5

Frå

S P E S I F I K K

til

G E N E R I S K

R E L I A B I L I T E T S T E O R I

Hans-Magne Eikeland

University of Oslo

Oslo, november 1967

FØREORD

Dette arbeidet er eit referat eg skreiv frå ei seminarrekkje eg hadde ved Pedagogisk Forskningsinstitutt vårsemestret og haustsemestret 1967.

Tittelen har kome til etterpå. Nemningane spesifikk og generisk reliabilitet er henta frå Lord og Novick: Statistical theories of mental test scores, som kom i bokform i 1968, og dei dekkjer omgrepet klassisk reliabilitet og det som på engelsk-amerikansk er kalla generalizability. Terminologien ser ikkje ut til å ha stabilisert seg i litteraturen. Sjølv synest eg det kan vera om å gjera at vi ved den terminologien vi tek i bruk, får fram kontinuiteten frå klassisk til moderne testteori. Det lukkast ikkje godt om vi bruker reliabilitet og generalizability. Eg såg gjerne at omgrepet reliabilitet også kunne brukast innanfor nyare testteori, men då sjølvsagt med eit tenleg modifiserande adjektiv for både klassisk og moderne teori. Spesifikk og generisk reliabilitet tykkjest vera godt brukande.

Det kan vera nyttig å ta med eit noko lengre sitat frå Lord og Novick's kapittel 8, for om mogleg å gjera det meir intuitivt klårt kva skilnaden på spesifikk og generisk testteori er. Vi siterer frå innleiinga til kapittel 8:

> Consider an examiner who has obtained one measurement on each of a number of people. If he is perfectly satisfied with his measurements, that is, if he feels that the score of each individual accurately represents the psychological variable that he is trying to measure, then he will see little reason to concern himself with mental test theory and he will proceed to use the scores as they are.
>
> On the other hand, the examiner may feel that these scores may slightly misrepresent the abilities of the individuals being measured. For example, he may feel that the scores might have been different if a different but equally satisfactory test had been used, or if the test had been administered at a different time or under different conditions. In this case, the examiner will be interested in something other than the test scores that he has at hand.....
>
> In Chapters 1 through 7, we assumed that the variable (ability) of immediate interest to the examiner can be defined as the

> expected value of the measurement he has obtained, the expectation being taken over the (hypothetical) set of all parallel measurements. In Chapter 7, we called this expected value the specific true score, to distinguish it from other kinds of true scores. The present chapter is concerned not only with situations where it is impossible for practical reasons to obtain parallel measurements, but also with situations where it is undesirable for logical reasons to define true score in terms of any single test form.
>
> The examiner who chooses to study the specific true score is in effect making the following assertion: "If I were allowed more testing time to obtain a single total score for each examinee, I would choose to administer a longer test made up of forms identical (or tau-equivalent) to the form actually administered, insofar as this were possible without having experimentally dependent errors of measurements." Most examiners would not really wish to utilize additional testing time in this way, however. They would feel that the "true score" in which they are interested has aspects not covered by the items in the test actually administered, aspects that they would wish to cover if additional testing time were available. If they were able to administer several additional test forms in the additional time, obtaining a single total score, it would not disturb them to know that some of these forms were a little more difficult than others or measured along slightly different dimensions, so long as they were sure that each form measured important aspects or manifestations of the psychological variable under study.....
>
> The simplest situation to consider is the one in which the examiner conceives of a pool or population of nominally parallel test forms and defines his true score as the expected score over this population of forms.....
>
> The notion of generic true score is implicit in any approach to the analysis of repeated measurements by analysis of variance components. The model studied in Chapter 7 is a special case in which the test forms effect is assumed to be zero. For many practical applications this simpler model is entirely adequate. The idea of using a generic true score has been developed and is strongly advocated by Cronbach and his associates; the reader might see Rajaratnam (1960), and also Cronbach, Gleser, and Rajaratnam (1963). They do not use the term generic but speak of generalizability, whereas we use the older term reliability.

I det opphavlege referatet var det eit 7. kapittel om ein kovariansmodell for G-estimering av ikkje-stratifiserte komposita. Dette har eg ikkje lenger med etter som det er skrive ut som eit sjølvstendig bidrag: Coefficient alpha and the expected variance-covariance matrix of random composite measurements, 1970.

I det same kapittel 7 var det peika på ei vidare modellutvikling for stratifiserte testar. Dette er gjort i The structure of generalizability theory for hierarchically stratified tests, 1972.

Referatet er elles i si opphavlege form. Forandringar burde sikkert gjerast. Med tanke på ei eventuell omarbeiding og á jour-føring er eg sjølvsagt takksam om dei som les dette referatet, kan koma med framlegg til forbetring, kritisera og peika på feil som måtte vera gjorde.

Av tekniske grunnar er det greske symbol, lite sigma, skriven ø og ikkje som det burde vore skrive, σ. Det er å vona at dette ikkje vil bli til bry for lesaren.

Oslo i januar 1973.

Hans-Magne Eikeland

INNHALDSLISTE

1. Innleiing.

1.1 Reliabilitet og validitet

Vi kan trygt seia at reliabilitet og validitet er dei to sentrale omgrep i testteorien. I klassisk teori er desse omgrep identifisert med interform korrelasjonar og testkriterium korrelasjonar (Cronbach, Rajaratnam, Gleser (1963), 137).

Vi har rekna og reknar framleis validitet viktigare enn reliabilitet, med rette når vi ser reint praktisk på testing. Dei fleste av oss veit nok at validiteten har ei øvre grense sett av reliabiliteten. Men dette er akademisk kunnskap, kontraintuitiv meir enn intuitiv. Det kan vera på sin plass i blant å minna om at reliabilitet er nødvendig for validitet, om enn ikkje nok.

I ein statistisk analyse av eit vel tilrettelagt psykologisk eller pedagogisk eksperiment vil ein signifikans fortelja oss at det er god grunn til å tru at det observerte resultat kan tilskrivast systematiske påverknader. Det ligg då påliteleg informasjon i resultatet, som vi reknar med har si forklaring i den uavhengige variable i eksperimentet og ikkje i andre ukontrollerte systematiske påverknader. Eit ikkje-signifikant resultat fortel oss at den observerte variasjon, ofte ein differense, kan vera såkalla feilvarians, ein varians frå tilfellelege og/eller ikkje-ønskte variasjonskjelder. Vi har såleis ikkje grunn til å tru at resultatet gjev oss påliteleg informasjon, i vår samanheng.

Parallellen frå eksperiment til test er klår: Reliabiliteten kan oppfattast som ei signifikansprøving av individuelle differensar. Er reliabiliteten "høg nok", reknar vi med at dei observerte differensane i stor grad skriv seg frå systematiske variasjonskjelder og at dei ikkje i særleg grad er bestemte av tilfellelege og ikkje-ønskte påverknader. Når ein reliabilitetskoeffisient gjev oss den informasjon at systematisk variasjon i stor grad kan forklara dei observerte individuelle differensane, er dette å oppfatta som eit klarsignal til å gå vidare og freista å finna den psykologiske meining i desse systematiske variasjonane. Då er vi over i valideringsproblematikken.

Det vi no har sagt om reliabilitet og validitet, er berre ein første grenseoppgang mellom desse to grunnleggjande omgrep.

1.2 Teikn på tidskifte i testteori

Ting tyder på at med den nyorientering som er i ferd med å skje i testteorien, vil reliabilitetsomgrepet, eller eit reformulert reliabilitetsomgrep, koma til å bli meir likestilt med validitetsomgrepet.

