Sonderdrucke aus

DIAGNOSTICA

Heft 3/1978

Diagnostica 1978, Band XXIV, Heft 3, S. 191–234

Zur Fairness psychologischer Intelligenztests: Ein unlösbares Trilemma zwischen den Zielen von Gruppen, Individuen und Institutionen?

Claus Möbus
Institut für Psychologie der Freien Universität Berlin
im Fachbereich 12 (Erziehungswissenschaften)

1. Einleitung

„Weißer Amerikaner klagt wegen Rassendiskriminierung. Bei Uni-Zulassung zugunsten von Anwärtern rassischer Minderheiten abgelehnt. Kann auch ein Weißer, besser gesagt ein ‚europäischer' Amerikaner das Opfer einer Rassendiskriminierung in den USA werden? Die Antwort lautet in den Augen des 37jährigen Studenten Allan Bake ‚ja', und die amerikanische Regierung hat dieser Tage bestätigt, daß ein weißer Bürger der Vereinigten Staaten nach ihrer Ansicht mit seiner Hautfarbe gelegentlich völlig legal Pech haben kann.

Bake hatte sich an der Medizinischen Fakultät der kalifornischen Universität in Davis eingeschrieben, war jedoch zurückgewiesen worden. An seiner Stelle akzeptierte die Universität mehrere Anwärter rassischer Minderheiten, obwohl sie im Aufnahmeverfahren schlechter abschnitten als er. Bake klagte auf ‚umgekehrte Diskriminierung', und der Oberste Gerichtshof der Vereinigten Staaten hat nun die schwierige Frage zu entscheiden, ob auch heute noch in den USA ein Bürger wegen seiner Hautfarbe zurückgesetzt werden darf.

Der Fall hat eine so erhebliche politische Brisanz, so daß sich das Justizministerium nach Beratungen mit Präsident Jimmy Carter mit einer schriftlichen Entscheidungshilfe – in der amerikanischen Rechtssprechung als ‚amicus curiae' bekannt – einschaltete. Dieses Schriftstück sagt nichts anderes, als daß rassische Minderheiten wie Neger oder Mexikaner in gewissen Situationen Vorrechte gegenüber der Mehrheit beanspruchen können.

Natürlich lehnt auch die Regierung eine umgekehrte Rassendiskriminierung offiziell ab und sie hält auch ein Quotensystem bei Universitäten oder in der Wirtschaft für nicht vertretbar. Sie verweist jedoch auf die besondere Verpflichtung des Staates, die Jahrzehnte unterdrückten Minderheiten in den USA zu fördern und ihnen gelegentlich auch zum Nachteil der Weißen Chancen einzuräumen, die sie aufgrund unverschuldeter mangelnder Vorbildung oder ebenso unverschuldeter Armut bisher nicht nutzen konnten.

Diesem Prinzip, das die kalifornische Universität offenbar befolgte, ist der nicht eben taufrische Student Bake zum Opfer gefallen. Ginge es allein nach der Regierung, er würde vermutlich eines höheren Zieles willen geopfert. Doch das letzte Wort hat der Gerichtshof. Man erwartet, daß er das Verfahren an eine untere Instanz zur nochmaligen Verhandlung zurückverweisen wird. Für Bake würde das mindestens zwei weitere Prozeßjahre bedeuten.

Der Fall wird von Bürgerrechtlern als das wichtigste Verfahren in einer Minderheitenfrage der letzten 20 Jahre gewertet. Sie verweisen – mit Recht – darauf, daß es in den USA beispielsweise noch längst nicht genügend farbige Ärzte gibt, und daß weiße Berufsvertreter dieser Gruppe sich oft nur unwillig der Minderheiten annehmen. Letztlich geht es nicht um Bake, sondern um die restlose Eingliederung der zu kurz Gekommenen in die Wohlstandsgesellschaft. Im Oktober soll die Entscheidung fallen." Washington.

Meldung der dpa vom 22. September 1977

„Richter zerbrechen sich die Köpfe. Klage wegen ‚umgekehrter Rassendiskriminierung'. WASHINGTON. (dpa) Im Herbst 1973 bewarb sich der weiße Amerikaner Allan Bake um Aufnahme als Medizinstudent an der Universität von Kalifornien in Davis. Er wurde abgelehnt. Auch ein zweiter Antrag half nichts. Aufgenommen wurden unter 100 Studenten 16 Schwarze, die schlech-

tere Prüfungsergebnisse hatten als der blonde, blauäugige Nachfahre norwegischer Einwanderer. Der gelernte Ingenieur klagte wegen umgekehrter Rassendiskriminierung. ... Jetzt hörte sich das Oberste Amerikanische Bundesgericht in Washington vor überfüllten Zuhörerbänken als letzte Instanz die Argumente der Anwälte von Bake und der Universität sowie eines Vertreters der US-Regierung an. Die Entscheidung der neun Richter, ob Institutionen zugunsten von Minoritäten die Rasse von Antragstellern in Erwägung ziehen dürfen, wird zwar nicht vor Ende des Jahres erwartet. Für die Entwicklung der Bürgerrechte gilt der Fall Bake aber schon jetzt als einer der wichtigsten seit dem höchstrichterlichen Verbot von rassisch getrennten Schulen im Jahre 1954.

Bakes Anwalt Reynold Colvin versuchte den Richtern klarzumachen, warum sein Mandant sich ungerecht behandelt fühlte: ‚Egal, wie man es dreht, Mr. Bake durfte die Schule aufgrund seiner Rassenzugehörigkeit nicht besuchen.'

Anwalt Archibald Cox, der Berühmtheit als Watergate-Ankläger erlangte, konterte im Namen der medizinischen Hochschule: ‚Jahrelang hat Rassendiskriminierung in den Vereinigten Staaten einige Minderheiten isoliert ... und sie zu schlechter Ausbildung verdammt ... jetzt wollen Schulen freiwillig die Zahl von Ärzten aus Minderheitsgruppen erhöhen und ein Modell setzen für die nächste Generation von Schwarzen.' Cox betonte, es würden kaum Farbige ein Medizinstudium beginnen können, wenn man bei der Auswahl ‚rassenblind' wäre. Auch der ehemalige schwarze Bundesrichter Wade McCree, der für die US-Regierung sprach, argumentierte für eine gewisse Sonderbehandlung: ‚Blind gegenüber der Rasse zu sein, heißt die Augen vor der Realität zu verschließen.'

Die Klage umgekehrter Diskriminierung hat viele Schwarze, die den Kampf um Gleichberechtigung noch lange nicht als beendet ansehen, zutiefst empört. Sie verweisen darauf, daß von 100 Ärzten in den USA nur zwei Schwarze sind. Sie betonen auch, daß im Rahmen von Minoritäten-Programmen aufgenommene Studenten nach Anfangsschwierigkeiten zum größten Teil ihr Bildungsdefizit wettmachen konnten.

Die dunkelhäutige Ärztin Toni Johnson-Chavis nannte noch ein weiteres Argument: ‚Die Weißen wollen nach dem Studium schnell Geld verdienen. Von armen Leuten, schwarzen Kranken, wollen sie dann nichts mehr wissen.' "

Meldung der dpa vom 14. Oktober 1977

Diese Meldungen machen deutlich, daß die Fragen nach Chancengleichheit, Diskrimination von Minoritäten oder anderen Gruppen von Personen, Fairness psychologischer Aussagen und Entscheidungen in Zukunft wohl noch drängender werden als man bisher angenommen hat.

Unter einem fairen Intelligenztest wurde bis 1968 meist ein Test verstanden, der jedem eine Chance einräumte, sich seiner individuellen Lerngeschichte entsprechend intelligent zu verhalten. Folglich sei ein unfairer Test so konstruiert, daß er Gruppen mit unterschiedlichen Lernerfahrungen diskriminiere. Der Test bevorzuge oder benachteilige Gruppen nur, weil die Auswahl seiner Items nach einem Intelligenzkonzept erfolge, das nur einer Gruppe gerecht werde („Itembias"; siehe hierzu: Eells, Davis, Havighurst, Herrick & Tyler, 1951; Cardall & Coffman, 1964; Angoff, 1972; Green & Draper, 1972; Merz, 1973, 1976; Möbus & Simons, 1975; Scheuneman, 1975; Veale & Foreman, 1975, 1976; Möbus, 1976; Rudner, 1977; Wolf, 1977). Da die Ausschaltung des Itembias am saubersten mit den Modelltests zum Rasch-Modell gelingt (wenn die Items gleiche Trennschärfe besitzen und die richtige Antwort nicht einfach geraten werden kann) und dieser Ansatz in deutschsprachigen Ländern weit bekannt ist, wollen wir uns hier auf den „Testbias" oder die „Testfairness" 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,

„... da *alle* geistigen Fähigkeiten sich erst in der unter bestimmten äußeren Bedingungen verlaufenden Tätigkeit des Individuums herausbilden (vgl. Lompscher, 1970; Keiser, 1969). Auch die sogenannten Handlungs- oder performance-Tests ... sind natürlich bildungsabhängig.

... Zweitens gehören die verbal-theoretischen Fähigkeiten zu den wesentlichsten und auch für die Lebensbewährung (vgl. a. Orlik, 1967a) sehr entscheidenden Kriterien der Intelligenz eines Menschen, so daß deren Eliminierung aus intelligenz-diagnostischen Verfahren letztlich zu einer einseitigen, verzerrten Widerspiegelung der intellektuellen Potenz führen würde.

Wir fordern also eine möglichst alle wesentlichen Dimensionen des im Alltag zu beobachtenden intellektuellen Leistungsverhalten berücksichtigende Leistungsdiagnostik. ..." (Guthke, 1972, s. 25 f.)

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 Getesten, 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, Humphreys, Kendrick & Wesman, 1975; Gross & Su, 1975; Möbus & Simons, 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; Simons & Möbus, 1976; Petersen, 1977; Gösslbauer, 1977).

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?

Dieses Konzept ist umfassender als z. B. das der „differentiellen Validität" (Saunders, 1956; Kirkpatrick, Ewen, Barrett & Katzel, 1968) oder der „dif-

ferentiellen Prädiktibilität" (Banas, 1964; Ghiselli, 1956, 1960, 1963; Hobert & Dunette, 1967), die beide als Störquellen und unerwünscht empfunden werden. Die differentielle Validität wird von uns als Nachteil angesehen, der durch entsprechende Hinzunahme weiterer Tests ausgeschaltetet werden muß: Für jede Gruppe ist ein gleich enger Zusammenhang zwischen Test und Kriterium zu fordern. Ist ein Test für eine Gruppe weniger valide, sollte durch Hinzunahme weiterer Variabler mittels optimierter Linearkombination eine neue „Testvariable" geschaffen werden, die innerhalb jeder Gruppe gleich hoch mit dem Kriterium korreliert (vgl. Simons & Möbus, 1975).

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 Fähigkeitstest nicht erklärt und vorhergesagt werden können, *nicht* ignoriert. Würde man diese Unterschiede verleugnen, würde das einer Diskriminierung der im Kriterium besseren Gruppe gleichkommen. Ferner wird die Gruppenzugehörigkeit bei Entscheidungen über Personen benutzt: *gruppenspezifische* Regressionen, cutoffs, Tests etc.; bzw. differentielle Validität. 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 Indivi-

dualismus 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 wird von mir im Gegensatz zu einigen Autoren durch die Einführung einer neuen Betrachtungsebene (Regressionskonzepte) wieder aufgewertet.

2. Klassische Testfairnesskonzepte

Ich will hier nur die klassischen Konzepte, die sich bis 1976 entwickelt haben, vorstellen. Meine Kritik ist sinngemäß auf die allerneuesten Entwicklungen übertragbar, weil auch hier cutoffs bestimmt werden. Ferner wirft der Einbezug von entscheidungstheoretischen und bayesianischen Vorstellungen in die neuesten Modelle (Petersen & Novick, 1976; Petersen, 1976) Probleme auf, die in diesem Rahmen aus Platzgründen nicht mehr behandelt werden können.

Zwischen 1968 und 1976 haben sich in der psychologischen Forschung neben dem „Identitätskonzept" (Darlington's 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

$$E(Y|x_i^*) = y^* = a_i + b_i x_i^* \quad (i = 1, ..., g) \tag{2.1}$$

gelten, wobei:

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

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

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. So kritisiert Thorndike (1971) an der Regressionsdefinition von Cleary, daß bei der Erfüllung der Annahme einer gemeinsamen Regression die Überlappung der Randverteilungen in Y größer ist als die Überlappung der Verteilungen in X.

"A larger percentage of total minor group would achieve the acceptable criterion level than would be accepted on the basis of the test." (Thorndike, 1971, S. 67)

Thorndike (1971) entdeckte eine Verletzung des „fair share"-Gesichtspunktes: Die Definition kann unfair gegenüber ganzen Gruppen sein, wenn bei Gruppen mit schlechten Testwerten keine Person akzeptiert wird, obwohl etliche Mitglieder dieser Gruppe im Kriterium erfolgreich gewesen wären. Ein anderer Einwand gegen Cleary wurde wegen der Unterstützung konservativer Akzeptanzpraktiken vorgebracht: Liegen eine differentielle Validität des Tests und gleiche Kriteriumsmittelwerte vor und werden nur wenige Probanden akzeptiert, dann liegt das erforderliche minimal zufriedenstellende Testergebnis (cutoff x^*) für die Gruppe mit niedriger Testvalidität höher als das entsprechende Testergebnis für die Gruppe mit höherer Testvalidität. Aus einer ähnlich konservativen Grundhaltung (Risikominimierung) scheint das Konzept von Einhorn & Bass geboren zu sein: cutoffs (x_i^*) müssen so gesetzt werden, daß die minimale Erfolgswahrscheinlichkeit für alle Gruppen gleich ist. Diese Prozedur ist in der Praxis noch stärker auf die Bedürfnisse einer Institution zugeschnitten als die Cleary-Definition. Darlington kritisierte an der Clearyschen Konzeption, daß der faire Test mit der ethnischen oder kulturellen Drittvariablen höher als das Kriterium korrelieren sollte: Die Randverteilungen der Gruppen überlappen sich im Test weniger als im Kriterium ($r_{zx} > r_{zy}$).

