Archiv für die gesamte Psychologie, Bd. 118, S. 47—72 (1966)

Aus dem Psychologischen Institut der Universität Bonn
(Direktor: Prof. Dr. H. THOMAE)

Über die Zuverlässigkeit von Verhaltensbeurteilungen durch Rating-Skalen*

Von

Hans Dieter Schmidt

(Angenommen am 18. Oktober 1965)

I.

"Though second-best, ratings have their uses." Dieser Satz G. W. ALLPORTS (1961, S. 421) kennzeichnet treffend das Problem der Anwendung von Rating-Skalen: Das „theoretische" Mißtrauen gegenüber subjektiven Beurteilungsskalen mit dem Niveau von Ordinalskalen (STEVENS, 1960) wird meist verdrängt durch die „praktische" Einsicht, ihre große Beliebtheit (GUILFORD, 1954) und ihren unbestrittenen Nutzen im Bereiche der angewandten Psychologie (z. B. KNAUFT, 1947) und verschiedenen Gebieten der Grundlagenforschung (vgl. die Längsschnittuntersuchungen THOMAES, 1955).

Die Motivationen zur Anwendung von Ratings in den Verhaltenswissenschaften lassen sich zumeist wie folgt beschreiben: 1. Man wünscht Verhaltensbeurteilungen in quantitativer Form, sei es aus Gründen der Datenverarbeitung, sei es, weil man überzeugt ist, daß die zu betrachtenden Merkmale sich — wenn auch noch so primitiv — skalieren lassen. 2. Man wünscht eigentlich keine subjektiven Beurteilungen, sondern erstrebt möglichst objektive Aussagen über menschliches Verhalten (z. B. Testdaten); die zu betrachtenden Eigenschaften sind jedoch entweder einer „objektiven" Erfassung grundsätzlich schwer zugänglich, oder es liegen (noch) keine besseren Verfahren zur Registrierung des betreffenden Verhaltens vor. In diesen Fällen würde man auf die Rating-Methode zurückgreifen.

---

* Zur Terminologie: Wegen der Kürze und Prägnanz sollen „rating" und „rating scale" mit „Rating" und „Rating-Skala" übersetzt werden; „Beurteilungsskala" ist ein zu umständliches Wort. „Rater" wird alternierend mit „Beurteiler", „Ratee" abwechselnd mit „Individuum", „trait" als „Eigenschaft" angegeben.

Daß die Rating-Methode als Hilfsmittel der Verhaltensbeurteilung Verwendung findet, bestimmen also nicht zuletzt die Anforderungen der Praxis. Daß sie sich dabei als relativ „empfindliches" Instrument erweist, dem man — zumindest bei Verwendung mehrerer Rater — genaue Beurteilungen zutraut, mag folgendes Experiment zeigen:

40 Studenten wurden nach Zufall in zwei Gruppen aufgeteilt. In getrennten Sitzungen, aber unter weitgehend konstanten äußeren Bedingungen, wurde den Versuchsgruppen A ($n = 18$) und B ($n = 22$) ein Film gezeigt, in dem eine 16jährige Schülerin die „Bauprobe" nach KLEMM ausführte.

Gruppe A wurde der Film mit der korrekten Geschwindigkeit von 24 Bildern pro Sekunde dargeboten (Dauer ca. 12 Min.), *Gruppe B* mit der zu langsamen Darbietungsgeschwindigkeit von 12 B/sec (Dauer ca. 27 Min.).

Beide Gruppen hatten das im Film gezeigte Verhalten durch 12 graphische Rating-Skalen (Linien von 120 mm Länge) zu beurteilen. Es wurde angenommen, daß die mittleren Ratings der Gruppen A und B erhebliche Beurteilungs-Unterschiede aufweisen würden.

Tatsächlich zeigten sich in allen verwendeten Skalen deutliche Mittelwertsunterschiede, die jedoch nur in drei Fällen („Antrieb", „Anregbarkeit", „Mitschwingungsfähigkeit" nach THOMAE, 1955) auf dem 1%-Niveau gesichert waren. Diese Eigenschaften scheinen eine starke Aktivitäts- bzw. Bewegungskomponente zu enthalten; daher wurde in einem weiteren Versuch die Hypothese aufgestellt, daß Eigenschaften, die offenbar eine starke Aktivitätskomponente enthalten, bei „zu schneller" Filmgeschwindigkeit im Rating höher, und Eigenschaften mit anscheinend starker Steuerungskomponente („Angepaßtheit", „Sicherheit", „Steuerung" nach THOMAE, 1955) niedriger beurteilt werden.

Zwei vergleichbare Gruppen von Studenten sahen einen Film von ca. 5 Minuten Dauer, auf dem ein 6jähriges Mädchen mit Spielmaterial des „Sceno-Tests" (v. STAABS) spielte. *Gruppe A* ($n = 12$) wurde der Film mit derjenigen Bildgeschwindigkeit dargeboten, mit der auch die Aufnahme erfolgt war (16 B/sec); *Gruppe B* ($n = 11$) sah den gleichen Film mit nur 4 B/sec größerer Geschwindigkeit. Insgesamt waren 10 Verhaltensmerkmale zu beurteilen. Es zeigte sich, daß bereits der geringfügige Unterschied in der Darbietungsweise des Films (4 B/sec) in der Beurteilung durch Rating-Skalen seinen deutlichen (quantitativen) Niederschlag fand: 3 von 10 Mittelwertsdifferenzen waren auf dem 5%-Niveau signifikant, bei 7 weiteren zeigten sich deutliche Tendenzen zu einer unterschiedlichen Beurteilung der Vp durch die Rater der Gruppen A und B. Erwartungsgemäß zeigten sich unter der Bedingung B ausnahmslos höhere „Aktivitäts"-Werte, während bei den ansonsten niedrigeren „Steuerungs"-Werten ein Merkmal eine Ausnahme bildete: „Angepaßtheit" war unter Bedingung B nicht signifikant erhöht.

Selbstverständlich ist der Grad der „Empfindlichkeit" eines Beurteilungsinstruments gegenüber Veränderungen der Reizkonstellation lediglich dazu geeignet, Evidenz für den Nutzen der Methode hervorzurufen. Entscheidend ist allein, wie zuverlässig und zutreffend sich menschliches Verhalten durch Rating-Skalen beurteilen läßt. Mit dem Problem der Zuverlässigkeit (Reliabilität) von Ratings sollen sich denn auch die folgenden Überlegungen und Experimente beschäftigen. Abhandlungen zur *Geschichte* und *Bedeutung* der Rating-Skalen finden sich bei ANDREGG (1951), GUILFORD (1954) und SCHMIDT (1965). Über die verschiedenen gebräuchlichen *Formen* von Rating-Skalen orientiert GUILFORD (1954), dessen Systematik teilweise von HASEMANN (1964) wiedergegeben wird.

II. Das Problem der Zuverlässigkeit von Ratings

Ehe man die Gültigkeit der Rating-Methode für bestimmte Zwecke prüft, muß die Frage entschieden werden, in welchem Maße man mit Hilfe von Rating-Skalen das, was man mit ihnen beurteilt, genau beurteilen kann. Dabei kann man sich ein Rating-Urteil ebenso wie einen Testscore als additiv zusammengesetzt aus einem „wahren" Wert und einer Fehlerkomponente vorstellen, die beide auf die gleiche Skala, nämlich die Rating-Skala bezogen sind. Der Grad der Zuverlässigkeit von Ratings ist dann durch den Anteil der „wahren" Varianz an der Gesamtvarianz bzw. durch das Ausmaß des Fehlens von Fehlervarianz bestimmt (zum Reliabilitätsbegriff vgl. GUILFORD, 1954; LIENERT, 1961).

Zur Schätzung der Reliabilität von Testdaten werden zumeist die hinlänglich bekannten drei Methoden (Retest-, Split-half- und Paralleltestmethode; vgl. LIENERT, 1961, S. 210ff.) angewendet. Die *Wiederholungsmethode,* deren schwacher Punkt ein möglicher Übungseffekt beim Probanden ist, scheint für die Schätzung der Reliabilität von Ratings noch weniger geeignet zu sein; die Gründe dafür dürften vor allem im Gegenstand der Beurteilung liegen (das zu beurteilende Verhalten ist zumeist „offener" und schwerer wiederholbar als das beim Test gezeigte Verhalten). Dennoch wurde die Rerating-Methode zuweilen verwendet, so bei TSCHECHTELIN (1944) und STOCKFORD u. BISSELL (1949). Auch die *Halbierungsmethode* und *Parallelverfahren* können — vorwiegend aus Gründen der Konstruktion — nicht in üblicher Weise zur Schätzung der Reliabilität von Ratings verwendet werden.

Stattdessen wird in nahezu allen Forschungsarbeiten über die Genauigkeit der Rating-Methode die sogenannte *Rater-Reliabilität* bestimmt, d. h. die Zuverlässigkeit von Ratings soll durch die Korrelation zwischen den Beurteilungen zweier oder mehrerer Rater ausgedrückt werden. Dies geschieht zuweilen durch die Angabe eines Korrelationskoeffizienten zwischen den Ratings zweier Beurteiler, teils durch Berechnung der Korrelation zwischen den mittleren Ratings zweier Rater-Gruppen, teils durch die Ermittlung des mittleren Korrelationskoeffizienten aus einer Matrix von *Interkorrelationen* mehrerer Rater (MINER, 1917; HOLLINGWORTH, 1922; CONKLIN u. SUTHERLAND, 1923; MARSH u. PERRIN, 1925; SHEN, 1925; KORNHAUSER, 1926; REMMERS, 1934; REYMERT u. KOHN, 1938; TIFFIN, 1942; CARTER, 1945; ALLPORT, 1949; STOCKFORD u. BISSELL, 1949; GUETZKOW, 1950; LEE u. BURNHAM, 1963). Auch in den Arbeiten von E. K. TAYLOR und seinen Mitarbeiten wurden Interrater-Korrelationen berechnet, zum Teil durch Korrelationen zweier Rater-Populationen, nämlich direkter und indirekter Vorgesetzter von Büroangestellten (TAYLOR u. HASTMANN, 1956; BARRETT, TAYLOR, PARKER u. MARTENS, 1958; TAYLOR, BARRETT, PARKER u. MARTENS, 1959; TAYLOR, PARKER u. FORD, 1959). A. W. BENDIG schätzte die Rater-Reliabilität durch den Grad, mit dem Beurteiler zwischen verschiedenen Reizen unterscheiden können (BENDIG, 1954a; 1954b; 1955a; 1955b; 1957a; 1957b; BENDIG u. SPRAGUE, 1954). Auch GUILFORD (1954) gibt für die Bestimmung der Reliabilität von Ratings Statistiken der subjektiven Übereinstimmung an. Als

geeignet werden die *Intraclass*-Korrelation nach EBEL (1951), die *Correlation-ratio*-Methode nach HOLLINGWORTH (1913) — ein mittleres Rangreihenkorrelationsmaß — und die Anwendung der *Spearman-Brown*-Formel auf die Rater, wie sie CLARK (1935) anwendete, vorgeschlagen. Zu ergänzen wäre hier, daß sich als wohl ökonomischstes Maß der Rater-Übereinstimmung, dessen Prüfung gegen H_0 über die Chi²-Verteilung erfolgt, der auf Rangplatzinformationen basierende *Konkordanzkoeffizient* von KENDALL (LIENERT, 1962) vorschlagen läßt, da dieses Maß eine Berechnung der Übereinstimmung mehrerer, von verschiedenen Beurteilern erstellter Rangreihen erlaubt.

