ISSN 1012-277X S.Afr.j.ind.eng.

S A Tydskrif vir Bedryfsingenieurswese, Vol 7, No 1, Junie 1993, pp.12-28

METING VAN DIE BETROUBAARHEID VAN DIE EVALUERING VAN BEDRYFSINGENIEURSWESE STUDENTE

DC Page
Departement Bedryfsingenieurswese
Universiteit van Stellenbosch
7600 STELLENBOSCH

ABSTRACT

Reliability of evaluation refers to the repeatibility of results when similar tests are repeated by the same student, or a similar evaluation procedure is applied. By measuring reliability an objective, quantitative measure of the quality of an evaluation is obtained. Poor formulation of tests can be diagnosed, and problems in the evaluation system identified. The subject is dicussed with reference to an analysis of performance in Industrial Engineering subjects at the University of Stellenbosch.

OPSOMMING

Betroubaarheid van evaluering verwys na die mate van herhaalbaarheid van resultate as soortgelyke toetse deur dieselfde student herhaal word, of 'n soortgelyke evalueringsprosedure toegepas word. Deur betroubaarheid te meet word 'n objektiewe syfermaatstaf van die kwaliteit van die evaluering verkry. Swak eienskappe van toetse kan gediagnoseer-, en probleme met die evalueringstelsel geidentifiseer word. Die onderwerp word bespreek aan die hand van 'n analise van die prestasie in Bedryfsingenieurswese vakke aan die Universiteit van Stellenbosch.

1 INLEIDING

Vervolgens word betroubaarheid eerstens gedefinieer, met besondere verwysing na die oorsprong van die betroubaarheidskoeffisient, wat dien as syfermaatstaf van betroubaarheid. Uit die model van prestasie, wat volledig behandel word in paragraaf 3, word die neiging om studente wat 'n ondergemiddelde punt in 'n toets verwerf, te onderevalueer, en die wat bo die gemiddelde presteer, te oorevalueer, geillustreer.

Die berekening van die betroubaarheid van individuele toetse, en van geweegde saamgestelde vakprestasiepunte, word behandel. Vergelykings vir optimale gewigte, wat die betroubaarheid van die saamgestelde prestasiepunt vir 'n vak sal maksimeer, word ook afgelei. Daar word in 'n afsonderlike paragraaf getoon hoekom die korrelasiekoeffisiente tussen elke paar evaluerings in 'n vak, of vrae in 'n vraestel, die sleutelveranderlike is wat betroubaarheid bepaal.

'n Paragraaf word ook gewy aan die nadele van sg. spoed- vs. kragtoetse. 'n Toets behoort nl. opgestel te word sodat ongeveer 90 persent van die klas genoeg tyd het om alle vrae te probeer beantwoord. Die belangrikste bevindings en voorstelle ter verhoging van betroubaarheid word opgesom in die finale paragraaf.

Praktiese resultate word aangehaal soos toepaslik. Dit volg uit 'n analise van 'n verteenwoordigende steekproef van bedryfsingenieurswese vakke aan die Universiteit van Stellenbosch vir die tweede semester 1991, en die eerste semester 1992. 'n Lotus toepassing is geskryf om eerstens die berekening van die betroubaarheid van individuele evaluerings te doen, en tweedens die betroubaarheid van die evaluering vir 'n vak te bereken.

Let daarop dat die gevolgtrekkings hier gemaak gegrond is op beperkte gegewens en dat veralgemening versigtig hanteer moet word. Die gedrag van die stelsel self word ongetwyfeld beinvloed deur die meting en meetprosedure, sodat die resultate onder 'n bepaalde evalueringsprosedure verkry nie sondermeer gebruik kan word om die resultate met 'n ander evalueringsprosedure te voorspel nie. Tweedens lewer die klein klasgroottes in meeste van die vakke uit 'n statistiese oogpunt probleme (dit het gewissel van 10 tot 97). Dit is egter belangrik om te begryp dat 'n klas die totale populasie is en dat die analise gaan om die die herhaalde meting van dieselfde student. Die klasgrootte is 'n gegewe en maak nie die statistiek ongeldig nie, dit verhoog net die variasie, of onsekerheid, dan.

2 BETROUBAARHEID EN GELDIGHEID

Dit is nodig om te onderskei tussen validiteit, of geldigheid, en betroubaarheid.

Die geldigheid van die inhoud van 'n evaluering gaan om die mate waartoe 'n evaluering, sê 'n toets, meet wat dit veronderstel is om te meet. Beteken 'n hoë toetspunt dat die student die vak goed magtig is, en 'n lae toetspunt onbevredigende prestasie t.o.v. die doelwitte met die vak? Of meet die toets dalk suiwer intellektuele vermoeë, met toevallige korrelasie tussen die toetspunt en die verlangde kennis, begrip, vaardighede, analise, sintese en evaluering wat in die besondere vak verwag word. Die geldigheid van 'n evaluering is die akkuraatheid waarmee gespesifiseerde gevolgtrekkings gemaak kan word op grond van die punte in die evaluering verwerf.

Die betroubaarheid van 'n meetstelsel is 'n skatter van die bestendigheid waarmee herhaalde metings gedoen kan word. Die presiesheid van die meetstelsel word beraam. In terme van puntetoekenning is betroubaarheid die bestendigheid waarmee 'n evaluering of toets differensieer tussen die prestasie van studente. (Thomas, 1986.) Die betroubaarheid (presisie) van meting gaan om die spreiding van resultate om die gemiddelde, terwyl geldigheid (akkuraatheid) gaan om die afwyking van die gemiddelde van die teikenwaarde.

