

Trierer Psychologische Berichte

Band 9 (1982) Heft 11

Joachim Funke

Schwierigkeiten bei der Beurteilung der
Evaluationsgüte im Rahmen von
Problemlösungsprozessen

Bericht 10 aus dem EIS-Projekt (Entwicklung informationsreduzierender und -generierender Strukturen als lebenslanger Prozeß), unterstützt von der Stiftung Volkswagenwerk (Az.: II/35450: I, II).

SCHWIERIGKEITEN BEI DER BEURTEILUNG DER EVALUATIONSGÜTE IM RAHMEN VON PROBLEMLÖSUNGSPROZESSEN

Joachim Funke
Universität Trier

Zusammenfassung

HUSSY & KÖHL (1982) haben kürzlich ein Verfahren zur Bestimmung der Evaluationsgüte vorgeschlagen, dessen unzulässige Anwendung auf einen bestimmten Datensatz die vorliegende Arbeit kritisiert.

Ausgehend von einem Modell der Informationsverarbeitung, in dem Evaluatoren als Bewertungsmaßstäbe bei Ist-Soll-Discrepanzen einen ablaufenden Prozeß an zentralen Punkten steuern, legen die Autoren Evaluationsbögen zur Erfassung solcher Komponenten vor, in denen der Proband drei verschiedene Vorgaben in eine Rangreihe bringen soll. Aus dieser Information werden nun über eine differenzierte Strategie Maße der Evaluationsgüte abgeleitet, die Auskunft über eventuell verwendete Evaluatoren liefern sollen.

Es wird nachgewiesen, daß die geplante Identifikation von Evaluatoren in der vorgestellten Form zu beliebigen Schlüssen führen muß, was weniger am Verfahren der Operationalisierung selbst als vielmehr an seiner mangelnden Angemessenheit an den Datensatz liegt. Eine einfachere Auswertung zeigt interessante Altersunterschiede im Konkordanz- und Präferenzverhalten verschiedener Altersgruppen.

Hinweise darauf, in welcher Richtung Verbesserungen des Erhebungsmaterials vorgenommen werden sollten, um eine Anwendung der theoriegeleiteten Maßentwicklung zu ermöglichen, schließen die Arbeit ab.

1 Problemstellung

Beurteilungs- und Bewertungsprozesse finden in gängigen Vorstellungen über den Ablauf kognitiver Prozesse (NEWELL & SIMON 1972; LÜER 1973; DÖRNER 1976; HUSSY 1979, 1982) bedeutungsvolle Stellenzuweisungen, ohne daß die empirische Basis zu ihrem Zugang als geklärt bezeichnet werden darf. So wird vielerorts die von MILLER, GALANTER & PRIBRAM (1960) propagierte TOTE-Einheit herangezogen, um die Abfolge von Entscheidungen bei einfachen wie komplexen Problemstellungen zu verdeutlichen: DÖRNER (1976, P. 40f.) verwendet sie bei der Beschreibung verschiedener Heuristiken, LOMPSCHER (1972, p. 38) verweist darauf, daß ein Denkschritt daraufhin "geprüft" wird, "ob er der Lösung nützt", HUSSY (1979, p. 27f.) legt mit dem Konzept des "Zentralen Prozessors" den Entwurf einer Ablaufsteuerung vor, der sich permanent auf Evaluationen gründet.

Diese Liste ist absolut unvollständig, denn fast jeder Autor, der sich mit Problemlöseprozessen oder besser Prozessen der Informationsverarbeitung schlechthin beschäftigt, kommt um die Frage nach den Evaluationsprozessen nicht umhin. Es soll nicht der Eindruck erweckt werden, daß eine Klärung dieses Problems in Sicht wäre; vielmehr möchte die Arbeit aufzeigen, welche Schwierigkeiten sich bereits bei der Erfassung globaler Evaluationskonzepte und deren Auswertung sowie Interpretation ergeben. Damit mag deutlicher werden, warum zwar ständig von Evaluationsprozessen (implizit oder explizit) gesprochen wird, die Konstrukterfassung und die "Evaluation der Evaluation" jedoch weit hinter ihren theoretischen Stellenwert zurückfällt. Ein Forschungsbereich, der differenzierte Aus-

sagen möglich machen könnte, aus verschiedenen Gründen jedoch nicht weit verbreitet ist, soll ausgeklammert bleiben, nämlich Studien zum Zusammenhang von Blickbewegungen und kognitiven Prozessen (siehe dazu WINTERHOFF 1980; PUTZ-OSTERLOH 1981; BEATTY 1982).

Zugleich versteht sich diese Arbeit als Plädoyer dafür, daß die Auswertung von Daten nicht unnötig kompliziert erfolgen sollte, wenn das zur Verfügung stehende Datenmaterial einer derartigen Auflösung nicht entspricht.

2 Datenrückgriff

Die im folgenden herangezogenen Daten einer Evaluationsstudie sind Bestandteil der Arbeiten im Forschungsprojekt "Entwicklung informationsreduzierender und -generierender Strukturen als lebenslanger Prozeß" (EIS-Projekt, unterstützt durch die Stiftung Volkswagenwerk, Az. II/35450). Probanden im Alter von 8, 10, 12, 15, 20, 40 und 60 Jahren nahmen an einem Versuch teil, bei dem sie unter anderem zwei Problemstellungen zur sequentiellen Zeichenvorhersage (Typ "HUSSY" ohne, Typ "WELTNER" mit externem Gedächtnis) in spielerischer Form bearbeiten sollten; genauere Details zu Form und Präsentation dieser Problemstellungen finden sich bei FUNKE & HUSSY (1979). Von Interesse ist in unserem Zusammenhang die Bearbeitung sogenannter Evaluationsbögen im Anschluß an die Problembearbeitung: die Probanden haben die schriftlich dokumentierten Spielleistungen von drei fiktiven Personen zu beurteilen und in eine Rangreihe zu bringen. Die angegebenen subjektiven Rangreihen bilden das Rohmaterial für unsere weitere Analyse. Es handelt sich hierbei um eine Re-Analyse des von HUSSY & KÖHL (1982)

verwendeten Datensatzes. Die Rohdaten der 227 Versuchspersonen aus dem Altersbereich der 8- bis 60jährigen vom ersten Meßzeitpunkt im Sommer 1979 finden sich in den Tabellen 1 und 2, jeweils getrennt nach den Problemarten HUSSY und WELTNER.

