pedagogiska
rapporter
umeå

nr 35 1973

MÄTPROBLEM I NORM- OCH KRITERIERELATERADE PROV

Några analyser och försök med tonvikt på reliabilitets- och diskriminationsmått

Ingemar Wedman

UNIVERSITETET OCH LÄRARHÖGSKOLAN I UMEÅ

RÄTTELSER

Scandinavian Journal of Educational Research

Sidan 28 formel (9):

$$\alpha = \frac{n\bar{c}_{ij}}{\bar{V}_i + (n-1)\bar{c}_{ij}}$$

Sidan 33 formel (30):

$$r'_{it} = r'_{pb} = \sqrt{\frac{n}{n-1}} \; \frac{r_{pb}S_t - \sqrt{p_i q_i}}{\sqrt{S_t^2 - \sum_i p_i q_i}}$$

MÄTPROBLEM I

NORM- OCH KRITERIERELATERADE PROV

NÅGRA ANALYSER OCH FÖRSÖK MED TONVIKT PÅ RELIABILITETS- OCH DISKRIMINATIONSMÅTT

INGEMAR WEDMAN

UMEÅ UNIVERSITET
PEDAGOGISKA INSTITUTIONEN
1973

PEDAGOGISKA

RAPPORTER

UMEÅ NR 35 1973

MÄTPROBLEM I NORM- OCH KRITERIERELATERADE PROV

NÅGRA ANALYSER OCH FÖRSÖK MED TONVIKT PÅ RELIABILITETS- OCH DISKRIMINATIONSMÅTT

INGEMAR WEDMAN

UNIVERSITETET OCH LÄRARHÖGSKOLAN I UMEÅ

FÖRORD

Det arbete som här redovisas har tillkommit genom mångas stöd. Ett speciellt tack vill jag ge min handledare professor Sten Henrysson. Hans stora kunnande och konstruktiva kritik, liksom hans positiva attityd har varit mig till ovärderlig hjälp. Hans engagemang för mitt fysiska välbefinnande har också varit betydelsefullt och i hög grad bidragit till att underlätta mitt arbete.

Mitt tack riktas även till docenterna Jarl Backman, Gösta Berglund, Hans-Magne Eikeland, Stig Fhanér och Ference Marton, som läst delar av mitt manuskript och givit mig många värdefulla synpunkter.

Till mina arbetskamrater står jag i stor tacksamhetsskuld. Ett särskilt erkännande vill jag ge Mats Hamrén, som lagt ned ett omfattande programmeringsarbete.

Till sist vill jag tacka min hustru Barbro för att inte alltför starkt ha betonat emancipationsfrågan under mitt arbete och vår son Jörgen för att ha anpassat sin dygnsrytm till min.

Detta arbete har finansierats med medel från Statens Råd för Samhällsforskning.

Umeå i april 1973

Ingemar Wedman

Wedman, I. Problems of measurement in norm- and criterion-referenced tests. Some analyses and investigations with emphasis on reliability and discrimination measures. Pedagogiska rapporter, Umeå, 1973, No 35.

SUMMARY

The present report is a summary discussion of a number of studies in connection with the following problems: a) analysis and empirical investigations of some test theoretical formulae and principles, b) differential scoring of multiple-choice questions and c) criterion-referenced tests: theoretical and empirical implications. In studies associated with the first problem is shown how some test theoretical formulae and principles can be derived from the information in the inter-item-covariance matrix as well as how this information can be used to explain empirically found relations. In the latter respect the standard error of measurement and various biserial correlation techniques are dealt with. Studies carried out within the second field of problems account for reliability and validity effects with differential scoring of multiple-choice questions. A priori as well as empirical weight systems have been used. Systematically ordered alternatives gave higher reliability with differential scoring compared to conventional scoring. At the same time, however, the dimensionality of the test changed with negative validity effects as a consequence. The trends of the results were the same for the two types of weight systems. In connection with the third field of problems the development of criterion-referenced tests, their qualities and restrictions are discussed. Three aspects are quite closely penetrated, viz. the defining of objectives, the homogeneity and the different cutting-scores of the tests. Furthermore theoretical as well as empirical problems of measurement techniques when evaluating data from criterion-referenced measurement are dealt with. In the empirical study especially low and moderate correlations between norm- and criterion-referenced discrimination indices were received, as well as between different criterion-referenced discrimination indices. Moreover, the results indicated that increased validity defined in terms of the difference between pre- and post-tests does not necessarily imply increased reliability when placing subjects above and below a fixed cutting-score.

Föreliggande rapport utgör en sammanfattande diskussion av följande arbeten:

(I) Henrysson, S., & Wedman, I. Analysis of the Inter-Item-Covariance-Matrix. Scandinavian Journal of Educational Research, 1972, 16, 25-35.

(II) Wedman, I. Mätningens standardfel och testlängd. Pedagogiska rapporter, Umeå, 1973, nr 31.

(III) Wedman, I. Relationer mellan biseriala diskriminationsindex. Pedagogiska rapporter, Umeå, 1971, nr 16.

(IV) Wedman, I. Reliabilitets- och validitetsstudier vid differentiell poängsättning av flervalsfrågor. Pedagogiska rapporter, Umeå, 1973, nr 32.

(V) Wedman, I. Kriterierelaterade prov: bakgrund, egenskaper och begränsningar. Pedagogiska rapporter, Umeå, 1973, nr 33.

(VI) Wedman, I. Reliabilitets-, validitets- och diskriminationsmått för kriterierelaterade prov. Pedagogiska rapporter, Umeå, 1973, nr 34.

I den framställning som följer kommer hänvisningar till ovanstående arbeten att göras i form av motsvarande romerska siffror.

INTRODUKTION

Mätningar med test eller prov har blivit allt vanligare i olika sammanhang. I takt med den ökade användningen av prov har också teorin bakom dessa utvecklats och forskningen för att förbättra provens kvalitet breddats.