Met hierdie ondersoek is gekonsentreer op betroubaarheid en geldigheidsoorwegings word nie verder bespreek nie.

3 DIE BETROUBAARHEIDSKOEFFISIENT

In die analise van die betroubaarheid van 'n evaluasie word die volgende basiese model gebruik:

$$x_i = t_i + e_i \quad \text{, waar} \qquad (1)$$

x_i = Waargenome punt wat die i-de student verwerf in 'n enkele toets (evaluering).

x_i kom uit die normaalverdeling $n(\mu,\sigma_x^2)$, waar

μ = Verwagte waarde van die waargenome punte van alle studente in die klas.

σ_x^2 = Ooreenstemmende variansie.

t_i = Onbekende "werklike" punt (universumpunt) wat die gemiddelde punt is wat die i-de student sou kry deur oneindig kere getoets te word op 'n oneindige aantal paralelle ekwivalente toetse.

t_i Kom uit die normaalverdeling $n(\mu,\sigma_t^2)$, met

μ = dieselfde verwagte waarde as x_i
(omdat die verwagte waarde van e_i = o).

σ_t^2 = Verwagte variansie van "werklike" punte van die klas.

Die "werklike" punt, t_i, word grotendeels deur twee faktore bepaal. Eerstens die vermoë van die i-de student, tweedens die kennis, of leereffek a.g.v. die i-de student se voorbereiding vir die evaluering onder beskouing.

e_i = Meetfout. Afwyking tussen die waargenome punt in die enkele toets en die "werklike" punt van die i-de student. Ontstaan hoofsaaklik a.g.v. meetfoute in die toets self. Dubbelsinnige vrae, ens. beteken dat die toets nie bestendige resultate lewer nie. Die student het egter ook 'n eie veranderlike prestasie in opeenvolgende toetssituasies waaroor die toetser nie beheer het nie, en hierdie veranderlikheid is verstrengel met die meetfout. Dit beteken prakties dat al is toetse perfek sal e_i nie nul wees nie.

e_i kom uit die normaalverdeling $n(0,\sigma_e^2)$, met

verwagte waarde = 0,

σ_e^2 = verwagte toevalsvariansie van die klas a.g.v. meetfoute, ens., met 'n enkele toets.

Uit die model (1) volg:

$$\sigma_x^2 = \sigma_t^2 + \sigma_e^2 \qquad (2)$$

Die betroubaarheidskoeffisient, die persentasie van die totale variasie in waargenome prestasie wat verklaar word deur die "werklike" verskil tussen studente, word nou gedefinieer as

$$\alpha = \sigma_t^2/\sigma_x^2 \qquad (3)$$

(2): $$= (\sigma_x^2 - \sigma_e^2)/\sigma_x^2 \qquad (4)$$

$$= 1 - \sigma_e^2/\sigma_x^2 \qquad (5)$$

ook volg $\sigma_e = \sigma_x\sqrt{(1-\alpha)}$ = standaardramingsfout (6)

Uit (6) kan betroubaarheidintervalle en ander waarskynlikhede vir waargenome punte bereken word.

Met die betroubaarheid (α), die gemiddelde ($\bar{x}$) en standaard afwyking (s_x) van 'n bepaalde toets bekend kan 'n skatter van die i-de student se "werklike" punt (t_i) as volg bereken word:

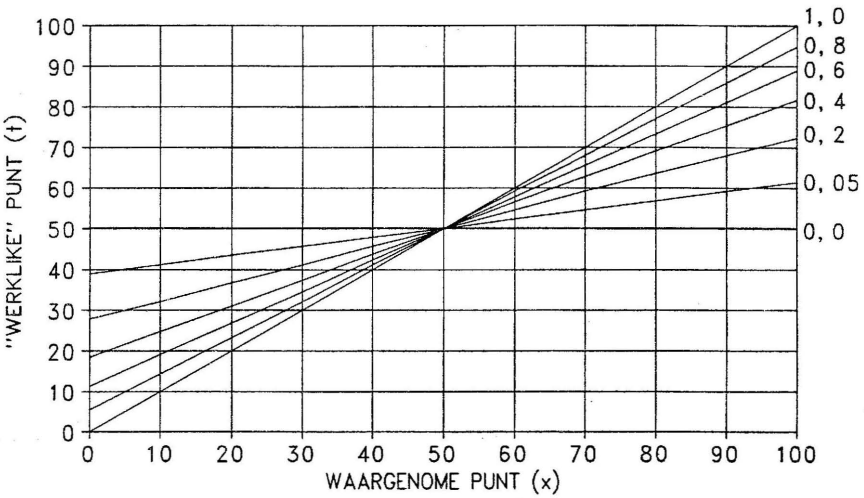
$$t_i = x + \sqrt{\alpha}(x_i - x) \qquad (7)$$

Die bedoeling is nie dat punte wat studente in toetse verwerf aangepas behoort te word na skatters van die "werklike" universum punte nie, maar 'n belangrike verskynsel, die sg. "regressie na die gemiddelde", word geillustreer. Met 'n enkel evaluering word die student met 'n punt onder die gemiddelde, se prestasie nl. gemiddeld onderskat, en die bo die gemiddelde oorskat, eweredig aan $\sqrt{\alpha}$ van die afwyking van die waargenome punt van die gemiddelde.