Tabelle 1

Häufigkeiten möglicher Rangreihen bei der HUSSY-Evaluation, getrennt nach Altersgruppen, für die drei Alternativen ABC

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
Zwischen- summe	19	38	34	32	25	17	24	189
andere ¹	5	4	4	3	1	4	3	24
fehlend	7	1	2	1	1	-	2	14
Summe	31	43	40	36	27	21	29	227

¹Unter die Kategorie "andere" fallen: 1 1 1 (19x),
1 2 1 (1x), 1 3 1 (1x), 2 1 1 (1x) und 2 1 2 (2x).

Die drei zu beurteilenden Alternativen (ABC) sollten in eine von sechs möglichen Rangreihen gebracht werden; unter der Kategorie "andere" sind Fälle mit Rangbindungen aufgeführt.

Tabelle 2

Häufigkeiten möglicher Rangreihen bei der WELTNER-Evaluation,
getrennt nach Altersgruppen, für die drei
Alternativen ABC

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
Zwischen- summe	24	43	36	31	25	19	23	201
andere ¹	3	-	2	4	1	1	5	16
fehlend	4	-	2	1	1	1	1	10
Summe	31	43	40	36	27	21	29	227

¹Unter die Kategorie "andere" fallen: 1 1 1 (10x),
1 2 1 (2x), 2 2 1 (1x) und 3 1 1 (1x).

3 Zur Frage der Auswertungsstrategie

Bevor der bisherige Auswertungsvorschlag von HUSSY & KÖHL (1982) skizziert wird, muß kurz auf die Reizvorlagen (die Evaluationsbögen) eingegangen werden. Abbildung 1 zeigt die drei zu beurteilenden Alternativen des WELTNER-Paradigmas. Der HUSSY-Evaluationsbogen ist analog dazu konstruiert und soll deswegen hier nicht vorgestellt werden; man findet ihn im Anhang zur Arbeit von HUSSY & KÖHL (1982).

Die Kästchen in Abbildung 1 stellen in ihrer Anordnung das aufzudeckende Muster dar, wobei die Ziffer in jeder Zelle an-

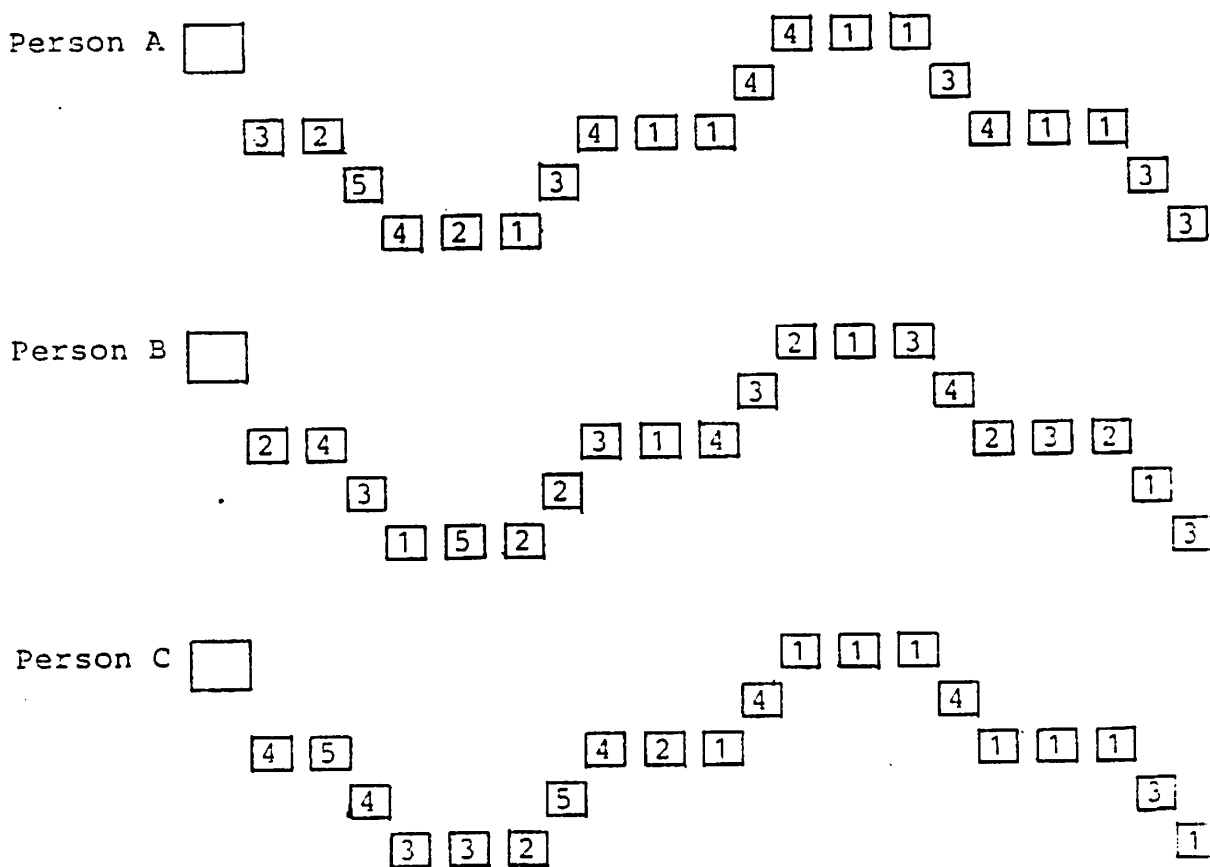


Abbildung 1. Vorlage zur WELTNER-Evaluation: die Zahl im Kästchen entspricht der Anzahl von Versuchen bis zum Finden der Lösung.

