit hoher Wahrscheinlichkeit irreführend. Man vergleiche dazu die einschlägige Literatur über Free-Response- vs. Forced-Choice-Verfahren (z.B. Palmer, 1978).

5. Die Beschreibung der psychologischen Variablen, die Rey *et al.* zur Verfügung stellen, macht ohnehin den Eindruck, als seien die Autoren eigentlich an diesen Variablen *per se* interessiert, und nicht daran, welche Rolle sie ggf. als sinnvoll gerechtfertigte oder wenigstens als wahrscheinliche Moderatoren spielen könnten. Dies zeigt erneut, wie

problematisch ihre Entscheidung ist, die Wahrscheinlichkeit des Auftretens irgendeines interpretierbaren Effekts dadurch zu vermindern, dass sie eine mutmaßliche ASW-„Energie", die sich im Experiment hätte zeigen können, über mancherlei Zusatzbedingungen all zu dünn ausgewalzt haben.

6. Fiktives Feedback!? Dies ist eine weitere methodologische „Sünde" in einem Psi-Experiment wie diesem. Außer in ganz seltenen, wohlbegründeten Ausnahmefällen verwenden professionelle Parapsychologen niemals fiktive Feedbackbedingungen. Feedback ist in der parapsychologischen Forschung, auch aus theoretischen Gründen, ein äußerst sensibles Thema, was um so mehr für fiktives, falsches Feedback gilt.

7. Extraversion, Neurotizismus, Selbstwirksamkeit, Einstellung, falsches Feedback (positiv / negativ / control) – eine sehr beachtliche Liste von Dingen, die aber mit einer vorgeblichen Replikation allesamt nicht das Geringste zu tun haben. Muss nicht selbst eine „konzeptuelle Replikation" wenigstens im Kern mit dem konsistent sein, was sie replizieren soll?

8. Fünf verschiedene Versuchsleiter mit Einstellungen zwischen „neutral" und „skeptisch" mit der Durchführung des Experiments zu betrauen, hätte als überprüfbare unabhängige Variable eine weit höhere Plausibilität gehabt als all jene zweifelhaften psychologischen Trait-Variablen. Im Versuchs-Design ist gerade dieser „Faktor" der Experimentatoren-Einstellung jedoch nicht systematisch berücksichtigt.

9. „Überzufällige Leistung" vs. „Gar nichts richtig": Das erteilte Feedback ist nicht nur falsch/fiktiv, vielmehr ist es auch nicht symmetrisch, da sich, so die Autoren, „auch die Trefferwahrscheinlichkeiten asymmetrisch verteilen". Dies ist eine ziemlich verschrobene Rechtfertigung. Weshalb haben sie nicht einfach „bessere" vs. „schlechtere" Leistungen zurückgemeldet? Was sie stattdessen tun, klingt ein wenig so, als hätten die Autoren zwar die richtigen Wörter verwendet, aber ihre Bedeutung nicht verstanden. Man könnte sogar argumentieren, dass ein normaler Proband die Quantifizierung eines solchen Feedbacks mutmaßlich gar nicht kennt oder versteht. Kümmert es ihn, ob er „falsch geraten" oder „signfikant falsch geraten" hat? Anzunehmen, in einem der beiden Fälle sei die demotivierende Wirkung stärker als im anderen, ist naiv. Falls aber ein Proband versuchte, diese Rückmeldungen kompetent zu deuten, müsste ihm das Feedback „Gar nichts richtig" als so unwahrscheinlich vorkommen, dass er es gerade nicht als Demotivation, sondern vielmehr als einen Ausweis ganz besonders ausgeprägter eigener Fähigkeiten interpretieren würde. Es ist befremdlich zu sehen, wieviel fehlgeleitete Gedankenarbeit in diese Spekulationen über falsches Feedback investiert worden ist. Ein Lächeln oder ein finsterer Blick, ein grünes vs. eines roten Lichts wäre als (De-) Motivator völlig ausreichend und allemal ergiebiger gewesen. All dies erweckt den Ein-

druck, als sei bei der Versuchsplanung blinder Aktionismus mit sinnvoller Aktivität verwechselt worden oder an deren Stelle getreten – etwa so, wie es Erving Goffman (1959) in *Presentation of Self in Everyday Life* vorführt.

10. 25 Trials per Deck gelten den Autoren als sichere Größe. Haben wir es hier vielleicht mit einer wiederholten Stichprobennahme zu tun? Sind es wirklich 5x5? Oder sind es, da auf Telepathie, Hellsehen und Präkognition getestet wurde, eher 3x25 oder 3x5x5? Es scheint, als seien die drei Modalitäten in Reihe getestet worden. Ist das ein solides experimentelles Design?

11. Selbstgefertigte Karten: Hmmm. Spricht das etwa für besondere Sorgfalt? Selbst entworfene Fragebogen: Weshalb wurden keine standardisierten und validierten Fragebogen verwendet, die reichlich zur Verfügung stehen, auch in deutscher Sprache (Schriever, 1998)? Wie wurden die Karten gemischt? Selbstgefertigte Karten lassen eine gründliche Durchmischung vermutlich kaum zu. Wie wurden die Ergebnisse protokolliert? Per Hand? Hat irgendjemand die korrekte Protokollierung überprüft? Sicherung gegen Betrug wird zwar behauptet, entsprechende Sicherheitsmaßnahmen waren jedoch augenscheinlich nicht definiert.

12. Wir haben fünf Versuchsleiter, teils offen, teils neutral, teils skeptisch gestimmt, die aber als Versuchsvariablen keine Rolle spielen (siehe Punkt [9]). Und die Versuchsleiter waren hinsichtlich der Resultate nicht blind?! Und ihnen waren die positiven vs. negativen Feedback-Gaben bekannt!

13. Ich kann nicht erkennen, dass irgendwo Maßnahmen zur Entdeckung und ggf. zur Korrektur von Protokollierungsfehlern beschrieben wären, wie sie etwa bei der Übertragung der Calls vom Protokollbogen (der wiederum nicht beschrieben wird) in Computerdateien oder Excel-Tabellen leicht vorkommen können. Ironischer Weise sind also gerade die entscheidenden Daten, die zur Beantwortung der grundlegenden Forschungsfrage dienen sollen, gegen fehlerhafte Aufzeichnung und Übertragung nicht gesichert, während andererseits – wiederum ironischer Weise – die sekundären psychologischen Messgrößen unmittelbar in Computerdateien eingegeben wurden.

14. Das grundlegende Problem dieser vorgeblichen Replikation: Die Ausgangsdaten sind „aufgeschlüsselt für die drei abhängigen Variablen, sowie zahlreiche weitere deskriptiv- und inferenzstatistische Berechnungen." So wenige Daten, so viel Analyse. Ohnehin geringe statistische Power, und dann noch verweht im Wind aufgeblasenen Designs und der Überanalyse.

15. Ich bezweifle, das die Powerberechnung stichhaltig ist (siehe auch den Abschnitt Poweranalyse im obigen Text). Nimmt man 25 Rateversuche mal 96 Probanden, erhält

man ein großes N, aber ein N von was genau? Binäre Wahlen dieser Art sollten nicht für Tests eines Effekts gehalten werden, der hier durch den „kleinen Effekt“ ($h$ = 0.2) nach Cohen repräsentiert wird. Diese Zahl ist eine Schätzung für das Gesamtexperiment. Betrachtet man tatsächlich die einschlägige Literatur und die von den Autoren angeführten Metaanalysen, dann erscheint die Powerabschätzung der Autoren als zweifelhaft. Der Wert einer Poweranalyse liegt in seiner prinzipiellen Logik – die hier jedoch ins Leere greift. Die Autoren haben richtig gerechnet, aber falsch geschlossen. Dies zeigt sich auch daran, dass sie zunächst die ungewöhnlichen Scores einer der Versuchspersonen (12, 9 und 10 Treffer) beschreiben, dann aber dazu übergehen, über Prozentsätze in verschiedenartigen Untergliederungen des experimentellen Designs zu reden. Das tatsächlich verwendete Maß ist also (und zwar angemesser Weise) doch der relative Anteil oder der Prozentsatz je Versuchsperson. Da dem so ist, muss die Power-Berechnung von N = 96, nicht von N = 25x96, ausgehen.

16. „Die Hypothese, dass Telepathie, Hellsehen und Präkognition bei der Auswahl von Zener-Karten zu überzufällig hohen Trefferraten führen, kann unter der Annahme eines sehr kleinen Effekts auf Basis dieser Untersuchung zurückgewiesen werden“, betonen die Autoren. Hier kommt leider eine klassische Fehlinterpretation der p-Wert-Statistik zum Ausdruck. Man kann korrekterweise behaupten, die Nullhypothese könne nicht zurückgewiesen werden oder die Alternativhypothese werde nicht gestützt – die hier zitierte Aussage aber unterstellt, wenn ich sie recht verstehe, die geprüfte Hypothese *könne* suspendiert oder zurückgewiesen werden, und das ist kein zulässiger Schluss. Die Ergebnisse dieses Experiments können nur als Rauschen beschrieben werden; es gibt keine verwertbare Evidenz für oder gegen die getestete Hypothese.

17. Da das Experiment keinen eigentlichen Effekt zeigt, ist auch der Versuch einer Einschätzung von Hypothese 2 zweifelhaft. Und dass sie “verworfen werden” müsse, ist wiederum nicht gerechtfertigt, wenn sich den Daten im Wesentlichen nicht mehr als Rauschen entnehmen lässt.