Die *Übereinstimmungskoeffizienten* zwischen Ratern, die immer wieder zur Bestimmung der Reliabilität von Ratings angegeben wurden, erreichen zwar selten die Höhe von Reliabilitätskoeffizienten, wie sie für Leistungstests erforderlich sind, zeigen aber zuweilen eine beachtliche Übereinstimmung zwischen den verschiedenen Rating-Beurteilungen an. Der mittlere Korrelationskoeffizient (Median), ermittelt aus 35 verschiedenen Reliabilitätsberechnungen von 16 Autoren aus den Jahren 1923 bis 1963 liegt bei 0,59, bei einer Schwankungsbreite von 0,04 bis 0,91. Die vorliegenden Ergebnisse sind sehr unterschiedlich, je nach Art des beurteilten Verhaltens und nach der Anzahl der Rater. Eine geraffte Übersicht über die Abhängigkeit der Koeffizienten von Rater-Zahl, Art der beurteilten Eigenschaften, Eigenart der Beurteiler, Form von Rating-Skalen und verschiedenen Situations-Bedingungen findet sich an anderer Stelle (SCHMIDT, 1965).

Gegen das geschilderte Vorgehen, nämlich die Zuverlässigkeit (Reliabilität) von Ratings durch die *Übereinstimmung* der Beurteiler (Interrater-Korrelation) auszudrücken, müssen nun schwere Bedenken erhoben werden:

1. Die Berechnung von Übereinstimmungen zwischen Ratern führt allenfalls zu Maßen der *Objektivität* von Ratings. Die Objektivität als „der mittlere Grad, in dem die Beurteilungen mehrerer Auswerter über die Testleistung einer Stichprobe von Probanden miteinander korrelieren" (LIENERT, 1961, S. 13) ist zwar eine wichtige Voraussetzung für die Güte eines Verfahrens, ist aber nicht unbedingt geeignet, das Verhältnis von echter Varianz und Fehlervarianz von Daten (also die Reliabilität des Verfahrens) zu schätzen. Wenn mehrere Personen über einen bestimmten Menschen ähnliche Urteile abgeben, so könnte dies ja auch daran liegen, daß alle Beurteiler einem ähnlichen Fehlurteil verfallen sind, insbesondere, da bestimmte Typen von Beurteilungsfehlern von Psychologen immer wieder beschrieben worden sind. Was durch erhöhte Konkordanz also erhöht wird, ist zunächst nur etwas, das man als „face reliability" bezeichnen könnte.

2. Es ist evident, daß Interrater-Korrelationen nicht unabhängig von der Interkorrelation der Rating-Skalen sind. Die Korrelation der einzelnen Eigenschaften untereinander geht aber offenbar zum großen Teil auf den prominentesten systematischen Beurteilungsfehler, den *Halo-Effekt* zurück; bei vielen Autoren gilt die Höhe der Interkorrelation sogar als direkte Schätzung dieses Fehlers. So zeigte sich in der Arbeit von GUILFORD, CHRISTENSEN, TAAFE u.

WILSON (1962), deren programmatischer Titel „Ratings should be scrutinized" dem Leitgedanken auch der vorliegenden Arbeit entspricht, daß Rater die verschiedenen zu beurteilenden Eigenschaften stets sehr stark miteinander verwechselten bzw. sich bei ihren Beurteilungen von einem oder zwei ausgewählten Faktoren leiten ließen (obgleich die acht Skalen bereits relativ unabhängigen Faktoren entsprachen). Die hohen Rating-Interkorrelationen (durchschnittlich 0,70 bei einer Streuungsbreite von 0,59 bis 0,89) werden von GUILFORD und seinen Mitarbeitern als durch systematische Rating-Fehler determiniert und als ungeeignet angesehen, einer Faktorenanalyse als Grundlage zu dienen: "The factors obtained from trait ratings are likely to reflect what is in the minds of the raters than basic traits of personality" (1962, S. 445).

Hohe Korrelationskoeffizienten können also nicht als Zeichen für die relative Abwesenheit von Beurteilungsfehlern angesehen werden — dies sollte aber die Funktion von Rating-Reliabilitätskoeffizienten sein —, da sie teilweise selbst auf Beurteilungsfehlern beruhen. (Setzt man sich über GUILFORDS Rat hinweg und faktorisiert Ratings gar, so ergibt sich stets das gleiche Bild: "Multiple scale ratings have always boiled down to two or three orthogonal factors"; TAYLOR u. HASTMANN, 1956, S. 186). Hohe Übereinstimmungen zwischen Ratern nutzen uns nichts: Wir wissen nicht, in welchem Maße sie auf dem Halo-Effekt beruhen.

Die folgende Überlegung muß sich hier anschließen, wenn man es nicht aufgeben will, systematische Fehler beim Rating weitgehend zu kontrollieren: Wenn es ein Maß der relativen Abwesenheit von systematischen Beurteilungsfehlern beim Rating (constant errors) gäbe, so wäre dieses Maß wahrscheinlich geeigneter zur Schätzung der Zuverlässigkeit von Ratings als die üblichen Maße der Rater-Übereinstimmung.

Aus der Psychologie des Beurteilens von Mitmenschen sind seit langem gewisse Fehlerarten bekannt, die immer wieder auftreten, wenn Beurteiler Individuen hinsichtlich einer Mehrzahl von Eigenschaften quantitativ beurteilen (nähere Beschreibung bei GUILFORD, 1954; GRAUMANN, 1960; HASEMANN, 1964; SCHMIDT, 1965). Während beim error of leniency, error of central tendency, contrast error und proximity error hauptsächlich gewisse Einstellungen und „sets" eine Rolle spielen dürften, scheinen der logical error und der Halo-Effekt vor allem durch eine allgemeine Tendenz zum Organisieren, beim Diagnostiker sicherlich auch durch das, was HÖRMANN (1964) als „Sucht nach Ganzheitlichkeit" bezeichnet, zustande zu kommen. Zieht man in Betracht, daß verschiedene Fehlerarten sich in ihrer Bedeutung überlappen und in der Praxis aufheben dürften, so wird die überragende Bedeutung des Halo-Effekts, der sich vom logischen Fehler nicht immer streng trennen läßt, deutlich. Dem Halo-Effekt, der operational der Tendenz des Raters entspricht, einen Ratee in allen Eigenschaften ähnlich zu beurteilen, wird auch in der Literatur der weitaus meiste Raum gewidmet. GUILFORD (1959) bezeichnet ihn auch als „Fehler der Wechselwirkung zwischen Beurteiler und Beurteiltem" (rater-ratee interaction error).

Wenn sich also insbesondere der Halo-Effekt, der zu den unecht hohen Korrelationen von beurteilten Eigenschaften führt, statistisch bestimmen ließe, so könnte die Angabe seines Betrages bzw. seine relative Abwesenheit, d. h. das

Ausmaß, in dem er vermieden wurde, als eine gute Schätzung der Zuverlässigkeit von Ratings gelten. Diese Schätzung wäre dann wahrscheinlich besser als die Angabe des Grades, in dem Beurteiler übereinstimmen, da diese Übereinstimmung selbst in hohem Maße auf dem Halo-Effekt beruhen kann.

Zu einem anderen Zweck, nämlich zur statistischen Korrektur von Original-Ratings, hatte GUILFORD (1954) bereits eine Methode, das Ausmaß des Halo-Effekts zu messen, vorgeschlagen[1]. Es handelt sich um eine Wechselwirkungsvarianzanalyse zwischen Ratern und Ratees, und zwar ohne Berücksichtigung der Unterschiede zwischen den beurteilten Eigenschaften. Unter der Voraussetzung, daß die Daten mehrerer Beurteiler über mehrere beurteilte Personen zur Verfügung stehen, läßt sich der Grad der statistischen Wechselwirkung zwischen beiden Variablen als Ausdruck des relativen Halo-Effekts unter verschiedenen Bedingungen ermitteln. Dieses Verfahren wurde, wenn auch nicht im Bereich der unmittelbaren Verhaltensbeurteilung, von JOHNSON u. VIDULICH (1956) mit Erfolg angewendet:

> In ihrer Arbeit „Experimental manipulation of the halo effect" schufen die Autoren für die Beurteilung von 5 bekannten Persönlichkeiten (Queen Elizabeth, Senator McCarthy, Sir W. Churchill, Mrs. E. Roosevelt, Papst Pius XII.) nach 5 Eigenschaften (Intelligenz, persönliche Erscheinung, Freundlichkeit, Mut, Nützlichkeit) „Maximalisierungs-Bedingungen" und „Minimalisierungs-Bedingungen" für das Auftreten des Halo-Effekts:
>
> Eine Gruppe von 18 Studenten beurteilte alle Individuen in jeweils *einer* Eigenschaft an jedem Experimentiertag (Minimalisierungs-Bedingung), eine gleich große Gruppe beurteilte an jedem Tag jeweils ein Individuum in *allen* geforderten Eigenschaften (Maximalisierungs-Bedingung). Die Varianz, die der Interaktion zwischen Ratern und Ratees entspricht und als Wirkung des Halo-Effekts bezeichnet werden kann, war in beiden Fällen gering; unter Maximalisierungsbedingungen war sie aber auf dem 1%-Niveau signifikant, während sie unter Minimalisierungsbedingungen nicht das 5%-Niveau erreichte.

