Kopf

Kon. S.5!

BERICHTE

aus dem PSYCHOLOGISCHEN INSTITUT der UNIVERSITÄT BONN

Band 14 (1988) Heft 4

Entwicklung eines Polynomial-Tests für die Ausreißer-Alternative und Anwendung auf ein kognitionspsychologisches Beispiel

Edgar Erdfelder & Joachim Funke

Inhaltsverzeichnis

Zusammenfassung

Wenn die Nullhypothese gleicher Wahrscheinlichkeiten von k disjunkten und exhaustiven Kategorien gegen die spezielle Alternativhypothese getestet werden soll, daß genau eine (unbekannte) Kategorie überfrequentiert ist, die anderen k-1 Kategorien dagegen gleichwahrscheinlich sind, so sind der Pearson-CHI^2- und der 2I-Test sowie deren exakte Varianten (Binomial- bzw. Polynomialtest) suboptimal. Es wird ein spezieller Polynomialtest zur Prüfung der "Ausreißer-Hypothese", wonach genau eine von k Kategorien überfrequentiert ist, abgeleitet und begründet. Für ausgewählte Stichprobenumfänge und $\alpha = .05$ wird gezeigt, daß der Test befriedigende Teststärkeeigenschaften (relativ zum exakten Pearson-CHI^2-Test) aufweist. Die Anwendung des Tests wird am Beispiel kognitionspsychologischer Hypothesen demonstriert, welche die Evaluation von Problemlöseleistungen durch Probanden verschiedener Altersgruppen zum Gegenstand haben.

Summary

Development of a polynomial test for the outlier-hypothesis and application to an example from cognitive psychology.

If the null hypothesis of equal probabilities for k disjunctive and exhaustive categories has to be tested against the alternative that exactly one (unknown) category is more probable than the remaining k-1 equally probable categories, the Pearson-CHI^2- and the likelihood-ratio-test as well as their exact counterparts (binomial and polynomial test) are suboptimal. Therefore, a special polynomial test is developed which tests the "outlier-hypothesis" that exactly one population proportion is greater than the remaining k-1 ones. For selected sample sizes and $\alpha = .05$ the power of the test is proved to be satisfactory (compared to the exact Pearson-CHI^2-test). The application of the test is demonstrated in the context of cognition, using the evaluation of problem solving capabilities by subjects of different age groups as an example.

Entwicklung eines Polynomial-Tests für die Ausreißer-Alternative und Anwendung auf ein kognitionspsychologisches Beispiel[1]

Edgar Erdfelder und Joachim Funke
Psychologisches Institut der Universität Bonn

1 Problem

Wenn anhand einer Stichprobe des Umfangs N die Nullhypothese (H_0) überprüft werden soll, daß allen k (disjunkten und exhaustiven) Kategorien einer qualitativen Zufallsvariablen (ZV) X gleiche Wahrscheinlichkeiten $p_1 = p_2 = \ldots = p_k$ zukommen, verwendet man bekanntlich im Falle $k = 2$ den Binomial- und im Falle $k > 2$ den Polynomialtest (vgl. Lienert 1973, p. 143 f.). Die Prüfung kann bei hinreichend großem N auch approximativ über die Pearson-Chi^2- oder die 2I-Statistik unter Bezugnahme auf die CHI^2-Verteilung als Prüfverteilung erfolgen.

Eine gemeinsame Eigenschaft aller genannten Prüfverfahren besteht darin, daß sie auf jede Abweichung von der Gleichverteilung gleichermaßen sensibel reagieren. Dies ist oft, aber nicht immer erwünscht. Wenn aufgrund theoretischer Vorüberlegungen spezielle Alternativhypothesen, d.h. bestimmte Formen der Abweichung von der Gleichverteilung abgeleitet werden können, sind "schärfere" (teststärkere) Prüfungen der H_0 denkbar, als sie mit den oben genannten Statistiken erreicht werden können. Dies ist zum Beispiel dann der Fall, wenn die spezielle Alternativhypothese von Interesse ist, daß genau eine der k Kategorien überfrequentiert ist. Kann eine derartige Nonzentralitätsstruktur (mit genau einem "Ausreißer" unter den Kategoriewahrscheinlichkeiten) abgeleitet werden, so sind die oben genannten Verfahren der Überprüfung der Gleichverteilungshypothese nicht optimal.

---

[1] Die Autoren danken Herrn Prof. Dr. J. Bredenkamp, Frau Dr. S. Mecklenbräuker und Herrn Dr. R. Steyer für die kritische Durchsicht der Erstfassung des Manuskripts. Das Manuskript der vorliegenden Arbeit wurde im Jahr 1983 abgeschlossen. Die Arbeit war nach dem Umzug der Autoren von Trier nach Bonn zunächst nicht auffindbar, wurde aber dann wiedergefunden und trotz der inzwischen verstrichenen Zeit für mitteilenswert befunden.

Wir werden im folgenden zunächst ein geeignetes Prüfverfahren für die "Ausreißer-Alternative" vorstellen, wobei wir uns auf den Fall des exakten Tests beschränken. Die Teststärke des Verfahrens wird mit der des exakten Pearson-CHI^2-Tests verglichen. Die Brauchbarkeit des Verfahrens im substanzwissenschaftlichen Kontext wird anschließend anhand einer exemplarischen Anwendung im Bereich der Denkpsychologie demonstriert. Ein Anhang mit detaillierten Hinweisen zur Durchführung des "Tests für die Ausreißer-Alternative" schließt die Arbeit ab.

2 Ableitung einer Teststatistik für die "Ausreißer-Alternative"

Seien $p_1, p_2, \ldots, p_i, \ldots, p_k$ die Wahrscheinlichkeiten für das Auftreten von k Kategorien in einer Population (mit $p_1 + p_2 + \ldots + p_k = 1$) und ergeben sich nach N-maliger Zufallsentnahme aus dieser Population genau n_1 Elemente aus der Kategorie 1, n_2 Elemente aus der Kategorie 2 usw. (mit $n_1 + n_2 + \ldots + n_k = N$), so beträgt die Wahrscheinlichkeit für dieses Ereignis gemäß der Polynomialwahrscheinlichkeitsfunktion (vgl. z.B. Lienert 1973, p. 156 f.):

$$p(n_1,n_2,\ldots,n_k|p_1,p_2,\ldots,p_k) = \frac{n!}{n_1!\, n_2! \ldots n_k!}\, p_1^{n_1}\, p_2^{n_2} \cdots p_k^{n_k} \qquad (1)$$

Nehmen wir nun an, es soll die Gleichverteilungs-Nullhypothese

$$H_0: \quad p_1 = p_2 = \ldots = p_i = \ldots = p_k = 1/k \qquad (2)$$

gegen die Ausreißer-Alternativhypothese

$$H_1: \begin{cases} p_1 = 1/k + E, \text{ und} \\ p_2 = \ldots = p_i = \ldots p_k = 1/k - E/(1-k) \end{cases} \qquad (3)$$

mit $0 < E < (1-1/k)$ getestet werden. H_1 behauptet eine Gleichverteilung der Kategorien 2, 3, ..., k und eine Überfrequentierung der Kategorie 1; die Konstante E ist dabei ein Maß für die Stärke des Ausreißer-Effekts in der Population. Wenn wir nun nach einer vernünftigen Prüfstatistik für das gegebene Testproblem suchen, d.h. nach einem Stichprobenkennwert, der in optimaler Weise zwischen H_1 und H_0 zu diskriminieren gestattet, so ist zu fordern, daß die Größe des Likelihood-Quotienten:

$$\underline{LQ} = \frac{p(n_1,n_2,\ldots,n_k|H_1)}{p(n_1,n_2,\ldots,n_k|H_0)} \qquad (4)$$

bzw. des entsprechenden Log-Likelihood-Quotienten:

$$\underline{LLQ} = \log \frac{p(n_1,n_2,\ldots,n_k|H_1)}{p(n_1,n_2,\ldots,n_k|H_0)} \qquad (5)$$

eine Funktion dieses Kennwerts ist. Wenn $\underline{LQ}$ bzw. $\underline{LLQ}$ "groß" ist, liegt insgesamt Evidenz zugunsten von H_1 vor; sind die Quotienten dagegen "klein", so liegt insgesamt Evidenz zugunsten von H_0 vor. Wir werden demzufolge zu analysieren haben, von welcher Statistik $\underline{LQ}$ bzw. $\underline{LLQ}$ für gegebenen Stichprobenumfang $\underline{N}$, gegebenes $\underline{k}$ und gegebenen Effekt $\underline{E}$ abhängen. Setzen wir (2) und (3) in (1) ein, so resultiert:

$$p(n_1,n_2,..,n_k|H_0) = \frac{\underline{N}!}{n_1!n_2!..n_k!} (1/\underline{k})^{n_1} (1/\underline{k})^{(\underline{N}-n_1)} \qquad (6)$$

und (7):

$$p(n_1,n_2,..,n_k|H_1) = \frac{\underline{N}!}{n_1!n_2!..n_k!} (1/\underline{k} + \underline{E})^{n_1} (1/\underline{k} - \underline{E}/(\underline{k}-1))^{(\underline{N}-n_1)}$$

