5 Die praktische Bedeutung der Testfairness als zusätzliches Kriterium zu Reliabität und Validität

Claus Möbus

1 Einleitung

Die Einführung in die o.a. Fragestellung läßt sich am besten an Hand eines konkreten Beispiels geben. In der Wochenzeitschrift „DIE ZEIT" vom 6.8.82 befand sich folgendes Stellenangebot.

Ist im Fachbereich 2 – Gesellschafts- und Geschichtswissenschaften – am Institut für Soziologie die Stelle eines/einer

Wissenschaftlichen Mitarbeiters(in) (BAT IIa)

(Kenn-Nr. 124)

für Aufgaben von begrenzter Dauer gemäß SR 2y BAT (maximal 5 Jahre) zu besetzen.

Zu den mit dieser Stelle verbundenen Aufgaben gehören die Mitarbeit in der Methodenausbildung, die Betreuung von Praktika der empirischen Sozialforschung, Studienberatung und Betreuung von Studien- und Diplomarbeiten. Fundierte Methodenkenntnisse und Forschungserfahrung werden erwartet. Berufsbezogene Schwerpunkte der Ausbildung von Soziologen in Darmstadt sind Stadt- und Sozialplanung sowie Organisation/Personalwesen.

Zweck des Arbeitsverhältnisses ist auch, sich unter Betreuung wissenschaftlich zu qualifizieren mit dem Ziel der Promotion.

Schwerbehinderte werden bei gleicher Qualifikation bevorzugt.

Bewerber(innen) mit einem abgeschlossenen Studium der Soziologie, die an einem befristeten Arbeitsverhältnis interessiert sind, werden gebeten, die Bewerbung mit den üblichen Unterlagen unter Angabe der Kenn-Nr. 124 an den Dekan des Fachbereichs 2, Residenzschloß, 6100 Darmstadt, zu senden.

Die Technische Hochschule Darmstadt

Figur 0: Anzeige aus der ‚ZEIT' vom 6.8.82

Interessant ist die Annonce aus drei Gründen. Sie enthält

a) eine Beschreibung von Antecedensbedingungen (Prädiktoren) auf denen Bewerber z.B. in eine Rangreihe bezüglich ihrer Qualifikation gebracht werden können. Dazu zählt der Satz: „Fundierte Methodenkenntnisse und Forschungserfahrung werden erwartet."
b) eine Beschreibung der Konsequenzen (Kriterien). Der Bewerber muß eine Qualifikation für bestimmte Kriterien aufbringen. Zu den Kriterien gehören die Aufgaben, die der Kandidat *in der Zukunft* (d.h. in den nächsten 5 Jahren) durchführen soll. Darauf verweisen die Textstellen: „. . . für Aufgaben von begrenzter Dauer . . . (maximal 5 Jahre) zu besetzen. Zu den mit dieser Stelle verbundenen Aufga-

ben gehören die Mitarbeit in der Methodenausbildung, die Betreuung von Praktika der empirischen Sozialforschung, Studienberatung und Betreuung von Studien- und Diplomarbeiten.

c) eine unvollständige Beschreibung der Selektionsprozedur für den Fall mehrerer Bewerber. Anscheinend wird davon ausgegangen, daß man die Bewerber (irgendwie) auf einem evtl. kombinierten Prädiktor „Qualifikation" in eine Reihe bringen kann. Der „Beste" auf dem Prädiktor wird dann selegiert, weil man annimmt, daß der aus Methodenkenntnissen und Forschungserfahrung kombinierte Prädiktor X *reliabel* und für das Kriterium Y *valide* ist.

„Reliabel" bedeutet in diesem Zusammenhang, daß der Bewerber nicht unregelmäßige Schwankungen in Methodenkenntnissen und Forschungserfahrung hat. Er sollte z.B. nicht an Arbeitsstörungen leiden.

Interessant wird der Text der Anzeige bei der Beschreibung der Selektionsprozedur für den Fall gleicher Qualifikation. Sollten *alle* Bewerber entweder nicht behindert oder aber behindert sein, wird über die Auswahl nichts gesagt. Vielleicht entscheidet das Los. Das wäre bei gleicher Qualifikation gerecht. Der Auswahlmodus wird aber etwas mehr präzisiert, wenn mindestens ein Bewerber behindert ist. In diesem Fall sagt der Text:

„Schwerbehinderte werden bei gleicher Qualifikation bevorzugt."

Es wird also einmal angenommen, daß man die Gleichheit der Qualifikation feststellen kann und daß zum anderen der Behinderte einen Bonus bekommen soll. Die Begründung für den Bonus fehlt ebenso wie der Hinweis, ob der Bonus so groß ist, daß ein nur unwesentlich schlechter qualifizierter Schwerbehinderter ebenfalls einem Nichtbehinderten vorgezogen wird. Dieser letzte Punkt c) ist das zentrale Thema der Fairnessdiskussion:

Werden Angehörige unterschiedlicher Gruppen (Behinderte, Ausländer, Frauen aber auch arbeitslose Akademiker, Hochbegabte etc.) bei psychologisch fundierten Entscheidungen *fair* behandelt?

Dabei setze ich hier den Fall voraus, daß es Kriterien und Antecedensbedingungen explizit gibt. Sonst dürfte psychologische Beratung unmöglich sein.

Auf den Schul- oder Bildungsbereich übertragen, wäre z.B. folgende Situation denkbar. Ein Psychologe trifft Entscheidungen: d.h. er rät Eltern bzw. Kindern und Jugendlichen zu bestimmten Kursen, Ausbildungszweigen etc. zu oder ab. Die Beratung erfolgt auf der Kenntnis von Antecedensbedingung X (z.B. Kognitive Ausstattung, die mit Tests oder dem Lehrerurteil gemessen wird), der Kenntnis des Kriteriums Y („welche Anforderungen setzt Kurs A voraus?") und nach der Prognose des Psychologen (meist im Sinne von Wahrscheinlichkeitsaussagen). In das Dilemma der Fairness kommt er, wenn Haupt- und Realschüler sich gemeinsam beraten lassen und bestimmte Qualifikationsgleichheiten auftauchen. So wird immer wieder beklagt, daß Realschüler allen anderen die Stelle wegnehmen, selbst wenn Hauptschüler „eine gute Chance" hätten. Anscheinend geben solche Institutionen Bewerbern mit gleicher Qualifikation einen Bonus, wenn er Realschüler ist und einen Malus, wenn er Hauptschüler ist.

Um die Frage der Fairness von psychologischen Entscheidungen aufzurollen, werden im *Kapitel 2* zuerst verschiedene Bedingungsmodelle des (schulischen) Lernens

referiert. Dann werden im *Kapitel 3* Reliabilität und Validität mit besonderer Berücksichtigung zeitlicher Zusammenhänge dargestellt. Anschließend erfolgt im *Kapitel 4* eine Übersicht verschiedener Fairnesskonzeptionen, die im *Kapitel 5* unter einem gemeinsamen Blickwinkel betrachtet werden. Im *Kapitel 6* werden dann praktische Anweisungen gegeben: „Wie kann ich größere Fairness erreichen?"

2 Bedingungsmodell der Schulleistung

Zentrales Problem der Pädagogischen Psychologie ist die Erklärung und Vorhersage von Leistungsverhalten in der Schule. Dazu wurden eine Reihe von Modellen entwickelt (Bloom 1976; Carroll 1973 und Harnischfeger & Wiley 1977). Eine Darstellung der Modelle, die die Parallelität der Ansätze betont, indet sich bei Opwis & Gold (1982) und ist in Figur 1 wiedergegeben.

Allen Modellen ist gemeinsam, daß schulbezogene und nichtschulbezogene Schülercharakteristika sowie Variablen des Lehrprozesses als die wichtigsten Voraussetzungen der Schulleistung angesehen werden. Jedoch sind die Variablen und deren Beziehungen untereinander (noch) nicht ausreichend präzisiert (s.a. Treiber 1982).

Speziellere Modelle des schulischen Lernens, die besonders das Lernen im Zeitverlauf berücksichtigen, sind die Modelle von Atkinson (1974) und das Lernkomponentenmodell von Gagne (1968, 1973).

Gagne's Lernkomponentenmodell betont die Rolle fachspezifischer Vorkenntnisse bei der Erklärung von Lernleistung. Die Bedeutung dieser lehrgangsinternen Variablen nimmt dabei im Laufe der Zeit immer weiter zu. Jedoch wirken aufgrund des Modells lernprozeßexterne Basisfähigkeiten und lernprozeßimmanente spezifische Vorkenntnisse immer zusammen bei der Genese schulischer Leistungen.

Auf der Basis der o.a. theoretischen Modelle werden laufend empirische Studien erstellt, die meist vier Zielrichtungen besitzen (s.a. Weinert & Treiber 1981).

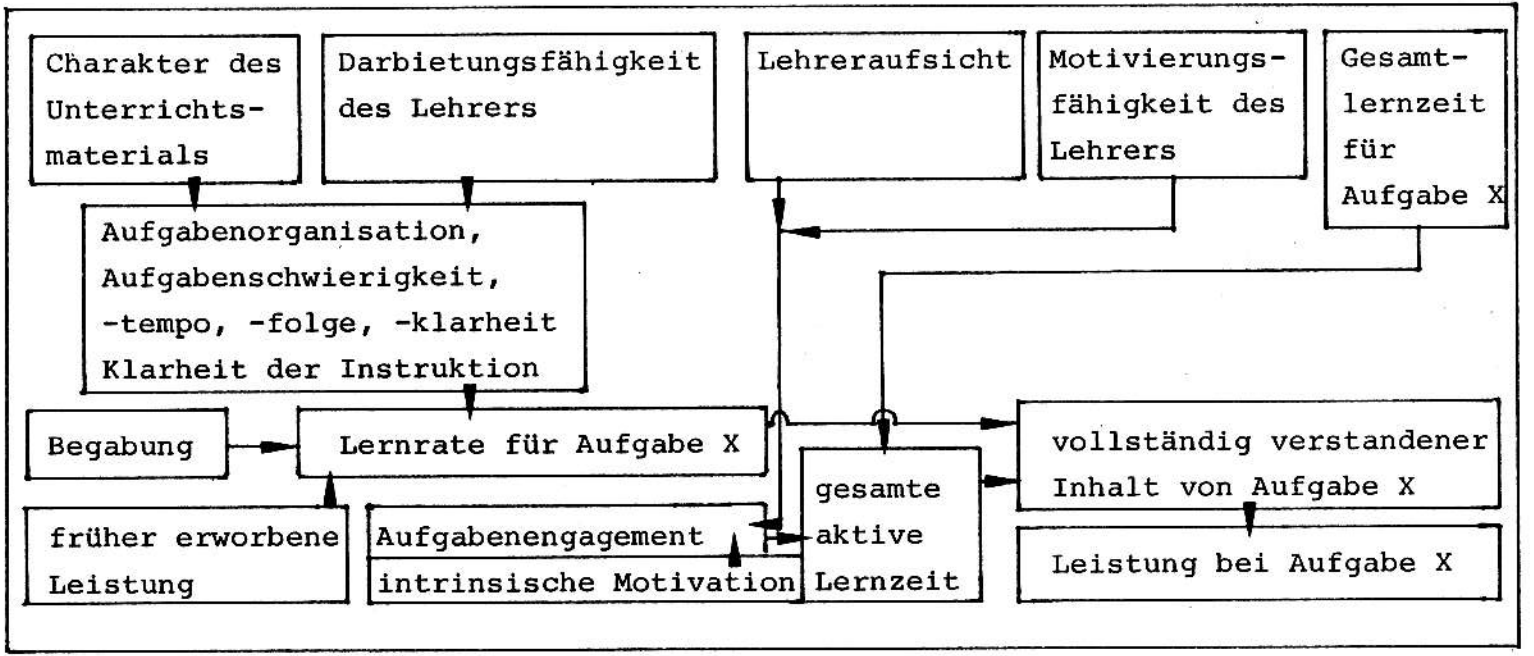
Schulleistungsunterschiede werden auf

1. *außerschulische Variable,* insbesondere individuelle Dispositionen (z.B. „Intelligenz")
2. *innerschulische Variable*
   (2a) die *wenig veränderbar* (z.B. „Lehrerpersönlichkeit") und
   (2b) die *veränderbar* sind (z.B. „Instruktionsmethode", „Lehrmaterialien" etc.)
3. *Wechselwirkungen* zwischen (1) und (2)

zurückgeführt. Bei Studien der Kategorie (1) werden Korrelationen zwischen Intelligenztestergebnissen als Prädiktoren und Schulleistungsindikatoren (Noten, Lehrerbeurteilungen, Schulleistungstests) als Kriterien zwischen 0.30 und 0.80 berichtet (Löschenkohl 1973).

Studien zur Kategorie (2a) weisen uneinheitliche, meist jedoch nicht vielversprechende Ergebnisse (Flanders 1970; Getzels & Jackson 1970; Solomon, Rosenberg & Bezdek 1964) auf.

Die Studien der Kategorie (2b) (Rosenshine & Furst 1973; Dunkin & Biddle 1974; Rosenshine 1976; Treiber 1980) gehören nach denen der Kategorie (3) (Flammer 1975, 1978; Cronbach & Snow 1977; Glaser 1977; Treiber 1981) zu den methodisch und inhaltlich anspruchsvollsten.



Figur 1: Schulisches Lernen (Das Modell von Harnischfeger & Wiley, 1976)

Jedoch sind auch hier die Befunde (noch) nicht so eindeutig, daß gezielt Einzelfallberatung allein aufgrund dieser Ergebnisse gemacht werden könnte. Eine Reihe bisher vernachlässigter Bedingungen schulischen Lernens wurden schon von Treiber & Weinert (1982, 264 ff.) zusammengetragen.

Für die praktische pädagogische Diagnostik scheinen im Moment jedoch zwei Variablenklassen die wichtigsten zu sein:

a) Schülerdispositionen
b) lernstandspezifische Vorkenntnisse

Zu letzterem Ergebnis ist neben Gagne (1962) u.a.a. Kleiter & Petermann (1977) und Zielinski (1980) gekommen.
„Die Wahrscheinlichkeit, daß die meisten Schüler mit Lernschwierigkeiten auch Vorkenntnislücken aufzuweisen haben, läßt es zweckmäßig erscheinen, die diagnostische Prozedur mit der Vorkenntnisprüfung zu beginnen" (Zielinski, 1980, S. 136).

Der Vorkenntnistest (z.B. lehrzielorientierter Test) läßt als Prädiktor eine „Prognose" auf die Vorkenntnisse des Schülers als Kriterium zu. Eine *echte* Prognose liegt aber nicht vor, weil keine Prognose *in die Zukunft* gemacht wurde. Ferner tritt hier kein Problem der Testfairness auf, weil keine *Plazierungsentscheidungen* gefällt werden.

Anders liegt der Fall beim allgemeinen Übertrittsverfahren, das im dreigliedrigen Schulsystem mit dem Ende der 4. Grundschulklasse in Gang gesetzt wird. Im Regelfall gibt der Grundschullehrer eine Empfehlung bezüglich des Übertritts ab. Die Empfehlung stützt sich auf folgende „Daten": (pädagogischer) Gesamteindruck, bisherige Schulleistungen, Abschneiden in speziellen, für die Übertrittsauslese entwickelten Probearbeiten in der 4. Klasse. Ist die Empfehlung des Grundschullehrers nicht eindeutig oder kollidiert sie mit dem Elternwunsch, wird ein Beratungslehrer um die psychologische Testung des Schülers gebeten.

Der Beratungslehrer steht nun vor vier Aufgaben:

a) *Testung* des Schülers
b) *echt zeitliche Prognose* der Schülerleistungen und des Schülerverhaltens in der 5. bzw. höheren Klassen
c) *Beratung* von Lehrern, Eltern und Schülern
d) Prüfung, ob das Vorgehen von a)–c) gegenüber dem Schüler und den Eltern einerseits und der aufnehmenden Schule andererseits *fair* ist.

Zum besseren Verständnis der Schritte a)–d) sind jetzt die Kapitel 3)–5) eingefügt. Dieses erscheint besonders deshalb notwendig, weil der Fairnessbegriff relativ neu ist, obwohl es schon einige deutschsprachige Publikationen hierzu gibt (Simons & Möbus 1979; Möbus 1978).

3.1 Parallelität, Reliabilität und Validität bei einem Meßzeitpunkt

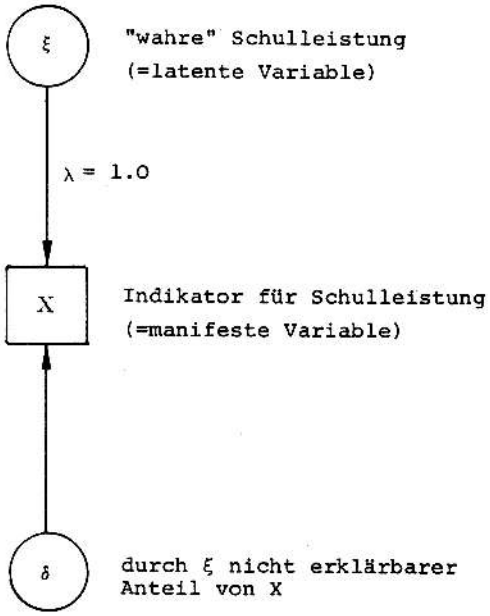
Obwohl „Reliabilität" und „Validität" im Gegensatz zur „Stabilität" gängige psychologische Konzepte sind, ist es für die Erörterung der Fairness psychologischer Entscheidungen (bzw. der „Testfairness) von Nutzen, alle drei Begriffe zu diskutieren.

Wir nehmen an, daß die „Schulleistung" von Schülern mit einem entsprechenden Test gemessen wird. Die „Schulleistung" ist *nicht direkt beobachtbar*, sondern nur über Indikatoren (z.B. Tests, Lehrerurteile etc.) *erschließbar*. Da sie nicht *direkt* beobachtbar ist, wird sie auch „Konstrukt" oder „latente Variable" genannt. Die Werte des Indikators dagegen sind direkt beobachtbar („Andreas hat 20 Punkte in dem xy-Test"). Der Indikator wird deshalb „manifeste Variable" genannt. Denken wir uns den Testwert X_i für den Schüler i jetzt zusammengesetzt aus der „Schulleistung" ξ_i (auch „wahrer Wert" genannt) und einem nicht erklärbaren Rest δ_i (auch als „Fehler" bezeichnet).

$$X_i = \lambda\xi_i + \delta_i \qquad (3.1)$$

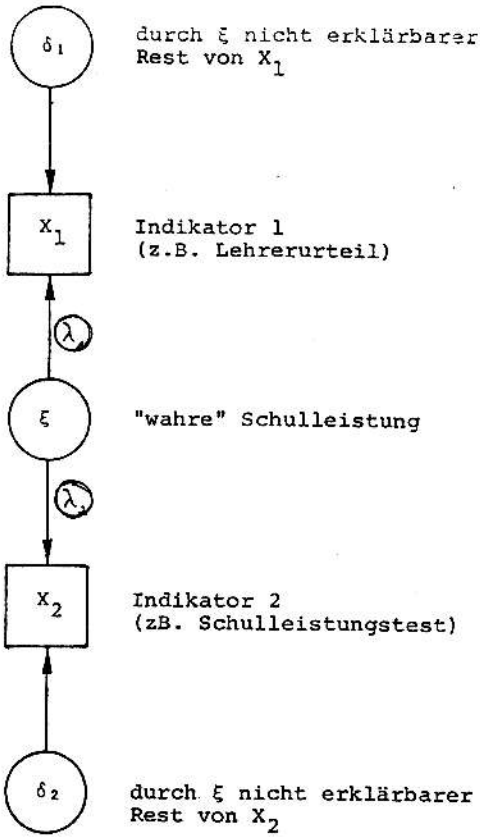
$(i = 1, 2, \ldots N)$
$(\lambda = 1.0 = \text{Skalenfaktor})$

Graphisch kann man (3.1) darstellen in einem Pfadmodell (Figur 2)



Figur 2: Meßwert X setzt sich aus dem „wahren" Wert und einem nicht erklärbaren Rest („Fehler") zusammen

Da man ξ und δ nicht kennt, hat (3.1) keinen praktischen Nutzen. Anders wird die Situation, wenn wir zu X einen strikt parallelen (Lord 1980) Test X' besitzen. Wir wollen X in X_1 und X' in X_2 umbenennen. Graphisch ist die Situation in Figur 3 dargestellt.



Figur 3: Zwei parallele Indikatoren einer latenten Variablen „Schulleistung"

Praktisch kann man einen parallelen Test evtl. durch die Split-half-Methode konstruieren: nach Schwierigkeit geordnete Aufgaben, werden abwechselnd Meßinstrument X_1 und X_2 zugeordnet. Hinweise hierzu finden sich bei Lienert (1969, 3. Auflage, S. 218ff.). Die Tests X_1 und X_2 sind nur parallel, wenn $\lambda_1 = \lambda_2$ und $\text{var}(\epsilon_1) = \text{var}(\epsilon_2)$. Dann ist nämlich $\text{var}(X_1) = \text{var}(X_2)$ (oder in anderer Schreibweise $\sigma^2_{X_1} = \sigma^2_{X_2} = \sigma^2_X$).