gibt, wie oft der fiktive Proband probiert hat, bis er das richtige Zeichen erkannte (der Rest der Folge war dabei natürlich nicht sichtbar). Offensichtlich kann die subjektive Bewertung verschiedenen Kriterien ("Evaluatoren") folgen: ein naheliegendes Kriterium würde den Probanden favorisieren, der die geringste mittlere Fehlerrate erzielt; jedoch weisen alle drei Personen insgesamt 51 Versuche in den 20 Zügen auf. Andere Kriterien könnten eine Beschränkung auf bestimmte Folgeausschnitte vorsehen (z.B. nur Endverhalten berücksichtigen)

oder die Personen nach der Anzahl fehlerloser Vorhersagen (also Kästchen mit "1") gruppieren; natürlich könnten auch andere (z.B. unsinnige oder zufällige) Kriterien Verwendung finden oder Kombinationen bisher genannter Kriterien. Diese Informationen liegen jedoch nicht vor; lediglich das "Endprodukt" der Evaluation, die subjektive Rangreihe, kann einer Analyse unterzogen werden.

3.1 Vergleich subjektiver und "objektiver" Rangreihen

HUSSY & KÖHL (1982) haben nun vorgeschlagen, die Alternativen in gleicher Weise mittels Transinformationsanalyse (vgl. HUSSY et al. 1981) zu bewerten, wie dies für die Analyse subjektiver Vorhersagen bereits gemacht wird. Die drei alternativen Folgen werden dabei als drei subjektive Vorhersagesequenzen betrachtet und erlauben somit die Berechnung ihrer Transinformation hinsichtlich der objektiven Folge. Die Evaluationsgüte (EG-Maß) eines bestimmten subjektiven Evaluationsvorschlags wird dann als Differenz zur besten Rangfolge bestimmt, wobei sich die "objektiv beste" Rangfolge aus der günstigsten Transinformation ergibt; dies erfolgt über einen Zwischenschritt, der Komponenten der Transinformation zu einem Verarbeitungsindex komprimiert und der hier nicht näher erläutert werden soll.

Die Strategie von HUSSY & KÖHL (1982) läßt sich also wie folgt zusammenfassen:

- 1) bestimme eine "objektive" Rangreihe der Evaluationsalternativen; diese kann sich auf globale oder differenzierte Aspekte der Folgen beziehen, es gibt also unterschiedliche

"objektive" Rangreihen je nach den zugrundegelegten Beurteilungskriterien ("Evaluatoren");

- 2) nehme die beste Rangreihe und berechne den Abstand zwischen ihr und allen möglichen anderen Rangreihen;
- 3) das so erhaltene EG-Maß kann interpretiert werden als die Stärke, mit der ein bestimmter Evaluationsaspekt, den man in Schritt (1) festgelegt hat, vom Probanden berücksichtigt wird.

Die Autoren schlagen eine Vielzahl von Differenzierungsmöglichkeiten vor: so die Berücksichtigung unterschiedlicher Folgenausschnitte (gesamte Folge, Anfangs-, Mittel- oder Endteil), den Einbezug unterschiedlicher Hypothesenniveaus (nur bei WELTNER-Evaluation möglich) oder die Unterscheidung verschiedener Komponenten der Transinformation (zeichenkorrekte versus trefferkorrekte; nur bei HUSSY-Evaluation möglich) oder auch die Bildung unterschiedlicher Verarbeitungsindizes (die Autoren führen insgesamt sieben davon auf). Den Vorteil der dadurch möglichen "Flexibilität" beschreiben die Autoren so:

"An diesen Überlegungen läßt sich auch eine weitere fruchtbare Eigenschaft des ... gewählten Weges zur Analyse des Evaluationsverhaltens demonstrieren. Man kann z.B. den Umfang der Endphase (...) solange vergrößern, bis sich eine bestimmte Rangreihe ... ergibt und von da aus Rückschlüsse auf Altersunterschiede bezüglich der Verwendung spezifischer Folgenausschnitte ziehen!" (HUSSY & KÖHL 1982, p. 50).

Dies besagt aber nichts anderes als: suche solange nach weiteren Differenzierungen, bis die subjektive Rangfolge als "Optimum" beurteilt wird, und schließe dann auf die Berück-

sichtigung dieses Aspekts durch den Beurteiler. Die Problematik dieser Vorgehensweise liegt in der Tatsache begründet, daß bei der Vielzahl der genannten Differenzierungsmöglichkeiten im Falle der vorliegenden Evaluationsbögen notwendigerweise mehrere Evaluatoren zu ein- und demselben Optimum führen müssen: es gibt bei drei Alternativen ja nur sechs mögliche Anordnungen, sofern keine gebundenen Ränge vorliegen; aus den oben genannten Aspekten läßt sich mehr als das Fünffache an potentiellen Evaluatoren bilden. Somit erübrigt sich die Interpretation der gewählten EG-Maße: es besteht keine eindeutige Relation mehr zwischen subjektiv gebildeter Rangfolge und einer wie auch immer hergestellten "objektiven" Rangfolge. Derartige Maßeigenschaften lassen den Operationalisierungsvorschlag zwar als sophistiziert, aber den vorliegenden Daten unangemessen erscheinen. Die Verwendung des Begriffs "objektive Rangreihe", von den Autoren im übrigen selbst problematisiert, täuscht über die Beliebigkeit der zur Interpretation herangezogenen Differenzierungsaspekte hinweg.

