

Geistheilertest der GWUP Wien, Januar und Februar 2005

Dr. Elisabeth und Dr. Ivo Ponocny

Die Redaktion der Sendereihe „HelpTV“ des österreichischen öffentlich-rechtlichen Rundfunks ORF2 ist an die GWUP-Wien herangetreten mit der Bitte, einen Beitrag zum Geistheilen mit ihnen zu gestalten. Man entschied sich für einen Test ähnlich dem, den die Schülerin Emily Rosa mit amerikanischen Wissenschaftlern durchgeführt hatte. Dort mussten die KandidatInnen durch einen Paravent greifen und jeweils entscheiden, ob eine Person auf der anderen Seite deren Hand über die linke oder die rechte Hand des(r) KandidatIn hielt. Die Resultate der amerikanischen Tests waren durchwegs negativ und konnten keine Trefferraten zeigen, die über die Ratewahrscheinlichkeit hinausgingen.

Im Oktober 2004 wurde in „help-TV“ zur Teilnahme am Versuch aufgerufen, wobei Personen/Geistheiler gesucht wurden, welche „menschliche Energieströme aus kurzem Abstand wahrnehmen können“. Es meldeten sich 39 KandidatInnen, deren Adressen von der „helpTV“-Redaktion an die GWUP übergeben wurden. Die KandidatInnen wurden telefonisch genau über die Bedingungen des geplanten Tests bzw. die zu erreichende Quote informiert und ihnen wurde geraten, einen entsprechenden privaten Vorversuch durchzuführen. Letztendlich unterzogen sich 21 von ihnen der Testung, welche an 5 verschiedenen Tagen zwischen 17. Jänner und 10. Februar 2005 stattfand.

Versuchsplan

Die Grundidee dieses Tests besteht darin, dass jemand, der bestimmte Eigenschaften in einem sog. Energiefeld eines(r) KlientIn wahrnehmen und auch beeinflussen will, zumindest spüren können sollte, ob sich dort überhaupt ein Energie ausstrahlender Körper befindet. Eigene Vorversuche ließen es als ratsam erscheinen, nicht nur auf der Seite des(r) KandidatIn ein Tuch vorzusehen, um die Löcher für die Arme blickdicht zu halten, sondern auch auf der Seite der Zielperson, da sonst durch Wärme oder Luftzug triviale physikalische Hinweisreize zu befürchten gewesen wären. Mit diesem zweiten Tuch hielten wir es für zulässig, einen Abstand zwischen den Händen von KandidatIn und Zielperson von nur ca. 10 Zentimetern vorzusehen. Unser Test verlangte eine Entscheidung, ob beide Hände der Zielperson über die des(r) Kandidaten gelegt wurden, oder ob dies gerade nicht der Fall war.

Das Testdesign wurde allen KandidatInnen telefonisch mitgeteilt und viele zeigten sich sehr zuversichtlich, diesen Test zu bestehen. Einige äußerten aber auch Bedenken, weil der Versuch sich von ihrer üblichen Tätigkeit deutlich unterschied. Dennoch kannten alle der teilnehmenden Personen das Design, und bestätigten das auch schriftlich. Etliche befanden es durchaus als ihren Fähigkeiten entsprechend wenn nicht sogar als ihre „Spezialität“. Dieser Punkt sei gleich an dieser Stelle erwähnt, da wir nachher mit der Kritik konfrontiert wurden, der Versuch sei völlig irrelevant und zeige nur unser Unverständnis auf. Zumindest vor der Bekanntgabe der Ergebnisse waren jedenfalls alle KandidatInnen selbst mit diesem Design durchaus einverstanden.

Von jedem(r) KandidatIn wurden im ersten Durchgang 20 Ja/Nein-Entscheidungen verlangt. Unser Angebot, die Testreihe früher zu beenden, wurde von niemandem in Anspruch genommen. Die tatsächliche Abfolge von Hände hinstrecken bzw. nicht hinstrecken wurde zufällig mithilfe von zwanzig roten bzw. gelben Jetons (mit Zurücklegen) gezogen (gelb für „ja“, rot für „nein“) und folgte somit genau einer Binominalverteilung mit zwanzig Versuchen und Erfolgswahrscheinlichkeit 50%. Individuelle Signifikanz eines(r) KandidatIn auf dem 5%-Niveau resultiert dabei bei