Mit Hilfe der Gleichungen (6) und (7) können wir nun $\underline{LLQ}$ nach Gleichung (5) wie folgt ausdrücken:

$$\underline{LLQ} = \log \left[\frac{\frac{\underline{N}!}{n_1!\, n_2!\, \ldots\, n_k!} (1/\underline{k} + \underline{E})^{n_1} (1/\underline{k} - \underline{E}/(\underline{k}-1))^{(\underline{N}-n_1)}}{\frac{\underline{N}!}{n_1!\, n_2!\, \ldots\, n_k!} (1/\underline{k})^{n_1} (1/\underline{k})^{(\underline{N}-n_1)}} \right] \qquad (8)$$

$$= n_1 * \log \frac{1/\underline{k} + \underline{E}}{1/\underline{k} - \underline{E}/(\underline{k}-1)} + \underline{N} * \log(1 - \underline{k}*\underline{E}/(\underline{k}-1)).$$

Die Ableitung (8) macht deutlich, daß n_1 - also die Stichprobenhäufigkeit für Kategorie 1 - die für die Größe von LLQ (und damit auch LQ) entscheidende Statistik ist: Für gegebenes N, k und E variiert LLQ lediglich mit n_1. Dieses Ergebnis wird kaum verwundern, da n_1 das quasi "natürliche" Stichproben-Pendant zur "Ausreißer-Proportion" ist.

Damit läßt sich nun leicht ein exakter Test für das oben dargestellte Testproblem konstruieren. Für einen erhaltenen Wert n_1^{emp} betrachte man die Menge aller möglichen Summenzerlegungen von N in die k ganzzahligen Komponenten $n_1, n_2, \ldots, n_k$ (mit $n_1 + n_2 + \ldots + n_k = N$), welche die Nebenbedingungen

$$n_1 \geq n_1^{emp} \tag{9}$$

und

$$n_j \geq 0 \quad (\text{für } 2 \leq j \leq k) \tag{10}$$

erfüllen.[2] Unter der Annahme der Gültigkeit der Nullhypothese läßt sich die Wahrscheinlichkeit für jede dieser k-komponentigen Partitionen nach (6) errechnen. Die Summe dieser Wahrscheinlichkeiten - also die sog. "Überschreitungswahrscheinlichkeit" - wird nun mit einem (zuvor festgelegten) Wert α verglichen. Ist die Überschreitungswahrscheinlichkeit kleiner oder gleich α, so wird H_0 abgelehnt, andernfalls wird H_0 beibehalten. Eine Ablehnung von H_0 impliziert allerdings noch nicht die Annahme von H_1, da Parameterkonstellationen denkbar sind, die keiner der beiden Hypothesen (2) und (3) entsprechen. Erst wenn man die zusätzliche Annahme macht, daß $p_2 = p_3 = \ldots = p_k$ (hierin unterscheiden sich H_0 und H_1 nicht), ist die Annahme der Alternativhypothese logisch äquivalent mit der Negation von H_0. Unter dieser Prämisse sind die Hypothesen des vorgestellten Tests aber mit denen des klassischen Binomialtests identisch:

$$H_0: p_1 = 1/k; \; H_1: p_1 > 1/k \tag{11}$$

---

[2] Algorithmisch ist die Generierung der Partitionsmenge nicht ganz einfach. Späth (1978, p. 62ff.) hat jedoch einige FORTRAN-Subroutinen vorgestellt, die in diesem Zusammenhang sehr hilfreich sind.

Der Polynomialtest für die Ausreißer-Alternative kann daher auch als Binomialtest durchgeführt werden, wenn - wie bisher behandelt- die Überfrequentierung für eine ganz bestimmte der k Kategorien geprüft werden soll. Hierzu ist lediglich die hypothetische "Ausreißer-Kategorie" mit der "Rest-Kategorie" zu kontrastieren, die aus der Zusammenfassung aller anderen k-1 Kategorien resultiert. Geprüft wird die H_0 p(Ausreißer-Kategorie) = $1/k$ bzw. p(Rest-Kategorie) = $(k-1)/k$ gegen die einseitige Alternativhypothese p(Ausreißer-Kategorie) > $1/k$.

Liegt demgegenüber der Fall vor, daß zwar eine Ausreißer-Alternative, nicht aber die Kategorie abgeleitet werden kann, in der die vermutete Überfrequentierung auftritt, ist ein nicht-kategoriegebundener (im folgenden auch k-kategoriell genannter) Test zu fordern, der die Ausreißer-Hypothese für eine beliebige der k Kategorien prüft:

$$H_1{}^k: \begin{cases} \exists j\,(1 \le j \le k): p_j = 1/k + E, \text{ und} \\ \forall i\,(i<>j,\ 1 \le i \le k): p_i = 1/k - E/(1-k), \end{cases} \qquad (12)$$

wobei wiederum $0 < E < 1 - 1/k$.

Die k-kategorielle Verallgemeinerung des Tests stößt allerdings auf keine großen Probleme: man sucht sich zunächst das größte n_j^{emp} der Stichprobe und berechnet für dieses n_j^{emp} die (einkategorielle) Überschreitungswahrscheinlichkeit. Darüberhinaus müssen aber auch noch diejenigen k-komponentigen Partitionen von N betrachtet werden, die zu einer mindestens so großen Frequenz wie n_j^{emp} in einer anderen Kategorie geführt hätten, sofern sie nicht schon betrachtet worden sind. Die zugehörigen Wahrscheinlichkeiten sind mit der zuerst berechneten (einkategoriellen) Überschreitungswahrscheinlichkeit aufzusummieren; das Resultat ist die Überschreitungswahrscheinlichkeit für die k-kategorielle Hypothese (12), die wiederum mit dem zuvor festgelegten α verglichen werden kann.

Auch beim k-kategoriellen Test ist natürlich die Ablehnung von H_0 nur dann der Annahme von H_1 äquivalent, wenn man eine Gleichverteilung bei den nicht überfrequentierten Zellen unterstellt. Leider ist es aber bei k-kategorieller Fragestellung nicht generell möglich, die Überschreitungswahrscheinlichkeiten für den Polynomialtest über die einfachere Binomialverteilung zu errechnen. Man muß daher den etwas umständlichen Weg über die Polynomialverteilung

gehen. Wie sich dieser für großes k und N ziemlich rechenaufwendige exakte Test algorithmisch am besten bewerkstelligen läßt, wird deshalb im Anhang an einem kleinen Beispiel demonstriert.

Wir werden nun zu analysieren haben, wie die Power des vorgeschlagenen Tests für verschiedene Effektstärken E ausfällt. Voraussetzung dafür ist, daß zunächst für jedes interessierende N ein "kritischer Wert" n_{krit} definiert wird, der von einer der empirischen Kategoriehäufigkeiten erreicht oder überschritten werden muß, damit H_0 abgelehnt werden kann. Wir wählen diesen Wert n_{krit} jeweils so, daß ein α von "ungefähr" .05 gesichert ist (vgl. im einzelnen Tabelle 1). Durch n_{krit} ist nun jeweils (für den ein- oder k-kategoriellen Test) eine k-komponentige Partitionsmenge von N definiert. Für jede dieser Partitionen ist wiederum die Polynomialwahrscheinlichkeit, hier jedoch unter einer der Alternativen nach (12) bzw. (3), zu berechnen. Die Summe dieser Wahrscheinlichkeiten ist die Power des Tests bei Gültigkeit der betreffenden Alternativhypothese. Für die interessierenden Stichprobenumfänge N aus unserem Beispiel (siehe dazu weiter unten) sind die Ergebnisse der Power-Analyse in Tabelle 1 aufgelistet. Wir sehen, daß die Power des 6-kategoriellen Tests schon ab $N \geq 19$ für $E \geq .40$ sehr erfreulich ausfällt.