Vi har hatt berre ein grunnleggjande teori i testing. Det er den Spearman og Brown, kvar for seg, utvikla i det første tiåret av dette hundreåret. Andre har ført denne utviklinga vidare, men vi kan ikkje seia å ha fått noko fundamentalt nytt.

Like fram til midt i 50-åra skjedde det ikkje noko radikal nyorientering. Men då tok ting til å skje. Det starta med validitetsomgrepet. Construct validity er ei nyskaping i testteorien og kom truleg som resultat av misnøye med så einsidig å knyta valideringa av ein test til eit kriterium, som ofte er eit mykje mangelfullt kriterium.

Construct validering kan kort og noko upresist karakteriserast som validering ved hypoteseprøving. Ei slik validering vil i mykje større grad enn tradisjonell validering bli eit samspel mellom teori, fantasi resonnering og observasjon. Validitetsomgrepet har fått noko meir spekulativt over seg, men det er framleis under streng empirisk kontroll.

Det er forunderleg kor lett det gjekk å få det nye validitetsomgrepet akseptert. Det skjedde faktisk før den teoretiske presentasjon av omgrepet. Technical Recommendations frå 1954 rådde til å ta i bruk construct validering. Den teoretiske utgreiing kom først i 1955 (Cronbach og Meehl (1955)). Med construct validering er testpsykologien vorten meir teoretisk enn før, og mindre operasjonistisk. Denne utviklinga ser ut til å gå igjen i det som skjer med reliabilitetsomgrepet just no.

I 1966 kom revidert utgåve av Technical Recommendations, no under nytt namn Standards, altså meir imperativt enn i 1954. Her skal visstnok alt i alt ikkje vera særleg mykje nytt i høve til TR 1954 (Ed Ps Ms (1966)). Likevel er det all grunn til å merka seg det som er sagt om reliabilitet. Som vi alle kjenner til er reliabilitetstypane stabilitet, ekvivalens, internal consistency og stabilitet og ekvivalens innarbeidde omgrep både teoretisk og praktisk. Desse omgrep er det i Standards gjort framlegg om å sløyfa, og det vil sikkert koma noko uventa på mange. Dette minner ikkje lite om det som skjedde med validitetsomgrepet i Technical Recommendations frå 1954. Den gongen vart construct validering rekommendert til praktisk bruk utan å vera førebudd i særleg grad. No blir tradisjonelle reliabilitetstermar tilrådd å takast ut av bruk, og mange vil nok synast at ei så vidt stor forandring er lite førebudd.

Grunnlaget for denne rekommendasjonen finn vi i den nyorientering vi i dei aller seinaste år har kunna merka innanfor reliabilitetsforskning. Denne nyorienteringa kan kanskje først og fremst tilskrivast Cronbach og hans medarbeidarar som frå 1963 og utetter har publisert ein del artiklar om generalizability. Det er eit nytt omgrep som representerer ei vesentleg reformulering av reliabilitet i tradisjonell forstand. Ser vi historisk på reliabilitetsomgrepet, vil vi likevel kunna finna at generalizability har røter langt tilbake.

Dei tendensar til nyorientering vi her har nemnt, kjem truleg til å få ei førebels avrunding i eit omfattande testteoretisk verk som er under førebuing av F.M. Lord og M.R. Novick ved Educational Testing Service. Boka får tittel "Statistical Theories of Mental Test Scores" og har ei tid vore tilgjengeleg i stensil.

Til no har Gulliksens "Theory of Mental Tests" frå 1950 vore standardverket i testteori. Med den boka må vi truleg kunna seia at klassisk testteori kulminerte i og med at Gulliksens bok ikkje berre byggjer på dei matematiske modellane som kan førast tilbake til Spearman og Brown, og som er restriktive i den forstand at dei set mykje strenge statistiske krav til data, men Gulliksen tek i tillegg også med ekstramatematiske restriksjonar.

Medan Cronbachs generalizability er ein ny teori som i stor grad byggjer på ikkje-restriktive krav til materialet og som i så måte poengterer eit brot med tradisjonell testteori, ser det ut til at Lord og Novick godtek både klassisk teori og generalizability som matematiske modellar og utviklar dei side om side. Det er grunn til å merka seg at Lord og Novick reknar med at deira bok vil koma til å avløysa Gulliksen som standardverk i testteori.

1.3 Plan

Vi skal i det som følgjer, sjå nærmare på det som har skjedd i testteorien i dei seinare år og som fører fram mot den nye teorien som er kalla generalizability. For å få perspektiv på denne utviklinga skal vi starta med nokre generelle målingsteoretiske synspunkt og ein historikk over reliabilitetsomgrepet før vi tek for oss den nyorientering som endar opp med ein ny reliabilitetsteori.

Det er grunn til å poengtera at denne utgreiinga ikkje tek sikte på å dekka Lord og Novick. Vi kjem nok til å låna ein god del synspunkt frå dei, men det blir spreitt og lite systematisk. Medan Lord og Novick prøver å integrera klassisk teori og generalizability, vil vår framstilling sikta mot det som skil generalizability frå klassisk teori.

2. Nokre generelle målingsteoretiske synspunkt

2.1 Deterministiske og probabilistiske modellar

I testteorien står vi framfor det problem å rekna ut om råd er, eller i alle høve å estimera i kor stor grad variasjonen i testskårane, den avhengige variable, kan forklarast ved ein eller fleire uavhengige variable (systematisk variasjon). For å koma ut av dette problemet må vi ha ein matematisk modell som, så langt råd er, er isomorf til våre empiriske observasjonar.

Vi skal sjå på to slike matematiske modellar: Den eine kallar vi ein deterministisk modell, den andre ein probabilistisk. Den deterministiske kan sjå slik ut i generell form: $x = f(s)$. Her er x ein avhengig observerbar variabel og s ein eller fleire uavhengige variable, observerbare eller ikkje observerbare, og f ein funksjon som relaterer x og s. Modellen seier at når vi veit verdien av s så veit vi og verdien av x. I dei fysiske vitskapane kan denne modellen vera realistisk nok. I mange høve er x praktisk talt bestemt av s, slik at det er berre lite av variasjonen i x igjen som ikkje er bestemt av s. Denne restvariasjonen kallar vi gjerne residualen. I gagnet kan ein slik modell vera tenleg, endå om vi i namnet sjeldan kan seia at ein deterministisk modell er den korrekte.

I psykometrien er ein deterministisk modell ikkje tenleg, fordi vi der korkje i namnet eller gagnet kan rekna med at våre empiriske observasjonar let seg forklara ved systematisk variasjon åleine. Vi ventar ein ikkje uvesentleg residual. Difor har vi bruk for ein modell som reknar med både systematisk og ikkje systematisk variasjon. Ein slik modell kallar vi ein probabilistisk modell. Vi kan skriva modellen i generell form: $x = f(s) + r$. I denne modellen har vi fått med ein r som kan karakteriserast som eit kompositum av effektar som ikkje har samanheng med den uavhengige variable.

Denne probabilistiske modellen krev ein teori om residualvariansen, ein feilteori seier vi gjerne. Vi siterer Coombs (1966):

"The process of constructing a correspondence between an empirical relational system and a numerical system is measurement. But there are implications in the formal numerical relational system which imply corresponding observations in the empirical relational system... As is well known this empirical implication is commonly violated so one has a correspondence which is not perfect. Those empirical observations which violate implications of the model are called errors .... Hence any application of a measurement theory requires an error theory which permits establishing a correspondence between the measurement theory and the empirical observations when the correspondence is imperfect and which simultaneously, then, describes the error".

Eit sentralt problem i testteoriens meir enn 60-årige historie er nettopp dette korleis ein adekvat feilteori skal formast ut.

