Sonderdruck aus:
Bericht über den 30. Kongreß der Deutschen Gesellschaft für Psychologie

GRUNDFRAGEN PSYCHOLOGISCHER DIAGNOSTIK: FAIRNESS UND VALIDITÄT

Claus Möbus

Psychologisches Institut
der
Universität Heidelberg

1. Einleitung

Zu den klassischen Gütekriterien psychologischer Tests (Objektivität,Reliabilität,Validität,Informationsgehalt der Daten in Bezug auf die Parameter,relative Effizienz) gesellte sich in den letzten Jahren ein neues Kriterium, das bis heute kontrovers behandelt wird (Darlington, 1971; Thorndike, 1971; Linn, 1973; Cole, 1973; Petersen & Novick, 1976). Dieses neue Kriterium bezieht sich stärker auf den Handlungs- und Entscheidungsaspekt der Diagnostik, weist enge Verbindungen zur Validitätsproblematik auf und ist in der amerikanischen Literatur unter den Begriffen 'Testfairness', 'Testbias' oder 'Selektionsbias' bekannt geworden. Für Testkonstrukteur und Testanwender stellt sich die Frage, ob mit Testanwendung, Prognose und Entscheidung bestimmte soziale, ethnische oder psychologische Gruppen systematisch diskriminiert werden. Diese Probleme finden auch ausserhalb der Universität grosse Beachtung. Zum Teil ist dieses Interesse auf einige Musterprozesse vor amerikanischen Gerichten zurückzuführen. Sie wurden von Personen angestrengt, die sich durch - auf psychologischen Tests basierenden - Entscheidungen benachteiligt fühlten (Ginger, 1974; Breland & Ironson, 1976).

2. Definitionen der Testfairness

In der fachspezifischen Diskussion haben sich eine Reihe von Definitionen herausgeschält, die sich alle mehr oder minder zu widersprechen scheinen. Der Vollständigkeit halber seien hier noch die Definitionen (5, 7) eingeführt. Wir unterscheiden: (1) das Identitätskonzept (Darlington's Fall 4); (2) das OLS-Regressionskonzept Y auf X (Cleary, 1968, Darlington's Fall 1); (3) das OLS-Regressionskonzept X auf Y (Darlington's Fall 3); (4) das Risikomodell Y auf X (Einhorn & Bass, 1971); (5) das Risikomodell X auf Y (Möbus & Simons, 1977); (6) das konverse Risikomodell Y auf X (Petersen & Novick, 1976); (7) das konverse Risikomodell X auf Y; (8) das Gleichwahrscheinlichkeitsmodell (Linn, 1973); (9) das konverse Gleichwahrscheinlichkeitsmodell (Petersen & Novick, 1976); (10) das konstante Verhältnismodell (Thorndike, 1971); (11) das konverse konstante Verhältnismodell (Petersen & Novick, 1976); (12) das bedingte Wahrscheinlichkeitsmodell (Cole, 1973); (13) das konverse bedingte Wahrscheinlichkeitsmodell (Petersen & Novick, 1976). Damit die vielleicht entstandene Verwirrung nicht vergrössert wird, wollen wir auf die neuen Vorschläge von Petersen & Novick (1976, das Schwellen-Nutzen-Modell), Cronbach (1976), Darlington (1976), Sawyer, Cole & Cole (1976), Novick & Petersen (1976) nicht eingehen, weil unsere Kritik in Teil 4 sinngemäss ebenfalls auf diese Definitionen zutrifft.

3. Bezugsrahmen für eine weitgehend einheitliche Darstellung der Definitionen (1) - (13)

Als Prämisse schicken wir die Forderung nach e i n e m für alle Gruppen geltenden cutoff auf dem Kriterium $Y=y_*$ und dem Test $X=x_*$ voraus. Gruppenspezifische cutoffs werfen

unlösbare Zuordnungsprobleme auf, wie wir sie von allen Persönlichkeitstypologien her kennen. Zusätzlich erschwerend würde sich der Wunsch vieler Personen auswirken, in eine Gruppe mit 'günstigerem' cut-off aufgenommen zu werden. Für die weiteren Betrachtungen wählen wir drei Darstellungsebenen: a) Akzeptanz-Erfolgsquoten, b) korrelative Zusammenhänge und Pfadmodelle, c) Regressionsmodelle. Für inhaltliche Begründungen sei auf die Originalliteratur und Möbus & Simons (1977) verwiesen.

a) Akzeptanz-Erfolgsquoten-Definitionen: Durch x_* und y_* kann der bivariate Kriterium-Testraum in die Quadranten I, II, III, IV zerlegt werden: I = $P(Y \geq y_*, X \geq x_*)$; II = $P(Y \geq y_*, X < x_*)$; III = $P(Y < y_*, X < x_*)$; IV = $P(Y < y_*, X \geq x_*)$. Die jeweilige Definition (i) ist erfüllt, wenn folgende Wahrscheinlichkeiten in a l l e n Gruppen gleich sind: (8): I/(I+IV); (9): III/(II + III); (10): C=(I+IV)/(I+II); (11): (II+III)/(III+IV).