In dieser Weise soll auch im experimentellen Teil der vorliegenden Arbeit verfahren werden. Wenn sich bei einer Varianzanalyse von Ratings eine erhebliche Restvarianz ergäbe, die statistisch der Interaktion zwischen Ratern und Ratees entspricht, dann wären die Beurteilungen in hohem Maße unter Beteiligung eines Halo-Effekts zustande gekommen; diesen Ratings käme eine geringere Zuverlässigkeit zu als solchen, bei denen die Fehlervarianz vermindert und somit der Betrag der „wahren" Varianz im Verhältnis zur Gesamtvarianz erhöht ist. Auf diese Weise ergibt sich zwar kein exaktes Maß der Zuverlässigkeit, kein Reliabilitätskoeffizient; es lassen sich jedoch vergleichende Aussagen über die Zuverlässigkeit von Ratings unter verschiedenen *Bedingungen* machen.

---

[1] Das Verfahren berücksichtigt nicht nur eine Kontrolle des Halo-Effekts, sondern auch eine statistische Korrektur der rater-trait-Interaktion und der ratee-trait-Wechselwirkung. Transformationen von Original-Ratings (vgl. auch SCHNEEWIND, 1965) sind aber sehr zeitraubend und heben damit einen der größten Vorzüge der Rating-Methode wieder auf.

III. Experimenteller Teil: Die Größe des Halo-Effekts unter verschiedenen Bedingungen

Experiment I

Es soll untersucht werden, ob sich die Wirkung des Halo-Effekts durch die Art der *Instruktion* der Rater vermindern läßt.

Methode

Eine Gruppe von 42 Studenten beobachtete nacheinander 3 Vpn (zwei weiblich, eine männlich) beim Zusammensetzen eines SCHULZschen Pumpwerks. Jeweils im Anschluß an die etwa halbstündlichen Verhaltensbeobachtungen nahmen die Beurteiler Ratings an Hand von 10 Rating-Skalen vor:

„Angepaßtheit", „Anregbarkeit", „Antrieb", „Mitschwingungsfähigkeit", „Sicherheit", „Steuerung", „Stimmung" in modifizierter Form nach THOMAE (1955), ferner „Äußere Erscheinung" („abstoßend-häßliches Äußeres" bis „überaus ästhetisches Äußeres"), „Körperliche Gesundheit" („Eindruck körperlichen Verfalls" bis „Eindruck körperlichen Bestzustandes") und „Originalität" („ohne jede Originalität, totale Einfallsarmut" bis „überaus große Originalität, Absonderlichkeit"). Die Skalen wurden auf DIN-A5-Blättern, in randomisierter Reihenfolge zu Blocks gebunden, dargeboten; es handelte sich um 9stufige, numerische Skalen; sie waren mit ihrem Titel benannt, schriftlich definiert und enthielten adjektivische Beschreibungen zu jedem der 9 Skalenpunkte.

Nach Zufall waren die 42 Beurteiler in 4 Gruppen aufgeteilt worden: *Gruppe A* ($n = 10$) erhielt zusätzlich zum Rating-Block eine Instruktion, die Informationen über die wichtigsten systematischen Beurteilungsfehler (constant errors) enthielt:

„Der Wert von Ratings hängt davon ab, ein wie gutes Instrument quantitativer Beobachtung und Beurteilung der menschliche Betrachter ist, d. h. wie objektiv und zuverlässig er beurteilen kann.

Ein *Fehler*, dem fast jeder Beurteiler (Rater) unterliegt, ist der sogenannte Halo-Effekt oder ‚Hof-Effekt'. Er besteht darin, daß der Beurteiler sämtliche Eigenschaftsbeurteilungen (Ratings), die er vorzunehmen hat, einem allgemeinen Eindruck unterordnet, den er von dem Beurteilten (Ratee) hat.

Das wirkt sich derart aus, daß die verschiedenen Eigenschaften, die beurteilt (geratet) werden sollen, in der Beurteilung alle sehr ähnlich bzw. gleichartig ausfallen. So kommt es vor, daß ein Ratee in allen Ratings immer sehr hoch oder sehr niedrig oder immer in der Mitte liegt.

Jeder Rater sollte unbedingt versuchen, diesen Fehler nach Möglichkeit zu vermeiden.

Ein weiterer Fehler, dem fast jeder Rater unterliegt, ist der sogenannte Leniency-Effekt oder ‚Milde'-Effekt. Er besteht darin, daß der Rater den Ratee höher bzw. günstiger einschätzt, als er eigentlich sollte. Zum Beispiel beurteilt jemand jemanden milder, wenn er ihn gut kennt oder ihn sympathisch findet. Auch das Gegenteil kommt vor, daß nämlich bestimmte Rater ihrer Ratees zu ‚hart' beurteilen; häufiger jedoch ist das zu „milde" Beurteilen.

Ein dritter Fehler, dem viele Rater unterliegen, ist der sogenannte error of central tendency oder ‚Fehler der zentralen Tendenz'. Er besteht darin, daß sich der Beurteiler scheut, extreme Urteile (Ratings) abzugeben, auch wenn dies eigentlich angebracht wäre;

auf diese Weise erhalten fast alle Individuen (Ratees) in allen zu beurteilenden Eigenschaften mittlere Beurteilungen (Zentrum der Skala).

Jeder Rater sollten unbedingt versuchen, diese Fehler nach Möglichkeit zu vermeiden."

Gruppe B ($n = 9$) erhielt lediglich denjenigen Teil der Instruktion, der sich auf den *Halo-Effekt* bezieht und dazu auffordert, Halo-Effekte beim Rating zu vermeiden.

Gruppe C ($n = 10$) erhielt ebenfalls einen Begleittext, doch enthielt dieser nur einige sehr allgemein gehaltene Wendungen in bezug auf die Schwierigkeit, objektiv zu urteilen; irgendwelche brauchbaren Anweisungen werden überhaupt nicht gegeben:

„Der Wert von Ratings hängt davon ab, ein wie gutes Instrument quantitativer Beobachtung und Beurteilung der menschliche Betrachter ist, d. h. wie objektiv und zuverlässig er beurteilen kann.

Leider zeigt es sich immer wieder, daß es menschlichen Beobachtern sehr schwer fällt, ganz objektiv zu urteilen, gerade bei Beurteilungen (Ratings), die den Beurteiler (Rater) vielleicht ein wenig überfordern. Der Mensch unterliegt — im Gegensatz etwa zu bestimmten Registrier-Maschinen — ständig bestimmten *Fehlern*, die zum größten Teil in ihm selbst begründet sind. Dies sollte sich jeder Beurteiler menschlicher Verhaltensweisen immer wieder vor Augen halten.

Deshalb sollte jeder Rater sich ganz besonders bemühen, objektiv zu urteilen und fehlerhafte Verzerrungen nach Möglichkeit zu vermeiden."

Gruppe D ($n = 13$) erhielt keinerlei zusätzliche Instruktion zu den Rating-Skalen.

Hypothese

Unter den vier Versuchsbedingungen A, B, C, D werden sich verschieden große Schätzungen des Halo-Effekts ergeben. Die Größe der Restvarianz zwischen Ratern und Ratees wird mit dem Grad der Instruktions-Information variieren; sie wird demnach bei Gruppe A am geringsten sein und bei der Kontrollgruppe (C) am größten.

Ergebnisse

Die Ratings der vier Beurteilergruppen über die drei Vpn wurden (unter Außerachtlassung der Unterschiede zwischen den 10 Skalen) vier verschiedenen Varianzanalysen unterzogen. Es wurde jeweils die Interaktionsvarianz zwischen Ratern und Ratees ermittelt. Jeder Beurteiler lieferte 3mal 10 Ratings, so daß der Wechselwirkungsvarianzanalyse 300 (A), 270 (B), 300 (C) und 390 (D) Ratings zugrundelagen. Bei diesem wie auch bei den folgenden Planversuchen war die Voraussetzung der Varianzenhomogenität zumindest näherungsweise, die der Normalität der Daten ausschließlich gegeben. Es ergaben sich folgende Resultate:

Gruppe A (Information über Halo, Leniency und Central tendency)

Quelle der Varianz	Quadratsumme	*df*	Varianz	*F*	*p*
Zw. Ratern (*R*)	108,944	9	13,104	5,655	< 0,01
Zw. Ratees (*I*)	65,890	2	32,945	14,218	< 0,01
Interaktion (*R* × *I*)	70,966	18	3,942	*1,701*	> *0,05*
Innerhalb	625,600	270	2,317		
Total	871,400	299			

Gruppe B (Information über den Halo-Effekt)

Quelle	Quadratsumme	*df*	Varianz	*F*	*p*
Zw. Ratern (*R*)	17,206	8	2,150	1,013	> 0,05
Zw. Ratees (*I*)	45,696	2	22,848	10,767	< 0,01
Interaktion (*R* × *I*)	37,839	16	2,364	*1,114*	> *0,05*
Innerhalb	515,700	243	2,122		
Total	616,441	269			

Gruppe C (allgemeingehaltene Erklärung)

Quelle	Quadratsumme	*df*	Varianz	*F*	*p*
Zw. Ratern (*R*)	16,288	9	1,803	1,402	> 0,05
Zw. Ratees (*I*)	59,540	2	29,770	23,149	< 0,01
Interaktion (*R* × *I*)	94,462	18	5,247	*4,080*	< *0,01*
Innerhalb	347,300	270	1,286		
Total	517,530	299			

Gruppe D (keine Information)

Quelle	Quadratsumme	*df*	Varianz	*F*	*p*
Zw. Ratern (*R*)	14,854	12	1,237	1,428	> 0,05
Zw. Ratees (*I*)	66,311	2	33,155	38,285	< 0,01
Interaktion (*R* × *I*)	114,835	24	4,784	*5,524*	< *0,01*
Innerhalb	304,190	351	0,866		
Total	500,190	389			

Unter den Bedingungen A und B zeigen sich nur unbedeutende, unter den Bedingungen C und D dagegen signifikante Wechselwirkungs-Anteile. Die größte Interaktion zeigt sich erwartungsgemäß bei der Kontrollgruppe. Abweichend von der Hypothese tritt jedoch unter der Bedingung B die geringste Restvarianz auf; hier zeigt sich auch eine wünschenswert unbedeutende Varianz zwischen den Beurteilern.