einer Trefferrate von mindestens 15 von 20 bzw. einer Trefferquote von mindestens 75% (sog. einseitige Testung). Mehr als 20 Durchgänge erschienen sowohl was die Zumutbarkeit für die KandidatInnen betraf als auch in Hinblick auf die für die Testung verfügbare Zeit als unrealistisch. Das 5%-Niveau entspricht dabei den üblichen wissenschaftlichen Standards, was hier bedeutet, dass im Falle reinen Ratens im Schnitt nur jeder 20. Durchgang irrtümlich als erfolgreich bewertet werden darf. Diese Sicherheit gilt aber nur für einen einzelnen Durchgang und nicht für die Gesamtheit aller Durchgänge; bei einer erwarteten Anzahl von 30 KandidatInnen wäre es sogar recht wahrscheinlich, dass mindestens einer der KandidatInnen auch unter Ratebedingungen den ersten Durchgang besteht. Aus diesem Grunde wurde festgelegt, dass im Falle eines positiv absolvierten ersten Durchgangs auch noch ein zweiter Durchgang ebenfalls mit einer Trefferrate von mindestens 75% zu absolvieren war, erst dann konnte der Test als erfolgreich bewertet werden.¹ Bei der Testung kam es in genau einem Fall zu einer Wiederholung.

Die Erfolgsaussichten aus dem Blickwinkel der KandidatInnen stellen sich dabei folgendermaßen dar: Bei 90% Trefferwahrscheinlichkeit, was einige KandidatInnen als ihre Quote vermutet hatten, besteht eine Wahrscheinlichkeit von über 99,9%, den ersten Durchgang zu bestehen (die sogenannte Macht oder Power), und ebenfalls eine über 99,9%-ige Wahrscheinlichkeit, beide Durchgänge zu bestehen. Der für klinische Studien übliche Standard von 80% Power für das Bestehen beider Durchgänge würde bei einer zugrundeliegenden Trefferwahrscheinlichkeit von 83% erreicht.

Versuchsablauf

Die Testung verlief folgendermaßen:

Vorbereitung (Team 1 = 2 Personen)

Aus dem Korb mit Jetons in zwei Farben werden von einer Person nacheinander 20 Jetons gezogen und jeweils gleich wieder hineingegeben.

Das Ergebnis jeden Zugs – Gelb für ja, Rot für nein - wird von einer weiteren Person in das Protokoll A eingetragen.

Dieses Protokoll wird der 2. Person des Teams 2) übergeben.

Test (Team 2 = 3 Personen)

Die Person 1 begleitet den Kandidaten zum Tisch, an dem beide Platz nehmen. Vor dem Kandidaten steht ein Paravent, der die Sicht auf die andere Seite verstellt.

Der Kandidat streckt seine Unterarme mit den Handflächen nach oben durch den Paravent, und über seine Handflächen wird ein Tuch gelegt. Seine Aufgabe ist es, zu erspüren, ob der Testpartner seine Handflächen im Abstand von 10 cm darüber hält - oder nicht.

Die Person 2. und der Testpartner nehmen am Tisch dem Kandidaten gegenüber Platz. Das Protokoll A wird dem Testpartner vorgelegt. Dieser geht nun nach den Eintragungen vor und streckt seine Handflächen über die des Kandidaten – oder nicht. Mit dem Wort „jetzt“ startet die Person 2. jeden Versuch, zur gleichen Zeit löst sie die Stoppuhr aus, die nach 20 Sekunden einen Klingelton abgibt. Der Kandidat hat noch weitere 10 Sekunden Zeit, sein Urteil – „ja, ich spüre die Handflächen“ / „nein, ich spüre nichts“ – abzugeben. Die Aussagen des Kandidaten werden von seiner Begleitperson in das Protokoll B eingetragen.

Als weitere Person ist der Versuchsleiter anwesend, um alle Abläufe zu überwachen.

¹ Bei $n=30$ Personen soll – nach allgemeingültigen wissenschaftlichen Standards – nur mit Wahrscheinlichkeit 5% mindestens eine Person unter Ratebedingungen 2 Durchgänge erfolgreich absolvieren können. Eine sichere Entscheidung mit 95% bei 30 Personen erfordert daher bei 2 Versuchsdurchgängen eine Irrtumswahrscheinlichkeit von $\sqrt[30]{,95} = (1 - \alpha) \Rightarrow \alpha = 0,0017$, daher ist bei einem Versuchsdurchgang eine maximale

Irrtumswahrscheinlichkeit von $\alpha = \sqrt{0,0017} = 0,0413$ zu fordern. Dies ist bei einer geforderten Trefferzahl von 15 erfüllt.

Die KandidatInnen hatten jederzeit die Möglichkeit, Versuchsdurchgänge zu wiederholen; da sie zwischendurch keine Rückmeldung über ihre bisherigen Trefferquoten bekamen.