De centrala begreppen inom testteorin är reliabilitet och validitet. Beträffande reliabilitetsteorin har den sedan länge varit relativt väl utvecklad (Gulliksen, 1967). Först på senare tid har nya angreppssätt presenterats. Från den klassiska reliabilitetsteorins utgångspunkt i korrelationen mellan parallella test (Gulliksen, 1967) betraktar Tryon (1957), Cronbach et al. (1963 och 1972) och Lord & Novick (1968) reliabiliteten i termer av generaliserbarhet till ett tänkt universum av uppgifter. Fortfarande är emellertid den klassiska reliabilitetsteorin den dominerande i provsammanhang och kännedom om den är också mer eller mindre en nödvändig förutsättning för förståelsen av de nya mer komplicerade modellerna.

Den klassiska testteorin tillsammans med de utvecklingar som skett på senare år har emellertid en begränsad täckning och tillämpning. De s k kriterierelaterade proven (Glaser, 1963) vilar på delvis andra förutsättningar.

Några problemområden inom testteori och uppgiftsanalys

På det matematiskt-deskriptiva planet kan olika utgångspunkter tas för att härleda och klargöra olika formler och principer inom testteori och uppgiftsanalys. En mycket fruktbar utgångspunkt för att beskriva många relationer har visat sig vara informationen i inter-item-kovariansmatrisen. Beträffande reliabilitetsteori finns detta utvecklat av t ex Cronbach (1951) och Tryon (1957). McLean & Tait (1954) har visat hur uppgiftsanalysdata på ett

enkelt sätt kan erhållas ur ovan nämnda matris. Nedan redovisas och diskuteras ytterligare ansatser med samma utgångspunkt (I, II, III).

Att med utgångspunkt från informationen i inter-item-kovariansmatrisen kunna härleda och klargöra olika former och principer inom testteori och uppgiftsanalys har både praktiska och teoretiska konsekvenser. Ur praktisk synvinkel möjliggörs därmed en förenkled undervisning i testteori och också förenklade rutiner för databehandling av provresultat. Ur teoretisk synvinkel skapas underlag för nya upptäckter liksom för förklaringar av empiriskt funna samband (II, III).

För att prov skall vara användbara krävs att de är reliabla. Uppgiftsanalysen, som vanligen inrymmer någon form av biserial korrelationsteknik, har bl a det syftet att ge information om vilken eller vilka uppgifter som fungerar mindre bra ur homogenitets- och reliabilitetssynpunkt. Sådana uppgifter brukar revideras eller utmönstras. I (III) jämförs olika biseriala korrelationstekniker beträffande vilka förutsättningar de baseras på och vilka estimat de ger upphov till.

Ett annat sätt som i reliabilitetshöjande syfte prövats på prov bestående av flervalsfrågor har varit att poängsätta samtliga svarsalternativ efter grad av riktighet (se t ex Stanley och Wang, 1968). I vanliga fall bedöms sådana frågor med en poäng om svaret är riktigt och noll poäng om svaret är felaktigt. Avsikten med differentiell poängsättning av denna typ av frågor är att höja andelen "sann" varians av den observerade variansen och därigenom få en reliabilitetsökning (Davis & Fifer, 1959). I (IV) har detta poängsättningsförfarande prövats.

I provsammanhang kan man skilja mellan norm- och kriterierelaterade prov (NRP respektive KRP) (Glaser, 1963; Pop-

ham & Husek, 1969; Glaser & Nitko, 1971). I ett normrelaterat prov jämförs varje individs resultat med resultaten från en normgrupp. I ett kriterierelaterat prov däremot jämförs individernas resultat med ett fastställt kriterium.

På senare tid har det riktats kritik mot att använda normrelaterade prov i undervisningssammanhang. Syftet borde vara att ta reda på vad eleverna kan och mindre att rangordna dem. Många problem återstår emellertid ännu att lösa beträffande de kriterierelaterade provens syfte och uppbyggnad (V).

I ett bestämt avseende är NRP och KRP lika. De vanliga provkvaliteterna reliabilitet och validitet är av betydelse i båda provtyperna. Oavsett vilken provtyp man utgår från strävar man efter att få dem så reliabla och valida som möjligt (Popham & Husek, 1969; Hambleton, 1972). I det sammanhanget kan man emellertid konstatera att den reliabilitetsteori som finns beskriven av t ex Gulliksen (1967) och Lord & Novick (1968) gäller främst för de normrelaterade proven. För de kriterierelaterade proven finns ännu ingen motsvarande teoretisk underbyggnad (Popham, 1971).

I (VI) diskuteras och prövas olika metoder som föreslagits för utvärdering av KRP. Denna studie skall ses mot bakgrund av det utvecklingsskede forskningen inom detta område befinner sig i.

Nedan följer en närmare beskrivning av de studier som är föremål för diskussion i denna rapport. Dessa studier kan inordnas under följande rubriker:

a. Analys och empiriska prövningar av några testteoretiska formler och principer (I-III).

b. Differentiell poängsättning av flervalsfrågor (IV).

c. Kriterierelaterade prov: teoretiska och empiriska implikationer (V-VI).

ANALYS OCH EMPIRISKA PRÖVNINGAR AV NÅGRA TESTTEORETISKA FORMLER OCH PRINCIPER

I tre rapporter, (I), (II) och (III), visas hur formler och principer inom testteori på ett enkelt sätt kan relateras till inter-item-kovariansmatrisen. I den första studien (I) ges en samlad beskrivning av vilka möjligheter inter-item-kovariansmatrisen erbjuder i detta avseende. I rapporterna (II) och (III) tillämpas denna metodik för att förklara empiriskt erhållna resultat.

Analys av inter-item-kovariansmatrisen

I den klassiska reliabilitetsteorin är korrelationen mellan parallella test ett uttryck för reliabiliteten (Gulliksen, 1967; Lord & Novick, 1968). Med samma utgångspunkt har Kuder & Richardson (1967) visat hur reliabiliteten kan bestämmas utifrån kännedom om resultaten från endast ett test. Andra har presenterat motsvarande metoder för att uppskatta reliabiliteten (se Cronbach et al., 1963, för en kortfattad resumé).

Med utgångspunkt i inter-item-kovariansmatrisen visas i (I) att Kuder-Richardsons reliabilitetskoefficient 20 (Cronbachs α-koefficient) under vissa antaganden är identisk med parallelltest- och "split-half"-koefficienterna.