Zur Kontrolle dieses Resultats wurden zusätzlich *Übereinstimmungskoeffizienten* berechnet, und zwar zunächst pro Versuchsbedingung und Ratingskala.

Die 4 Mittelwerte der KENDALLschen Konkordanzmaße (W) für jeweils 10 Rating-Skalen betragen

für A: 0,219
für B: 0,253
für C: 0,254
für D: 0,281

Berechnet man — getrennt nach Ratee — die Konkordanz der Beurteiler hinsichtlich aller 10 Eigenschaften für die beiden Bedingungen B und D, so ergibt sich folgendes Bild:

	Bedingung B (Erläuterung des Halo-Effekts)	*Bedingung C* (Rating ohne Zusatzinstrukt.)
Ratee 1	0,1312	0,1832 *
Ratee 2	0,1632	0,3179 ***
Ratee 3	0,0684	0,2129 *

(* $p < 0{,}05$; ** $p < 0{,}01$; *** $p < 0{,}001$.)

Demnach zeigt sich, daß zwischen denjenigen Beurteilern, die am stärksten dem Halo-Effekt unterliegen, auch die größte Übereinstimmung herrscht!

Experiment II

Es soll geklärt werden, ob unter einem gewissen seelischen Druck (stress) zuverlässigere Ratings zustande kommen bzw. ob besonders leistungsmotivierte Rater weniger einem Halo-Effekt unterliegen. Ferner soll erforscht werden, ob sich Anhaltspunkte dafür ergeben, daß bestimmte Arten von Eigenschaften mit geringerem Halo-Effekt beurteilt werden als andere.

Methode

28 Studenten beobachteten, wie 4 Vpn (zwei männliche und zwei weibliche) nacheinander ca. 30 Minuten lang an einer mechanisch-technischen Aufgabe, dem MOEDEschen Montagebrett arbeiteten. (Wie bei allen Versuchen dieser Art, wurde ihre Kommunikation mit dem Versuchsleiter in die Beurteilung miteinbezogen.) Mit graphischen Rating-Skalen („Millimeter-Skalen" von 120 mm Länge, deren Enden mit adjektivischen Verhaltensbeschreibungen „verankert" waren) sollten vier verschiedene Arten von Eigenschaften beurteilt werden:

a) objektiv meßbare Merkmale: „Körpergröße", „Körpergewicht",

b) somatisch-äußerliche Merkmale: „Äußere Erscheinung", „Körperliche Gesundheit",

c) Verhaltenseigenschaften: „Antrieb", „Anpassungsfähigkeit",

d) sogenannte Wesenseigenschaften: „Gefühlsbestimmtheit", „Originalität".

Nach Zufall wurden die Beurteiler in zwei gleich große Gruppen ($n = 14$) aufgeteilt, die sich durch die Art der schriftlichen *Instruktion,* die sie erhielten, unterschieden.

Gruppe A erhielt eine Instruktion, die geeignet erschien, einen gewissen Druck auf die Beurteiler auszuüben und sie zu größeren Leistungen anzuspornen:

Der Sinn dieses Versuchs besteht darin, zu erforschen, wie gut bzw. *zutreffend* jeder einzelne von Ihnen schätzen bzw. beurteilen kann. (Arbeiten Sie deshalb bitte ganz für sich.) Um die Genauigkeit Ihrer Beurteilungen festzustellen, wurden über die zu beurteilenden Versuchspersonen bereits objektive Daten und Testdaten erhoben. Wir wollen am Schluß des Versuchs gemeinsam feststellen, wie gut jeder einzelne von Ihnen die „objektiven" Daten getroffen hat, und diese mit den Resultaten, die Sie im Fragebogen-Experiment erzielten, in Beziehung setzen. (Der letzte Satz bezog sich auf einen kurz zuvor ausgefüllten Attitude-Fragebogen.)

Außerdem wurden die Mitglieder der Gruppe A nochmals aufgefordert, ihren Namen anzugeben und auf einer graphischen Rating-Skala anzukreuzen, „ein wie guter Beurteiler menschlicher Eigenschaften" sie zu sein glaubten.

Daß die Instruktion A tatsächlich so etwas wie Stress bewirkte, zeigte eine *Inhaltsanalyse von Selbstbeobachtungs*protokollen der Gruppe A. Danach gaben 11 Rater an, sich um „peinliche Genauigkeit" bemüht zu haben, 7 befürchteten irgendwelche Folgen für ihr Fortkommen im Studium; 4 Beurteiler fühlten sich durch die forcierte Namensnennung beunruhigt, 3 gaben an, es möglichst „wie die anderen" gemacht haben zu wollen; 3 weitere bemühten sich, vor allem extreme Urteile zu vermeiden.

Gruppe B erhielt eine Instruktion von ähnlicher Länge, aber mit ganz anderem Inhalt:

Der Sinn dieses Versuchs besteht darin, das Verhalten der Versuchsperson zu *schätzen,* nicht aber etwa exakt zu messen. Deshalb gibt es auch keine „richtigen" oder „falschen" Lösungen bei Ihrer Aufgabe, da keine objektiven Daten von den Versuchspersonen vorliegen. Sie sollen die Vpn einfach so einschätzen, wie Sie es für richtig halten. Sie werden aber gebeten, ganz für sich zu arbeiten. Geben Sie bitte nun noch auf der folgenden Linie durch Ankreuzen an, in welchem Maße Sie gerne menschliche Eigenschaften beurteilen ...

Da es sich beim graphischen Rating von Eigenschaften um eine relativ wenig komplexe Leistung handelt, war zu vermuten, daß die Ratings der Gruppe A allgemein fehlerfreier, und dort, wo es sich nachprüfen läßt, exakter ausfallen würden:

Hypothesen

1. Die Beurteiler der Gruppe A werden weniger dem Halo-Effekt unterliegen als die der Gruppe B.
2. Die Merkmale der Kategorie a), deren Ausprägunggrade objektiv meßbar sind, werden unter der Bedingung A genauer geschätzt als unter der Bedingung B.

3. Je „äußerlicher" Eigenschaften sind, desto zuverlässiger lassen sie sich mit Rating-Skalen einschätzen, d. h. das Ausmaß des relativen Halo-Effekts wird bei den Merkmalen b) geringer sein als bei denjenigen der Kategorien c) und d).

Ergebnisse

Zur Prüfung der ersten Hypothese wurde für jede der beiden Versuchsgruppen eine Wechselwirkungsvarianzanalyse zwischen Ratern und Ratees (14 × 4) gerechnet.

Bedingung A (Rating unter Druck)

Quelle	Quadratsumme	*df*	Varianz	*F*	*p*
Zw. Ratern (*R*)	18 993	13	1 416,000	1,770	= 0,05
Zw. Ratees (*I*)	8 632	3	2 877,333	3,486	< 0,05
Interaktion (*R* × *I*)	13 988	39	358,923	*0,434*	> *0,05*
Innerhalb	323 469	392	825,176		
Total	365 092	447			

Bedingung B (Rating in entspannter Situation)

Quelle	Quadratsumme	*df*	Varianz	*F*	*p*
Zw. Ratern (*R*)	25 114	13	1 931,846	2,440	< 0,01
Zw. Ratees (*I*)	4 316	3	1 438,666	1,817	> 0,05
Interaktion (*R* × *I*)	15 882	39	407,230	*0,514*	> *0,05*
Innerhalb	310 325	392	791,645		
Total	355 637	447			

Da in beiden Fällen nur eine minimale Restvarianz für die Wechselwirkung gefunden wurde, kann *Hypothese 1* nicht gehalten werden. Dennoch zeigt sich eine gewisse Überlegenheit der Ratings der Gruppe A, da hier in viel stärkerem Maße als bei Gruppe B die Unterschiede *zwischen* den beurteilten *Individuen* gegenüber denen zwischen den *Beurteilern* zur Gesamtvarianz beitragen.

Zur Prüfung der *Hypothese 2* wurden arithmetische Mittel und Standardabweichungen der Ratings von Körpergrößen und Körpergewicht den tatsächlichen Maßen der vier Beurteilten gegenübergestellt (die Ratings auf den lediglich an den Enden mit Zahlen verankerten Skalen waren zuvor in genaue Maße transformiert worden).

Auch *Hypothese 2* kann demnach nicht eindeutig bestätigt werden; die gefundenen Unterschiede in den Schätzleistungen sind statistisch unbedeutend. Es zeigen sich jedoch in jedem Fall deutliche Tendenzen zu einer genaueren Schätzleistung durch die unter Druck arbeitende Gruppe (A). Besonders bemer-

kenswert ist, daß die mittleren Ratings nur minimal von den objektiv gemessenen Werten der Vpn abweichen, und zwar

in Gruppe A um durchschnittlich 4,0 cm und 1,2 kg,
in Gruppe B um durchschnittlich 5,9 cm und 1,6 kg.

Körpergröße

Vp	echter Wert (cm)	Schätzung durch Gruppe A		Schätzung durch Gruppe B	
		M	*s*	*M*	*s*
1	178,3	172,4	6,0	170,7	5,8
2	169,5	166,4	5,2	164,4	5,2
3	176,7	171,1	5,2	169,5	4,2
4	166,3	164,5	5,8	162,5	5,4

Körpergewicht

Vp	echter Wert (kg)	Schätzung durch Gruppe A		Schätzung durch Gruppe B	
		M	*s*	*M*	*s*
1	66,3	67,1	6,6	65,6	5,8
2	59,8	59,3	5,1	60,4	6,3
3	71,5	70,9	7,9	72,8	5,7
4	55,6	58,4	6,5	59,5	4,4

Unter beiden Versuchsbedingungen wurde die Körpergröße durchgängig leicht *unter*schätzt, und zwar von Gruppe B durchweg stärker als von Gruppe A.

Um *Hypothese 3* zu prüfen, wurden die Ratings der Gruppe B getrennt nach 3 Gruppen von jeweils zwei Rating-Skalen einer Varianzanalyse zwischen Ratern und Ratees unterzogen.