2.2 Måling per definisjon

Ein viktig ting å merka seg er dette: Ein eigenskap ved fysiske ting er som regel handfast og kan målast direkte. Denne form for måling blir gjerne kalla fundamental måling (Torgerson (1958)). Annleis er det med dei psykologiske eigenskapar som vi ønskjer å måla. Dei er alt anna enn handfaste. Ein psykologisk eigenskap, ein dimensjon eller eit trekk om vi vil, er i første omgang eit hypotetisk construct, ein definert dimensjon. Denne dimensjonen må eksplikerast. I vår samanheng vil det seia å definera åtferd som vi reknar med kan spegla av denne tenkte dimensjonen. Endeleg må vi fram til konkrete testsampel (items) som representerer vårt definerte åtferdsunivers. Vi seier då at det hypotetiske construct er operasjonelt definert ved testsampelet. Av dette skulle det gå fram at det vi kallar måling i psykologien i grunnen ikkje er måling i det heile. Denne form for måling blir ofte nemnt indikering. Torgerson (1958) bruker termen "measurement by fiat" eller "measurement by definition". Det knyter seg mange refleksjonar til måling per definisjon. Torgerson har ein del tankevekkjande synspunkt som vi siterer (Torgerson (1958), 23-35):

"There is little we can say about measurement by fiat, since it depends so heavily on the intuition of the particular experimenter. One thing should be emphasized, however: there is certainly nothing wrong or logically incorrect with the procedure ..... It has led to a great many results of both practical and theoretical importance. For example, a major share of the results of the field of mental testing and of the quantitative assessment of personality traits has depended upon measurement by fiat. Measurement of morale, effeiciency, drives, and emotion, as well as most sociological and economic indices, is largerly measurement of this type".

"In all these cases, one or more observable properties are selected which on a priori grounds are judged to be related to the concept of interest. A measure of the observable property itself or of a simple or weighted sum of several such observable properties is taken as the measure of the concept of interest".

"The major difficulty with measurement by fiat is the tremendous number of ways in which such defined scales can be constructed. We might measure the strength of food drive by the number of hours of food deprivation, by the amount of shock an animal is willing to take in order to reach food, by the amount of weight lost during a particular period of deprivation, and so on".

"In the field of mental testing, the possibilities are enormous. We have only to consider that, since any single arithmetic problem can be considered to be a indicant of arithmetic ability, any combination of any number of arithmetic items, presented orally or written, can be taken as the defined measure of this ability. Each is a separate explication of an initial concept of aritmetic ability. Although subsequent investigations may establish that many lead to virtually the same result and hence may be considered to be equivalent operational definitions of the

same concept, many will also lead to quite different results, in which case they are operational definitions of different concepts. The same state of affairs occurs as well in measurement of attitudes and personality traits, sociological and economical indices, and the like".

"Since there are so many possibilities, since such scales come se cheap, the confidence in any particular explication of this type can be expected to be low. As a result we cannot always blame the theoretician for rejecting the explication rather than his model when the experimental results do not go in the direction indicated".

Desse refleksjonane omkring måling per definisjon har viktige implikasjonar i reliabilitetsforskning, og vi skal sjå nærmare på dei ved seinare høve.

2.3 Konstans

I fysisk måling kan vi rekna med at det som skal målast i stor grad er uforanderleg eller at dei vilkår som fører til forandring, er kjente slik at desse kan haldast under kontroll. Difor er det god meining i repeterte målingar. I psykometrien derimot er "gjenstanden" for måling alltid meir eller mindre foranderleg. Det er difor vanskeleg for oss å få tak i ein målingsfeil som går på presisjonen i å måla ein konstant "gjenstand". Ein målingsfeil i psykometrien må i praksis bli ein kombinasjon av målingsfeil og funksjonsfluktuasjon. Berre reint hypotetisk kan vi definera oss fram til det som med rette måtte bera namnet målingsfeil.

2.4 Eksperimentell independens

Suksessive målingar i fysiske vitskapar kan seiast å vera uavhengige i den forstand at ei første måling ikkje nødvendigvis verkar inn på ei andre måling. I psykometrien kan vi berre reint teoretisk postulera uavhengige repeterte målingar, men vi har ikkje særleg god grunn til å tru at dette er sant. Både minne om føregåande måling og dette at ei måling kan føra til forandring av det som skal målast ein andre gong, gjer at postulatet om eksperimentell independens ikkje er særleg plausibelt.

Når vi såleis korkje kan rekna med konstans eller eksperimentell independens, er det forståeleg at vi ved repeterte målingar ikkje kan få ein uhilda målingsfeil og såleis heller ikkje ein reliabilitet av måle-instrumentet per se. Vi kan ikkje observera ein tests reliabilitet i den meining som er vanleg når det er tale om instrument i fysisk måling. Ellers er omgrepet målingsfeil mykje uklårt.

3. Klassisk testteori.

3.1. Det tradisjonelle reliabilitetsomgrepet

Ein historisk studie av reliabilitetsomgrepet frå Spearman og Brown til i dag vil nokså sikkert enda opp med eit totalinntrykk av omgrepet som er alt anna enn klårt. Den enkelte kan skriva greitt og forståeleg om problemet ut frå sin definisjon av omgrepet. Men den meining reliabilitetsomgrepet får, ber merke etter utgangspunktet. Når så utgangspunkta er mange, blir det også mange meiningar om reliabilitet.

Tryon skreiv i 1957: "If an investigator should invent a new psychological test and then turn to any recent scholarly work for guidance on how to determine its reliability (Tryon viser til Guilfords Psychometric Methods som døme), he would confront such an array of different formulations that he would be unsure about how to proceed. After fifty years of psychological testing, the problem of discovering the degree to which an objective measure of behavior reliably differentiates individuals is still confused" (Tryon (1957), 229).

Det er likevel misvisande, vilkårslaust, å seia at reliabilitetsomgrepet er uklårt. Usemje om reliabilitetsomgrepet gjeld først og fremst innhald og ikkje form.

Vi har tidlegare sagt at reliabiliteten går på den systematiske variasjon i dei observerte individuelle differensane. Å finna den systematiske varians er eit spørsmål om å kunna dekomponera totalvariansen eller å finna fram til variansstrukturen. Det er reliabilitet på eit formalt plan. Når vi vidare skal definera kva vi legg i systematisk varians, då er det spørsmål om innhald, og då er det rom for mange definisjonar.

Reliabilitet i klassisk meining har ein syntaktisk definisjon som er bunden av det testteoretiske grunnlag. Har vi godteke eit teoretisk utgangspunkt, må vi også godta visse konsekvensar av dette utgangspunktet. Vi har før sagt at det teoretiske grunnlaget i testpsykologien har vore det same i 60 år. Det vil seia at den grunnleggjande modellen går igjen endå om synet på innhaldet i reliabilitetsomgrepet kan variera.

Innhald i reliabilitetsomgrepet impliserer eksperimentelle framgangsmåtar til å skaffa test-data til vegar. Ein formell modell skal realiserast. Ei form skal fyllast med innhald. Det er på dette semantiske eller operasjonelle plan det har så lett for å bli usemje om reliabilitetsomgrepet. Kvar fyller forma med sitt innhald, for di strukturmodellen kan passa på ulike eksperimentelle framgangsmåtar.

Vi trur det er viktig å poengtera dette: Det syntaktiske reliabilitetsomgrepet er ein konsekvens av testeorien og er så langt godt som eintydig. Det semantiske reliabilitetsomgrepet derimot, har vore og er uklårt.

I tur og orden skal vi sjå på reliabilitetsomgrepet ut frå denne systematiseringa.