Dit is die maklikste om te begryp met verwysing na 'n toets met 'n betroubaarheid van nul ($\alpha = 0$). In die geval behoort alle studente dieselfde "werklike" universum punt $t_i = \bar{x}$ te kry omdat die variansie van die toetsresultate (σ_x^2) gelyk aan die foutevariansie (σ_e^2) is, en die variansie van die werklike punt (σ_t^2) dus gelyk aan nul moet wees (uit (2)). Enige student wat nou in die toets 'n punt onder die gemiddelde verwerf (wat dus suiwer toevallig is) se prestasie word dus onderskat omdat dit gelyk aan die gemiddelde behoort te wees. Om dieselfde rede word alle studente wat 'n punt bo die gemiddelde verwerf se prestasie oorskat. By 'n betroubaarheid van 1,0 (die foute variansie σ_e^2 is nou nul) geld die ander uiterste, nl. dat studente se werklike punte gelyk is aan die waargenome punt in die enkele toets. Soos reeds vermeld is dit onmoontlik en sal die betroubaarheid altyd kleiner as 1,0 wees, en dus tussen hierdie twee uiterstes val. Hoe minder betroubaar die evaluering is, hoe groter is die verwagte onder- of oor-evaluering dus. Hierdie verskynsel word verder geillustreer in fig. 3.1 waar die "werklike" prestasie teenoor die waargenome punt in 'n enkeltoets gestip word vir verskillende waardes van α (uit (7)).

'n Tweede belangrike verskynsel wat hierdeur geillustreer word is dat die standaardafwyking van waargenome punte altyd groter is as die "werklike" standaardafwyking. Hoe minder betroubaar 'n toets is, hoe groter is die standaardafwyking van die waargenome punt. Met gradering, waar simbole toegeken word op grond van persentiele bereken in ooreenstemming met die gemiddelde en standaardafwyking van die evaluering, maak dit nie saak nie. In 'n situasie waar waargenome punte sonder enige manipulering gebruik word om saamgestelde prestasie te bereken het dit ernstige gevolge. Veronderstel byvoorbeeld die standaardafwyking is baie groot a.g.v. 'n onbetroubare toets. Die prestasie van die swak student, wie se punt juis baie afwyk van die gemiddelde, mag nou in so 'n mate onderskat word dat hy moontlik onregverdiglik mag druip. Terselfdertyd kom onverdiende uitstaande prestasie voor.

Fig 3.1 – "Werklike"– vs. waargenome punt vir verskillende betroubaarhede



4 METING VAN BETROUBAARHEID

4.1 BETROUBAARHEID VAN INDIVIDUELE EVALUERINGS

4.1.1 Teoretiese agtergrond

Die betroubaarheidskoeffisient (α) kan op verskeie wyses bepaal word vir 'n bepaalde toets. Die oorkoepelende metodiek is reeds in 1937 deur Kuder en Richardson (1937) afgelei, maar die basiese "gesplete halwes" Spearman-Brown metode dateer volgens Cronbach (1951) reeds uit 1910! Slegs die beginsels waarop die metodes en formules gebaseer is word hier bespreek aangesien die detailafleiding daarvan omslagtig is en maklik in die oorspronklike bronne nageslaan kan word. Dieselfde formules is ook oor die jare vanaf verskillende uitgangspunte en met ver skillende aannames deur verskillende skrywers afgelei.

Die volgende eenvoudige formule word algemeen gebruik om die betroubaarheidskoeffisient (α) te bereken.

$$\alpha = [n/(n-1)](1 - \sum_{i=1}^{n} s_i^2/s_x^2) \quad , \text{waar} \tag{8}$$

n = aantal vrae in die toets
s_i^2 = variansie van die klas se punte vir die i-de van die n toetsvrae
s_x^2 = variansie van die totale toetspunt

Hierdie formule volg uit die sg. "gesplete halwes" metode, waar die resultate van 'n toets bloot in twee ewegroot helftes verdeel word en die kwadraat van die korrelasiekoeffisient (r^2) tussen die prestasie in die twee helftes vir die groep studente wat die vraestel geskryf het, aangepas vir die lengte van die toets, dan dien as skatter van α. (Dit is aanvanklik slegs vir veelvoudige keuse tipe toetse met ongeweegde 0 of 1 antwoorde, wat maklik verdeel kan word, ontwikkel.)

$$\alpha = 2r^2/(1+r^2) \tag{9}$$

Die totale toets is naamlik tweemaal so lank as die twee helftes, en vandaar (9) wat volg uit die volgende algemene "Spearman-Brown" formule vir α vir 'n toets wat L maal so lank is as die een waaruit r^2 bepaal is (Mosier, 1943).

$$\alpha = Lr^2/[1+(L-1)r^2] \tag{10}$$

Die probleem met die gesplete helftes metode is dat r^2 afhanklik is van hoe die toets gesplit word. Prakties sal elke ander verdeling in twee helftes 'n ander waarde vir r^2 gee. Die gemiddelde van die korrelasiekoeffisiente van elk van die moontlike maniere waarop 'n toets gesplit kan word, word deur die algemene formule (8) hierbo genoteer gegee.

Die bekende KR20 formule (11) vir veelvoudige keuse vrae is bloot (8) met die resultaat vir elke item binomiaalverdeel met verwagtingswaarde p_i en variansie $p_i(1-p_i)$. p_i Is die fraksie studente wat vraag i korrek beantwoord het.