b) Korrelations- und Pfadmodelle: Wenn wir die kulturelle oder ethnische Drittvariable mit Z bezeichnen, können - unter einigen Verteilungsannahmen - folgende Darstellungen gewählt werden: (1): $r_{xz} = 0$; (2, 4, 6): $r_{yz.x} = 0$ (Z "beeinflusst" das Kriterium Y via X); (3, 5, 7): $r_{xz.y} = 0$ (Z "beeinflusst" X nur via Y: Validierung an einem kulturabhängigen Krit.); (10, 11 bei C=1): $r_{yz} = r_{xz}$ (Y und X werden gleich stark von Z "beeinflusst"). Es ist dabei interessant, dass (10, 11 bei C=1) das geometrische Mittel von (2, 4, 6) und (3, 5, 7) ist.

c) Regressionsmodelle: der cut-off wird mit Hilfe folgender, allen Gruppen gemeinsamen, Regression bestimmt: bei (2, 4, 6) durch OLS-Regression Y auf X; bei (10,11) durch orthogonale Regression (wenn $c = s_{x_i}/s_{y_i} = 1$ für alle Gruppen i); bei (8, 9, 12, 13) durch nichtlineare Regression zwischen Y und X. Natürlich muss die Gleicheit der gruppenspezifischen Regressionen getestet werden (Gulliksen & Wilks, 1950).

4. Evaluation der Definitionen

Verlassen wir die deskriptiven Modelle in 3c. Wenn wir eine Fehlertheorie und damit Konstrukte einführen (Fehler-in-den-Variablen und Fehler-in-der-Gleichung; Johnston, 1972), erkennt man die Abhängigkeit der Wahl des jeweiligen Regressionsmodells vom Verhältnis der Fehlervarianzen in Y und X. Die OLS-Regression Y auf X setzt die Messfehlerfreiheit von X voraus. Für die OLS-Regression X auf Y gilt das Entsprechende. Bei gleichen Störvarianzen in Y und X ist bei $s_y = s_x$ die orthogonale Regression, die der Hauptkomponente entspricht, angebracht (Malinvaud, 1970). Der allgemeinere Fall wird von Schönfeld (1969), van de Geer(1971), Isaac(1970), Werts et al.(1973) behandelt. Wegen unrealistischer Messfehlerannahmen und unserer Konstruktorientierung können wir (2)-(7) aussondern. Die Modelle (8,9,12,13) setzen die Gleichheit nichtlinearer Regressionen (unter komplizierten Fehlerannahmen) voraus und sind nach Petersen & Novick (1976) logisch inkonsistent. Das Modell (10) ist bei C=1 mit (11) konsistent und macht relativ realistische Fehlerannahmen, wenn Y gut operationalisiert und reliabel ist. Ferner sind nach diesem Modell Y und X gleichberechtigt, die Prognose ist richtungsunabhängig und die Gleichheit von orthogonaler Regression mit der Hauptkomponente legt die Auswahl eindimensionaler (bzw. kongenerischer) Tests nahe.

5. Neue Vorschläge

3c und 4 zeigen die enge Verbindung von Testfairness und Gleichheit struktureller Parameter des Prädiktionssystems. Wir wollen hier zwischen Item- und Systembias unterscheiden und folgende Definition einführen:

Ein Prognosesystem mit dem Test als Prädiktor ist dann fair, wenn die strukturellen Parameter dieses Systems (=Strukturkoeffizienten der Prognosegleichung im Sinne von Neyman & Scott, 1948) für alle relevanten ethnischen und sozialen Gruppen gleich sind und daher die interindividuelle Variation in der Kriteriumsprognose nur auf die Variation der inzidentellen Personenparameter zurückzuführen ist.

Bei der praktischen Testkonstruktion muss zuerst der itembias beseitigt werden. Das kann man mit den stichprobenunabhängigen logistischen Testmodellen erreichen. Wir verwenden hierbei ihre spezifische Objektivität (Rasch, 1966). Im Falle e i n e r Dimension, die durch Kriteriumsitems (KI) u n d Testitems (TI) operationalisiert wird, kann es nach dieser Definition nur Itembias geben. Die Konstruktion eines fairen Tests hätte dann folgendermassen abzulaufen:

1. Bildung eines Pools von KI und TI. Die TI sollen die gleiche latente Dimension wie die KI messen; jedoch sollten sie leichter applizierbar sein, um die Testkonstruktion zu rechtfertigen.
2. Bildung homogener Itempools, die sowohl KI als auch TI enthalten. Die TI sollten in der Überzahl sein.
3. Schätzung aller Itemparameter an einer Stichprobe, die alle relevanten Gruppen umfasst.
4. Modelltest der spezifisch objektiven Messprozedur (Andersen, 1973; Fischer, 1974). Lässt sich die Modellgültigkeit nicht zurückweisen, bedeutet die Gleichheit der Strukturparameter zwischen den relevanten Gruppen T e s t f a i r n e s s gegenüber ihren Mitgliedern. Darüber hinaus sollte das Modell auch bei anderen Gruppierungen beibehalten werden können.
5. Ausschluss modellunverträglicher TI und KI.