Korreliert man nun die beiden parallelen Tests X_1 und X_2, erhält man die *Reliabilität* als Varianzverhältnis:

$$\text{korr}(X_1, X_2) = \rho(X_1, X_2) = \frac{\text{Kovarianz}(X_1, X_2)}{[\text{Varianz}(X_1)]^{1/2}\,[\text{Varianz}(X_2)]^{1/2}} =$$

$$= \frac{\text{kov}(X_1, X_2)}{[\text{Var}(X_1)]^{1/2}\,[\text{Var}(X_2)]^{1/2}} = \frac{\text{kov}(\xi+\delta_1, \xi+\delta_2)}{[\text{Var}(X_1)]^{1/2}\,[\text{Var}(X_2)]^{1/2}} =$$

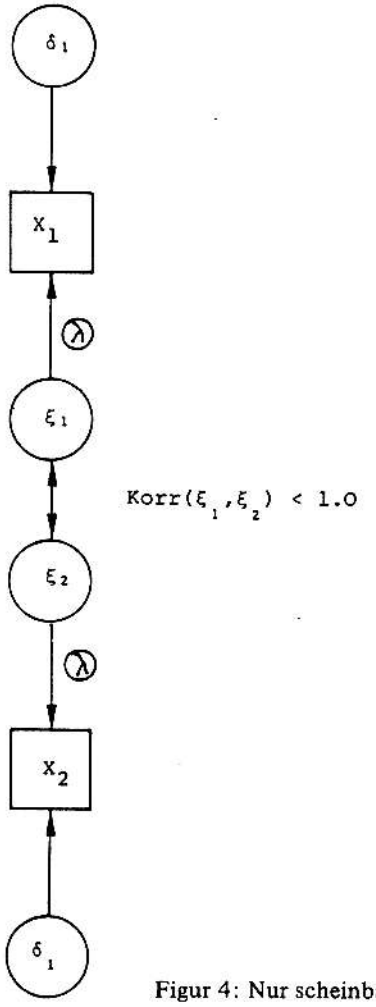
$$= \frac{\overset{=0}{\text{kov}(\xi, \xi)} + \overset{=0}{\text{kov}(\xi, \delta_2)} + \overset{=0}{\text{kov}(\delta_1, \xi)} + \text{kov}(\delta_1, \delta_2)}{\text{Var}(X)} = \frac{\text{Var}(\xi)}{\text{Var}(X)}$$

= Reliabilität

Die Terme im Zähler rechts sind Null, weil die latente Variable ξ unabhängig von den δ_1, δ_2 sein soll. Zudem sollen auch δ_1 und δ_2 voneinander unabhängig sein (im Pfaddiagramm in Figur 3 fehlen Pfeile zwischen diesen Variablen).

Der Bruch $0 \leq \frac{\mathrm{Var}(\xi)}{\mathrm{Var}(X)} \leq 1$, der über die Korrelation paralleler Tests gemessen werden kann, drückt das Verhältnis von wahrer Varianz zur gesamten Varianz aus. Sollte die Reliabilität im Idealfall gleich 1 sein, kann man bei parallelen Tests erwarten, daß alle Personen im Test X_1 und X_2 dieselben Meßwerte besitzen, bzw. die Testwerte entsprechen dann genau den unbekannten Werten ξ_i auf der latenten Skala. Ist dagegen die Reliabilität gleich Null, nutzt einem die Kenntnis der Testwerte X_1, X_2 nichts bei der Vorhersage von ξ.

In der Praxis steckt man aber immer noch in einem Dilemma. Man muß die Parallelität voraussetzen, um die Reliabilität bestimmen zu können. Es gibt aber in dieser Situation keinen sauberen Test für die Parallelität. Die Gleichheit der Varianzen $\sigma^2_{X_1} = \sigma^2_{X_2}$ ist nur notwendig aber nicht hinreichend, da auch folgendes Pfaddiagramm denkbar wäre (Figur 4):

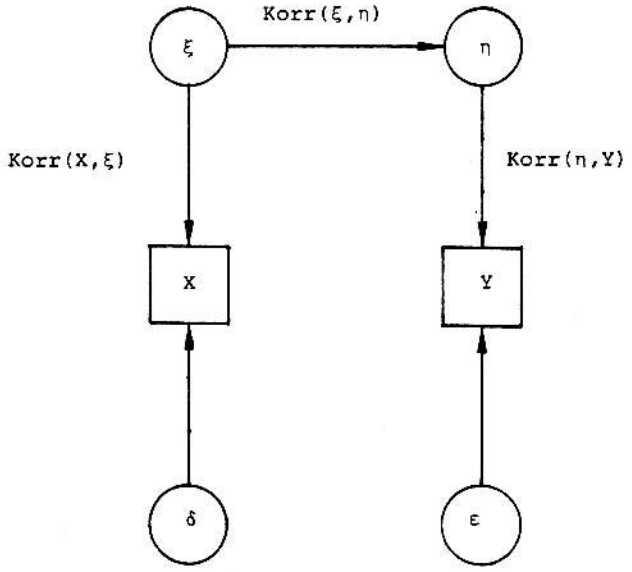


Figur 4: Nur scheinbar parallele Indikatoren X_1 und X_2

In Kapitel 3.3 wird gezeigt, wie man die Parallelität von Indikatoren prüfen kann.

Kommen wir nun zur *Validität*. Darunter versteht man die Korrelation des Tests (Prädiktors) mit einem Kriterium Y:

$$(3.3) \qquad \rho(X,Y) = \mathrm{korr}(X,Y) = \frac{\mathrm{kov}(X,Y)}{[\mathrm{Var}(X)]^{1/2}\,[\mathrm{Var}(Y)]^{1/2}}$$



Figur 5: Die korrelative Brücke der Validität r_{xy}

Die Validität läßt sich aufspalten, wie ein Blick auf Figur 5 und (3.4) zeigt: Die Validität setzt sich aus einer korrelativen „Brücke" zusammen, in die die Reliabilität von X und Y und die Korrelation der Konstrukte $\rho(\xi,\eta)$ hineinspielt.

$$(3.4) \qquad \mathrm{korr}(X,Y) = \underbrace{\mathrm{korr}(X,\xi)}_{\text{Reliabilität}(X)^{1/2}} \cdot \mathrm{korr}(\xi,\eta) \cdot \underbrace{\mathrm{korr}(\eta,Y)}_{\text{Reliabilität}(Y)^{1/2}}$$

$$= \frac{\mathrm{kov}(\xi+\delta,\xi)}{[\mathrm{Var}(X)]^{1/2}\,[\mathrm{Var}(\xi)]^{1/2}} \cdot \frac{\mathrm{kov}(\xi,\eta)}{[\mathrm{Var}(\xi)]^{1/2}\,[\mathrm{Var}(\eta)]^{1/2}} \cdot \frac{\mathrm{kov}(\eta,\eta+\epsilon)}{[\mathrm{Var}(\eta)]^{1/2}\,[\mathrm{Var}(Y)]^{1/2}}$$

$$= \frac{\mathrm{kov}(\xi,\eta)}{[\mathrm{Var}(X)]^{1/2}\,[\mathrm{Var}(Y)]^{1/2}} \cdot \frac{\mathrm{kov}(\xi+\delta,\xi)\,\mathrm{kov}(\eta,\eta+\epsilon)}{[\mathrm{Var}(\xi)]^{1/2}\,[\mathrm{Var}(\xi)]^{1/2}\,[\mathrm{Var}(\eta)]^{1/2}\,[\mathrm{Var}(\eta)]^{1/2}}$$

$$= \frac{\mathrm{kov}(\xi,\eta)}{[\mathrm{Var}(X)]^{1/2}\,[\mathrm{Var}(Y)]^{1/2}} \cdot \frac{\mathrm{Var}(\xi)\cdot\mathrm{Var}(\eta)}{\mathrm{Var}(\xi)\,\mathrm{Var}(\eta)}$$

$$= \frac{\mathrm{kov}(\xi+\delta,\eta+\epsilon)}{[\mathrm{Var}(X)]^{1/2}\,[\mathrm{Var}(Y)]^{1/2}} = \rho(X,Y) = \text{Validität}$$

Validität kann es nur geben, wenn X und Y Reliabilität besitzen.

Ähnlich, wie man die Validität im Sinne einer korrelativen Brücke interpretieren kann, ist die Zerlegung der Reliabilität möglich, wie man an Figur 3 und (3.5) sieht:

$$(3.5)\qquad \mathrm{korr}(X_1,X_2) = \underbrace{\mathrm{korr}(X_1,\xi)}_{\text{Reliabilität}(X)^{1/2}} \cdot \underbrace{\mathrm{korr}(\xi,X_2)}_{\text{Reliabilität}(X)^{1/2}}$$

$$= \frac{\mathrm{kov}(\xi+\delta_1,\xi)}{[\mathrm{Var}(X_1)]^{1/2}\,[\mathrm{Var}(\xi)^{1/2}} \cdot \frac{\mathrm{kov}(\xi,\xi+\delta_2)}{[\mathrm{Var}(\xi)]^{1/2}\,[\mathrm{Var}(X_2)]^{1/2}}$$

$$= \frac{\mathrm{Var}(\xi)\,\mathrm{Var}(\xi)}{\mathrm{Var}(X)\,\mathrm{Var}(\xi)} = \frac{\mathrm{Var}(\xi)}{\mathrm{Var}(X)} = \text{Reliabilität}$$

3.2 Reliabilität, Stabilität und Validität bei zwei Indikatoren und zwei Meßzeitpunkten

Wir wollen den Fall zweier Zeitpunkte und einer latenten Variablen betrachten. Diese Situation ist gegeben, wenn man die Schulleistung im 5. Jahr aufgrund der Kenntnisse der Schulleistung im 4. Jahr vorhersagen will. Das einfachste realistische Pfadmodell ist in Figur 6 aufgeführt.

Wir haben jeweils für jeden Zeitpunkt die Forderung nach strikter Parallelität fallengelassen (d.h. u.a. $\mathrm{Var}(X_1) \neq \mathrm{Var}(X_2)$ und $\mathrm{Var}(Y_1) \neq \mathrm{Var}(Y_2)$). Dennoch sollen X_1 und X_2 sowie Y_1 und Y_2 jeweils von einer latenten Variablen ξ und η abhängen. X_1 und X_2 sind kongenerische Tests (haben den gleichen Ursprung ξ). Ähnliches gilt für Y_1 und Y_2 mit Ursprung η.

Uns interessiert die Möglichkeit der Vorhersage von Zeitpunkt t–1 auf t: „Kann man von Testwerten in X_1 und X_2 zum Zeitpunkt (t–1) auf Testwerte in Y_1 und Y_2 zum Zeitpunkt t schließen?" Es wird dabei aufgrund der Überlegungen in Kapitel 3.1 klar, daß das nicht der Fall sein kann, wenn die Reliabilität von X oder Y Null ist. Zusätzlich muß aber auch die Korrelation $\rho(\xi,\eta)$ der latenten Variablen ξ (Schulleistung in Klasse 4) und η (Schulleistung in Klasse 5) hoch sein. Eine ideale Prognosemöglichkeit wäre gegeben, wenn perfekte Reliabilitäten mit einer perfekten Korrelation $\rho(\xi,\eta) = 1.0$ einhergehen würden.

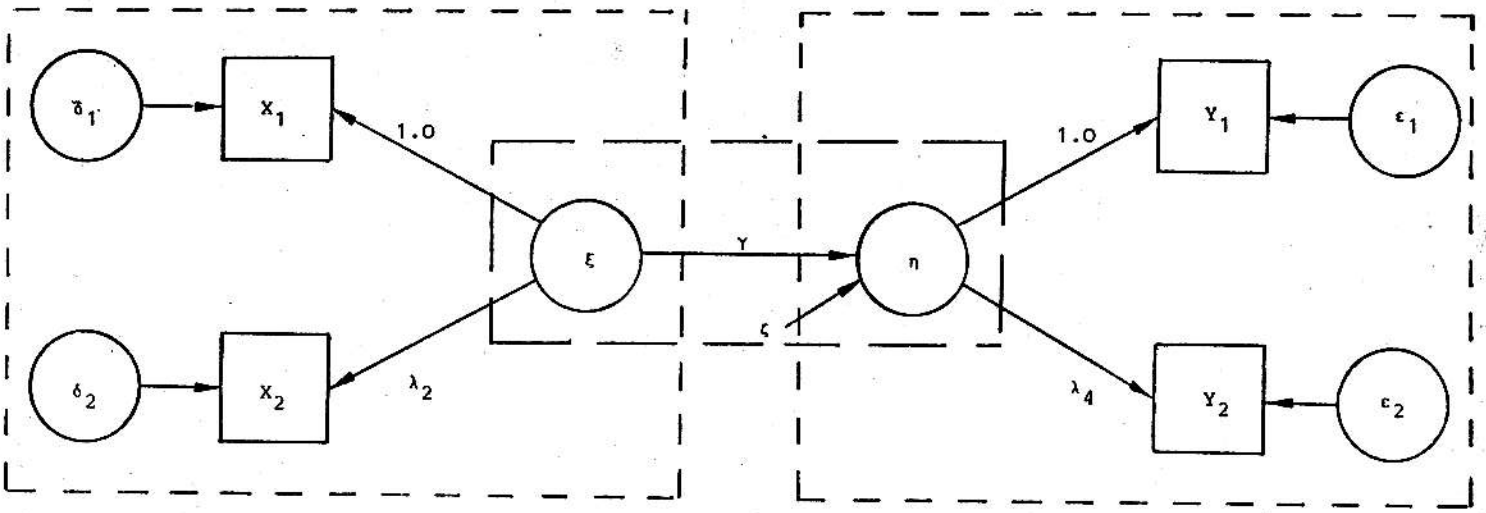
Kommen wir jetzt zu den Begriffen *„Validität"* und *„Stabilität"*. Unter Validität (Fischer, 1968, S. 35) versteht man, wie schon gezeigt, die Korrelation eines Tests X mit einem Kriterium Y.

$$(3.6)\qquad \rho(X,Y) = \frac{\mathrm{kov}(X,Y)}{[\mathrm{Var}(X)]^{1/2}\,[\mathrm{Var}(Y)]^{1/2}} = \text{Validität von X bezüglich Y}$$

Wir wollen die Terminologie aber noch etwas präzisieren und an X und Y Zeitindices anhängen, so daß die Validität sich auf eine bestimmte Zeitspanne $\Delta t = t - (t-1)$ bezieht.

$$\rho(X_{t-1},Y_t) = \text{Validität bezüglich der Zeitspanne } \Delta t$$

Solch eine zeitbezogene Validität ist nur möglich, wenn das Konstrukt ξ zeitlich stabil ist (d.h. $\rho(\xi,\eta) \approx 1.0$) (vgl. auch Figur 5).



Figur 6: Das einfachste Modell einer längsschnittlichen Betrachtung: 1 Gruppe von Personen, 1 latente Variable zu jedem Zeitpunkt, 2 Indikatoren für jede latente Variable, 2 Meßwellen und *unkorrelierte* Meßfehler

Wir wollen nun anhand des Pfadmodells in Figur 6 untersuchen, ob es uns möglich ist, folgende Fragen zu beantworten:

a) Sind die Tests X_1, X_2, Y_1, Y_2 reliabel
b) Ist das Konstrukt ξ zeitlich stabil, d.h. $\rho(\xi,\eta) = 1.0$
c) Sind die Vorhersagen von X_1, X_2 auf Y_1, Y_2 über die Zeitdistanz Δt prinzipiell möglich?

3.2.1 Annahme unkorrelierter Meßfehler: $\rho(\delta,\epsilon) = 0$

Zur Skalenfixierung der latenten Variablen nehmen wir an, daß sie in der Metrik der Indikatoren X_1, Y_1 gemessen sind ($\lambda_1 = \lambda_3 = 1$). Die Gleichungen des Meßmodells stehen in (3.7). Dahinter verbirgt sich die Annahme, daß die X_1, X_2 und Y_1, Y_2 jeweils kongenerische Indikatoren darstellen. Variable zwischen denen im Pfadmodell (Figur 6) keine Pfeilbeziehung besteht, werden als unabhängig angenommen. So sind z.B. δ_1 und ϵ_1 unkorreliert. Die Gleichungen des Längsschnittmodells lauten:

(3.7)
$$X_1 = \xi + \delta_1$$
$$X_2 = \lambda_2\xi + \delta_2$$
$$Y_1 = \eta + \epsilon_1$$
$$Y_2 = \lambda_4\eta + \epsilon_2$$

Die Indikatoren sind als Abweichungen vom jeweiligen Mittelwert gemessen.

Die „unerklärbaren Anteile" δ, ϵ beinhalten nicht nur Meßfehler sondern auch indikatorspezifische Anteile, die durch die latente Variable oder das Konstrukt (z.B. „Intelligenz", „Schulleistung") nicht erklärbar ist. Nach dem theoretischen Modell (3.7) bauen sich die Kovarianzen zwischen den Indikatoren aus den verschiedenen Parametern des Modells auf (s. (3.8)).

(3.8a)
$$\begin{bmatrix} \mathrm{var}(X_1) & \cdot & \cdot & \cdot \\ \mathrm{kov}(X_2,X_1) & \mathrm{var}(X_2) & \cdot & \cdot \\ \mathrm{kov}(Y_1,X_1) & \mathrm{kov}(Y_1,X_2) & \mathrm{var}(Y_1) & \cdot \\ \mathrm{kov}(Y_2,X_1) & \mathrm{kov}(Y_2,X_2) & \mathrm{kov}(Y_2,Y_1) & \mathrm{var}(Y_2) \end{bmatrix}$$

(3.8b)
$$\begin{bmatrix} \mathrm{Var}(\xi) + \mathrm{var}(\delta_1) & \cdot & \cdot & \cdot \\ \lambda_2\mathrm{var}(\xi) & \lambda_2^2\mathrm{var}(\xi) + \mathrm{var}(\delta_2) & \cdot & \cdot \\ \mathrm{kov}(\xi,\eta) & \lambda_2\mathrm{kov}(\xi,\eta) & \mathrm{var}(\eta) + \mathrm{var}(\epsilon_1) & \cdot \\ \lambda_4\mathrm{kov}(\xi,\eta) & \lambda_2\lambda_4\mathrm{kov}(\xi,\eta) & \lambda_4\mathrm{var}(\eta) & \lambda_4^2\mathrm{var}(\eta) + \mathrm{var}(\epsilon_2) \end{bmatrix}$$

Mittels 10 verschiedener Kovarianzen in (3.8a) müssen 9 Parameter identifiziert werden ($\lambda_2, \lambda_4, \mathrm{var}(\xi), \mathrm{var}(\eta), \mathrm{kov}(\xi,\eta), \mathrm{var}(\delta_1), \mathrm{var}(\delta_2), \mathrm{var}(\epsilon_1)$ und $\mathrm{var}(\epsilon_2)$. Aus den Parametern $\mathrm{kov}(\xi,\eta), \mathrm{var}(\xi)$ und $\mathrm{var}(\eta)$ läßt sich dann die Korrelation zwischen ξ und η berechnen (3.20).

Die Identifikation des Modells (3.7) kann in folgenden Schritten erfolgen:

(3.9) Kovarianz der Konstrukte $\mathrm{kov}(\xi,\eta) = \mathrm{kov}(Y_1,X_1)$

(3.10) Regressionskoeffizient („Ladung") im Meßmodell der 1. Welle:

$$\lambda_2 = \frac{\lambda_2\mathrm{kov}(\xi,\eta)}{\mathrm{kov}(\xi,\eta)} = \frac{\mathrm{kov}(Y_1,X_2)}{\mathrm{kov}(Y_1,X_1)}$$

(3.11) Regressionskoeffizient („Ladung") im Meßmodell der 2. Welle:

$$\lambda_4 = \frac{\lambda_4 \text{kov}(\xi,\eta)}{\text{kov}(\xi,\eta)} = \frac{\text{kov}(Y_2,X_1)}{\text{kov}(Y_1,X_1)}$$

Jetzt liegt eine einschränkende Bedingung vor, die das Modell überidentifiziert: das Produkt der in (3.9–3.11) berechneten Parameter muß gleich der beobachteten Kovarianz $\text{kov}(Y_2,X_2)$ sein:

(3.12) $$\lambda_2\lambda_4\text{kov}(\xi,\eta) \stackrel{!}{=} \text{kov}(Y_2,X_2)$$

Gilt diese Gleichheitsforderung nicht, wenn (3.8a) eine Populationsmatrix ist, stimmt das Modell oder eine seiner Annahmen nicht.

(3.13) Varianz des 1. Konstrukts:

$$\text{var}(\xi) = \frac{\lambda_2 \text{var}(\xi)}{\lambda_2} = \frac{\text{kov}(X_1,X_2)}{\lambda_2}$$

(3.14) Varianz des 2. Konstrukts:

$$\text{var}(\eta) = \frac{\lambda_4 \text{var}(\eta)}{\lambda_4} = \frac{\text{kov}(Y_2,Y_1)}{\lambda_4}$$

(3.15) Meßfehlervarianz oder spezifische Varianz von X_1:

$$\text{var}(\delta_1) = \text{var}(X_1) - \text{var}(\xi)$$

(3.16) Meßfehlervarianz oder spezifische Varianz von X_2:

$$\text{var}(\delta_2) = \text{var}(X_2) - \lambda_2^2\text{var}(\xi)$$

(3.17) Meßfehlervarianz oder spezifische Varianz von Y_1:

$$\text{var}(\epsilon_1) = \text{var}(Y_1) - \text{var}(\eta)$$

(3.18) Meßfehlervarianz oder spezifische Varianz von Y_2:

$$\text{var}(\epsilon_2) = \text{var}(Y_2) - \lambda_4^2\text{var}(\eta)$$

Regressionskoeffizient der strukturellen Gleichung $\eta = \gamma\xi + \zeta$

(3.19) $$\gamma = \frac{\text{kov}(\xi,\eta)}{\text{var}(\xi)} = \frac{\text{kov}(Y_1,X_2)}{\text{kov}(X_2,X_1)}$$

Es stellt sich nun die Frage, ob γ den maximal möglichen Wert angenommen hat, bzw. die Korrelation der Konstrukte 1 ist:

(3.20a) $$\text{Korr}(\xi,\eta) = \frac{\text{kov}(\xi,\eta)}{[\text{var}(\xi)^{1/2}\,[\text{var}(\eta)]^{1/2}} = \frac{\sqrt{\text{kov}(Y_2,X_2)}\,\sqrt{\text{kov}(Y_1,X_1)}}{\sqrt{\text{kov}(X_2,X_1)}\,\sqrt{\text{kov}(Y_2,Y_1)}}$$

(3.20b) $$= \frac{\sqrt{\text{kov}(Y_1,X_2)}\,\sqrt{\text{kov}(Y_2,X_1)}}{\sqrt{\text{kov}(X_2,X_1)}\,\sqrt{\text{kov}(Y_2,Y_1)}}$$

Perfekte Stabilität des Konstruktes ξ im korrelativen Sinne würde vorliegen, wenn (3.20) gleich 1 wäre. Dabei sind natürlich Niveauänderungen im Sinne von Mittelwertsverschiebungen nicht ausgeschlossen.