Tabelle 1: Power des k-kategoriellen Polynomialtests für die Ausreißer-Alternative bei k=6, α ungefähr 0.05, für verschiedene Stichprobengrößen und Effektstärken.

N	n_{krit}	exaktes α	Effektstärke E				
			.10	.20	.30	.40	.50
17	8	.021	.065	.261	.582	.852	.973
19	8	.047	.124	.399	.735	.934	.993
23	9	.052	.151	.487	.826	.972	.998
24	9	.070	.190	.551	.867	.982	1.00
25	10	.028	.112	.441	.807	.970	.998
31	11	.053	.197	.625	.925	.995	1.00
32	11	.069	.234	.675	.944	.997	1.00
34	12	.040	.183	.632	.935	.996	1.00
36	12	.067	.252	.722	.964	.999	1.00
38	13	.039	.201	.683	.957	.998	1.00
43	14	.048	.249	.763	.979	1.00	1.00

Um einen Eindruck davon zu erhalten, um welchen Betrag die Power des Tests für die Ausreißer-Alternative größer als die des bekann-

ten Chi²-Anpassungstests ist, wurde für N = 17 und α = .021 eine entsprechende Power-Analyse für den exakten Test durchgeführt.[3] Für E = .10 bis E = .50 beträgt die Teststärke hier .055, .219, .513, .799 und .954. Die Power unseres Tests ist also durchweg größer, z.T. sogar um mehr als 5%. Der Aufwand der Testkonstruktion hat sich demnach gelohnt.

3 Datenrückgriff

Zur Demonstration des vorgestellten Tests greifen wir auf eine Untersuchung zurück[4], in der es um die Bewertung der Problemlösegüte fiktiver Spieler ging. Probanden im Alter von 8 (N=31), 10 (N=43), 12 (N=40), 15 (N=36), 20 (N=27), 40 (N=21) und 60 Jahren (N=29) nahmen im Sommer 1979 an einem Versuch teil, bei dem sie unter anderem zunächst selbst zwei Problemstellungen zur sequentiellen Zeichenvorhersage in spielerischer Form bearbeiten sollten. Abbildung 1 zeigt jeweils ein Beispiel für die Präsentation beider Programme auf dem Monitor eines Tischrechners. Abbildung 1a demonstriert einen Zwischenstand beim sog. Weltner-Paradigma: vom Probanden ist schrittweise vorherzusagen, in welche von fünf möglichen Richtungen sich das begonnene geometrische Muster fortsetzen wird. Im Falle einer falschen Vorhersage probiert der Proband solange weiter, bis er die richtige Vorhersage getroffen hat. Die richtig gelösten Vorhersagen bleiben auf dem Monitor präsent und bilden somit einen externen Speicher der nach einer bestimmten Regel festgelegten Zeichenfolge. Dieser externe Speicher fehlt dagegen bei der Bearbeitung des sog. Hussy-Paradigmas (vgl. Abbildung 1b): aus einer Zeichenmatrix ist das Element anzugeben, das im nächsten Zug aufblinken wird. Dem Probanden ist jeweils nur eine Vorhersage erlaubt. Er muß die "richtigen" Zeichen (d.h. die tatsächlich auf-

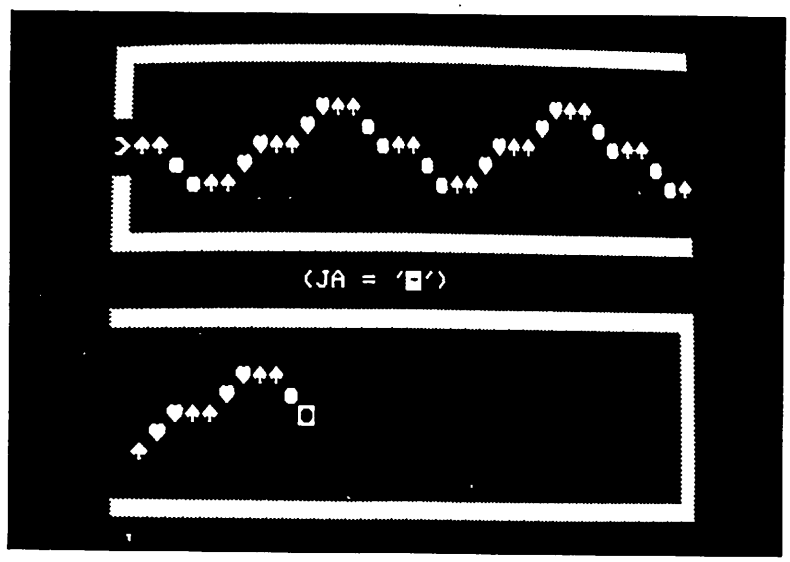
---

[3] Der approximative Test über die Chi²-Verteilung ist bei N=17 und k=6 nicht zulässig, da die Erwartungswerte für die einzelnen Frequenzen deutlich unter 5 liegen.

[4] Die im folgenden herangezogenen Daten entstammen dem Forschungsprojekt "Entwicklung informationsreduzierender und -generierender Strukturen als lebenslanger Prozeß" (Leiter: Prof. Dr. W. Hussy und Dr. habil. A. von Eye), unterstützt von der Stiftung Volkswagenwerk.

blinkenden Symbole) intern speichern, wenn er die Regel, nach der die Zeichenfolge aufgebaut ist, erkennen will.

a)



b)

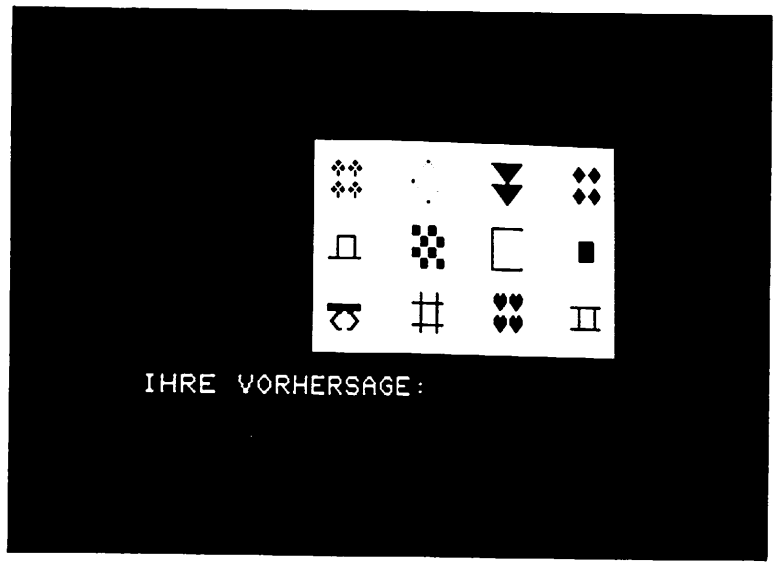


Abbildung 1: Präsentation von zwei Vorhersageproblemen auf dem Monitor: das (a) Weltner- und (b) Hussy-Paradigma.

Weitere Details interessieren in diesem Zusammenhang nicht (vgl. dazu Funke & Hussy, 1979). Wir richten unser Augenmerk auf die sog. Evaluationsbögen, welche die Probanden im Anschluß an die jeweiligen Vorhersagefolgen zu bearbeiten hatten. Aufgabe dabei war es, die schriftlich dokumentierten Vorhersageleistungen von drei fiktiven "Problemlösern" auf ihre Lösungsnähe hin zu beurteilen und

in eine Rangreihe zu bringen. Abbildung 2 zeigt die drei zu beurteilenden Alternativen des Evaluationsbogens zum Aufgabentyp "Weltner".