I samma studie (I) visas också hur uppgiftsanalysdata kan erhållas ur nämnda matris. Tre korrelationsmått behandlas i detta avseende (se också III), nämligen (a) sambandet mellan en uppgift och det totala provet, (b) sambandet mellan en uppgift och de övriga (n-1) uppgifterna i provet och (c) sambandet mellan en uppgift och den generella faktor som provet mäter.

Det speciellt nya i denna uppsats (I) är ett nytt sätt att bestämma α och mätningens standardfel, S_{EM}. Om man utför

en faktoranalys enligt centroidmetoden på den matris som definierar variansen, V_t, och därefter summerar uppgifternas laddningar i den första faktorn erhålls

$$\sum_i a_{i1} = \sqrt{\sum_j \sum_i c_{ij}} \qquad (1)$$

där a_{i1} = uppgift i:s laddning i den första faktorn
c_{ij} = kovariansen mellan uppgifterna i och j.

Motsvarande operation på matrisen, som istället för varianser i diagonalen innehåller kommunaliteter,

$$h_i^2 = \frac{\sum_j c_{ij}}{n-1} \qquad (i \neq j) \qquad (2)$$

där h_i^2 = kommunaliteten för uppgift i,
ger följande uttryck:

$$\sum_i a'_{i1} = \sqrt{\frac{n}{n-1}} \sqrt{\sum_j \sum_i c_{ij}} \qquad (i \neq j) \qquad (3)$$

där a'_{i1} = uppgift i:s laddning i den första faktorn.

Om (1) och (3) kvadreras erhålles nämnare och täljare i uttrycket för reliabiliteten, α, enligt Kuder-Richardsons formel 20, d v s

$$\alpha = \frac{(\sum_i a'_{i1})^2}{(\sum_i a_{i1})^2} \qquad (4)$$

Detta uttryck (4) möjliggör urval av uppgifter i syfte att maximera α.

Med samma faktoranalytiska utgångspunkt visas också att mätningsfelet, S_{EM}, kan uttryckas på följande sätt:

$$S_{EM} = \sqrt{(\sum_i a_{i1})^2 - (\sum_i a'_{i1})^2} \qquad (5)$$

Mätningens standardfel och testlängd

I tidigare arbeten har relationen mellan mätningens standardfel, S_{EM}, och testlängd (n) behandlats (Lord, 1957 och 1959; Swineford, 1959; Gardner, 1970). I en omfattande empirisk studie fann Lord (1959) följande samband mellan S_{EM} och testlängd (n):

$$S_{EM} = .432\sqrt{n} \qquad (6)$$

S_{EM} bestämdes i detta fall utifrån följande relation

$$S_{EM} = S_t\sqrt{1-KR20} \qquad (7)$$

där S_t = standardavvikelsen i totalpoäng
KR20 = Kuder-Richardsons reliabilitetskoefficient 20.

Utifrån antaganden om uppgifternas svårighetsgrad och interkorrelationer i vanligen använde prov kunde Swineford (1959) på analytisk väg estimera ett liknande samband som i (6) mellan S_{EM} och testlängd. Gardner (1970) kom fram till en motsvarande relation utifrån antaganden om uppgifternas svårighetsgrad och relationen mellan standardavvikelse och testlängd.

I (II) visas med utgångspunkt i inter-item-kovariansmatrisen hur S_{EM} kan uttryckas i termer av genomsnittlig svårighetsgrad hos uppgifter, $(\overline{pq})$, och antal uppgifter (n). Följande samband gäller approximativt mellan S_{EM} och testlängd:

$$S_{EM} \doteq \sqrt{\overline{pq}}\sqrt{n} \qquad (8)$$

Liknande relation har diskuterats av Lord (1957 och 1959) och Swineford (1959). Ovanstående approximationsformel är i det sammanhanget att betrakta som maximaluttryck.

Uttrycket i (8) möjliggör en enkel förklaring av Lords (1959) empiriskt erhållna resultat. För provkonstruktionsändamål har (8) sitt givna värde. Med kännedom om detta samband kan man a priori approximativt uppskatta felet i mätningarna.

Relationer mellan biseriala diskriminationsindex

Vid uppgiftsanalys ingår oftast beräkningar av sambandet mellan resultatet på de enskilda uppgifterna och resultatet på provet. Dessa beräkningar utförs vanligen med biserial korrelationsteknik.

Om korrelationen beräknas mellan respektive uppgift och det totala testet (IT) kommer sambandet att överskattas, eftersom uppgiften ifråga också ingår i det totala provet och uppgiftens unika varians därigenom kommer att bli gemensam för uppgiften och provet.

Olika formler för att korrigera denna överskattning har presenterats. Två olika förfaranden har härvidlag redovisats. Det ena innebär att sambandet beräknas mellan respektive uppgift och de återstående (n-1) uppgifterna (Zubin, 1934; Guilford, 1954; Henrysson, 1963). Henryssons (1963) formel (IR_1) har visat sig vara den mest relevanta om man önskar en koefficient enligt denna princip (Henrysson, 1963; Cureton, 1966; Wolf, 1967).

En nackdel med att beräkna sambandet mellan respektive uppgift och de övriga uppgifterna är att kriteriet blir olika för de olika uppgifterna. För att eliminera denna brist har Henrysson (1963) och Cureton (1966) redovisat alternativa tekniker med syfte att endast eliminera den del av variansen som ger upphov till överskattningen, d v s den unika delen av respektive uppgifts varians från det totala provet.

Curetons (1966) teknik (IR_3) innebär att man i det totala provet ersätter ifrågavarande uppgift med en ekvivalent uppgift. Med Henryssons (1963) teknik (IR_2) elimineras den unika variansen på så sätt att medelvärdet av uppgiftens kovarianser med de övriga uppgifterna insätts i inter-item-kovariansmatrisens diagonal.