Reliabilitätsschätzungen der 4 Indikatoren:

(3.21)
$$\mathrm{Rel}(X_1) = \frac{\mathrm{var}(\xi)}{\mathrm{var}(X_1)} \qquad \mathrm{Rel}(Y_1) = \frac{\mathrm{var}(\eta)}{\mathrm{var}(Y_1)}$$

$$\mathrm{Rel}(X_2) = \frac{\lambda_2^2 \mathrm{var}(\xi)}{\mathrm{var}(X_2)} \qquad \mathrm{Rel}(Y_2) = \frac{\lambda_4^2 \mathrm{var}(\xi)}{\mathrm{var}(Y_2)}$$

3.2.2 Annahme korrelierter Fehler: $\rho(\delta, \epsilon) \neq 0$

Im Gegensatz zu (3.9–3.21) ist in vielenLängsschnittstudien die Korrelation einiger Fehlervariabler anzunehmen: Die Indikatoren X_1 und Y_1 sowie X_2 und Y_2 korrelieren auch dann noch, wenn man die latenten ξ und η Variablen konstant hält. Die Korrelation, die trotz dieser Partialisierung noch zu beobachten ist, kann auf die für Längsschnittstudien typischen Effekte (z.B. Erinnerungseffekte) zurückzuführen sein. Unter der Annahme dieser Korrelation kann man dieses Modell *graphisch* in einem geänderten Pfadmodell abbilden (s. Figur 7).

Bei einem Modell mit korrelierten Meßfehlern ändert sich die Kovarianzmatrix (3.8b) des Modells (3.7) an zwei Stellen zu:

(3.22)
$$\begin{bmatrix} \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \mathrm{kov}(\xi,\eta) + \mathrm{kov}(\delta_1,\epsilon_1) & \cdot & \cdot & \cdot \\ \cdot & \lambda_2\lambda_4\mathrm{kov}(\xi,\eta) + \mathrm{kov}(\delta_2,\epsilon_2) & \cdot & \cdot \end{bmatrix}$$

Da die Zahl der unbekannten Parameter größer als die Zahl der Varianzen bzw. Kovarianzen ist, kann das Modell nicht mehr identifiziert werden. Zur Sicherstellung der Identifikation muß eine zusätzliche Annahme gemacht werden. Es muß entweder λ_2, λ_4, $\mathrm{var}(\xi)$, $\mathrm{var}(\eta)$, $\mathrm{kov}(\xi,\eta)$, $\mathrm{kov}(\delta_1,\epsilon_1)$ oder $\mathrm{kov}(\delta_2,\epsilon_2)$ gleich einem von Null verschiedenen Wert gesetzt werden. Dann ist das Modell gerade identifiziert. Die sinnvollste Zusatzannahme kann gemacht werden, wenn im Meßmodell der 1. Welle nicht kongenerische sondern τ-äquivalente Tests (Lord & Novick, 1968, S. 47) vorliegen. In diesem Fall kann man $\lambda_2 = 1$ setzen. Wie sich weiter unten zeigt, bleiben aber erfreulicherweise die Koeffizienten γ, $\rho_{\xi\eta}$ und die Reliabilitäten der Indikatoren Y_1 und Y_2 invariant gegenüber einer Fixierung von λ_2. Das Modell läßt sich nun ähnlich wie in (3.9–3.21) identifizieren:

(3.23) Kovarianz der Konstrukte $\mathrm{kov}(\xi,\eta) = \mathrm{kov}(Y_1,X_2)$

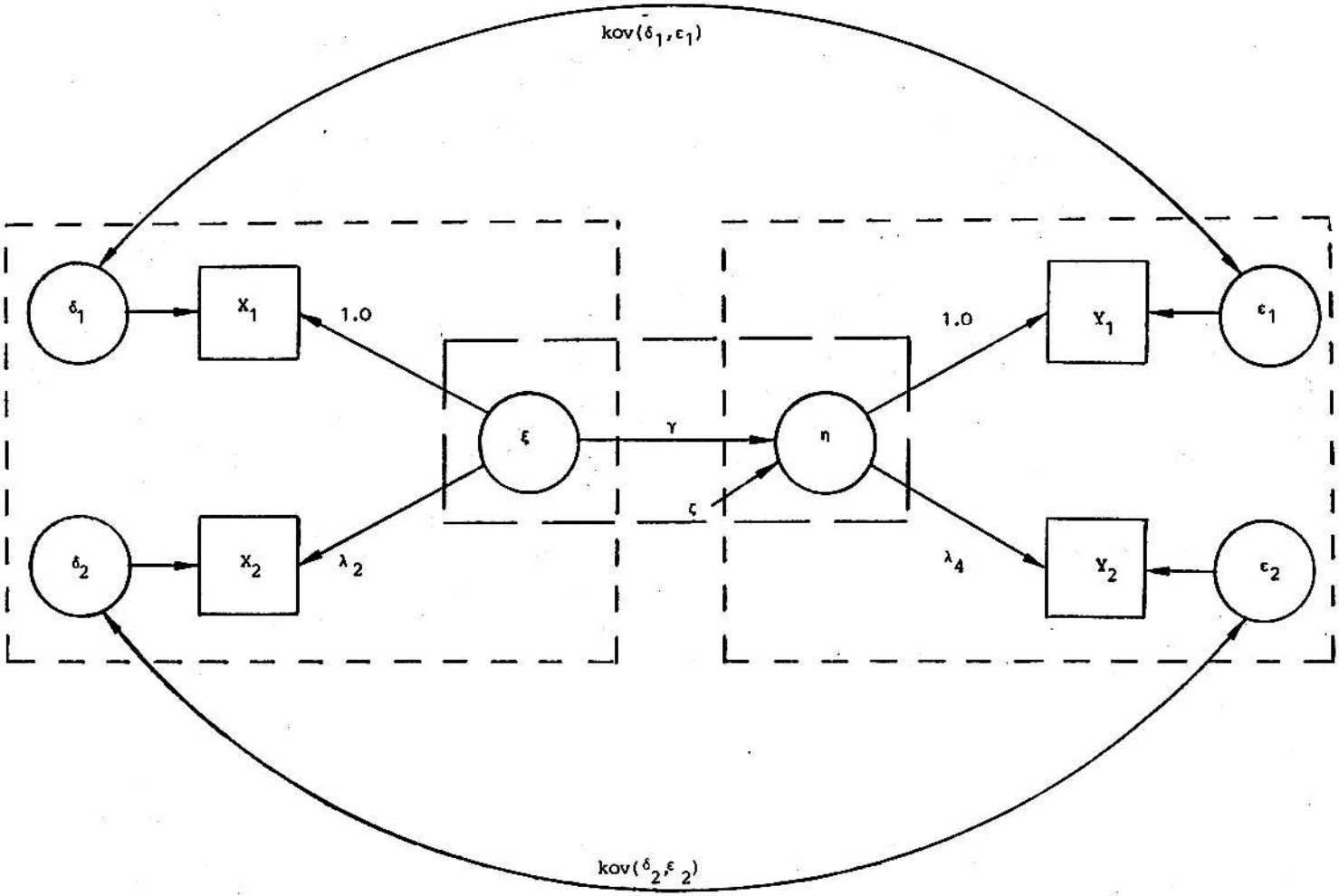
(3.24) Regressionskoeffizient („Ladung") im Meßmodell der 2. Welle:

$$\lambda_4 = \frac{\lambda_4\mathrm{kov}(\xi,\eta)}{\mathrm{kov}(\xi,\eta)} = \frac{\mathrm{kov}(Y_2,X_1)}{\mathrm{kov}(Y_1,X_2)}$$

(3.25) Varianz des 1. Konstrukts:
$\mathrm{var}(\xi) = \mathrm{kov}(X_2,X_1)$

(3.26) Varianz des 2. Konstrukts:
$\mathrm{var}(\eta) = \mathrm{kov}(Y_2,Y_1)/\lambda_4$

(3.27) Meßfehlervarianz oder spezifische Varianz von X_1:
$\mathrm{var}(\delta_1) = \mathrm{var}(X_1) - \mathrm{var}(\xi)$



Figur 7: Das einfachste Modell der Längsschnittanalyse mit korrelierten Fehlern

(3.28) Meßfehlervarianz oder spezifische Varianz von X_2:
$\text{var}(\delta_2) = \text{var}(X_2) - \text{var}(\xi)$

(3.29) Meßfehlervarianz oder spezifische Varianz von Y_1:
$\text{var}(\epsilon_1) = \text{var}(Y_1) - \text{var}(\eta)$

(3.30) Meßfehlervarianz oder spezifische Varianz von Y_2:
$\text{var}(\epsilon_2) = \text{var}(Y_2) - \lambda_4^2 \text{var}(\eta)$

(3.31) Kovarianz der spezifischen Anteile des 1. Indikators über die Zeit (Autokorrelation der Meßfehler)
$\text{kov}(\delta_1, \epsilon_1) = \text{kov}(Y_1, X_1) - \text{kov}(Y_1, X_2)$

(3.32) Kovarianz der spezifischen Anteile des 2. Indikators über die Zeit (Autokorrelation der Meßfehler)
$\text{kov}(\delta_2, \epsilon_2) = \text{kov}(Y_2, X_2) - \text{kov}(Y_2, X_1)$

Regressionskoeffizient der strukturellen Gleichung $\eta = \gamma\xi + \zeta$

$$(3.33) \qquad \gamma = \frac{\text{kov}(\xi,\eta)}{\text{var}(\xi)} = \frac{\lambda_2 \text{kov}(\xi,\eta)}{\lambda_2 \text{var}(\xi)} = \frac{\text{kov}(Y_1, X_2)}{\text{kov}(X_2, X_1)}$$

An (3.33) sieht man, daß die Wahl des λ_2 irrelevant für die Bestimmung von γ ist!
Korrelation der Konstrukte (= *Stabilitätsschätzung* des 1. Konstrukts):

$$(3.34) \qquad \rho_{\xi\eta} = \text{korr}(\xi,\eta) = \frac{\sqrt{\text{kov}(Y_1, X_2)}\,\sqrt{\text{kov}(Y_2, X_1)}}{\sqrt{\text{kov}(X_2, X_1)}\,\sqrt{\text{kov}(Y_2, Y_1)}}$$

Die Korrelation (3.34) ist unabhängig von den Regressionsparametern λ_2, λ_4 der Meßmodelle der 1. und 2. Welle.

Reliabilitäten der 4 Indikatoren:

$$(3.35) \qquad \text{Rel}(X_1) = \frac{\text{var}(\xi)}{\text{var}(X_1)} \qquad \text{Rel}(Y_1) = \frac{\text{var}(\eta)}{\text{var}(Y_1)}$$

$$\text{Rel}(X_2) = \frac{\text{var}(\xi)}{\text{var}(X_2)} \qquad \text{Rel}(Y_2) = \frac{\lambda_4^2 \text{var}(\eta)}{\text{var}(Y_2)}$$

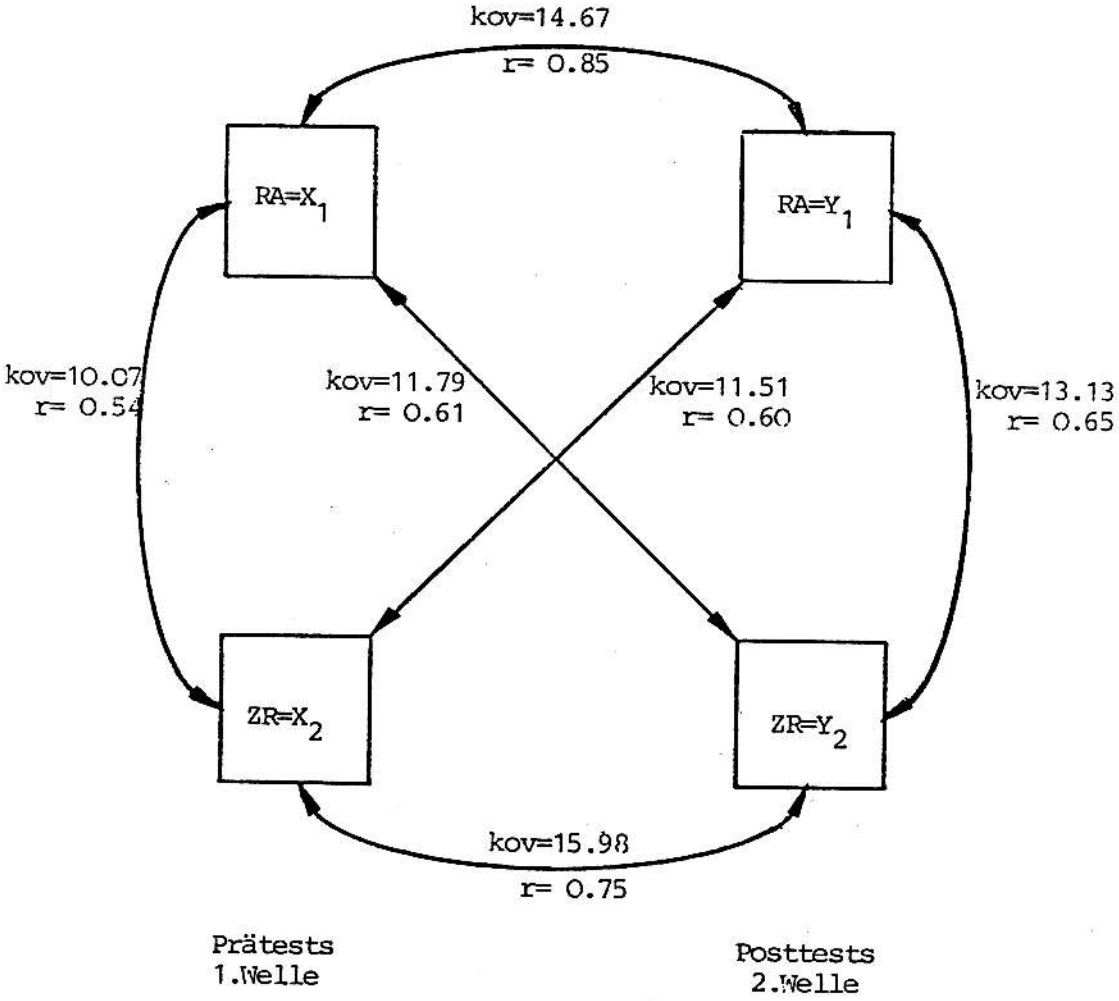
Die Fixierung von λ_2 wirkt sich nur auf die Reliabilitätsschätzung von X_2 aus.

Korrelation der Fehler:

$$(3.36) \qquad \text{korr}(\delta_i, \epsilon_i) = \frac{\text{kov}(\delta_i, \epsilon_i)}{\sqrt{\text{var}(\delta_i)}\,\sqrt{\text{var}(\epsilon_i)}} \qquad (i = 1, 2)$$

Tab 1: Mittelwerte, Kovarianzen und Korrelationen der vier Indikatoren

$\bar{X}$	Kovarianzen				Korrelationen			
X_1 8.98	16.81				1.000			
X_2 9.55	10.07	20.43			0.540	1.000		
Y_1 10.19	14.67	11.51	17.80		0.850	0.600	1.000	
Y_2 11.38	11.79	15.98	13.13	22.56	0.610	0.750	0.650	1.000

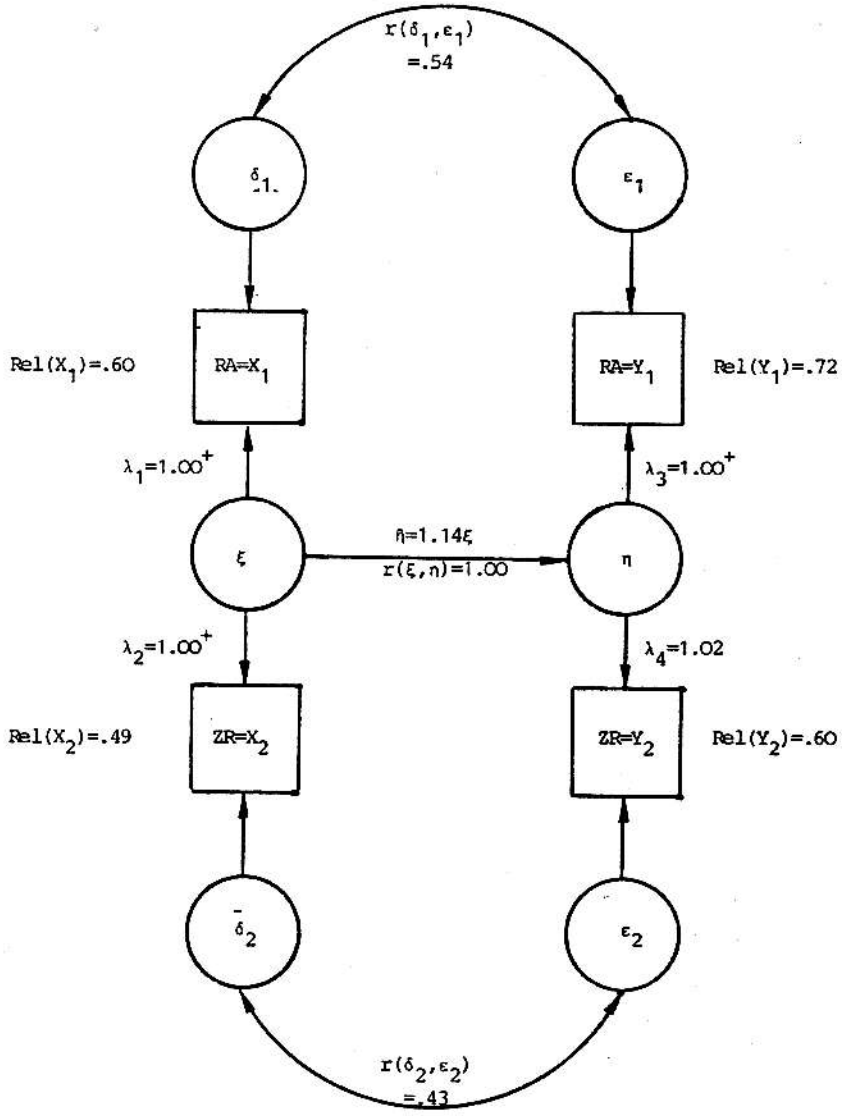


Figur 8a: Kovarianzen und Korrelationen für ein einfaches „Fehler-in-den-Variablen & Fehler-in-der-Gleichung" (2 IST-Untertests)

Die in (3.9–3.36) allgemein gehaltenen Gedanken sollen jetzt an einem Beispiel konkretisiert werden.

In einer Untersuchung von Simons & Möbus (1977) zur Auswirkung eines Intelligenztrainings auf Entscheidungen in Rehabilitationsprogrammen wurden 9 IST-Untertests (Form A und B) zu zwei Zeitpunkten (Prä- und Posttests) gemessen. In einer späteren Auswertung dieser Daten wurden u.a. Faktorladungen nach der Maximum-Likelihood-Methode geschätzt und PROMAX-rotiert.

Aus dieser Faktorenstruktur ist ersichtlich, daß man eventuell die Untertests RA (Rechenaufgaben) und ZR (Zahlenreihen) als kongenerische Tests ansehen könnte. Ich habe diesen Subaspekt herausgegriffen (s. Figur 8a und Tab. 1). Die dabei zugrunde liegende Forschungshypothese lautet: Durch das Training wird das RA und ZR „unterliegende" Konstrukt strukturell nicht verändert. RA und ZR messen im Vor- und Nachtest dasselbe. Dabei sind Niveauerhöhungen zugelassen.

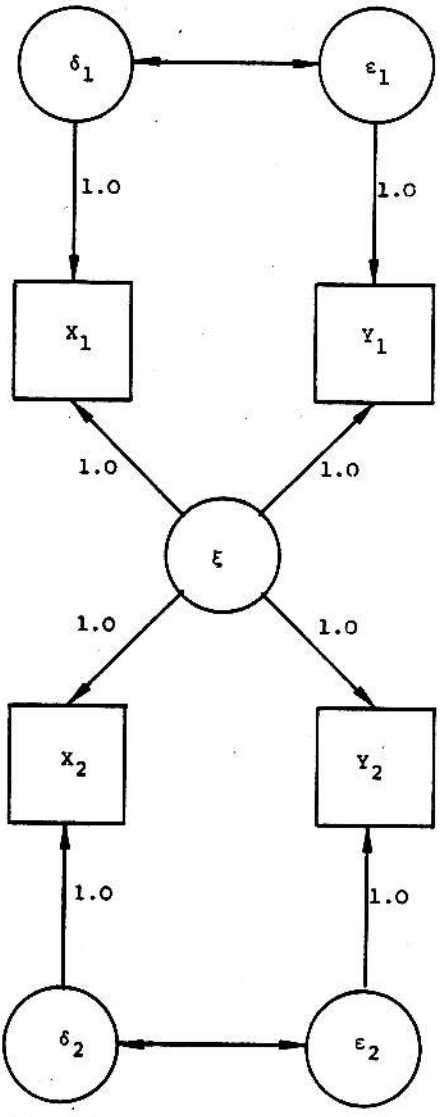


+zur Sicherung der Parameteridentifikation a priori gleich 1.00 gesetzt (d.h. es wird angenommen, daß X_1 und X_2 ξ-äquivalent sind)

Figur 8b: Kombiniertes „Fehler-in-den-Variablen & Fehler-in-der-Gleichung" Modell mit korrelierten Fehlern für die IST-Untertests RA (Rechenaufgaben) und ZR (Zahlenreihen) für 2 Meßzeitpunkte

Die Parameter„schätzungen" für das Modell mit *korrelierten* Fehlern sind graphisch in Figur 8b und die Berechnungen nach (3.9–3.20) sind unter (3.9′–3.20′) ausführlich dargestellt.

Zuerst sollen die Parameterschätzungen des Modells mit *unkorrelierten* Fehlern bestimmt werden:



Figur 9: Streng parallele Prädiktor- und Kriteriumsindikatoren mit evtl. korrelierten Meßfehlern

(3.9)′ $\hat{kov}(\xi,\eta) = 14.67$

(3.10)′ $\hat{\lambda}_2 = 11.51/14.67 = 0.785$

(3.11)′ $\hat{\lambda}_4 = 11.79/14.67 = 0.804$

(3.12)′ $0.785 * 0.804 * 14.67$ soll gleich 15.98 sein; tatsächlich ist aber das Produkt links gleich $9.26 \ll 15.98$

Das Modell scheint sich nicht sonderlich gut an die Daten anzupassen, da die theoretisch vorhergesagte Kovarianz 9.26 nur 57% der empirisch beobachteten 15.98 beträgt.