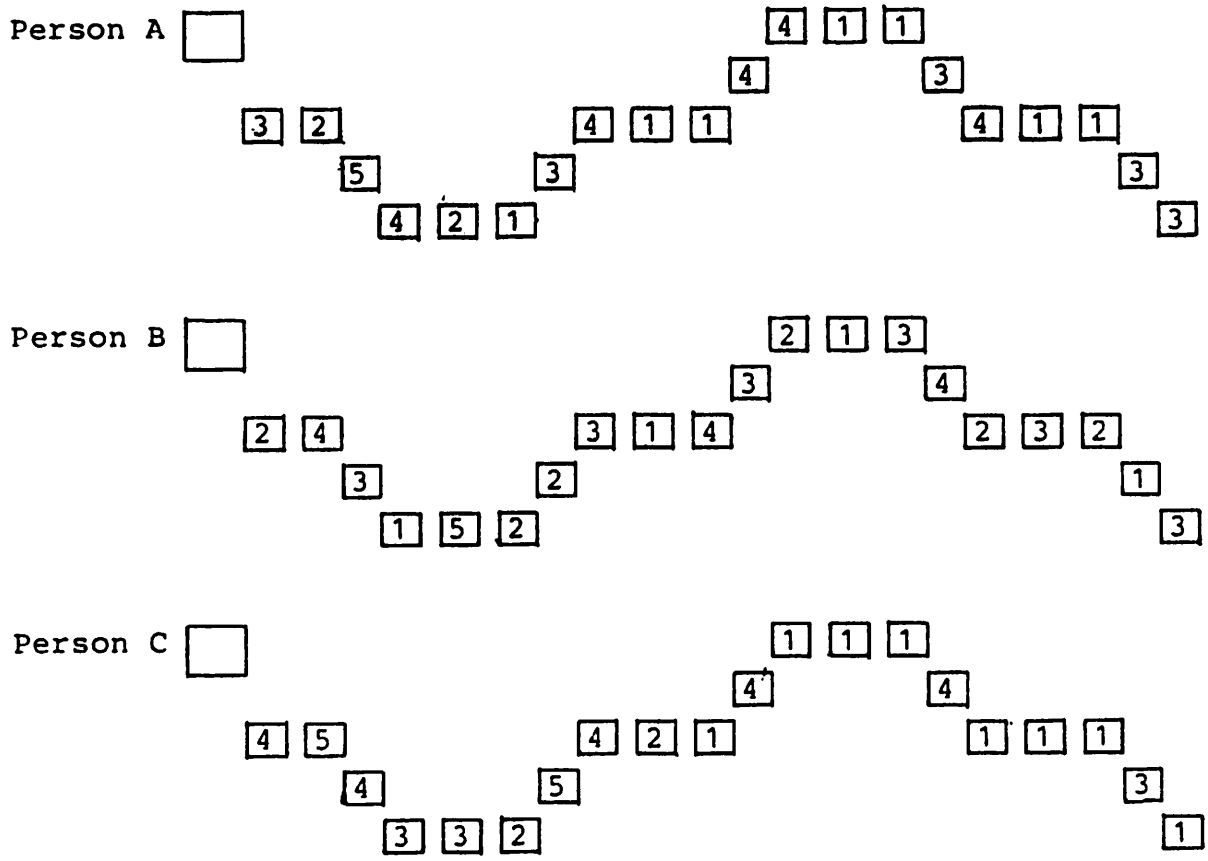


Abbildung 2: Vorlage zur Weltner-Evaluation.

Die Anordnung der Kästchen in Abbildung 2 stellt das bereits demonstrierte aufzudeckende Muster dar, wobei die Ziffer in jeder Zelle angibt, wie oft der fiktive Problemlöser probiert hat, bis er das richtige Zeichen erkannt hat. Man beachte, daß alle drei "Problemlöser" jeweils 51 Versuche in 20 Durchgängen aufweisen; eine Rangierung nach dem naheliegenden Kriterium (oder "Evaluator") der geringsten mittleren Fehlerrate ist damit nicht möglich. Zwangsläufig muß daher die subjektive Bewertung anderen Evaluatoren folgen, z.B. "durchschnittliche Fehlerrate in einem bestimmten Folgenausschnitt", "Anzahl fehlerloser Durchgänge" etc. Es sei ausdrücklich betont, daß die tatsächlich vergebene Rangreihe keine eindeutigen Rückschlüsse auf den benutzten Evaluator zuläßt; immer lassen sich ohne Schwierigkeiten mehrere Begründungen für eine bestimmte Rangierung der "Problemlöser" finden. Diese Feststellung ist insofern

wichtig, als sie impliziert, daß nicht der Evaluationsvorgang selbst (d.h. die vom Probanden herangezogenen inhaltlichen Bewertungskriterien), sondern nur dessen Ergebnis einer empirischen Analyse unterzogen werden kann (vgl. Funke, 1982).

Dies gilt auch für den Bogen zur Hussy-Evaluation, der analog zum Weltner-Bogen konzipiert wurde. Ausgangspunkt war hierbei eine aus Symbolen bestehende Zeichenmatrix. Die Vorhersagereihen dreier fiktiver Problemlöser zu je 20 Zeichen wurden zusammen mit der richtigen Folge in vertikaler Anordnung auf dem Papier abgetragen. Um den Probanden den Vergleich mit der richtigen Folge zu erleichtern, wurden die richtigen Vorhersagen jeweils mit einem Häkchen markiert. Bei allen drei "Problemlösern" wurde die Anzahl richtiger Vorhersagen auf acht fixiert, so daß auch bei diesem Evaluationstypus das naheliegende Kriterium der geringsten mittleren Fehlerrate keine Rangierung erlaubt. Den Evaluationsbogen zum Aufgabentyp "Hussy", auf dessen Darstellung hier aus Platzgründen verzichtet wird, findet man samt Instruktion im Anhang der Arbeit von Funke & Hussy (1980) wie auch im Anhang der Arbeit von Hussy & Köhl (1982a). Der Datenanalyse konnten 189 instruktionsgemäße Bearbeitungen des Hussy-Evaluationsbogens und 201 korrekte Bearbeitungen des Weltner-Bogens zugrunde gelegt werden.

4 Hypothesen

Allein die Tatsache, daß bei beiden Evaluationsbögen nicht von einer eindeutigen Relation zwischen der vom Probanden vergebenen Rangreihe und dem kognitiven Evaluationsvorgang, der zu der betreffenden Rangreihe führte, ausgegangen werden kann, verbietet die Interpretation von Altersunterschieden bezüglich der vergebenen Rangreihen im Sinne von Unterschieden in der Evaluationsgüte (vgl. auch Funke, 1982; anderer Ansicht sind Hussy & Köhl, 1982a, p. 9f, 1982b, p.7-8). Da die Vorhersagefolgen der drei fiktiven Problemlöser jeweils so konstruiert wurden, daß extreme Zielnähe bzw. Zielferne bei keinem von ihnen nahegelegt wird (z.B. nur Treffer oder nur Fehler in den letzten zehn Durchgängen), muß man sich fragen, ob es überhaupt sinnvoll ist, von einer "objektiv richtigen" Rangreihe auszugehen. Die Bewertung der drei Problemlösungsalternativen wird zusätzlich durch den Sachverhalt erschwert, daß

die rationalste Problemlösungsstrategie häufig eine Falsifikationsstrategie ist. Sei a eine Hypothese über die Regel, nach der die Zeichenfolge aufgebaut ist; seien ferner b_i empirische Ereignisse, die durch a impliziert werden und non-b_i die dazu komplementären empirischen Ereignisse, welche im Widerspruch zu a stehen. Durch "Anhäufen" von empirischen Ereignissen b_i wird man dann selbstverständlich nie Klarheit darüber bekommen, ob a wahr ist oder nicht. Tritt dagegen ein Ereignis non-b_i ein, so kann a über den modus tollens falsifiziert werden; außerdem muß eine neue Hypothese gesucht werden, die die bisherige Zeichenfolge erklären kann. Nun soll mit diesem Beispiel keineswegs gesagt werden, daß sich der "beste" Problemlöser dadurch auszeichnet, daß er im Problemlösungsprozeß ausschließlich Vorhersagen vom Typ non-b_i macht, die - wenn die Hypothese a zutrifft - natürlich sämtlich falsch wären; er wird sich nach der Falsifizierung der von ihm in Erwägung gezogenen "Konkurrenzhypothesen" selbstverständlich dazu entschließen, richtige Vorhersagen im Sinne von Hypothese a zu machen. Mit dem Beispiel sollte lediglich aufgezeigt werden, daß "Fehler" (in unserem Beispiel also falsche Vorhersagen) - speziell in der Anfangsphase des Problemlösungsprozesses - nicht notwendig "Lösungsferne" bedeuten, sondern ganz im Gegenteil eine vernünftige Strategie reflektieren können, der Lösung näher zu kommen. Die tatsächliche Lösungsnähe der drei fiktiven Problemlöser muß daher nicht notwendig mit einem bestimmten Aspekt der Vorhersagefolge korrespondieren.