Anschliessend könnten mit den TI Prognosen getroffen werden:

1. Vorgabe der TI an eine neue Gruppe von Personen.
2. Schätzung der Personenparameter (vgl. Fischer, S.239) bei festgehaltenen TI-Parametern. Die Anzahl der TI sollte gross sein (Verringerung von Stichprobenfehlern, feine Abstufung der Personenparameterskala.
3. Prognose der Lösungs- oder Reaktionswahrscheinlichkeiten in den KI.

Stehen dagegen die Konstrukte nicht in perfekter Beziehung zueinander, kann Systembias eintreten, der ebenfalls ausgeschaltet werden muss. Die Parameter des rekursiven oder nichtrekursiven Prognosesystems müssen gruppenunabhängig sein. ML-Schätzungen und Likelihood-ratio Tests für diesen Aspekt der Fairness finden sich z.B. bei Jöreskog (1971,1973) und Rock et al. (1976). Diese Strukturmodelle können dabei auch mit den Personenparametern der logistischen Testmodelle berechnet werden. Ihr Hauptvorteil liegt in der Analysemöglichkeit komplizierter dynamischer Wechselbeziehungen zwischen Y und X. Gerade dieser Aspekt ist trotz bekannter Ergebnisse (zB. Rosenthaleffekt) in der Fairnessdebatte kaum berücksichtigt worden. So kann ein feedback von Y auf X vorliegen, wenn die Aufnahme in ein Programm, Schule, Hochschule seinerseits wieder einen positiven Effekt auf die Prädiktoren hat. Liegen die wechselseitigen Veränderungen zwischen den Messzeitpunkten, führen einfache Regressionsmodelle zu irrtümlichen Schlussfolgerungen. Es müssen - ähnlich wie in ökonometrischen Modellen - simultane Strukturmodelle eingeführt werden (Jöreskog, 1973).

6. Literatur

Andersen, E.B. A goodness-of-fit test for the Rasch model, Psychometrika, 38, 1973, 1

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

Cleary, T.A. Test bias: Prediction of grades of Negro and white students in integrated colleges. Journal of Educational Measurement, 1968, 5, 115-124

Cole, N.S. Bias in selection. Journal of Educational Measurement, 1973, 10, 237-255

Cronbach, L.J. Equity in selection - where psychometrics and political philosophy meet.

Journal of Educational Measurement,1976,13, 31-42

Darlington,R.B. Another look at "cultural fairness". Journal of Educational Measurement, 1971,8,71-82

Darlington,R.B. A defense of "rational" personnel selection, and two new methods. Journal of Educational Measurement, 1976,13,43-52

Einhorn,H.J. & Bass,A.R. Methodological considerations relevant to discrimination in employment testing. Psychological Bulletin,1971,75,261-269

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

Isaac,P. Linear regression, structural relations, and measurement error. Psychological Bulletin,1970,74,213-218

Ginger,A.F.(ed). DeFUNIS versus ODEGAARD and the University of Washington (3 vols.). Dobbs Ferry,N.Y.:Oceana Publications,Inc.,1974

Gulliksen,H. & Wilks,S.S. Regression tests for several samples. Psychometrika,1950,15, 91-114

Jöreskog,K.G. Simultaneous factor analysis in several populations. Psychometrika,1971, 36,409-426

Jöreskog,K.G. A general method for estimating a linear structural equation system. in: A.S.Goldberger & O.D.Duncan (eds.) Structural equation models in the social sciences. N.Y.:Seminar Press,1973, 85-112

Linn,R.L. Fair test use in selection. Review of Educational Research,1973,43,139-161

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

Möbus,C. & Simons,H. Zur Fairness psychologischer Intelligenztests gegenüber ethnischen, sozialen und psychologischen Gruppen, Diagnostica,1977 (im Druck)

Neyman,J. & Scott,E.L. Consistent estimates based on partially consistent observations. Econometrica,1948,16,1-12

Novick,M.R. & Petersen,N.S. Towards equalizing educational and employment opportunity. Journal of Educational Measmt. ,1976,13,77-88

Petersen,N.S. & Novick,M.R. An evaluation of some models for culture-fair selection. Journal of Educational Measmt. ,1976,13,3-30

Rasch,G. An informal report on a theory of objectivity in comparisons. Proceedings of of the NUFFIC international summer session in science at "Het Oude Hof", The Hague, 14.-28. Juli,1966

Rock,D.A.,Werts,C.E.,Linn,R.L. & Jöreskog,K.G. A maximum likelihood solution to the errors in variables and errors in equation models. Multivariate Behavioral Research (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

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

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

Werts,C.E.,Linn,R.L. & Jöreskog,K.G. Another perspective on "linear regression, structural relations, and measurement error", Educational and Psychological Measurement,1973, 33,327-332