Schließlich klammern HUSSY & KÖHL (1982) das Problem gebundener Ränge unter Verweis auf "nicht-instruktionsgemäßes Verhalten" (p. 47) aus der Auswertung aus, obwohl diese Kategorie alterstypische Eigenschaften besitzt. Eine an Alterseffekten interessierte Analyse des Evaluationsverhaltens sollte folglich derartige Abweichungen berücksichtigen.

Als Alternative zu den bisher vorgeschlagenen Möglichkeiten der Beurteilung subjektiver Rangreihen schlagen wir nachfolgend eine einfache, auf Häufigkeitsinformationen beruhende Datenanalyse vor, die den Nachteil mangelnder Komplexität durch gesicherte Eindeutigkeit in der Aussage ausgleicht und Fehlinterpretationen vermeiden hilft.

3.2 Alternative Vorgehensweise

Die von uns vorgeschlagene Auswertungsstrategie verzichtet auf das Konzept einer objektiv besten Rangfolge, da hierüber keine Einigkeit herbeigeführt werden kann. Die Betrachtungsweise wechselt deswegen von der Ebene individueller Information zur Ebene aggregierter Daten: nicht mehr die Evaluation einer einzelnen Person findet Beachtung, sondern das Evaluationsverhalten ganzer Altersgruppen rückt in den Mittelpunkt des Interesses. Die Datenanalyse orientiert sich nunmehr an den zwei nachfolgend präzisierten Fragen.

Zunächst interessiert das Maß der Übereinstimmung bei der Beurteilung möglicher Alternativen in verschiedenen Altersgruppen. Unter der Annahme, daß mit zunehmendem Alter die Beurteilung homogener erfolgt, da geeignete Evaluatoren "problemadäquater" ausfallen dürften, erwarten wir vor allem für die Gruppen der 8- und 10jährigen Probanden eine eher unsystematische Beurteilung. Dieser Punkt bezieht sich auf die Homogenität des Urteilsverhaltens in verschiedenen Altersbereichen unabhängig von den präferierten Alternativen; es handelt sich also um einen strukturellen Aspekt.

Weiter richtet sich das Interesse darauf, welche Rangreihen über- bzw. unterfrequentiert sind. Das Problem der objektiven Rangfindung wird hier in die Fragestellung aufgelöst, welche Rangreihen in den einzelnen Altersgruppen zentrale Bedeutung besitzen. Damit wird ein Schritt in Richtung auf inhaltliche Interpretationen der möglichen Rangreihen vorgenommen, ohne deswegen die Verwendung spezifischer Evaluatoren postulieren zu müssen. Man kann diese Fragestellung auch als die nach der Richtung des Urteilsverhaltens bezeichnen; damit wird ein

inhaltlicher Aspekt beim Vergleich der Altersgruppen berücksichtigt.

3.3 Verwendete Methoden

Verfahren zur Feststellung von Urteilerhomogenitäten finden sich unter dem Titel "Boniturmethode" zum Beispiel bei LIENERT (1978, Kapitel 10). Ein bekanntes Maß ist der Konkordanzkoeffizient W von Kendall, der aus den Spaltensummen vorgegebener Bonituren (= Evaluatoren) die beobachtete Rangdevianz bestimmt und an der maximal möglichen relativiert. Der Koeffizient W , der zwischen Null (= minimale Konkordanz) und Eins (= maximale Konkordanz) variiert, läßt sich als W' auch im Falle von Rangbindungen für jede Altersgruppe bestimmen und ist damit für die Beantwortung unserer ersten Frage indiziert. W' läßt sich asymptotisch über die Chi-Quadrat-Verteilung bei $N-1$ Freiheitsgraden auf Signifikanz prüfen.

Um Aufschluß über "zentrale Tendenzen" bei der Rangvergabe zu erhalten, wird in jeder Altersgruppe die Häufigkeitsverteilung der Rangfolgen gegen eine Gleichverteilungsannahme mittels der $2I$ -Statistik von KULLBACK (1959; vgl. BLÖSCHL 1966) geprüft. Die Prüfgröße $2I$, gedacht für Kontingenztafeln mit geringen Erwartungswerten, ist asymptotisch (das heißt für großes N) wie Chi-Quadrat verteilt und kann analog bei $k-1$ Freiheitsgraden inferentiell beurteilt werden. Da einige Zellen nullfrequentiert waren, wurde in diesen Fällen zusätzlich KUL'S (1963) konservative Nullenkorrektur eingesetzt (vgl. LIENERT 1978, p. 474f.).

3.4 Hypothesen

Hinsichtlich der Urteilshomogenität wird eine zunehmende

Konkordanz im Alter von 8 bis 20 Jahren auf der Basis allgemeiner Leistungssteigerung im kognitiven Bereich sowie sozialisationsbedingter Anpassung im Bereich evaluativer Strukturen erwartet. Für 40- und 60jährige werden ähnlich wie bei 8- und 10jährigen verstärkt idiosynkratische Entscheidungen postuliert, die zu verminderter Konkordanz führen.

Die Erwartungen über Unterschiede zwischen beiden Paradigmen lassen sich so formulieren: bei insgesamt ähnlichen Altersunterschieden kann bei der WELTNER-Evaluation ein erhöhtes Niveau der Urteilshomogenität gegenüber den HUSSY-Ergebnissen erwartet werden, da die Aufgabenstellung im ersten Fall als subjektiv leichter anzusehen ist; darauf deutet bereits der geringere Prozentsatz nicht-instruktionsgemäßen Verhaltens bei der WELTNER-Evaluation hin (vgl. Tabellen 1 und 2).