"The mind boggles at the vision of a private school using admissions 'tests' so culture-bound that in effect automatically admits all whites and automatically rejects all blacks, with its principal admitting to a low correlation between race and academic aptitude, and yet pointing out that the test is culturally fair by what is currently the most widely accepted definition of cultural fairness." (Darlington, 1971, S. 75)

Für einen gegebenen Mittelwertsunterschied (Überlappungsbereich der Gruppenrandverteilungen) im Kriterium muß der Mittelwertsunterschied (Überlappungsbereich) im Test umso größer werden, je invalider der Test ist, um nach Cleary fair zu sein.

Es kann daher unter Umständen der Fall eintreten, daß zwei Tests die gleiche Mittelwertsdifferenz (Überlappung der Randverteilungen im Test) aufweisen, daß aber der invalidere der nach Cleary faire Test ist.

Diese Argumentationskette wurde von Linn (1971) noch fortgeführt: Ist ein Test nach Cleary fair und wird seine Reliabilität erhöht, verbessert sich sowohl seine Validität; die Regression von Y auf X wird steiler und damit die Überlappung der Randverteilungen der Gruppen auf X kleiner (r_{xz} erhöht sich). Dagegen bleibt die Überlappung der Randverteilungen auf dem Kriterium aber gleich (r_{yz} bleibt gleich). Durch das Ansteigen der gruppenspezifischen Regressionen haben die Gruppen aber keine gemeinsame Regression mehr, so daß der Test nach Cleary unfair wird, wenn man an der gemeinsamen Regression und an einem für alle Gruppen gültigen minimal zufriedenstellendem Testergebnis (cutoff x*) festhält. Dieser Gedankengang läßt sich noch auf die Spitze treiben: Die Summe (oder das Mittel) zweier fairer Tests ergibt einen unfairen Test (Darlington, 1971, S. 75).

Bisher nahmen alle Überlegungen zum Cleary-Konzept die Fairness des Kriteriums an. Dieses bezweifelte Darlington (1971):

"Suppose we knew the exact criterion scores of a white man and a black man. What would the difference between their criterion scores have to be in order for them to be regarded as equally desirable candidates for selection?" (Darlington, 1971, S. 79)

Da Cleary diese Überlegung außer Acht ließ, glaubte Darlington sein Bonus/Malusmodell („culture-modified criterion") einführen zu müssen (s. a. unten).

Die stärkste Ähnlichkeit mit Cleary's Definition hat

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

Einhorn & Bass geben keine direkte Difinition, sondern beschreiben eine Prozedur, mit der Diskrimination durch Testsverhindert 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 necissarily 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:

$$P[Y > y^* \mid X = x_i^*] = Z \quad \text{für } i = 1, \ldots g \tag{2.2}$$

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 Applikanten aus) ist die Risiko- und Kostenminimierung 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 Va-

lidität) stammen (so z. B. bei durch Lehrer gut geförderte Klassen; siehe Simons & Möbus, 1975, oder bei Weißen) dadurch benachteiligt, daß die cutoffs x_i^* im Gegensatz zu anderen Gruppen (z. B. schlecht geförderte Klassen oder Neger) 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 ($Y-kZ$[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 ($Y-kZ$)." (Darlington, 1971, S. 80)

Dieser Ansatz wird bis heute von einer Reihe von Autoren begrüßt:

"... Darlington's formulation of selection fairness recognizes explicitly that the variable which is traditionally considered to be the criterion (e. g. college grade point average) is not the only criterion. Group membership or culture is also the part of the criterion (Petersen & Novick, 1976, S. 23).

Andererseits dürfen aber auch Zweifel an der praktischen Nützlichkeit des Konzeptes angemeldet werden. Die Bestimmung des Maluskoeffizienten k ist subjektiv. Y und Z besitzen nicht gleiche Skalen, so daß die Substraktion etwas artifiziell erscheint. Liegen mehrere kulturelle Variable vor, muß das modifizierte Maluskriterium ($Y-k_1Z_1-k_2Z_2-k_3Z_3...$) vorhergesagt werden. Dabei müssen die Maluskoeffizienten k_i vorher festgelegt werden. Da aber die z_j in der Regel untereinander korrelieren, dürfen die k_j nicht auf die *generelle* Wichtigkeit einer Variablen Z_j bezogen werden, sondern nur auf die partielle Wichtigkeit unter Konstanthaltung aller anderen Variablen. Diese Problematik ist identisch mit der Schwierigkeit, Regressionsgewichte bei korrelierten Prädiktoren zu interpretieren.

2 Z = kulturelle oder ethnische Drittvariable (z. B. dichotom für die Variable Weiß/Schwarz oder kontinuierlich für die Variable SES).

Eine ursprünglich positiv gemeinte Bemerkung von Linn (1973) zeigt aber, daß das culture-modified criterion nichts anderes als eine verkappte und gemilderte Form des ‚Identitätskonzeptes' ist.

"... If we want to predict Y–kZ, it is quite natural and appropriate to use Z as one of the predictors. In this situation, what is desired is a test which would maximize the multiple correlation of X and Z with the criterion variable Y–kZ. Fred Lord (personal communication) has shown that this multiple correlation is maximized when the partial correlation between X and Y with Z partialled out is maximized, this is a nice result in that it does not depend on the value of k." (Linn, 1973, S. 157)

Nun zeigt sich aber, daß nach dem Identitätskonzept $r_{xz} = 0.0$ bzw. $r_{yx.z} = r_{yx}$ sein soll. Das moderne Konzept „subjektive Regression" hat sich als gemilderte Version des wohlbekannten Identitätskonzeptes entpuppt!

Eine ebenfalls kritische Betrachtung stellten Hunter & Schmidt (1976) an:

"... Thus, a cynic might well assume that in practice Darlington's definition of *culturally optimum* would simply lead to the selection of a value of k that would make whatever test was being used appear to be culturally optimum. ... What is the upshot of Darlington's suggestion? From a mathematical point of view, adding or subtracting a constant to the criterion is exactly equivalent to adding or subtracting constants to the predictor, and this in turn is equivalent to using different cut-off scores for the two groups. Thus in the last analyses, Darlington's method is simply an esoteric way of setting quotas, and hence the expense of constructing culturally optimum tests to do this way is a waste of time and money." (Hunter & Schmidt, 1976, S. 1066)

In dem oben erwähnten Artikel stellt Darlington noch ein anders 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 McClelland'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:

$$E_i (X \mid Y) = E (X \mid Y) \quad \text{für alle } i = 1, ..., g \tag{2.3}$$

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

$$P_i [X > x^* \mid Y] = P [X > x^* \mid Y] \quad \text{für alle } i = 1, ..., g \tag{2.4}$$

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 homogen 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 $r_{xz.y} \neq 0.0$, dann besteht auf dem Test 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, auf 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.

Trotz der offenkundigen Plausibilität der Definition d) regte sich auch hier Widerspruch. Besonders Hunter & Schmidt (1976) kritisierten die ihrer Meinung nach unlogische Auspartialisierung einer zeitlich später zu erbringenden Kriteriumsleistung Y aus einem früher anfallenden Testergebnis X. Dieses Argument ist aber unserer Ansicht nach nur z. T. stichhaltig, da der Einsatz eines Tests nur gerechtfertigt ist, wenn er „stellvertretend" für das Kriterium steht und die langfristige Prognose auf Grund eines statusorientierten Variablenansatzes prinzipiell nur dann möglich ist, wenn die Reaktionsfreiheit des Individuums stark eingeengt wird. Im Gegensatz zu Hunter & Schmidt sind wir der Ansicht, daß zumindest im Fall der konkurrenten Validität nichts gegen die Partialisierungsrichtung einzuwenden ist.

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:

$$K = \frac{P_g\,[X_g > x_g^*]}{P_g\,[Y_g > y^*]} = \frac{P_g\,[X_h > x_h^*]}{P_h\,[Y_h > y^*]} \quad \text{für alle } g, h = 1, \ldots G \tag{2.5}$$

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

Die Definition (2.5) stellt eine Abkehr vom validitätsmaximierenden ‚Individualismus' und eine Hinwendung zu einem gruppenorientieren 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 group 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:

$$(2.6) \qquad K = P[X_g > x_g^* \mid Y_g > y^*] = P[X_h > x_h^* \mid Y_h > Y^*]$$

für alle $g, h = 1, ..., G$

Dabei ist:

$$(2.7) \qquad P[X>x^* \mid Y>y^*] = \frac{P[X>x^*]}{P[Y>y^*]} P[Y>y^* \mid X>x^*]$$

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 Applikanten 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).

"... In many situations the rights of potentially successful applicants to fairness should be of primary concern. Under the conditional probability model such applicants are guranteed equal chance of selection regardless of group membership. This procedure places the burden of improving prediction on the selecting institution ... If a test tends to work poorly in predicting criterion scores for minority members, under the conditional probability model the selecting institution must compensate for the poor predictors by selecting more, not fewer, minority students. It should be noted that the conditional probability model does not eliminate or de-emphasize the use of tests or other predictor variables in the selection process. In fact, under the model the selecting institution and the potentially successful applicant both benefit from the use of better tests or other predictors within each group. However, use of the model should provide an important step in the direction of insuring minority group members, that their rights are not being subverted by tests chosen by and constructed by majority group members." (Cole, 1973, S. 253 f.)

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 Applikanten 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:

$$(2.8) \qquad P[Y_g \geq y^* \mid X_g \geq x_g^*] = P_h[Y_h \geq y^* \mid X_h \geq x_h^*]$$

Einige Autoren setzten unzulässigerweise d) mit f) und a) mit g) gleich. Dabei wurde außer Acht gelassen, daß bei f) und g) in die jeweiligen bedingten Wahrscheinlich-

keiten Abschnitte von Randverteilungen eingehen (z. B. $P[\ldots \mid X \geqslant x^*]$ oder $P[\ldots \mid Y \geqslant y^*]$), nicht aber wie bei a) und d) nur bestimmte Ereignisse (z. B. $P[\ldots \mid X = x^*]$ oder $P[\ldots \mid Y = y^*]$). 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:

$$P[X_g > x_g^* \mid Y = y^*] = P[X_h > x_h^* \mid Y = y^*] \tag{2.9}$$

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

i) konverse Modelle

Betrachtet man bei a) – h) auf dem Kriterium statt der Erfolgs- die Mißerfolgswahrscheinlichkeiten und auf dem Test statt der Selektions- die Ablehnungswahrscheinlichkeiten, kommt man zu den konversen Modellen (bzw. Definitionen). Es ist der Verdienst von Petersen & Novick (1976), gezeigt 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.

3. Bezugsrahmen für eine weitgehend einheitliche Darstellung der Definitionen a) – h)

Die Definitionen können auf vier Ebenen diskutiert werden: (3.1) Akzeptanz/Erfolgsquoten, (3.2) korrelative Zusammenhänge und Pfadmodelle, (3.3) Regressionsmodelle und (3.4) Selektionscharakteristikkurven (SCK). Dabei gehe ich (wie schon 1976) 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 „Schwarze" vs. „Weiße" wegen der Rassenmischung 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.

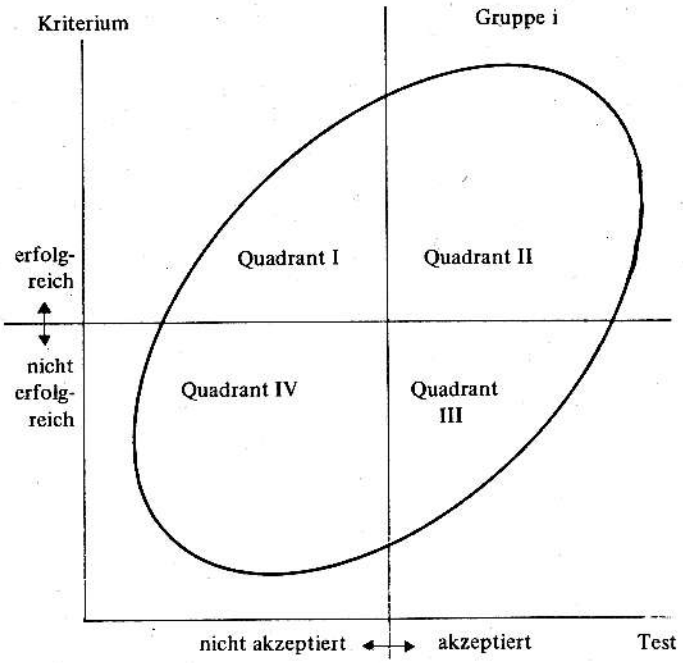
3.1 Akzeptanz-Erfolgsquoten

Betrachten wir die Verhältnisse im kombinierten Kriteriums-Prädiktorraum, können wir durch entsprechende cut-offs 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 cut-offs, sind die Testwerte zwischen den Gruppen vergleichbar (vgl. a. Figur 1).

Figur 1

Durch cut-offs aufgeteilter Test-Kriteriumsraum

Der Anteil von Personen in den Quadranten I und III sollte möglichst niedrig sein, da er die prädiktive Validität des Tests stört.