18. Kehren wir nochmals zu den auffälligen Resultaten der einen Versuchsperson zurück. Die Autoren führen augenscheinlich einige Tests *post hoc* durch, um so festzustellen, ob sich die ungewöhnliche Trefferleistung fortsetzt. In einem Run ist sie weiterhin auffällig, in einem zweiten nicht. Dann setzt die Deutungsarbeit ein. Die Daten für diese Vp ergaben gemäß Protokoll eine Wahrscheinlichkeit 2/100000, nach den *Post-hoc*-Tests aber erfahren wir, dass sie bei 5/100000 liege – worauf uns eine Interpretation im Sinne von von Lucadous „Absinkungseffekt“ angeboten wird (Lucadou, 1997). Nun mal halblang! Anhand solchen Datenmaterials kann man sich vielleicht über Biersorten verständigen, aber eine solche Diskussion hat in einem wissenschaftlichen Bericht nichts zu suchen.

Man denke: Gerade einmal 3 Mini-Tests („Runs" in Rhines Terminologie) im eigentlichen Protokoll weisen ungewöhnliche Prozentsätze aus. Zwei weitere Runs, außerhalb des Protokolls, zeigen uneinheitliche Resultate (zwei Zahlenwerte, die angesichts der hier untersuchten hypothetischen Phänomene auf einer verschwindend geringen Stichprobe beruhen). Das ist wirklich nicht der Stoff für eine akademische Diskussion. Was wir hier vor uns haben, ist günstigenfalls die Demonstration einer Experimentiertechnik. Wir haben gar keinen deutbaren Datenbestand, sondern lediglich ein paar Stichproben.

19. In der Diskussion behaupten die Autoren wiederum (gestützt auf eine angeblich „ausreichende Power", die ich für ein ernstes Missverständnis halte), dass die Nullhypothese akzeptiert werden könne. So sollte man keine Inferenzstatistik betreiben. Das Experiment zeigt nur Rauschen. Evidenz stellt es weder für die Nullhypothese noch für die Alternativhypothese zur Verfügung. Hätte das Experiment tatsächlich die Power, die die Autoren ihm zuschreiben, und bestünden seine Resultate dennoch nur in Rauschen, dann könnte ich die Behauptung akzeptieren, dass die Nullhypothese nicht zurückgewiesen werden könne. Aber es wäre auch dann unzulässig zu behaupten, es gebe kein Phänomen der Art, wie es die Alternativhypothese beschreibt. Eben das aber, so glaube ich, behaupten die Autoren.

20. Es freut mich, dass der Diskussionsteil des Beitrags von Rey *et al.* immerhin eine unverblümte Auflistung der Unzulänglichkeiten ihres Experiments enthält. Zugleich bin ich jedoch überrascht, dass das Experiment angesichts dieser Sorten von Problemen, die ja alle bestens literaturbekannt sind, dennoch in der beschriebenen Weise durchgeführt worden ist. Und ich bin ein wenig ratlos angesichts des Umstands, dass es das Referee-Verfahren überstanden hat und zur Publikation angenommen worden ist. Gemessen an den Qualitätsstandards, die die Parapsychologie über Jahrzehnte hinweg entwickelt hat, hat es in der einschlägigen Literatur eigentlich nichts zu suchen.

(Aus dem Amerikanischen von Gerd H. Hövelmann)

## Literatur

Goffman, E. (1959). *Presentation of Self in Everyday Life*. New York: Doubleday.

Lucadou, W. von (1997). *Psi-Phänomene. Neue Ergebnisse der Psychokinese-Forschung*. Frankfurt/M.: Insel.

Palmer, J. (1978). Extrasensory perception: Research findings. In Krippner, S. (Ed.), *Advances in Parapsychological Research. Volume 2: Extrasensory Perception* (pp. 59-243). New York & London: Plenum Press.

Rhine, J.B., & Pratt, J.G. (1954). A review of the Pearce-Pratt distance series of ESP tests. *Journal of Parapsychology*, 18, 165-177.

Schriever, F. (1998). Die Skala zur Erfassung paranormaler Überzeugungen (SEPÜ). *Zeitschrift für Parapsychologie und Grenzgebiete der Psychologie*, 40-41, 95-133.

Suitbert Ertel[8]

## Ein Forschungsbericht mit Versäumnissen

Die „konzeptuelle Replikationsstudie zu Experimenten zur Außersinnlichen Wahrnehmung" von Dr. Rey und seinen studentischen Mitarbeiterinnen hat didaktischen Wert. Denn eine Reihe von Versäumnissen in ihr, die im Folgenden aufgedeckt und besprochen werden, veranlassen mich auszuführen, wie man Experimente zur ASW mit mehr Aussicht auf Erfolg planen kann.

### *Versäumte begriffliche Präzision*

Nach Rey *et al.* sollten mit Hilfe von Zenerkarten Fähigkeiten in der Außersinnlichen Wahrnehmung für die Bereiche Telepathie, Hellsehen und Präkognition untersucht werden: *„Auf Grund von paranormalen Fähigkeiten der Probanden [wird] eine überzufällig hohe Trefferzahl [erwartet]"*. Soll sich die ASW-Fähigkeit also unter den Testbedingungen dieses Experiments in einer überzufällig hohen Trefferzahl auswirken? Doch im ersten Satz der Diskussion lesen wir: *„In der vorliegenden Untersuchung wurden keinerlei überzufällige Trefferraten bei der Auswahl von Zenerkarten verzeichnet und somit ASW nicht nachgewiesen"*. Ging es denn in diesem Experiment eher um einen „Nachweis" von ASW-Fähigkeit? Um einen solchen zu erbringen, muss man experimentell ungleich viel mehr als in dieser Studie auf die Beine stellen. Deren dürftige Bedingungen eignen sich eher dazu, einen Nachweis NICHT zu erbringen. Oder wollten die Untersucher nur wissen, ob die ASW-Fähigkeit auch unter diesen dürftigen Bedingungen noch Wirkungen zeigt, aber ohne dass ein möglicherweise negatives Ergebnis zur Frage der Existenz von ASW etwas beitragen könnte? Das aber hätten sie sagen müssen, und sie hätten es sich verkneifen müssen zu sagen, dass ein Nachweis zur Existenz von ASW nicht erbracht wurde. Dieser Zusatz allein lässt auf eine ziemliche Voreingenommenheit des Teams bzw. seines verantwortlichen Leiters gegenüber ASW schließen.

Den Formulierungen des Berichts mangelt es an begrifflicher Klarheit. Die Befundlage zur Existenz von ASW sei „uneindeutig", wird nur gesagt, mit Hinweis auf einige sich widerspre-

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

chende Forschungsergebnisse in der Ganzfeldforschung. Doch ein Ganzfeld-Experiment wird nicht geplant, sondern ein Experiment aus früherer Zeit, das man schon lange wegen methodischer Einwände durch Ganzfeld-Experimente ablösen wollte. Wegen der „uneinheitlichen Befundlage" in Experimenten aus neuerer Zeit wird ein Experiment aus älterer Zeit methodisch aufgewärmt, ohne dass die umfangreiche methodische Diskussion, die dem vorausging, auch nur gestreift wird.

### *Versäumte Erfassung der Psi-Fähigkeit*

Ich nehme an, die jungen lerneifrigen Psychologiestudentinnen um Dr. Rey hätten nichts gegen die Forderung einzuwenden, dass eine Variable wie Psi-Fähigkeit ebenso ernst genommen werden sollte wie Intelligenz, Musikalität oder Kreativität etc. Denn wenn ein Einfluss dieser uns geläufigeren Fähigkeitsvariablen z.B. auf den Schulerfolg in verschiedenen Schulfächern geprüft werden soll, dann muss man die Ausprägungsgrade der Intelligenz, Musikalität, Kreativität usw., die interindividuell verschieden sind, mit standardisierten Tests operationalisieren. Doch eine Operationalisierung der Psi-Fähigkeit, welche Einfluss nehmen soll auf das Voraussagen von Zener-Symbol-Serien, wird von Rey *et al.* nicht diskutiert.

Dieses Versäumnis findet man allerdings in der parapsychologischen Forschungsszene generell. Seit langem ist es mein Anliegen, Forscherkollegen davon zu überzeugen, dass ein Fortschritt in der experimentellen Parapsychologie erst nach Einführung von Psi-Tests zu erwarten ist, welche psychometrischen Ansprüchen genügen und diagnostische Differenzierungen zwischen psi-begabteren und weniger begabten Personen erlauben.

### *Versäumte Berücksichtigung der relativen Seltenheit von Menschen mit nachweisbarer Psi-Fähigkeit*

Zum Ergebnis meiner eigenen Psi-Forschung, in der ich erste Schritte zur Operationalisierung von Psi-Fähigkeit unternommen habe (Ertel, 2005a, 2005b), gehört der Befund, dass Psi-Fähigkeit, z.B. in der studentischen Population, mit der ich es fast ausschließlich zu tun habe, nicht normalverteilt ist wie etwa Intelligenz und Kreativität. Nur bei ca. 15-20 Prozent getesteter Personen, die nicht ausgelesen waren, fand ich signifikante Abweichungen von der Zufallserwartung, obgleich ich mit den Testpersonen innerhalb von zwei Stunden pro Person 360 Trials eines Forced-Choice-Tests durchführte, die auf zwei bis drei Testsitzungen verteilt wurden. Ich verwendete dafür den sogenannten „Ball Selection Test", den ich als ein ökonomisches Testinstrument für Psi-Fähigkeit eingeführt hatte. Für diesen benötigt man außer genügend Zeit lediglich einen Beutel mit Pingpongbällen.