Hinsichtlich der Präferenz einzelner Rangreihen soll - als Konsequenz aus dem unter 3.2 Gesagten - keine begründete Vorhersage gemacht werden. Jedoch wird bei jungen wie alten Probanden als Ausweg aus der für sie unklaren Bewertungssituation ein Ausweichen in nicht regelgemäßes Beurteilen bzw. in die Kategorie "123" erwartet, also eine Rangreihe, die der Vorgabesequenz entspricht; diese Hypothese stützt sich auf Beobachtungen der Versuchsleiter.

Die Ausbildung von Präferenzen sollte - will man Hypothesen über Unterschiede zwischen beiden Evaluationstypen formulieren - im WELTNER-Teil insgesamt leichter fallen als bei der HUSSY-Evaluation. Neben einer kürzeren und zugleich leichteren Instruktion wird die HUSSY-Evaluation im Unterschied zum WELTNER-Bogen durch ihre spezifische Präsentationsform erschwert: das sequentielle Aufleuchten einzelner Zeichen der

Symbolmatrix wird in eine stark veränderte Form, ein räumliches Untereinander gebracht und verlangt somit vom Probanden eine zusätzliche kognitive Operation, die der ursprünglichen Problemvorgabe fremd ist und das zur Beurteilung erforderliche Vorstellungsvermögen beeinträchtigt.

4 Ergebnisse der vereinfachten Datenanalyse

Entsprechend den zwei gestellten Fragen nach altersspezifischen (1) Konkordanzen und (2) Präferenzen soll die Ergebnisdarstellung vorgenommen werden.

4.1 Altersspezifische Urteilshomogenität

In beiden Evaluationsbereichen wurde für jede Altersgruppe Kendall's W' bestimmt und inferentiell gemäß der Chiquadrat-Verteilung beurteilt. Die Tabellen 3 und 4 enthalten die Ergebnisse dieser Berechnungen.

Tabelle 3

Beurteilungskonkordanzen in sieben Altersgruppen für die HUSSY-Evaluation

Parameter	Altersgruppe						
	08	10	12	15	20	40	60
Vpn	24	42	38	35	26	21	27
Kendall's W'	.099	.013	.014	.135	.191	.049	.094
E(W)	.042	.024	.026	.028	.038	.048	.037
χ^2	4.78	1.07	1.05	9.44*	9.92*	2.08	5.08

* $p < 0.05$, $df = 2$

Tabelle 4

Beurteilungskonkordanzen in sieben Altersgruppen für die
WELTNER-Evaluation

Parameter	Altersgruppe						
	08	10	12	15	20	40	60
Vpn	28	43	38	36	26	20	28
Kendall's W'	.053	.010	.058	.164	.294	.427	.061
E(W)	.036	.023	.026	.028	.038	.050	.036
χ^2	2.96	0.88	4.45	11.81*	15.28*	18.00*	3.43

* $p < 0.05$, $df = 2$

Neben dem Wert für W' befinden sich in den Tabellen Angaben über den Stichprobenumfang, den dadurch bedingten Erwartungswert für W sowie die Chiquadrat-Statistik für W'. Abbildung 2 veranschaulicht die Altersunterschiede in der Urteilshomogenität; die fett markierten Punkte repräsentieren signifikante Beurteilerkonkordanzen ($p < 0.05$).

Betrachtet man zunächst die Ergebnisse der HUSSY-Evaluation, ergeben sich in der Altersgruppe 15 und 20 statistisch bedeutsame Konkordanzen: offensichtlich nehmen die Personen dieser Altersgruppen eine jeweils weitgehend einheitliche Bewertung der drei Alternativen vor. Erstaunlicherweise ist der Trend in den jüngeren Altersgruppen nicht kontinuierlich ansteigend; ein "wannenförmiger" Verlauf mit dem Tiefpunkt bei 10- und 12jährigen läßt sich beobachten. Ein ähnlicher "Einbruch" liegt - zumindest tendenziell - auch bei den 40jährigen vor, die mit einem W'-Wert von 0.05 weit unterhalb der Konkordanz der 60jährigen liegen.

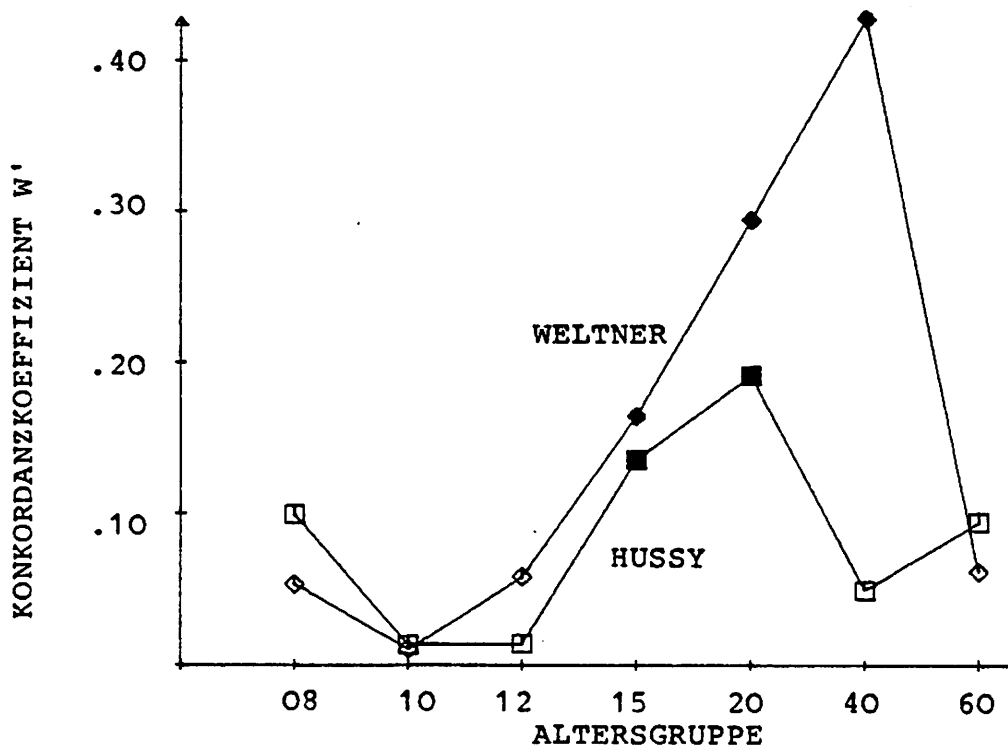


Abbildung 2. Konkordanz W' in sieben Altersstufen für zwei Paradigmen (gefüllte Punkte markieren auf dem 5%-Niveau signifikante Homogenität).