(3.13)′ $10.07/0.785 = 12.82$

(3.14)′ $13.13/0.800 = 16.41$

(3.15)′ $16.81 - 12.82 = 3.99$

(3.16)′ $20.43 - 7.90 = 12.53$

(3.17)′ $17.80 - 16.41 = 1.39$

(3.18)′ $22.56 - .804^2 \cdot 16.41 = 11.95$

(3.19)′ $\hat{\gamma} = 11.51/10.07 = 1.14$

(3.20)′ $$\rho_{\xi\eta} = \frac{\sqrt{11.51}\sqrt{11.79}}{\sqrt{10.07}\sqrt{13.13}} \approx 1.0$$

Es ist interessant zu beobachten, daß, wenn man die Stabilität $\rho(\xi,\eta)$ des Konstrukts ξ nicht über (3.20b) oder (3.34) sondern über die klassische „Verdünnungsformel" der Testtheorie (coefficient of attenuation; Lord 1980, S. 7) schätzt und dabei für die Reliabilität $r_{XX'}$ von X die Korrelation $r_{X_1X_2} = .54$ und für die Reliabilität $r_{YY'}$ von Y die Korrelation $r_{Y_1Y_2} = .65$ einsetzt, unsinnige Ergebnisse auftreten:

$$\text{Verdünnungsformel: } r_{XY} = \rho_{\xi\eta} \cdot [r_{XX'}]^{1/2} \cdot [r_{YY'}]^{1/2}$$

wobei: $r_{XX'} := r_{X_1X_2} = .54$

$r_{YY'} := r_{Y_1Y_2} = .65$

$r_{XY} := r_{X_1Y_1} = .85$

wobei := bedeutet „wird ersetzt durch"

Nach der Verdünnungsformel würde sich für die „wahre" Stabilität ein Wert von:

$$\rho_{\xi\eta} = .85 / \left([.54]^{1/2}[.65]^{1/2}\right) = 1.43$$

ergeben. Eine so hohe Korrelation ist natürlich unsinnig. Der Grund hierfür liegt nicht allein darin, daß die Reliabilitäten zu niedrig geschätzt wurden, sondern daß in die Korrelation r_{XY} sowohl die Stabilität $\rho_{\xi\eta}$ als auch die Korrelation der Meßfehler $\text{korr}(\delta_1,\epsilon_1)$ und $\text{korr}(\delta_2,\epsilon_2)$ (s.a. Figur 7 und 9 eingegangen sind. Die Korrelation der Meßfehler, die die Validität r_{XY} in einem längsschnittlichen Zusammenhang hochdrückt, wird in der Verdünnungsformel nicht berücksichtigt. Der gleiche Effekt ist bei Verwendung von (3.20a) zu beobachten.

3.3 Der Test als „Stellvertreter" für das Kriterium: Strikt parallele Prädiktor- und Kriteriumsindikatoren

Die Bestimmung der Parameter des Strukturgleichungsmodells nach der in 3.2.1 und 3.2.2 dargestellten Methode weist zwei Mängel auf:

a) Ist das Modell überidentifiziert (mehr Kovarianzen als Parameter), bietet das Rückrechnen der Parameter aus den Kovarianzen keine statistisch saubere Schätzmethode.

b) Die wichtige Stabilitätskorrelation $\rho(\xi,\eta)$ läßt sich mit den oben erwähnten Methoden nicht daraufhin abtesten, ob deren Schätzung signifikant von 1.0 abweicht. Üblich, aber wenig sinnvoll, ist der Signifikanztest gegen Null.

Diese Gründe führen zur Maximum-Likelihoodschätzung der Parameter. Als Daten

benötigen wir die Varianz-Kovarianzmatrix S der vier Indikatoren X_1, X_2, Y_1 und Y_2.

$$S = \begin{bmatrix} \mathrm{Var}(X_1) & \cdot & \cdot & \cdot \\ \mathrm{Kov}(X_2,X_1) & \mathrm{Var}(X_2) & \cdot & \cdot \\ \mathrm{Kov}(Y_1,X_1) & \mathrm{Kov}(Y_1,X_2) & \mathrm{Var}(Y_1) & \cdot \\ \mathrm{Kov}(Y_2,X_1) & \mathrm{Kov}(Y_2,X_2) & \mathrm{Kov}(Y_2,Y_1) & \mathrm{Var}(Y_2) \end{bmatrix}$$

Die Matrix S sei eine Schätzung der Populationskovarianzmatrix Σ.

Wir formulieren nun zwei Modelle. Das erste Modell nennen wir *idiographisches Modell.* Es wird tatsächlich nicht geschätzt, sondern dient nur zur Verdeutlichung einiger Gedankengänge. Das idiographische Modell enthält soviele unbekannte Parameter, daß ein beliebiges S unter seiner Kontrolle erzeugt sein könnte. Bei geeigneter Parameterschätzung kann aus dem idiographischen Modell eine Matrix $\hat{\Sigma}_i$ berechnet werden, der exakt S entspricht. Die Anpassung an die Daten ist perfekt! Insbesondere ist auch die Korrelation $\rho(\xi, \eta)$ unbestimmt gelassen worden.

Das zweite Modell *(„eingeschränktes Modell")* enthält weniger Parameter. Es enthält auch die Hypothese, daß $\rho(\xi,\eta) = \lambda_1 = \lambda_2 = \lambda_3 = \lambda_4 = 1.0$ ist (das bedeutet τ-Äquivalenz) und daß zusätzlich die Fehlervarianzen $\mathrm{Var}(\delta_1) = \mathrm{Var}(\delta_2) = \mathrm{Var}(\epsilon_1) = \mathrm{Var}(\epsilon_2)$ alle gleich sind (das bedeutet „strikte Parallelität"). Dieses Modell erzeugt eine Matrix Σ_e (3.22), die gegenüber Σ_i eine eingeschränkte Struktur besitzt. Liegen statt der Parameter Schätzungen vor, erhält man aus dem Modell $\hat{\Sigma}_e$. Mit einem χ^2-Test überprüft man die Abweichungen von S und $\hat{\Sigma}_e$ auf Signifikanz und damit die Hypothese, ob die Stichprobendaten S aus einer Population mit Matrix Σ_e oder Σ_i gezogen wurden.

Die Annahme strikter Parallelität läßt sich in folgendes Pfad- und Strukturmodell kleiden:

$$\begin{aligned} X_1 &= \xi + \delta_1 \\ X_2 &= \xi + \delta_2 \\ Y_1 &= \xi + \epsilon_1 \\ Y_2 &= \xi + \epsilon_2 \end{aligned} \qquad \text{mit gleichen Fehlervarianzen}$$

Die Populationskovarianzmatrix strikt paralleler Test besitzt dabei folgende Struktur:

$$\Sigma_e = \begin{bmatrix} \mathrm{Var}(\xi) + \mathrm{Var}(\delta_1) & \cdot & \cdot & \cdot \\ \mathrm{Var}(\xi) & \mathrm{Var}(\xi) + \mathrm{Var}(\delta_2) & \cdot & \cdot \\ \mathrm{Var}(\xi) + \mathrm{Kov}(\delta_1,\epsilon_1) & \mathrm{Var}(\xi) & \mathrm{Var}(\xi) + \mathrm{Var}(\epsilon_1) & \cdot \\ \mathrm{Var}(\xi) & \mathrm{Var}(\xi) + \mathrm{Kov}(\delta_2,\epsilon_2) & \mathrm{Var}(\xi) & \mathrm{Var}(\xi) + \mathrm{Var}(\epsilon_2) \end{bmatrix} \quad (3.22)$$

Das Computerprogramm LISREL (Jöreskog & Sörbom, 1981) schätzt die Parameter $\mathrm{Var}(\xi)$, $\mathrm{Var}(\delta_1) = \mathrm{Var}(\delta_2) = \mathrm{Var}(\epsilon_1) = \mathrm{Var}(\epsilon_2)$, $\mathrm{Kov}(\delta_1,\epsilon_1)$ und $\mathrm{Kov}(\delta_2,\epsilon_2)$. Aus den Schätzungen wird dann die Matrix $\hat{\Sigma}_e$ (3.22) zurückgerechnet. Ein Anpassungstest prüft dann die Haltbarkeit des Modells (d.h. die Annahme strikter Parallelität).

Anpassungsfunktion:

$$F = [\ln |\hat{\Sigma}_e| + \mathrm{Spur}\,\{S \cdot \hat{\Sigma}_e^{-1}\} + \ln |S| - p]$$

$p = 4$ = Zahl der Indikatoren

Die Nullhypothese (Hypothese strikter Parallelität) wird verworfen, wenn:

$$\chi^2 = N \cdot F > \chi^2_{1-\alpha, df = p_{idi} - p_{ein}} = 10 - 4 = 6$$

p_{idi} = unabh. Parameter im idiographischen Modell (hier 10)

p_{ein} = unabh. Parameter im eingeschr. Modell (hier: 4)

Ist die Anpassung gut ($\chi^2 < \chi^2_{1-\alpha, df}$), wird die Hypothese $\lambda_1 = \lambda_2 = \lambda_3 = \lambda_4 = \rho(\xi, \eta) = 1.0$ beibehalten und die Indikatoren gelten als strikt parallel. *Damit können die X-Indikatoren als Stellvertreter für die Y-Indikatoren angesehen werden.*

Die Analyse der Daten aus Tabelle 1 mittels LISREL ergab für die Annahme strikter Parallelität ohne korrelierte Meßfehler einen chi-Quadratwert von $\chi^2_{df=8} = 60.85$, der hochsignifikant war, so daß das Modell abgelehnt werden mußte. Läßt man dagegen die Korreliertheit der Meßfehler zu (nach 3.22 und Figur 9), erhalten wir ein chi-Quadrat von $\chi^2_{df=6} = 6.93$, das selbst auf dem Signifikanzniveau von $\alpha = 0.30$ nicht signifikant wäre. Es läßt sich also das Modell aus Figur 9 für diese Daten halten.

Näheres zum Hypothesenprüfen mittels LISREL findet sich bei Möbus & Nagl (1983).

Später in der Fairnessdebatte wird es von Wichtigkeit sein, die Gleichheit von Parametern *zwischen* Gruppen zu prüfen. Auch das kann mit LISREL gemacht werden. Dazu wird die Anpassungsfunktion für jede Gruppe j berechnet und über alle Gruppen aufsummiert.

4 Überblick über Fairnesskonzeptionen

4.1 Einführung in die Terminologie

Zunächst soll eine begriffliche Abklärung erfolgen. Man unterscheidet einerseits den „*Item*bias" von dem „Bias" oder der „Unfairness" im *Test* oder in der psychologischen *Entscheidung* (Beratung).

Unter „Itembias" versteht man die Ungerechtigkeit die durch Aufgaben mit unterschiedlicher Schwierigkeit für verschiedene Bevölkerungsgruppen (Petersen 1980; Scheuneman 1980) auftreten. So würde z.B. zur Prüfung der Feinmotorik eine Häkelaufgabe im Test eine Benachteiligung von Jungen oder Männern bedeuten. Das Item hätte einen geschlechtspezifischen Bias. Der Itembias läßt sich nur bei der Neukonstruktion eines Tests beseitigen. Man verwendet dazu das in Fischer (1974) ausführlich dargestellte Rasch-Modell.

Wegen der beiden letztgenannten Gründe wollen wir uns auf die Fairness des Tests und der psychologischen Entscheidung beschränken. Bis zu der Untersuchung von Cleary (1968) schloß man von unerwünschten Unterschieden im Kriterium (z.B. Schulnoten) auf die Unerwünschtheit von Gruppenunterschieden im Test (sogenanntes „Identitätskonzept" der Testfairness). Jede Mittelwertsdifferenz im Test zwischen sozialen Schichten und ethnischen Gruppen wurde auf die diskriminierende Konstruktion des Tests zurückgeführt. Konsequent ging man daran, die „Unfairness" der traditionellen Intelligenztests durch die Konstruktion von „culture-fair"- oder „status free"-Tests auszuschalten. Die Nivellierung der Mittelwertsunterschiede gelang jedoch auch hier nicht vollständig. Darüber hinaus scheint es aber auch geradezu unsinnig, zu

glauben, Populationsmittelwerte, die die intellektuelle Lerngeschichte der Angehörigen unterschiedlicher Populationen widerspiegeln, sollten durch den Testkonstrukteur ausgeschaltet werden. Dieser Ansicht ist z.B. Guthke (1972, S. 25f.).

Es steht hier nicht so sehr die Fairness des Test- (bzw. Item-)Inhaltes sondern seines *Gebrauches* zur Diskussion: „Sind die auf seiner Basis angestellten Schlußfolgerungen fair gegenüber Individuen, Gruppen und Institutionen? Da die Wünsche der Getesteten, Beratungsuchenden, gesellschaftlich relevanten Gruppen und Testverwendern nur schwer miteinander in Einklang zu bringen sind, wird auch das Thema Testbias oder Testfairness bis heute in der Fachliteratur kontrovers behandelt (Cleary 1968; Darlington 1971; Einhorn & Bass 1971; Thorndike 1971; Jensen 1973; Linn 1973; Cole 1973; Cleary, Humphereys, Kendrick & Wesman 1975; Gross & Su 1975; Breland & Ironson 1976; Cronbach 1976; Darlington 1976; Hunter & Schmidt 1976; Linn 1976; Möbus 1976; Novick & Petersen 1976; Petersen & Novick 1976; Sawyer, Cole & Cole 1976; Somons & Möbus 1976; Gösslbauer 1977; Möbus 1978; Petersen 1980; Novick 1980; Scheuneman 1980; Jensen 1980 und die Erwiderung von 28 Diskutanten).

Einigkeit konnte jedoch darin erzielt werden, daß bei Prognose und Entscheidungsbildung die Beziehung zwischen Test und Kriterium auf ihre Fairness überprüft werden muß. Es verlagert sich also die Betrachtung von Mittelwertsunterschieden im Test auf die Analyse der Beziehung zwischen Kriterium und Test: Kann für mehrere Populationen eine gemeinsame Beziehung akzeptiert werden oder müssen mehrere getrennte Beziehungen angenommen werden, die populationsspezifische Aussagen treffen? Führen gleiche Testwerte für Angehörige verschiedener Gruppen zu gleichen Entscheidungen oder werden trotz gleicher Prädiktorwerte Personen unterschiedlich behandelt werden müssen, um Testfairness zu erzielen?

Da der Test nur sinnvoll eingesetzt werden kann, wenn er eine „Stellvertreterfunktion" erfüllt, ist ihm ein Kriterium an Wichtigkeit vorgeordnet. Zu diesem Kriterium wird ein Test konstruiert, der dann später bei Prognosen etc. „stellvertretend" für das Kriterium die Beurteilung der Personen übernimmt. Mit den Termini der Testtheorie kann man die Situation auch anders schildern. *Test und Kriterium sollten Indikatoren ein- und desselben Konstrukts sein.* Im Laufe der Testfairnessdebatte haben sich eine ganze Reihe von möglichen Validierungskriterien ergeben, die eine beachtliche Bandbreite besitzen. Sie reichen vom Genotyp (bei Jensen) über Schulleistungen bis zu berufsspezifischen Kompetenzen (bei McClelland). Ist z.B. der Test am Genotyp validiert (hohes h^2), kommt ihm nach Jensen erst das Etikett „kulturfrei" oder „fair" zu.

Bei der Betrachtung der verschiedenen Fairnesskonzepte werden nach Hunter & Schmidt (1976) *drei ethische Grundpositionen* bezogen, die für die widersprüchliche Beurteilung der Fairnessdefinitionen verantwortlich sind: „unqualified individualism", „qualified individualism", „fair share". Die Anhänger dieser drei verschiedenen Positionen verstehen auch dementsprechend etwas anderes unter „Diskriminierung" bzw. „Fairness". Unter der ersten Position werden Gruppenunterschiede im Kriterium, die durch Prädiktoren nicht erklärt und vorhergesagt werden können, nicht ignoriert. Würde man diese Unterschiede verleugnen, würde das einer Diskriminierung der im Kriterium besseren Gruppen gleichkommen. Ferner wird die Kenntnis der Gruppenzugehörigkeit bei Entscheidungen über Personen benutzt: *gruppenspezifische* Regressionen, cutoffs, Tests etc. Nach der zweiten Position werden Gruppenunterschiede (z.B. zwischen Status-, ethnischen oder Geschlechtsgruppen) ignoriert. Auch werden

keine Informationen über die Gruppenzugehörigkeit bei der Entscheidung benutzt: *ein* cutoff, *eine* Regression, *ein* Test für *alle* Gruppen bzw. *keine* differentielle Validität. Eine Beachtung von Gruppendifferenzen würde die im Kriterium schlechtere Gruppe unfair behandeln. Die Vertreter der dritten Gruppe sehen dann eine Diskriminierung gegeben, wenn bei Selektionen (z.B. auch bei Aufnahme in Förderungs- oder Rehabilitationsprogramme etc.) die gesellschaftlich relevanten Gruppen nicht mit angemessenen Quoten („fair share") beteiligt sind. Es ist klar, daß die Maximierung der Validität für die Anhänger der zweiten und dritten Position nur eine untergeordnete Rolle spielt.

Ich neige insofern dem qualifizierten Individualismus zu, indem ich im Gegensatz zu fast allen Autoren auf diesem Gebiet gruppenspezifische cutoffs im Test sowie die differentielle Validität ablehne, jedoch wie die Anhänger des unqualifizierten Individualismus eventuelle Gruppenunterschiede in Test und Kriterium nicht leugnen will. Darüber hinaus bietet die „fair-share"-Auffassung einige attraktive Aspekte, was sich besonders in der Analyse der Thorndikeschen Position niederschlägt. Thorndike (1971) wird von mir im Gegensatz zu einigen Autoren durch die Einführung einer neuen Betrachtungsebene (Regressionskonzepte) wieder aufgewertet.

4.2 Klassische Testfairnesskonzepte

Zwischen 1968 und heute haben sich in der psychologischen Forschung neben dem „Identitätskonzept" (Darlingtons Fall IV) eine Reihe von Definitionen der Testfairness herausgeschält, die sich z.T. widersprechen und nur im – wohl unrealistischen – Fall der perfekten Validität des Tests identisch sind.

a) die Definition von Cleary & Anastasi (OLS[1]-Regressionskonzept von Y auf X):

„A test is biased for members of a subgroup of the population if, in the predicition of a criterion for which the test was designed, consistent nonzero errors of prediction are made for members of the subgroup. In other words, the test is biased if the criterion score predicted from the common regression line is consistently too high or too low for members of the subgroup. With this definition of bias, there may be a connotation of „unfair", particularly if the use of the test produces a prediction that is too low." (Cleary 1968, S. 115)

„In the psychometric sense, test bias refers to overprediction or underprediction of criterion measures. If a test consistently underpredicts criterion performance for a given group, it shows unfair discrimination of „bias" against the group." (Anastasi 1968, S. 559)

Demnach ist der Test fair, wenn die Gleichungen

$$(4.1) \qquad E(Y|x_{i*}) = y^* = a_i + b_i x_{i*} \qquad (i = 1, \ldots, g)$$

gelten, wobei:

y^* = minimal zufriedenstellendes Abschneiden auf dem Kriterium (Kriteriums-cutoff),

1 OLS-Regression: damit ist die ordinary least squares regression (= Methode der kleinsten Quadrate) gemeint.

x^*_i = minimal zufriedenstellendes Abschneiden (Niveau, cutoff) auf dem Test für Gruppe i,

a_i = Regressionskonstante für Gruppe i,

b_i = Regressionsgewicht für Gruppe i,

g = Zahl der Gruppen.

Sind die gruppenspezifischen Regressionen gleich (weichen nicht signifikant voneinander ab), kann die gemeinsame Regression zur Bestimmung eines einzigen – für alle Gruppen gültigen – cutoffs x^* herangezogen werden.

Eine empirische Arbeit aus dem deutschsprachigen Raum, die sich dieses Konzepts bedient, findet sich bei Wottawa & Amelang (1980).

Diese Definition blieb bis 1971 ohne Widerspruch. Dann aber wurden im gleichen Jahr vier Alternativen angeboten: das Modell gleichen Risikos (Einhorn & Bass 1971), das OLS-Regressionskonzept von X auf Y (Darlington's Fall III), das Modell des modifizierten Kriteriums (Darlington 1971) und das konstante Verhältnismodell von Thorndike (Thorndike 1971). Cleary's Definition weist zwar Vorteile auf (Maximierung der Validität, der Gesamttrefferquote und der erwarteten Kriteriumsleistung, Minimierung der Schätzfehler pro Gruppe), jedoch glaubten Einhorn & Bass, Darlington und Thorndike etliche Mängel entdeckt zu haben, die Neuformulierungen notwendig erscheinen ließen. Eine Zusammenfassung der Kritik an Cleary von Thorndike (1971), Darlington (1971), Linn (1971) findet sich bei Möbus (1978, S. 196f.).

Die stärkste Ähnlichkeit mit Cleary's Definition hat

b) das Modell gleichen Risikos von Einhorn & Bass (1971)

Einhorn & Bass geben keine direkte Definition, sondern beschreiben eine Prozedur, mit der Diskrimination durch Tests verhindert werden kann:

„. . . One approach to the problem of the test discrimination which is advocated here, involves the use of separate cutting scores on the predictor tests so that the criterion means, the different regression lines, and standard errors of estimate are used to arrive at a nondiscriminatory procedure . . ." (Einhorn & Bass 1971, S. 266) und „. . . This method is nondiscriminatory since there is no over- or underprediction of criterion scores. In addition, the probability of success for persons selected from the different groups is necessarily the same, since the standard errors of estimate have been taken into account in making selection decisions . . ." (Einhorn & Bass 1971, S. 268).

Formal läßt sich dieses Vorgehen fassen in:

$$(4.2) \qquad P[Y>y^* \mid X = x^*_i] = Z \text{ für } i = 1, \ldots g$$

Die cutoffs x^*_i müssen so gesetzt werden, daß die minimale Erfolgswahrscheinlichkeit (positive Korrelation von X und Y vorausgesetzt) für alle Gruppen einem bestimmten Wert Z entspricht. Die Prozedur ist umfassender als die Cleary-Definition, bezieht sie doch die differentielle Prädiktibilität (verschiedene Residualstreuungen bzw. Standardschätzfehler) mit in das Kalkül ein.