Äußere Erscheinung; Körperliche Gesundheit *(R = 14; I = 4)*

Quelle	Quadratsumme	*df*	Varianz	*F*	*p*
Zw. Ratern (*R*)	13 000	13	1 000,000	3,031	$< 0{,}01$
Zw. Ratees (*I*)	4 692	3	1 564,000	4,714	$< 0{,}01$
Interaktion ($R \times I$)	13 489	39	345,872	*1,048*	$> 0{,}05$
Innerhalb	18 474	56	329,893		
Total	46 655	111			

Antrieb; Anpassungsfähigkeit *(R = 14; I = 4)*

Quelle	Quadratsumme	*df*	Varianz	*F*	*p*
Zw. Ratern (*R*)	10 437	13	802,846	1,683	$> 0{,}05$
Zw. Ratees (*I*)	16 628	3	5 560,666	11,658	$< 0{,}01$
Interaktion ($R \times I$)	14 450	39	370,513	0,777	$< 0{,}05$
Innerhalb	26 711	56	476,982		
Total	68 226	111			

Gefühlsbestimmtheit; Originalität *(R = 14; I = 4)*

Quelle	Quadratsumme	*df*	Varianz	*F*	*p*
Zw. Ratern (*R*)	26 781	13	2 060,077	4,441	$< 0{,}01$
Zw. Ratees (*I*)	13 270	3	4 423,333	9,535	$< 0{,}01$
Interaktion ($R \times I$)	16 131	39	413,715	0,892	$> 0{,}05$
Innerhalb	25 978	56	463,893		
Total	82 160	111			

Hypothese 3 muß damit zurückgewiesen werden; es zeigen sich keine Unterschiede in der Schätzung des relativen Halo-Effekts zwischen den drei Arten von beurteilten Eigenschaften. Berücksichtigt man das wünschenswerte Verhältnis der *Zwischen*-Varianzanteile zueinander (zwischen den Beurteilern: unbedeutend! zwischen den Beurteilten: sehr signifikant! Wechselwirkung: unbedeutend!), so scheint es, als ließen sich die als „Verhaltenseigenschaften" bezeichneten Merkmale „Antrieb" und „Anpassungsfähigkeit" am zuverlässigsten beurteilen.

Experiment III

Es soll untersucht werden, wie groß der Halo-Effekt beim Rating mit verschiedenen Skalen-*Formen* ist. Erneut wird die Frage einer Verminderung dieses Rating-Fehlers durch Instruktion gestellt.

Methode

29 Studenten wurden nach Zufall in 6 Gruppen aufgeteilt. In einem 3×3-Versuchsplan wurden zwei Merkmale variiert:

a) die *Form* der Rating-Vorlage
   1. Millimeter-Rating (gerade Linie ohne Unterteilung, 120 mm),
   2. Abschnitts-Rating (gerade Linie, 9fach unterteilt, 120 mm),
   3. Numerisches Rating (9 Stufen);

b) die *Instruktion* an die Rater
   1. schriftliche Instruktion mit Erläuterung des Halo-Effekts wie in Experiment I, Bedingung B,
   2. mündliche Instruktion mit Erläuterung des Halo-Effekts (5—10 Minuten lang),
   3. Kontrollgruppe: Keinerlei besondere Rater-Instruktion.

Die Beurteiler verteilten sich wie folgt auf die einzelnen Versuchsbedingungen:

	Graphische Rating-Skala		*Numerische*
(*N* = 29)	Millimeter-Rating	Abschnitts-Rating	Rating-Skala
Mündliche Erläuterung	2	4	4
Schriftliche Erläuterung	2	4	3
Keine Erläuterung	4	2	4

Zu beurteilen waren 4 Vpn (2 männliche und zwei weibliche), die etwa 10 Minuten lang mit der „Einsteckprobe" nach STEIN beschäftigt waren und sich dabei mit dem Versuchsleiter unterhielten. Vorgegeben waren 6 Rating-Skalen, mit denen die Merkmale „Anpassungsfähigkeit", „Antrieb", „Gefühlsbestimmtheit", „Mitschwingungsfähigkeit" und „Originalität" sowie „Bewegungsstil" („äußerst fließend, geschmeidig" bis „äußerst eckig, ungelenk") beurteilt werden sollten.

Hypothesen

1. Bei den verschiedenen Skalenvorlagen wird der relative Halo-Effekt eine unterschiedliche Größe annehmen.

2. Die schriftliche und die mündliche Erläuterung des Halo-Effekts werden den Halo-Effekt bei den betreffenden Ratern gegenüber dem bei der Kontrollgruppe vermindern.

Ergebnisse

Zur Prüfung der ersten Hypothese wurden drei Wechselwirkungsvarianzanalysen gerechnet.

Graphisches Rating: Millimeter-Score *(R = 8; I = 4)*

Quelle	Quadratsumme	*df*	Varianz	*F*	*p*
Zw. Ratern (*R*)	4 936	7	705,143	2,316	< 0,05
Zw. Ratees (*I*)	1 640	3	546,666	1,796	> 0,05
Interaktion (*R* × *I*)	17 574	21	836,857	*2,749*	*< 0,01*
Innerhalb	48 711	160	304,444		
Total	72 861	191			

Graphisches Rating: Abschnitts-Score *(R = 10; I = 4)*

Quelle	Quadratsumme	*df*	Varianz	*F*	*p*
Zw. Ratern (*R*)	6 547	9	727,444	1,501	> 0,05
Zw. Ratees (*I*)	24 175	3	8 058,333	16,632	< 0,01
Interaktion (*R* × *I*)	14 938	27	553,259	*1,142*	*> 0,05*
Innerhalb	96 900	200	484,500		
Total	142 560	239			

Numerisches Rating *(R = 11; I = 4)*

Quelle	Quadratsumme	*df*	Varianz	*F*	*p*
Zw. Ratern (*R*)	26 031	10	2 603,100	3,967	< 0,01
Zw. Ratees (*I*)	13 915	3	4 638,333	7,069	< 0,01
Interaktion (*R* × *I*)	18 089	30	602,967	*0,919*	*> 0,05*
Innerhalb	144 359	220	656,177		
Total	202 394	263			

Damit wurde *Hypothese 1* bestätigt. Am größten ist die Interaktionsvarianz zwischen Beurteilern und Beurteilten bei derjenigen Gruppe, die Millimeter-Skalen benutzte; bei den beiden anderen Gruppen ist der Halo-Effekt unbedeutend. Die günstigste Verteilung der Varianzanteile zeigen die Ratings mit der 9stufigen graphischen Skala („Abschnitts-Rating"), da hier ein beträchtlicher Varianzanteil lediglich auf die Differenzen zwischen den beurteilten Individuen zurückgeht.

Hypothese 2 wurde analog derjenigen des Experiments I aufgestellt. Ob die dort gefundenen Resultate erhärtet und noch näher erläutert werden können, sollten Varianzanalysen der drei Gruppen mit verschiedener Instruktion zeigen.

Schriftliche Instruktion der Rater *(R = 9; I = 4)*

Quelle	Quadratsumme	*df*	Varianz	*F*	*p*
Zw. Ratern (*R*)	5 370	8	671,250	1,019	> 0,05
Zw. Ratees (*I*)	5 102	3	1 700,666	25,801	< 0,01
Interaktion (*R* × *I*)	4 935	24	205,625	*0,312*	*> 0,05*
Innerhalb	118 603	180	658,903		
Total	134 037	215			

Mündliche Instruktion der Rater *(R = 10; I = 4)*

Quelle	Quadratsumme	*df*	Varianz	*F*	*p*
Zw. Ratern (*R*)	5 913	9	657,000	1,081	> 0,05
Zw. Ratees (*I*)	20 915	3	6 971,666	11,475	< 0,01
Interaktion (*R* × *I*)	11 092	27	410,815	*0,676*	*> 0,05*
Innerhalb	121 512	200	607,560		
Total	159 432	239			

Keine Instruktion *(R = 10; I = 4)*

Quelle	Quadratsumme	*df*	Varianz	*F*	*p*
Zw. Ratern (*R*)	20 905	9	2 322,777	5,790	< 0,01
Zw. Ratees (*I*)	21 521	3	7 173,666	17,880	< 0,01
Interaktion (*R* × *I*)	13 227	27	489,888	*1,221*	*> 0,05*
Innerhalb	80 241	200	401,205		
Total	135 894	239			

Unter der mündlichen und schriftlichen Instruktion zum Halo-Effekt tritt eine völlig unbedeutende R × I-Restvarianz auf; bei der Kontrollgruppe erreicht ihr Betrag jedoch nahezu die 5%-Grenze. Wollte man Nuancen berücksichtigen, so zeigt sich der geringste Einfluß eines Halo-Effekts bei der Gruppe von Beurteilern, die *schriftliche* Erläuterungen über die Gefahr dieses Fehlers erhalten hatte. Jedenfalls kann *Hypothese 2* als im wesentlichen bestätigt gelten.

Experiment IV

In diesem Versuch sollen wiederum verschiedene Skalen-*Formen* untersucht werden. In Anlehnung an ein Experiment von MADDEN u. BOURDON (1963) soll entschieden werden, ob der Hinweis auf eine Verteilung von Verhaltensmerkmalen gemäß der *Normalverteilung*, wie er in vielen amerikanischen Rating-Instruktionen enthalten ist und von einigen Autoren gefordert wird, zu zuverlässigeren Ratings führt, oder nicht.

Methode

27 Studenten wurden nach Zufall in 3 Gruppen eingeteilt, von denen jede mit einer andern Form von Rating-Skalen das Verhalten von 4 Vpn (zwei männlichen und zwei weiblichen) zu beurteilen hatte.

Die zu beobachtenden Vpn unterzogen sich nacheinander einer Prüfung mit zwei Untertests des HAWIE beim gleichen Versuchsleiter; es handelte sich um den Untertest „Allgemeines Wissen" aus dem Verbalteil und um das „Bilderordnen" aus dem Handlungsteil dieses Verfahrens. Jede Vp konnte auf diese Weise etwa 20 Minuten lang beobachtet werden.

Sechs Merkmale waren zu beurteilen: „Anpassungsfähigkeit", „Antrieb", „Irritierbarkeit", „Originalität", „Sicherheit" und „Steuerungsfähigkeit"; sie wurden schriftlich erläutert.