Zieht man im Vergleich dazu die Ergebnisse der WELTNER-Evaluation heran, bestätigt sich der Wanneneffekt bei 10- und 12jährigen, nicht aber der Abfall im Alter von 40 Jahren; ganz im Gegenteil zu den HUSSY-Befunden handelt es sich bei dem W' -Wert von 0.43 um die höchste überhaupt vorfindbare Konkordanz. Der Zuwachs an Konkordanz von 20 zu 40 Jahren liegt im Bereich des von 15 zu 20 beobachteten Wertes!

Dieser überraschende und nicht hypotesenkonforme Befund macht deutlich, daß die Ähnlichkeit von Evaluationskriterien bei 20- und 40jährigen über verschiedene Evaluationsaufgaben hinweg in Zweifel gezogen werden muß. Die Altersgruppen "8-15"

sowie "60" besitzen dagegen in beiden Aufgaben ähnliche (nämlich fehlende) Urteilshomogenität.

4.2 Altersspezifische Präferenzen

Die mit dem 2I-Test festgestellten Abweichungen von der Annahme gleichverteilter Rangvergaben gehen aus Tabelle 5 hervor.

Tabelle 5

Werte der 2I-Statistik getrennt für Altersgruppen und Problemtypen

Problemtyp	Altersgruppe						
	08	10	12	15	20	40	60
HUSSY	8.90	17.91**	13.85*	17.93**	16.83**	2.73	9.56
WELTNER	13.96*	8.89	10.31	14.60*	19.73**	21.99**	16.23**

* $p < 0.05$, $df = 5$

** $p < 0.01$, $df = 5$

Für die HUSSY-Evaluation bestehen bei 10-, 12-, 15- und 20-jährigen Personen spezifische Präferenzen. Während 10jährige eine Rangreihe präferieren, in der Alternative A in 22 von 38 Fällen auf den 2. Platz gesetzt wird (Rangreihen 213 und 231), die beiden Extrembeurteilungen jedoch schwanken (was zu einer niedrigen Konkordanz führt, vgl. Abbildung 2), bilden sich bei 15- und 20jährigen klare Präferenzen für die Rangreihe 231 aus (47% bzw. 52%). Die 12jährigen entscheiden sich stärker für die Rangfolge 312, wengleich 231 die nächst häufiger frequentierte Beurteilung darstellt. Auffällig ist die Tatsache, daß es unter 8-, 40- und 60jährigen nicht zu einer Bevorzugung einzelner Rangreihen kommt. Die Hypothese, daß im

Fall fehlender Beurteilungskriterien die Abfolge 123 wahrscheinlicher wird, läßt sich für diesen Aufgabentyp statistisch nicht stützen, obzwar deren Besetzung (7 von 19 Wahlen) die Beobachtungen der Versuchsleiter zumindest tendenziell decken.

Bei der WELTNER-Evaluation sind es 8-, 15-, 20-, 40- und 60jährige Probanden, bei denen bedeutsame Abweichungen von einer Gleichverteilung auftreten (vgl. Tabelle 5). Mit Ausnahme der 40jährigen entfallen die jeweils meisten Wahlen in jeder Altersgruppe auf die Evaluation 231 (vgl. Tabelle 2); die bei 40jährigen stärkere Präferenz besteht in der Rangreihe 321, d.h. in einer veränderten Bewertung der Alternativen A und B. Über den ersten Platz für Alternative C herrscht auch dort kein Zweifel. Obgleich 10- und 12jährige Probanden in dieser Evaluationsaufgabe keine statistisch auffälligen Abweichungen von einer Gleichverteilung aufweisen, erkennt man, daß bei gezielter Kollabierung des Datenkörpers (etwa derart, daß nur die Frequenzen ausgezählt werden, mit der eine bestimmte Alternative einen bestimmten Rangplatz erhält) sehr wohl weitere Effekte deutlich werden könnten, auf deren Nachweis hier verzichtet wird.

Untersucht man die Abweichungen von einer Beurteilung nach Zufall zwischen den beiden Altersgruppen, kann man nicht von einem konsistenten Verhalten sprechen: während 8-, 40- und 60jährige Personen im HUSSY-Teil eher nach Zufall urteilen, im WELTNER-Teil jedoch klare Präferenzen entwickeln, gilt das Umgekehrte für die 10jährigen. Der erste Befund - keine Präferenzen bei HUSSY, wohl aber bei WELTNER - entspricht der formulierten Hypothese unter Bezug auf die unterschiedliche

Schwierigkeit der Bögen. Der gegenteilige Befund überrascht dagegen und erlaubt keine befriedigende Erklärung.

5 Diskussion der Befunde

Die gefundenen Ergebnisse lassen sich wie folgt zusammenfassen. Erstens kommt es im Alter von 15 und 20 Jahren zu homogenen Beurteilungen der zu bewertenden Alternativen in beiden Evaluationsbögen. Darüberhinaus weisen 40jährige im WELTNER-Bogen übereinstimmende Bewertungen auf. Zweitens finden sich Abweichungen von einer nach Zufall erfolgenden Beurteilung bei 15- und 20jährigen in beiden Aufgaben, bei 10- und 12jährigen nur im HUSSY-Teil und bei 8-, 40- und 60jährigen nur im WELTNER-Teil. Schließlich bestehen zwischen beiden Evaluationstypen - je nach Altersgruppe - beträchtliche Unterschiede, die aufgabenspezifische Evaluatoren widerspiegeln; die Altersunterschiede sind beachtlich, weswegen die Auswertung von Evaluationsdaten ohne Einbezug des Altersfaktors unsinnig wäre.