Positiv an dem Vorgehen sind die ersten Ansätze entscheidungstheoretischer Überlegungen: So die Steuerung des Risikos bei Fehlentscheidungen. Negativ (jedenfalls vom Standpunkt des Individuums bzw. des Bewerbers aus) ist die Risiko- und Kosten-

minimierung der Institution auf Kosten des Individuums (dem Bewerber wird keine „Bewährungschance" gegeben). Ferner werden Personen, die aus Gruppen mit großer nicht vorhersagbarer Heterogenität (großer Standardschätzfehler, niedrige Validität) stammen (so z.B. bei durch Lehrer gut geförderten Klassen; siehe Simon & Möbus 1976,) dadurch benachteiligt, daß die cutoffs x_i^* im Gegensatz zu anderen Gruppen (z.B. schlecht geförderte Klassen) wesentlich höher liegen.

Das Modell gleichen Risikos ist mit dem Cleary-Konzept verwandt, da bei gleichen gruppenspezifischen Standardschätzfehlern und Regressionen die beiden Modelle identische Ergebnisse liefern können (bei Z = 0.5).

Mit den beiden Definitionen a) und b) hat ferner das Modell des kultur-modifizierten Kriteriums von Darlington (1971) Ähnlichkeit.

c) das Modell des kultur-modifizierten Kriteriums (Darlington 1971)

Darlington ließ sich – über Cleary hinausgehend – von zwei Überlegungen leiten: 1. Es kann einen Bias im Kriterium geben, d.h. die Mittelwertsunterschiede im Kriterium sind durch gesellschaftliche Gegebenheiten verfälscht bzw. verzerrt. Dieser Bias muß beseitigt werden bzw. es muß angegeben werden, welche Mittelwertsdifferenz wünschenswert, sinnvoll und vertretbar ist. 2. Es kann sinnvoll sein, gesellschaftlichen Gruppen für bisher „erlittene" Benachteiligung einen Bonus einzuräumen, d.h. die wünschenswerte Differenz zwischen den Mittelwerten auf Y zu verringern.

„. . . that the term ‚cultural fairness' be replaced in public discussions by the concept of ‚cultural optimality'. The question of whether a test is culturally optimum can be divided in two: a subjective, policy-level question concerning the optimum balance between criterion performance and cultural factors (operationalized in our equations as the optimum value of k), and a purely empirical question concerning the test's correlation with the culture-modified criterion variable $(Y\text{-}kZ^2)$ and whether that correlation can be raised. Anyone who objects to a test on the latter ground can be invited to construct a test correlating higher with $(Y\text{-}kZ)$." (Darlington 1971, S. 80)

Für positive und negative Kritik am kulturmodifizierten Kriterium sei wieder auf Möbus (1978, S. 198f.) verwiesen. Dort findet sich auch ein Hinweis, der die Beziehung zum Identitätskonzept betrifft.

In dem oben erwähnten Artikel stellt Darlington noch ein anderes Konzept vor, das von ihm wohl mehr als Vervollständigung seiner Ideen angesehen wurde, dem wir aber grundsätzlichere Bedeutung einräumen wollen. Dieses Konzept wird interessant, wenn man sich den Argumenten McClellands's (1973) anschließt und Tests durch Kriteriumssampling zur Kompetenzprüfung (statt Intelligenzprüfung) konstruiert. Es ist

d) das OLS-Regressionskonzept von X auf Y (Darlington's Fall III)

„. . . we define a test X as culturally fair only if Z (culture) has no direct effect on X, independent of Y (criterion)." (Darlington 1971, S. 75)

Das bedeutet aber, daß die Regressionen von X auf Y gleich sind:

$$(4.3) \qquad E_i(X|Y) = E(X|Y) \text{ für alle } i = 1, \ldots, g$$

[2] Z = kulturelle oder ethnische Drittvariable (männl./weibl., Status etc.).

Die typischen Testwerte $E_g(X|Y)$ sollen innerhalb jeder Kriteriumsgruppe gleich sein. Die nichtvalide Varianz soll unabhängig von der kulturellen Drittvariablen Z sein.

Eine andere Formulierung wäre (bei gleichen gruppenspezifischen Standardschätzfehlern): Für jede Erfolgsgruppe (d.h. für jeden Kriteriumswert) soll die Akzeptanzwahrscheinlichkeit (= Selektionswahrscheinlichkeit) für alle kulturellen Gruppen gleich sein

(4.4) $\quad P_i[X > x^* | Y] = P[X > x^* | Y]$ für alle $i = 1, \ldots, g$

Diese Definition läßt einen Test als fair zu, der fehlerbehaftet ist. Die Fehler müssen Zufallsfehler und unabhängig von anderen identifizierbaren Variablen sein. Sehen wir das Kriterium als einen unendlich langen homogenen Test an und nehmen wir entsprechend den Vorschlägen von McClelland (1973) ein Kriteriumssampling für die Items des Tests vor, sind die Items eine Zufallsstichprobe aus dem langen Test. Der Testwert hängt nur vom wahren Wert und einem Zufallsfehler ab. Der kurze Test erfüllt dann das Fairnesskriterium, da man innerhalb jeder Kriteriumsgruppe nicht mehr auf individuelle Testwerte schließen kann (auch dann nicht, wenn man die Gruppenzugehörigkeit kennt): Für jede Kriteriumsgruppe liegen identische Testwertverteilungen vor. Diese bedingten Testwerte sind unabhängig von allen anderen Variablen, d.h. man kann innerhalb dieser Erfolgsgruppe bei Kenntnis der Testwerte nicht auf die Gruppenzugehörigkeit schließen.

Eine etwas anders geartete Interpretation, die die Fähigkeitsstrukturen stärker ins Blickfeld rückt, könnte folgendermaßen aussehen. Um im Kriterium erfolgreich zu sein, bedarf es mehrerer Fähigkeiten. Ist die partielle Korrelation $r_{xz \cdot y} \neq 0.0$, dann besteht in den Testergebnissen zwischen den kulturellen Gruppen ein größerer Unterschied als er auf Grund des Unterschieds auf dem Kriterium zu erwarten wäre. Folglich muß der Test Fähigkeiten messen, die zwar nicht relevant für das Kriterium sind, in denen sich aber die Gruppen unterscheiden. Daher ist der Test diskriminierend und unfair.

Hängt die Gruppenvariable Z (Kultur, Rasse etc.) mit dem Kriterium zusammen (die Gruppenrandverteilungen auf Y überlappen sich nur teilweise), würde der Kurztest nach Cleary unfair sein, weil die OLS-Regressionen von Y auf X für die Gruppen verschieden sind. Nach Cleary würde die Annahme einer gemeinsamen OLS-Regression von Y auf X die im Kriterium niedrigere Gruppe bevorzugen.

Die Kritik hierzu von Hunter & Schmidt (1976) ist wieder bei Möbus (1978, S. 200) dargestellt.

Im gleichen Jahr kam dann von Thorndike (1971) eine weitere Definition hinzu, die als erste den „fair-share"-Gesichtspunkt betont und nicht soviel Wert auf die Validitätsmaximierung legt:

e) Thorndike's konstantes Verhältnismodell

„. . . An alternate definition (of fairness) would specify that the qualifying scores on a test should be set at levels that will qualify applicants in the two groups in proportion to the fraction of the two groups reaching a specified criterion performance." (Thorndike 1971, S. 63)

Formal läßt sich die Definition darstellen als:

$$(4.5) \qquad K = \frac{P_g[X_g > x_g^*]}{P_g[Y_g > y^*]} = \frac{P_h[X_h > x_h^*]}{P_h[Y_h > y^*]} \quad \text{für alle} \quad g,h = 1, \ldots, G$$

Ist $K = 1$, liegt, wie wir später noch sehen werden, ein wichtiger Spezialfall der Thorndikeschen Definition vor.

Die Definition (4.5) stellt eine Abkehr vom validitätsmaximierenden „Individualismus" und eine Hinwendung zu einem gruppenorientierten Quotensystem dar: Berufschancen werden nach einem Gruppenschlüssel verteilt. Dabei operiert man mit dem Test so, als ob er exakt dem Kriterium entsprechen würde, d.h. als ob der Test perfekt valide wäre.

Thorndike ging von der Beobachtung aus, daß bei der cutoff-Bestimmung nach Cleary eine Gruppe im Test keine Selektionschance haben könne, obwohl einige Mitglieder durchaus Erfolg im Kriterium hätten, wenn sie zugelassen würden.

Thorndike's Argumentation ist – wie später immer wieder ignoriert wird – *konstruktorientiert und gilt daher für uns als richtungsweisend.* Unter der Voraussetzung gleicher gruppenspezifischer Streuungen $s_g(X) = s_g(Y)$ und $s_h(X) = s_h(Y)$ sind nach Thorndike gleiche Mittelwertsdifferenzen fair (Der Test als „Spiegelbild" des Kriteriums):

„. . . Here the minor group differs as much from the major on the criterion as it does on the predictor test. Under these circumstances, the two groups differ as much with respect to the factors that are unique to the test as they do with respect to its unique nonvalid variance, as well as with respect to its valid variance. If we were to use any specified critical test score and apply it to both groups, a smaller percentage of the minor group than of the major group would be accepted, but this smaller percentage would in this case exactly parallel the smaller percentage reaching a specified level of performance on the criterion measure." (Thorndike 1971, S. 68)

Ein offensichtlicher Nachteil der Prozedur scheint zu sein, daß die Kriteriumsleistungen der Ausgewählten niedriger sind als dieses nach Cleary der Fall wäre. Ist z.B. die Quote der Erfolgreichen in Gruppe A gleich 16% und in Gruppe B gleich 48%, so sollten nach Thorndike die Quoten der Akzeptierten 1: 3 sein (z.B. 10% : 30%). Ist aber die Validität nicht perfekt, sind die selegierten Personen oft nicht diejenigen, die erfolgreich wären. Man könnte Thorndike vorwerfen, er würde blind einem Quotenmechanismus auf Kosten der individuellen Leistung (besonders bei den Mitgliedern besserer Gruppen) das Wort reden.

„. . . one must attempt to explain to applicants with objectively higher qualifications why they were not admitted – a rather difficult task and from the point of individualism an unethical one." (Hunter & Schmidt 1976, S. 1059)

Zwei Jahre später kamen dann noch zwei weitere Quotenmodelle von Cole (1973) und Linn (1973) hinzu.

f) das bedingte Wahrscheinlichkeitsmodell von Cole (1973)

„. . . The basic principle of the conditional probability selection model is that for both minority and majority groups whose members can achieve a satisfactory criterion score ($Y > y^*$) there should be the same probability of selection regardless of group membership." (Cole 1973, S. 240)

Formal läßt sich die Definition darstellen als:

(4.6) $K = P[X_g > x_g^* | Y_g > y^*] = P[X_h > x_h^* | Y_h > y^*]$
für alle $g, h = 1, \ldots, G$

Dabei ist:

(4.7) $$P[X>x^*|Y>y^*] = \frac{P[X>x^*]}{P[Y>y^*]} P[Y>y^* | X>x^*]$$

mit: $P[X>x^*|/P[Y>y^*]$ als Thorndike's Bruch.

Cole's Argumentation ist hauptsächlich gegen den Subjektivismus des Darlington-Modells gerichtet und versucht die Wichtigkeit des (unveränderten) Kriteriums mit der Fairness für Minoritätsgruppen zu verbinden. Sie ist der Ansicht, nicht objektivierbare Ermessensspielräume zu vermeiden und trotzdem den in Test und Kriterium schlechteren Gruppen größere Selektionschancen einzuräumen. Dabei geht sie davon aus, daß es im Interesse eines Bewerbers ist, selegiert zu werden und eine Chance der Bewährung zu bekommen. Diesem Ziel steht das Regressionsmodell (Cleary) oder das Modell gleichen Risikos (Einhorn & Bass) entgegen. Bei Cole liegen die cutoffs im Prädiktor für die benachteiligten Gruppen niedriger als bei den anderen Ansätzen (Ausnahme: bei $k < 0$ im culture modified criterion von Darlington).

Von allen Autoren (bis auf Darlington's Fall III) schwächt Cole die Bedeutung des Ergebnisses eines nicht perfekt validen Tests am meisten ab. Entscheidend ist vielmehr der mögliche Erfolg im Kriterium.

Eine dazu spiegelbildliche Auffassung vertritt Linn (1973). Petersen & Novick (1976) nennen sie:

g) das Gleichwahrscheinlichkeitsmodell (Linn 1973)

In der normalen diagnostischen Situation steht als Information über den Bewerber nicht seine zukünftige Entwicklung (Erfolg oder Mißerfolg im Kriterium) sondern sein gegenwärtig beobachtbarer Prädiktorwert zur Verfügung.

„. . . Thus from one point of view, it would seem reasonable to propose a definition of culture fair selection based on the conditional probability of success given selection. One might argue that all applicants who are selected should be guaranteed an equal, or fair chance of being successful, regardless of group membership." (Petersen & Novick 1976, S. 13).

Formal läßt sich das Modell beschreiben als:

(4.8) $P[Y_g \geq Y^* | X_g \geq x_g] = P_h[Y_h \geq y^* | X_h \geq x_h^*]$

Einige Autoren setzten unzulässigerweise Modell d) mit f) und a) mit g) gleich. Dabei wurde außer Acht gelassen, daß bei f) und g) in die jeweiligen bedingten Wahrscheinlichkeiten Abschnitte von Randverteilungen eingehen (z.B. $P[\ldots|X \geq x^*]$ oder $P[\ldots|Y \geq y^*]$), nicht aber wie bei a) und d) nur bestimmte Ereignisse (z.B. $P[\ldots|X=x^*]$ oder $P[\ldots|Y=y^*]$). Auf jeden Fall ist a) ein Spezialfall von g) und d) von f).

Es bereitet keine Schwierigkeit, das Modell von Einhorn & Bass durch die Vertauschung von X und Y in

h) das Modell gleichen Risikos (X auf Y)

umzuwandeln. Danach müßten die cutoffs x_i^* so gesetzt werden, daß die Selektionswahrscheinlichkeit am cutoff y^* für alle Gruppen gleich ist:

(4.9) $$P[X_g > x_g^* | Y = y^*] = P[X_h > x_h^* | Y = y^*]$$

Dieses Konzept wäre etwas allgemeiner als Darlingstons's Fall III.

i) konverse Modelle

Betrachtete man bei a)–h) auf dem Kriterium statt der Erfolgs- die Mißerfolgswahrscheinlichkeiten und auf dem Test statt der Selektions- und Ablehnungswahrscheinlichkeiten, kommt man zu den konversen Modellen (bzw. Definitionen). Es ist der Verdienst von Petersen & Novick (1976), gezeigt zu haben, daß durch die Umkehr der Betrachtungsweise bei den Definitionen e), f), g) im allgemeinen andere cutoffs notwendig werden und damit diese Modelle intern widersprüchlich sind.

„. . . Specifically, the conditioning process *must* be on the specific value X observed on the person and not on the marginal distribution of X (Thorndike), the conditional distribution of X given y (Cole) or the event $X > x$ (Linn). The three models mentioned above do not use the correct probability." (Petersen & Novick 1976, S. 25)

Allerdings muß zur Verteidigung Thorndike's gesagt werden, daß dieser Vorwurf *nicht* zutrifft, wenn der wichtige Spezialfall $K = 1$ gilt.

5 Bezugsrahmen für eine weitgehend einheitliche Darstellung

Die Definitionen können auf vier Ebenen diskutiert werden: (5.1) Akzeptanz/Erfolgsquoten, (5.2) korrelative Zusammenhänge und Pfadmodelle, (5.3) Regressionsmodelle und (5.4) Selektionscharakteristikkurven (SCK). Dabei gehen wir von der Forderung nach *einem* für alle Gruppen geltenden cutoff auf dem Kriterium $Y = y^*$ und dem Test $X = x^*$ aus. Gruppenspezifische Cutoffs werfen unlösbare Zuordnungsprobleme auf. So sind ja selbst die Gruppen „ausländische Schüler" vs. „einheimische Schüler" wegen unterschiedlicher Integrationsgrade der ausländischen Eltern nicht klar abgrenzbar. Würde man alle Variablen berücksichtigen wollen, deren Ausprägungen kulturelle Benachteiligung bedeuten könnten, müßte eine Unzahl von Gruppen und damit gruppenspezifische Cutoffs bestimmt werden. Darüber hinaus würde die Angabe von Cutoffs für psychologische Typen (z.B. intro-extravertierte Personen) die gleichen Probleme aufwerfen, wie sie von Persönlichkeitstypologien hinlänglich bekannt sind. Gruppenspezifische Cutoffs würden verschiedene Entscheidungen bei gleichen Testleistungen bedeuten und Rassismus und Gruppenegoismen eher fördern als abbauen.

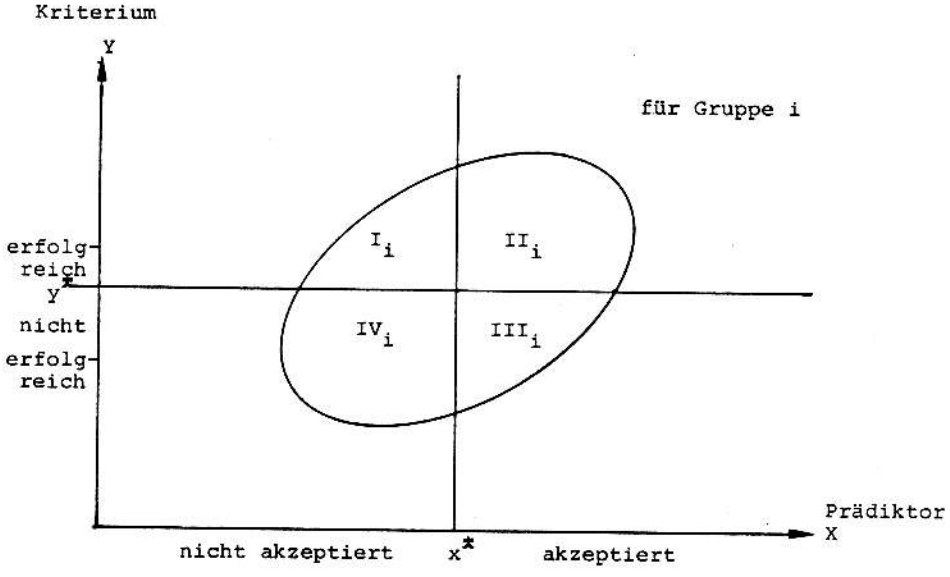
Nun lehnen die meisten Psychologen Cutoffs aus ethischen Gründen ab. Normalerweise bedeutet der Cutoff einen kritischen Wert auf dem Prädiktor X, den eine Person nicht unterschreiten darf, wenn sie für irgendeine besondere Beratung, Behandlung, Schulung akzeptiert werden soll.

Die Forderung nach *einem* für alle Gruppen geltenden Cutoff bedeutet aber nicht notwendigerweise, daß man einen kritischen Wert bestimmt, der über Wohl und Wehe entscheidet. Vielmehr sollte die Forderung ausführlicher formuliert werden als „Forderung nach *einem* für alle Gruppen geltenden *beliebigen* Cutoff". Dieses bedeutet, daß identische Testwerte für alle Gruppen gleiche Konsequenzen (Beratungen, Schulungen, Förderungen etc.) zur Folge haben.

5.1 Akzeptanz-Erfolgsquoten

Eine Möglichkeit, die verschiedenen Fairnesskonzepte einheitlich zu diskutieren, bietet der Kriteriums-Prädikatorraum. Dieser Raum wird durch die manifesten Indikatoren X und Y, wie wir sie im Kapitel 3 kennengelernt haben, aufgespannt. Zu beobachten ist dabei, daß der Kriteriums-Prädiktorraum *nicht* durch die latenten Variablen ξ und η aufgespannt wird, was u.U. sinnvoll wäre.

Betrachten wir die Verhältnisse im kombinierten Kriteriums-Prädiktorraum, können wir durch entsprechende Cutoffs im Prädiktor und Kriterium zwischen Akzeptierten und Nichtakzeptierten bzw. zwischen Erfolgreichen und Nichterfolgreichen trennen. Wir können also den Variablenraum in Quadranten zerlegen. Decken sich die gruppenspezifischen Cutoffs, sind die Testwerte zwischen den Gruppen vergleichbar (vgl. a. Figur 10).



Figur 10: Test-Kriteriumsraum, der durch Cutoffs aufgeteilt wird (für Gruppe i)

Oberstes Ziel sollte es sein, die Validität zu erhöhen. Das geschieht durch die Maximierung der Zahl der richtigen Prognosen und Minimierung der Zahl der Fehlprognosen (einfachstes Nutzenmodell). Zu den Ausweitungen und Fehlergewichtungen in den Expected Utility Modellen sei auf Petersen & Novick (1976) und Petersen (1977) verwiesen.

Es sollte also folgendes erreicht werden:

5.1.1 $\sum_{i=1}^{G} II_i + \sum_{i=1}^{G} IV_i = \max!$ Summe aller sich in den Quadraten II und IV befindlichen Personen.

$\sum_{i=1}^{G} I_i + \sum_{i=1}^{G} III_i = \min!$ Summe aller sich in I und III befindlichen Personen

i = Gruppenindex, G = Anzahl der Gruppen

Nach LINN's Equal Probability Model sollten die Cutoffs so gewählt werden, daß der Prozentsatz der Erfolgreichen unter den Akzeptierten in allen Gruppen gleich ist:

(5.1.2) $II_i/(II_i+III_i) \stackrel{!}{=} II_j/(II_j+III_j)$

wobei: i, j = Gruppenindices sind.

Dagegen bedeutet Fairness nach THORNDIKE, daß

(5.1.3) $(I_i+II_i)/(II_i+III_i)$

für alle Gruppen gleich ist. Im Sonderfall K = 1 sollte

(5.1.4) $(I_i+II_i) \stackrel{!}{=} (II_i+III_i)$ bzw. $I_i = III_i$

für jede Gruppe i gelten: Nach Maximierung der Trefferquote soll das Verhältnis von Leistungsüber- und -unterschätzung in jeder Gruppe gleich sein.