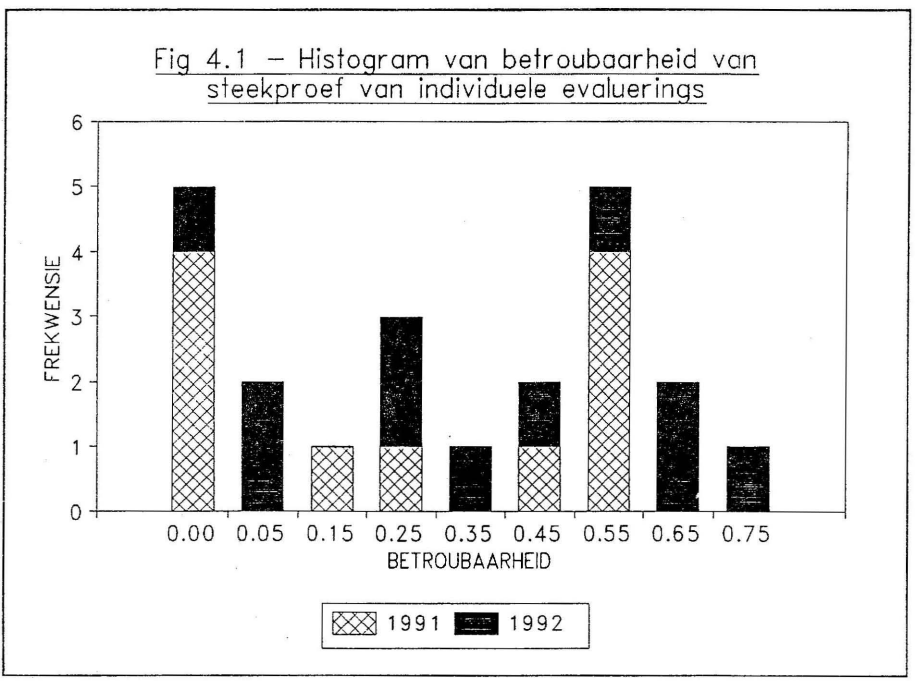
$$\alpha = [n/(n-1)](1-[\Sigma p_i(1-p_i)]/s_x^2) \tag{11}$$

4.1.2 Praktiese resultate

Die volgende resultate is verkry met die meting van betroubaarhede van 'n steekproef van derde- en finalejaars- bedryfsingenieurswese vakke aan die Universiteit van Stellenbosch. Die resultate is in twee groepe verdeel, nl. tweede semester 1991 en eerste semester 1992, en word opgesom in tabel 4.1. Histogramme word getoon in fig. 4.1.

Tabel 4.1 - Betroubaarhede van individuele evaluerings in 'n steekproef van Bedryfsingenieursvakke aan die Universiteit van Stellenbosch

PERIODE	BETROUBAARHEID		
	GEMIDDELDE	STANDAARDAFW.	AANTAL
2de semester 1991	0,28	0,25	11
1ste semester 1992	0,37	0,28	11
TOTAAL	0,33	0,26	22



Die betroubaarhede verkry het gewissel van negatief tot positief (met negatiewe korrelasie tussen die punte in individuele vrae verwerf is dit moontlik om numeries negatiewe betroubaarheid te verkry). Betroubaarheid is egter gelyk aan 0 gestel wanner dit negatief bereken, omdat 'n negatiewe waarde niksseggend is.

Alhoewel nie werklik statisties beduidend nie (die gemiddelde betroubaarheid neem beduidend toe op 'n ongeveer 75 persent vertroupeil), is daar tog 'n merkbare verbetering, wat gevolg het uit die rapportering van resultate vir 1991 aan dosente

voor aanvang van die 1992 kursusse. Die elementêre voorstelle aan dosente gerig ter verbetering van formulering van toetse hang primêr om die gebruik van krag-, eerder as spoedtoetse, soos in paragrawe 2.5 en 5 bespreek.

4.2 BETROUBAARHEID VAN GEWEEGDE, SAAMGESTELDE PUNT

4.2.1 Teoretiese agtergrond

Die vraag ontstaan wat die betroubaarheid is van die finale prestasiepunt wat aan 'n student in 'n vak toegeken word, wat gebaseer word op 'n geweegde samestelling van meer as een toets- en/of ander evalueringspunt.

Veronderstel die saamgestelde prestasiepunt vir die i-de student word bereken deur

$$X_i = W_1x_{i1} + W_2x_{i2} + \ldots W_vx_{iv} + \ldots W_nx_{in} \qquad (12)$$

, waar X_i = saamgestelde prestasiepunt van die i-de student
W_v = gewig van die v-de toets/evaluering,
x_{iv} = punt deur die i-de student in die v-de toets/evaluering verwerf.

Soortgelyk aan (5) word die betroubaarheid van die totale prestasiepunt, α_x , gegee deur (13):

$$\alpha_x = 1 - s_{ex}^2/s_x^2 \quad , \text{ waar} \qquad (13)$$

s_{ex}^2 = Foutvariansie van die saamgestelde punt X
s_x^2 = Totale variansie van X

Omdat die foutvariansies van die evaluerings onafhanklik is (nie korreleer nie) geld:

$$s_{ex}^2 = \Sigma W_v^2 s_{ev}^2$$

Maar uit (6) volg:

$$s_{ev}^2 = s_v^2(1 - \alpha_v)$$

waar s_v^2 = variansie van die v-de evaluering
α_v = betroubaarheid van die v-de evaluering

sodat: $s_{ex}^2 = \Sigma W_v^2 s_v^2(1 - \alpha_v) = \Sigma W_v^2 s_v^2 - \Sigma W_v^2 s_v^2 \alpha_v$

$$\text{in (13):} \quad \alpha_x = 1 - \frac{(\sum_{v=1}^{n} W_v^2 s_v^2 - \sum_{v=1}^{n} W_v^2 s_v^2 \alpha_v)}{S_x^2} \qquad (14)$$

Vergelyking 14 gee die betroubaarheid van die totale prestasiepunt in terme van die statistieke van elkeen van die evaluerings, v.