Auswertung (Team 3=2 Personen)

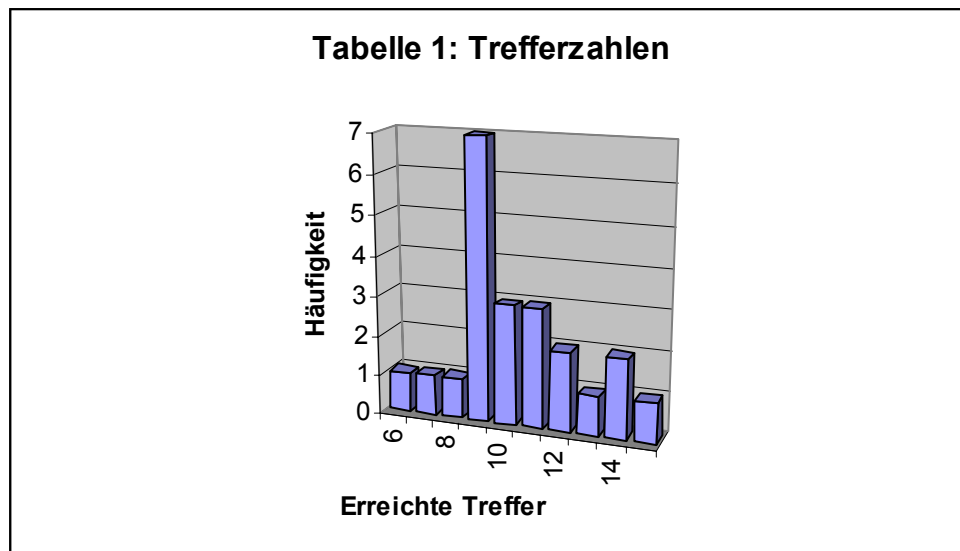
Die Protokolle A und B werden verglichen, die Treffer gezählt und mit Farbstift in Protokoll C eingetragen.

Testwiederholung: Erreicht der Kandidat bei 20 Versuchen 15 Treffer (=75 % Erfolg) wird der Test wiederholt. Der Ablauf entspricht dem oben beschriebenen.

Eine Kandidatin ließ sich auch auf einen **Wiedererkennungstest** ein, welcher folgendermaßen ablief: Die Kandidatin verbrachte mit 4 weiblichen Zielpersonen ein paar Minuten jeweils allein, um sich auf diese einzustimmen. Für jede Person wählte die Kandidatin nicht den richtigen Namen „nicknames“, z.B. das Wort „Wald“. Dies senkte zusätzlich das Risiko beträchtlich, die Personen, die sie eben erst kennen gelernt hatte, zu verwechseln.) Die Zielpersonen zogen sich daraufhin in ein Zimmer zurück, welches durch einen blickdichten Vorhang im Türstock abgedeckt war. Gemäß eines Zufallsalgorithmus (Zufallszahlengenerator) stellte sich jeweils eine der Zielpersonen knapp hinter den Vorhang, und die Kandidatin versuchte, sie richtig zu benennen. Die Kandidatin erriet bei 20 Versuchen 5 Mal die Zielperson, das entspricht genau dem Erwartungswert von $20/4=5$ und somit auch genau der Ratewahrscheinlichkeit. Dieser Versuch wurde aufgrund des abweichenden Designs gesondert behandelt und nicht gemeinsam mit den anderen Tests ausgewertet.

Ergebnisse

Tabelle 1 gibt eine Übersicht über die von den einzelnen KandidatInnen erreichten Trefferzahlen. Der offizielle Versuch umfasste pro KandidatIn einen Durchgang. Nur für den einen Kandidaten, welcher 15 Treffer erreichte und damit das Signifikanzkriterium beim ersten Durchgang erfüllte, gab es noch einen zweiten offiziellen Durchgang. (Bei den Kandidatinnen mit 14 Treffern wurden ebenfalls sicherheitshalber weitere Durchgänge ausgeführt, aber als außer Konkurrenz gewertet und führten auch zu keinen Signifikanzen mehr.) Der Kandidat mit 15 Treffern im ersten Versuch erreichte 7 Treffer im zweiten. Diesen Versuch annullierte er aber selber noch vor Bekanntgabe seiner Trefferquote. In dem von ihm selbst autorisierten Versuch an einem weiteren Tag erreichte er 12 Treffer und verfehlte somit das Kriterium für statistische Signifikanz. Somit konnte kein einziger der 21 Kandidaten ein signifikantes Resultat erzielen.