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:

(3.1.1)

$$\Sigma II_i + \Sigma IV_i = \max!$$

Summe aller sich in den Quadranten II und IV befindlichen Personen.

$$\Sigma I_i + \Sigma III_i = \min!$$

Summer aller sich in I und III befindlichen Personen.

i = Gruppenindex

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:

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

wobei: i, j = Gruppenindices sind.

Dagegen bedeutet Fairness nach Thorndike, daß

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

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

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

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

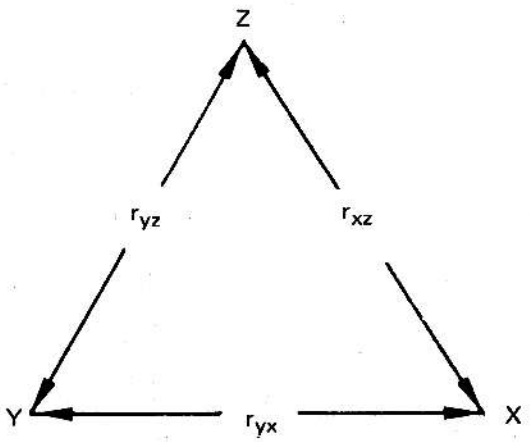
Nach Cole ist ein Test fair, wenn der Prozentsatz der Akzeptierten unter den Erfolgreichen in allen Gruppen gleich ist:

(3.1.5) $$II_i/(I_i + II_i) \stackrel{!}{=} II_j/(I_j + II_j)$$

3.2 Korrelative Beziehungen und Pfadmodelle

Einige Fairnesskonzepte lassen sich im korrelativen Netz zwischen Test, Kriterium und ethnischer bzw. soziologischer Drittvariabler 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 2).

Figur 2
Korrelative Beziehungen zwischen Test X, Kriterium Y und Drittvariabler Z



wobei: r_{yz} = 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).

Unter den schon erwähnten Annahmen können wir die Testfairnesskonzepte in ein korrelationsstatistisches Bezugssystem bringen, in dem r_{yz} und 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 nicht 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:

(3.2.1) $r_{xz} = f(r_{yz}, r_{yx})$

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:

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

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, ..., j, ..., G) gleich sind. Damit sind auch die bedingten Erwartungswerte $E(Y_{z_i}|X) = E(Y_{z_i}|X)$ einander gleich. Aus diesem Grund muß $r_{yz.x} = 0.0$ sein. Daraus ergibt sich:

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

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

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

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 (3.2.4).

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

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

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

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

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.

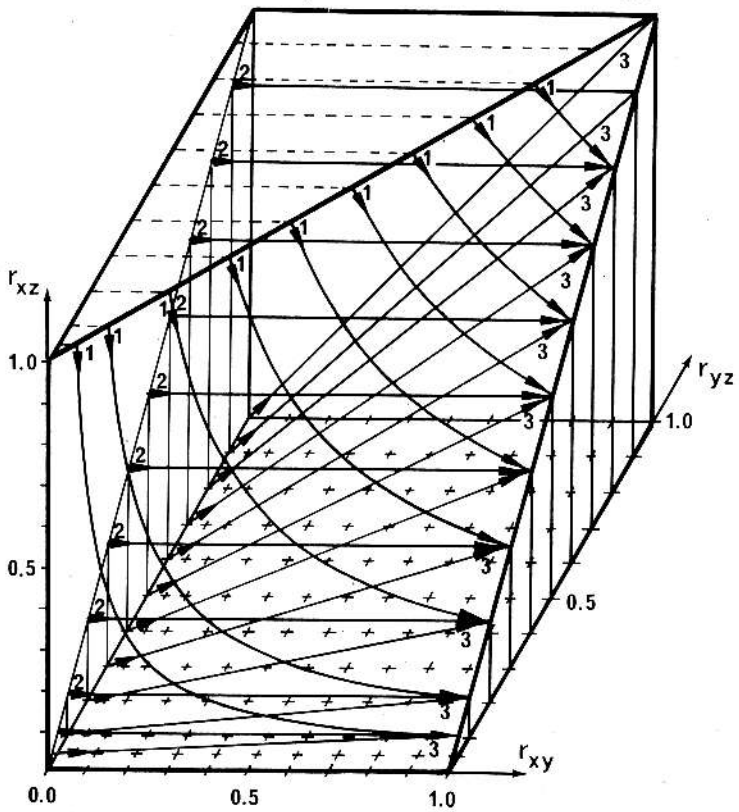
Auch das altbekannte Identitätskonzept läßt sich korrelativ darstellen:

(3.2.7) $$r_{xz} = 0.0 \text{ bzw. } r_{y(x.z)} = r_{yx} = r^1_{yx.z} \sqrt{1 - r^2_{yz}}$$

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.

Figur 3

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 3 auf. Je nach Fairnessdefinition muß der

3 siehe Linn (1973, S. 157). Linn führt Lord an: Darlingtons subjektive Regression wird maximiert, wenn $r_{yx.z}$ maximiert wird. Damit ist natürlich evident, daß Darlingtons subjektive Regression ein verkapptes modifiziertes Identitätskonzept ist.

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 (Thorndike's Definition) oder eine Linie 1 (Cleary's Definition) stoßen. Die Höhen der Linien 1, 2, 3 entsprechen den Funktionswerten der Fairnessdefinitionen (3.2.2), (3.2.4) und (3.2.6). Das Identitätskonzept fordert $r_{xz} = 0.0$ für alle r_{yx} und r_{yz}, so daß wir die Fußebene von Figur 3 nicht verlassen müssen. Die Definitionen stimmen bei perfekter Validität ($r_{xy} = 1.0$) oder kultureller Unabhängigkeit des Kriteriums überein.

Die Zusammenhänge zwischen den Definitionen sind in Figur 3 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 3).

Auch hier zeigt sich die Mittelstellung der Thorndikeschen Formulierung. Bildet man nämlich das geometrische Mittel $\bar{r}$ aus den Definitionen bzw. Forderungen von Cleary und Darlington (Fall III), erhält man die Konzeption von Thorndike in korrelativer Darstellung. Für Darlington dagegen scheint Thorndike's Mittelstellung wohl mehr nur einen faulen Kompromiß darzustellen, wenn er schreibt:

"... Definition 2 represents a compromise position midway between these two ... the value of r_{xz} specified is the geometric mean of these values specified by definitions 1 and 3. As in so many cases, however, a compromise may end up satisfying nobody; psychometricans are not in the habit of agreeing on important definitions or theorems by compromise." (Darlington, 1971, S. 76 f.)

Wir werden weiter unten zeigen, daß das Thorndikesche Konzept doch fundierter ist als es bisher von Darlington, Linn und Flaugher gesehen wurde.

Setzt sich die Fähigkeit, im Kriterium erfolgreich zu sein, aus Teilfähigkeiten zusammen, kann es u. U. zu einem *unvollständigen Kriteriumssampling* kommen: Der Test mißt das Kriterium nur unvollständig.

Benötigt man zur Bewältigung des Kriteriums (z. B. erfolgreicher Besuch des Mathematikunterrichts) die Fähigkeiten A und B und unterscheiden sich zwei Gruppen nur in A, so ist ein Test, der A mißt (X_A) nach Cleary ein fairer Test ($r_{yz.x} = 0.0$), jedoch nach Thorndike und Darlington's Fall III ein unfairer Test ($r_{xz} \neq r_{yz}$) bzw. $r_{xz.y} \neq 0$. Unterscheiden sich die beiden Gruppen sowohl in A und B, so ist der Test, der A mißt (X_A) nach Cleary jetzt zu einem unfairen Test ($r_{yz.x} \neq 0$) geworden. Nach Darlington's Fall III ist der Test aber dann fair, wenn die Mittelwertsdifferenz auf X_A (im Test gemessen) in z-Werten so groß ist, wie die Differenz auf der nicht erfaßten Variablen X_B ($r_{xz.y} = 0.0$). Ist die Differenz auf X_B (im Test nicht erfaßt) kleiner als die Differenz auf X_A (im Test erfaßt), ist der Test auch nach Darlington's Fall III unfair und eventuell fair nach Thorndike. Ist dagegen die Differenz auf X_B größer als die Differenz auf X_A (was besonders bei der Konstruktion von sogenannten ‚culture-fair'-Tests vorkommt), ist der Test nach allen drei Definitionen unfair gegenüber der besseren Testwertgruppe.

Man sieht also, daß die Fairness des Tests nicht nur von den Gruppendifferenzen auf X und Y abhängt, sondern auch von den Differenzen auf den nicht erfaßten – für das Kriterium aber wichtigen – Prädiktoren.

Hunter & Schmidt (1976) gaben der nicht verfaßten Variablen X_B, die das Abschneiden im Kriterium wesentlich mitbeeinflußt, eine andere Deutung, die auf nichtkognitive Einflüsse abzielt:

"... Thus, there will inevitably be a large random component due to the situational factors in performance that is not measurement error, but that has the same effect as measurement error in that it sets an upper bound on the validity of any predictor test battery even if it includes biographical information on family income, stability, and the like. ... Because on the average blacks are assumed to be unlucky as well as lower in ability, the racial difference in college achievement in this model will be greater than the predicted by ability alone, and hence the regression lines of college performance compared with ability will not be equal. The black regression line will be lower. Thus according to Cleary, the test is biased against whites. According to Thorndike, the test may be approximately fair (perhaps slightly biased against blacks). According to Darlington, the test could be either fair or unfair: If the racial difference on luck were about the same in magnitude as the race difference on the ability test, then the test would be fair; but if the race difference on luck were less than the difference on ability, then the test would be unfair to blacks. That is, the Darlington assessment of the fairness of the test would not depend on the validity of the test in assessing ability, but on the relative harshness of the personal-ecconomic factors determining the amount of luck accorded the two groups." (Hunter & Schmidt, 1976, S. 1063)

3.3 Die Beziehungen zwischen Fairnesskonzepten und Kriterium-Test-Regressionen

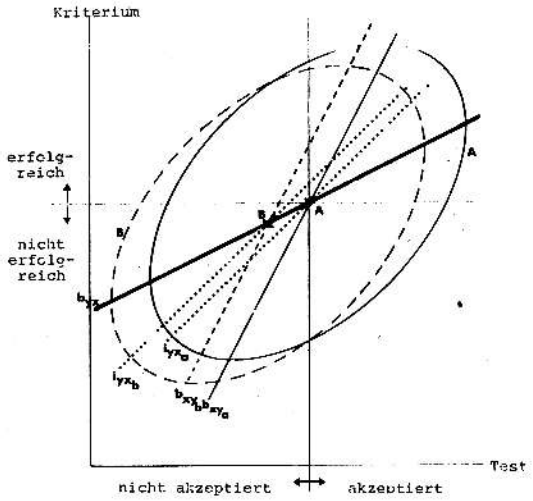
Wir betrachten dabei nur die Fälle, in denen auf Kriterium und Prädiktor nur jeweils gleiche gruppenspezifische cutoffs $Y = y^*$ bzw. $X = x^*$ gelten; d. h. Entscheidungen über Personen gelten für einen bestimmten Testwert für alle Personen ohne Ansehen der Gruppenzugehörigkeit.

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 4.

Auch Thorndike's Definition läßt eine regressionsorientierte Deutung zu, die bisher aber noch nicht entdeckt worden ist. 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 weist die orthogonale Regression enge Beziehungen zur Hauptkomponentenanalyse auf (s. Anhang A und Figur 5 und 12).

Figur 4

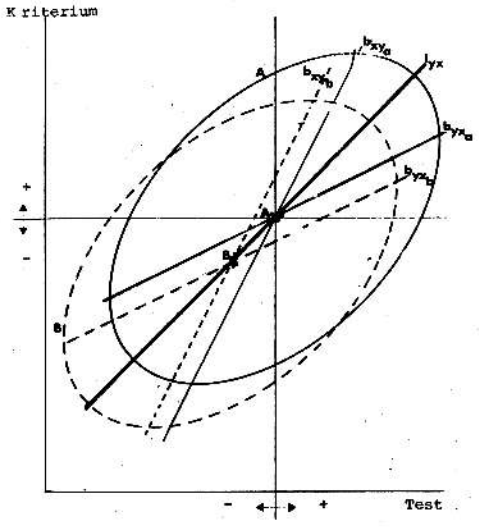
Situation, in der nach Cleary derselbe cut-off im Test für beide Gruppen A und B fair ist: Die Regressionen von Y auf X sind für beide Gruppen gleich



A, B = Centroide der Gruppen A und B im Kriteriums-Prädiktorraum; b_{YX_a} = Regression der Gruppe A mit Y als Kriterium und X als Prädiktor; i_{YX_a} = interdependente oder orthogonale Regression in Gruppe A.

Figur 5

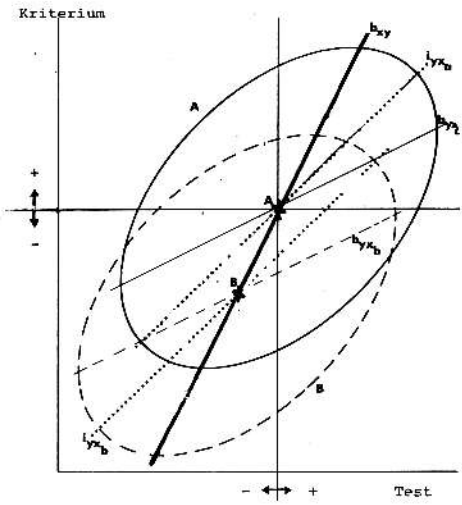
Situation, in der nach Thorndike derselbe cut-off im Test für beide Gruppen A und B fair ist: Die interdependenten oder orthognalen Regressionen sind für beide Gruppen A und B gleich



Die Definition von Darlington's Fall III (Spezialfall von Cole'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 6).