Selbst wenn sich zeigen ließe, daß mehrere Probanden, die man als Evaluations-Experten für diesen Aufgabentyp zu akzeptieren bereit ist, sämtlich dieselbe Rangreihe für die drei Alternativen vergeben, so sollte man nicht kurzschlüssig diese Rangreihe zur Norm erheben. Tversky & Kahneman (1971) u.a. haben in einem anderen Zusammenhang klar gezeigt, daß auch die mehrheitliche Meinung von sog. Experten de facto falsch sein kann.

Diese Überlegungen legen es nahe, einen theoretischen Zugang zum Problem der Altersunterschiede bei der Evaluation von Problemlösungsalternativen zu wählen, der nicht auf inhaltliche, sondern auf <u>strukturelle</u> Aspekte der Beurteilung und Bewertung in bestimmten Altersgruppen abzielt. Wir gehen davon aus, daß interindividuelle Varianz in den Evaluationsurteilen prinzipiell zweierlei Quellen entstammen kann, nämlich (a) unterschiedlichen Bewertungskriterien

("Evaluatoren") und (b) der richtigen oder falschen Anwendung eines Bewertungskriteriums. Um es konkreter auf die in dieser Arbeit untersuchten Paradigmen zu beziehen: es ist unmittelbar evident, daß zwei Probanden, die unterschiedliche Bewertungskriterien zugrunde legen, zu unterschiedlichen Rangreihen für die drei Alternativen gelangen können; sofern zwei Probanden dasselbe Bewertungskriterium zugrunde legen, ist eine Übereinstimmung der vergebenen Rangreihen ebenfalls nicht zwangsläufig, da der eine das Kriterium richtig und der andere es falsch anwenden kann.

Wir vermuten, daß sowohl die Wahl der Bewertungskriterien als auch die Richtigkeit/Falschheit ihrer Anwendung zum einen von der Aufmerksamkeit in der gegebenen Situation, zum anderen aber auch von dem abhängen, was man gemeinhin als "Urteilsvermögen" (reasoning) bezeichnet. Einiges scheint dafür zu sprechen, daß dieser nichtverbale Aspekt menschlicher Intelligenz tendenziell eine umgekehrt u-förmige Beziehung zum chronologischen Alter aufweist: neben einem steilen Anstieg des "Urteilsvermögens" bis zum Alter von ca. 15-20 Jahren ist ein weniger deutlicher, aber dennoch auch in Längsschnittstudien reliabel feststellbarer Abfall ab dem Alter von etwa 30-40 Jahren zu konstatieren (vgl. Botwinick, 1977). Bezogen auf das hier vorgestellte Paradigma hieße dies, daß die Bewertungskriterien, anhand derer die Rangreihenvergabe vorgenommen wird, in mittleren Altersgruppen eher kollektiven, in jüngeren und älteren Altersgruppen dagegen eher idiosynkratischen Charakter aufweisen. Die genaue Abgrenzung der Altersgruppen, in denen einheitliche bzw. unterschiedliche Bewertungskriterien zu erwarten sind, dürfte in erster Linie von der Schwierigkeit der Evaluationstätigkeit abhängen. Gleiches gilt für die Korrektheit der Anwendung von Bewertungskriterien: die Schwierigkeit der Evaluationstätigkeit sollte determinieren, welcher Ausprägungsgrad der Variablen "Urteilsvermögen" für eine richtige Anwendung der Bewertungskriterien ausreicht.

Betrachten wir nun die Evaluationsurteile, die nach unseren theoretischen Vorüberlegungen sowohl von der Wahl der Bewertungskriterien als auch von deren Anwendung abhängen, so sind für mittlere Altersgruppen - je nach Schwierigkeitsgrad - altershomogene, für jüngere und ältere Personen dagegen innerhalb einer Altersgruppe streuende Rangreihen zu erwarten.

Nach diesen eher allgemeinen Überlegungen zu Altersunterschieden werden wir im folgenden versuchen, unsere Hypothesen bezüglich struktureller Alterseffekte zu präzisieren. Dabei werden wir den verschiedenen Aufgabentypen (d.h. Weltner und Hussy) Rechnung zu tragen haben. Es darf davon ausgegangen werden, daß die Evaluation zum Hussy-Paradigma schwieriger als die zum Weltner-Paradigma ist. Nicht nur das erhöhte Zeicheninventar, sondern auch die Darstellungsform auf dem Evaluationsbogen im Vergleich zur vorher bearbeiteten Problemstellung spricht für eine derartige Unterscheidung in leichtere vs. schwierigere Evaluation. Der um eine Drittel erhöhte Anteil nicht-instruktionsgemäßen Verhaltens (d.h. Fälle von Rangbindungen) bei der Hussy-Evaluation (insgesamt 24 im Vergleich zu 16 bei Weltner) mag als weiterer Beleg für diese Annahme gelten.

Akzeptiert man die unterschiedliche Schwierigkeit beider Bögen im geschilderten Sinn, so sollte man erwarten, daß der in den vorangestellten theoretischen Überlegungen angesprochene Altersseffekt deutlicher bei der schwierigeren Hussy-Evaluation auftritt. Wir werden demzufolge zwei Hypothesen, H(Hussy) und H(Weltner), zu unterscheiden haben:

H(Hussy): Für 15- und 20jährige Probanden wird beim Hussy-Typ die Überfrequentierung genau einer Rangordnungsalternative prognostiziert. Demgegenüber sollten die Rangreihen-Präferenzen für die anderen Altersgruppen gleichverteilt sein.

H(Weltner): Beim Weltner-Typ, der geringere Anforderungen an die kognitive Leistungsfähigkeit stellt, kann demgegenüber eine Ausweitung des Bereichs altershomogener Urteile auf benachbarte Altersgruppen - also die 12- und 40-jährigen - erwartet werden. Gleichverteilte Präferenzen werden demnach für die 8-, 10- und 60jährigen postuliert; die Überfrequentierung genau einer Rangreihe soll dagegen für die 12-, 15-, 20- und 40jährigen charakteristisch sein.

Es sei nochmals betont, daß wir keine Hypothese darüber aufstellen können und wollen, welche Rangreihe von den mittleren Altersgruppen jeweils präferiert wird. Nicht bestritten wird allerdings,

daß eine post-hoc-Analyse der Daten in bezug auf die Art der jeweils präferierten Rangreihe heuristisch fruchtbar sein kann (dies zeigen die Arbeiten von Hussy & Köhl, 1982a, 1982b).

5 Anwendung des Tests für die Ausreißer-Alternative

Beide Hypothesen, H(Hussy) und H(Weltner), implizieren offensichtlich, daß zwischen den Variablen "vergebene Rangreihe" und "Altersgruppe" für jeden Aufgabentyp stochastische Abhängigkeit vorliegen muß. Weder für den Hussy-Typ noch für den Weltner-Typ darf die Verteilung der Rangreihen über die Altersgruppen konstant sein, wenn die genannten Hypothesen zutreffen. Die Nullhypothese der stochastischen Unabhängigkeit kann bei den gegebenen Stichprobenumfängen approximativ via Perason-CHI^2- oder 2I-Statistik (vgl. Lienert, 1973) gegen die Alternativhypothese der stochastischen Abhängigkeit getestet werden, wobei wir uns allerdings bewußt sind, daß die Power der Tests bei "kleinen" Abweichungen von der Nullhypothese unbefriedigend ist. Für "kleine", "mittlere" und "große" Abweichungen von der Nullhypothese im Sinne von Cohen (1977, p. 224-225) ergibt sich die Power des Chi^2-Tests wie in Tabelle 2 dargestellt.[5]

Tabelle 2: Power des Chi^2-Tests für die Hussy- und Weltner-Daten bei $\alpha = 0.05$ und df = 30.