3.1.1. Syntaktiske definisjonar

Eit sentralt omgrep i klassisk reliabilitetsteori er sann skåre (true score). På grunn av "unøyaktig" måling tenkjer vi oss at den observerte skåre, X, redusert med ein feilskåre, E, gjev oss den sanne skåre, T. Dette kan vi skriva slik:

$$X = T + E \qquad (F1)$$

Den sanne skåre kan ikkje observerast og er difor ikkje sjeldan sett på med skepsis. "The concept of true score appears to raise some philosophical problems because often the true score cannot be directly measured. Certainly direct measurement is necessary in science; generally, however, scientists do not insist that all concepts in a science must be directly measureable. Rather,..... it is sufficient that all concepts be related syntactically to other directly measureable concepts" (Lord and Novick (1966), 37).

Vi skal her skilja mellom det vi kan kalla to ulike tradisjonar i syn på sann skåre. Ein kan det høva å kalla ein Spearman-Yule tradisjon; ein annan ein Brown-Kelley tradisjon (Tryon (1957), 230. Ghiselli (1964), 219).

3.1.1.1. Spearman-Yule tradisjonen

Det syn på sann skåre denne tradisjonen representerer, har nyss fått namnet det platoniske synspunktet (Sutcliffe (1965),). Synspunktet går tilbake til Spearman og er ført vidare av Yule, m.a. i hans lærebok i statistikk frå 1922.

Spearmans grunnleggjande konstruksjon i den kjente 1910-artikkelen postulerer ein konstant skåre som for same person går igjen frå test til test når testane er tenkt å måla det same. Spearman seier: "x_1, x_2, ... $x+d_1$, $x + d_2$, where x is the underlying regular measurement, while the ds are superimposed accidental components" (Spearman (1910), 289). Dette synet på sann skåre er truleg treffande karakterisert som eit platonisk synspunkt. Det byggjer på et postulat som korkje kan verifiserast eller falsifiserast.

Spearmans utgangspunkt i ein definisjon av sann skåre og feilskåre fører til konsekvensar for dei statistiske eigenskapar ved testane. Når sann skåre er gjennomgåande over testar og feilskåren er definert som ein slumpskåre, får dette følgjande konsekvensar:

$$\left.\begin{aligned} &\rho_{TE} = \rho_{EE'} = 0 \\ &\sigma_X^2 = \sigma_{X'}^2 \qquad (X = T + E\ ,\ X' = T + E') \\ &\rho_{XX'} = \rho_{XX''} = \rho_{X'X''} = \cdots\cdots \end{aligned}\right\} \qquad (F2)$$

Dette seier at feilskåren er ukorrelert med sann skåre og med feilskåre på ein annan test, at variansen er den same frå test til test, endeleg at interkorrelasjonane mellom slike testar er alle like. Testar med lik varians og like interkorrelasjonar kallar vi parallelle testar.

Den observerte skårevarians for ein test kan etter Spearmans definisjon skrivast slik:

$$\sigma_X^2 = \frac{\Sigma x^2}{N} = \frac{\Sigma (t + e)^2}{N} = \sigma_T^2 + \sigma_E^2 \qquad (F3)$$

Dette vil seia at total testvarians (observert skårevarians) er ein sum av to komponentar, variansen av sanne skårar og variansen av feilskårar eller feilvarians. Med berre ein test er det uråd å bestemma kor stor den sanne eller, om vi vil, den systematiske variansen er, difor heller ikkje kor stor del av den observerte varians som kan tilskrivast sytematisk varians. Vi har tidlegare sagt at forholdet mellom systematisk og observert varians kan stå som definisjon av reliabilitet. Med to testar, som per konsekvens av definisjonen av sann skåre og feilskåre blir parellelle, kan vi ved korrelasjon koma fram til eit estimat av den ikkje-observerbare sanne skårevarians ved observerbare storleikar.
Vi skriv:

$$\rho_{XX'} = \frac{\Sigma xx'}{N\sigma_X\sigma_{X'}} = \frac{\Sigma (t+e)\ (t+e')}{N\sigma_X\sigma_{X'}}$$

$$= \frac{\Sigma t^2/N + \Sigma te/N + \Sigma te'/N + \Sigma ee'/N}{\sigma_X\sigma_{X'}} \qquad (F4)$$

Det første uttrykket i teljaren i (F4) er sann skårevarians, dei tre andre er kovariansuttrykk og blir alle null etter (F2). $\sigma_X\sigma_{X'}$ i nemaren i (F4) blir etter (F2) lik σ_X^2, altså lik observert skårevarians.

Etter dette kan korrelasjonen mellom parallelle testar skrivast:

$$\rho_{XX'} = \frac{\sigma_T^2}{\sigma_X^2}\ , \qquad (F5)$$

dvs. korrelasjonen mellom to testar, her parallelle per konsekvens, gjev oss forholdet mellom sann skårevarians og observert skårevarians i <u>ein</u> slik test.

(F5) kan også skrivast slik:

$$\sigma_T^2 = \sigma_X^2 \rho_{XX'} \qquad \text{(F6)}$$

Etter (F6) blir sann skårevarians i ein test lik produktet av observert skårevarians i testen med korrelasjonen mellom parallelle testar.

3.1.1.2. Brown-Kelley tradisjonen.

Det platoniske element i synet på sann skåre i Spearman-Yule tradisjonen har gjort det vanskeleg for mange å godta omgrepet. Det let seg høyra å tala om den sanne vekt av ein stein eller den sanne avstand mellom to punkt. "This conception of true score does not, however, generally provide a satisfactory axiomatic basis for psychological theories since these theories are typically based on unexplicated, inexact constructs" (Lord. and Novick (1966), 39)

Spearman startar med å definera sann skåre og feilskåre, medan Brown tek utgangspunkt i ein definisjon av parallelle testar. Dette utgangspunktet blir vidare systematisert av Truman Kelley, m.a. i den kjente statistikkboka hans frå 1924 (Tryon (1957, 231). Dette synet postulerer at parallelle testar har lik varians og like interkorrelasjonar. Saman med dette utgangspunktet går eit meir operasjonelt syn på sann skåre. Ghiselli byggjer ein av sine reliabilitetsmodellar på eit liknande utgangspunkt, som han kallar "an eclectic concept of true scores and parallel tests". Han seier: "For some the notion of random error and the assumptions involved in the theory of true and error scores are too restrictive and tenuous, and therefore they prefer to approach the matter of reliability of measurement from a similar but more eclectic point of view. In this concept true scores are not conceived of as some quality inherent in the individual, but are merely taken as the average of an individual's score over an infinite number of parallel tests. Again true scores are an intellectual construct since we could never obtain scores of an individual over an infinite number of tests, but the construct is different from that of true scores in the concept of true and error scores" (vår Spearman-Yule tradisjon) (Ghiselli (1964), 230).

Trass i at omgrepet sann skåre kan te seg som eit urealistisk og mystisk omgrep slik at vi gjerne såg vi kunne greia oss det forutan, ser det likevel ut til at vi må ha det med i ei eller anna form. Om vi tenkte å koma oss unna omgrepet ved først å definera parallelle testar, tok vi feil; for sann skåre blir no ein konsekvens av definisjonen av parallelle testar. Men omgrepet er ikkje lenger platonisk. No er det rett og slett ein aritmetisk middelverdi. Lord og Novick seier at deira synspunkt "regards the notion of true score when given proper definition

as a very useful one conceptually and holds that many important practical results can be obtained by basing a theory of measurement on this concept. This is not metaphysics; we do not intend to produce a theory of measurement containing innumerable statements that are incapable of practical verification. The notion of true score is used because it yields tangible implications that can be verified in actual practice"
(Lord and Novick (1966), 37-38).