I (III) redovisas en empirisk studie, där både okorrigerade och korrigerade biseriala korrelationsmått jämförs. De senare utgjordes av Henryssons (1963) koefficient (IR_1) avseende sambandet mellan respektive uppgift och de (n-1) återstående uppgifterna och de av Henrysson (1963) och Cureton (1966) redovisade koefficienterna (IR_2 respektive IR_3). De fyra korrelationskoefficienterna prövades och jämfördes på två olika prov.

För samtliga uppgifter oavsett prov erhölls relationen $IT>IR_2>IR_1$. Resultatet visade vidare att differensen mellan de olika koefficienterna minskade med ökad testlängd. I båda proven var den överskattning man erhöll med IT relativt IR_2 mindre än .10 vid 20 uppgifter.

Beträffande IR_3 kunde i stort sett ingen skillnad noteras relativt IR_1. Med hänsyn till syftet med denna formel (IR_3) var resultatet anmärkningsvärt. Med utgångspunkt från inter-item-kovariansmatrisen kunde man emellertid visa att IR_1 och IR_3 är approximativt lika, vilket innebär att IR_3 inte är ett lämpligt sambandsmått om man önskar en koefficient som enbart eliminerar ifrågavarande uppgifts unika varians från det totala provet.

DIFFERENTIELL POÄNGSÄTTNING AV FLERVALSFRÅGOR

Prov bestående av flervalsfrågor har blivit accepterade i många olika sammanhang. Bedömningen av dessa frågor har i huvudsak varit sådan att rätt svar belönats med en poäng och ett felaktigt svar med noll poäng.

En invändning som rests mot detta bedömningssätt har varit att ingen hänsyn tas till partiella kunskaper och att man därigenom bortser från relevant svarsinformation (Davis & Fifer, 1959; Ramsey, 1968; Hendrickson, 1971 och andra). På grund av att svarsalternativen oftast representerar olika grader av riktighet relativt det korrekta alternativet, har man istället menat att val av olika svarsalternativ också borde medföra olika poäng i bedömningen. Härigenom skulle man erhålla högre reliabilitet i proven, eftersom andelen "sann" varians av observerad varians då skulle öka (Davis & Fifer, 1959; Hendrickson, 1971; se också Guttman & Schlesinger, 1967) och som en följd härav också högre validitet (Hendrickson, 1971).

I (IV) redovisas tre delförsök, där olika former av differentiell poängsättningsteknik tillämpats och jämförts med konventionell dikotom poängsättning. I första hand har reliabilitetsaspekten behandlats i dessa försök.

I det första delförsöket som var av explorativ karaktär tillämpades modifierade empiriska vikter (Davis & Fifer, 1959). Det empiriska underlaget till dessa vikter utgjordes av varje alternativs biseriala korrelation med totalpoängen. Ingen positiv reliabilitetseffekt kunde konstateras med differentiell poängsättning jämfört med konventionell. En tänkbar förklaring till det "negativa" resultatet finns att söka i provets utformning. Det prov som användes var ursprungligen avsett för konventionell bedömning.

I det andra delförsöket prövades både empiriska vikter och a priori-vikter (se Stanley & Wang, 1968). Underlaget till de förra utgjordes av varje alternativs biseriala korrelation med totalpoängen. Två uppsättningar a priori-vikter användes. Underlaget till båda dessa uppsättningar var 22 individers rangordning av svarsalternativens riktighet relativt huvudordet.

Två synonymordprov används i detta försök. Proven var identiska med avseende på huvudord och helt korrekt alternativ. Endast distraktorerna skilde sig åt i de två proven. Det ena var avsett för konventionell bedömning och hade utformats enligt konventionella normer. Det andra hade utformats i syfte att erhålla en gradering av alternativens riktighet relativt huvudordet.

Beträffande det "graderade" synonymordprovet befanns alla tre differentiella poängsättningstekniker medföra högre reliabilitet än konventionell poängsättning. Bland de förra gav a priori-vikterna det bästa resultatet, vilket kan betraktas som anmärkningsvärt i den meningen att endast mycket grova vikttilldelningar prövades.

Denna effekt elimineras emellertid om jämförelserna istället görs med den reliabilitet som erhölls vid konventionell bedömning av det konventionellt utformade provet (som endast poängsattes på detta sätt). Det är mot denna bakgrund sannolikt att en av anledningarna till de olikheter i resultat som framkommit i den tidigare forskningen kan ha berott på olikheter i provens utformning. Konstruktionssättet syns ha en avgörande inverkan på reliabilitetseffekterna vid olika poängsättningsförfaranden.

Samma försöksuppläggning som i det andra delförsöket användes i det tredje delförsöket. Två viktiga tillägg gjordes emellertid. För det första användes mer stringenta viktsystem både vad gäller empirisk och a priori poängsättning. För det andra studerades också validiteten.

Beträffande empiriska vikter användes dels faktiskt erhållna biseriala korrelationer mellan respektive svarsalternativ och totalpoäng och dels medelvärden i totalpoäng för de som valt respektive alternativ. De senare representerar s k guttmanvikter (Guttman, 1941; se också Hendrickson, 1971). A priori-vikterna erhölls från skatt-

ningar enligt parvis jämförelseteknik (Edwards, 1957). Från dessa bedömningar utformades två slags viktuppsättningar, det ena baserat på rangordning mellan svarsalternativen (jämför med den andra delundersökningen) och det andra baserat på subjektiva distanser mellan svarsalternativen. Tre viktuppsättningar av vardera slaget bestämdes och tillämpades, var och en erhållen från tre olika individgrupper.

Det största intresset vid differentiell poängsättning har knutits till reliabilitetsfrågan. På senare tid har emellertid validitetsaspekten uppmärksammats i allt högre grad. Flera anmärkningsvärda resultat har också redovisats i detta avseende. Hendrickson (1971), Hendrickson & Green (1972) och Reilly & Jackson (1972) fann alla en samtidig reliabilitetsökning och validitetsminskning vid differentiell poängsättning.