$W_v s_v$ staan bekend as die sg. "effektiewe gewigte" wat studente se relatiewe punte tot mekaar in 'n geweegde totaal bepaal (Clift, 1981, p.159). Mosier (1943), soos verder ontwikkel deur Conger (1980) toon dat 'n stel "optimale" gewigte om die maksimum totale betroubaarheid α_x' te lewer, verkry kan word deur (14) parsieël te differensieer m.b.t. die gewigte W_v , gelyk te stel aan nul, en op te los vir die gewigte. Dit lewer die volgende stel optimale gewigte W_v' vir die algemene geval:

$$W_v' = \sqrt{\alpha_v} / [s_v(1-\alpha_v)] \qquad (15)$$

4.2.2 Praktiese resultate

Voorbeelde van die analise van die betroubaarheid van vakprestasie word getoon in tabel 4.1. Vak A het byvoorbeeld uit drie evaluerings bestaan met betroubaarhede 0,56; 0,52 en 0,28 en gewigte 0,30; 0,40 en 0,30 respektiewelik. Die totale betroubaarheid vir die vak was 0,63 , terwyl die maksimum moontlike betroubaarheid vir die vak 0,73 sou wees as die evaluerings geweeg sou word in die verhouding 0,44; 0,38 en 0,17. Laasgenoemde "optimale weging" is van akademiese belang en is interessant heidshalwe by die tabel ingesluit. So 'n wegingsmetode waar die relatiewe gewigte van evaluerings nie vooraf bekend is nie is in die praktyk onaanvaarbaar.

Dit is nie altyd sinvol om die betroubaarheid van 'n vak te bepaal nie, weens twee hoofredes. Eerstens is daar gewoonlik groepprojekte in 'n vak, waar 'n student se individuele prestasie nie werklik deur die groepprestasie weerspieël word nie. Tweedens is dit selde moontlik om die betroubaarheid van elke komponent van die saamgestelde vakprestasie te bereken. Slegs 'n enkele punt word aan sommige evaluerings toegeken, sodat (8) nie toegepas kan word om betroubaarheid te bereken nie. Evaluerings soos huiswerkopdragte, praktika en blitstoetse kan ook nie saamgroepeer word as vrae in 'n enkele evaluering nie omdat dit verskillende vermoëns van die student meet. Die betroubaarheid van vak C, waar die betroubaarhede van twee evaluerings nie bereken kon word nie omdat dit uit enkelvrae bestaan het, illustreer die verskynsel. Let daarop dat die betroubaarheid van 'n vak wel groter as nul kan wees al is die betroubaarhede van individuele evaluerings almal nul, (14) verwys. Met hoë korrelasies tussen pare evaluerings kan 'n hoë

betroubaarheid selfs verkry word soos geillustreer deur vak C, waar twee van die evaluerings betroubaarhede van 0,00 gehad het, en die totale betroubaarheid van 0,73 steeds meer is as die betroubaarhede van beide die ander twee evaluerings.

Tabel 4.1 - Voorbeelde van Betroubaarheid van vakprestasie

EVALUERINGS		1	2	3	4	TOTAAL
VAK A						
Betroubaarhede	(α_v)	0,56	0,52	0,28		0,63
Gewigte	(W_v)	0,30	0,40	0,30		1,00
Optimale gewigte	(W_v')	0,44	0,38	0,17		1,00
Optimale Betroubaarh.	(α_x')					0,73
VAK B						
Betroubaarhede	(α_v)	0,68	0,33	0,43		0,67
Gewigte	(W_v)	0,30	0,40	0,30		1,00
Optimale gewigte	(W_v')	0,48	0,26	0,26		1,00
Optimale Betroubaarh.	(α_x')					0,77
VAK C						
Betroubaarhede	(α_v)	0,58	0,68	0,00	0,00	0,73
Gewigte	(W_v)	0,35	0,40	0,15	0,10	1,00
Optimale gewigte	(W_v')	0,40	0,60	0,00	0,00	1,00
Optimale Betroubaarh.	(α_x')					0,78

5 DIE SLEUTELVERGELYKING

Die sleutel tot begrip van betroubaarheid is vervat in (16), die totale variansie van 'n saamgestelde punt. Omdat prestasies van dieselfde student in verskillende evaluerings wel onderling afhanklik is word die totale variansie van 'n geweegde saamgestelde punt X, (12) verwys, gegee deur:

$$S_x^2 = \Sigma W_v^2 s_v^2 + 2 \sum_{k=2}^{n} \sum_{j=1}^{k-1} W_j W_k \text{cov}(j,k)$$

$$= \Sigma W_v^2 s_v^2 + 2 \sum_{k=2}^{n} \sum_{j=1}^{k-1} W_j W_k s_j s_k r_{jk} \quad , \text{ waar} \qquad (16)$$

$\text{cov}(j,k)$ = kovariansie tussen die j-de en k-de evaluerings
= $s_j s_k r_{jk}$
s_j en s_k = standaardafwykings van die j-de en k-de evaluerings.
r_{jk} = korrelasiekoeffisient tussen die j-de en k-de evaluerings.