Effekt	Cohens W	Hussy-Daten (N_H = 189)	Weltner-Daten (N_W = 201)
"klein"	.10	.08	.09
"mittel"	.30	.58	.62
"groß"	.50	.99	1.0

Es ist klar ersichtlich, daß bei den gegebenen Stichprobenumfängen erst sehr deutliche stochastische Abhängigkeiten mit großer Wahrscheinlichkeit durch die Tests aufgedeckt werden können. Wir

---

[5] Da Cohen (1977) keine Power-Werte für df=(6-1)x(7-1)=30, α=0.05 und N_H=189 bzw. N_W=201 tabelliert hat, wurden die hier angegebenen Power-Werte mittels eines einfachen, aber recht genauen BASIC-Programms berechnet, das auf einer Arbeit von Milligan (1979) basiert. Ein Ausdruck dieses Programms kann beim Erstautor angefordert werden.

erwarten jedoch, daß der in H(Hussy) und H(Weltner) postulierte "Zentrierungseffekt" der Präferenzen in den mittleren Altergruppen stark ist, insbesondere bei den 15- bis 20jährigen. Die Tests auf stochastische Unabhängigkeit/Abhängigkeit sind somit durchaus faire Prüfinstanzen für diese Hypothesen.

Tabelle 3: Häufigkeiten möglicher Rangreihen bei der (a) Hussy-Evaluation sowie (b) Weltner-Evaluation, getrennt nach Altersgruppen, für die drei Alternativen ABC.

(a) Hussy-Evaluation

Rangreihe	Altersgruppe							Gesamt
	08	10	12	15	20	40	60	
1 2 3	7	5	2	2	2	1	4	23
1 3 2	2	5	6	2	2	2	5	24
2 1 3	3	11	6	4	2	3	2	31
2 3 1	5	11	8	15	13	4	9	65
3 1 2	1	6	11	4	3	3	3	31
3 2 1	1	-	1	5	3	4	1	15
Summe	19	38	34	32	25	17	24	189

Anmerkung: Pearson-Chi^2 = 48.92, 2I = 45.29 bei df = 30.

(b) Weltner-Evaluation

Rangreihe	Altersgruppe							Gesamt
	08	10	12	15	20	40	60	
1 2 3	6	5	3	2	1	-	2	19
1 3 2	2	4	9	5	1	1	-	22
2 1 3	4	10	3	3	1	1	6	28
2 3 1	9	13	11	13	11	7	8	72
3 1 2	-	4	7	2	4	1	1	19
3 2 1	3	7	3	6	7	9	6	41
Summe	24	43	36	31	25	19	23	201

Anmerkung: Pearson-Chi^2 = 51.73, 2I = 50.62 bei df = 30.

Wenn, wie erwartet, sowohl für die Hussy- als auch für die Weltner-Daten auf stochastische Abhängigkeit erkannt werden kann, stellt sich jedoch das Folgeproblem, ob die Struktur der stochastischen Abhängigkeit auch von der Art ist, die in den Hypothesen

behauptet wurde. Dies fair zu prüfen erlaubt uns nun der der vorgestellte Test für die Ausreißer-Alternative.

Die verwertbaren[6] Evaluationsurteile von insgesamt 227 Versuchspersonen aus dem Altersbereich der 8- bis 60jährigen sind in den Tabellen 3a und 3b getrennt nach den Problemarten "Hussy" und "Weltner" zusammengestellt. Die jeweils drei zu beurteilenden Alternativen (ABC) erlauben 3! = 6 mögliche Rangreihen, deren altersspezifische Frequentierungen in den Tabellen 3a und 3b aufgeführt sind. Eine Rangreihe der Art '231' bedeutet bespielsweise, daß der "Problemlöser" A auf den zweiten, B auf den dritten und C auf den ersten Platz verwiesen wird.

Zunächst muß die Frage beantwortet werden, ob in den sieben ausgewählten Altersgruppen gleiche Präferenzverteilungen für die Hussy- und die Weltner-Daten vorliegen oder nicht. Die Hypothese der stochastischen Abhängigkeit zwischen den Altersgruppen einerseits und den sechs möglichen Rangfolgen andererseits konnte mittels $\underline{Chi^2}$- bzw. $\underline{2I}$-Test für beide Evaluationsbereiche bestätigt werden: der für $\alpha=0.05$ bei df=30 kritische Tabellenwert von 43.77 wurde jeweils knapp überschritten (vgl. die Angaben in den Tabellen 3a und 3b).

Um nun gemäß H(Hussy) und H(Weltner) Antwort auf die Frage nach alterstypischen Präferenzen zu geben, führten wir für jede Altersgruppe den beschriebenen 6-kategoriellen Polynomialtest für die Ausreißer-Alternative durch. Tabelle 4 enthält die Ergebnisse der Berechnungen.

Tabelle 4: Überschreitungswahrscheinlichkeiten des 6-kategoriellen Tests für die Ausreißer-Alternative bei den zwei Evaluationsaufgaben "Hussy" und "Weltner", getrennt nach Altersstufen.

Evaluation	Altersgruppe						
	08	10	12	15	20	40	60
"Hussy"	.167	.244	.111	<u>.000</u>[1]	<u>.000</u>	.992	.070
"Weltner"	.070	.118	.169	<u>.005</u>	<u>.007</u>	<u>.011</u>	.164

[1] Unterstrichen sind die Überschreitungswahrscheinlichkeiten $p < .05$.

[6] Fälle mit Rangbindungen wurden in dieser Arbeit ausgeschlossen.

Die Ergebnisse in Tabelle 4 bestätigen zunächst unsere Hypothese, daß im Hussy-Bogen 15- bis 20jährige Probanden eine spezifische Rangreihe präferieren, während für alle anderen Altersgruppen die Nullhypothese beibehalten werden kann. Bei einem liberalen α-Risiko von 10% müßte allerdings auch bei den 60jährigen ein Zentrierungseffekt unterstellt werden.

Im Weltner-Bogen ist die mit H(Weltner) postulierte Ausdehnung auf 12- und 40jährige nur partiell bestätigt: die 12jährigen zeigen keine signifikante Überfrequentierung genau einer Rangreihe - womit der erste Bestandteil unserer Hypothese keine Bestätigung findet. Bei den 40jährigen zeigt sich dagegen der erwartete Zentrierungseffekt. Wäre α=.10 zugrundegelegt worden, so müßten bei diesem Aufgabentyp die 8jährigen zu den Gruppen mit homogener Urteilsstruktur gezählt werden.

6 Diskussion der Befunde

Die Daten stützen unsere globale Ausgangshypothese, daß verschiedene Altersgruppen sich hinsichtlich des Evaluationsverhaltens strukturell unterscheiden: in den "mittleren" Altersgruppen liegt eine signifikante Tendenz zu (interindividuell) homogenen Urteilen vor, während jüngere und ältere Probanden "diffus", d.h. interindividuell heterogen urteilen.

Nicht vollständig bestätigt wurden dagegen unsere spezielleren Hypothesen bezüglich unterschiedlicher Alterseffekte bei "leichten" und "schweren" Evaluationstätigkeiten. Zwar ist der Bereich altershomogener Urteile bei der leichteren Weltner-Aufgabe gegenüber der schwereren Hussy-Aufgabe breiter, nicht jedoch - wie von uns prognostiziert - für jüngere und ältere Probanden in gleicher Weise.

Das Ausbleiben eines signifikanten Zentrierungseffekts für die 12jährigen bei der Weltner-Aufgabe kann nicht auf die zu geringe Power des Tests zurückgeführt werden. Mit N = 36 handelt es sich hier um die drittgrößte der analysierten Stichproben überhaupt, so daß die Power eher größer als bei den anderen Tests ausfällt (vgl. auch Tabelle 1). Wenn also überhaupt ein Zentrierungseffekt im Sinne von H(Weltner) bei dieser Gruppe vorliegt, dann ist er so

klein, daß ihm keine praktische Bedeutsamkeit zugesprochen werden kann.