Die drei Rating-Skalenformen waren sämtlich numerische 9-Punkte-Skalen. Sie unterschieden sich wie folgt:

Form A: Neben den untereinandergeschriebenen Ziffern von 1 (oben) bis 9 (unten) und den danebenstehenden Deskriptionen „Weit überdurchschnittlich" (9) bis „Weit unterdurchschnittlich" (1) befand sich ein graphisches Schaubild, das eine aus 9 mit den entsprechenden Häufigkeits-Prozentangaben versehene Normalverteilung zeigte. Die Blockdarstellung gab z. B. an, daß 20% der Population den Wert „5" erreichen würden, aber nur 4% die Werte „1" oder „9".

Form B: Ziffern mit Deskriptionen wie in Form A, aber kein Schaubild oder sonstiger Hinweis auf eine Normalverteilung.

Form C: Hier war lediglich gefordert, Ziffern von 1 bis 9 zuzuordnen. Es wurden weder eine Zahlenreihe noch Deskriptionen oder ein Schaubild geboten.

Hypothesen

1. Die Ratings mit der Form A werden sich insgesamt eher einer Normalverteilung annähern als die mit den beiden anderen Formen.

2. Die Ratings mit der Form A werden weniger dem Halo-Effekt unterliegen als die mit den beiden anderen Skalenformen.

Ergebnisse

Zur Prüfung der ersten Hypothese wurde der Grad der Normalität sämtlicher Ratings der drei Gruppen statistisch überprüft. Bei der Prüfung mit dem KOLMOGOROFF-SMIRNOFF-Test für die Güte der Anpassung (LIENERT, 1962) wurde ein kritisches Signifikanzniveau von 20% für erforderlich angesehen (vgl.

LIENERT, 1962, S. 322). Die folgende Tabelle zeigt die Größe des kritischen Anpassungswertes D bei den Ratings der drei Versuchsgruppen.

Rating-Skala	*M*	*s*	*D*	*p*	
Form A	4,85	1,74	0,0780	$< 0{,}20$	
Form B	5,17	1,73	0,0500	$> 0{,}20$	($D_{0.20} = 0{,}0776$!)
Form C	5,05	1,84	0,0615	$> 0{,}20$	

Das Ergebnis widerspricht der Erwartung: Bei annähernd gleichen Streuungen sind lediglich diejenigen Ratings, denen *keine* Normalitäts-Instruktion zugrunde lag, hinreichend gut der Normalverteilung angenähert. *Hypothese 1* wird daher zurückgewiesen.

Ob die Beurteiler der Gruppe A trotzdem relativ fehlerfreier beurteilten, zeigt die Prüfung der zweiten Hypothese durch den Vergleich dreier Wechselwirkungsvarianzanalysen.

Rating-Form A (Normalverteilung, Ziffern, Deskriptionen ($R = 8$; $I = 4$)

Quelle	Quadratsumme	*df*	Varianz	*F*	*p*
Zw. Ratern (R)	121,979	7	17,426	8,618	$< 0{,}01$
Zw. Ratees (I)	312,083	3	104,028	51,448	$< 0{,}01$
Interaktion ($R \times I$)	143,917	21	6,853	*3,389*	*< 0,01*
Innerhalb	323,500	160	2,022		
Total	901,479	191			

Rating-Form B (Ziffern, Deskriptionen) ($R = 8$; $I = 4$)

Quelle	Quadratsumme	*df*	Varianz	*F*	*p*
Zw. Ratern (R)	38,167	7	5,452	2,114	$< 0{,}05$
Zw. Ratees (I)	39,500	3	13,167	5,105	$< 0{,}01$
Interaktion ($R \times I$)	73,916	21	3,520	*1,365*	*> 0,05*
Innerhalb	412,667	160	2,579		
Total	566,250	191			

Rating-Form C (keine Vorlage) ($R = 9$; $I = 4$)

Quelle	Quadratsumme	*df*	Varianz	*F*	*p*
Zw. Ratern (R)	96,333	8	12,042	4,505	$< 0{,}01$
Zw. Ratees (I)	71,532	3	23,844	8,920	$< 0{,}01$
Interaktion ($R \times I$)	70,593	24	2,941	*1,100*	*> 0,05*
Innerhalb	481,167	180	2,673		
Total	719,625				

Danach ergibt sich, daß der Halo-Effekt nur bei der Gruppe A eine statistisch bedeutende Rolle spielt. Unter den beiden anderen Bedingungen tritt keine signifikante Restvarianz auf. *Hypothese 2* wird damit ebenfalls zurückgewiesen.

Die Größe des Halo-Effekts in Abhängigkeit von den Eigenschaften der Rater

Die folgenden Untersuchungen sind Auswertungen des Versuchsmaterials der bisher geschilderten Experimente. Ihnen kommt — insbesondere wegen der durch die Auswahl der Rater bedingten geringen Stichprobengröße — lediglich der Status von „pilot studies" zu.

a) Geschlecht der Rater

Bei den Experimenten I und II wurde die Größe der Rater-Ratee-Interaktion für Beurteiler männlichen und weiblichen Geschlechts getrennt berechnet. Es ergaben sich dabei keine wesentlichen Unterschiede zwischen männlichen und weiblichen Ratern, sondern lediglich leichte *Tendenzen* zugunsten einer gewissen Überlegenheit der *männlichen* Beurteiler, einen Halo-Effekt zu vermeiden (I) und bei größerer Konformität der Rater Urteile zwischen den Ratees differenzieren zu können (II).

b) Intelligenz der Rater

Die 29 Beurteiler des Experiments III wurden einer Intelligenzprüfung mit einer Kurzform des IST-Amthauer (Lienert u. Leuchtmann, 1958), bestehend aus den Untertests „Satzergänzen", „Analogien" und „Gemeinsamkeiten" unterzogen (reine Testzeit 21 Minuten). Auf Grund der Testresultate wurden je sechs (entsprechend 20%) Rater mit den höchsten und niedrigsten Testwerten ausgewählt; durch Randomisierung zu Beginn des Versuchs streuten sie relativ gleichmäßig über die verschiedenen Versuchsbedingungen. Die IST+-Gruppe (Median der Standardwerte: 115) bestand ebenso wie die IST—-Gruppe (Median der Standardwerte: 102) aus je 3 männlichen und weiblichen Beurteilern gleichen Alters. Für jede der beiden Gruppen wurde eine Wechselwirkungsvarianzanalyse zwischen Ratern und Ratees berechnet. In beiden Fällen ergaben die Varianzanalysen ein ähnliches Bild; geringfügige *Tendenzen* zugunsten der „hoch intelligenten" Gruppe (völlig unbedeutender Halo-Effekt), aber auch zugunsten der „niedrig intelligenten" Gruppe (signifikanter Unterschied zwischen Ratees, nicht aber zwischen Ratern), lassen sich kaum interpretieren. Die Homogenität des Bildungsniveaus der Beurteiler-Gruppen, die geringe Größe der Stichproben und die relative Grobheit des hier angewendeten Intelligenz-Testverfahrens legen es nahe, einen entsprechenden Versuch neu zu planen.

c) Extraversion (EYSENCK)

Die 28 Rater des Experiments II füllten einen Fragebogen aus, der u. a. die Skala „Extraversion" des Maudsley Personality Inventory (MPI) von EYSENCK (1959) enthielt. Aus dem oberen und unteren Viertel der Extraversionsskala wurden jeweils ein männlicher und ein weiblicher Beurteiler der Gruppen A und B ausgewählt, so daß zwei Extremgruppen von je 4 Beurteilern verglichen werden konnten. Entgegen der Erwartung zeigten sich keinerlei Unterschiede zwischen Beurteilern mit hohen und niedrigen Punktwerten auf der Extraversionsskala.

d) Neurotizismus (EYSENCK)

Die Beurteiler des Experiments III füllten einen Fragebogen mit den „Neurotizismus"-Items des MPI aus. Danach wurden die acht (entsprechend 25 %) Beurteiler mit den höchsten Scores (über 15) und den niedrigsten Werten (weniger als 9) voneinander getrennt und ihre Ratings Varianzanalysen unterzogen. Es wurde angenommen, daß sich Unterschiede in der Größe des relativen Halo-Effekts zeigen würden. Es ergaben sich folgende Resultate:

Rater mit hohen „Neurotizismus"-Werten ($R = 8$; $I = 4$)

Quelle	Quadratsumme	*df*	Varianz	*F*	*p*
Zw. Ratern (R)	9 145	7	1 306,429	2,790	$< 0{,}01$
Zw. Ratees (I)	14 362	3	4 787,333	10,222	$< 0{,}01$
Interaktion ($R \times I$)	16 891	21	804,333	*1,717*	*< 0,05*
Innerhalb	74 932	160	468,325		
Total	115 330	191			

Rater mit niedrigen „Neurotizismus"-Werten ($R = 8$; $I = 4$)

Quelle	Quadratsumme	*df*	Varianz	*F*	*p*
Zw. Ratern (R)	17 671	7	2 524,429	3,972	$< 0{,}01$
Zw. Ratees (I)	10 881	3	3 627,000	5,707	$< 0{,}01$
Interaktion ($R \times I$)	10 418	21	496,095	*0,781*	*> 0,05*
Innerhalb	101 687	160	635,544		
Total	140 657	191			

Entsprechend der Erwartung ergab sich ein Unterschied zwischen beiden Rater-Gruppen. Während die Interaktionsvarianz ($R \times I$) der Gruppe mit hohem „Neurotizismus" signifikant zur Gesamtvarianz beiträgt, ist diejenige der Gruppe mit niedrigen „Neurotizismus"-Werten völlig unbedeutend.

IV. Diskussion

In einer Erörterung des Problems der Zuverlässigkeit von Verhaltensbeurteilungen mit Rating-Skalen hatte sich gezeigt, daß die statistische Bestimmung der Größe des Halo-Effekts eine gute Schätzung der Rating-Zuverlässig-

keit ergebe. Daher wurde eine Reihe von Experimenten geplant und ausgeführt, in denen dieser Betrag unter verschiedenen Bedingungen ermittelt werden sollte.

Zur Eigenart dieser Versuche ist vor allem zu bemerken, daß die verwendeten Rating-Skalen hauptsächlich zur Schätzung einiger nach THOMAE (1955) modifizierter Eigenschaften dienten, die sich nach den Erfahrungen in diesen Experimenten gut zur Beschreibung des Verhaltens von Probanden in testähnlichen Situationen zu eignen scheinen. Die Modifikation bestand vor allem darin, ursprünglich qualitativ gemeinte Verhaltenszuordnungen („Gesamteindruck" von THOMAE) durch Rating-Skalen zu ersetzen, die eine *Kontinuität* des betreffenden Merkmals suggerieren. Das Ergebnis waren annähernd normal verteilte Daten.