Das Rey-Team hat mit 75 Trials pro Person (das sind nur ca. 20% der Trialzahl einer Ball-Test-Prüfung) drei Personen mit signifikanten[9] Trefferzahlen gefunden (mit 24, 24 und 31 Treffern, das sind nur 3% von 96 Personen). Doch nehmen wir einmal an, dass die beobachtete Trefferverteilung bei den Rey-Testpersonen mit gleicher Effektstärke wiederkehren würde, wenn 360 Trials statt nur 75 absolviert würden (also bei 360/75 = 4.8 mal so viele Trials und Treffer). Dann würden, wenn man die Trefferzahlen von Rey hochrechnet, 18 von 96 Personen signifikante Trefferzahlen über dem Erwartungswert erzielt haben, das sind 20% der Testpersonen von Rey *et al.*[10] Diese Trefferleistung würde sich größenordnungsmäßig in das Leistungsspektrum einordnen, das ich selbst mit meinen Studenten regelmäßig repliziere (Ertel, 2009). Was das Forscherteam Rey versäumt hat, war, durch eine vielfache Vergrößerung der Trialzahl (möglichst auf das Fünffache) den Probanden Gelegenheit zu geben, ihre ASW-Fähigkeit statistisch hinreichend zum Ausdruck kommen zu lassen.

Trotz der Seltenheit der Psi-Fähigkeit ist in der Parapsychologie ein Denkfehler weit verbreitet, den man korrigieren könnte, wenn man aus dieser Tatsache, über die auch andere Forscher gelegentlich berichten, richtige Schlussfolgerungen zöge. Immer noch testet man – wie das Rey-Team – unausgelesene Stichproben von Personen und verarbeitet die Daten für diese Stichproben aggregativ. Dabei macht man den Fehler vorauszusetzen, dass, wenn es Psi-Fähigkeit gibt, diese wie andere Fähigkeiten bei allen Probanden mehr oder weniger ausgeprägt sein müsse. Alle getesteten Personen hätten zum Summenwert der Stichprobe mehr oder weniger etwas beizutragen. Doch da ca. 80% der getesteten Personen, die unausgelesen einer Population entnommen werden, nichts von dieser Fähigkeit besitzen oder diese Fähigkeit, wenn sie sie haben sollten, unbewusst dauerhaft unterdrücken, werden durch das Aggregieren der Trefferzahlen die ASW-erhöhten Trefferbeiträge von etwa 20% einer so entnommenen Stichprobe ausgedünnt. Ihre ASW-Leistung kann sich für die Stichprobe insgesamt nicht mehr genügend auswirken, wofür die Untersuchung von Rey *et al.* ein Beispiel ist. Die Trefferquote der Probandin mit 31 Treffern, $p = .000004$, die ins Aggregat der Stichprobe eingespeist wurde, wirkte sich auf das Ergebnis der Stichprobe insgesamt statistisch nicht merklich aus. Ein simples Gedankenexperiment zeigt den Fehler auf: Würde man einen Test zur Messung der Fähigkeit des absoluten Gehörs entwickeln und eine unausgelesene Stichprobe damit testen, dann würde diese Fähigkeit, die im Massentest bei nur vereinzelten Personen Ausreißerwerte hervorbringt, beim Aufsummieren der Werte aller getesteten Personen keine signifikanten Indikatoren mehr hervorbringen. Man würde dann kaum Chancen haben, das absolute Gehör als Phänomen zu

9 Die Berechnung von Signifikanz für diesen Zweck erfolgt ohne Berücksichtigung der sonst üblichen Aufgabe, dass *p*-Werte zum Hypothesen-Testen verwendet werden. Korrekturen an *p* à la Bonferroni erübrigen sich.

10 Unter den Probanden des Rey-Teams gab es sechs mit 18 Treffern, zwei mit 19, zwei mit 20, drei mit 21, drei mit 22, zwei mit 24 und eine mit 31 Treffern (siehe die Abbildung 1 bei Rey *et al.*).

entdecken, wenn man von diesem nicht auf andere Weise, ohne Statistik, sichere Kunde hätte.

### *Versäumte Berücksichtigung von Psi Missing*

Dieses Versäumnis haben Rey *et al.* selbst erwähnt: Eine „überzufällig unterdurchschnittliche Trefferrate der Probanden (‚psi-missing') blieb unberücksichtigt". Beim Leser kann hier der Verdacht entstehen, als seien die Untersucher von überzufällig vielen Psi-Missing-Fällen überrascht worden: *„Erst im Verlauf der Untersuchung hat sich herausgestellt, dass eine hinreichende Anzahl an Probanden für die Studie rekrutiert werden konnte, die auch eine zweiseitige Testung unter Berücksichtigung von ‚psi-missing' mit akzeptabler Teststärke erlaubt hätte".* Wenn die Überlegungen hinsichtlich Teststärke keinen Hinderungsgrund ergaben, warum hat man dann keine zweiseitige Auswertung vorgenommen, bei der auch Psi Missing zu Buche schlägt? Dass man das ursprünglich nicht vor hatte, wäre als Hinderungsgrund nicht hinzunehmen. Je öfter ich den schillernden Satz mit den „rekrutierten" Studenten lese (die waren doch schon „rekrutiert"?), die eine „hinreichende Anzahl" hatten (wofür war diese „hinreichend", und welche Kriterien wurden angelegt?), umso weniger kann ich den Verdacht unterdrücken, dass die unerwarteten Hinweise darauf, dass Psi in Gegenrichtung mit im Spiele war, der Versuchsleitung zu psi-verdächtig erschienen. Konnte Rey mit einem Ergebnis dieser Untersuchung, über das er nach einer Veröffentlichung Diskussionen in seinem akademischen Umfeld zu erwarten waren, vielleicht besser leben, wenn ASW-Effekte jeder Art ausblieben? Vielleicht aber legt Rey uns noch ein entlastendes Versuchsprotokoll im Original mit Zahlen im Detail vor.

### *Versäumte Mitteilung über Varianzen*

Eine Überprüfung der Trefferzahlen auf individueller Ebene ist uns Lesern des Berichts nicht möglich, da nur die über die drei Bedingungen summierten Trefferzahlen mitgeteilt werden (in Abbildung 1 bei Rey *et al.*). Auf einer Seite hätte man die Trefferzahlen der 96 Personen – getrennt für die Bedingungen Telepathie, Hellsehen und Präkognition – auflisten können. Es kann sein, dass unter einer Bedingung nur Treffer-Überhänge vorkamen, unter einer anderen Bedingung vielleicht viele Psi-Missing-Fälle. Die Trefferzahlen können erfahrungsgemäß unter wechselnden Bedingungen auch sehr schwanken, vor allem bei den Psi-Begabten. Beim Aufsummieren der Trefferzahlen geht die durch den Wechsel der Bedingungen möglicherweise hervorgerufene Varianz verloren.

Auch sollte man noch prüfen, ob und wie hoch die unter den drei Bedingungen gewonnenen Trefferzahlen untereinander korrelieren, was aber ohne Datendifferenzierung nicht möglich ist. Vielleicht korrelieren die Treffer der Telepathiebedingung mit denen der Hellsehbedingung eher positiv; mit den Treffern der Präkognitionsbedingung könnten die Treffer der beiden

anderen Bedingungen eher nicht oder negativ korrelieren, wenn auch wegen der geringen Trialzahl alles nur andeutungsweise. Wenn man mehr als nur andeutungsweise Korrelationen unter den drei Bedingungen feststellen würde, egal, ob positive oder negative, dann wären diese ein Indiz für das Vorhandensein von ASW.

Warum führt der in der Methodendarstellung beschriebene dreifaktorielle Versuchsplan, bei dem man für die Datenauswertung eine Varianzanalyse erwartet, nicht auch zum Einsatz der Varianzanalyse? Wechselwirkungen zwischen wirksamen Faktoren könnten durch eine Varianzanalyse zutage treten. Warum erfährt der Leser darüber nichts? Wurde eine Varianzanalyse vielleicht durchgeführt, ohne dass deren Ergebnisse mitgeteilt werden? Man hat Grund, den Verdacht auf Fehlentscheidungen der Versuchsleitung, die tendenziös motiviert sein könnten, nicht fallen zu lassen.

## *Schlussfolgerung*

Das Experiment von Rey *et al.* leidet vor allem daran, dass die geforderten 75 Trials pro Person ganz und gar nicht genügten, um die Existenz von ASW-Fähigkeiten nachzuweisen oder auch nur, um über die Wirkungen dieser Fähigkeit irgendetwas zu erfahren. Die Tatsache der Seltenheit von ASW sollte alle ASW-Forscher dazu veranlassen, dies in ihren Versuchsdesigns zu berücksichtigen. Dazu gehört an erster Stelle das Screenen unausgelesener Stichproben von Probanden und das Selektieren der Psi-Begabten. Dem vorausgehen sollte eine Fortsetzung derjenigen Forschung, die die Entwicklung eines Psi-Begabungstest mit psychometrischen Qualitäten zum Ziel hat, meinen Ansatz in dieser Richtung könnte man fortführen, so dass in Zukunft Psi-Fähigkeit ähnlich wie Intelligenz etc. mit standardisierten Methoden erfasst werden kann. Mit meinen Publikationen darüber (s. auch Ertel, 2007, 2008, 2010a, 2010b) kann man sich einarbeiten. Leistungen von Probanden bei beliebigen anderen Psi-Aufgaben sollten mit den Ergebnissen, die zuvor mit standardisierten Psi-Screening-Tests gewonnen werden, mehr oder weniger voraussagbar werden.