Figur 6

Situation, in der nach Darlington's 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



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 4. diskutieren und bewerten, wollen wir auf der vierten Betrachtungsebene die Selektionscharakteristikkurven (SCK) der verschiedenen Konzepte betrachten.

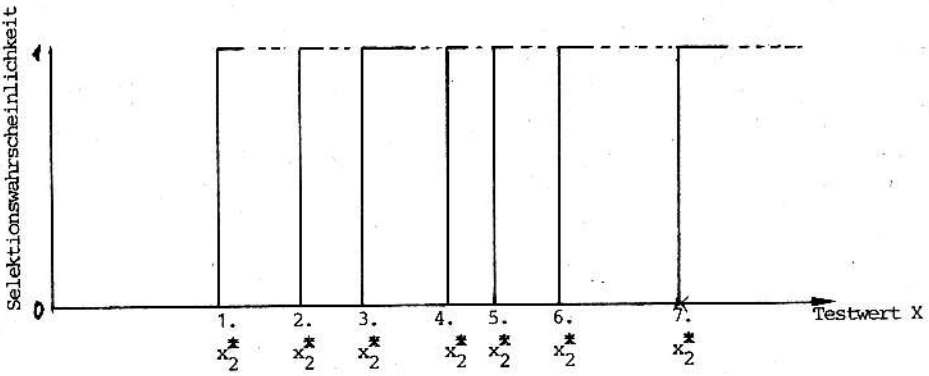
3.4 Selektionscharakteristikkurven und Fairnesskonzepte

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 Applikanten 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_1^*, x_2^* zu, können die cutoffs der deprivierten Gruppe je nach Fairnesskonzept in eine Rangreihe gebracht werden (s. a. Figur 7). Der cutoff x_2^* 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.

Figur 7

Typische Selektionscharakteristikkurven für die verschiedenen Fairnessdefinitionen



Die Figur stellt eine **mögliche** Rangreihe von cutoffs für eine unterpriviligierte Gruppe 2 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.

4. 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 wir die jeweiligen Modellannahmen und Fehlertheorien untersuchen. Sind im Regressionsmodell Spezifikations- („errors in equations"; „shock model") und Meßfehler („errors in variables") berücksichtigt (vgl. a. Griliches, 1974)? Zur Untersuchung dieser beiden Fragen wollen wir uns die Annahmen des klassischen Regres-

sionsmodells[4], auf der Cleary's Definition basiert, in Erinnerung rufen (vgl. a. Schönefeld, 1969, S. 53 ff.):

Für das klassische lineare Regressionsmodell der Matrixgleichung

(4.1) $y = X\beta + \epsilon$

mit y, ϵ als Zufallsvektoren, X als Beobachungswertmatrix und β als Regressionsparametervektor sind folgende Annahmen zu treffen:

A1: Die Wahrscheinlichkeitsverteilung des Zufallsvektors ϵ hat für die gegebene Beobachtungswertmatrix X und alle a priori zulässigen Parameterwerte β folgende Eigenschaften:

A1.1 $E(\epsilon) = 0$

A1.2 $E(\epsilon\epsilon') = \sigma^2 1$ die Kovarianzmatrix der Fehlervariablen ist eine Skalenmatrix.

A2: Die Beobachtungswertmatrix ist exogen determiniert und fest vorgegeben. Sie hat einen Rang, der ihrer Spaltenzahl entspricht (Zeilenzahl = Personenzahl).

A3: Über die festliegenden Werte der Regressionsparameter ist keine andere als die in der Beobachtungswertmatrix (yX) enthaltene Information verfügbar.

A4: Die Größe X ist *exakt* beobachtbar und enthält daher *keine* Meß- oder Beobachtungsfehler.

Die Annahmen können gelockert werden (s. a. die fünf Modelle von Press, 1972, S. 187–282). Am meisten wird den Psychologen A4 stören müssen, weil die mangelnde Reliabilität der psychologischen Tests in direktem Widerspruch zu dieser Forderung steht. Ignoriert man die Meßfehlerbelastetheit der Prädiktoren, können die Parameter im klassischen Regressionsmodell nicht mehr konsistent geschätzt werden (Johnston, 1963, S. 148 ff.).

Im folgenden Abschnitt soll der Effekt der Meßfehlerbelastung im bivariaten Fall kurz demonstriert werden. Der deskriptiven Regression

(4.2) $Y = b_0 + b_1 X + e$

steht die Linearbeziehung zwischen den Konstrukten oder latenten Variablen gegenüber:

(4.3) $\tau_y = \beta_0 + \beta_1 \tau_x + \epsilon_\tau$

Fällt der Fehlerterm weg, liegt eine funktionale Beziehung zwischen den latenten Variablen vor. Die Meßwerte denken wir uns zusammengesetzt aus latenten „wahren" Anteilen und Meßfehlern, die z. T. auf nicht gelungene Operationalisierungen zurückzuführen sind:

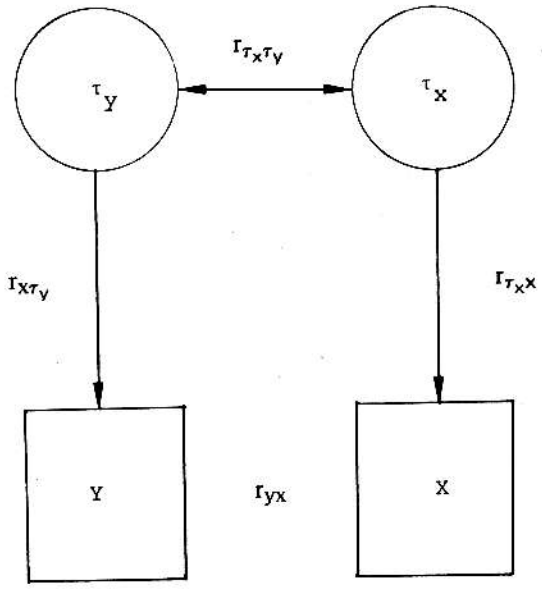
(4.4) $Y = \tau_y + \epsilon_y \qquad X = \tau_x + \epsilon_x$

Die Hoffnung, die Meßfehlervarianz des Prädiktors ungestraft ignorieren zu können, wird schon im bivariaten Fall enttäuscht werden müssen (vgl. a. Figur 8). Der Regressionskoeffizient der deskriptiven Regression b_1 ist bei nichtperfekter Reliabilität des Prädiktors X immer kleiner als der Regressionsparameter der „wahren" latenten Beziehung:

(4.5) $$b_1 = r_{xy}\frac{\sigma_y}{\sigma_x} = (r_{x\tau_x} r_{\tau_x\tau_y} r_{\tau_y y})\frac{\sigma_y}{\sigma_x} = \left(\frac{\sigma_{\tau_x}}{\sigma_x} r_{\tau_x\tau_y} \frac{\sigma_{\tau_y}}{\sigma_y}\right)\frac{\sigma_y}{\sigma_x}$$

4 Press (1972) unterscheidet fünf Regressionsmodelle: 1. Functional Regression, 2. Conditional Regression, 3. Stochastic Regression, 4. Errors in the Variables, 5. Random Parameters.

Figur 8
Beziehungen zwischen Test X und Kriterium Y über die Konstrukte τ_X und τ_Y



Erweitern mit $(\sigma_{\tau_X}/\sigma_{\tau_X})$ und zusammenfassen:

$$\frac{\sigma^2_{\tau_X}}{\sigma^2_X} r_{\tau_X\tau_Y} \frac{\sigma_{\tau_Y}}{\sigma_{\tau_X}} = \beta_1 \rho_{XX'} \qquad \text{(wobei: } \rho_{XX'} \text{ = Reliabilität von X)} \tag{4.6}$$

Die Regressionskonstante der deskriptiven Regression b_0 wird umso größer, je unreliabler der Prädiktor ist:

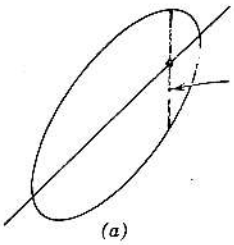
$$b_0 = \mu_Y - b_1\mu_X = \mu_{\tau_Y} - b_1\mu_{\tau_X} = \mu_{\tau_Y} - \rho_{XX'}\beta_1\mu_{\tau_X} \tag{4.7}$$

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_X 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 beobachtbaren Testwerte um ihre wahren Werte. Liegt die Standardabweichung der beobachtbaren 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 Applikanten.

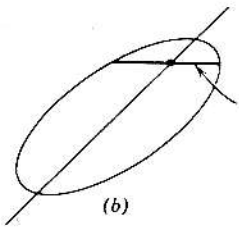
Die den Fairnesskonzepten unterliegenden drei Regressionskonzepte unterscheiden sich nur bezüglich ihrer Fehlerannahmen und sind ineinander überführbar. Geht man von einer „wahren" linearen Beziehung zwischen den latenten Variablen τ_X und τ_Y aus, kann man je nach Fehlerannahmen verschiedene bivariate Datensätze erzeugen (Figur 9). Umgekehrt kann man von *einem* Datensatz je nach Fehlerannahmen auf verschiedene „wahre" (hier: lineare) Zusammenhänge zwischen τ_X und τ_Y schließen.

Figur 9

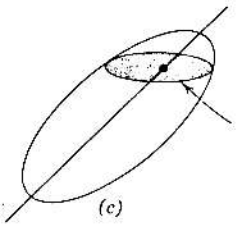
Eine einzige „wahre" Beziehung $\tau_Y = \alpha + \beta\tau_X$ kann je nach Fehlerannahme unterschiedliche Rohwertverteilungen erzeugen:



a: Y ist fehlerhaft:
Alle Personen auf der senkrechten Linie besitzen dieselben wahren Werte τ_Y und τ_X.



b: X ist fehlerhaft:
Alle Personen auf der waagrechten Linie besitzen dieselben wahren Werte τ_Y und τ_X.



c: Y und X sind fehlerhaft:
Alle Personen innerhalb der Ellipse können dieselben wahren Werte τ_Y und τ_X besitzen.

Nehmen wir jetzt den stochastischen Zusammenhang (4.3) mit meßfehlerbehafteten manifesten Variablen (4.4) an. Die Störgrößen ϵ_X, ϵ_Y, ϵ_τ seien voneinander unabhängig verteilt mit Varianzen $\sigma^2_{\epsilon_X}$, $\sigma^2_{\epsilon_Y}$, $\sigma^2_{\epsilon_\tau}$. Der Störterm $(\epsilon_Y + \epsilon_\tau)$ hat ebenfalls Erwartungswert Null und Varianz $\sigma^2_{\epsilon_{Y\tau}} = (\sigma^2_{\epsilon_Y} + \sigma^2_{\epsilon_\tau})$. Das Verhältnis λ_1 soll bekannt sein:

$$\lambda_1 = \sigma^2_{\epsilon_X} / (\sigma^2_{\epsilon} + \sigma^2_{\epsilon_\tau}) = \sigma^2_{\epsilon_X} / \sigma^2_{\epsilon_{\tau Y}} \tag{4.8}$$

Setzen wir den wahren Wert für Y aus (4.4) in (4.3) ein, haben wir den Kriteriumswert Y in latente Komponenten zerlegt:

$$Y = \beta_0 + \beta_1 \tau_X + \epsilon_\tau + \epsilon_Y \tag{4.9}$$

Setzen wir zusätzlich X in (4.9) ein, haben wir e aus (4.2) zerlegt:

(4.10) $$Y = \beta_0 + \beta_1 X + (\epsilon_\tau + \epsilon_y - \beta_1 \epsilon_x)$$

Danach bilden wir cov (X, Y) und cov (Y, Y) = var (Y):

(4.11a) $$\text{cov}(X, Y) = \beta_1 \text{var}(X) + \text{cov}(\tau_x + \epsilon_x, \epsilon_\tau + \epsilon_y - \beta_1 \epsilon_x)$$

(4.11b) $$\text{cov}(Y, Y) = \beta_1 \text{cov}(X, Y) + \text{cov}(\beta_1 \tau_x + \epsilon_\tau + \epsilon_y, \epsilon_\tau + \epsilon_y - \beta_1 \epsilon_x)$$

und erhalten zwei Gleichungen mit drei Unbekannten, von denen wir eine wegen (4.8) eliminieren können:

(4.12a) $$m_{xy} \triangleq \beta_1 m_{xx} - \beta_1 \sigma^2_{\epsilon_x} = \beta_1 m_{xx} - \beta_1 \sigma^2_{\epsilon_{\tau y}} \lambda_1$$

(4.12b) $$m_{yy} \triangleq \beta_1 m_{xy} + \sigma^2_{\epsilon_{\tau y}} = \beta_1 m_{xy} + 1\sigma^2_{\epsilon_{\tau y}} \lambda_2$$

mit $\lambda_2 = 1 = (\sigma^2_{\epsilon_{\tau y}} / \sigma^2_{\epsilon_{\tau y}})$ und $m_{xy} = \text{cov}(X, Y)$ sowie $m_{xx} = \text{var}(x)$.