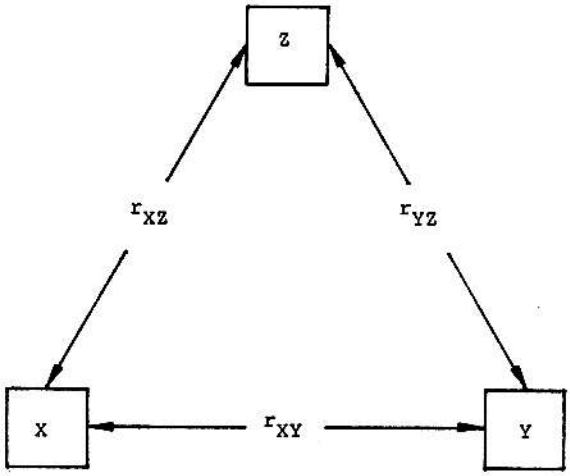
Nach COLE ist ein Test fair, wenn der Prozentsatz der Akzeptierten unter den Erfolgreichen in allen Gruppen gleich ist:

(5.1.5) $II_i/(I_i+II_i) = II_j/(I_j+II_j)$

5.2 Korrelative Beziehungen und Pfadmodelle

Einige Fairnesskonzepte lassen sich im korrelativen Netz zwischen Prädiktor (bzw. Test), Kriterium und ethnischer bzw. soziologischer Drittvariabler (d.h. Gruppenvariabler) verdeutlichen. Unter einigen einschränkenden Annahmen (Gleichheit der Cutoffs und von Prädiktor- und Kriteriumsstreuungen in allen Gruppen; keine differentielle Validität des Tests; bivariate Normalverteilungen in allen Gruppen; Polung der Variablen, so daß alle Korrelationen positiv sind) kann man, wie Darlington (1971) gezeigt hat, eine korrelative Darstellung der verschiedenen Definitionen wählen. Beschränkt man sich auf die Betrachtung der drei Variablen X (= Test), Y (= Leistungsvariable) und Z (= kulturelle oder ethnische Drittvariable, die die mögliche Unfairness von X bedingt), können wir die Korrelationen r_{yx}, r_{xz} und r_{yz} beobachten (vgl. a. Figur 11).

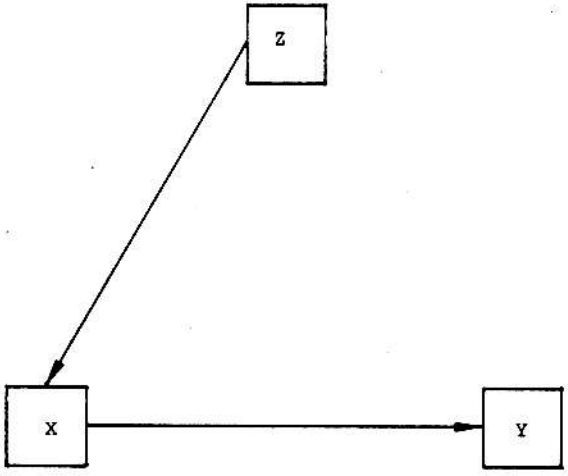
wobei: r_{yx} = Validität des Tests (sollte möglichst groß sein),
r_{xz} = Maß für die kulturelle oder ethnische Abhängigkeit des Tests (*kurzfristig* durch den Testkonstrukteur variierbar),
r_{yz} = Maß für die kulturelle oder ethnische Abhängigkeit des Kriteriums (nur *langfristig* durch Intervention veränderbar).



Figur 11: Korrelatives Netzwerk zwischen Prädiktor X (z.B. Schulleistungstest), Kriterium Y und Drittvariabler Z (z.B. Schicht, Geschlecht, Kind mit Deutsch als Muttersprache / Kind mit anderer Muttersprache)

Unter den schon erwähnten Annahmen können wir die Testfairnesskonzepte in ein korrelationsstatistisches Bezugssystem bringen, in dem r_{yz}, r_{yx} als unabhängiger angesehen werden als r_{xz}. Die Forderung nach hoher Validität sollte natürlich solange wie möglich aufrecht erhalten werden. r_{yz} ist für den Testkonstrukteur kurzfristig veränderbar. Sollte eine Veränderung wünschenswert sein (vgl. a. die Argumentation von Guthke), kann dieses nach allen Erfahrungen in den Sozialwissenschaften nur langfristig erfolgen. Wir haben also:

(5.2.1) $\quad r_{xz} = f(r_{yz}, r_{yx})$



Figur 12: Pfadmodell der Variablen, die dem Cleary'schen Konzept zugrundeliegt

Je nach Definition der Testfairness muß der Test einen bestimmten Grad an kultureller oder ethnischer Abhängigkeit besitzen, um als fair zu gelten: *Cleary's Konzept fordert* das Verschwinden der Korrelation zwischen Kriterium und kultureller Drittvariabler nach Ausschaltung des Einflusses der Testvariablen:

(5.2.2) $r_{xz} = (r_{yz}/r_{yx})$ bzw. $r_{yz \cdot x} = r_{y(z \cdot x)} = 0.0$

Der Forderung (5.2.2) liegt das Pfadmodell von Figur 12 zugrunde.

Die korrelative Brücke zwischen Z und Y kommt *nur* über X zustande:

$$r_{zy} = r_{zx} * r_{xy}$$

Dividiert man beide Seiten der obigen Gleichung durch r_{xy}, erhält man (5.2.2).

Z trägt nicht zu einer verbesserten Schätzung von Y bei, wenn man X kennt oder konstant hält. Es gibt keine Mittelwertsunterschiede pro Testwertgruppe. Cleary's Definition ist nämlich erfüllt, wenn die Regressionen von Y auf X in jeder Gruppe $Z_i (i=1, \ldots, j, \ldots, G)$ gleich sind. Damit sind auch die bedingten Erwartungswerte $E(Y_{z_i}|X) = E(Y_{z_j}|X)$ einander gleich. Aus diesem Grund muß $r_{yz \cdot x} = 0.0$ sein. Daraus ergibt sich:

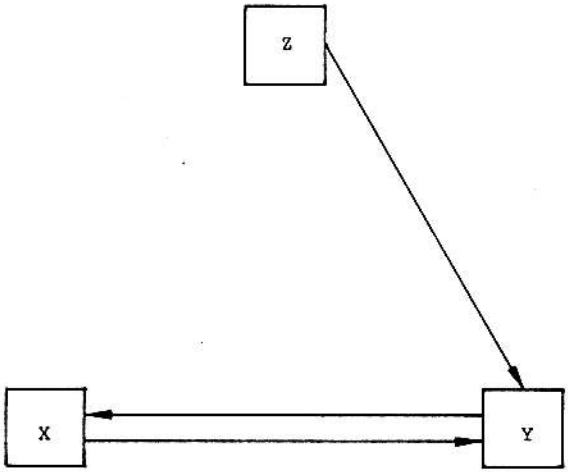
(5.2.3) $$r_{yz \cdot x} = \frac{r_{yz} - r_{yx} r_{xz}}{\text{Nenner}} \stackrel{!}{=} 0.0 \text{ bzw. } r_{xz} = \frac{r_{yz}}{r_{yx}}$$

Thorndike's Konzept fordert (bei K = 1), daß kulturelle Abhängigkeit von Kriterium und Test gleich sein sollten:

(5.2.4) $r_{xz} = r_{yz}$

Thorndike's Definition ist bei Variablen mit gleichen gruppenspezifischen Varianzen nur dann erfüllt, wenn die Mittelwertsdifferenz auf X und Y für alle Gruppenpaare gleich ist. Damit gilt (5.2.4).

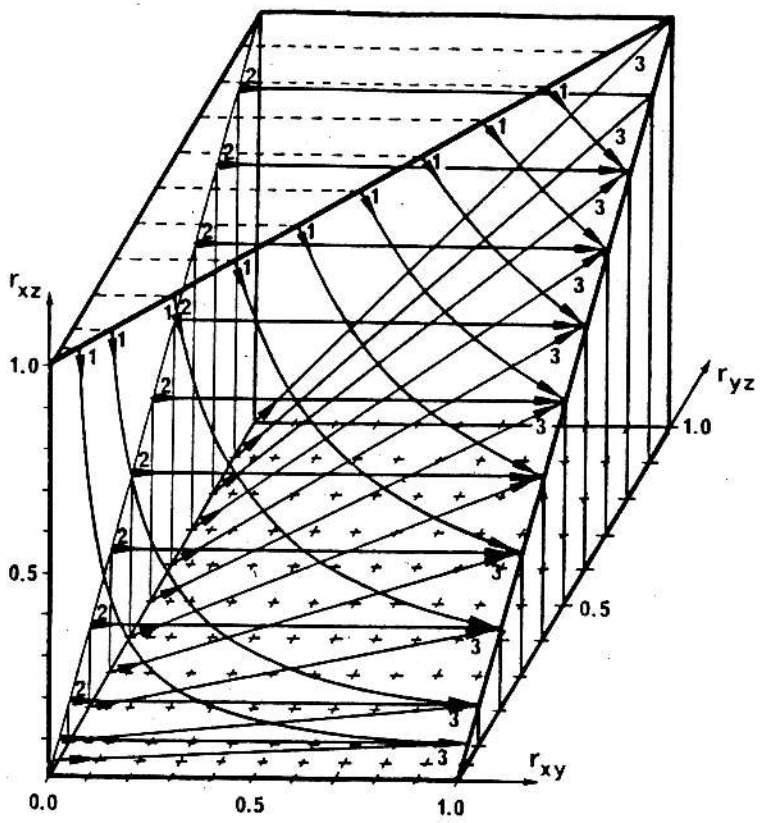
Nach der Definition von *Darlington (Fall III)* sollte die Korrelation zwischen Test und kultureller Drittvariabler verschwinden, wenn man Y kennt oder konstant hält:



Figur 13: Pfadmodell der Variablen, die dem Darlington'schen Konzept (Fall III) zugrundeliegt

(5.2.5) $r_{xz} = r_{yz}r_{yx}$ bzw. $r_{xz \cdot y.} = r_{x(z \cdot y)} = 0.0$

Die korrelative Brücke Z und X kommt *nur* über Y zustande. Der Forderung (5.2.5) liegt das Pfadmodell von Figur 13 zugrunde.



Figur 14: Definitionsspezifische Zusammenhänge zwischen der kulturellen Abhängigkeit des Kriteriums r_{YZ}, der Validität des Tests r_{YX} und der kulturellen Abhängigkeit des Tests r_{XZ}. Die vom Testkonstrukteur nicht kurzfristig veränderbare Korrelation r_{YZ} und die zu maximierende Korrelation r_{YX} spannen die Grundfläche der Figur 14 auf. Je nach Fairnessdefinition muß der Test – unter den Rahmenbedingungen, die ihm durch r_{YZ} und r_{YX} auferlegt sind– eine bestimmte kulturelle Abhängigkeit r_{XZ} aufweisen, um als fair angesehen zu werden. Diesen bestimmten Wert r_{XZ} suchen wir, indem wir vom Punkt (r_{XY}, r_{YZ}) in der Ebene senkrecht und parallel zur Achse r_{XZ} emporwandern bis wir auf eine Linie 3 (Darlington's Fall III), eine Linie 2 (Thorndikes's Definition) oder eine Linie 1 (Cleary's Definition) stoßen. Die Höhen der Linien 1, 2, 3 entsprechen den Funktionswerten der Fairnessdefinitionen (5.2.2), (5.2.4) und (5.2.6). Das Identitätskonzept fordert $r_{XZ} = 0.0$ für alle r_{YX} und r_{YZ}, so daß wir die Fußebene von Figur 14 nicht verlassen müssen. Die Definitionen stimmen bei perfekter Validität $(r_{xy} = 1.0)$ oder kultureller Unabhängigkeit des Kriteriums überein.

Ist nämlich die Regression von X auf Y für jede Gruppe Z_i gleich, sind damit auch die bedingten Erwartungswerte $E(X_{z_i}|Y) = E(X_{z_j}|Y)$ einander gleich. Für diesen Fall ist dann:

$$(5.2.6) \qquad r_{xz \cdot y} = \frac{r_{xz} - r_{yx}r_{yz}}{\text{Nenner}} \stackrel{!}{=} 0.0 \text{ bzw. } r_{xz} = r_{yx}r_{yz}$$

Die kulturelle Abhängigkeit eines fairen Tests r_{xz} dürfe nur über den Umweg über das Kriterium zustande kommen. Der Test X unterliege keinem direkten Einfluß von Z (d.h. der Test stellt keinen Indikator für Z dar).

Natürlich lassen sich diese Kausalannahmen nicht mit den Mitteln der Korrelationsrechnung überprüfen, jedoch bilden sie die Basis für die zu fordernden Korrelationsmuster.

Auch das altbekannte *Identitätskonzept* läßt sich korrelativ darstellen:

$$(5.2.7) \qquad r_{xz} = 0.0 \text{ bzw. } r_{y(x \cdot z)} = r_{yx} = r_{yx \cdot z}\sqrt{1 - r_{yz}^2}$$

Die Ausschaltung der Drittvariablen darf keinen Einfluß auf die Validität des Tests haben bzw. die Kenntnis der Drittvariablen hilft nicht bei der Vorhersage der Testergebnisse X. Vertritt man ernsthaft das Identitätskonzept, dürften damit die meisten gesellschaftlich relevanten Kriterien- und Prädiktorbereiche einer testpsychologischen Untersuchung und Prognose entzogen sein.

Die Zusammenhänge zwischen den Definitionen sind in Figur 14 dargestellt. Nur in den beiden unrealistischen Fällen perfekter Validität ($r_{yx}=1$) und der kulturellen und ethnischen Unabhängigkeit des Kriteriums ($r_{yz}=0.0$) stimmen die Definitionen überein. Für alle anderen Validitätsbereiche $0 < r_{yx} < 1$ mit der kulturellen Abhängigkeit des Kriteriums $r_{yz} > 0$ widersprechen sich die Konzepte. Die geforderte kulturelle Abhängigkeit des Tests ist bei Cleary am größten und beim Identitätskonzept ($r_{xz}=0.0$) am niedrigsten (s.a. Figur 14).

Auch hier zeigt sich die Mittelstellung der Thorndikeschen Formulierung. Bildet man nämlich das geometrische Mittel $\bar{r}$ aus den Korrelationen in den Definitionen bzw. Forderungen von Cleary und Darlington (Fall III), erhält man die Konzeption von Thorndike in korrelativer Darstellung.

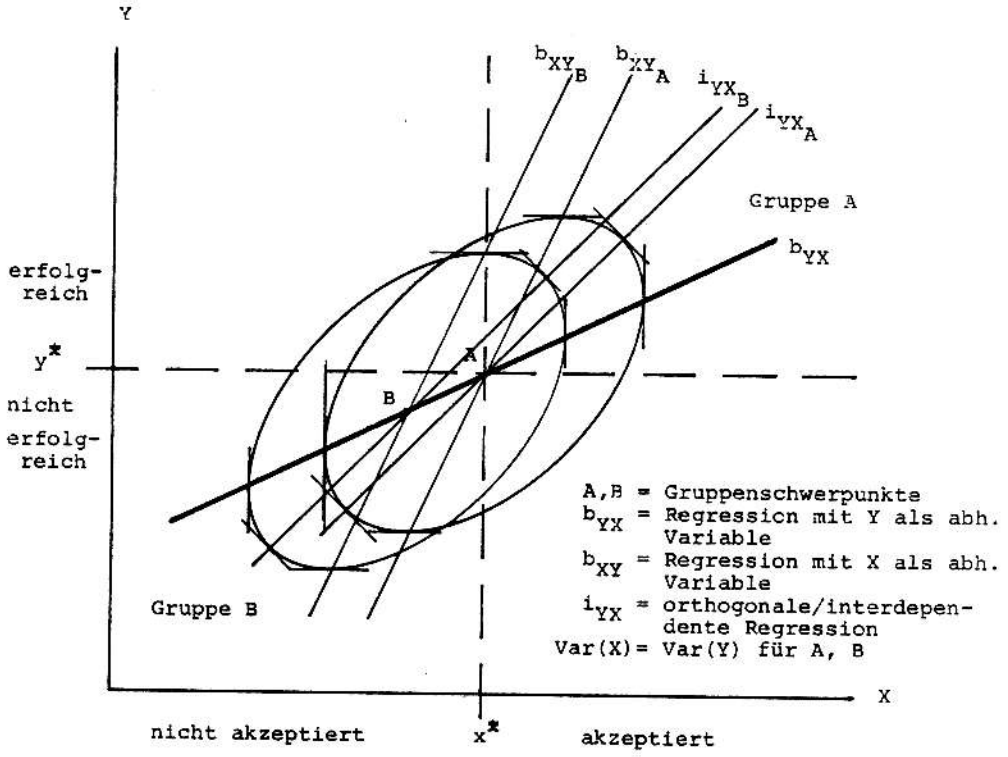
Setzt sich die Fähigkeit, in einem Kriterium erfolgreich zu sein, aus mehreren Teilfähigkeiten zusammen, kann es u.U. zu einem *unvollständigen Kriteriumssampling* kommen: Der Test mißt das Kriterium nur unvollständig. Sich daraus ergebende Konsequenzen sind bei Möbus (1978, S. 208f.) dargestellt.

5.3 Die Beziehungen zwischen Fairnesskonzepten und Kriterium-Test-Regressionen

Wir betrachten dabei nur die Fälle, in denen auf Kriterien und Prädiktor nur jeweils gleiche gruppenspezifische cutoffs $Y = y^*$ bzw. $X = x^*$ gelten; d.h. Entscheidungen über Personen gelten für einen bestimmten Testwert für alle Personen ohne Ansehen der Gruppenzugehörigkeit.

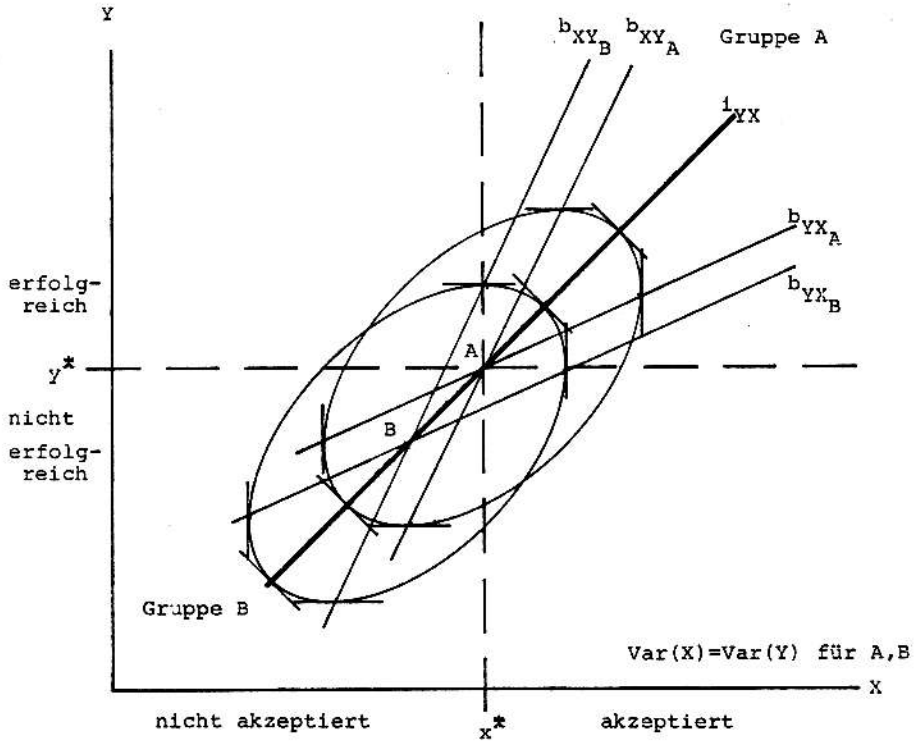
Die Definition von Cleary ist in jedem Fall erfüllt, wenn die Regressionskoeffizienten der Regression von Y auf X (= Test) für alle Gruppen nicht signifikant voneinander abweichen. Darüber hinaus ist sie aber auch erfüllt, wenn sich zwar die gruppenspezifischen Regressionslinien untereinander unterscheiden aber dennoch nicht signifikant von der *gemeinsamen* Regression abweichen.

Zu den Regressionskoeffizienten rechnen wir auch die Konstante, die als Gewicht einer Dummyvariablen mit dem konstanten Wert 1 angesehen werden kann. Die Annahme einer gemeinsamen Regression und eines Cutoffs x* ist nicht mehr erfüllt, wenn es zu einem „Strukturbruch" (s.a. Schneeweiss 1971) gekommen ist. Eine graphische Darstellung dieser Definition für den Sonderfall eines einzigen Cutoffs findet sich in Figur 15.



Figur 15: Situation, in der nach Cleary derselbe Cutoff im Test für beide Gruppen A und B fair ist: Die Regressionen von Y auf X (dicke Linie b_{YX}) sind für beide Gruppen gleich.

Auch Thorndike's Definition läßt eine regressionsorientierte Deutung zu. Die Definition ist für den wichtigen Spezialfall K = 1 erfüllt, wenn die interdependente bzw. orthogonale Regression zwischen Y und X oder X und Y für alle Gruppen bis auf Zufallsfehler gleich ist (bei $s_x = s_y$). Bei der orthogonalen Regression werden die Fehler nicht parallel zur Y-Achse, sondern orthogonal zur Regressionslinie oder Regressionshyperfläche gemessen. Dabei ist die orthogonale Regressionslinie der gemeinsame Faktor (im Sinne der Faktorenanalyse) von X und Y (vgl. a. Möbus 1978, Anhang A). Dieses kann man auch Figur 16 entnehmen.