Zur Ermittlung der Psi-Fähigkeit kann man den Testpersonen das Testmaterial mit Instruktion als Hausaufgabe mitgeben. Beim Einsatz von Heimtests geht es nicht um den Nachweis von Psi-Fähigkeit, sondern um ein ökonomisches Auswählen psi-begabter Personen für weitere Psi-Experimente, bei denen stärkere Kontrolle ausgeübt werden kann. Chris French, führender Skeptiker aus London, hat mit zwei seiner Studenten meinen Ball-Test eingesetzt, der unter Kontrolle (nicht als Heimtest) durchgeführt wurde. Durch eine Auswertung der in den studentischen Prüfungsarbeiten mitgeteilten Trefferscores, die mit 360 Trials pro Person erzielt wurden (mit N = 40 unausgelesenen Teilnehmern), wurden hochsignifikante ASW-Effekte ermittelt ($p$ = .0003; vgl. Ertel, 2010a). Der zeitliche Aufwand zur Gewinnung zuverlässiger

Psi-Fähigkeitswerte im Einzeltest ist nicht viel größer als der zeitliche Aufwand von Intelligenz-Einzeltests, mit denen man in der Regel zuverlässige Intelligenzindikatoren (IQs) ermittelt.[11]

### Literatur

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

Ertel, S. (2005b). The ball drawing test: Psi from untrodden ground. In Thalbourne, M.A., & Storm, L. (Eds.), *Parapsychology in the Twentieth Century: Essays on the Future of Psychical Research* (S. 90-123). Jefferson NC & London: McFarland.

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. (2009). Replikation von ASW-Heimtest-Ergebnissen im Labor. Zur Validierung der Ball- und Perlentests. *Zeitschrift für Anomalistik*, 9, 108-139.

Ertel, S. (2010a). Psi in a skeptic's lab? Ball selection results replicated. *Journal of Scientific Exploration*, 24, 581-598.

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

Sharp, V., & Clark, C.C. (1937). Group tests for extra-sensory perception. *Journal of Parapsychology*, 1, 123-142.

Wolfgang Helfrich[12]

## Ein klares „Jein"

Es ist bekannt, dass auch sehr umfangreiche Psi-Experimente ohne signifikantes Ergebnis bleiben können. Ein krasses Beispiel dieser Art sind die dreifachen Psychokinese-Experimente an binären Zufallsgeneratoren (in Princeton, Freiburg und Gießen), wie beschrieben in Jahn *et al.* (2000).

11 Die Möglichkeit einer Durchführung des Ball-Tests in Gruppen wurde noch nicht geprüft. Ich zweifle, ob man damit Erfolg haben wird. Schon Sharp & Clark (1937) berichteten: „*Individual tests gave higher scores than the group tests*". Ein hemmender Einfluss durch unkontrollierte soziale Faktoren könnte zu groß werden.

12 Dr. Wolfgang Helfrich ist pensionierter Professor für Physik an der Freien Universität Berlin. E-Mail: helfrich@physik.fu-berlin.de.

Das Gesamtergebnis von Rey *et al.* ist jedoch hochsignifikant, wenn man die Werte des einen herausragenden Probanden mit einer Zufallswahrscheinlichkeit von $2 * 10^{-6}$ gelten lässt. Die Zufallswahrscheinlichkeit des Gesamtergebnisses ergibt sich aus diesem Wert, indem man ihn durch die Zahl der Probanden (96) dividiert. Sie ist immer noch weit unterhalb der Signifikanzgrenze von $5 * 10^{-2}$.

**Literatur**

Jahn, R.G., Dunne, B.J., Bradish, G., Dobyns, Y., Lettieri, A., Nelson, R.D., Mischo, J., Boller, E., Bösch, H., Vaitl, D. Houtkooper, J., & Walter, B. (2000). Mind/Machine Interaction Consortium: PortREG replication experiments. *Journal of Scientific Exploration*, 34, 499-555.

Ulrich Timm[13]

## Drei statistische Anmerkungen

Nachdem die Autoren selbst auf einige Schwächen ihrer experimentellen Methodik hingewiesen haben, möchte ich mich auf drei Bemerkungen zur statistischen Auswertung beschränken:

**1)** Die Autoren geben für jeden ihrer drei primären Signifikanztests eine geschätzte *Teststärke* von mehr als .99 an, d.h. bei ihrem Design sollte ein real existierender Effekt in über 99% aller Fälle signifikant werden. Tatsächlich resultierte bei sämtlichen mit der Vpn-Gesamtheit durchgeführten Tests keine Signifikanz. Ein konsequent denkender Leser könnte daraus schließen, dass hier mit überwältigender statistischer Sicherheit die *Nichtexistenz* von Telepathie, Hellsehen und Präkognition in vergleichbaren Experimenten belegt worden sei. Diese Folgerung wäre jedoch unberechtigt, da die angegebene Teststärke aus drei Gründen stark überschätzt sein dürfte:

a) Die zu erwartende Effektstärke wird viel zu hoch angesetzt; sie sollte aus dem Durchschnitt *aller* bisherigen Experimente – und nicht nur der erfolgreichen – geschätzt werden.

b) Das verlangte Signifikanzniveau α ist mit .05 zu niedrig und sollte mindestens .01 betragen.

13 Dr. Ulrich Timm ist Psychologe, Parapsychologe und Mitglied der Parapsychological Association. Er war ab 1965 langjähriger Mitarbeiter des Instituts für Grenzgebiete der Psychologie und Psychohygiene sowie Redakteur und zeitweise Mitherausgeber der *Zeitschrift für Parapsychologie und Grenzgebiete der Psychologie*. E-Mail: ultimm@web.de.

c) In der benutzten Formel wird die hohe intra- und interexperimentelle Variabilität von Psi-Resultaten nicht berücksichtigt und stattdessen den Treffern die Varianz einer Binomialverteilung zugeschrieben, was nur unter $H_0$ (z.B. im Signifikanztest) korrekt wäre.

Die Teststärke ist also deutlich niedriger anzusetzen, und es ist sogar sinnvoll, über ihre Erhöhung durch *Vergrößerung des Stichprobenumfangs* nachzudenken. Da die Vpn-Zahl mit N = 96 bereits vorbildlich hoch ist, wäre in erster Linie eine Erhöhung der Zahl der Einzeltrials pro Vp, die nur n = 3•25 beträgt, zu empfehlen gewesen. Wegen der Gefahr eines intrapersonellen Declines sollte n aber nicht extrem hoch sein. Als wünschenswertes Minimum kann man die Zahl n = 250 ansehen, die genau derjenigen Menge von Trials entspricht, die auf den klassischen Protokollbögen J.B. Rhines (aufgeteilt in 25 „runs") Platz hatten. Bei noch größerem n bietet sich eine Aufteilung in mehrere Sitzungen an.

**2)** Es ist auffällig, dass die Autoren nirgendwo ein *Gesamtresultat* für die Summe aller Trefferzahlen präsentieren. Insbesondere werden die drei Versuchsbedingungen (bzw. Psi-Modalitäten) Telepathie, Hellsehen und Präkognition demonstrativ getrennt behandelt. Hier zeigt sich eine prinzipielle Problematik, die entsteht, wenn gewisse Regeln der Normalwissenschaft in eine Parawissenschaft übertragen werden: In der Parapsychologie gilt nämlich nach wie vor eine *übergeordnete* Nullhypothese ($H_0$), welche die Nichtexistenz *beliebiger* Psi-Effekte unter *beliebigen* Versuchsbedingungen postuliert. (In operationaler Definition handelt es sich schlicht um alle Effekte und Bedingungen, denen das gemeinsame Merkmal zukommt, im Untersuchungsprogramm der Parapsychologen vorzukommen.) Dieser generellen $H_0$ steht in einem Standard-Psi-Experiment die generelle $H_1$ gegenüber, dass *mindestens einer* der darin geprüften Psi-Effekte unter *mindestens einer* der Bedingungen aufgetreten ist. Deshalb sind diese Effekte in Bezug auf die übergeordnete $H_0$ als gleichartig anzusehen und die zugehörigen Resultate müssen aggregiert und einem *globalen Signifikanztest* unterzogen werden.[14]

Nur wenn der Globaltest signifikant ist, dürfen auch die Einzelresultate in normaler Weise auf Signifikanz geprüft werden, was meistens missachtet wird. Andernfalls muss eine *wesentliche Korrektur* erfolgen. Wäre etwa bei insignifikantem Globaltest *nur einer* von k Einzeltests signifikant, so wäre die normal berechnete Irrtumswahrscheinlichkeit *P* ungültig. Sie müsste durch die viel höhere Wahrscheinlichkeit *P´* ersetzt werden, in (mindestens) *irgendeinem* von k Einzeltests eine Signifikanz zu erreichen. *P´* ist bei unabhängigen Einzelresultaten ungefähr gleich *k•P*, so dass eine ähnliche Wirkung wie bei der sog. Bonferroni-Korrektur eintritt und die Signifikanzchancen rapide absinken. (Die Berechnung von *P´* aus dem niedrigsten Einzel-*P*

14 Wegen der hohen Variabilität ist allerdings die Testung der einfachen Treffersumme, die meist einer Summierung der z-Werte von Teilresultaten äquivalent ist, nicht optimal. Deshalb plädiere ich seit Langem für die Summierung bestimmter nichtlinearer Funktionen von z – wie $z^2$ oder lnP –, die u.a. zu einer stärkeren Gewichtung extremer Einzelresultate führen.

kann als ein ungewöhnlicher Globaltest angesehen werden, der nur auf dem günstigsten aller Einzelresultate basiert.) Wenn *P´* aber trotzdem hoch signifikant bleibt, ist dem das gleiche interpretatorische Gewicht beizumessen wie bei einem konventionellen Globaltest.