På motsvarande sätt som i det andra delförsöket erhölls i denna genomgående högre reliabilitet vid differentiell poängsättning jämfört med konventionell poängsättning om jämförelserna görs på det "graderade" provet. Effekterna var vidare störst då de stringenta viktsystemen tillämpades. Den poängsättning som baserades på guttmanvikter resulterade i högre reliabilitet än den som baserades på biseriala korrelationsvikter. Beträffande a priori-vikter erhölls klart bättre resultat då poängsättningen baserades på relativa distansvikter än då den baserades på rangordningsvikter.

Om emellertid jämförelserna görs med reliabiliteten för det konventionella provet dikotomt rättat blir bilden delvis en annan (jämför med det andra delförsöket). Endast då det differentiella poängsättningsförfarandet baserades på guttmanvikter och relativa distansvikter utifrån elevernas egna skattningar erhölls i detta fall positiva reliabilitetseffekter.

I validitetshänseende uppvisar resultaten en anmärkningsvärd homogenitet. Den reliabilitetsökning man i många fall kan konstatera har inte resulterat i en motsvarande validitetsökning. Snarare tycks det omvända förhållandet råda. Det tycks alltså som om den reliabilitetsökning man under vissa betingelser kan förvänta inte motsvaras av en validitetsökning. Uttryckt i andra termer innebär det att dimensionaliteten ändras i provet då differentiell poängsättning tillämpas.

Det bör i detta sammanhang nämnas att de resultat som ovan har redovisats för differentiell poängsättning baserat på empiriska viktsystem i samtliga fallen representerar korsvaliderade resultat. De empiriska viktuppsättningar som bestämts utifrån resultaten för en grupp individer har tillämpats på en annan grupps prestationer. För att möjliggöra meningsfulla jämförelser mellan konventionell och differentiell poängsättning är detta förfaringssätt nödvändigt (se t ex Cureton, 1950).

Sammanfattningsvis kan man av resultaten i (IV) konstatera att differentiell poängsättning av flervalsfrågors svarsalternativ kan resultera i positiva reliabilitetseffekter relativt konventionell bedömning. Två faktorer syns ha en avgörande inverkan på storleken av denna effekt, nämligen utformningen av provet och slaget av viktsystem. Om flervalsproven utformas med "graderade" eller ordnade svarsförslag kan man påräkna vissa reliabilitetsvinster med differentiell poängsättning jämfört med konventionell poängsättning. Emellertid måste man observera att en ökning i reliabilitet på detta sätt inte behöver innebära att också validiteten ökar.

Beträffande reliabilitets- och validitetsaspekterna bör den fortsatta forskningen inom detta område närmast inriktas mot att systematiskt kontrollera och helst också

variera svarsalternativens riktighet för att därigenom erhålla mer precis information angående effekterna av differentiell poängsättning.

KRITERIERELATERADE PROV: TEORETISKA OCH EMPIRISKA IMPLIKATIONER

I (V) och (VI) behandlas kriterierelaterade prov. I (V) ges en allmän bakgrund till denna provtyp och olika definitioner presenteras. Vidare behandlar samma rapport (V) egenskaper och begränsningar hos kriterierelaterade prov. I den andra rapporten (VI) presenteras en empirisk studie, där olika tekniker för utvärdering av kriterierelaterade prov studerats. I denna studie ges också en sammanfattning och analys av forskningsläget vad beträffar den statistiska behandlingen av data från kriterierelaterade mätningar.

Kriterierelaterade prov: bakgrund, egenskaper och begränsningar

De s k kriterierelaterade proven (Glaser, 1963) har tillkommit för att ge besked om vad eleverna kan och har sitt ursprung i den kritik som riktats mot att i undervisningssammanhang använda prov konstruerade för att differentiera mellan individer (se t ex Nitko, 1970; Glaser & Nitko, 1971).

Innebörden i kriterierelaterade prov är ännu oklar, men gemensamt för många definitioner är att uppgifterna skall representera preciserade mål. I denna fråga finns emellertid ett grundläggande problem. Det innebär stora svårigheter att entydigt bryta ned komplexa mål (se t ex Eisner, 1969; Ward, 1970; Ebel, 1972). Om vissa mål inte kan beskrivas i entydiga termer, så innebär det en begränsad möjlighet att använda sig av kriterierelaterade prov på ett meningsfullt sätt.

Två andra utmärkande problem i anslutning till kriterierelaterade prov utgörs av provens homogenitet och fastställande av kravnivåer. Beträffande provens homogenitet varierar uppfattningarna högst avsevärt. Enligt vissa uppfattningar (Davis, 1970; Kriewall, 1972) skall de kriterierelaterade proven mäta ett fåtal och långt preciserade mål. Följden av detta synsätt blir att ett stort antal prov måste konstrueras för till och med relativt små avsnitt och moment.

Andra är mer "liberala" i sina krav på homogeniteten i kriterierelaterade prov och menar att dessa kan mäta mer än ett preciserat mål (Popham & Husek, 1969; Cox, 1971; Fremer, 1972).

Gemensamt för kriterierelaterade prov är att de skall ge information om huruvida eleverna har uppnått vissa preciserade mål. För detta krävs en fixerad kravnivå. Olika förslag har redovisats angående hur dessa skall fastställas. De varierar från mer eller mindre intuitiva procentgränser (Lindvall & Cox, 1970) till gränser baserade på olika statistiska resonemang (Emrick, 1971; Fhanér, 1972). Gemensamt för dem alla är emellertid att de på ett eller annat sätt är grundade på värderingar, vars konsekvenser på undervisningssituationen kan vara olika. Ett närmare studium av problemet med att fastställa kravnivåer är av bl a den anledningen ytterst angeläget (se också Hambleton & Gorth, 1971; Airasian & Madaus, 1972).

I (V) berörs också kortfattat uppgiftskonstruktion och användningsområden av kriterierelaterade prov. Beträffande uppgiftskonstruktion har den sedan lång tid tillbaka baserats på vaga och intuitiva "regler" (Nitko, 1970). "Item-form-tekniken" (Osburn, 1968) innebär emellertid en konstruktionsteknisk förändring. Med den kan man automatiskt generera uppgifter från i förväg preciserade regler. Ännu så länge har den emellertid prövats i mycket

begränsad omfattning (se Hively et al., 1968; Fremer & Anastasio, 1969; Ferguson, 1971).