(4.12a) und (4.12b) lassen sich kompakt als Matrixgleichung schreiben:

(4.13) $$(\mathbf{M} - \sigma^2_{\epsilon_{\tau y}} \cdot \Lambda_1)\ (\hat{\beta}) = (\mathbf{0})$$

bzw.

$$(\mathbf{M} - \Lambda_2)\ (\hat{\beta}) = (\mathbf{0})$$

mit
- $\mathbf{M}$ = Matrix der zentralen Produkt-Momente (= Varianz-Kovarianz-matrix)
- $\hat{\beta}'$ = $(\hat{\beta}_1, -1)$ = Regressionsparametervektor mit $\hat{\beta}_2 = -1$
  $\hat{\beta}_0$ wird geschätzt: $\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$
- Λ_1 = diag (λ_1, λ_2)
- Λ_2 = diag $(\sigma^2_{\epsilon_x}, \sigma^2_{\epsilon_{\tau y}})$ } Diagonalmatrizen
- $\mathbf{0}$ = Nullvektor

An (4.12) und (4.13) lassen sich die den Fairnessdefinitionen zugrunde liegenden Regressionskonzepte verdeutlichen:

a) bei $\sigma^2_{\epsilon_x} = 0$ (Meßfehlerfreiheit des Prädiktors) genügt (4.12a), denn $\lambda_1 = 0$.

(4.14) $$\hat{\beta}_1 = \frac{m_{xy}}{m_{xx}}$$ Der Parameter wird durch die klassische Regression von Y auf X geschätzt.

Cleary's Regressionskonzept beruht auf der Vorstellung eines perfekt reliablen Prädiktors. Ist dagegen die Meßfehlervarianz von X bekannt, kann man aus $(M - \Lambda_2) \cdot \hat{\beta} = 0$ in (4.13) die Schätzung

(4.15) $$\hat{\beta}_1 = \frac{m_{xy}}{m_{xx} - \sigma^2_{\epsilon_x}}$$ herleiten.

b) bei $\sigma^2_{\epsilon_{\tau y}} = 0$ (beim stochastischen Zusammenhang der wahren Werte) oder bei $\sigma^2_{\epsilon_y} = 0$ (beim funktionalen Zusammenhang) genügt (4.12b) zur Schätzung von β_1. λ_1 ist jetzt ∞.

(4.16) $$\hat{\beta}_1 = \frac{m_{yy}}{m_{xy}}$$ Der Parameter wird durch die klassische Regression von X auf Y geschätzt.

Darlington's Fall III beruht auf der Vorstellung eines störtermfreien Kriteriums ($\rho_{yy'} = 1$) und perfekter Vorhersagbarkeit von Y bei Kenntnis der τ_X.

c) bei $\sigma^2_{\epsilon_X} = \sigma^2_{\epsilon_{\tau Y}}$ ist $\lambda_1 = 1$ und damit $\Lambda_1 = I$ (I = Einheitsmatrix) bzw. Λ_2 eine Skalarmatrix, deren unbekannter Skalarfaktor sich aus der Matrix, die dann gleich I wird, herausziehen läßt. Normieren wir den Parametervektor $\beta' = (\beta_1, -1)'$ in (4.13) auf $\beta'\beta = 1$ und setzen $\sigma^2_{\epsilon_{\tau Y}} = \mu$, erhalten wir das gleiche Eigenwertproblem wie bei der orthogonalen Regression oder der Hauptkomponentenanalyse (s. a. Anhang A1):

$$(4.17) \qquad (\mathbf{M} - \mu \cdot \mathbf{I})\ (\hat{\beta}) = (\mathbf{0}) \quad \text{mit } \mu = \min!$$

Wir können jetzt Thorndike's implizites Regressionskonzept näher erklären: Ist der Zusammenhang zwischen τ_Y und τ_X funktional und sind die Meßfehlervariablen $\sigma^2_{\epsilon_X} = \sigma^2_{\epsilon_Y}$ oder ist der Zusammenhang zwischen den latenten Variablen stochastisch und sind die Störvarianzen $\sigma^2_{\epsilon_X} = \sigma^2_{\epsilon_{\tau Y}}$, entspricht Thorndike's implizites Regressionskonzept der orthogonalen Regression. Wir haben diese Regression auch interdependente Regression genannt, weil sie Prediktor und Kriterium gleichwertig behandelt: Die Regressionslinie der orthogonalen Regression bleibt von der Wahl des Kriteriums in ihrer Lage im Variablenraum unberührt. Die Regressionslinie ist dabei identisch mit der ersten Hauptkomponente, während die Normale der Regressionslinie der zweiten Hauptkomponente entspricht. Soll wie im klassischen Fall auch hier das Kriterium den Regressionskoeffizienten 1 (vgl. (4.10)) bekommen, sind alle Komponenten des zweiten Eigenvektors der Varianz/Kovarianzmatrix durch die Komponente des Kriteriums zu dividieren. Für den Fall zweier standardisierter Variabler beträgt die so normierte Steigung $\beta_1 = 1$ (also die 45°-Gerade). Der Regressionsvergleich beschränkt sich dann auf den Vergleich der Konstanten: Sind die populationsspezifischen Regressionen parallel versetzt?

Eine gruppenspezifische Standardisierung der Prädiktor- und Kriteriumsvariablen (z-Werte) würde in jedem Falle zu einem fairen Test im Sinne von Thorndike für K = 1 führen. So bräuchte man bei Verwendung von IQ-Werten nur noch die Kriterienwerte für jede Altersgruppe auf z-Werte zu transformieren, um einen nach Thorndike fairen Test zu erhalten (für K = 1). Die nachträgliche statistische Ausschaltung gruppenspezifischer Differenzen in den Varianzen und Mittelwerten der Prädiktor- und/oder Kriteriumsvariablen ist unseres Erachtens jedoch unzulässig, wenn gerade die Berücksichtigung der faktischen Ungleichheit der Gruppen entscheidend für eine faire Prognose ist. So führt z. B. die Verwendung des IQ-Maßes, was ja einer Standardisierung des Prädiktors pro Altersgruppe gleichkommt, bei verschiedenen Kriteriumsmittelwerten für die Altersgruppen nach Cleary, Thorndike und Darlington III zu einer *unfairen* Testanwendung: Die altersspezifischen Regressionslinien können nicht zur Deckung gebracht werden! Somit bringt die Einführung des IQ-Maßes einen Effekt, den man durch den IQ gerade abwenden wollte.

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 Reliabilität sondern nur *eine* Validität und *eine* Reliabilität.

5. Neue Vorschläge

5.1 Multiple und multivariate Erweiterung der klassischen regressions-orientierten Fairnesskonzepte

Für den Fall mehrerer Prädiktoren und/oder Kriterien lassen sich die bisher diskutierten Konzepte verallgemeinern. Hierzu griffen Möbus & Simons (1975) z. T. auf in der Ökonometrie entwickelte Vorstellungen zurück: a) Mehrere meßfehlerfreie Prädiktoren stellen eine multiple Erweiterung des Cleary-Konzeptes dar und führen methodisch auf die multiple Regression; b) gleiche Meßfehlervarianzen aller Variabler und funktionale Beziehung zwischen den latenten Werten führen auf die orthogonale Regression (s. Anhang A) oder das Fehler-in-den-Variablen-Modell (Anhang B), während c) ungleiche Meßfehlervarianzen in diesem Fall immer in das Fehler-in-den-Variablen-Modell münden (s. Anhang B und Schönfeld, 1971, II, S. 115 ff.); kommt d) noch eine stochastische Beziehung zwischen den latenten Variablen hinzu, müssen wir das kombinierte „Fehler-in-den-Variablen- & Fehler-in-der-Gleichung-Modell heranziehen (s. a. Anhang C und Van de Geer, 1971, S. 234, S. 272). Diese Modelle weisen für die empirische Arbeit des Psychologen eine Reihe von Nachteilen auf: erhöhte Anzahl von Annahmen über die Fehler; nicht voll befriedigend ausgebaute statistische Theorie zur Hypothesenprüfung.

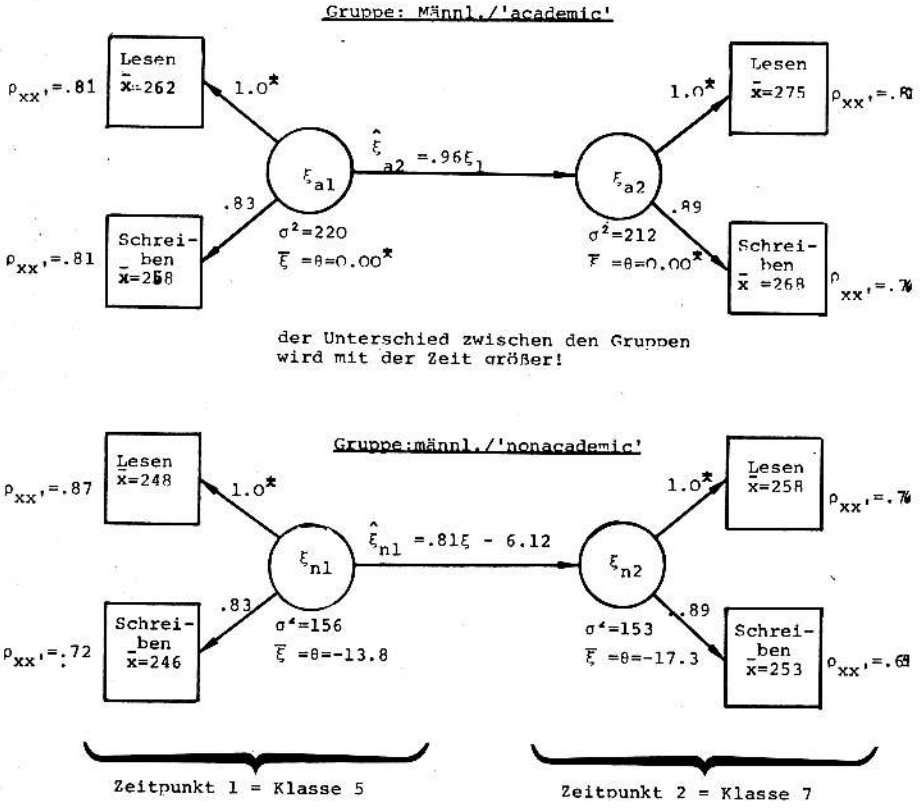
Einen Ausweg scheint die in der Soziologie und Psychologie entwickelte Vorstellung multipler Indikatoren für latente Variablen zu bieten: Nicht *eine* sondern mehrere manifeste Variablen sind Indikatoren für jeweils *ein* Konstrukt. Wir verlagern damit die Betrachtung der gruppenspezifischen Beziehungen zwischen manifesten Variablen (Test vs. Kriterium) auf die Analyse gruppenspezifischer Prognosesysteme im Raum der latenten Variablen (bzw. Konstrukte). Dieses Vorgehen entspricht unserer Überzeugung, daß nicht die Kenntnis von **Testwerten** sondern die geschätzten Fähigkeitswerte (= Werte der **latenten** Variablen) Basis von Entscheidungen sein sollten.

Ermöglicht wird der statistische Vergleich von gruppenspezifischen Mittelwerten, Regressionen und Veränderungen im Konstruktraum durch eine Verknüpfung von la-

tenten Strukturmodellen mit klassischen parametrischen Testverfahren (s. Sörbom, 1974; 1976). Unsere Überlegungen sind am Beispiel der ETS-Growth-Study „ETS longitudinal Study of Academic Prediction and Growth" von Anderson & Maier (1963) mit zwei Gruppen, zwei Meßzeitpunkten und jeweils zwei Indikatoren für eine latente Variable dargestellt. Indikatoren stellen die Formen des Sequential Test of Educational Progress (STEP) „Lesen" und „Schreiben" für Klasse 5 und 7 (= Meßzeitpunkte) und die Gruppen „academic curriculum" bzw. „nonacademic curriculum" dar.

Das Modell ist graphisch in Figur 10 dargestellt. Die für die Fairnessdiskussion relevanten Ergebnisse finden sich in Figur 11.

Figur 10
Pfadmodell für die ETS Growth Study
zur vergleichenden Analyse von Veränderungen der „Verbalfähigkeit"



Die mit * bezeichneten Parameter sind während der Schätzung festgehalten worden (restricted parameters).