Für das vorliegende Experiment scheinen diese Feststellungen auf den ersten Blick bedeutungslos zu sein, da weder der (fehlende, aber leicht zu berechnende) Globaltest noch irgendein Einzeltest signifikant ist. Aber dieser Eindruck erweist sich als falsch, wenn man auf die hierarchisch niedrigere Ebene der *Resultate pro Vp* übergeht. Hier erreicht eine der 96 Vpn eine sensationelle Trefferzahl mit *P*<.000005! Um das korrigierte *P´* zu erhalten, könnte man also *P* mit dem Korrekturfaktor k=96 multiplizieren. Das genügt aber nicht. Denn neben diesen 96 Teilresultaten sind auf mehreren Ebenen des Experiments noch zahlreiche andere Teilresultate definierbar, die auch unter $H_0$ irgendwann ein extremes *P* erreichen könnten (z.B. 3 Psi-Modalitäten, 288 Runs, 3 Feedback-Bedingungen, 5 Versuchsleiter). Hinzu kommen Korrelationen der Psi-Scores mit diversen biosozialen Variablen (hier immerhin drei) und eventuell die Prüfung von Sondereffekten (Psi-Missing, Decline, Displacement usw.), die den Untersuchern häufig erst *post hoc* einfallen. Solche zusätzlichen Auswertungen führen aber zwingend zu k‘ *verschiedenen* Globalwerten P oder P‘, deren niedrigster erneut auszuwählen und (approximativ) mit einem zusätzlichen Korrekturfaktor k‘ zu multiplizieren wäre. Im vorliegenden Fall sollte daher mindestens mit k‘=5 und k•k‘=480 gerechnet werden, so dass für die Spitzen-Vp ein *P´*<.0025 resultiert. Das Resultat dieser erfolgreichsten aller 96 Vpn wird erst dadurch von einer vieldeutig anekdotischen auf eine statistisch relevante Ebene gehoben und erweist sich dort immer noch als hoch signifikant!

Falls die Autoren auch nichtstatistische Täuschungsmöglichkeiten ausschließen könnten (was sie offensichtlich nicht tun), dürfte also die ASW zumindest bei *einer* ihrer Vpn als hochgradig gesichert und mit der gleichen Sicherheit als *prinzipiell existent* angesehen werden. Ihrer Feststellung „in der vorliegenden Untersuchung wurden keinerlei überzufällige Trefferraten [...] verzeichnet“ ist jedenfalls nicht zuzustimmen. Sie gilt nur für die konventionelle Testung der einfachen Treffersummen über alle Vpn. Diese erweist sich immer wieder als ein zu enges Korsett, das durch alternative statistische Methoden aufgebrochen werden muss.

**3)** Schließlich sei erwähnt, dass die mit einem geschlossenen Kartensatz („closed deck“) unter $H_0$ erlangten Trefferzahlen nicht (wie die Autoren sagen) exakt binomialverteilt sind, sondern eine sog. *Matching-Verteilung* besitzen. Diese ist allerdings gut durch die Binomialverteilung approximierbar, solange die verschiedenen Kartensymbole ungefähr gleich häufig gewählt werden. Mit zunehmender Ungleichheit der Wahlhäufigkeiten wird aber der Binomialtest immer konservativer, weil er die damit verbundene Abnahme der Treffervarianz nicht berücksichtigt. (Wenn 25mal das gleiche Symbol gewählt wird, wird die Varianz sogar 0, weil stets 5 Treffer resultieren!) Es ist dann vorteilhafter und genauer, mit der verringerten Varianz

einen z-Test durchzuführen. Die Formel für diese Varianz ist in dem von den Autoren zitierten Aufsatz von Burdick & Kelly (1977: 89) zu finden.

**Literatur**

Burdick, D.S., & Kelly, E.F. (1977). Statistical methods in parapsychological research. In Wolman, B.B. (Ed.), *Handbook of Parapsychology* (S. 81-130). New York: Van Nostrand Reinhold.

Gerd H. Hövelmann[15]

## Methodologie oder politisches Kalkül?

### Anmerkungen zum Kommentar von Wolfgang Ambach

Dr. Wolfgang Ambach hat das erste Viertel seines Kommentars zur experimentellen Studie von Dr. Günter Daniel Rey und seinen studentischen Mitarbeiterinnen der redaktionellen Vorbemerkung gewidmet, die jener Arbeit als Fußnote beigegeben war. Als verantwortlicher Redaktionsleiter möchte ich zu einigen seiner Ausführungen und Vermutungen kurz Stellung nehmen. Redaktionelle Entscheidungsfindungen im Nachhinein öffentlich zu kommentieren, entspricht eigentlich nicht den Gepflogenheiten. Da Herrn Ambachs mitunter sehr treffliche Anmerkungen und Beobachtungen aber dennoch in Teilen eine falsche Spur legen, deren Verfolgung auch bei möglichen künftigen Gelegenheiten zu Irritationen führen könnte, sei diese Ausnahme gestattet.

„Der Leser erfährt", so fasst Ambach die redaktionelle Vorbemerkung zunächst zusammen, „(a) dass hier ein offenbar veraltetes Forschungsparadigma zum Einsatz kam und noch dazu in dilettantischer Weise umgesetzt wurde. Weiterhin erfahren wir, (b) dass eine Nichtveröffentlichung […] Vorwürfen der Datenselektion Vorschub geleistet hätte, was man durch den Entschluss zur Veröffentlichung vermied. Schließlich nehmen wir zur Kenntnis, (c) dass der aufmerksamkeitswürdige Umstand, dass die Studie im Rahmen der studentischen Ausbildung an einer deutschen Universität durchgeführt wurde, weiterer Grund für die Annahme des Artikels war."

Herrn Ambachs weitergehenden Ausführungen und Überlegungen zu seinen Unterpunkten (b) und (c) kann ich fast uneingeschränkt beipflichten. Seine Deutungsversuche zu Punkt (a) scheinen mir jedoch in mehrfacher Hinsicht problematisch. Auf diese möchte ich deshalb

15 Gerd H. Hövelmann, M.A., studierte Philosophie, Linguistik, Literaturwissenschaft und Psychologie, war von 1984 bis 1993 wissenschaftlicher Mitarbeiter am Institut für Philosophie der Universität Marburg und ist seither selbständig. Er ist der Redaktionsleiter der *Zeitschrift für Anomalistik*.

in vier Punkten erwidern. Herr Ambach argumentiert:

> „Ad (a): Während das Redaktionskollegium das Aufgreifen eines historischen Ansatzes an sich vielleicht noch mehrheitlich gutgeheißen hätte, wurde der Bruch mit früher zu diesem Ansatz etablierten Standards offenbar sehr bemängelt. Für die damals üblichen aufwendigen Vorkehrungen gegen Täuschung und sensorische Lecks gab und gibt es sicherlich gute Gründe. Dennoch möchte ich im Rückblick zu bedenken geben, dass auch die frühere Maximierung von Abschirmung und Kontrolle zu keinem Zeitpunkt geeignet war, die Möglichkeit artifizieller Einflüsse gänzlich (und vor allem: in aller Augen!) auszuschließen. Gleichzeitig waren die Optimierungsversuche – so vermute ich – nicht nur durch forscherische Neugier motiviert, sondern auch durch das Bestreben, den Kampf gegen die Überzeugungsgegner, die Ungläubigen, eines Tages doch gewinnen zu können. Rey et al. haben mit dieser Tradition einfach gebrochen; sie waren einfach nur neugierig, ohne gleich wasserdichte Beweise liefern zu wollen. Die Autoren kämpften eindeutig nicht den gleichen Kampf wie die Forscher in der Rhine'schen Tradition, und sie hatten offenbar auch nicht das gleiche imaginierte Gegenüber. Ich frage mich, inwieweit die Kontroverse um die Publikationswürdigkeit des Artikels, soweit sie sich auf die Manipulationsvorkehrungen bezieht, tatsächlich wissenschaftlich motiviert ist, und in wie weit sie die verschiedenen Traditionszugehörigkeiten der Autoren und der einzelnen Reviewer widerspiegelt."

(1) Wolfgang Ambach glaubt, wie er an anderer Stelle sagt, dass die redaktionelle Vorbemerkung „weniger über den Artikel selbst" aussagt, „als über die Bedingungen und das Spannungsfeld, denen die Forschung zu unkonventionellen Fragestellungen […] generell ausgesetzt ist." Und er vermutet, dass es wohl weniger methodologische als vielmehr politische Gründe gewesen seien, die Anlass zu redaktionellen Bedenken gegeben hätten. Dies trifft entschieden nicht zu, was auch leicht plausibel zu machen ist, wenn wir uns eine vergleichbare Ausgangslage in irgendeiner anderen wissenschaftlichen Disziplin vergegenwärtigen. Womit hätten wir zu rechnen, wenn beispielsweise in Ambachs eigenem Fach, der Psychophysiologie, eine empirische Arbeit zur Veröffentlichung eingereicht würde, die sich (a) eines experimentellen Designs bediente, das bereits vor rund einem halben Jahrhundert als unergiebig verabschiedet worden ist, das (b) zudem alle wesentlichen Kontroll-, Überprüfungs- und Sicherungsmaßnahmen ignorierte, die dieses Design immerhin einstmals ausgezeichnet haben, und das (c) Teile seiner Begrifflichkeit und seiner begleitenden methodischen Veranstaltungen (hier z.B. Feedback) in einer Weise einsetzte, die den bisherigen Gepflogenheiten und Verabredungen widerspräche? Die sich zwangsläufig einstellenden Bedenken hinsichtlich der Angemessenheit der Verfahren, der potentiellen Orientierungsleistung und nicht zuletzt der methodischen Stringenz einer solchen Studie hätten eine Veröffentlichungsvorlage erwartbar in jeder wissenschaftlichen Disziplin und im Urteil jeder beliebigen wissenschaftlichen Zeitschriftenredaktion an die Grenze der Publikationsfähigkeit – und mehrheitlich wohl noch ein Stück darüber hinaus – getra-

gen. Entsprechende Publikationsvorbehalte sind unter solchen Umständen also nicht nur nicht aufsehenerregend, sondern der selbstverständliche Normalfall.