Für die 12jährigen scheint somit der von uns behauptete Schwierigkeitsunterschied zwischen der Weltner- und der Hussy-Aufgabe nicht relevant zu sein; zumindest drückt er sich im Urteilsverhalten nicht aus. Eine mögliche Folgerung aus diesem Ergebnis könnte lauten, daß altershomogene Evaluationsurteile generell erst ab einer bestimmten Altersstufe (z.B. 15 Jahre) zu erwarten sind, unabhängig davon, ob die Evaluationstätigkeit als "schwer" oder "leicht" eingeschätzt wird. Schwierigkeitseffekte würden sich dann erst in den höheren Altersstufen niederschlagen. Diese neue Hypothese stünde im Einklang mit kognitiven Entwicklungstheorien, die einen "Sprung" der kognitiven Entwicklung in der Pubertät behaupten (z.B. Piaget, 1966). Selbstverständlich bedarf diese Hypothese einer erneuten Überprüfung an einem unabhängig zu gewinnenden Datensatz. Dabei sollte sowohl die Schwierigkeitsvariable als auch die Altersvariable in dem interessierenden Bereich feiner abgestuft werden.

Die "Diffusion" des Evaluationsurteils in den höheren Altersgruppen, welche - wie prognostiziert - bei der schwierigeren Hussy-Aufgabe eher festzustellen ist als bei der leichteren Weltner-Aufgabe, führen wir auf nachlassende kognitive Leistungsfähigkeit zurück, die für verschiedene Leistungsbereiche bei querschnittlicher Datenanalyse bereits belegt ist (vgl. Hussy & Funke, 1982). Allerdings kann nicht ausgeschlossen werden, daß sich hierunter zumindest zum Teil auch Kohorteneffekte verbergen.

Abschließend wollen wir uns der Frage zuwenden, von welcher Art die Rangreihen sind, die in den verschiedenen Altersgruppen präferiert werden. Die Antwort auf diese Frage wird durch die Tabellen 5a und 5b erleichtert, in die die Überschreitungswahrscheinlichkeiten für Binomialtests auf Abweichung einer Zellenproportion vom Erwartungswert unter der Nullhypothese (also $p^j = 1/6$) eingetragen sind. Auch diese Tests wurden für jede Altersgruppe und jeden Aufgabentyp getrennt durchgeführt.

Diese Analyse hat für uns - wie weiter oben bereits angedeutet- ausschließlich explorativen Charakter. Inferenzstatistisch sind die Binomialtests wenig brauchbar, da ihre Power bei den gegebenen Stichprobenumfängen und wegen der aufgrund wiederholter Tests an denselben Daten notwendigen α-Justierung sehr gering ausfällt. Legt

man ein justiertes α^* von .008 zugrunde, resultieren daher sehr konservative Tests der Nullhypothese.

Tabelle 5: Überschreitungswahrscheinlichkeiten der Binomialtests für die Überfrequentierung der Rangreihen bei der (a) Hussy-Evaluation und (b) Weltner-Evaluation, getrennt nach Altersgruppen ($p > .50$ wurden ausgespart).

(a) Hussy-Evaluation

Rangreihe	Altersgruppe						
	08	10	12	15	20	40	60
1 2 3	.03	-	-	-	-	-	-
1 3 2	-	-	-	-	-	-	.37
2 1 3	-	.04	-	-	-	-	-
2 3 1	.20	.04	.19	.00[1]	.00	.31	.01
3 1 2	-	-	.02	-	-	-	-
3 2 1	-	-	-	-	-	-	.31
N	19	38	34	32	25	17	24

1 Unterstrichen sind die Überschreitungswahrscheinlichkeiten $p < \alpha^*$, wobei $\alpha^* = \alpha/k = .05/6 = .008$.

(b) Weltner-Evaluation

Rangreihe	Altersgruppe						
	08	10	12	15	20	40	60
1 2 3	.20	-	-	-	-	-	-
1 3 2	-	-	.13	-	-	-	-
2 1 3	-	.17	-	-	-	-	.17
2 3 1	.01	.02	.03	.00[1]	.00	.03	.03
3 1 2	-	-	.40	-	-	-	-
3 2 1	-	-	-	.41	.11	.00	.17
N	24	43	36	31	25	19	23

[1] Unterstrichen sind die Überschreitungswahrscheinlichkeiten $p < \alpha^*$, wobei $\alpha^* = \alpha/k = .05/6 = .008$.

Immerhin zeigt das Muster der Überschreitungswahrscheinlichkeiten deutlich, daß die Rangreihe "231" bei allen Altersgruppen und Aufgabentypen zu den "Favoriten" zählt. Was zeichnet diese Rangreihe aus? Legt man bei der Weltner-Evaluation als Kriterium für die Güte der Leistung z.B. das Kriterium "Anzahl der Treffer im

ersten Versuch" fest, so resultiert genau diese Rangreihe. Bei der Hussy-Evaluation ergibt sich diese Rangreihe bespielsweise dann, wenn man als Kriterium die Anzahl der Treffer in den letzten acht Versuchen zugrundelegen würde. Macht man die zusätzliche Annahme, daß die Anzahl der Treffer im ersten Versuch und speziell die Treffer in der Schlußphase ein guter Prädikator für zu antizipierenden zukünftigen Leistungen des fiktiven Probanden sind, so sind die Kriterien "vernünftig" zu nennen. Damit ist natürlich noch nicht gesagt, daß die Versuchspersonen, die die Rangreihe "231" vergeben haben, explizit oder implizit von diesen Kriterien ausgegangen sind: es lassen sich sicher auch andere - vielleicht weniger "vernünftige" - Begründungen für diese Rangreihe finden (z.B. "Anzahl aufeinanderfolgender Treffer"). Wir vermuten eher, daß nicht die Art des Kriteriums für die zu vergebende Rangreihe maßgeblich ist, sondern vielmehr die Anzahl möglicher Begründungen für eine bestimmte Rangreihe: je mehr Begründungen für eine bestimmte Rangreihe geliefert werden können, umso "plausibler" erscheint sie im Kontext der anderen möglichen Rangreihen. Die von uns prognostizierten und empirisch bestätigten Alterseffekte wären in diesem Zusammenhang derart zu interpretieren, daß die Anzahl möglicher Begründungen bei den mittleren Altersgruppen zwischen den sechs möglichen Rangreihen differenziert, während dies bei jüngeren und älteren Personen nicht der Fall ist. Dies könnte zum einen darauf beruhen, daß jüngeren und älteren Probanden "sinnvolle" Begründungen für bestimmte Rangreihen entgehen, zum anderen aber auch darauf, daß sie Begründungen gelten lassen, die von 15- bis 20jährigen als "unsinnig" verworfen werden würden. Für die Rangreihe "123" ist z.B. folgende Begründung denkbar, aber sicher nicht sinnvoll: "A ist am besten, B ist am zweitbesten und C ist am schlechtesten, weil die Guten immer über/vor den Schlechten stehen". Unterstellt man, daß 8jährige diese Begründung gelten lassen, ältere Personen aber nicht, so hat man eine Erklärung für den zunächst verblüffenden Sachverhalt, daß die Rangreihe "123" für die 8jährigen bei beiden Aufgabentypen offenbar recht attraktiv ist (vgl. Tabelle 5a und 5b).

Stimmt unsere Vermutung, so müßte sich empirisch eine positive Korrelation zwischen der Häufigkeit bestimmter Evaluationsurteile und der Anzahl möglicher Begründungen für diese Urteile feststellen lassen. Auch diese Hypothese wäre anhand neuer Daten zu prüfen.

Literatur

Botwinick, J. (1977). Intellectual abilities. In J.E. Birren & K.W. Schaie (Eds.), Handbook of the psychology of aging (pp. 580-605). New York: Van Nostrand Reinhold.

Cohen, J. (1977). Statistical power analysis for the behavioral sciences. Second edition. New York: Academic Press.

Funke, J. (1982). Schwierigkeiten bei der Beurteilung der Evaluationsgüte im Rahmen von Problemlösungsprozessen. Trierer Psychologische Berichte, 9, Heft 11.

Funke, J. & Hussy, W. (1979). Informationsverarbeitende Strukturen und Prozesse: Analysemöglichkeiten durch Problemlöseparadigmen. Trierer Psychologische Berichte, 6, Heft 8.

Funke, J. & Hussy, W. (1980). Informationsverarbeitende Strukturen und Prozesse: Entwicklung modellbezogener Meßinstrumente zur Erfassung von Aspekten der Informationsverarbeitung bei Kindern. Trierer Psychologische Berichte, 7, Heft 2.

Hussy, W. & Funke, J. (1982). Informationsverarbeitende Strukturen und Prozesse: Ergebnisüberblick zu einer Querschnittsanalyse an acht- bis sechzigjährigen Personen. Trierer Psychologische Berichte, 9, Heft 10.