Demgegenüber scheint die Auswahl der Rater (Studierende der Psychologie) kein besonderes Problem darzustellen, denn sie entspricht dem Fall des sogenannten *Experten-Rating*; auch in anderen Versuchen und Forschungsprogrammen wird mit besonders instruierten Studenten oder auch Lehrern gearbeitet. Dennoch dürfte die *Schulung* der Rater, der man häufig so gut wie alles zutraut (vgl. DRIVER, 1942; ANDREGG, 1951; SCHMIDT, 1965) eine wichtige Variable sein. Dem Resultat CHAUFFARDS (1948), der bei einem Vergleich eigener Rating-Ergebnisse mit denen von BONNARDEL (1946) fand, daß erfahrene Psychologen feiner differenzieren können und weniger dem Halo-Effekt unterliegen, steht allerdings eine Reihe von Befunden gegenüber (zusammengestellt bei HASEMANN, 1964), nach denen das Studium der Psychologie die Fähigkeit zum Beurteilen von Mitmenschen eher verschlechtert; das Vertrauen in den „erfahrenen Psychologen" wurde auch durch den Bericht von COHEN (1962) erschüttert. Nach alledem scheint es erforderlich zu sein, die Frage, wer der „zuverlässigere" Rater ist, in einem eigenen Experiment zur Entscheidung zu bringen.

Nach den deutlichen Resultaten der Experimente I und III scheint jedoch die Vermutung nicht unberechtigt zu sein, daß vor allem *kurzfristige Einstellungen* für das Ausmaß, in dem Ratings einem Halo-Effekt unterliegen, entscheidend sind. Im einzelnen wurde gefunden, daß sich der Halo-Effekt auf ein Minimum reduzieren läßt, wenn

1. den Beurteilern eine Instruktion mit einer Erläuterung des Halo-Effekts und einer Warnung vor seinen Auswirkungen gegeben wird,
2. diese Instruktion schriftlich gegeben wird, so daß sie den Ratern während der Beurteilung ständig gegenwärtig ist.

Außerdem scheinen besonders angespornte Rater etwas bessere Schätzleistungen aufzuweisen (II).

In bezug auf die Verwendung verschiedener *Formen* von Rating-Skalen zeigte sich, daß zwischen *graphischen* und *numerischen* Skalen nur ein äußerst geringer Unterschied in der Anfälligkeit gegenüber einem Rating mit Halo-Effekt besteht (V); immerhin waren Ratings auf einer ununterbrochenen Linie (Millimeter-Rating) nicht so fehlerfrei wie diejenigen auf einer 9fach unterteilten Strecke. Mit Bezeichnungen wie „durchschnittlich" usw. versehene nume-

rische Skalen zeigten sich gegenüber einfachen numerischen Skalen in bezug auf die Vermeidung des Halo-Effekts beim Rating nicht überlegen (IV). Skalenformen, die zu einer *Normalverteilung* von Ratings aufforderten, führten eher zu einem verstärkten Halo-Effekt, der jedoch nicht etwa durch eine tatsächlich größere Normalität dieser Ratings zustande kam. Dieser Befund entspricht einem Resultat von CANNON, OLSON u. SPANOCON (1961) nach dem Ratings, die man in eine symmetrische, vor allem normale Verteilung zwang, nicht signifikant reliabler als „freie" Ratings waren. Für eine eingehende Diskussion des Problems der *Form* von Rating-Skalen muß auf ein weiteres Experiment und die ausführlichere Erörterung bei SCHMIDT (1965; S. 41—43, 48—52, 81—83) verwiesen werden. Das Resultat, daß mit graphischen (Abschnitts-)Skalen ähnliche Beurteilungen (und Fehler) zustandekommen wie mit numerischen Systemen, legt die Empfehlung für die Praxis nahe, *graphische* Rating-Skalen (die beliebter sind) eher bei ungeübten Beurteilern, *numerische* Skalen (die teilweise ökonomischer sind) eher bei ausgesprochenen Experten-Ratings zu verwenden. Dennoch reichen die Ergebnisse dieser Arbeit nicht aus, um beide Formen als in jedem Falle austauschbar zu bezeichnen; dies müßte — insbesondere für verschiedene Situationen der Praxis — jeweils erst empirisch nachgewiesen werden.

In diesem Zusammenhang darf ein Ergebnis von TAYLOR, PARKER u. FORD (1959) nicht übersehen werden: Die Form einer Rating-Skala übte auf die Rater-Reliabilität einen geringeren Einfluß aus als gewisse *Situations*-Bedingungen. Unterscheidet man bei der vorliegenden Arbeit einmal zwischen drei verschiedenen Arten von Reaktionen des Beurteilers, nämlich Reaktionen auf das Verhalten des *Probanden*, auf die Situation des *Raters* einschließlich der ihm gegebenen Anweisungen, und schließlich auf das ihm vorliegende *Rating-System*, so scheint nach den vorliegenden Ergebnissen die Syntax des Ratings eine vergleichsweise weniger wichtige Rolle zu spielen. Dieser Eindruck müßte allerdings durch eine Reihe von weiteren experimentellen Entscheidungen erhärtet werden.

Die Versuche, die sich auf das Rating verschiedener Arten von *Eigenschaften* und auf Persönlichkeitseigenschaften der *Beurteiler* bezogen, sind eher geeignet, die Formulierung weiterer Hypothesen herauszufordern, als dazu, bereits jetzt zusammenhängend diskutiert zu werden.

Wie zu Beginn erwähnt wurde, stehen der Anwendung von Ratings eine Reihe theoretischer Bedenken entgegen. Der Rating-Vorgang würde zwar der Definition des Messens nach CAMPBELL (1940) und STEVENS (1960) entsprechen, doch erfolgt die Zuordnung von Zahlen zu den Aspekten von Gegenständen (CAMPBELL) bzw. den Gegenständen selbst (STEVENS) gemäß einem subjektiven Maßstab. Am Ende einer methodenkritischen Arbeit müßte daher rückblickend gefragt werden, ob man mit Rating-Skalen menschliches Verhalten überhaupt einigermaßen zuverlässig beurteilen kann.

Die Formulierung dieser Frage scheint deshalb delikat zu sein, weil Rating-Skalen seit je auch ohne Reliabilitätsprüfungen angewendet wurden und ständig angewendet werden. Zudem wurde als Reliabilität von Ratings fast ausschließlich

nur eine wichtige Voraussetzung derselben, nämlich die Objektivität der Beurteiler untersucht. Nach den vorliegenden Ergebnissen und Überlegungen kann die Frage unter der Bedingung mit „ja" beantwortet werden, daß bestimmte Bedingungen berücksichtigt werden. Ratings müssen jedoch noch unter wesentlich mehr Bedingungen untersucht werden, als es in dieser Arbeit geschah. Es ist zu empfehlen, Rating-Skalen nicht zu verwenden, ohne zuvor zu ermitteln, wie sich der Halo-Effekt reduzieren läßt.

Daß Ratings unter gewissen Voraussetzungen zuverlässige Beurteilungen ergeben, bedeutet zweierlei auf keinen Fall:

1. Wenn man Grund zu der Annahme hat, mit einer Rating-Skala lasse sich „Anpassungsfähigkeit" zuverlässig beurteilen, so bedeutet dies keineswegs, daß es tatsächlich „Anpassungsfähigkeit" ist, das hier relativ zuverlässig beurteilt wird. Im ungünstigsten Fall könnte es etwas ganz anderes sein als Anpassungsfähigkeit. Obwohl Reliabilität und *Validität* von subjektiven Beurteilungen nicht unabhängig voneinander sind, müßte die Frage, ob Rater das beurteilen, was sie beurteilen sollen, eigens beantwortet werden. Die Bestimmung der Validität von Ratings trifft aber auf zwei Schwierigkeiten: Rating-Methoden werden nicht nur häufig verwendet, weil keine besseren Verfahren (die als Validierungskriterien dienen könnten) vorhanden sind, sondern sie sind selbst eines der wohl am häufigsten verwendeten Validitätskriterien für andere Verfahren!

2. Es gibt kein allgemein zuverlässiges Rating-„Rezept". Sicherlich ließe sich eine ideale (fiktive) Rating-Situation schildern, die alle Ergebnisse dieser Arbeit (und anderer) berücksichtigen würde. Aber der Wert dieses Konglomerats von Bedingungen, die sich alle als zuverlässigkeitsfördernd erwiesen haben, ist als Ganzes keineswegs nachgewiesen. Der Nutzen der gleichzeitigen Beherzigung aller Resultate von einschlägigen Experimenten könnte sich als Artefakt erweisen. Dies würde fatal einem methodischen Mißverständnis gleichen, bei dem ein Erzieher die experimentellen Resultate des Kapitels „Educational Psychology" der „Psychological Abstracts" in komprimierter Form verwenden würde, um mit einer konkreten Erziehungsschwierigkeit in einer ganz bestimmten Schulklasse und einer ganz bestimmten Situation fertig zu werden.

Die Zuverlässigkeit von Rating-Skalen sollte deshalb experimentell kontrolliert werden, ehe man diese Beurteilungsinstrumente anwendet, um aus den mit ihnen gewonnenen Daten Schlüsse zu ziehen.

Zusammenfassung

In einer Erörterung des Problems der Zuverlässigkeit von Verhaltensbeurteilungen mit Rating-Skalen wurde gezeigt, daß die gebräuchlichen Methoden zur Bestimmung der Reliabilität von Ratings (vor allem die Berechnung der Rater-Übereinstimmung) kaum geeignet sind, die Zuverlässigkeit von Rating-Beurteilungen zu schätzen. Es wurde vorgeschlagen, die relative Zuverlässigkeit von Ratings durch die Bestimmung des Ausmaßes zu schätzen, in dem systematische Beurteilungsfehler, vor allem der Halo-Effekt, vermieden werden; dazu dient eine von GUILFORD (1954) beschriebene varianzanalytische

Technik. Zur Verwirklichung des vorgeschlagenen methodischen Konzepts wurden Experimente geplant und ausgeführt, bei denen mehrere Rater mehrere Personen in testähnlichen Situationen beobachteten und nach verschiedenen Eigenschaften beurteilten. Dabei wurde eine Reihe von Annahmen über die Zuverlässigkeit von Ratings unter verschiedenen Bedingungen überprüft. Bei einer Diskussion der Ergebnisse wurde auf die Notwendigkeit weiterer methodenkritischer Forschungen hingewiesen; von einer unkontrollierten Anwendung der Rating-Methode wurde abgeraten.