Man kan sammanfattningsvis notera att de krav som från många håll framförts på kriterierelaterade mätningar i undervisningssammanhang är förenade med en rad problem. För att lösa dessa fordras i fortsättningen stora forskningsinsatser.

Reliabilitets-, validitets- och diskriminationsmått för kriterierelaterade prov

En viktig skillnad mellan norm- och kriterierelaterade mätningar består i att man i det senare fallet inte kan utgå från att man kommer att erhålla variation mellan individer i resultaten (Popham & Husek, 1969). Det kan man däremot vid normrelaterade mätningar, eftersom man där redan i konstruktionsfasen förutsätter stabila interindividuella variationer i den egenskap provet skall mäta.

På grund av att man inte kan utgå från interindividuella variationer i resultaten vid kriterierelaterade mätningar kan heller inte de vanliga reliabilitets-, validitets- och diskriminationsmåtten tillämpas i dessa sammanhang. På senare tid har emellertid analoga mått utvecklats för att passa ett kriterierelaterat mätförfarande.

Beträffande validiteten i KRP, så är det främst innehållsvaliditeten som är av intresse (Popham & Husek, 1969; Hambleton, 1972). Skattningar av kongruensen mellan mål och uppgifter är ett sätt att mäta innehållsvaliditeten (Dahl, 1971). Ett annat, indirekt sätt, är att jämföra resultaten från för- och eftermätning. Ozenne (1971) har i variansanalytiska termer utvecklat ett mått, ε, som ger ett numeriskt uttryck för skillnaden mellan för- och eftermätning.

Kring reliabilitetsaspekten har ett stort intresse knutits. Flera av de mått som härvidlag har presenterats representerar motsvarigheter till mått baserade på den klassiska reliabilitetsteorin. Två egenskaper är speciellt utmärkande för många av dessa reliabilitetsmått. Det ena är det beslutsorienterade synsättet, det andra är skalnivån.

Ivens' (1970) förslag innebär att reliabiliteten mäts i termer av proportionen överensstämmande poäng mellan två mätningar. Hambleton (1972) däremot menar att reliabiliteten skall uttryckas i proportionen individer som vid två mättillfällen kommer över eller under en fixerad kriteriegräns. Livingstons (1972) mått, k^2_{Tx}, är en utveckling av Hambletons (1972) mått i det att hänsyn tas till hur långt över eller under en fixerad kriteriegräns individen hamnar. I själva verket är k^2_{Tx} en vanlig produktmomentkorrelationskoefficient där avvikelsepoängen definieras i relation till en fixerad kriteriegräns istället för i relation till medelvärdet i fördelningen av erhållna poäng.

Flera olika diskriminationsindex har också presenterats. Till de mest intressanta och också de som oftast använts hör (a) differensen uttryckt i p mellan för- och eftermätning (Cox & Vargas, 1966), (b) differensen uttryckt i p mellan expert- och icke-expertgrupp (Levin & Marton, 1971; Marton, 1973) och (c) Pophams (1971) chi^2, beräknade på resultaten från för- och eftermätning.

I den explorativa studie som redovisas i (VI) var syftet i första hand att pröva och jämföra olika diskriminationsindex för KRP och hur urvalet av uppgifter med ledning av dessa påverkade reliabiliteten och validiteten. Den huvudsakliga jämförelsen av diskriminationsmått avsåg (a) och (b) ovan och det normrelaterade differensmåttet. Det senare definierades som differensen uttryckt i p mellan en

hög- och låggrupp på resultaten från en eftermätning. Som reliabilitetsmått användes de av Ivens (1970), Hambleton (1972) och Livingston (1972) föreslagna måtten (se ovan). Validiteten mättes i termer av Ozennes (1971) sensitivitetsindex, ε, (se ovan). Prövningarna företogs på 285 elever i årskurs åtta på grundskolans högstadium och på 20 ämneslärarkandidater. De senare utgjorde expertgrupp.

Resultaten indikerade tydliga skillnader mellan norm- och kriterierelaterade diskriminationsmått. Det indikerar att ett urval av uppgifter enligt dessa två typer av diskriminationsmått leder till olika provsammansättningar.

Å andra sidan var skillnaderna också stora beträffande de två ovan nämnda kriterierelaterade diskriminationsmåtten. Det resultatet är speciellt intressant eftersom dessa mått framförts som alternativa mått i kriterierelaterade sammanhang.

Om man också i andra studier kan visa att sambandet mellan dessa två mått är lågt måste nästa steg bli att närmare undersöka vilken typ av uppgifter som de två måtten premierar för att därefter se vilken effekt en utmönstring eller revidering av uppgifter kan tänkas ha med avseende på de uppställda målen. I förlängningen av detta resonemang kommer också relationen mellan kort- och långsiktiga mål in i bilden (V, se också Marton, 1973).

Oavsett vilket diskriminationsmått urvalet av uppgifter baserades på, så påverkades reliabiliteten, definierad i termer av överensstämmande resultat vid två mättillfällen, positivt. Effekterna var dock relativt små. Vidare antyder resultaten att en höjning av validiteten, definierad som skillnaden mellan för- och eftermätning, inte nödvändigtvis behöver innebära att besluten om individernas placeringar ovan och under en fixerad kriteriegräns blir säkrare.

SLUTORD

För att på ett meningsfullt sätt kunna använda prov krävs att dessa ger både tillförlitlig och relevant information. Den testteoretiska forskningens syfte är att direkt eller indirekt ge underlag för att erhålla sådan information.

Beträffande ovan diskuterade analytiska angreppssätt på testteoretiska problem utifrån informationen i inter-item-kovariansmatrisen kan man notera att det på ett mycket fruktbart sätt möjliggjort många klargöranden och förenklingar. Inte minst ur praktisk synvinkel syns detta analysförfarande innebära många fördelar. Närmast till hands i det avseendet ligger en förbättrad undervisning på området och ett utvecklande av mer flexibla databehandlingsrutiner.