Auch in Gesamtbilanz konnte kein in irgendeiner Weise auffälliges Ergebnis erzielt werden: Von den absolvierten 22 mal 20 = 440 Einzeldurchgängen waren 227 Treffer und 213 falsch, das entspricht einer Trefferquote von 51%, was im Bereich der Zufallsschwankung um den unter Ratebedingungen erwarteten Wert von 220 liegt (50%). Fast jedes vierte Mal würde auch unter reinen Ratebedingungen ein Ergebnis mit ebenso vielen oder noch mehr Treffern resultieren (p -Wert: 0.23). Die Quoten der KandidatInnen unterschieden sich daher statistisch in keiner Weise von einer Zufallsreihe mit Ratewahrscheinlichkeit 0.5. Mit 95%-prozentiger Sicherheit liegt die tatsächlich zugrundeliegende Quote zwischen 45% und 55%, ist also mit der nötigen (und den KandidatInnen besprochenen) Quote von 75% keinesfalls verträglich.

Auch weitere Auswertungen zeigten keinerlei Auffälligkeiten: So weichen auch die Verteilungen der korrekten und nicht korrekten Ja-Antworten sowie der korrekten und nicht korrekten Nein-Antworten nicht signifikant von den Zufallserwartungen ab. Insbesondere beging jede(r) KandidatIn Fehler in beide Richtungen, d.h. in jedem Durchgang wurden sowohl vorhandene Hände nicht erfüllt bzw. wurde auch in jedem Durchgang irrtümlich das Vorhandensein von Händen angegeben. Nur eine Kandidatin gab – trotz mehrfachen Nachfragens seitens der Versuchsleitung – jedes Mal die Antwort „Ja“.

Diskussion

Besondere Fähigkeiten der KandidatInnen konnten durch die Testserie in keinsten Weise nachgewiesen werden. Das Design war so gewählt, dass ein positiver Ausgang des Experiments ein klares Indiz für das Vorhandensein von Phänomenen geliefert, welche sich im Rahmen der etablierten Naturgesetze kaum erklären ließen. Der negative Ausgang des Experiments lässt aber zwei Erklärungsmöglichkeiten zu: Es kann daran liegen, dass besondere Fähigkeiten der KandidatInnen gar nicht existieren, oder der Fehler liegt am Experiment selbst. Insbesondere könnte die Methode der Messung ungeeignet sein, was etwa schon aufgrund technischer Details wie etwa dem über den Händen liegenden Tuch der Fall sein könnte. Tatsächlich hat einer der Kandidaten das Baumwolltuch kritisiert, konnte aber in einer Replikation des Versuchs mit einem von ihm akzeptierten Seidentuch auch kein besseres Resultat liefern. Eine andere Kandidatin wiederum wählte explizit das Baumwolltuch, weil sich das Seidentuch nicht eignen würde. Es könnte aber prinzipiell die vorgenommene Art der Testung einfach an der Natur der Fähigkeiten der KandidatInnen vorbeigehen. Tatsächlich wurde der Unterschied zwischen der Testsituation und der üblichen Tätigkeit der KandidatInnen oft angesprochen. Eine

systematische Beobachtung bzw. Dokumentation, verbunden mit einer entsprechenden Evaluierung, wäre tatsächlich wünschenswert, aber natürlich mit riesigem Aufwand verbunden und bedürfte wohl eines zahlungskräftigen Financiers. Weitere Störfaktoren sind noch denkbar, z.B. Nervosität oder die Präsenz zu vieler Personen bei der Testung.

Dennoch wirft der Testaufgang die Frage auf, warum eine Aufgabe nicht gelöst werden kann, die für den Außenstehenden doch einfacher wirkt als eine Diagnose oder gar eine Heilung. Auch den KandidatInnen selbst war offenbar nicht selbstverständlich, dass sie die Aufgabe nicht lösen würden. Einige waren vor der Testung sogar ausgesprochen zuversichtlich. Kritik dahingehend, dass nur mit der Materie nicht vertraute Personen so ein Design für passend halten könnten, kann also entschieden entgegengetreten werden.

Solange angeblich paranormale Phänomene nur in Kontexten auftreten, die auch nicht-paranormale Erklärungsmöglichkeiten für die Beobachtungen zulassen, können die von Geistheilern behaupteten besonderen Fähigkeiten niemals nachgewiesen werden. Das hier beschriebene Experiment war ein Versuch, die paranormalen Anteile der Tätigkeit der Geistheilern zu isolieren und gezielt zu beobachten. Dieser Versuch ist einmal mehr gescheitert.