By berekening van die betroubaarheid van individuele evaluerings volgens (8) bereken, α, gee (16) die variansie van die totale toetspunt, s_x^2. v, j en k Verwys hier na die vraag nommer.

By die betroubaarheid van die saamgestelde prestasiepunt in 'n vak wat volgens (14) bereken word, is (16) die variansie van die totale prestasiepunt. v, j en k Verwys in die geval na die evaluerings waaruit die prestasiepunt saamgestel word.

In beide gevalle word betroubaarheid verhoog deur hierdie totale variansie te verhoog, wat bereik word deur r_{jk} te maksimeer. In 'n gegewe situasie is die gewigte W_v ,die variansies s_v^2 en standaardafwykings s_j en s_k in werklikheid gegewe en slegs r_{jk} varieer. Hoe nader r_{jk} neig na 1,0 hoe groter is die betroubaarheid.

By 'n toets (individuele evaluering) is die uitgangspunt dus dat die verskillende vrae almal dieselfde kennis en vermoë van die student meet. Hoe meer die relatiewe prestasie van die onderskeie studente in die verskillende vrae ooreenstem, hoe hoër is die betroubaarheid. Veronderstel 'n toets bestaan byvoorbeeld uit 10 vrae wat elkeen 10 punte tel. Vir 100 persent betroubaarheid sal 'n bepaalde student in elkeen van die vrae presies dieselfde punt uit 10 moet verwerf. Die korrelasie tussen die prestasie van die klas in elke paar vrae is dan 1,0 . Dit kom daarop neer dat dit vir 'n betroubaarheid van 1,0 'n vereiste is dat

$$s_j = s_k = s \text{ en } \text{cov}(j,k) = s_j.s_k = s^2 \qquad (\text{vir } \alpha = 1{,}0)$$

(Gelyke standaardafwykings in elke vraag impliseer ook dat die punte waaruit elke vraag tel gelyk moet wees. Berekende betroubaarheid word dus verlaag wanneer puntetole van vrae baie varieer.)

By die saamgestelde prestasiepunt vir 'n vak geld dieselfde uitgangspunt, nl. dat elk van die evaluerings dieselfde kennis en vermoë van die student meet. Ooglopend is dit 'n oorvereenvoudiging aangesien die relatiewe vermoë van 'n individuele student om 'n toets te beantwoord byvoorbeeld anders mag wees as sy vermoë om 'n praktiese opdrag uit te voer. Vir 'n groep studente behoort die korrelasies egter steeds hoog positief te wees, en negatiewe korrelasies sou byvoorbeeld nie verwag word nie.

Tabel 5.1 toon byvoorbeeld die korrelasietabel van 'n vak wat volledig ge-analiseer is. Die korrelasiekoeffisient, r_{ij} tussen die punte wat elk van die 20 studente verwerf het in elke paar evaluering, word getoon. Hier is nie veel verrassings nie, behalwe dat die korrelasies nie hoog is nie, met soos verwag die grootste korrelasie tussen die twee toetse. Uit die groep analises is hierdie 'n voorbeeld van 'n "goeie" vakevaluering, soos later bespreek. Tabel 5.2 toon 'n tipiese voorbeeld van 'n "swakker" korrelasietabel. Daar is 'n positiewe korrelasie tussen die twee toetse

(by alle vakke was hierdie korrelasie positief). Die ander korrelasies is egter laag en selfs relatief groot negatief (tussen die projek en die finale toets). Een verklaring vir die negatiewe en lae positiewe korrelasies is dat die verskillende evaluerings verskillende tipes van aanleg, verskillende kwaliteite, soos praktiese vaardighede vs. begrip en kennis, meet. 'n Ander verklaring is dat swak studente besondere aandag aan take, projekte en huiswerk gee omdat dit hulle 'n geleentheid bied om hulle gemiddelde punt te verbeter. Studente mag ook minder aandag gee aan werk as hulle in een evaluering goed gedoen het, omdat hulle slegs belangstel om 'n vak te slaag, en die minimum doen om deur te kom. Laastens mag die evaluerings bloot swak wees en nie kennis en vermoë nie, maar toevallige prestasie, meet. Wat betref die "goed" of "swak" interpretasie van 'n korrelasietabel is die skrywer van mening dat dit verkieslik is dat alle korrelasies positief moet wees, en verder hoe groter hoe beter. Clift (1981,p.114) meen selfs dat in die geval van negatiewe korrelasies "... it is not logical to combine them into a somewhat meaningless average". Wanneer daar negatiewe korrelasies is moet dit logies verklaarbaar wees (op grond van verskillende tipes van aanleg wat byvoorbeeld ter sprake mag wees), en konsekwent wees met die res van die korrelasies. Tabelle 5.3 en 5.4 toon addisionele voorbeelde van 'n "goeie" en 'n "swak" korrelasietabel.