In Klasse 5 werden die Konstrukte ξ_{a1}, ξ_{n2} durch die Indikatoren x_{a1}, x_{a2} für die „academic"-Gruppe und x_{n1}, x_{n2} für die „nonacademic"-Gruppe erfaßt. Für die Konstrukte der zweiten Welle ξ_{a2}, ξ_{n2} gilt das entsprechende. Das Meßmodell für eine Gruppe ist:

(5.1.1)
$$\begin{aligned} x_1 &= \mu_1 + \lambda_1\xi_1 + \epsilon_1 \\ x_2 &= \mu_2 + \lambda_2\xi_1 + \epsilon_2 \\ y_1 &= \mu_3 + \lambda_3\xi_2 + \delta_1 \\ y_2 &= \mu_4 + \lambda_4\xi_2 + \delta_2 \end{aligned}$$

Meßmodell für Klasse 5

Meßmodell für Klasse 7

oder in Matrixdarstellung:

(5.1.2)
$$z = \mu + \Lambda\,\xi + \gamma$$

mit: $z' = (x_1, x_2, y_1, y_2)$

$\mu' = (\mu_1, \mu_2, \mu_3, \mu_4)$

$\gamma' = (\epsilon_1, \epsilon_2, \delta_1, \delta_2)$

$\xi' = (\xi_1, \xi_2)$

$$\Lambda = \begin{bmatrix} 1 & 0 \\ \lambda_2 & 0 \\ 0 & 1 \\ 0 & \lambda_4 \end{bmatrix}$$

μ enthält die Lokalisationsparameter, Λ die Skalierungsparameter (= unnormierte „Ladungen") und γ die Meßfehlervariablen. Durch die Normierung von $\lambda_1 = \lambda_3 = 1$ werden die Maßstäbe von ξ_1, ξ_2 festgelegt. Die Erwartungswerte der Indikatoren bestimmen sich nach (5.1.3)

(5.1.3)
$$E(x_i) = \mu_i + \lambda_i\theta_1 \quad \text{mit} \quad \theta_1 = E(\xi_1)$$
$$E(y_j) = \mu_j + \lambda_j\theta_2 \quad \text{mit} \quad \theta_2 = E(\xi_2)$$

oder in Matrixdarstellung:

(5.1.4)
$$E(z) = \mu + \Lambda\theta$$

Die Varianz/Kovarianzmatrix der Indikatoren ist unter der Annahme der wechselseitigen Unabhängigkeit der Fehler und der Unabhängigkeit der Fehler von ξ darstellbar als (s. 5.1.5):

(5.1.5)
$$\Sigma = \Lambda\,\Phi\,\Lambda' + \Psi$$

wobei:

$$\Sigma = \begin{bmatrix} \sigma^2_{\xi_1} + \sigma^2_{\epsilon_1} & \cdot & \cdot & \cdot \\ \lambda_2\sigma^2_{\xi_1} & \lambda_2^2\sigma^2_{\xi_1} + \sigma^2_{\epsilon_2} & \cdot & \cdot \\ \sigma_{\xi_1\xi_2} & \lambda_2\sigma_{\xi_1\xi_2} & \sigma^2_{\xi_2} + \sigma^2_{\delta_1} & \cdot \\ \lambda_4\sigma_{\xi_1\xi_2} & \lambda_4\lambda_2\sigma_{\xi_1\xi_2} & \lambda_4\sigma^2_{\xi_2} & \lambda_4^2\sigma^2_{\xi_2} + \sigma^2_{\delta_2} \end{bmatrix}$$

$$\Phi = \begin{bmatrix} \sigma^2_{\xi_1} & \sigma_{\xi_1\xi_2} \\ \sigma_{\xi_1\xi_2} & \sigma^2_{\xi_2} \end{bmatrix}$$

Kovarianzmatrix der Konstrukte

$$\Psi = \begin{bmatrix} \sigma^2_{\epsilon_1} & & & \\ & \sigma^2_{\epsilon_2} & & \\ & & \sigma^2_{\delta_1} & \\ & & & \sigma^2_{\delta_2} \end{bmatrix}$$

Varianz/Kovarianzmatrix der Meßfehler

Die Lokalisationsparameter und die Konstruktmittelwerte sind ohne weitere Restriktion nicht schätzbar. Es ist jedoch möglich, Mittelwertsdifferenzen der Konstrukte zu ermitteln, wenn die gleichen Indikatoren mindestens zwei Gruppen vorgelegt werden:

(5.1.6) $z^i = \mu + \Lambda\xi^i + \gamma^i$ Meßmodell für Gruppe i

(5.1.7) $\Sigma^i = \Lambda\Phi^i\Lambda' + \Psi^i$ Kovarianzmatrix der Indikatoren für Gruppe i

(5.1.8) $E(z^i) = \mu + \Lambda\theta^i$ Erwartungswertvektor der Indikatoren für Gruppe i

Dabei sollen μ und Λ für alle Gruppen gelten. Die Unbestimmtheit in (5.1.8) kann beseitigt werden durch Nullsetzung der Konstruktmittelwerte der Gruppe i: $\theta^i = 0$. Gruppe i kann dann als Referenzgruppe angesehen werden. Es ist jetzt möglich, Unterschiede zwischen den Gruppen in den Kovarianzmatrizen und den Konstruktmittelwerten zu analysieren.

Analog zu den klassischen Fairnesskonzepten (außer der Identitätsdefinition) interessieren in diesem Zusammenhang aber nur die Unterschiede in den Regressionen der latenten Variablen. Sind die Regressionen der latenten Posttestkonstrukte auf die latenten Prätestkonstrukte für die Gruppen identisch? Gilt (5.1.9) oder (5.1.10)?

(5.1.9) $\xi_2^i = \alpha^i + \beta^i\xi_1^i + \zeta^i$ $(i = 1, ...)$

(5.1.10) $\xi_2^i = \alpha + \beta\,\xi_1^i + \zeta^i$ wobei: ζ = Gleichungsfehler

Die Regressionsparameter der latenten Variablen lassen sich nach (5.1.11) berechnen:

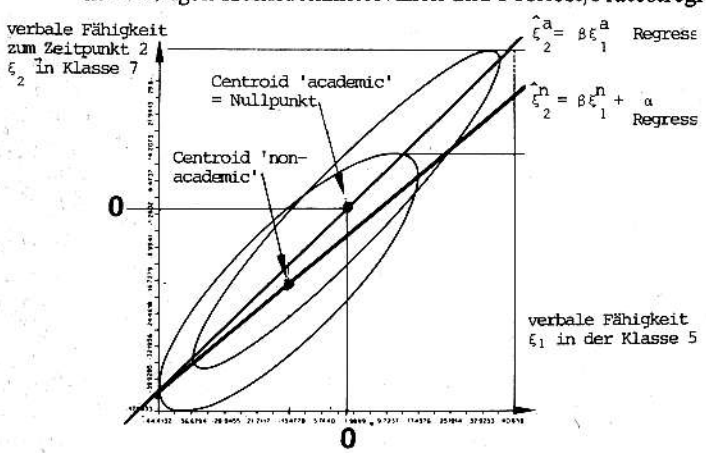
(5.1.11) $\beta^i = \sigma^i_{\xi_1,\xi_2} / \sigma^{2\,i}_{\xi_1}$

$\alpha^i = \theta_2^i - \beta\theta_1^i$

Die Maximum-Likelihoodschätzungen mit dem Programm COFAMM von Sörbom & Jöreskog (1976) finden sich in Tabelle 1 und Figur 11.

Figur 11

Raum der latenten Variablen „verbale Fähigkeit" in Klasse 5 und 7
bei amerikanischen Jungen mit ‚academic' und ‚nonacademic' Curriculum
mit 95%igen Konfidenzintervallen und Posttest/Prätestregressionen



Die Regressionen sind für die beiden Gruppen:

„Jungen mit akademischem Curriculum" $\xi_{a2} = .9636\xi_{a1}$

„Jungen mit nonakademischem Curriculum" $\xi_{n2} \quad .8116\xi_{n1} - 6.1174$

Tabelle 1

Daten und Maximum-Likelihood Schätzungen eines Pfadmodells zur Prognose der verbalen Fähigkeit in Klasse 7 auf der Basis von Testergebnissen aus Klasse 5 (s. Sörbom, 1976)

Mittelwerte für STEP-Untertests „Lesen" und „Schreiben"			
		Jungen ‚academic'	‚nonacademic'
STEP „Lesen"	Klasse 5	262.2	248.6
STEP „Schreiben"	Klasse 5	258.8	246.9
STEP „Lesen"	Klasse 7	275.6	258.5
STEP „Schreiben"	Klasse 7	269.1	253.3

Kovarianzmatrix für Jungen ‚academic'					
STEP „Lesen"	Klasse 5	281.3			
STEP „Schreiben"	Klasse 5	184.2	182.8		
STEP „Lesen"	Klasse 7	216.7	171.7	283.3	
STEP „Schreiben"	Klasse 7	198.4	153.2	208.8	246.1

Kovarianzmatrix für Jungen ‚nonacademic'					
STEP „Lesen"	Klasse 5	174.5			
STEP „Schreiben"	Klasse 5	134.5	161.9		
STEP „Lesen"	Klasse 7	129.8	118.8	228.4	
STEP „Schreiben"	Klasse 7	102.2	97.7	136.1	180.5

	Lokalisationsparameter		Matrix Λ		Diagonale der Matrix Ψ	
		μ	ξ_1	ξ_2	‚academic'	‚nonacademic'
STEP „L"	Kl. 5	262.4	1.00	0.00	50.1	23.2
STEP „S"	Kl. 5	258.7	0.84	0.00	36.5	42.8
STEP „L"	Kl. 7	275.7	0.00	1.00	51.7	65.7
STEP „S"	Kl. 7	268.9	0.00	0.89	57.8	67.3

	Kovarianzmatrix der latenten Variablen Φ				Mittelwerte der latenten Variablen θ	
	‚academic'		‚nonacademic'		‚academic'	‚nonacademic'
VF 5	220.0		156.5		0.00	−13.8
VF-7	212.1	233.6	126.9	153.7	0.00	−17.3

VF 5 = Verbale Fähigkeit in Klasse 5
VF 7 = Verbale Fähigkeit in Klasse 7

Da die Regressionen signifikant voneinander verschieden sind, ist ein Prädiktorsystem, das ein für alle Gruppen gemeinsames nomologisches Netz der Konstrukte annimmet, unfair. Das bedeutet mit anderen Worten, daß die Konsequenzen, die sich für Personen mit denselben Fähigkeitsparametern aber unterschiedlichen Gruppenzugehörigkeiten verbinden, gruppenspezifisch sind und daher verschieden sein können, wenn der minimal zufriedenstellende Konstruktwert (cutoff auf ξ_2) für alle Gruppen gleich ist.

5.2 Eine Fairnessdefinition

Wir wollen hier wie Möbus (1976) zwischen Item- und Systembias unterscheiden und folgende Definition übernehmen (s. Möbus & Simons, 1975).

„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.

5.3 Ein probabilistisches Validitäts- und Fairnesskonzept mit multiplen Kriterien

Bei der praktischen Testkonstruktion muß zuerst der Itembias beseitigt werden. Hierzu sind eine Reihe von Methoden vorgeschlagen worden (s. a. Rudner, 1977). Am befriedigendsten sind jedoch die stichprobenunabhängigen logistischen Testmodelle zur Beseitigung des Itembias. Sie lassen sich jedoch ebenfalls bei der praktischen Konstruktion eines fairen Tests verwenden. Wir haben die notwendigen Schritte hierzu angegeben (Möbus & Simons, 1975):

a) Validierungs- und Konstruktionsphase

a1) Bildung eines Pools von Kriteriumsindikatoren (KI) und Testitems (TI). Die TI sollen die gleiche latente Dimension wie die KI messen, jedoch leichter applizierbar sein, um die Testkonstruktion zu rechtfertigen (der Test substituiert das Kriterium).

a2) Bildung homogener Subpools. Subpools, die nur TI enthalten, sind unbrauchbar. Subpools, die nur KI enthalten, fordern die weitere Suche nach TI. Für die Testkonstruktion sind nur die aus KI und TI vermischten Pools brauchbar.

a3) Schätzung aller Itemparameter der KI und TI an einer Population, die alle relevanten ethnischen und sozialen Gruppen einschließt. Schätzalgorithmen hierzu gibt Fischer (1974) an.

a4) Modelltest der spezifisch objektiven Meßprozedur (s. a. Fischer, 1974, Kap. 15). Läßt sich die Modellgültigkeit nicht zurückweisen, bedeutet die Gleichheit der Itemparameter (struk-

turelle Parameter) in allen relevanten ethnischen und sozialen Gruppen *Testfairness* gegenüber diesen Gruppen. Die Strukturparameter sollen bei den spezifisch objektiven Verfahren aber auch noch in allen anderen Personengruppierungen gleich sein, so daß der Test auch gegenüber diesen Gruppen fair ist, wenn er den Modelltest passiert hat.

a5) Ausschluß aller modellunverträglicher TI und KI. Bleibt kein KI in der Meßprozedur enthalten, hat der Test seine Validität eingebüßt.

b) Prognosephase

b1) Vorgabe der TI an eine zu untersuchende Gruppe von Personen.

b2) Schätzung der Personenparameter (vgl. Fischer, S. 239) bei festgehaltenen TI-Itemparametern (gewonnen in a3). Da der Modelltest die Annahme der Populationsunabhängigkeit (bzw. Stichprobenunabhängigkeit) von Item- und Personenparametern nicht zurückgewiesen hat, müssen die Personenparameterschätzungen auf der Basis der TI annähernd mit denen auf der Basis KI + TI übereinstimmen. Die Übereinstimmung wird dabei durch Stichprobenfehler und die verringerte Anzahl von Rohwertsummen eingeschränkt. Daher sollte die Anzahl der TI gegenüber der Anzahl der KI groß sein.

b3) Die Personenparameter aus b2 werden in die Modellgleichungen bei festgehaltenen KI-Itemparametern (aus a3) eingesetzt und die theoretisch zu erwartenden Lösungs- und Reaktionswahrscheinlichkeiten für die KI berechnet.

b4) Evaluation der Prognosen an Hand der Lösungs- oder Reaktionswahrscheinlichkeiten, die prognostiziert werden.

Dieses Vorgehen könnte dazu führen, daß man für jede Gruppe andere TI einbezieht, die trotz ihrer gruppenspezifischen Relevanz im Sinne der ökologischen Validität (Pawlik, 1975) gemeinsam mit den KI dieselbe latente Dimension messen. Das Resultat wäre dann nicht nur die *faire* Messung einer spezifischen Kompetenz, sondern diese Kompetenz würde auch in einer für die jeweilige Gruppe angemessenen, maßgeschneiderten Weise erfaßt.

Stehen die dem Prädiktor und Kriterium unterliegenden Konstrukte in perfekter Beziehung (Homogenität) zueinander, ist sensu Thorndike Reliabilität, Kriteriumsvalidität, Konstruktvalidität und Fairness identisch.

Tragen wir wünschenswerte Eigenschaften eines fairen Tests zusammen (F1 – F8),[5] sehen wir, daß das Vorgehen nach den logistischen Testmodellen diesen Vorstellungen weit entgegenkommt, wenn die Kriteriums- und Prädiktorkonstrukte homogen sind.

---

5 F1: Zu einem Kriterium wird theoriegeleitet ein Test entwickelt.

F2: Das Kriterium wird durch mehrere Kriteriumsindikatoren (KI) beschrieben.