Med utgangspunkt i parallelle testar og eit operasjonelt syn på sann skåre skal vi no sjå korleis vi kan estimera den sanne skårevarians for endå ein gong å freista koma fram til eit mål på reliabilitet. Vi kan framleis ikkje rekna ut nokon sann skårevarians etter som vi aldri har tilgjengeleg eit uendeleg tal testskårar.

Vi tenkjer oss at vi for kvar person kan få tak i den sanne skåre ved å summera alle k skårane i universet av skårar og så dividera med k. Vi er vidare interessert i å finna variansen til desse sanne skårane som no er uttrykt ved aritmetiske middelverdier. Vi er likevel klår over at dette ikkje let seg gjera direkte: Middelverdiane er ikkje tilgjengelege, vi må få eit uttrykk for dei ved kjente storleikar.
Det let seg lett visa at kvar persons middelverdi uttrykt i avvikskåre frå total middelverdi kan skrivast som ein sum av avviksskårar på dei enkelte testane. Altså,

$$\bar{x} = \frac{x_1 + \ldots + x_k}{k} \qquad \text{(F7)}$$

Når vi kvadrerer begge siden av (F7) og summerer over personar, får vi ein kvadratsum:

$$\Sigma \bar{x}^2 = \frac{\Sigma (x_1 \ldots\ldots + x_k)^2}{k^2} \qquad \text{(F8)}$$

Utviklar vi (F8) og dividerer med N-1, får vi variansen til middelverdiane, som i dette tilfelle er variansen til dei sanne skårane:

$$\sigma_T^2 = \frac{\Sigma \bar{x}^2}{N-1} = \frac{1}{k^2} \frac{(\Sigma x_1^2 + \ldots + \Sigma x_k^2 + 2 \Sigma x_1 x_2 + \ldots + 2 \Sigma x_{k-1} x_k)}{N-1} \qquad \text{(F9)}$$

Ved å dividera kvar lekk inni parentesen i (F9) med N-1 får vi ut ei rekkje variansar og kovariansar:

$$\sigma_T^2 = \sigma_{\bar{x}}^2 = \frac{1}{k^2} \left(\sigma_1^2 + \ldots + \sigma_k^2 + 2\sigma_1\sigma_2 \rho_{12} + \ldots + 2\sigma_{k-1}\sigma_k \rho_{(k-1)k}\right) \qquad \text{(F10)}$$

Alle variansane i parantesen i (F10) er per definisjon like, det er og alle standardavvik og alle korrelasjonar. Difor kan (F10) skrivast slik, etter som det er k variansar og k (k-1) kovariansar i ein varians-kovarians matrise:

$$\sigma_T^2 = \frac{1}{k^2}(k\sigma_X^2 + k(k-1)\sigma_X^2 \rho_{XX'})$$

$$= \frac{1}{k}\sigma_X^2 + (\frac{k-1}{k})\sigma_X^2 \rho_{XX'} \qquad \text{(F11)}$$

Når vi reknar k for uendeleg, blir $\frac{1}{k} = 0$ og $\frac{k-1}{k} = 1$. Difor kan no (F11) skrivast:

$$\sigma_T^2 = \sigma_X^2 \rho_{XX'} \qquad \text{(F12)}$$

Med utgangspunkt i ein definisjon av parallelle testar finn vi at den sanne skårevarians, som ikkje er observerbar, kan estimerast ved produktet av observert skårevarians og observert korrelasjon mellom parallelle testar.

Vi har no vist at med to ulike teoretiske utgangspunkt har vi kome fram til eit og same estimat av sann skårevarians.

(F12) kan også skrivast:

$$\rho_{XX'} = \frac{\sigma_T^2}{\sigma_X^2}, \qquad \text{(F5)}$$

dvs. korrelasjonen mellom to testar, her parallelle per definisjon, gjev oss den sanne skårevarians i høve til observert skårevarians. Som vi ser, er vi framme ved (F5).

Konklusjonen på denne utleiing av syntaktiske definisjonar av reliabilitet med utgangspunkt i det vi har kalla ein Spearman-Yule tradisjon og ein Brown-Kelley tradisjon på sann skåre, blir då at begge utgangspunkt endar med same syntaktiske definisjon av reliabilitet.

"Whereas Spearman based his development on the true-score-plus-error assumption, Brown began by defining parallel tests. His approach leads to the same reliability theory as Spearman's. What is postulated by one is derived by the other; such small logical distinction between the theories as were once matters for contention no longer seem important. Whichever starting place is chosen, the true score turns out to be the limit of the mean observed score as the number of tests becomes indefinitely large"(Cronbach, Rajaratnam, Gleser (1963), 138).

Både Spearman-Yule tradisjonen og Brown-Kelley tradisjonen definerer reliabilitet som korrelasjonen mellom parallelle testar;den første per konsekvens,den andre per definisjon. Med parallelle testar forstår vi her at minst to konkrete parallelle testar finst,slik at reliabiliteten kan reknast ut ved å korrelera to slike testar.

Det er ein annan tradisjon som også definerer reliabilitet som korrelasjonen mellom parallelle testar,men som berre krev at vi har ein konkret test for hand.Denne testen må då vera eit kompositum,vi kallar,ein test samansett av minst to komponentar (t.d.halvtestar,subtestar,items),og vi må setja dei same statistiske krav til komponentane og relasjonane mellom dei som vi tidlegare sette til dei konkrete testane og relasjonane mellom dei.Dette vil seia at vi no krev parallelle komponentar.Dersom vi kjenner desse statistiske eigenskapane ved komponentane,kan vi estimera ein parallell-test korrelasjon mellom dette kompositum og eit hypotetisk parallelt kompositum.

Det er denne tradisjonen som i særleg grad er interessant når vi skal prøva å dra utviklingslinene fram til ein reformulert reliabilitetsteori,generalizability.Tradisjonen fører tilbake til Spearman og Brown som kvar for seg i eitt og same nummer av British Journal of Psychology (1910)(den eine sluttar sin artikkel på side 295,den andre tek til på side 296) utvikla denne kjente Spearman-Brown formelen ut frå teoretiske rammeverk godt som like,som vi har sett. Seinare har vi fått nye formlar,slike som Flanagan,Rulon, Guttman,Kuder-Richardson,Hoyt og Cronbachs alpha,som alle høyrer til denne tradisjonen,men som ikkje alle gjer like restriktive krav gjeldande som Spearman-Brown.