Tabel 5.1 - Korrelasietabel Vak 1

	Huiswerk	Finale toets
Toetsweek	0,22	0,33
Huiswerk		0,14

Tabel 5.2 - Korrelasietabel Vak 2

	Huiswerk	Toetsweek	Finale toets
Projek	0,19	-0,01	-0,26
Huiswerk		0,07	-0,10
Toetsweek			0,22

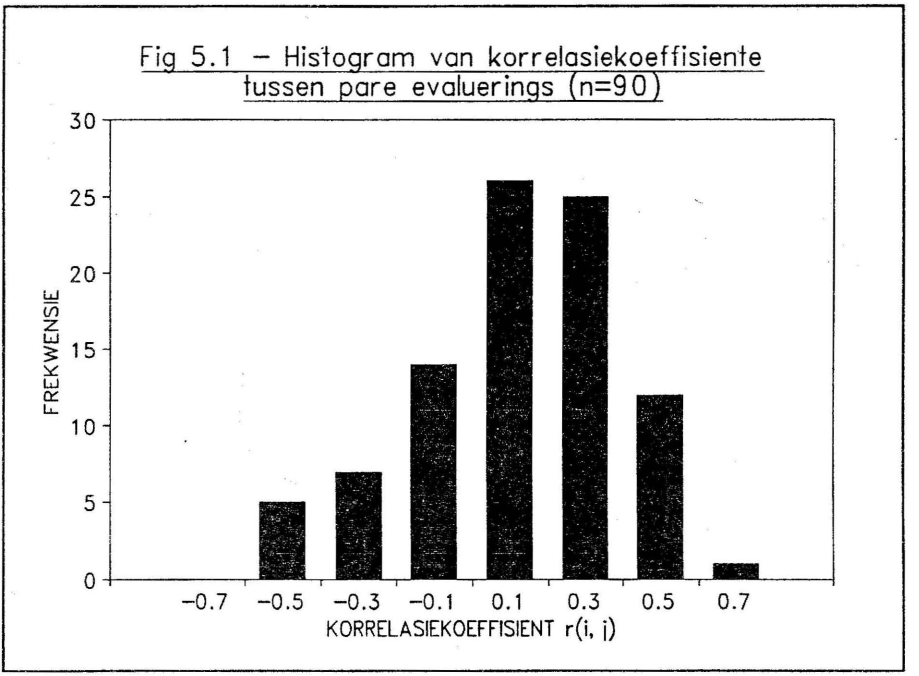
Tabel 5.3 - Korrelasietabel Vak 3 ("goed")

	Take & Prak	Toetsweek	Finale toets
Tutoriale	0,39	0,13	0,57
Take & Prak		0,02	0,08
Toetsweek			0,39

Tabel 5.4 - Korrelasietabel Vak 3 ("swak")

	Klastoets	Opdragte	Toetsweek	Finale toets
Tutoriale	-0,15	0,18	-0,45	0,34
Klastoets		-0,25	0,35	-0,17
Opdragte			-0,28	0,55
Toetsweek				0,08

'n Histogram van die korrelasiekoeffisiente tussen pare evalue rings wat verkry is met die analise word in fig. 5.1 getoon. Die gemiddelde korrelasie was 0,14 met 'n standaardafwyking van 0,28.



6 SPOED VS. KRAGTOETSE

Die gesplete helfte tipe benadering hier gevolg tot betroubaarheidsberekening is slegs toepaslik op sg. "kragtoetse" waar daar genoeg tyd gelaat word vir meeste

studente om alle vrae te probeer beantwoord. Met "spoedtoetse", waar almal nie voldoende tyd het om alle vrae te beantwoord nie, kan toets-hertoets metodes waar volledige soortgelyke vraestelle byvoorbeeld beantwoord word, gebruik word. In die klaskamersituasie is dit egter totaal onprakties.

Vanselfsprekend is 'n vraestel baie selde 'n suiwer kragtoets. Dit is 'n kombinasie van krag en spoed, hopelik oorwegend krag-, maar met beperkte tyd sodat sommige studente nie geleentheid kry om alle vrae te beantwoord nie. Uit die oogpunt van betroubaarheidsberaming is dit goed om vrae van maklik- na moeilik te rangskik en studente so in te lig sodat hulle die vrae in volgorde van moeilikheidsgraad beantwoord. Sommige studente se vermoë raak dan op lank voor die einde van die toets, maar dit het 'n kleiner invloed op die berekende betroubaarheid omdat die variansies van vraagresultate kleiner behoort te wees. 'n Ander eenvoudige hulpmiddel is om gedurende die toets met gereelde tussenposes af te kondig wat die oorblywende tyd is, sodat studente herinner word om tyd aan alle vrae te spandeer.

Hopkins (1981, p.146) som die nadele verbonde aan spoedtoetse uitstekend op, en dit is gepas om hom verbatim aan te haal.

"Some examinees have a test-taking set that causes them to work slowly and carefully; others have a tendency to work quickly and with less caution (Guilford & Lacey, 1947). The correlation between ability and working rate on tests have been shown to be very low (Tate, 1948; Ebel, 1954; Hopkins, 1964b). Some examinees respond more slowly than others irrespective of item difficulty or test content (Bennett & Doppelt, 1956; Davidson & Carroll, 1945, Tate, 1948). . . . Teacher-made and standardized tests (Boag & Neild, 1962; Kahn, 1968; Knapp, 1960) frequently have inadequate time limits; this allows the irrelevant effects from the speed- vs. -accuracy response set to contaminate the validity of test scores. Several studies have found that tests may measure different mental functions when administered under power and speed conditions (Lord, 1956; Myers, 1960; Mollenkopf, 1960). Older people tend to work more slowly, a factor that led to a gross overestimation of the degree of mental decline with age in some earlier studies (Lorge, 1952).

Except for those educational objectives for which speed of response is an important objective (e.g., typing, reading), tests should be constructed and administered so that virtually all examinees (perhaps 90%) complete the examination."