Vor dem Hintergrund eines Verarbeitungsmodells, das kognitive Steuerungsprozesse vom Vorhandensein adäquater Evaluatoren abhängig macht, geben die vorgelegten Ergebnisse neue Impulse für eine "Theorie der Evaluatorentwicklung". Im Alter von 8 Jahren besitzen die Kinder noch keine Evaluatoren für schwierige Bewertungen. Diese lassen sich jedoch schon bei 10jährigen beobachten, auch wenn die mangelnde Konkordanz bei gleichzeitig vorhandener, klarer Präferenzstruktur darauf hinweist, daß - eventuell durch Übergänge zwischen verschiedenen Entwicklungsständen bedingt - zwei Subgruppen mit

gegenläufiger Urteilstendenz existieren, sowohl für leichte wie auch schwere Evaluationen. Mit 12 Jahren herrscht im einfachen Fall keine Einigkeit, während die schwierigere Evaluation deutliche, wenngleich andere Präferenzen hervorbringt als sie bei den übrigen Altersgruppen zu beobachten sind. Bei 15- und 20jährigen sehen wir die gelungene Ausbildung von Evaluatoren in einem konkordanten Urteilsverhalten mit klarer Präferenz bestätigt, die von 40jährigen nur noch im leichten Fall übertroffen wird. Für die schwierige Evaluation lassen sich weder bei 40- noch bei 60jährigen konkordante Urteile finden, auch Präferenzen gibt es nicht mehr. Die 60jährigen haben bei der einfacheren Evaluation zwar noch Präferenzen, aber diese sind nicht mehr einheitlich in der Altersgruppe verteilt, so daß auch hier ähnlich wie bei den 10jährigen die Vermutung einer Subgruppen-Struktur naheliegt.

Zwei mögliche Einwände gegen eine derartige Entwicklungsannahme seien kurz genannt, die momentan nicht zu widerlegen sind. Zum einen kann die Existenz eines Kohorteneffekts zu den gezeigten Ergebnissen geführt haben, zum anderen ist natürlich klar, daß in einem Prozeßmodell Evaluatoren nicht als statische Größen angesehen werden dürfen. Die Dynamik von Evaluationsprozessen in verschiedenen Altersbereichen haben wir nicht berücksichtigt.

6 Abschließende Bemerkungen

Die anfangs geäußerte Kritik am Auswertungsvorschlag von HUSSY & KÖHL (1982) soll angesichts unserer Befunde konstruktiv ausgedehnt werden. Die alternative Auswertung entspricht

zwar dem vorliegenden Datenmaterial, bietet aber nicht die wünschenswerte Tiefe der Interpretation in Hinblick auf die Erfassung potentiell verwendeter Evaluatoren. Wir schlagen deshalb vor, das Konzept der EG-Maße durch die Entwicklung eines geeigneten Evaluationsbogens zu ergänzen: durch die Erhöhung der Alternativenzahl, Kürzung des zu beurteilenden Folgenausschnitts und der vorherigen Bestimmung eindeutiger "objektiver" Rangreihen mag es gelingen, klarere Hypothesen über verwendete Evaluatoren zu bilden. Schon ein Bogen mit sechs Alternativen erlaubt 720 instruktionsgemäße Rangreihen und damit selbst bei umfangreichen Evaluator-Annahmen wohl genügend Fälle für eine eindeutige Zuordnung von Rangreihe und postuliertem Evaluator; den Versuch, aus dem Universum der nach Konzeption der EG-Maße möglichen Evaluationskriterien eine Subgruppe "markanter" Evaluationsaspekte auszuwählen, um durch diese (im Prinzip beliebige) Beschränkung den Umfang von zu bestimmenden "objektiven" Rangreihen zu begrenzen, lehnen wir dagegen ab, da der Einsatz nicht berücksichtigter Evaluatoren durch den Probanden nicht auszuschließen ist. Im übrigen bleibt festzuhalten, daß die Evaluationsaspekte, die unter transinformationsanalytischer Perspektive entwickelbar sind, keine erschöpfende Auswahl aus der Menge "natürlicher" Evaluatoren darstellen.

Ein weiterer konstruktiver Vorschlag bezieht sich auf eine Verbesserung des Skalierungsverfahrens zur Erfassung subjektiver Rangreihen. Mit der Methode des (eventuell unvollständigen) Paarvergleichs könnten die gewünschten Informationen bereits auf Rohdaten-Seite präziser erfaßt werden als es bisher

gelingt (vgl. den Einsatz dieser Methode bei SYDOW 1970).

Eine weitere Möglichkeit besteht darin, die vom Probanden vermuteten Rangreihen in Form von Aussagen etwa der Art "A ist dreimal besser als B" formulieren zu lassen.

Schließlich muß gefragt werden, ob nicht gänzlich andere, stärker prozeßorientierte Erhebungsverfahren zum Evaluationsverhalten (wie etwa Blickbewegungen, Entscheidungszeiten) dichter an das interessierende Phänomen heranführen. Darin sehen wir einen wichtigen Schritt der Weiterentwicklung.

Die hier vorgeschlagene Alternative - Bestimmung von Konkordanzen und Präferenzen innerhalb psychologisch sinnvoll unterscheidbarer Gruppen - bleibt natürlich auch bei neu zu entwickelnden Evaluationsbögen als Analyseform bestehen; ihre Interpretation kann durch den Rekurs auf EG-Maße verbessert werden. Diese waren im Fall des vorliegenden Evaluationsbogens nicht sinnvoll zu bestimmen, weshalb die frequenzorientierte Auswertung als mögliche Alternative den Vorzug erhielt.