(2) Herr Ambach „möchte […] im Rückblick zu bedenken geben, dass auch die frühere Maximierung von Abschirmung und Kontrolle zu keinem Zeitpunkt geeignet war, die Möglichkeit artifizieller Einflüsse gänzlich (und vor allem: in aller Augen!) auszuschließen". Damit hat er völlig recht. Eben dies waren ja die Gründe, weshalb das betreffende experimentelle Paradigma vor mehr als einem halben Jahrhundert verabschiedet worden ist. Wenn man aber der wohl begründeten Auffassung ist, dass die betreffenden Verfahren bereits in ihrer methodisch strengsten Form nicht dazu geeignet waren, die ihnen einstmals zugedachte Beantwortung empirischer Forschungsfragen zu leisten, wie kann dann die Vernachlässigung gerade jener Faktoren, die ihre methodische Rigidität ausgemacht haben, Anlass dazu geben, ein „Loblied auf die Naivität" anzustimmen? Und wie kann man die heutige Wiederbelebung dieses Paradigmas dann für einen, wie es bei Ambach an späteter Stelle heißt, „respektablen Ansatz" halten, „einen experimentellen Klassiker neu mit Leben zu füllen und ihn zur Grundlage für ein sauber durchgeführtes Lehrexperiment zu machen"?

(3) Herr Ambach fragt sich des weiteren, ob die methodische Optimierung dieses alten experimentellen Paradigmas seinerzeit nicht (mindestens) auch durch einen „Kampf gegen die Überzeugungsgegner, die Ungläubigen" motiviert war, und ihn beschleicht der Verdacht, dass auch die Diskussion um die Publikationsfähigkeit der vorliegenden Arbeit nicht so sehr „tatsächlich wissenschaftlich motiviert" als vielmehr am Feindbild eines ungläubigen „imaginierten Gegenüber[s]" orientiert gewesen sei. Diese Vermutung macht sich zweier Unterstellungen schuldig, die durch die tatsächlichen Verhältnisse nicht gedeckt sind. Zum einen unterstellt sie nämlich, dass „die Parapsychologen" bekanntlich die „Psi-Gläubigen", ihre imaginierten oder tatsächlichen „Überzeugungsgegner" hingegen die „Ungläubigen" seien. Diese propagandistische Vorstellung hat international, insbesondere aber im deutschen Sprachraum eine veritable Tradition. Auch ist nicht zu bestreiten, dass es, zumal in den 1970er und 1980er Jahren, angesichts einer Gegnerschaft gegen die parapsychologische Forschung, die sich weniger durch ihre Kenntnis als durch ihre Militanz auszeichnete,[16] durchaus sorgsame bis ängstliche Bemühungen gegeben hat, dieser Opposition nur ja nicht auch noch mittels unzulänglich konzipierter experimenteller Verfahren in die Hände zu spielen. Auch wenn sich modernere und durchaus besser informierte Kritiker der Parapsychologie (etwa Alcock, 1987; Hergovich, 2001) nach wie vor gerne dieses liebgewonnenen Stereotyps des notorisch glaubensversessenen Parapsychologen und seines rational-zweifelnden Gegenübers bedienen, sind die tatsächlichen Verhältnisse jedoch um sehr Vieles komplizierter (vgl. z.B. Hövelmann, 1987, 1988; Blackmo-

16 Sehr instruktive Beispiele wird der Interessent u.a. bei Prokop & Uhlenbruck (1975), Schäfer (1978), Glowatzki (1980), Wimmer (1979, 1980) und Prokop & Wimmer (1987) finden.

re, 1989; Zingrone, 2006; Hövelmann & Michels, 2011). Insbesondere zeigen seit Jahrzehnten ausnahmslos alle Umfragen unter Mitgliedern der Parapsychological Association bei der Frage, ob „Psi" (was immer darunter konkret verstanden werden mag) existiere oder experimentell nachgewiesen sei, Zustimmungsraten, die zwischen gerade einmal 10 und höchstens 30 Prozent liegen (vgl. z.B. May, 2009). [17] Wer die parapsychologische Fachliteratur und ihre Vorläufer seit dem Ende des 19. Jahrhunderts im Detail kennt, der kann eine Diskussion um dieses Forschungsgebiet, die sich einer Dichotomisierung in vermeintlich „Gläubige" und vermeintlich „Ungläubige" bedient, nur für grotesk unangemessen halten. Im übrigen möchte ich grundsätzlich darauf bestehen, dass wir unser Augenmerk weniger dem zuwenden, was der eine oder andere vielleicht glauben möchte, sondern vielmehr dem, was wir konsensfähig wissen können.

(4) Das oben zitierte Argument von Wolfgang Ambach enthält implizit noch eine zweite Unterstellung (oder sagen wir: eine Mutmaßung), die hier bestritten werden muss: den Verdacht nämlich, dass auch die redaktionelle „Kontroverse um die Publikationswürdigkeit" des Artikels von Rey und Mitarbeiterinnen den unter Punkt (3) erörterten Bedingtheiten der parapsychologischen Forschungstradition und einer ihnen entspringenden politischen Strategie geschuldet sein könnte. Diese Mutmaßung mag legitim sein, aber sie trifft nicht zu, denn sie würde voraussetzen, dass die *Zeitschrift für Anomalistik* entweder selbst ein parapsychologisches Publikationsorgan sei oder doch wenigstens eines, das eine eigene Position in den Diskussionen um parapsychologische Forschungsthemen verträte. Dies ist jedoch entschieden nicht der Fall, und jede Parteilichkeit der Redaktion – oder auch nur Versuche einschlägigen Taktierens – stünden in einem eklatanten und ganz und gar unentschuldbaren Widerspruch zur Publikationspolitik dieser Zeitschrift, die gerade in der genannten Hinsicht programmatisch einer strikten Neutralität verpflichtet ist. Nicht politisches Kalkül, sondern allein die Qualitäten der eingereichten Arbeit haben Anlass zu redaktionellen Vorbehalten hinsichtlich ihrer Publikationsfähigkeit gegeben.

## Literatur

Alcock, J.E. (1987). Parapsychology: Science of the anomalous or search for the soul? *Behavioral and Brain Sciences*, 10, 553-565.

Blackmore, S.[J.] (1989). What do we really think? A survey of parapsychologists and skeptics. *Journal of the Society for Psychical Research*, 55, 251-262.

Glowatzki, G. (1980). Erwiderung zu „Parapsychologie – Scharlatanerie oder Wissenschaft?". *Schweizerische Rundschau Medizin (PRAXIS)*, 69, 547-549.

17 Dabei sind die Schwankungen eher unterschiedlichen Formulierungen der betreffenden Frage geschuldet, als dass sie eine Entwicklung in die eine oder andere Richtung andeuteten.

Hergovich, A. (2001). *Der Glaube an Psi. Die Psychologie paranormaler Überzeugungen*. Bern: Verlag Hans Huber.

Hövelmann, G.H. (1987). "Please wait to be tolerated": Distinguishing fact from fiction on both sides of a scientific controversy. *Behavioral and Brain Sciences*, 10, 592-593.

Hövelmann, G.H. (1988). Parapsychologists and skeptics – problems of identification: Some personal comments evoked by J.C. Jacobs. *SRU Bulletin*, 13, 125-132.

Hövelmann, G.H., & Michels, H. (Eds.) (2011). *Legitimacy of Unbelief: The Collected Papers of Piet Hein Hoebens*. Eindhoven: Synchronicity Research Unit.

May, E.C. (2009). Facing the challenges of parapsychology. In Roe, C.A., Kramer, W., & Coly, L. (Eds.), *Utrecht II: Charting the Future of Parapsychology. Proceedings of an International Conference held in Utrecht, The Netherlands, October 16-18, 2008* (S. 224-238). New York: Parapsychology Foundation.

Prokop, O., & Uhlenbruck, G. (1975). Über Wissenschaftskriminalität. Manifestationen von Dunkelfällen, Scharlatanerie und Außenseitertum in Teilbereichen der Medizin. *DDR-Medizin-Report*, 4, 966-985.

Prokop, O. & Wimmer, W. ([2]1987). *Der moderne Okkultismus. Parapsychologie und Paramedizin. Magie und Wissenschaft im 20. Jahrhundert*. 2., überarb. u. erw. Aufl. Stuttgart & New York: Gustav Fischer Verlag.

Schäfer, H. (1978). Parakriminologie – Glossierende Anmerkungen zur DKG-Tagung. *Kriminalistik*, 32, 363-365.

Wimmer, W. (1979). Okkultismus und Rechtsordnung. Die Methoden der Parapsychologie in kriminalistischer und juristischer Sicht. *Archiv für Kriminologie*, 164, 1-16.

Wimmer, W. (1980). Hexenwahn an Universitäten? *Zeitschrift für Allgemeinmedizin*, 56, 1390-1400.

Zingrone, N.L. (2006). *From Text to Self: The Interplay of Criticism and Response in the History of Parapsychology*. Dissertation. Edinburgh: University of Edinburgh.

## *Autorenantwort:*

Günter Daniel Rey

## Zur Teststärke und anderen Kritikpunkten

Zunächst möchte ich mich ganz herzlich für die zahlreichen und vielfältigen Kommentare zum Beitrag „Konzeptuelle Replikationsstudie zu Experimenten zur außersinnlichen Wahrnehmung" bedanken. Mein Dank gilt insbesondere all den wissenschaftlichen Experten auf diesem Gebiet, die sich die Zeit genommen haben, die von mir und anderen Novizen konzipierte und sicherlich mit erheblichen Mängeln behaftete Arbeit in so ausführlicher Form kritisch zu beleuchten. Bedanken möchte ich mich auch bei den anonymen Gutachtern für ihre kritischen Kommentare während der Begutachtungsphase.