Dosente volg somtyde die strategie om uitermate lang- en/of moeilike vraestelle op te stel om die gemiddelde prestasie in 'n toets laag te hou. Dit skep dan die geleentheid om punte na bo aan te pas deur dit uit 'n kleiner totaal te laat tel. Die gemiddelde prestasiepunt word so na 'n verlangde vlak verstel. Hierdie is 'n uiters

swak werkwyse en 'n normale vraestel met moeilike en makliker vrae-, en van normale lengte, wat dan liewer baie streng nagesien word, sodat punte op dieselfde wyse aangepas kan word, is verkieslik. Vraestelle moet nog te lank, en nog onredelik moeilik wees.

7 BESPREKING

Daar is bevind dat die berekening van die betroubaarheid van evaluering 'n uiters nuttige oefening is en dat dit 'n beduidende invloed het op die wyse waarop toetse opgestel word. Dit gee 'n objektiewe syfermaatstaf van die kwaliteit van evaluering in terme van die bestendigheid waarmee gedifferensieer word tussen studente. Swak eienskappe van 'n toets kan gediagnoseer-, en probleme in die evalueringstelsel kan geidentifiseer word. Die eenvoudige reëls, nl. krag i.p.v. spoedtoetse; rangskikking van vrae van maklik na moeilik; herinnering van studente aan die oorblywende beskikbare tyd gedurende die afneem van 'n toets; streng merk, eerder as 'n onredelik moelike toets, om die gemiddelde prestasie te beheer; lei tot beter evaluering in die algemeen, en tot hoër betroubaarheid. Soos genoem in paragraaf 5 word die berekende betroubaarheid ook verlaag wanneer puntetole waaruit vrae tel baie varieer. Hierdie probleem kan verminder word deur vrae te verdeel in onderafdelings, met punte vir elke onderafdeling. Elke onderafdeling word dan as 'n afsonderlike vraag gebruik vir berekening van betroubaarheid.

Die formules wat gebruik word is eenvoudig. Slegs vergelykings (8) en (14) en die korrelasie-koeffisiente tussen punte verwerf in pare van vrae of evaluerings is van belang. Verder is dit goed om die konstruksie van 'n histogram van totale punte as deel van die program in te sluit. Daar moet buitendien rekord gehou word van punte en al addisionele werk is om die punte per vraag in te voer.

By die interpretasie van resultate moet die fundamentele uitgangspunt by die berekening van betroubaarhede in gedagte gehou word, nl. dat t.o.v. 'n toets aanvaar word dat elke vraag dieselfde basiese vermoë, vaardighede en kennis van 'n student meet, en dat hoë betroubaarheid volg uit 'n hoë korrelasie tussen punte in individuele vrae. By die berekening van die betroubaarheid van geweegde vakprestasie geld 'n soortgelyke aanname t.o.v. die evaluerings waaruit die prestasiesyfer opgebou word. Hierdie aannames is nie altyd geldig nie. 'n Betroubaarheid van 0,00 bly egter swak en dui waarskynlik op 'n tekortkoming in die opstel van toetse, of in die evalueringstelsel.

Ten slotte is die belangrikheid van die begrip van analise van kovariansie op 'n elementêre vlak geillustreer. Met die meeste ingenieurs statistiek, uitgeslote regressie-analise, word die aanname gemaak dat veranderlikes lineêr onafhanklik is, met die gepaardgaande sentrale limietbenadering van sommering van variansies as skatter van die variansie van die som. Tydens toepassing fouteer ingenieurs

waarskynlik gereeld deur oorvereenvoudigde aannames t.o.v. lineêre onafhanklikheid te maak. Die bedryfsingenieur kry gereeld te doen met probleme waar verskillende eienskappe en prestasies van dieselfde groep mense van die veranderlikes is wat 'n rol speel. Dit is juis hier waar lineêre afhanklikheid verwag kan word en die bedryfsingenieur behoort toegerus te wees om hierdie tipe statistiek te kan hanteer. By simulasie verdien die simulasie van lineêr afhanklike veranderlikes byvoorbeeld meer aandag.

VERWYSINGS

[1] Clift JC, Imrie BW: Assessing Students, Appraising Teaching, Croom Helm, London, 1981.

[2] Conger AJ: "Maximally Reliable Composites for Unidimensional Measures", Educational and Psychological Measurement, 40, 1980, pp.367-375.

[3] Cronbach LJ: "Coefficient Alpha an the Internal Structure of Tests", Psychometrika, 16(3), Sept. 1951, pp.297-334.

[4] Hopkins KD, Stanley JC: Educational and Psychological Measurement and Evaluation (6th ed), Prentice-Hall, Englewood Cliffs, 1981.

[5] Hoyt C: "Test Reliability Estimates by Analysis of Variance", Psychometrika, 6(3), Jun. 1941, pp.153-160.

[6] Kuder GF, Richardson MW: "The Theory of the Estimation of Test Reliability", Psychometrika, 2(3), Sept. 1937, pp.151-160.

[7] Mosier CI: "On the Reliability of a Weighted Composite", Psychometrika, 8(3), Sept. 1943, pp.161-168.

[8] Page DC: Die statistiese aspekte van die evaluering van senior bedrysingenieurswese studente aan die Universiteit van Stellenbosch, ongepubliseerde navorsingsverslag, Dept. Bedrysingenieurswese, Universiteit van Stellenbosch, 1992, 50 pp.

[9] Thomas CR: "Examination Reliability and Reliability-weighted Composite Scores", Engineering Education, Jan. 1986, pp.227-231.