Literatur

- BEATTY, J. 1982. Task-evoked pupillary responses, processing load, and the structure of processing resources. Psychological Bulletin 91, 276-292.
- BLÖSCHL, L. 1966. Kullbacks 2I-Test als ökonomische Alternative zur Chiquadrat-Probe. Psychologische Beiträge 9, 379-391.
- DÖRNER, D. 1976. Problemlösen als Informationsverarbeitung. Stuttgart: Kohlhammer.
- FUNKE, J. & HUSSY, W. 1979. Informationsverarbeitende Strukturen und Prozesse: Analysemöglichkeiten durch Problemlöseparadigmen. Trierer Psychologische Berichte 6, Heft 8.
- HUSSY, W. 1979. Informationsverarbeitende Strukturen und Prozesse: Versuch einer allgemein- und entwicklungspsychologischen Modellbildung. Trierer Psychologische Berichte 6, Heft 6.
- HUSSY, W. 1982. Komplexe menschliche Informationsverarbeitung: Das SPIV-Modell. Sprache und Kognition (im Druck).
- HUSSY, W. & KÖHL, W. 1982. Informationsverarbeitende Strukturen und Prozesse: Erfassung des kognitiven Aspekts der Evaluation. Trierer Psychologische Berichte 9, Heft 3.
- KU, H.H. 1963. A note on contingency tables involving zero frequencies and the 2I-test. Technometrics 5, 398-400.
- KULLBACK, S. 1959. Information theory and statistics. New York: Wiley.
- LIENERT, G.A. 1978. Verteilungsfreie Methoden in der Biostatistik. Band 2. Meisenheim: Hain.
- LOMPSCHER, J. (Ed.) 1972. Theoretische und experimentelle Untersuchungen zur Entwicklung geistiger Fähigkeiten. Berlin: Volk und Wissen.

- LÜER, G. 1973. Gesetzmäßige Denkabläufe beim Problemlösen.
Ein empirischer Beitrag für eine psychologische Theorie
der Entwicklung des Denkens. Weinheim: Beltz.
- MILLER, G.A., GALANTER, E. & PRIBRAM, K.H. 1960. Plans and the
structure of behavior. New York: Holt, Rinehart and Winston.
- NEWELL, A. & SIMON, H.A. 1972. Human problem solving. Englewood
Cliffs, N.J.: Prentice-Hall.
- PUTZ-OSTERLOH, W. 1981. Über die Beziehung zwischen Testin-
telligenz und Problemlöseerfolg. Zeitschrift für Psycho-
logie 189, 79-100.
- SYDOW, H. 1970. Metrische Charakterisierung von subjektiven
Problemzuständen in Lösungsprozessen. Zeitschrift für
Experimentelle und Angewandte Psychologie 17, 233-244.
- WINTERHOFF, P. 1980. Zum Zusammenhang von Blickbewegungen
und sprachlich-kognitiven Prozessen - ein Überblick.
Psychologische Rundschau 31, 261-276.

<u>Inhalt</u>	Seite
1 Problemstellung	2
2 Datenrückgriff	3
3 Zur Frage der Auswertungsstrategie	5
3.1 Vergleich subjektiver und "objektiver" Rangreihen	7
3.2 Alternative Vorgehensweise	10
3.3 Verwendete Methoden	11
3.4 Hypothesen	11
4 Ergebnisse der vereinfachten Datenanalyse	13
4.1 Altersspezifische Urteilshomogenität	13
4.2 Altersspezifische Präferenzen	16
5 Diskussion der Befunde	18
6 Abschließende Bemerkungen	19
Literatur	22
Inhalt	24

Anschrift des Verfassers

Joachim Funke
Universität Trier
Fachbereich I - Psychologie

5500 Trier

Bisher erschienene Hefte des Jahrganges 9 (1982)

- BREDENKAMP, J. 1982. Psychophysikalische Analysen zur Erklärung des Entstehens und rätselhaften Verschwindens von Wahrnehmungstäuschungen. Trierer Psychologische Berichte 9, Heft 1.
- BARTUSSEK, D. 1982. Modelle der Testfairness und Selektionsfairness. Trierer Psychologische Berichte 9, Heft 2.
- HUSSY, W. & KÖHL, W. 1982. Informationsverarbeitende Strukturen und Prozesse: Erfassung des kognitiven Aspekts von Evaluationen. Trierer Psychologische Berichte 9, Heft 3.
- KRAMPEN, G. & MORY, M. 1982. Zur kognitiv-emotionalen Verarbeitung schlechter Zensuren bei Schülern. Trierer Psychologische Berichte 9, Heft 4.
- KRAMPEN, G. & VON EYE, A. 1982. Informationstheoretische Modelle kausaler Attribution. Trierer Psychologische Berichte 9, Heft 5.
- BRANDTSTÄDTER, J. & VON EYE, A. 1982. Aussagenlogische Analyse von Kontingenztafeln: I. Methodologische Vorüberlegungen. Trierer Psychologische Berichte 9, Heft 6.
- BECKER, J.H. 1982. Zur Versorgung praktisch tätiger Psychologen mit fachlichen Informationen. Informationsgewohnheiten und -bedürfnisse von Psychologen in Arbeitsämtern, Marktforschungsinstituten und Technischen Überwachungsvereinen. Trierer Psychologische Berichte 9, Heft 7.
- DAHL, U., MONTADA, L. & SCHMITT, M. 1982. Hilfsbereitschaft als Personmerkmal. Trierer Psychologische Berichte 9, Heft 8.
- DALBERT, C. & MONTADA, L. 1982. Vorurteile und Gerechtigkeit in der Beurteilung von Straftaten. Eine Untersuchung zur Verantwortlichkeitsattribution. Trierer Psychologische Berichte 9, Heft 9.

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