F3: KI und TI sind homogen; d. h. messen die gleiche latente Dimension (kongenerisch).

F4: Gleiche Testwerte führen in allen relevanten Populationen zu gleichen Vorhersagen; d. h. alle Parameter (Strukturparameter im Sinne von Neyman & Scott), auf denen Vorhersagen beruhen, müssen in allen Populationen gleich sein: gleiche Testwerte führen zu gleichen latenten Werten (= incidentelle Parameter) und diese dann zu gleichen Vorhersagen.

F5: KI und TI gelten als meßfehlerbehaftet.

F6: Es gibt keine unabhängigen Variablen im experimentalpsychologischen Sinne: Prognosen sind richtungsunabhängig (jedenfalls bei konkurrenter Validität im zeitlichen Querschnitt).

F7: Die Vorgabe der TI erlaubt die Messung der den KI als auch den TI unterliegenden latenten Dimension und damit die Prognose des Verhaltens in den KI („Konzept der multiplen Kriterien").

F8: Das gesamte Meßmodell ist empirisch prüfbar.

Die Forderung nach der Integration von Validität und Reliabilität, wie sie in F1 – F7 zum Ausdruck kommt, ist wegen der starken Reliabilitätsorientierung der Testtheorie nicht in ausreichendem Maße beachtet worden. Weit verbreitet ist die Ansicht, ein Test habe nur *eine* Reliabilität, aber *mehrere* Validitäten. Demgegenüber billigt unser Konzept dem Test nur *eine* Validität zu, weil der Test, mit dem wir prognostizieren wollen, *zum* Kriterium und nicht unabhängig davon konstruiert ist. Validität und Reliabilität sind *identische* Begriffe geworden. Darüber hinaus lassen sich bei uns auch die Kriteriums- und die Konstruktvalidität nicht mehr voneinander trennen, da *alle* Indikatoren (TI + KI) *ein* Konstrukt messen.

Stehen dagegen die Konstrukte nicht in perfekter Beziehung zueinander, was wohl in der psychologischen Praxis die Regel ist, kann Systembias eintreten, der ebenfalls ausgeschaltet werden muß. Die Parameter des rekursiven oder nichtrekursiven Prognosesystems müssen gruppenunabhängig sein. Dazu eignet sich (zumindest was den wichtigen rekursiven Fall betrifft) die Maximum-Likelihoodschätzungen von Sörbom (1975). Die Methode wurde in (5.1) ausführlich dargestellt.

5.3 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 die Utility 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 Applikanten, 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 **Prozeß**diagnostik 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 Selektionscharakteristikkurven (SCK), z. B. in der Art von Ogiven oder der logistischen Funktion. Ein möglicher Ansatz wäre die Gleichsetzung von folgenden Wahrscheinlichkeiten:

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

bzw.

$$P(\text{Erfolg} \mid \xi) \stackrel{!}{=} P(\text{Selektion} \mid \xi)$$

wobei: ξ = Fähigkeitsparameter eines Bewerbers

Die bedingten Erfolgswahrscheinlichkeiten können an Hand einer vorangehenden Validierungsstudie berechnet werden.

Allerdings hat die randomisierte Zuweisung den Nachteil einer niedrigeren Gesamttrefferquote. Diese negative Eigenschaft verliert sich mit einer immer höheren Validität des Tests. Ein nicht zu unterschätzender Nachteil der randomisierten Zuweisung besteht in ihrem „Lotterie"-charakter, der sie damit für viele als unakzeptabel erscheinen läßt. Jedoch muß man dem entgegenhalten, daß die cutoff-Methode bei nicht perfekter Validität eine Pseudopräzision der psychologischen Diagnostik vortäuscht. Deutlich wird der Gegensatz der randomisierten und nicht randomisierten Zuweisung bei dem Extrem des nichtvaliden Tests ($r_{yx} = 0.0$).

Nimmt man an, daß 50 % der Gruppe Erfolg hat, wird man bei Neubewerbungen 50 % selegieren. Der cutoff liegt am Centroid $x^* = \bar{x}$. Bei der nichtrandomisierten Zuweisung werden *alle* Personen, die über dem cutoff liegen, akzeptiert, obwohl die Erfolgswahrscheinlichkeit für *jeden* Testwert X nur bei .50 liegt. Das Lotteriespiel wird zugunsten einer pseudopräzisen Testselektion verschoben auf die mögliche und ungewisse Bewältigung des Kriteriums. Bei der randomisierten Zuweisung wird die Invalidität des Instruments offengelegt, indem auf *jedem* Test- oder Fähigkeitsniveau 50 % akzeptiert werden.

Der Ansatz der randomisierten Zuweisung ist sehr flexibel und läßt Utilityüberlegungen weiten Raum. Die Verbesserung der kognitiven Fähigkeiten einer Person werden durch erhöhte Chancen der Person belohnt. Auf diese Weise werden für Personen Anreize geschaffen, die bei der cutoff-Methode nicht in dem Maße vorhanden sind. Personen, die beim ersten Mal abgewiesen wurden, können darüber hinaus erhöhte Akzeptanzchancen erhalten etc.

Wir glauben, daß dieser Ansatz – trotz seiner Problematik – so viele Vorteile besitzt, daß er diskussionswürdig im Rahmen der Fairnessdebatte erscheint, wenn man nicht ganz vom Selektions- oder Counsellinggedanken ablassen möchte.

6. Schlußwort

Es ist deutlich geworden, daß *vor* der fairen Verwendung eines Tests eine gründliche Analyse der Prädiktor-, Kriteriums- und Gruppenvariablen unter testtheoretischen Gesichtspunkten durchgeführt werden muß. Das Ergebnis dieser Analyse entscheidet über die Präferenz einer Fairnesskonzeption.

Kritisch ist dabei die Stabilität der Merkmale, die bisher immer stillschweigend vorausgesetzt wurde. Sind Merkmale veränderungsfähig, sprechen viele Argumente gegen die cutoff-Methode. Als Alternative bietet sich die randomisierte Zuweisung an, deren Nachteil in der Vernachlässigung der Validitätsmaximierung in der für Nichtfachleute schweren logischen Durchschaubarkeit und ihrem „Lotteriecharakter" liegt.

Zusammenfassung

Ziel der vorliegenden Arbeit ist es, die in letzter Zeit häufig aufgegriffene Frage, wann ein Test gegenüber ethnischen und kulturellen Gruppen *fair* ist, einer Beantwortung näher zu bringen. Dazu werden die bekanntesten und miteinander konkurrierenden Testfairnesskonzepte in einem einheitlichen Rahmen dargestellt und ein wesentlicher Teil ihrer Widersprüche auf bestimmte Annahmen eines kombinierten „Fehler-in-den-Variablen- & Fehler-in-der-Gleichung"-Modells zurückgeführt. Die sich daraus ergebenden Konsequenzen für die Konstruktion eines fairen Tests werden inhaltlich und methodologisch diskutiert. Daraus entwickeln sich mehrere Vorschläge zur Gestaltung fairer Tests und zur Überwindung des für eine Prozeß- und Entwicklungsdiagnostik inadäquaten cutoff-Konzepts.

Summary

The purpose of this article is to discuss the non-bayesian definitions of testbias. It can be shown that most contradictory issues are due to unrealistic error assumptions in the underlying regression models. It is necessary to introduce equation *and* measurement errors. By leaving classical test theory and using instead latent trait theory, it is further shown that (with the help of the Rasch-model) it is possible to avoid item *and* testbias.

In the last part of the paper we discuss random and nonrandom admission strategies in the context of testbias. It is argued that nonrandom strategies which are based on cutoff-orientated decisions are antagonistic to the idea of man as a learning system.

Literatur

Anastasi, A.: Culture-fair testing. Educational Horizons, 1964, **43**, 26–30.

Anastasi, A.: Psychological Testing. N. Y.: Macmillan, 1968[3].

Anderson, S. B. & Maier, M. H.: 34000 Pupils and how they grew. Journal of Teacher Education, 1963, **14**, 212–216.

Armor, D. J.: Theta reliability and factor scaling. In: H. L. Costner (Hg.), Sociological methodology. San Francisco: Jossey-Bass, 1974. 17–50.

Banas, P. M.: An investigation of transsituational moderators. Univ. of Minnesota. Unveröff. Diss. 1964.

Breland, H. M. & Ironson, G. H.: DeFUNIS reconsidered: a comparative analysis of alternative admissions strategies. Journal of Educational Measurement, 1976, **13**, 89–99.

Butler, P. M.: Alpha-maximized factor analysis and its relation to alpha and canonical factor analysis. Psychometrika, 1968, **33**, 335–346.

Cardall, C. & Coffman, W. R.: A method for comparing performance of different groups on the items in a test (R. M. 64–61). Princeton: Educational Testing Service, 1964.

Cleary, T. A.: Test bias: Prediction of grades of negro and white students in integrated colleges. J. Educ. Measurement, 1968, **5**, 115–12.

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. J. Educ. Measurement, 1973, **10**, 237–255.

Cronbach, L. J.: Equity in selection – where psychometrics and political philosophy meet. J. Educ. Measurement, 1976, **13**, 31–42.

Darlington, R. B.: Another look at "cultural fairness". J. Educ. Measurement, 1971, 8, 71–82.

Darlington, R. B.: A defense of "rational" personnel selection and two new methods. J. Educ. Measurement, 1976, **13**, 43–52.

Eells, K., Davis, A., Havighurst, R. J., Herrick, V. E. & Tyler, R. W.: Intelligence and cultural differences. Chicago: Univ. of Chicago Press, 1951.

Einhorn, H. J. & Bass, A. R.: Methodological considerations relevant to discrimination in employment testing. Psychol. Bulletin, 1971, **75**, 261–269.

Fischer, G.: Einführung in die Theorie psychologischer Tests. Bern: Huber 1974.

Ghiselli, E. E.: Moderating effects and differential reliability and validity. J. Appl. Psychol., 1963, XLVII, 81–86.

Gösslbauer, J. P.: Tests als Selektionsinstrumente – fair oder unfair? Psychologie und Praxis, 1977, **21**, 95–111.

Goslin, D. A.: Standardized ability tests and testing. Science, 1968, **159**, 851–855.

Green, D. R. & Draper, J. F.: Exploratory studies of bias in achievement tests. Monterey: CTB/McGraw Hill, 1972.

Griliches, Z.: Errors in variables and other unobservables. Econometrica, 1974, **42**, 971–998.

Gross, A. L. & Su, W.: Defining a "fair" or "unbiased" selection model: a question of utilities. J. Appl. Psychol., 1975, **60**, 345–351.

Guthke, J.: Zur Diagnostik der intellektuellen Lernfähigkeit. Berlin: Verlag Erziehung und Bildung, 1972 (Stuttgart: Klett, 1976).

Hobert, X. X. & Dunnette, X. X.: Development of moderator variables to entrance the prediction of managerial effectiveness. J. Appl. Psychol., 1967, **51**, 50–64.

Hunter, J.E. & Schmidt, F.L.: Critical analysis of the statistical and ethical implications of various definitions of test bias. Psychol. Bulletin, 1976, **83**, 1053–1071.

Jensen, A.R.: Another look at culture-fair testing. In: A.R. Jensen: Educational Differences. London: Methuen & Co., 1973.

Johnston, J.: Econometric methods. New York: Wiley, 1963.

Kirkpatrick, J.J., Ewen, R.B., Barrett, R.S. & Katzel, R.A.: Testing and fair employment. New York: N.Y. Univ. Press, 1968.

Linn, R.L.: Fair test use in selection. Rev. Educ. Res., 1973, **43**, 139–161.

Linn, R.L.: In search of fair selection procedures. J. Educ. Measurement, 1976, **13**, 53–58.

Lord, F.M.: Some relations between Guttman's principal components of scale analysis and other psychometric theory. Psychometrika, 1958, **23**, 291–296.

Malinvaud, E.: Statistical methods of econometrics. Amsterdam: North-Holland, 1970.

McClelland, D.C.: Testing for competence rather than for 'intelligence'. Amer. Psychologist, 1973, **28**, 1–14.

Merz, W.R.: Factor analysis as a technique in analyzing test bias. Paper presented at the annual meeting of the California Educational Research Association. Los Angeles, 1973.

Merz, W.R.: Estimating bias in test items utilizing principle component analysis and the general linear solution. Paper presented at the annual meeting of the American Research Association. San Francisco, April 1976.

Möbus, C.: Grundfragen psychologischer Diagnostik: Fairness und Validität. In: W.H. Tack (Hg.): Bericht über den 30. Kongreß der Deutschen Gesellschaft für Psychologie in Regensburg, 1976, Band 2, 17–20. Göttingen: Hogrefe, 1977.

Möbus, C. & Simons, H.: Zur Fairness psychologischer Intelligenztests gegenüber ethnischen und sozialen Gruppen: Kritik klassischer Konzepte. Bericht aus dem Psychologischen Institut der Universität Heidelberg, Nr. 2, Oktober 1975.

Novick, M.R. & Peterson, N.S.: Towards Equalizing Educational and Employment Opportunity. Journal of Educational Measurement, 1976, **13**, 77–88.

Pawlik, K.: Ist die Umwelt Persönlichkeit– Zum Verhältnis von Ökologie, Differentieller Psychologie und Persönlichkeitsforschung. Vortrag auf dem Symposion über ökologische Fragestellungen, Schloß Heinsheim/Neckar, Oktober 1975.

Petersen, N.S.: An Expected Utility Model for "Optimal" Selection. Journal of Educational Statistics, 1976, **1**, 333–358.