Differentiell poängsättning av flervalsfrågors svarsalternativ representerar ett mera direkt sätt att försöka förbättra kvaliteten i prov. Vad som i det sammanhanget är särskilt intressant att observera är den avgörande inverkan svarsalternativens utformning tycks ha på reliabiliteten och validiteten. Om proven utformas med graderade svarsalternativ tycks ur reliabilitetssynpunkt ett differentiellt poängsättningsförfarande vara att föredra framför ett dikotomt. Omvändningen syns emellertid gälla om validiteten tas som utgångspunkt för jämförelsen. Denna faktor kan sannolikt också delvis förklara de varierande resultat som erhållits i tidigare undersökningar med differentiell poängsättning. För att närmare utreda effekterna av differentiell poängsättning fordras att "graderingsfaktorn" ingående studeras.

Fortsatta studier av differentiell poängsättning av flervalsfrågors svarsalternativ är inte enbart av mättekniskt intresse. Också ur pedagogisk synvinkel är detta angeläget. Av kritiken mot flervalsfrågor att döma finns det

anledning tro att systematiskt utformade svarsalternativ och en förfinad poängsättning skulle medföra en mer positiv inställning till denna frågetyp.

Tillkomsten av de på senare tid mycket omdiskuterade kriterierelaterade proven ställer andra krav på utvärderingsförfarande än vad de normrelaterade proven gör. Den normrelaterade provfilosofins utgångspunkt i stabila interindividuella differenser har inte sin motsvarighet i den kriterierelaterade provfilosofin.

För att tillgodose de krav som ett kriterierelaterat provförfarande implicerar måste ökade forskningsinsatser kanaliseras till hur dessa prov skall konstrueras och utformas. Härvidlag är det av central betydelse att närmare undersöka (a) vilka effekter olika grader av målprecisering får på möjligheterna att dra slutsatser om vad individerna kan, (b) på vilket sätt provens homogenitet påverkar tolkningsmöjligheterna och (c) hur kravgränserna skall fastställas.

De forskningsinsatser som på senare tid ägnats de kriterierelaterade proven och hur dessa skall utvärderas har resulterat i flera intressanta tekniker och angreppssätt för att bedöma och förbättra kvaliteten i kriterierelaterade prov. Många av dessa baseras emellertid på olika förutsättningar och får olika konsekvenser på provens utformning. Innebörden av detta måste närmare utredas.

Mot bakgrund av de resultat som erhölls i ovan diskuterade undersökning (VI) är det särskilt angeläget att under andra betingelser studera och jämföra reliabilitets- och validitetseffekter av uppgiftsurval med ledning av differenserna uttryckta i proportioner mellan (a) resultaten på en för- och eftermätning för en och samma grupp och (b) resultaten från en expertgrupp och icke-expertgrupp. I det senare avseendet bör olika expertgrupper användas.

Detta arbete har genomförts med medel från Statens Råd för Samhällsforskning.

REFERENSER

Airasian, P. W., & Madaus, G. F. Criterion-Referenced Testing in the Classroom. NCME Reports, 1972, 3(4).

Cox, R. C. Evaluative Aspects of Criterion-Referenced Measures. In W. J. Popham (ed.): Criterion-Referenced Measurements. Englewood Cliffs, New Jersey: Educational Technology Publications, 1971.

Cox, R. C., & Vargas, J. S. A Comparison of Item Selection Techniques for Norm-Referenced and Criterion-Referenced Tests. Paper read at the Annual Meeting of the National Council on Measurement in Education, Chicago, Illinois, 1966.

Cronbach, L. J. Coefficient Alpha and the Internal Structure of Tests. Psychometrika, 1951, 16, 297-334.

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. Theory of Generalizability: A Liberalization of Reliability Theory. The British Journal of Statistical Psychology, 1963, XVI, 137-163.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles. New York: John Wiley & Sons, 1972.

Cureton, E. E. Validity, Reliability, and Baloney. Educational and Psychological Measurement, 1950, 10, 94-96.

Cureton, E. E. Corrected Item-Test Correlations. Psychometrika, 1966, 31, 93-96.

Dahl, T. Toward an Evaluative Methodology for Criterion-Referenced Measures: Objective-Item Congruence. CSE Report. Center for the Study of Evaluation, Los Angeles, 1971, No 15.

Davis, F. B. Criterion-Referenced Tests. In Educational Records Bureau: Testing in Turmoil: A Conference on Problems and Issues in Educational Measurement. The Thirty-Fifth Annual Conference of Educational Records Bureau, New York, 1970.

Davis, F. B., & Fifer, G. The Effect on Test Reliability and Validity of Scoring Aptitude and Achievement Test with Weights for every Choice. Educational and Psychological Measurement, 1959, 19, 159-170.

Ebel, R. Some Limitations of Criterion-Referenced Measurement. In K. D. Hopkins, & J. C. Stanley (eds.): Perspectives in Educational and Psychological Measurement. Englewood Cliffs, New Jersey: Prentice-Hall, 1972.

Edwards, A. L. Techniques of Attitude Scale Construction. New York: Appleton-Century-Crofts, 1957.

Eisner, E. W. Instructional and Expressive Educational Objectives: Their Formulation and Use. In W. J. Popham, E. W. Eisner, H. J. Sullivan, & L. L. Tyler (eds.): Instructional Objectives. AERA Monograph 3. Chicago: Rand McNally, 1969.

Emrick, J. A. An Evaluation Model for Mastery Testing. Journal of Educational Measurement, 1971, 8, 321-326.

Ferguson, R. L. Computer Assistance for Individualizing Measurement. Learning Research and Development Center, University of Pittsburgh, 1971, No 8.

Fhanér, S. Estimation and Decision Making in Achievement Testing: An Item Sampling Model. Göteborg Psychological Reports, 1972, 2, No 15.

Fremer, J. Criterion-Referenced Interpretations of Survey Achievement Tests. Test Development Memorandum, 1972. TDM-72-1. Princeton, New Jersey, Educational Testing Service.

Fremer, J., & Anastasio, E. J. Computer-Assisted Item Writing - I. (Spelling Items). Journal of Educational Measurement, 1969, 6, 69-74.

Gardner, P. L. Test Length and the Standard Error of Measurement. Journal of Educational Measurement, 1970, 7, 271-273.