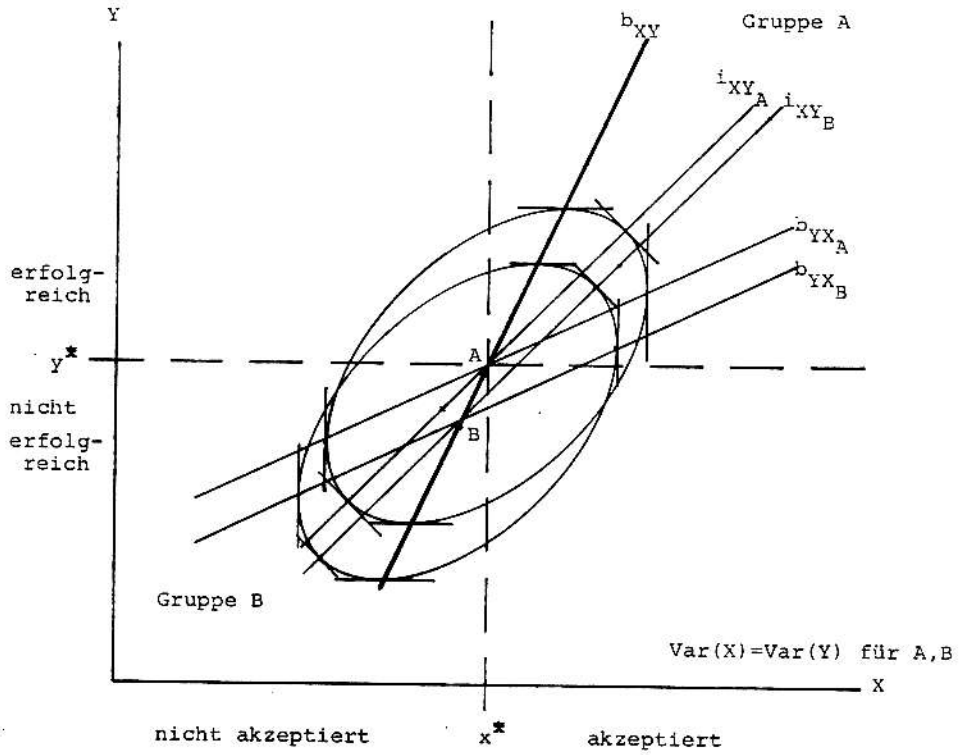


Figur 16: Situation, in der nach Thorndicke derselbe Cutoff im Test für beide Gruppen A und B fair ist: Die Regressionen i_{XY} (orthogonale/interdependente) sind für beide Gruppen gleiche (dicke Linie i_{YX})

Die Definition von Darlington's Fall III (Spezialfall von Coles's Definition ist – invers zu Cleary – dann erfüllt, wenn die Regressionskoeffizienten der Regression von X auf Y (Y = Kriterium, das jetzt die Rolle des Regressors spielt) nicht signifikant in den Gruppen voneinander abweichen (vgl. a. Figur 17).

Auch hier zeigt sich wieder Thorndike's Mittelstellung zwischen den relativ extremen Positionen von Cleary (bzw. Linn) auf der einen und Darlington III bzw. Cole) auf der anderen Seite.

Bevor wir diese deskriptiven Regressionsansätze in 5.5 diskutieren und bewerten, wollen wir auf der vierten Betrachtungsebene die Selektionscharakteristikkurven (SCK) der verschiedenen Konzepte betrachten.

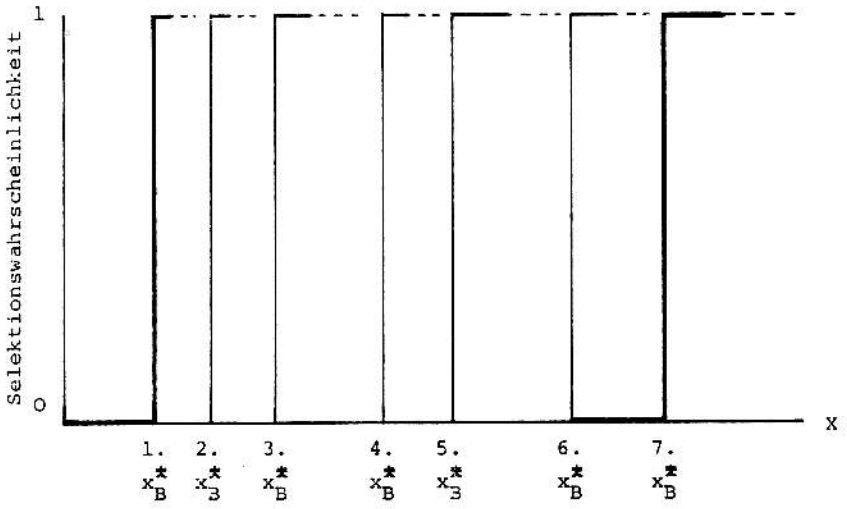


Figur 17: Situation, in der nach Darlington (Fall III) derselbe Cutoff im Test für beide Gruppen A und B fair ist: Die Regressionen von X auf Y sind für beide Gruppen gleich (dicke Linie b_{XY})

5.4 Selektionscharakteristikkurven und Cutoffs

5.4.1 Selektionscharakteristikkurven

Wir wählten den Begriff SCK in Anlehnung an den aus der Testtheorie bekannten Begriff der Itemcharakteristikkurve (ICK). Während mit der ICK die Abhängigkeit der Wahrscheinlichkeit einer richtigen Antwort vom Fähigkeitsniveau eines Probanden beschrieben wird, soll mit der SCK die Abhängigkeit der Selektionswahrscheinlichkeit eines Probanden von seinem Testwert oder Fähigkeitsniveau beschrieben werden. Dabei zeigt sich, daß die SCKs aller bisher behandelten Ansätze eine Form aufweisen, wie man sie von den Guttman-Items (s.a. Fischer 1974) her kennt. Am Cutoff $X = x^*$ springt die Selektionswahrscheinlichkeit von 0 auf 1. Nimmt man zwei Gruppen an (eine priviligierte und eine deprivierte) und lassen wir gruppenspezifische Cutoffs x_A^*, x_B^* zu, können die Cutoffs der deprivierten Gruppe je nach Fairnesskonzept in eine Rangreihe gebracht werden (s.a. Figur 18). Der Cutoff x_B^* für die deprivierte Gruppe liegt bei solch ‚konservativen' Definitionen, wie sie Einhorn & Bass sowie Cleary vorschlugen, am höchsten und bei solch ‚risikofreudigen', wie sie Darlington's Fall III und Cole's Ansatz darstellen, am niedrigsten.



Die Figur stellt eine mögliche Rangreihe von Cutoffs für eine unterpriviligierte Gruppe B in Abhängigkeit von der jeweiligen Fairnessdefinition dar. Je nach besonderen Gegebenheiten des Datensatzes und subjektiver Kriterien (z.B. bei Einhorn & Bass bzw. bei Darlington's culture modified criterion) können allerdings Rangplatzvertauschungen und Verschiebungen vorkommen.

1. Cutoff nach Darlington's culture modified criterion
2. Cutoff nach Darlington's Fall III
3. Cutoff nach Cole
4. Cutoff nach Thorndike
5. Cutoff nach Linn
6. Cutoff nach Cleary
7. Cutoff nach Einhorn & Bass

Figur 18: Rangreihe der Selektionscharakteristikkurven für die verschiedenen Fairnessdefinitionen

5.4.2 Kritik an der Verwendung kritischer Grenzwerte (Cutoffs)

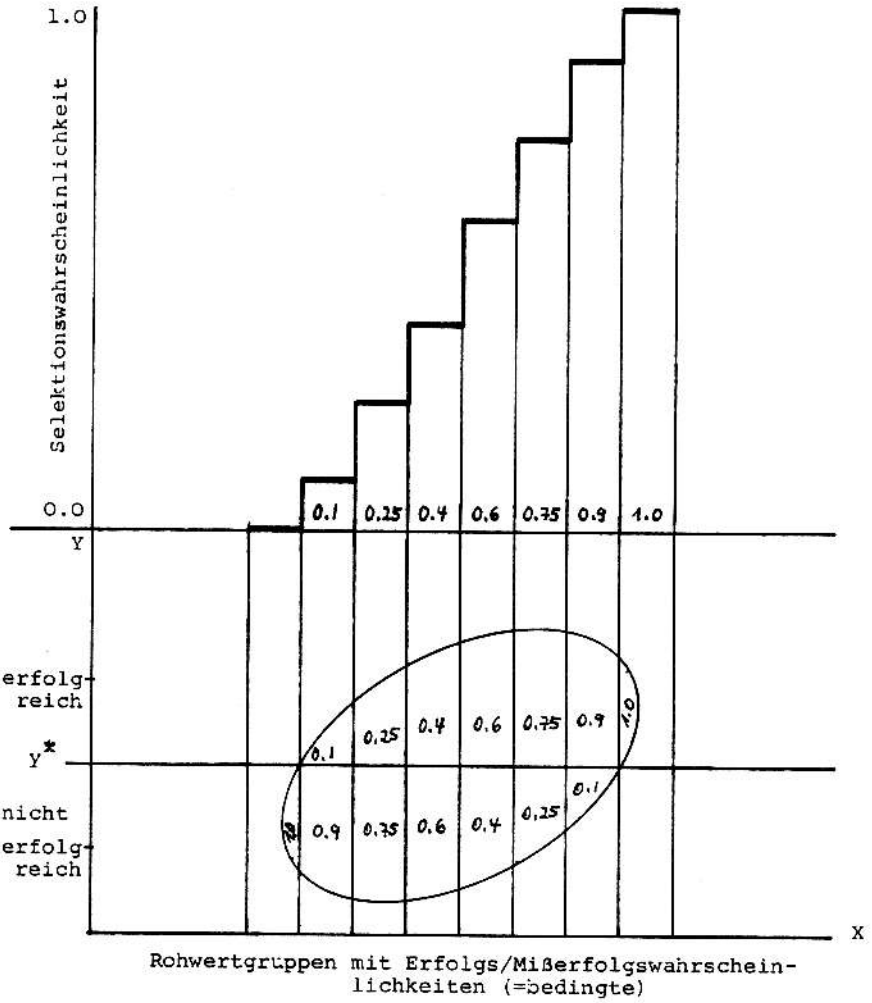
Die gerechte Bestimmung und Anwendung kritischer Grenzwerte ist bislang ein Hauptziel *aller* Fairnesstheoretiker gewesen. Man kann aus verschiedenen Gründen Zweifel an der Richtigkeit dieser Zielsetzung anmelden.

a) Die Aufteilung des Kriteriums in zwei Gruppen ($y<y^*$ und $y\geqslant y^*$) ist nur gerechtfertigt, wenn dichotome Entscheidungen gefällt werden (bestanden/nicht bestanden). Im allgemeinen steigt aber der Nutzen eines Kriteriumswertes mit seiner Ausprägung kontinuierlich an. Ein ähnliches Argument gilt auch für den Testwert X.
b) Die Cutoff-Methode ist nur fair gegenüber besseren Probanden, wenn X und Y perfekt reliabel sind. Je unreliabler die Variablen sind, desto häufiger fällt eine Entscheidung aufgrund fehlerhafter Meßwerte, die bei Kenntnis der latenten Variablen nicht gefallen wäre: d.h. Cutoffs sollten eher bei den latenten Variablen gesetzt werden.
c) Die Cutoff-Methode ist unfair gegenüber schlechteren Probanden, weil sie im Sinne einer Statusdiagnostik von einem trait-Konzept und nicht im Sinne einer Pro-

zeßdiagnostik von der Wandelbarkeit einer Person ausgeht. Wie Simons und Möbus (1977) zeigten, genügt eine Testwiederholung mit einem Paralleltest oder ein kurzes Training (3 Stunden), um die Leistung so zu verbessern, daß 60% der mit Hilfe von Cutoffs getroffenen Entscheidungen revidiert werden müssen.

Personen, die trotz erheblicher Verbesserungen im Test unter dem Cutoff bleiben, werden trotz gleichzeitiger Erhöhung ihrer Erfolgsmöglichkeiten nicht mit einer erhöhten Selektionswahrscheinlichkeit belohnt: Die Cutoff-Methode ist relativ insensitiv gegenüber Veränderungen bzw. Verbesserungen im kognitiven Bereich. Dabei nimmt die Wahrscheinlichkeit einer Entscheidungsrevision mit der Distanz der Personenwerte vom Cutoff ab.

Eine Möglichkeit, kognitive Verbesserungen mit erhöhten Selektionschancen (allerdings auf Kosten der Gesamttrefferquote) zu belohnen, bieten andere Formen der



Figur 19: Selektionswahrscheinlichkeiten in Abhängigkeit von der bedingten Erfolgswahrscheinlichkeit

Selektionscharakteristikkurven (SCK), z.B. in der Art von Ogiven oder der logistischen Funktion. Ein möglicher Ansatz wäre die Gleichsetzung von folgenden Wahrscheinlichkeiten (s.a. Figur 19).

$$P(Y>y^*|X=x) \stackrel{!}{=} P(\text{Selektion}|X=x)$$

5.5 Bewertung und Diskussion der verschiedenen Regressionskonzepte

Wir glauben, daß eine Diskussion der den Definitionen unterliegenden Regressionskonzepte von allen Betrachtungsweisen die aufschlußreichste ist. Dabei fassen wir Cleary, Einhorn & Bass sowie Darlington (culture modified criterion) zur Gruppe der Definitionen zusammen, die der OLS-Regression Y auf X verpflichtet sind, während Darlington III und das Modell gleichen Risikos (h) zur Gruppe der OLS-Regressionskonzepte X auf Y gezählt werden. Thorndike (für den Spezialfall: $K = 1$, $s_x = s_Y$) ist der orthogonalen Regression zuzuordnen.

Die faire Anwendung eines Tests setzt bei der Forderung nach *einem* Cutoff $X = x^*$ u.U. die Gleichheit einer von der jeweiligen Definition abhängigen Regression in allen relevanten Populationen voraus. Die Regressionsparameter müssen populationsunabhängig sein, wenn man gleichen Meßwerten gleiche Entscheidungen zuordnen will (gleicher Cutoff x^* für alle Gruppen).
Zur Bewertung und Auswahl einer bestimmten Regression und damit eines Fairnesskonzeptes müssen die jeweiligen Modellannahmen und Fehlertheorien untersucht werden.

Dazu wenden wir uns den Beispielen in Figur 20 zu.
Gehen wir davon aus, daß die „wahre" Regressionsbeziehung zwischen 2 latenten Variablen ξ, η lautet:

$$E(\eta|\xi) = \lambda\xi + \alpha$$

Wobei ξ, η kongenerisch mit unterschiedlichen Mittelwerten und α, λ (= Regressionsparametern) sind.

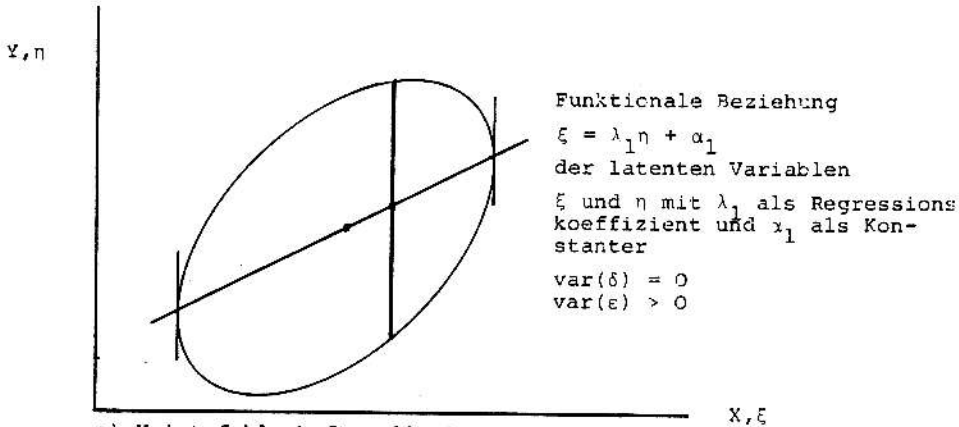
Alle Personen liegen mit ihren latenten Werten auf einer geraden Linie.

Nun treffen wir drei verschiedene Annahmen über die auftretenden Fehler im Regressionsmodell

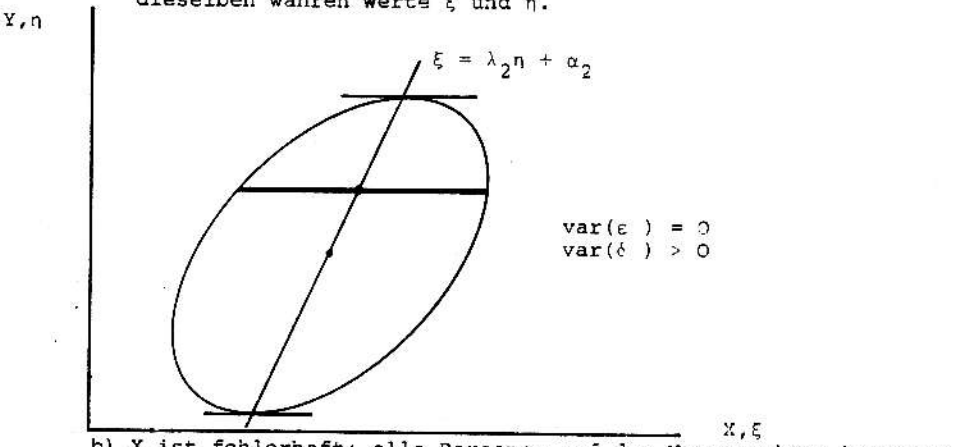
a) Y läßt sich nicht exakt durch ξ vorhersagen: es schleicht sich ein „Gleichungsfehler" ζ ein.
   η läßt sich nicht exakt beobachten: es schleicht sich ein Meßfehler ϵ' ein.
   ξ läßt sich exakt durch X beobachten.
   Die Regression ist

$$Y = \lambda\xi + \alpha + \epsilon \quad (\epsilon=\epsilon'+\zeta)$$

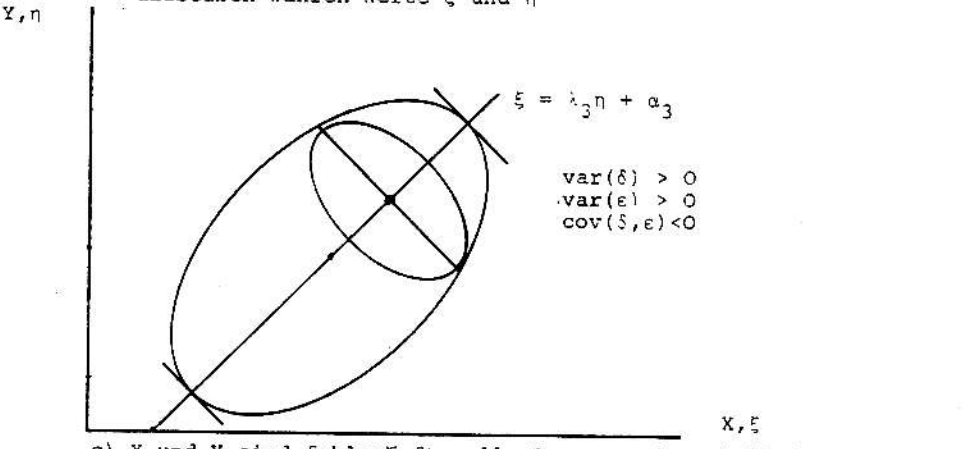
   Die Situation ist in Figur 20a dargestellt. Will man λ und α schätzen, kann man die Parameter konsistent *nur* mit der OLS-Regression von Y auf X schätzen.
b) Ist dagegen X nur fehlerhaft meß- und vorhersagbar und Y exakt beobachtbar (Figur 20b), lassen sich die Regressionsparameter konsistent *nur* mit der OLS-Regression von X auf Y schätzen.
c) Sind dagegen X *und* Y fehlerhaft meß- und vorhersagbar, trifft die Situation in Figur 20c zu. In einem besonderen Spezialfall (gleiche Fehlervarianzen auf beiden



a) Y ist fehlerhaft: alle Personen auf der Senkrechten besitzen dieselben wahren Werte ξ und η.



b) X ist fehlerhaft: alle Personen auf der Waagerechten besitzen dieselben wahren Werte ξ und η



c) X und Y sind fehlerhaft: alle Personen innerhalb der Ellipse können dieselben wahren Werte ξ und η besitzen

Figur 20: *Verschiedene* „wahre" Beziehungen zwischen den latenten Variablen ξ und η können je nach Fehlerannahmen die *gleiche* bivariate Häufigkeitsverteilung der Indikatoren X und Y erzeugen

Variablen) lassen sich die Regressionsparameter mit der orthogonalen Regression schätzen (s. Möbus 1978, Anhang A). In allen anderen Fällen werden die Regressionsparameter nach dem Strukturgleichungsansatz in Kapitel 3 geschätzt. Dazu benötigt man aber für X und Y jeweils mindestens 2 Indikatoren X_1, X_2, Y_1, Y_2.

Ignoriert man die Meßfehler des Prädiktors, unterschätzt man den Zusammenhang der latenten Variablen, benachteiligt Personen mit besseren Prädiktorwerten und bevorzugt Personen mit schlechteren Prädiktorwerten in der Population. Fatal ist der Einbezug der ϵ in die Prognose, wie man noch in einer anderen Argumentationskette zeigen kann (s.a. Hunter & Schmidt 1976): Ist der perfekt reliable Test nach Cleary bei Beibehaltung einer für alle Gruppen gültigen Regression fair und haben zwei Personen A und B die Testwerte $x_A = 115$ und $x_B = 110$ (= wahre Werte), ist die Bevorzugungswahrscheinlichkeit $P[A>B] = 1.00$. Ist dagegen die Reliabilität nur .50, dann variieren die beobachteten Testwerte um ihre wahren Werte. Liegt die Standardabweichung der beobachteten Werte bei 15, haben die Differenzen einen Mittelwert 5 und eine Standardabweichung von 21. Die Wahrscheinlichkeit einer negativen Differenz steigt von 0.00 auf $P[x_A<x_B] = .41$ und die Selektionswahrscheinlichkeit für Vp A fällt von 1.00 auf 0.59. Der unreliable Test benachteiligt also stark die besseren Probanden.

Cleary's Regressionskonzept beruht auf der Vorstellung eines perfekt reliablen Prädiktors, Darlington's Fall III dagegen auf der Fiktion eines meßfehlerfreien Kriteriums bei perfekter Vorhersagbarkeit bei Kenntnis von ξ.