Im Folgenden gehe ich auf einige ausgewählte Aspekte näher ein, die in den Kommentaren aufgegriffen wurden. Zunächst komme ich der Bitte von Herrn Ambach gerne nach, die Entstehungsgeschichte des Experiments und mögliche Folgestudien näher zu beleuchten. Es folgt eine Diskussion über den in mehreren Kommentaren aufgeführten Aspekt der Teststärke und andere ausgewählte Kritikpunkte.

### *Die Entstehungsgeschichte (mit persönlichen Anmerkungen)*

Zum Zeitpunkt des Experiments im Sommer 2007 war ich – nach Abschluss meines Diplomstudiums Psychologie im Jahr 2006 – ungefähr seit einem Jahr als wissenschaftlicher Mitarbeiter in der Abteilung für Allgemeine Psychologie und Methodenlehre an der Universität Trier tätig. Ich betreute damals zwei studentische Kleingruppen im Rahmen eines Experimentalpraktikums. Die Studierenden befanden sich im Grundstudium Psychologie und hatten im vorangegangenen Wintersemester bei mir den ersten Teil des Experimentalpraktikums absolviert. Dort hatten wir verschiedene Gestaltungsempfehlungen zu multimedialen und interaktiven, elektronischen Lernumgebungen experimentell überprüft – ein Forschungsgebiet, dem ich bis heute treu geblieben bin. In Trier war es damals üblich, dass der Dozent das Thema des Experimentalpraktikums im Wintersemester festlegte, während es im Sommersemester meist von den Studierenden bestimmt wurde. Offen gestanden bin ich mir nicht mehr sicher, wer eine parapsychologische Untersuchung bei einem gemeinsamen Abschlusstreffen der beiden Kleingruppen vorschlug. Ich kann mich nur noch daran erinnern, dass der Vorschlag schnell breite Unterstützung fand. Während eine Kleingruppe ein Würfelexperiment durchführte, widmete sich die andere Gruppe dem Kartenexperiment. Im Anschluss versuchte ich gemeinsam mit

den Studierenden, die beiden Arbeiten zu veröffentlichen. Während das Würfelexperiment, in dem keinerlei signifikante Ergebnisse festgestellt werden konnten, in einer parapsychologischen Fachzeitschrift von den Gutachtern abgelehnt wurde, konnte das Kartenexperiment veröffentlicht werden.

Seit diesen beiden Studien im Sommer 2007 habe ich keine weiteren parapsychologischen Studien durchgeführt und die Versuchsperson mit den extrem hohen Trefferzahlen im Kartenexperiment nicht weiter untersucht. Stattdessen habe ich mich in den letzten Jahren intensiv mit der Gestaltung elektronischer Lernumgebungen befasst und zu diesem Thema promoviert. Zu verwandten Forschungsthemen habe ich im Anschluss meine kumulative Habilitation erstellt, die ich Anfang 2011 offiziell eingereicht habe. Die Effektgrößen liegen in diesem Forschungsgebiet tatsächlich häufig über 0.2 (Cohen's *d*). Ob ich jemals wieder eine parapsychologische Fragestellung untersuchen werde, kann ich nicht vorhersagen. Wer weiß, vielleicht eines Tages erneut im Rahmen eines Experimentalpraktikums?

## *Teststärke*

Bevor ich inhaltlich auf die Kommentare zur Teststärke eingehe, möchte ich eine persönliche Anmerkung vorwegschicken. Ich habe in meiner nunmehr fünfjährigen Arbeit als wissenschaftlicher Mitarbeiter im Fach Psychologie diverse Methodenseminare sowie eine Methodenvorlesung zum Thema „Multivariate Verfahren" gehalten. In vielen dieser Veranstaltungen bin ich detailliert auf das Thema Stichprobenumfangsplanung und Teststärkeberechnung eingegangen. Zudem habe ich ein kleines PC-Programm zur Teststärkenbestimmung erstellt. Ich glaube daher, dass ich mich mit diesem Thema vergleichsweise gut auskenne, obgleich ich selbstverständlich kein Experte auf diesem Gebiet bin.

In zahlreichen Lehrbüchern (z.B. Bortz & Döring, 2006) wird darauf verwiesen, dass die Teststärke durch den Stichprobenumfang, das Signifikanzniveau und eine angenommene Effektgröße bestimmt werden kann. Diese drei Variablen sind für die Teststärkenbestimmung zwar erforderlich, aber keineswegs ausreichend. Ich habe in Abb. 1 zu skizzieren versucht, welche weiteren Faktoren die Teststärke unter anderem beeinflussen.[1] Die meisten dieser Faktoren wirken sich durch eine Veränderung der zentralen und/oder nonzentralen Verteilung (in Abb. 1 als „Verteilungen" zusammengefasst) auf die Teststärke aus. Neben den in Abb. 1 aufgeführten Variablen beeinflussen vermutlich noch zahlreiche weitere Variablen die Teststärke.

1 Auf Anfrage (GuenterDanielRey@web.de) sende ich gerne jedem Interessenten eine detaillierte Beschreibung zu, wie die in Abb. 1 aufgeführten Variablen die Teststärke m.E. beeinflussen.

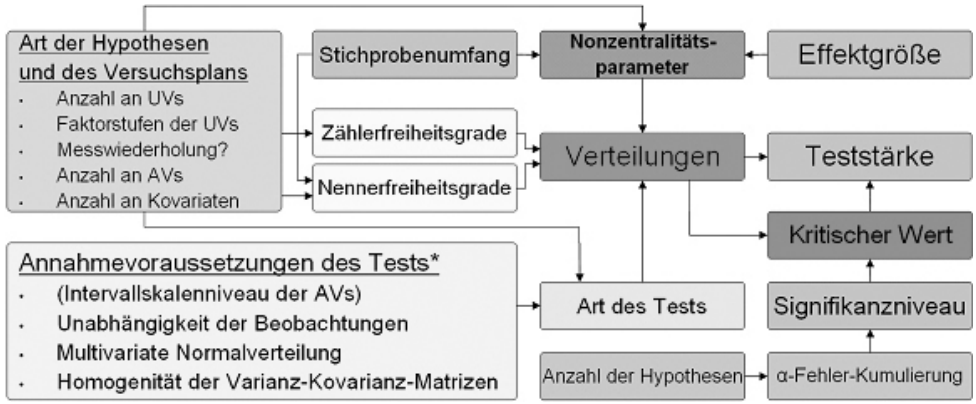


**Abb. 1**: Darstellung der Einflussfaktoren, welche unter anderem die Teststärke der Untersuchung beeinflussen (* Hier beispielhaft: Annahmevoraussetzungen der multivariaten Varianzanalyse).

Die Teststärke ist folglich im Kontext zahlreicher Variablen zu betrachten. Vor diesem Hintergrund sollte man meines Erachtens zu einem nicht signifikanten Ergebnis nicht die pauschale Behauptung aufstellen, dass die Teststärke ausgereicht bzw. nicht ausgereicht habe. Stattdessen ist unter anderem die Wahl des statistischen Tests, das Signifikanzniveau und vor allem die angenommene Effektgröße anzugeben.[2] Erst im Rahmen dieser Variablen kann man die angegebene Höhe der Teststärke einordnen und beispielsweise behaupten, dass die Teststärke bei Verwendung eines bestimmten statistischen Tests für ein bestimmtes Signifikanzniveau und einen bestimmten angenommenen Effekt hinreichend bzw. nicht hinreichend war. Insofern ist der vorgebrachte Einwand selbstverständlich richtig, dass die Teststärke in der vorliegenden Studie für eine kleinere, angenommene Effektgröße möglicherweise nicht mehr ausreicht, um die Nullhypothese begründet anzunehmen. Allerdings trifft dieser Einwand für sämtliche Teststärken zu, die jemals in empirischen Untersuchungen berichtet worden sind.

Darüber hinaus mag es durchaus zutreffen, dass der angenommene empirische Effekt in der vorliegenden Arbeit bei der Berechnung der Teststärke deutlich zu hoch angesetzt wurde. Jedoch besteht hier ein grundlegendes Problem. Die für die Berechnung der Teststärke bzw. des Stichprobenumfangs essentiell wichtige Effektgröße ist meines Erachtens generell a priori kaum zuverlässig abzuschätzen. Dies trifft vor allem, aber nicht ausschließlich auf neue Forschungsfragen zu. Selbst wenn man in einem Forschungsgebiet auf Metaanalysen mit Effektgrößenangaben zurückgreifen kann, besteht das Problem, dass die dort berichteten Effektgrößen sich bereits durch geringfügige Änderungen (zum Beispiel eine andere Stich-

2 Außerdem sollte angegeben werden, auf welche Effektgröße man sich bezieht (z.B. Cohen's *d*). Da es diverse Effektgrößen gibt, sind Aussagen wie „eine Effektgröße von 0.2" uneindeutig.

probenzusammensetzung oder andere Operationalisierungen) erheblich von den Effektgrößen in der neuen Studie unterscheiden können. Insofern weiß man im Vorfeld einer Studie in aller Regel nicht genau, wie stark der Effekt ausfallen wird, was die Stichprobenumfangsplanung in ganz erheblichem Maße erschwert. Um diesem Problem zu begegnen, schlage ich statt der herkömmlichen Stichprobenumfangsplanung im Vorfeld einer Studie einen anderen Ansatz zur Ermittlung des Stichprobenumfangs vor, den ich als „*intermediate Stichprobenumfangsplanung*" bezeichne. Nach diesem Ansatz wird der erforderliche Stichprobenumfang auf Basis des empirisch ermittelten Effekts *während* der Durchführung der Studie geschätzt. Im ersten Schritt erfolgt zunächst eine fortlaufende Ermittlung der Effektgröße während der Untersuchung auf Basis der bis dahin erhobenen Versuchspersonen. Im zweiten Schritt wird die Stichprobenumfangsplanung mit Hilfe des empirisch ermittelten Effekts *einmalig* durchgeführt, sobald sich die Effektgröße stabilisiert hat.[3] Vielleicht kann man diesen Ansatz weiter ausarbeiten und nicht nur in der Parapsychologie als Alternative zur herkömmlichen Stichprobenumfangsplanung einsetzen.