Summary

Methods for measuring the reliability of behaviour-ratings were discussed. It was shown that usual measures (above all rater-reliability) are not adequate enough to estimate the reliability of ratings. A suggestion was made to estimate relative reliability by estimating the tendency of avoiding halo variance according to a technique designed by GUILFORD (1954). Within this concept experiments with several raters and ratees were planned and executed. Assumptions about the reliability of ratings under different conditions were tested. A discussion showed how necessary it is to do further investigations.

Résumé

En discutant des méthodes à estimer la réliabilité de ratings de comportement humain on pouvait faire voir que des mésures usuelles (surtout celle de la réliabilité des juges) ne sont pas suffisamment adéquates pour estimer la réliabilité de ratings. On proposait à estimer la réliabilité rélative en estimant la tendance à éviter l'effet « halo » suivant une méthode présenté par GUILFORD (1954). Pour réaliser ce concept on faisait des expériments avec plusieurs « raters » et « ratees ». On prouvait des suppositions concernant la réliabilité de ratings aux différents conditions expérimentelles. Une discussion faisait voir la nécessité d'autres recherches méthodiques.

Literatur

ALLPORT, G. W., Persönlichkeit. Stuttgart 1949.

ALLPORT, G. W., Pattern and growth in personality. New York 1961.

ANDREGG, N. B., A critical study of graphic rating scales. Doctorial Diss. Mich. State Coll. University Microfilms Publ. No. 2702, Ann Arbor, Michigan 1951.

BARRETT, R. S., E. K. TAYLOR, J. W. PARKER and S. L. MARTENS, Rating scale content: I. Scale information and supervisory ratings. Pers. Psychol. 1958, *11*, 333—346.

BENDIG, A. W., Reliability and the number of rating scale categories. J. appl. Psychol. 1954a, *38*, 38—40.

BENDIG, A. W., Reliability of short rating scales and the heterogeneity of the rated stimuli. J. appl. Psychol. 1954b, *38*, 167—170.

BENDIG, A. W., Rater reliability and the heterogeneity of the scale anchors. J. appl. Psychol. 1955a, *39*, 37—39.

BENDIG, A. W., Rater reliability and "judgmental fatigue". J. appl. Psychol. 1955b, *39*, 451—454.

BENDIG, A. W., The comparative reliability of double and single rating scales. J. gen. Psychol. 1957a, *57*, 197—201.

BENDIG, A. W., Rater reliability and the heterogeneity of clinical case histories. J. gen. Psychol. 1957b, *57*, 203—207.

BENDIG, A. W., and J. SPRAGUE, Rater experience and the reliability of case history ratings of adjustment. J. consult. Psychol. 1954, *8*, 207—211.

BONNARDEL, R., Étude sur l'évaluation de l'aptitude professionnelle de la maîtrise subalterne et sur les jugements analytiques sur différents aspects du comportement de l'homme. Le travail humain 1946, 178—191.

CANNON, D., H. C. OLSON and SPANOCON, Span of control. II. Effect on reliability of free and forced distributions in rating. Hum.RRO res. Memo., Subtask Spanocon, Task 1961, 11—28.

CAMPBELL, N. R., with others. Final report. Advanc. Sci., No. 2, 1940, 331—349 (nach STEVENS, 1960).

CARTER, G. C., Student personalities as instructors see them. Studies in Higher Education: Lafayette: Purdue University 1945.

CHAUFFARD, C., Essai d'une étude objective du comportement au cours des tests. Le travail humain 1948, 175—190.

CLARK, E. L., Spearman-Brown formula applied to ratings of personality traits. J. educ. Psychol. 1935, *26*, 522—555.

COHEN, R., Die Psychodynamik der Testsituation. Diagnostica 1962, *8*, 3—12.

CONKLIN, E. S., and J. W. SUTHERLAND, A comparison of the scale of values method with the order of merit method. J. exp. Psychol. 1923, *6*, 44—57.

DRIVER, R. S., Training as a means of improving employee performance ratings. Personnel 1942, *18*, 364—370.

EBEL, R. L., Estimation of the reliability of ratings. Psychometrika 1951, *16*, 407—424.

EYSENCK, H. J., Das „Maudsley Personality Inventory" (MPI). Göttingen 1959.

FRÖHLICH, W. D., Kleine Einführung in die Forschungsstatistik. 4. Aufl., Bonn 1964.

GRAUMANN, C. F., Eigenschaften als Problem der Persönlichkeitsforschung. In: LERSCH-THOMAE (Hrsg.), Handbuch der Psychologie, Band 4: Persönlichkeitsforschung und Persönlichkeitstheorie. Göttingen 1960.

GUETZKOW, H., Unitizing and categorizing problems in coding qualitative data. J. clin. Psychol. 1950, *6*, 47—58.

GUILFORD, J. P., Psychometric methods. 2nd ed. New York–Toronto–London 1954.

GUILFORD, J. P., Fundamental statistics in psychology and education. 3rd ed. New York–Toronto–London 1956.

GUILFORD, J. P., Personality. New York–Toronto–London 1959.

GUILFORD, J. P., R. P. CHRISTENSEN, G. TAAFFE and R. C. WILSON, Ratings should be scrutinized. Educ. psychol. Measmt. 1962, *22*, 439—447.

HASEMANN, K., Verhaltensbeobachtung. In: HEISS–GROFFMANN–MICHEL (Hrsg.), Handbuch der Psychologie, Band 6: Psychologische Diagnostik. Göttingen 1964.

HÖRMANN, H., Aussagemöglichkeiten psychologischer Diagnostik. Z. exp. angew. Psychol. 1964, *11*, 353—390.

HOLLINGWORTH, H. L., Experimental studies in judgment. New York 1913.

HOLLINGWORTH, H. L., Judging human character. New York 1922.

JOHNSON, D. M., and R. N. VIDULICH, Experimental manipulation of the halo effect. J. appl. Psychol. 1956, *40*, 130—134.

KNAUFT, E. B., A classification an evaluation of personnel rating methods. J. appl. Psychol. 1947, *31*, 617—625.

KORNHAUSER, A. W., Reliability of average ratings. J. Pers. Res. 1926, *5*, 309—317.

LEE, H. E., and LUCY E. BURNHAM, Correlation between peer ratings and behavior patterns. Technical Rep. No. 2, Graduate School of Business, Stanford University, prep. for Group Psychological Branch, Office of Naval Res. Nonr-225 (62), (1963).

LIENERT, G. A., Testaufbau und Testanalyse. Weinheim/Bergstr. 1961.

LIENERT, G. A., Verteilungsfreie Methoden in der Biostatistik. Meisenheim am Glan 1962.

LIENERT, G. A., und LEUCHTMANN, Die Möglichkeiten einer Kurzform des IST-Amthauer. Psychol. Prax. 1958, *2*, 177—182.

MADDEN, J. M., and R. D. BOURDON, Effects on judgment of variations in rating scale format. USAF Personnel Research Laborat., Tech. Docum. Rep., No. 63-2, 1963.

MARSH, S. E., and F. A. C. PERRIN, An experimental study of the rating scale technique. J. abnorm. soc. Psychol. 1925, *19*, 383—399.

MINER, J. B., The evaluation of a method for finely graduated estimates of ability. J. appl. Psychol. 1917, *1*, 123—133.

MITTENECKER, E., Planung und statistische Auswertung von Experimenten, 4. Aufl. Wien 1963.

PARKER, J. W., E. K. TAYLOR, R. S. BARRETT and S. L. MARTENS, Rating Scale Content: III. Relationship between supervisory- and self-rating. Pers. Psychol. 1959, *12*, 49—63.

REMMERS, H. H., Reliability and halo effect of high school and college student's judgments of their teachers. J. appl. Psychol. 1934, *18*, 619—630.

REYMERT, M. L., and H. A. KOHN, The Mooseheart Graphic Rating Scale for Housemothers and Housefathers. J. appl. Psychol. 1938, *22*, 288—294.

SCHMIDT, H. D., Die Beurteilung des menschlichen Verhaltens durch Rating-Skalen. Phil. Diss., Bonn 1965.

SCHNEEWIND, K. A., Eine non-parametrische Methode zur individuellen Intervallbildung von Ratingskalen. Z. exp. angew. Psychol. 1965, *12*, 302—315.

SHEN, E., The reliability coefficient of personal ratings. J. educ. Psychol. 1925, *16*, 232—236.

SIEGEL, S., Nonparametric statistics for the behavioral sciences. New York–Toronto–London 1956.

STEVENS, S. S., Mathematics, measurement, and psychophysics. In STEVENS, S. S. (ed.), Handbook of experimental psychology. New York–London 1960.

STOCKFORD, L., and H. W. BISSELL, Factors involved in establishing a merit-rating scale. Personnel 1949, *26*, 94—118.

TAYLOR, E. K., R. S. BARRETT, J. W. PARKER and S. L. MARTENS, Rating scale content: II. Effect of rating on individual scales. Pers. Psychol. 1959, *12*, 247—266.

TAYLOR, E. K., and HASTMAN, Relation of format and administration to the characteristics of graphic ratings. Pers. Psychol. 1956, *9*, 181—206.

TAYLOR, E. K., J. W. PARKER and G. L. FORD, Rating Scale Content: IV. Predictability of structured and unstructured scales. Pers. Psychol. 1959, *12*, 247—266.

THOMAE, H., Der psychologische Gesamteindruck. In: COERPER–HAGEN–THOMAE (Hrsg.), Deutsche Nachkriegskinder. Stuttgart 1955.

THOMAE, H., Beobachtung und Beurteilung von Kindern und Jugendlichen. (2. Aufl.), Basel 1957.

TIFFIN, J., Industrial Psychology. New York 1942.

TSCHECHTELIN, S. M. A., A 22-point personality rating scale. J. Psychol. 1944, *18*, 3—8.

Anschrift des Verfassers:

Dr. H. D. SCHMIDT, Psychologisches Institut der Universität Bonn,
53 Bonn, Am Hof 4.