3.1.1.3.1. Dei originale Spearman-Brown formlane

a) Reliabiliteten til eit kompositum med to komponentar (Split-half formelen)

Tradisjonen med å korrelera ein konkret test med ein hypotetisk parallell test tok utgangspunkt i at den konkrete test kan delast i to parallelle halvtestar.Spearman og Brown ville predikera korrelasjonen med ein hypotetisk test med parallelle halvtestar 3 og 4 når vi berre har

data frå ein test med halvtestane 1 og 2.Dei starta med følgjande krav:

$$\left.\begin{aligned} &\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 \\ &\rho_{12} = \rho_{13} = \rho_{14} = \rho_{23} = \rho_{24} = \rho_{34} \end{aligned}\right\} \qquad \text{(F13)}$$

(F13) postulerer halvtestar med lik varians og like korrelasjonar mellom halvtestane,intratest som intertest.Vi korrelerer no dei to komposita,den første ein konkret test, den andre ein hypotetisk parallell test:

$$\rho_{(1+2)(3+4)} = \frac{\Sigma(x_1+x_2)(x_3+x_4)}{N\sigma_{(1+2)}\sigma_{(3+4)}} \qquad \text{(F14)}$$

Når vi multipliserer parantesane i (F14),summerer og etterpå dividerer med N,får vi ut fire kovariansar:

$$\rho_{(1+2)(3+4)} = \frac{\rho_{13}\sigma_1\sigma_3+\rho_{14}\sigma_1\sigma_4+\rho_{23}\sigma_2\sigma_3+\rho_{24}\sigma_2\sigma_4}{\sigma_{(1+2)}\sigma_{(3+4)}} \qquad \text{(F15)}$$

Vi ser av (F15) at ingen av kovariansane i teljaren kan bestemmast etter som det berre er kovariansar mellom konkrete og hypotetiske halvtestar.Men frå (F13) veit vi at alle variansar og kovariansar er like.Det vil seia at alle kovariansar i (F15) er like med den bestemmelege kovariansen $\rho_{12}\sigma_1\sigma_2$.Ergo kan (F15) skrivast:

$$\rho_{(1+2)(3+4)} = \frac{4\rho_{12}\sigma_1\sigma_2}{\sigma^2_{(1+2)}} \qquad \text{(F16)}$$

Men variansen i nemnaren i (F16) kan også skrivast som ein sumvarians.Altså,

$$\rho_{(1+2)(3+4)} = \frac{4\rho_{12}\sigma_1\sigma_2}{\sigma_1^2+\sigma_2^2+2\rho_{12}\sigma_1\sigma_2} \qquad \text{(F17)}$$

Med utgangspunkt i (F13) kan no (F17) reduserast slik:

$$\rho_{(1+2)(3+4)} = \frac{4\rho_{12}\sigma_1\sigma_2}{2\sigma_1^2+2\rho_{12}\sigma_1^2} = \frac{2\rho_{12}}{1+\rho_{12}} \qquad \text{(F18)}$$

Dette er Spearman-Browns split-half formel som gjev oss reliabiliteten til ein test av dobbel lengd med utgangspunkt i korrelasjonen mellom dei to parallelle halvtestane.Vi vil i vår samanheng gjerne poengtera at det er korrelasjonen mellom eit konkret test kompositum med komponentane 1 og 2

med eit hypotetisk parallelt test kompositum med komponentane 3 og 4.Denne korrelasjonen gjev oss då etter det vi har sett tidlegare,reliabiliteten til den konkrete testen.

b) Reliabiliteten til eit kompositum med k komponentar (Den generelle Spearman-Brown formelen)

(F18) er eit spesialtilfelle av ein generell formel.Denne generelle Spearman-Brown formel gjev oss reliabiliteten når vi forlengjer ein test k gonger,eller sagt på ein annan måte som er betre i vår samanheng: Vi korrelerer eit kompositum $(X_1+..+X_k)$ med eit hypotetisk parallelt kompositum $(X_1'+..+X_k')$ og set same krav til komponentane og relasjonen mellom dei som i (F13).Det vil seia:Alle komponentvariansar,intratest som intertest,er like.Det same gjeld alle komponentkorrelasjonar.

$$\rho_{(1+..+k)(1'+..+k')} = \frac{\Sigma\,(x_1+..+x_k)(x_1'+..+x_k')}{N\sigma_{(1+..+k)}\sigma_{(1'+..+k')}} \qquad \text{(F19)}$$

Når vi multipliserer parantesane i teljaren i (F19),får vi k^2 produktsummar.Når desse k^2 produktsummane blir dividert med N,får vi like mange kovariansar som alle er like,men ingen av dei kan bestemmast.Men vi veit at desse intertest kovariansane er like med intratest kovariansane,dei er definert like,slik at vi kan bestemma intertest kovariansane med ein intratest kovarians frå den konkrete testen.Nemnaren i (F19) er no produktet av to like standardavvik og kan skrivast som sumvariansen til den konkrete testen.Altså,

$$\rho_{(1+..+k)(1'+..+k')} = \frac{k^2\rho_{ij}\sigma_i^2}{k\sigma_i^2+k(k-1)\rho_{ij}\sigma_i^2} \qquad \text{(F20)}$$

I (F20) står fotskrift i og j for to komponentar frå det konkrete test kompositum. Etter reduksjon kan (F20) skrivast slik:

$$\rho_{(1+..+k)(1'+..+k')} = \frac{k\rho_{ij}}{1+(k-1)\rho_{ij}} \qquad \text{(F21)}$$

(F21) gjev oss korrelasjonen mellom eit konkret test kompositum med k parallelle komponentar og eit like langt hypotetisk parallelt kompositum.Altså blir (F21) reliabiliteten til det konkrete test kompositum.

3.1.1.3.2. Andre split-half formlar

Vi har utvikla SB (Spearman-Brown) split-half formel (F18) og generell SB formel (F21) på ortodokst klassiske vilkår. Etter kvart vart det klårt at det let seg gjera å utvikla alternative formlar som ikkje er fullt så restriktive i sine krav som SB. Vi skal her sjå på tre alternative split-half formlar,på deira forhold til SB og på deira innbyrdes forhold.

a) Flanagans formel

Etter Cronbach((1951)300) kan eit litt meir liberalt krav til ekvivalens (parallellitet) spesifiserast slik:

$$\left.\begin{array}{l} \sigma^2_{(1+2)} = \sigma^2_{(3+4)} \\ \rho_{12}\sigma_1\sigma_2 = \rho_{13}\sigma_1\sigma_3 = \rho_{14}\sigma_1\sigma_4 = \rho_{23}\sigma_2\sigma_3 = \rho_{24}\sigma_2\sigma_4 = \rho_{34}\sigma_3\sigma_4 \end{array}\right\} \quad \text{(F22)}$$

(F22) postulerer like kompositumvariansar og like kovariansar mellom komponentar,intratest som intertest eller intrakompositum som interkompositum.Denne måten å definera ekvivalens på kan førast tilbake til John Flanagan (Cronbach(1951)300).Etter denne definisjonen av parallelle komponentar vil Flanagans split-half formel bli lik (F17),altså lik SB split-half på eit visst steg i utleiinga.Dette impliserer at SB split-half formel må vera eit spesialtilfelle av Flanagans formel.Når vi i tillegg til krava i (F22) også kan stetta kravet $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2$, då blir Flanagan lik SB splithalf.Difor kan vi seia at Flanagan er ein meir generell formel enn SB split-half.Men vi bør då ha lagt merke til at vi har lempa litt på dei klassiske krav.

b) Rulons formel

I ein artikkel i Harvard Educational Review i 1939 gjer Rulon greie for ein relativt enkel framgangsmåte til å rekna split-half reliabilitet.Rulon reknar ut standardavviket til differensane $X_{1g} - X_{2g}$ og tek dette standardavviket som eit estimat av standardfeilen til den totale testskåren.