Petersen, N.S.: Bias in the Selection Rule – Bias in the Test. 3rd International Symposion on Educational Testing, Leiden, 27.–30.6.1977 (in Press).

Petersen, N.S. & Novick, M.R.: An Evaluation of some Models for Culture-fair Selection. Journal of Educational Measurement, 1976, **13**, 3–31.

Press, J.S.: Applied Multivariate Analysis. N.Y.: Holt, Rinehart & Winston, 1972.

Rudner, L.M.: Efforts Toward the Development of Unbiased Selection and Assessment Instruments. 3rd International Symposion on Educational Testing, Leiden, 27.–30.6.1977 (in press).

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.

Scheunemann, J.: A new Method of Assessing Bias in Test Items. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC. April 1975.

Schneeweiss: Ökonometrie. Würzburg: Physica Verlag, 1971.

Schönfeld, P.: Methoden der Ökonometrie, Bd. I/II. Berlin: Franz Vahlen, 1969.

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, Nr. 8, Juni 1977.

Sörbom, D.: A General Method for Studying Differences in Factor Means and Factor Structure between Groups. The British Journal of Mathematical and Statistical Psychology, 1974, **27**, 229–239.

Sörbom, D.: A Statistical Model for the Measurement of Change in True Scores. In: de Gruijter & van der Kamp (eds.): Advances in Psychological and Educational Measurement, 1976, 159–170, N.Y.: Wiley.

Sörbom, D. & Jöreskog, K.: COFAMM Computer Program. National Educational Resources, Inc., 215 Kenwood Avenue, Ann Arbor, Michigan 48103.

Späth, H.: Algorithmen für multivariable Ausgleichsmodelle. München: Oldenbourg, 1974.

Thorndike, R.L.: Concepts of Culture-Fairness. Journal of Educational Measurement, 1971, **8**, 63–70.

Van de Geer, J.P.: Introduction to Multivariate Analysis for the Social Sciences. San Francisco: Freeman, 1971.

Veale, J.R. & Foreman, D.I.: Cultural Validity of Items and Tests: A new Approach, Score Technical Report. Iowa City, Iowa: Westingshouse Learning Corporation/Measurement Research Center, 1975.

Wolf, B.: Intelligenzaufgaben und Sozialer Status. Weinheim: Beltz, 1977.

Wonnacott, R.J. & Wonnacott, T.H.: Econometrics. N.Y.: Wiley, 1970.

Anschrift des Verfassers: Prof. Dr. Claus Möbus
Institut für Psychologie der FU Berlin im FB 12
Dietrich-Schäfer-Weg 6–10
1000 Berlin 41

Anhang A

Die orthogonale bzw. interdependente Regression

Die in der klassischen Regression verwendete Geradengleichung $x_2 = x_1 \tan \alpha + c$ kann nach einigen Umformungen in die Richtungskosinusgleichung (A.1), die der Hesseschen Normalform entspricht, überführt werden: $x_2 \cos \alpha + x_1 \cos \beta - p = 0$ (vgl. a. Figur 12), wobei:

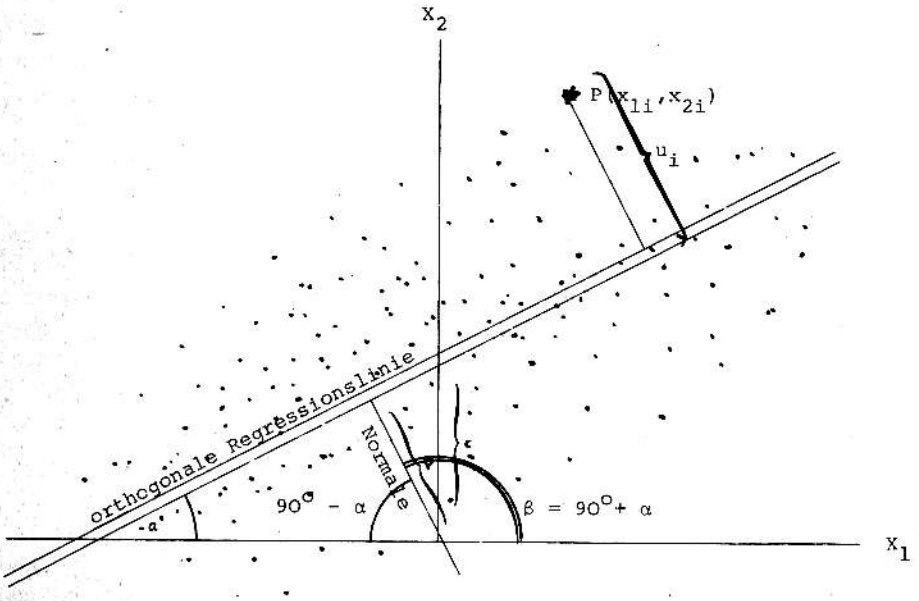
$-p = -c \cos \alpha = b_0$

$\cos \alpha$ = Richtungskosinus (Normale, x_2) = b_1

$\cos \beta$ = Richtungskosinus (Normale, x_1) = b_2

Figur 12

Die orthogonale Regression im Zwei-Variablen-Fall



Die Residuen werden nicht wie bei der klassischen Regression parallel zur Y-Achse sondern orthogonal zur Regressionslinie gemessen.

Zur Vorhersage von x_1 oder x_2 muß durch den betreffenden Regressionskoeffizienten dividiert werden, so daß das Kriterium das Gewicht 1 erhält. Nun werden empirische Meßwerte nicht exakt auf der Regressionsgeraden liegen. Bei der orthogonalen Regression wird der senkrechte Abstand u_i des Punktes $P(x_{1i}, x_{2i})$ als Residuum angesehen:

(A.1) $\quad b_0 + x_{1i}b_1 + x_{2i}b_2 = u_i$ bzw.: $X_e b = u$

wobei: X_e = eine erweiterte Rohdatenmatrix X von der Ordnung (Personen, Variable + 1) ist: die erste Spalte von X_e besteht aus Komponenten mit dem Wert 1;

b_e' = (b_0, b_1, b_2) vollständiger Parametervektor;

u' = Zeilenvektor mit den Residuen.

Das Regressionsproblem besteht jetzt in der Bestimmung des Koeffizientenvektors b, so daß die Fehlerquadratsumme $Q_0 = u'u = b'X_e'X_e b$ unter der Nebenbedingung $b_1'b_1 = 1$ ein Minimum erreicht:

(A.2) $$Q = b'X_e'X_e b - \lambda (b_1'b_1 - 1) = \min!$$

wobei: λ = ein Lagrangescher Multiplikator;
b_1' = der um b_0 reduzierte Parametervektor b.

Die partiellen Ableitungen von Q nach b gleich Null setzend, erhält man beim Übergang zu Abweichungswerten:

(A.3a) $$\hat{b}_0 = 0$$

(A.3b) $$(M - \frac{\lambda}{N} \cdot I)(\hat{b}_1) = (0) \qquad \text{(einfaches Eigenwertproblem)}$$

wobei: M = Varianz/Kovarianzmatrix.

Eine nichttriviale Lösung liegt dann vor, wenn die Koeffizientendeterminante gleich Null ist. Von den verschiedenen Eigenwerten wählen wir den kleinsten, da $\lambda = Q_0$ ist. (A.3b) entspricht (4.17). Die Schätzung der Parameter der orthogonalen Regression ist identisch mit der Parameterschätzung im „Fehler-in-den-Variablen"-Modell, wenn gleiche Störvarianzen vorliegen.

Die praktische Berechnung der orthogonalen Regression bei z-Werten erübrigt sich, weil die Regressionsgerade bei positiver Korrelation der 45 °-Geraden durch den Ursprung entspricht. Bei Abweichungswerten müssen wir die Eigenwerte und Eigenvektoren der Varianz/Kovarianzmatrix bestimmen. Dieses geschieht am einfachsten mit einem Programm zur Hauptkomponentenanalyse. Eine andere Methode wurde von Späth (1974) vorgeschlagen, der auch ein FORTRAN-Programm angibt. Im Gegensatz zur Hauptkomponentenanalyse sind wir am kleinsten Eigenwert und dem dazugehörigen Eigenvektor interessiert. Das ist dennoch nicht erstaunlich, denn die orthogonale Regressionslinie minimiert die Störvariation Q_0 (korrespondiert zur Varianzaufklärung der 2. Hauptkomponente) und maximiert die Varianzaufklärung der 1. Hauptkomponente (korrespondiert zur Störvarianz der Normalen oder 2. Hauptkomponente).

Faßt man die Variablen als Teile eines Kompositum auf, das den latenten Trait messen soll, der sowohl dem Kriterium als auch dem Prädiktor unterliegt, kann man das rein intuitiv formulierte deskriptive Maß für die Güte der Anpassung Q_0 überführen in ein Maß für die Reliabilität (interne Konsistenz) dieses Kompositums (s. Lord, 1958; Butler, 1968; Malinvaud, 1970; Armor, 1974).

Anhang B

Das „Fehler-in-den-Variablen"-Modell

Zur Vereinfachung der Betrachtungen seien die Variablen standardisiert (z-Werte). Die Meßwerte lassen sich in latente Komponenten zerlegen:

(B.1) $$Z = T_z + E_z$$

mit: T_z = Matrix der wahren Werte;
E_z = Matrix der Meßfehler (Ordnung, Personen, Variablen).

Für jede Person seien die Fehler unabhängig vom Personenindex und von den wahren Werten mit dem Nullvektor als Erwartungswertvektor und regulärer Varianz/Kovarianzmatrix Ω^2 verteilt; ferner sei die Beziehung zwischen den wahren Werten fehlerfrei (Schönfeld, 1971, II, S. 115 ff.):

(B.2) $T_z\beta_z = 0$ mit: β_z = Regressionsparametervektor

Zur Schätzung der Parameter wird die gewogene Meßfehlerquadratsumme Q unter der Nebenbedingung (B.2) und $\beta'\Omega^2\beta = 1$:

(B.3) $$Q = \text{Spur}\,[(Z - T_z)\,\Omega_z^{-2}\,(Z - T_z)']\,/\,n + 2\lambda' T_z\beta_z\,/n - \mu\,(\beta_z'\Omega_z^2\beta_z - 1) = \min!$$

wobei: λ' = Zeilenvektor mit Lagrangeschen Multiplikatoren.

Im ersten Schritt werden bei festgehaltenem β die inzidentellen Parameter T_z eliminiert, damit im zweiten Schritt die Regressionsparameter geschätzt werden können. Nach der Berechnung partieller Ableitungen und einigen Umformungen erhält man:

(B.4) $(R - \mu\Omega^2)\,\beta_z = 0$ mit: $\mu = \hat{\beta}_z' R \hat{\beta}_z$

wobei: R = Korrelationsmatrix.

Das allgemeine Eigenwertproblem (B.4) läßt sich in ein einfaches transformieren. Sind alle Meßfehlervarianzen gleich, geht das „Fehler-in-den-Variablen"-Modell in die orthogonale Regression über.

Anhang C

Das kombinierte „Fehler-in-den-Variablen- & Fehler-in-der-Gleichung"-Modell

Liegen z-Werte vor, lassen sich diese nach (B.1) zerlegen. Die Fehler sollen unabhängig verteilt sein mit Erwartungswert Null und Varianz/Kovarianzmatrix $\Omega^2 = \text{diag}\,(\sigma_{\epsilon_x}^2)$. Die Korrelationsmatrix R kann dann zerlegt werden:

(C.1) $R = Z'Z/n = R^0 + \Omega^2$

Die Regressionsgleichung zwischen den latenten Variablen hat den Charakter einer Strukturgleichung (Van de Geer, 1971):

(C.2) $T_z\beta_z = v_z$

wobei: β_z = Spaltenvektor mit Parametern;
v_z = Spaltenvektor mit Gleichungsfehlern.

Wir minimieren die Fehlerquadratsumme v'v/n unter einer Nebenbedingung, die in ähnlicher Weise im Diskriminanzanalysenkriterium auftaucht, wenn man gewillt ist, Ω^2 mit der gepolten Binnenvarianzmatrix W zu vergleichen. Man müßte W (im Gegensatz zur Zwischenvarianzmatrix) als Meßfehlervarianzmatrix ansehen, die die Gruppenzugehörigkeit der Personen verschleiert (vgl. Van de Geer, S. 272, S. 234).

(C.3) $Q = \beta_z' T_z' T_z \beta_z - \lambda\beta_z'\Omega_z^2\beta_z = \min!$

Setzt man die partielle Ableitung von Q nach dem Vektor β_z' gleich Null, erhält man das allgemeine Eigenwertproblem:

(C.4) $(R^0 - \lambda\Omega_z^2)\,\hat{\beta}_z = 0$

Mit Hilfsvariablen ist es möglich (C.4) in ein gewöhnliches Eigenwertproblem zu überführen:

$$(C.5) \qquad (\Omega_z^{-1} R^0 \Omega_z^{-1} - \lambda \cdot I)\, \hat{q}_z = 0$$

wobei: $q_z = \Omega_z \beta_z$

Bei gleichen Meßfehlervarianzen $\Omega^2 = \sigma^2 I$ vereinfacht sich (C.4) zu:

$$(C.6) \qquad (R^0 - \mu I)\, \hat{\beta}_z = 0$$

Bei gleichen Meßfehlervarianzen vereinfacht sich das „Fehler-in-den-Variablen-und-in-der-Gleichung"-Modell auf die Berechnung des kleinsten Eigenvektors der reduzierten Korrelations- oder Kovarianzmatrix. Auf die Möglichkeit, (C.4) mit Hilfe der kanonischen Faktorenanalyse zu lösen, verweist Van de Geer (1971).