Hussy, W. & Köhl, W. (1982a). Informationsverarbeitende Strukturen und Prozesse: Erfassung des kognitiven Aspekts von Evaluationen. Trierer Psychologische Berichte, 9, Heft 3.

Hussy, W. & Köhl, W. (1982b). Empirische, methodische und theoretische Beiträge zur Analyse der Entwicklung evaluativer Strukturen und Prozesse im kognitiven Bereich. Trierer Psychologische Berichte, 9, Heft 13.

Lienert, G. A. (1973). Verteilungsfreie Methoden in der Biostatistik. Band 1. Meisenheim: Anton Hain.

Milligan, G.W. (1979). A computer program for calculating power of the chi-square test. Educational and Psychological Measurement, 39, 681-684.

Piaget, J. (1966). Psychologie der Intelligenz. 2. Auflage. Zürich: Rascher.

Späth H. (1978). Ausgewählte Operations Research-Algorithmen in FORTRAN. München: Oldenbourg.

Tversky, A. & Kahneman, D. (1971). Belief in the law of small numbers. Psychological Bulletin, 76, 105-110.

Anhang

Berechnung des "Polynomialtests für die Ausreißer-Alternative".

Seien $k = 3$ Merkmale und $N = 6$ Beobachtungen gegeben, von denen $n_1 = 4$ das Merkmal 1, $n_2 = 2$ das Merkmal 2 und $n_3 = 0$ das Merkmal 3 aufweisen. Die Nullhypothese (2) soll gegen die (einkategorielle) Alternativhypothese (3) getestet werden.- Man betrachte die Menge der 3-komponentigen Summenzerlegungen von 6 in ganze Zahlen n_j mit $n_1 \geq 4$ sowie $n_2 \geq 0$ und $n_3 \geq 0$ und berechne die zugehörigen Polynomialwahrscheinlichkeiten nach (6). Das Ergebnis zeigt Tabelle A1.

Tabelle A1: Durchführung des einkategoriellen Polynomialtests für die Ausreißer-Alternative.

Nr.	Partition(n_1,n_2,n_3)		$p(n_1,n_2,n_3 \mid H_0)$
1	(6, 0, 0)	$(6!/6!)\ (1/3)^6$	$= 1.37\ 10^{-3}$
2	(5, 1, 0)	$(6!/(5!\ 1!))\ (1/3)^6$	$= 8.23\ 10^{-3}$
3	(5, 0, 1)	$(6!/(5!\ 1!))\ (1/3)^6$	$= 8.23\ 10^{-3}$
4	(4, 2, 0)	$(6!/(4!\ 2!))\ (1/3)^6$	= .020558
5	(4, 1, 1)	$(6!/4!\ 1!\ 1!))\ (1/3)^6$	= .04115
6	(4, 0, 2)	$(6!/(4!\ 2!))\ (1/3)^6$	= .02058
		Summe	= .10014

Auf dem Niveau $\alpha = .05$ ist das Ergebnis nicht signifikant, so daß H_0 beibehalten wird.[7]

Wenn dagegen die Nullhypothese (2) gegen die 3-kategorielle Alternativhypothese (11) getestet werden soll, ist lediglich die Teilmenge obiger Partitionen zu betrachten, deren Elemente die Relation $n_1 \geq n_2 \geq n_3$ erfüllen. Für jede dieser Partitionen berechne man die Polynomialwahrscheinlichkeiten nach (6) und multipliziere das Ergebnis mit der Anzahl möglicher Permutationen der drei Komponenten, wie in Tabelle A2 dargestellt. Auch das dann resultierende Ergebnis (Überschreitungswahrscheinlichkeit = .30042) wäre natürlich bei $\alpha = .05$ nicht signifikant.

---

[7] Dasselbe Ergebnis hätte man erzielt, wenn man die Überschreitungswahrscheinlichkeit über den Binomialtest für $N = 6$ Versuche, $k \geq 4$ Erfolge und p(Erfolg) = 1/3 bestimmt hätte.

Tabelle A2: Durchführung des 3-kategoriellen Polynomialtests für die Ausreißer-Alternative.

Nr.	Partition (n_1, n_2, n_3)	$p(n_1, n_2, n_3 \mid H_0)$	mögliche Permutationen	$p(n_1, n_2, n_3 \mid H_0)$ * Anzahl Permutat.
1	(6, 0, 0)	$1.37\ 10_{-3}$	= 3	$4.12\ 10_{-3}$
2	(5, 1, 0)	$8.23\ 10_{-3}$	3! = 6	.04938
3	(4, 2, 0)	.02058	3! = 6	.12346
4	(4, 1, 1)	.04115	= 3	.12346
			Summe =	.30042

Auf den ersten Blick mag es so scheinen, als ob die Überschreitungswahrscheinlichkeit für den k-kategoriellen Test zwangsläufig gleich dem Produkt von k und der einkategoriellen Überschreitungswahrscheinlichkeit sein muß. In unserem Beispiel trifft dies tatsächlich zu, da p(3-kategoriell) = .30042 = 3 * p(einkategoriell) = 3 * .10014. Man kann sich jedoch leicht klarmachen, daß sich ein ganz anderes Bild ergibt, wenn z.B. zwei "größte" Stichprobenfrequenzen $n_i^{emp} = n_j^{emp}$ vorliegen. Hier würde k * p(einkategoriell) eine Überschätzung von p(k-kategoriell) darstellen, weil die Wahrscheinlichkeiten für einige Partitionen doppelt gezählt würden. Es empfiehlt sich daher in jedem Fall die Durchführung des exakten Tests nach dem oben dargestellten Verfahren.

Die in der Arbeit aufgeführten Ergebnisse wurden mittels eines BASIC-Programmes gewonnen; ein Ausdruck dieses Programmes kann beim Erstautor angefordert werden.

Berichte aus dem Psychologischen Institut der Universität Bonn

Die "Berichte aus dem Psychologischen Institut der Universität Bonn" (ISSN 0931-024X) gibt es seit 1975. Die ersten vier Jahrgänge bestehen aus 21 fortlaufend numerierten Heften. Ab Jahrgang 5 (1979) beginnt die Heftzählung in jedem Jahr bei Heft 1. In den Jahren 1983 und 1985 sind keine "Berichte" erschienen. Eine Übersicht über die zuletzt publizierten Hefte gibt nachfolgende Liste. - Seit 1984 besteht am Institut eine zweite Reihe unter dem Titel "Bonner Methoden-Berichte".

Band 10 (1984)

Heft 1: Schmitz, P.G. (1984). Personality factors as determinants of the Spiral-After-Effect (SAE).
Heft 2: Rüppell, H. & Rüschstroer, H. (1984). GIN & CHIPS. Ein prozessorientiertes Curriculum zur Ausbildung der produktiven Intelligenz.

Band 12 (1986)

Heft 1: Funke, J. (1986). Ein Forschungsprogramm zur subjektiven Repräsentation dynamischer Kleinsysteme: Aufbau und Anwendung von Wissen in Abhängigkeit von Person- und Systemmerkmalen.
Heft 2: Bredenkamp, J. (1986). Dürfen wir psychologische Hypothesen statistisch prüfen?
Heft 3: Funke, J., Fahnenbruck, G. & Müller, H. (1986). DYNAMIS - Ein Computerprogramm zur Simulation dynamischer Systeme.

Band 13 (1987)

Heft 1: Fahnenbruck, G., Funke, J. & Müller, H. (1987). Wissensdiagnose bei dynamischen Systemen.
Heft 2: Müller, H., Funke, J., Fahnenbruck, G. & Rasche, B. (1987). Über die Auswirkungen verschiedener Aktivitätsanforderungen auf Wissen und Können im Kontext dynamischer Systeme.

Band 14 (1988)

Heft 1: Müller, H., Funke, J. & Rasche, B. (1988). Wechselseitige Abhängigkeiten: Zum Einfluß von Nebenwirkungen und Eigendynamik auf die Bearbeitung dynamischer Systeme.
Heft 2: Fahnenbruck, G., Funke, J. & Rasche, B. (1988). Vorwissensverträglichkeit, Steuerbarkeit, Steueranforderung und Darbietungsform als Determinanten der Bearbeitung dynamischer Systeme.
Heft 3: Erdfelder, E. (1988). The empirical evaluation of deterministic developmental theories.