Wir sind der Ansicht, daß mit Meßfehlern behafteten Variablen nur dann prognostiziert werden darf, wenn die Prognoseverfahren eine entsprechende Fehlertheorie besitzen. Cleary's und Darlington's (Fall III) Fairnessdefinitionen gehen aber von unrealistischen Annahmen aus. Zumindest sind diese Annahmen in nichtexperimentellen Untersuchungen zweifelhaft. Dahingegen scheint Thorndike's Konzept einige interessante Denkanstöße zu geben: 1) Prädiktor und Kriterium sollen gleichberechtigte Indikatoren einer latenten Dimension sein (faktorielle Validität). 2) Die Trennung zwischen endogenen und exogenen Variablen ist weitgehend aufgehoben, zumal Prädiktor und Kriterium auch von der kulturellen Drittvariablen abhängig sein können. Die manifesten Indikatoren der latenten Dimension unterscheiden sich jetzt hauptsächlich nach dem Grad ihrer kurzfristigen Manipulierbarkeit bei der Testkonstruktion. Während die Kriteriumsitems fest ausgewählt sind (z.B. „die Anforderungen des Berufs oder des Studiums"), können oder müssen die Testitems solange ausgetauscht werden bis sie die gleiche latente Dimension messen. 3) In diesem Sinne konvergieren Reliabilität und Validität des Tests: Alle Items (Kriteriums- wie auch Prädiktoritems) messen die gleiche latente Dimension. Ein Test hätte demnach nicht mehr *viele* Validitäten und *eine* Reliablilität sondern nur *eine* Validität und *eine* Reliabilität.

Greifen wir die Gedanken des Abschnitts 3 auf, müssen die Indikatoren X_1, X_2, Y_1, Y_2 *streng parallel* sein mit gleichen gruppenspezifischen Parametern (aber unterschiedlichen gruppenspezifischen Mittelwerten):

$var_i(\epsilon_1) = var_i(\epsilon_2) = var_i(\delta_1) = var_i(\delta_2)$ $(i=1,\ldots,G)$ so daß X_1, X_2, faire Prädiktoren für Y sind.

Gruppenspezifische Mittelwertsunterschiede (z.B. $X_{1i}-X_{1j}$) zwischen Gruppen i und j bleiben auf allen Indikatoren gleich. Eventuell sind noch Meßfehlerkovarianzen zugelassen, wenn die X und Y – Indikatoren zu unterschiedlichen Zeitpunkten

erhoben wurden. Wir wollen jetzt eine Fairnessdefinition geben, die die o.a. Gedanken zusammenfaßt (Möbus 1978, S. 223).

„Ein Prognosesystem mit dem Test als Prädiktor ist dann fair, wenn die im Sinne von Neyman & Scott strukturellen Parameter dieses Systems (= Strukturkoeffizienten der Prognosegleichung) für alle relevanten ethnischen und sozialen Gruppen gleich sind und daher die interindividuelle Variation in der Kriteriumsprognose nur auf die Variation der incidentellen Personenparameter zurückzuführen ist." (Möbus & Simons 1975, S. 16)

Diese Forderung ist intuitiv einleuchtend: Der Test soll sich gegenüber den Gruppen neutral verhalten (= Gleichheit der strukturellen Parameter der verschiedenen Gruppen). Prognoseunterschiede sollen nur durch individuelle Merkmale (= incidentelle Parameter) erklärbar sein. Das schließt natürlich nicht aus, Fehlentscheidungen zu gewichten und Cutoffs entsprechend der statistischen Entscheidungstheorie zu bestimmen, wie dieses von Gross & Su (1976), Petersen & Novick (1976) Sawyer, Cole & Cole (1976) und Petersen (1977) vorgeschlagen wurde. Allerdings sind diese Ansätze nach unserer Überzeugung nicht zufriedenstellend, was auf ihre Fixierung an der Cutoff-Bestimmung und ihren mangelnden Konstruktbezug zurückzuführen ist.

6 Folgerungen für die Praxis

Will man psychologische Entscheidungen oder Beratungen möglichst fair gegenüber Gruppen, Individuen und Institutionen mit Hilfe psychologischer Prädiktoren (u.a. Tests) durchführen, sollte man folgende Punkte berücksichtigen:

1. Inhaltliche Festlegung:

1.1 Was ist Kriterium Y (relevantes Verhalten), das durch den Psychologen vorhergesagt wird? Auch bei psychologischen Beratungen wird stillschweigend eine Prognose des Kriteriums Y durchgeführt!

1.2 Ist das Kriterium Y inhaltlich festgelegt, wird dazu ein Prädiktor X inhaltlich bestimmt. X und Y sollten Indikatoren einer gemeinsamen latenten Variablen ξ (z.B. ξ = Schulleistung) sein (X als Stellvertreter für Y).

1.3 Welchen Gruppen gegenüber sollen die psychologischen Entscheidungen fair sein?

2. Methodische Festlegung

2.1 Bin ich in der Lage zu X und Y jeweils ca. 30 *Einzelindikatoren* (z.B. Testaufgaben) selbst zu formulieren, kann die Testkonstruktion, die Validierung, die Beseitigung von Item- und Testbias und evtl. spätere Prognosen elegant mit dem probabilistischen Testmodell von Rasch durchgeführt werden. Eine ausführliche Beschreibung des Vorgehens findet sich bei Möbus, (1978, Kap. 5.2, 5.3). Dazu ist aber ein erheblicher Arbeitsaufwand, der Zugang zu entsprechenden Computerprogrammen voraussetzt, notwendig. Allerdings ist in letzter Zeit eine „Papier-und-Bleistift-Methode" von Wright & Stone (1979) bekannt geworden, so daß der Zugang zum Computer nicht mehr unerläßlich ist.

2.2 Trifft 2.1 nicht zu, d.h. liegen nicht so viele Einzelindikatoren vor, wählen wir den Weg über die klassische Testtheorie:

2.2.1 Für das Kriterium Y wählen wir mindestens 2 Indikatoren Y_1 und Y_2 (z.B. Ergebnisse von 2 Schulleistungstests oder Lehrerurteile von 2 Lehrern zum Zeitpunkt t)

2.2.2 Für den Prädiktor X wählen wir ebenfalls zwei Indikatoren X_1 und X_2 (bei Zeitpunkt t−1: echte Prognose; bei Zeitpunkt t: unechte Prognose)

2.2.3 Mit dem Computerprogramm Lisrel (Jöreskog & Sörbom 1981) testen wir eine Reihe von Annahmen, die immer einschränkender werden. Je einschränkender die haltbare Annahme ist, desto leichter läßt sich die gewünschte Fairness erzielen:

2.2.3.1 Sind X_1, X_2, X_3, X_4 kongenerisch und die Parameter gleich in allen Gruppen?

2.2.3.2 Sind X_1, X_2, X_3, X_4 τ-äquivalent d.h. $\lambda_1 = \lambda_2 = \lambda_3 = \lambda_4 = 1.0$ und die Parameter gleich in allen Gruppen?

2.2.3.2 Sind zusätzlich die Fehlervarianzen $\text{var}(\delta_1)$, $\text{var}(\delta_2)$, $\text{var}(\epsilon_1)$, $\text{var}(\epsilon_2)$ gleich, besitzen alle Indikatoren gleiche Reliabilität und sind streng parallel. Dieses ist das anzustrebende Ideal (s.a. Kap. 3.3). Die Tests sind fair im Sinne von Thorndike. Mittelwertsunterschiede zwischen den Gruppen auf X und Y sind gleich. Wenn sich Cutoffs nicht vermeiden lassen, sollten sie nach der Formel (4.5) oder (5.1.3) gesetzt werden.

2.2.3.3 Trifft 2.2.3.2 nicht zu, entscheidet das Verhältnis der Fehlervarianzen, welche Fairnessdefinition anzuwenden ist.

2.2.3.4 Trifft 2.2.3.1 nicht zu, sollte von einer Verwendung des Tests X_1, X_2 für die Beratung, Prognose Entscheidung des Psychologen hinsichtlich des Kriteriums Y abgesehen werden.

7 Zusammenfassung

Ziel der vorliegenden Arbeit war es, zu zeigen, wie sich die Reliabilität, Validität und Fairness als Gütemerkmale ergänzen. Dabei ergab sich, daß das vorherzusagende Verhalten und das der Prognose zugrunde liegende Verhalten von *einem* einzigen Konstrukt abhängen soll. Erreichen läßt sich das am einfachsten, wenn das der Prognose dienende Verhalten durch Kriteriumssampling bestimmt wird.

Fair im Sinne unserer Definition ist eine psychologische Entscheidung dann, wenn Tests sich neutral gegenüber Gruppen (= Gleichheit der strukturellen Parameter) und differenzierend gegenüber Individuen (= Ungleichheit der inzidentellen Parameter) verhalten.

Fairness läßt sich über die *elaborierte* Methode des Rasch-Modells (Möbus 1978, S. 223f.) oder über die „Quick & Dirty" Methode der klassischen Testtheorie (strikt parallele Indikatoren X_1, X_2, Y_1, Y_2; Gleichheit gruppenspezifischer Varianzen $\text{var}(\xi)$, $\text{var}(\epsilon)$, $\text{var}(\delta)$) und Bestimmung des Cutoffs (wenn sie unvermeidlich sind) nach Thorndike's Methode erreichen.

Anastasi, A.: Culture-fair Testing, Educational Horizons. 1964, 43, 26–30.
Anastasi, A.: Psychological Testing. New York: Macmillan 1968[3].
Atkinson, J.W.: Motivational Determininants of Intellectual Performance and Cumulative Achievement. In: Atkinson & Raynor (eds): Achievement and Performance. Washington, D.C.: Winston 1974, 389–410.
Bloom, B.S.: Human Characteristics and School Learning. New York: Mc Graw Hill 1976.
Breland, H.M. & Ironson, G.H.: DeFunis reconsidered: A Comparative Analysis of Alternative Admission Strategies. Journal of Educational Measurement 1976, 13, 89–99.
Carroll, J.B.: Ein Modell des schulischen Lernens. In: Edelstein & Hopf (Hrsg.): Bedingungen des Bildungsprozesses. Stuttgart: Klett 1973, 234–250.
Cleary, T.A.: Test Bias: Prediction of Grades of Negro and White Students in Integrated Colleges. Journal of Educational Measurement 1968, 5, 115ff.
Cleary, T.A., Humphreys, L.G., Kendrick, S.A. & Wesman, A.: Educational Uses of Tests with Disadvantaged Students. American Psychologist 1975, 30, 15–41.
Cole, N.S.: Bias in Selection. Journal of Educational Measurement 1973, 10, 237–255.
Cronbach, L.J.: Equity in Selection – Where Psychometrics and Political Philosophy Meet. Journal of Educational Measurement 1976, 13, 31–42.
Cronbach, L.J. & Snow, R.E.: Aptitude and Instructional Methods: A Handbook for Research on Interactions. New York: Irvington Press 1977.
Darlington, R.B.: Another Look at „Culture Fairness". Journal of Educational Measurement 1971, 8, 71–82.
Darlington, R.B.: A Defense of „RATIONAL" Personnel Selection and two New Methods. Journal of Educational Measurement 1976, 13, 43–52.
Dunkin, M.J. & Biddle, B.J.: The Study of Teaching. New York: Holt, Rinehart & Winston 1974.
Einhorn, H.J. & Bass, A.R.: Methodological Considerations Relevant to Discrimination in Employment Testing. Psychological Bulletin 1971, 75, 261–269.
Fischer, G.: Einführung in die Theorie psychologischer Tests. Bern: Huber 1974.
Fischer, G.: Psychologische Testtheorie. Bern: Huber 1968.
Flammer, A.: Wechselwirkungen zwischen Schülermerkmalen und Unterrichtsmethoden – eine zeronnene Hoffnung? In: Mandl, H. & Krapp, A. (Hrsg.): Schuleingangsdiagnose. Göttingen: Hogrefe 1978, 113–120.
Flammer, A.: Individuelle Unterschiede im Lernen. Weinheim: Beltz 1975.
Flanders, N.A.: Analyzing Teaching Behavior. Reading, Mass.: Addison-Wesley 1970.
Gagne, R.M.: The Acquisition of Knowledge. Psychological Review 1962, 69, 355–365. Deutsch in: Hofer, M. & Weinert, F.E. (Hrsg.): Reader zum Funkkolleg Pädagogische Psychologie 2, Frankfurt: Fischer 1973, 106–123.
Gagne, R.M.: Learning Hierarchies. Educational Psychologist 1968, 6, 1–9.
Gagne, R.M.: Die Bedingungen menschlichen Lernens. Hannover: Schroedel 1977[3].
Getzels, J.W. & Jackson, W.P.: Merkmale der Lehrerpersönlichkeit. In: Ingenkamp, K. (Hrsg.): Handbuch der Unterrichtsforschung, Bd. II, Weinheim: Beltz 1970, 1353–1526.
Glaser, R.: Adaptive Education: Individual Diversity and Learning, New York: Holt, Rinehart & Winston 1977.
Gösslbauer, J.P.: Tests als Selektionsinstrumente – fair oder unfair? Psychologie und Praxis 1977, 21, 95–111.
Gross, A.L. & Su, W.: Defining a „Fair" or „Unfair" Selection Model: A Question of Utilities. Journal of Applied Psychology 1975, 60, 345–351.
Guthke, J.: Zur Diagnostik der intellektuellen Lernfähigkeit, Berlin: Verlag Erziehung und Bildung 1972. (Oder: Stuttgart: Klett 1976).
Harnischfeger, A. & Wiley, D.E.: Teaching-Learning Processes in Elementary School: A Synoptic View. In: Curriculum Inquiry 1976, 6, 5–43.

Harnischfeger, A. & Wiley, D.E.: Kernkonzepte des Schullernens. Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie 1977, 9, 207–230.
Hunter, J.E. & Schmidt, F.L.: Critical Analysis of the Statistical and Ethical Implications of Various Definitions of Test Bias, Psychological Bulletin 1976, 83, 1053–1071.
Jensen, A.R.: Another Look at Culture-fair Testing. In: Jensen. Educational Differences. London: Methuen 1973.
Jensen, A.R.: Précis of Bias in Mental Testing. The Behavioral and Brain Sciences 1980, 3, 325–371.
Jöreskog, K.G.: Statistical Estimation of Structural Models in Longitudinal-Developmental Investigations. In: Nesselroade & Baltes (eds.): Longituinal Research in the Study of Behavior and Development. New York: Academic Press 1979, 303–351.
Jöreskog, K.G. & Sörbom, D.: Lisrel V: Analysis of Linear Structural Relationships by Maximum Likelihood and Least Squares Methods, User's Guide, University of Uppsala, Department of Statistics, P.O.Box 513, S-751 20 Uppsala. Sweden: 1981.
Kleiter & Petermann, F.: Abbildung von Lernwegen. München: Oldenbourg 1977.
Lienert, G.A.: Testaufbau und Testanalyse. Weinheim: Beltz 1969[3].
Linn, R.L.: Fair Test Use in Selection. Review of Educational Research 1973, 43, 139–161.
Linn, R.L.: In Search of Fair Selection Procedures. Journal of Educational Measurement 1976, 13, 53–58.
Lord, F.M.: Applications of Item Response Theory to Practical Testing Problems, Hillsdale, N.J.: Erlbaum 1980.
Lord, F.M. & Novick, M.R.: Statistical Theories of Mental Test Scores, Reading, Mass.: Addison-Wesley 1968.
Löschenkohl, E.: Gibt es einen allgemeinen faßbaren Zusammenhang zwischen Schulleistung und Intelligenz? Psychologie in Erziehung und Unterricht 1973, 20, 145–155.
McClelland, D.C.: Testing for Competence Rather Than for Intelligence. American Psychologist 1973, 28, 1–14.
Möbus, C.: Grundfragen psychologischer Diagnostik 1976: Fairness und Validität. In: Tack (Hrsg.): Bericht über den 30. Kongreß der Deutschen Gesellschaft für Psychologie in Regensburg, Bd. 2, 17–20, Göttingen: Hogrefe 1977.
Möbus, C.: Zur Fairness psychologischer Intelligenztests: Ein unlösbares Trilemma zwischen den Zielen von Gruppen, Individuen und Institutionen? Diagnostica 1978, 24, 191–234.
Möbus, C. & Nagl, W.: Messung, Analyse und Prognose von Veränderungen 1983. In: Bredenkamp & Feger (Hrsg.): Hypothesenprüfung, Bd. 5 der Serie Forschungsmethoden der Psychologie in der Enzyklopädie der Psychologie, 239–470, Göttingen: Hogrefe.
Möbus, C. & Simons, H.: Zur Fairness psychologischer Intelligenztests gegenüber ethnischen und sozialen Gruppen: Kritik klassischer Konzepte. Bericht aus dem Psychologischen Institut der Universität Heidelberg 1975, Nr. 2.
Novick, M.R.: Policy Issues of Fairness in Testing. In: van der Kamp, Langerak & de Gruijter (eds.): Psychometrics for Educational Debates. Chichester: Wiley 1980, 123–137.
Novick, M.R. & Petersen, N.S.: Towards Equalizing Educational and Employment Opportunity. Journal of Educational Measurement 1976, 13, 77–88.
Opwis, K. & Gold, A.: Vergleichende Bedingungsanalyse schulischer Leistungen: Die Analyse von Zeitreihen von Querschnitten mittels linearer Strukturgleichungssysteme, Diplomarbeit am Psychologischen Institut der Universität Heidelberg 1982.
Petersen, N.S.: Bias in the Selection Rule – Bias in the Test, in: van der Kamp, Langerak & de Gruijter (eds.): Psychometrics for Educational Debates. Chichester: Wiley 1980, 103–122.
Petersen, N.S. & Novick, M.R.: An Evaluation of Some Models for Culture-fair Selection. Journal of Educational Measurement 1976, 13, 3–31.

Rosenshine, B.: Classroom Instruction. In: Gage (ed.): The Psychology of Teaching Methods. Chicago: University of Chicago Press 1976.
Rosenshine, B. & Furst, N.: The Use of Direct Observation to Study Teaching. In: Travers (ed.): Second Handbook of Research on Teaching. Chicago: Rand McNally 1973, 122–183.
Sawyer, R.L., Cole, N.S. & Cole, J.W.L.: Utilities and the Issue of Fairness in a Decision Theoretic Model for Selection. Journal of Educational Measurement 1976, 13, 59–76.
Scheuneman, J.: Latent-Trait Theory and Item Bias. In: van der Kamp, Langerak & de Gruijter (eds.): Psychometrics for Educational Debates. Chichester: Wiley 1980, 139–151.
Schneeweiss: Ökonometrie. Würzburg: Physica 1971.
Simons, H. & Möbus, C.: Untersuchungen zur Fairness von Intelligenztests. Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie 1976, 8, 1–12.
Simons, H. & Möbus, C.: Veränderungen von Berufschancen durch Intelligenztraining. Bericht aus dem Psychologischen Institut der Universität Heidelberg 1977, Nr. 8.
Simons, H. & Möbus, C.: Testfairness. In: Klauer (Hrsg.): Handbuch der Pädagogischen Diagnostik 1979, 187–197.
Solomon, D., Rosenberg, L. & Bezdek, W.E.: Teacher Behavior and Student Learning. Journal of Educational Psychology 1964, 55, 23–30.
Thorndike, R.L.: Concepts of Culture-Fairness. Journal of Educational Measurement 1971, 8, 63–70.
Treiber, B.: Qualifizierung und Chancenausgleich in Schulklassen. Frankfurt: Lang 1980.
Treiber, B.: Aptitude-Treatment-Interaction. In: Schiefele & Krapp (Hrsg.): Handlexikon Pädagogische Psychologie. München: Ehrenwirth 1981.
Treiber, B.: Lehr- und Lernzeiten im Unterricht. In: Treiber & Weinert (Hrsg.): Lehr–Lernforschung. München: Urban & Schwarzenberg 1982, 12–36.
Treiber, B. & Weinert, F.E.: Lehr–Lernforschung. München: Urban & Schwarzenberg 1982.
Weinert, F.E. & Treiber, B.: Unterrichtsqualität und Leistungszuwachs bei Formen direkter Instruktion im Mathematikunterricht 5. Hauptschulklassen. Forschungsantrag bei der DFG 1981 (zitiert nach Opwis & Gold 1982).
Wright, B.D. & Stone, M.H., Best Test Design: Rasch Measurement, Chicago: Mesa Press, 1979
Wottawa, H. & Amelang, M.: Einige Probleme der „Testfairness" und ihre Implikationen für Hochschulzulassungsverfahren. Diagnostica 1980, 26, 198–220.
Zielinski, W.: Lernschwierigkeiten. Verursachungsbedingungen, Diagnose, Behandlungsansätze. Stuttgart: Kohlhammer 1980.

Horn/Ingenkamp/Jäger (Hrsg.)

TESTS UND TRENDS 3

Jahrbuch der Pädagogischen Diagnostik

Die Jahrbücher stellen neuere Entwicklungen kritisch dar und erörtern die Konsequenzen für die praktische Diagnostik.
Mit Beiträgen von:
F. Schott/H.-J. W. Wieberg/K.-E. Neeb: Probleme Kriterienbezogener Leistungsmessung, H. Probst: Testverfahren zur Diagnostik spezifischer Lernvoraussetzungen, U. Raatz/Chr. Klein-Braley: Ein neuer Ansatz zur Messung der Sprachleistung, K. A. Heller: Diagnostische Ausbildung und Tätigkeit von Beratungslehrern, C. Möbus: Die praktische Bedeutung der Testfairness als zusätzliches Kriterium zu Reliabilität und Validität.
Testbesprechungen – Verzeichnis deutschsprachiger Schultests.

Beltz

Ralf Horn, Dipl.-Psych., Leiter der Beltz Test Gesellschaft, Weinheim

Karlheinz Ingenkamp, Dr. phil., Professor für Pädagogik und Leiter des Zentrums für empirische pädagogische Forschung der Erziehungswissenschaftlichen Hochschule Rheinland-Pfalz, Landau.

Reinhold S. Jäger, PD Dr. phil., Leiter der Unterabteilung Testforschung und Testentwicklung des Deutschen Instituts für Internationale Pädagogische Forschung, Frankfurt.

CIP-Kurztitelaufnahme der Deutschen Bibliothek

Tests und Trends : Jahrbuch d. Pädag. Diagnostik. - Weinheim, Basel : Beltz
Erscheint jährl.
1981 ff.

Umschlaggestaltung: E. Warminski, Frankfurt/M.
Printed in Germany

ISBN 3 407 54638 6