Rulons formel blir gjerne skriven slik:

$$\rho_{(1+2)(3+4)} = 1 - \frac{\sigma_d^2}{\sigma_X^2} \qquad \text{(F23)}$$

Vi skal her visa at Rulons formel er identisk med Flanagans. Etter som σ_d^2 er ein differensevarians og σ_X^2 ein sumvarians, kan (F23) skrivast:

$$\rho_{(1+2)(3+4)} = 1 - \frac{\sigma_1^2+\sigma_2^2-2\rho_{12}\sigma_1\sigma_2}{\sigma_1^2+\sigma_2^2+2\rho_{12}\sigma_1\sigma_2},$$

$$= \frac{(\sigma_1^2+\sigma_2^2+2\rho_{12}\sigma_1\sigma_2)-(\sigma_1^2+\sigma_2^2-2\rho_{12}\sigma_1\sigma_2)}{\sigma_1^2+\sigma_2^2+2\rho_{12}\sigma_1\sigma_2} \qquad \text{(F24)}$$

Ved reduksjon kan (F24) skrivast:

$$\rho_{(1+2)(3+4)} = \frac{4\rho_{12}\sigma_1\sigma_2}{\sigma_1^2+\sigma_2^2+2\rho_{12}\sigma_1\sigma_2} \qquad \text{(F17)}$$

Vi er tilbake til (F17), som er Flanagans formel. Rulons krav er såleis dei same som Flanagans. Dermed blir også SB split-half eit spesialtilfelle av Rulon.

c) Guttmans formel

Guttmans split-half formel ser slik ut:

$$\rho_{(1+2)(3+4)} = 2\left(1 - \frac{\sigma_1^2+\sigma_2^2}{\sigma_X^2}\right) \qquad \text{(F25)}$$

Vi skal visa at også denne split-half formelen er identisk med Flanagan, under ortodokst klassiske vilkår også identisk med SB split-half (F18). (F25) kan skrivast som følgjer:

$$\rho_{(1+2)(3+4)} = 2\left(\frac{(\sigma_1^2+\sigma_2^2+2\rho_{12}\sigma_1\sigma_2)-(\sigma_1^2+\sigma_2^2)}{\sigma_1^2+\sigma_2^2+2\rho_{12}\sigma_1\sigma_2}\right) \qquad \text{(F26)}$$

Ved reduksjon av (F26) får vi:

$$\rho_{(1+2)(3+4)} = \frac{4\rho_{12}\sigma_1\sigma_2}{\sigma_1^2+\sigma_2^2+2\rho_{12}\sigma_1\sigma_2} \qquad \text{(F17)}$$

Som vi ser, er vi endå ein gong tilbake til Flanagans formel (F17).

3.1.1.3.3. Kuder-Richardsons formel 20

Den minste komponent i eit test kompositum blir gjerne kalla eit item.Dersom vi gjer ekvivalenskrav gjeldande for items,får vi eit kompositum samansett av parallelle items. Med dette utgangspunkt let det seg gjera å estimera ein korrelasjon mellom eit test kompositum med k items med eit hypotetisk kompositum som også har k items.Både intra- og intertest items er definert parallelle.Det vil då seia at alle items har same varians og alle interkorrelasjonar er like.Det har vore mykje diskutert kva som eigenleg er dei opphavlege Kuder-Richardson krav.Cronbach,Rajaratnam og Gleser konkluderer slik i deira korte omtale av KR (Kuder-Richardson):"Thus the original derivation(s) of (both KR21 and) KR20 assumed equal item means,equal item variances,and singlefactoredness of items"(CRG(1963)140). Vi ønskjer her å utvikla KR20 på strengt klassiske krav for å ha denne utviklinga som ei referanseramme for seinare diskusjon.

Eit test kompositums totalvarians kan skrivast som ein sum av variansar og kovariansar:

$$\sigma_X^2 = \Sigma\sigma_i^2 + 2\Sigma\rho_{ij}\sigma_i\sigma_j \qquad \text{(F27)}$$

Når ekvivalenskrav blir gjort gjeldande,kan (F27) skrivast slik etter som vi har k variansar og k(k-1) kovariansar:

$$\sigma_X^2 = k\sigma_i^2 + k(k-1)\rho_{ij}\sigma_i^2 \qquad \text{(F28)}$$

Vi ønskjer no å isolera ρ_{ij} for å finna korrelasjonen,og vi får følgjande fasong på (F28):

$$\rho_{ij} = \frac{\sigma_X^2 - k\sigma_i^2}{k(k-1)\sigma_i^2} \qquad \text{(F29)}$$

(F29) gjev oss korrelasjonen mellom parallelle items,som per definisjon er reliabiliteten til eitt item.Vi har for hand eit test kompositum med k items og ønskjer å estimera korrelasjonen med eit hypotetisk parallelt kompositum.Det kan vi greia ved å nytta den generelle SB formelen:

$$\rho_{(1+..k)(1'+..+k')} = \frac{k\rho_{ij}}{1+(k-1)\rho_{ij}} \qquad \text{(F21)}$$

Vi set no inn (F29) i (F21) i staden for ρ_{ij} og får følgjande uttrykk:

$$\rho_{(1+..+k)(1'+..+k')} = \frac{k(\sigma_X^2-k\sigma_i^2)/k(k-1)\sigma_i^2}{1+(k-1)(\sigma_X^2-k\sigma_i^2)/k(k-1)\sigma_i^2}$$

$$= \frac{k(\sigma_X^2-k\sigma_i^2)/k(k-1)\sigma_i^2}{k(k-1)\sigma_i^2+(k-1)(\sigma_X^2-k\sigma_i^2)/k(k-1)\sigma_i^2}$$

$$= \frac{(\sigma_X^2-k\sigma_i^2)/(k-1)\sigma_i^2}{k\sigma_i^2+(\sigma_X^2-k\sigma_i^2)/k\sigma_i^2}$$

$$= \frac{(\sigma_X^2-k\sigma_i^2)/(k-1)\sigma_i^2}{\sigma_X^2/k\sigma_i^2}$$

$$= \frac{(\sigma_X^2-k\sigma_i^2)k/(k-1)}{\sigma_X^2}$$

$$= (\frac{k}{k-1})(\frac{\sigma_X^2-k\sigma_i^2}{\sigma_X^2})$$

$$\rho_{(1+..+k)(1'+..+k')} = (\frac{k}{k-1})(1-\frac{k\sigma_i^2}{\sigma_X^2}) \qquad \text{(F30)}$$

(F30) er KR20 i ei litt meir generell form enn den opphavlege KR20.Vi har ikkje avgrensa vår utleiing til berre å gjelda items som er skåra dikotomt,slik Kuder og Richardson gjorde.Den opphavlege KR20 blir skriven slik:

$$r_{tt} = \rho_{(1+..+k)(1'+..+k')} = (\frac{k}{k-1})(1-\frac{\Sigma p_i q_i}{\sigma_X^2}) \qquad \text{(F31)}$$

I (F31) er $\Sigma p_i q_i$ summen av alle itemvariansane når items blir skåra 1 eller 0.

3.1.1.3.4. Cronbachs alpha

KR20 i original utleiing og form byggjer på svært restriktive krav til data,som vi ser.Det er heilt urealistisk å rekna med at vi skal kunna stetta kravet om parallelle items.Heldigvis er det gong etter gong vist at KR20 byggjer på krav som er tilstrekkelege men ikkje nødvendige.(Sjå t.d. Tryon(1957))

Cronbach lanserte i 1951 ein generell KR20 som han har kalla alpha.Cronbach kom ikkje med noko ny ut-leiing av formelen.Han tek alpha for gjeven,og med det utgangspunktet prøver han å gje alpha meining.Cronbachs alpha ser slik ut:

$$\text{Alpha} = \left(\frac{k}{k-1}\right)\left(1 - \frac{\Sigma\sigma_i^2}{\sigma_X^2}\right) \qquad \text{(F32)}$$

I meir eklektisk testteori vil vi no nesten alltid finna KR20 skriven som (F32).Vi kan også koma fram til alpha ad variansanalytisk veg.Det var Hoyt(1941) som først viste at vi kan utvikla ein variansanalytisk formel som er identisk med KR20.Det hadde vore heilt på sin plass å ta med Hoyt i dette oversynet.Vi skal likevel venta med variansanalysen til eit seinare høve.Her gjer vi berre merksam på at Hoyt høyrer til i denne tradisjonen vi no har for oss.