Neben diesen grundsätzlichen Anmerkungen möchte ich noch auf zwei ausgewählte Kritikpunkte zur Teststärke näher eingehen. Zunächst möchte ich zur Teststärkenbestimmung von Herrn Nelson folgendes anmerken: Bei einem One-Sample *t*-test mit $n = 96$, einer Effektgröße von 0.2 (Cohen's *d*) und einem Signifikanzniveau von 0.05 komme ich bei einseitiger Testung mit dem PC-Programm GPower 3.1.2 auf eine Teststärke von 61.8% (bei zweiseitiger Testung auf 49.2%) statt der von Herrn Nelson berichteten Teststärke von 62.4%. Damit will ich keineswegs behaupten, dass die Teststärke von Herrn Nelson falsch berechnet wurde. Mich würde lediglich interessieren, worauf diese geringen Unterschiede zurückzuführen sind. Der zweite Kritikpunkt, auf den ich näher eingehen möchte, stammt ebenfalls von Herrn Nelson (sein Punkt 16): Meines Erachtens erfolgte in der vorliegenden Studie keine klassische Fehlinterpretation des *p*-Werts. Die Nullhypothese wurde nicht allein auf Grundlage der ermittelten *p*-Werte angenommen, sondern unter Berücksichtigung der ermittelten Teststärke und zwar nur für den im Artikel beschriebenen Signifikanztest sowie die im Artikel aufgeführte, angenommene Effektgröße und das dort genannte Signifikanzniveau. Für andere Kennwerte (z.B. für eine geringere Effektgröße) oder andere Signifikanztests treffe ich keinerlei Aussagen. Zu diesem Punkt darf ich auch auf meine allgemeinen Ausführungen zur Teststärke (siehe oben) sowie auf den treffenden Kommentar von Herrn Ambach zur Annahme der Nullhypothese verweisen.

### *Weitere Kritikpunkte*

In den Kommentaren zum Artikel wurden diverse weitere Aspekte kritisch beanstandet. Viele Kritikpunkte mögen sicherlich zutreffen. Auf einige ausgewählte Punkte möchte ich hier näher eingehen:

3 Auf Anfrage sende ich gerne jedem Interessenten eine Beschreibung zu, wie man mittels Bootstrap ermitteln könnte, ob sich die Effektgröße stabilisiert hat.

1. Die Verwendung eines selbst konstruierten Fragebogens zur Erfassung paranormaler Einstellungen halte ich rückblickend für einen Fehler. Stattdessen hätten wir auf einen bereits bewährten und validierten Fragebogen zurückgreifen sollen. Dieser Fehler ist meines Erachtens wie folgt entstanden. Wie oben bereits aufgeführt, erforsche ich Fragestellungen zu multimedialen und interaktiven Lernumgebungen. In diesem Forschungsgebiet werden sowohl die Lernmaterialien als auch die Lernleistungstests für die jeweilige Studie in aller Regel gesondert konzipiert. Dieser Umstand führte wohl dazu, dass auch der Fragebogen zur Erfassung paranormaler Einstellungen eigenständig konstruiert wurde. Ähnlich verhält es sich mit den selbst erstellten Zener-Karten.
2. Nachträglich möchte ich Herrn Ambach ausdrücklich zustimmen, dass es „nicht erlaubt ist, die Wertigkeit und die Einschränkungen einer Studie (und damit die Tragweite möglicher Ergebnisse) zeitlich nach Kenntnisnahme der Ergebnisse festzulegen". Der von Herrn Ambach angesprochene Argumentationsfehler ist der unreflektierten und vermutlich missverstandenen Übernahme eines Kommentars eines anonymen Gutachters geschuldet.
3. Maßnahmen zur Entdeckung und zur Korrektur von Fehlern bei der Übertragung der Werte in die Excel-Tabellen (siehe Punkt 13 von Herrn Nelson) wurden vorgenommen. Derartige Maßnahmen werden in dem Forschungsgebiet, in dem ich tätig bin, in aller Regel nicht gesondert benannt, sondern als selbstverständlich betrachtet. Daher wurden die getroffenen Maßnahmen auch in dieser Studie nicht explizit aufgeführt.
4. Dem Kritikpunkt von Herrn Ertel, dass es dem Artikel bisweilen an begrifflicher Klarheit mangelt, muss ich leider zustimmen. Ich halte den entsprechenden Kommentar von Herrn Ertel für wichtig und angemessen und darf mich für diesen Hinweis bedanken. Zurückweisen möchte ich allerdings seine Vermutung, dass wir die Untersuchung voreingenommen durchgeführt haben.
5. Herr Ertel beanstandet des Weiteren, dass Psi-Missing ursprünglich nicht berücksichtigt und nachträglich nicht statistisch überprüft wurde. Im Artikel wurde bereits darauf hingewiesen, dass Psi-Missing unberücksichtigt blieb, weil wir im Vorfeld davon ausgegangen sind, dass die Anzahl an Versuchspersonen für eine zweiseitige Testung und die im Artikel aufgeführten weiteren Variablen (angenommener Effekt, Signifikanzniveau usw.) nicht für eine akzeptable Teststärke ausgereicht hätte. Auch aus heutiger Sicht halte ich die abschließend dann doch erreichte Versuchspersonenzahl von 96 für ein experimentalpsychologisches Praktikum für außerordentlich hoch; im Vorfeld war damit jedoch nicht zu rechnen gewesen. Bei dieser Gelegenheit darf ich mich nochmals für die überaus engagierte Arbeit meiner Mitautorinnen bedanken. Eine nachträgliche inferenzstatistische Überprüfung von Psi-Missing wäre zwar möglich gewesen, aber meines Erachtens sollte man nur a priori aufgestellte Hypothesen inferenzstatistisch testen. Gleiches gilt für die von Herrn Ertel aufgeführten Korrelationen zwischen den

Trefferzahlen unter den drei Versuchsbedingungen. Das von Herrn Ertel erbetene „entlastende Versuchsprotokoll im Original mit Zahlen im Detail" kann ich gerne in Form der Rohdaten und Auswertungen vorlegen, falls der Kollege die Daten und Analysen im Detail überprüfen möchte. Bereits mit Hilfe der Angaben im Artikel kann jedoch abgeschätzt werden, dass Psi-Missing in dem Experiment wohl keine Rolle gespielt hat. Im Artikel steht explizit, dass unter der Telepathie-Bedingung 493 Treffer erzielt wurden, beim Hellsehen hingegen 490 Treffer. Der Erwartungswert lag bei 480 Treffern. Demnach kann für diese beiden Bedingungen Psi-Missing bereits deskriptivstatistisch ausgeschlossen werden. Lediglich im Präkognitions-Versuch wurden mit 456 weniger als die erwarteten 480 Treffer bei 2400 Versuchen beobachtet. Dieser Wert unterscheidet sich aber nicht signifikant vom Erwartungswert. Über die Binomialverteilung kann die Signifikanzprüfung mit diesen Daten nachvollzogen werden.

6. Im Zusammenhang mit der fehlenden inferenzstatistischen Überprüfung von Psi-Missing vermutet Herr Ertel, dass ich mit ausbleibenden ASW-Effekten in meinem akademischen Umfeld „besser leben" könne als mit signifikanten ASW-Effekten. Die Vermutung, dass ich signifikante Psi-Missing-Effekte aus diesem Grund zurückhalte, möchte ich zurückweisen. Wie bereits erwähnt, kann sich jeder selbst davon überzeugen, dass derartige Effekte nicht auftraten. Zudem erscheint mir die Argumentation von Herrn Ertel an dieser Stelle unschlüssig. Hätte ich Angst vor Konsequenzen bezüglich der Veröffentlichung parapsychologischer Ergebnisse, hätte ich eine solche Untersuchung gar nicht erst begonnen, geschweige denn publiziert. Wie viele meiner Kolleginnen und Kollegen aus der Psychologie heutzutage parapsychologische Forschung skeptisch betrachten und wie viele dieser Forschung aufgeschlossen gegenüberstehen, vermag ich nicht zu beurteilen. In meinem unmittelbaren Arbeitsumfeld habe ich die diesbezügliche Haltung bisher überwiegend als neutral bis aufgeschlossen empfunden. Ein befreundeter Kollege, der damals ebenfalls in der Abteilung für Allgemeine Psychologie und Methodenlehre in Trier tätig war, hat im Übrigen zeitgleich zu meinen Experimenten eine parapsychologische Fragestellung zum Remote-Staring-Effekt untersucht.

Abschließend darf ich mich abermals für alle Kommentare bedanken. Ich hoffe, dass ich mit meinem abschließenden Kommentar noch einige nützliche Zusatzinformationen zur Studie zur Verfügung stellen und einige Unklarheiten beseitigen konnte.

## Literatur

Bortz, J., & Döring, N. ([4]2006). *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler*. Berlin: Springer.