Glaser, R. Instructional Technology and the Measurement of Learning Outcomes: Some Questions. American Psychologist, 1963, 18, 519-521.

Glaser, R., & Nitko, A. J. Measurement in Learning and Instruction. In R. L. Thorndike (ed.): Educational Measurement. Second Edition. Washington: American Council on Education, 1971.

Guilford, J. P. Psychometric Methods. New York: McGraw-Hill, 1954.

Gulliksen, H. Theory of Mental Tests. Sixth Printing. New York: John Wiley & Sons, 1967.

Guttman, L. The Quantification of a Class of Attributes: A Theory and Method of Scale Construction. In Paul Horst (ed.): The Prediction of Personal Adjustment. New York: Social Science Research Council, 1941.

Guttman, L., & Schlesinger, I. M. Systematic Construction of Distractors for Ability and Achievement Test Items. Educational and Psychological Measurement, 1967, 27, 569-580.

Hambleton, R. K. Towards a Theory of Criterion-Referenced Tests. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, 1972.

Hambleton, R. K., & Gorth, W. P. Criterion-Referenced Testing: Issues and Applications. Technical Reports, 1971, No 13. School of Education, University of Massachusetts, Amhurst.

Hendrickson, G. F. An Assessment of the Effect of Differentially Weighting Options of a Multiple-Choice Objective Test Using a Guttman Weighting Scheme. The John Hopkins University: Center for Social Organization of Schools, Working Paper Number 6, 1971.

Hendrickson, G. F., & Green, B. F., Jr. Comparison of the Factor Structure of Guttman-Weighted vs. Rights-Only-Weighted Tests. Paper presented at the annual meeting of the AERA, Chicago, Illinois, 1972.

Henrysson, S. Correction of Item-Total Correlation in Item Analysis. Psychometrika, 1963, 28, 211-218.

Hively II, W., Patterson, H. L., & Page, S. H. A "Universe-Defined" System of Arithmetic Achievement Tests. Journal of Educational Measurement, 1968, 5, 275-290.

Ivens, S. H. An Interpretation of Item Analysis, Reliability, and Validity in Relation to Criterion-Referenced Tests. Unpublished doctoral dissertation. Florida State University, 1970.

Kriewall, T. E. Aspects and Applications of Criterion-Referenced Tests. Technical paper No 103. A paper presented at the annual meeting of the American Educational Research Association, Chicago, Illinois, 1972.

Kuder, G. F., & Richardson, M. W. The Theory of the Estimation of Test Reliability. In W. A. Mehrens, & R. L. Ebel (eds.): Principles of Educational and Psychological Measurement. Chicago: Rand McNally, 1967.

Levin, L., & Marton F. Provteori och provkonstruktion. Stockholm: Almqvist & Wiksell, 1971.

Lindvall, C. M., & Cox, R. C. The IPI Evaluation Program. AERA Monograph 5. Chicago: Rand McNally, 1970.

Livingston, S. A. Criterion-Referenced Applications of Classical Test Theory. Journal of Educational Measurement, 1972, 9, 13-26.

Lord, F. M. Do Test of Same Length have the Same Standard Errors of Measurement? Educational and Psychological Measurement, 1957, XVII, 510-521.

Lord, F. M. Tests of the Same Length do have the Same Standard Error of Measurement. Educational and Psychological Measurement, 1959, XIX, 233-239.

Lord, F. M., & Novick, M. R. Statistical Theories of Mental Test Scores. Reading, Massachusetts: Addison-Wesley, 1968.

Marton, F. Evalueringsteori och metodik. I G. Handal, L-G. Holmström och O. B. Thomson (red.): Universitetsundervisning. Malmö: Studentlitteratur, 1973.

McLean, A. G., & Tait, A. T. A Procedure for Analyzing a Test and Maximizing its Reliability. Journal of Experimental Education, 1954, 22, 273-278.

Nitko, A. J. Criterion-Referenced Testing in the Context of Instruction. Paper presented at the Educational Records Bureau - National Council on Measurement in Education Symposium, "Criterion-Referenced Measures: Pros and Cons", New York, 1970.

Osburn, H. G. Item Sampling for Achievement Testing. Educational and Psychological Measurement, 1968, 28, 95-104.

Ozenne, D. G. Toward an Evaluative Methodology for Criterion-Referenced Measures: Test Sensitivity. CSE Report. Center for the Study of Evaluation, Los Angeles, 1971, No 72.

Popham, W. J. Indices of Adequasy for Criterion-Referenced Test Items. In W. J. Popham (ed.): Criterion-Referenced Measurement. Englewood Cliffs, New Jersey: Educational Technology Publications, 1971.

Popham, W. J., & Husek, T. R. Implications of Criterion-Referenced Measurement. Journal of Educational Measurement, 1969, 6, 1-9.

Ramsey, J. O. A Scoring System for Multiple-Choice Test Items. The British Journal of Mathematical and Statistical Psychology, 1968, 21, 247-250.

Reilly, R. R., & Jackson, R. Effects of Item Weighting on Validity and Reliability of Shortened Forms of the GRE Aptitude Tests. Paper presented at the annual meeting of the AERA, Chicago, Illinois, 1972.

Stanley, J. C., & Wang, M. D. Differential Weighting. A Survey of Methods and Empirical Studies. New York: College Entrance Examination Board, 1968.

Swineford, F. Note on "Test of the Same Length do have the Same Standard Error of Measurement". Educational and Psychological Measurement, 1959, XIX, 241-242.

Tryon, R. C. Reliability and Behavior Domain Validity: Reformulation and Historical Critique. Psychological Bulletin, 1957, 54, 229-249.

Ward, J. On the Concept of Criterion-Referenced Measurement. British Journal of Educational Psychology, 1970, 40, 314-323.

Wolf, R. Evaluation of Several Formulae for Correction of Item-Total Correlations in Item-Analysis. Journal of Educational Measurement, 1967, 4, 21-26.

Zubin, J. The Method of Internal Consistency for Selecting Test Items. Journal of Educational Psychology, 1934, 25, 345-356.