Det er verdt å merka seg at på same måte som alpha er generell i høve til den originale KR20,så er denne formelen også generell i høve til Guttmans formel(F25).Dersom vi set k = 2,blir (F32) lik (F25).

Vi nemnde at det lenge har vore diskutert kva som er dei nødvendige og tilstrekkelege vilkår for utleiing av KR20 eller alpha,som vi no held oss til.Novick og Lewis har i ein artikkel i Psychometrika (1967) gjeve oss løysinga på dette problemet,men det ligg utanfor vår ramme å gjera greie for det her.

3.1.1.3.5. Generell Spearman-Brown som eit spesialtilfelle av alpha

Vi skal til slutt i dette oversynet som gjeld Spearman-Brown tradisjonen,visa at den generelle SB formelen (F21) er eit spesialtilfelle av alpha (F32).Med det får vi og vist kor homogen denne tradisjonen er.Dette poenget har kome for lite fram i tidlegare litteratur.

Vi skriv KR20 som alpha:

$$\text{Alpha} = \left(\frac{k}{k-1}\right)\left(1 - \frac{\Sigma\sigma_i^2}{\sigma_X^2}\right) = \left(\frac{k}{k-1}\right)\left(\frac{\sigma_X^2 - \Sigma\sigma_i^2}{\sigma_X^2}\right) \qquad \text{(F30)}$$

$(\sigma_X^2 - \Sigma\sigma_i^2)$ i (F30) er totalvariansen minus summen av itemvariansane.Men denne differensen er lik summen av alle kovariansane.(Sjå (F27)) På Flanagans vilkår,alle kovariansar like,kan då (F30) skrivast slik:

$$\text{Alpha} = \left(\frac{k}{k-1}\right)\left(\frac{k(k-1)\rho_{ij}\sigma_i\sigma_j}{\Sigma\sigma_i^2+k(k-1)\rho_{ij}\sigma_i\sigma_j}\right)$$

$$= \frac{k^2\rho_{ij}\sigma_i\sigma_j}{\Sigma\sigma_i^2+k(k-1)\rho_{ij}\sigma_i\sigma_j} \qquad \text{(F33)}$$

Stoggar vi ei stund ved (F33),vil vi kunna finna likskap med Flanagans formel (F17).Vi må ha lov å seia at (F33) er ein generell Flanagan .Dermed blir (F17) eit spesialtilfelle av (F33).Det skjer når k = 2.

Dersom vi går vidare no,og gjer ortodokst klassiske krav gjeldande på (F33),kan vi skriva:

$$\text{Alpha} = \frac{k^2\rho_{ij}\sigma_i^2}{k\sigma_i^2+k(k-1)\rho_{ij}\sigma_i^2} = \frac{k\rho_{ij}}{1+(k-1)\rho_{ij}} \qquad \text{(F21)}$$

Som vi ser,er vi med utgangspunkt i alpha og klassiske krav tilbake til (F21),den generelle SB.

Vi har i dette oversynet vore mest interessert i å visa korleis vi kan estimera ein korrelasjon mellom eit konkret test kompositum med frå 2 til k komponentar med eit hypotetisk parallelt test kompositum også med frå 2 til k komponentar.Vi har vidare vore interessert i å visa dette med basis i ein systematisk testteori,her den klassiske testteorien.

Vi har likevel merka oss at somme av dei formlane vi har teke for oss,på eit visst steg i utviklinga kan stå som sjølvstendige formlar med basis i noko lempelegare krav enn dei ortodokst klassiske Spearman-Brown krav.Vi augnar her spiren til ein liberaliseringstendens som meir og meir gjer seg gjeldande ved å vilja lempa på dei strenge krav som klassisk teori set til data,urealistiske krav som dei er.

Det er denne liberaliseringstendensen vi i særleg grad ønskjer å følgja vidare.

3.2. Semantiske definisjonar

Våre syntaktiske definisjonar går alle ut på at reliabilitet er korrelasjonen mellom parallelle testar.Vi bør ha det klårt for oss at vi så langt berre har definert reliabilitet statistisk.Omgrepet "parallelle testar" er til no definert ved statistiske kriteria,og berre ved slike.Reliabilitetsomgrepet har ikkje fått seg tillagt innhald og meining.Det skal vi freista gjera i det som følgjer.

Dei syntaktiske definisjonane legg ikkje restriksjonar på innhaldet i parallelle testar.Innhaldet kan vera identisk, innhaldet kan vera meir eller mindre ulikt,innhaldet kan vera heilt ulikt,og vi har framleis parallelle testar så sant desse testane har lik varians og like interkorrelasjonar.Difor må vi godta at parallelle testar kan implisera minst to applikasjonar av same test eller av meir eller mindre innhaldsmessig ulike testar. Det kan vera grunn til å minna om at parallelle testar ut frå teorien berre set syntaktiske krav og ikkje semantiske. Dette punktet blir lite presisert i litteraturen,stundom kan vi få inntrykk av at også semantiske krav må gjerast gjeldande på parallelle testar.Gulliksen seier:"In addition to satisfying these objective and quantitative criteria (equal means,variances,and intercorrelations),parallel tests should also be similar with respect to test content,item types,instructions to students,etc."(Gulliksen(1950),14) Til dette kommenterer Cronbach,Rajaratnam og Gleser:"This restriction on content is nowhere embodied in the mathematical model of classical theory; nothing in the classical mathematical assumptions prohibits each test from having some unique psychological content.Thus a series of compositions on diverse topics,used to appraise writing ability,would not have uniform content;it might nonetheless conform to the mathematical model."(Cronbach,Rajaratnam og Gleser(1963),139)

Dei syntaktiske definisjonane som byggjer på korrelasjon mellom konkrete parallelle testar,kan heller ikkje leggja nokon restriksjon på det eksperimentelle design når det gjeld tidsdimensjonen.Her må vi stå fritt om vi vil applisera parallelle testar så å seia samtidig eller med stort eller lite tidsintervall mellom.

Vi må etter dette kunna definera parallelle testar semantisk

med utgangspunkt i minst to dimensjonar samstundes:

1) grad av innhaldsmessig likskap mellom parallelle testar,
2) tidsintervallet mellom applikasjonar av parallelle testar.

I prinsippet har vi her for oss to kontinuerlege dimensjonar som gjev oss eit utal av kombinasjonar til eksperimentelt design.I praksis kan vi dikotomisera dei to dimensjonane til samtidig/ikkje samtidig og like testar/ulike testar.Vi sit då igjen med ein firefeltstabell som gjev oss dei fire kombinasjonane til ulike eksperimentelle design for prallelltest-korrelasjon.

	Like testar	Ulike testar
Ikkje samtidig	(a)	(b)
Samtidig	(c)	(d)

Fig.1 Eksperimentelle design for parallelltest-korrelasjon

Dei fire eksperimentelle design er utgangspunkt for parallell-testkorrelasjon

(a) når same test blir applisert med tidsintervall mellom,
(b) når ulike testar blir applisert med tidsintervall mellom,
(c) når same test blir applisert (minst to gonger) til same tid,
(d) når ulike testar blir applisert til same tid.

Alle desse fire kombinasjonane kan tenkjast når vi opererer med konkrete parallelle testar og når vi ikkje tolkar samtidig alt for trongt.Berre kombinasjonen (d) er tenkjeleg når vi korrelerer konkret test med hypotetisk parallell test.I dette tilfelle må vi tenkja oss ein samtidig applikasjon av dei to testane.Utan denne restriksjonen ville vi ikkje kunna estimera reliabilitet med mindre konstanskravet var stetta.

Vi har brukt omgrepet innhald utan å presisera kva det står for i vår samanheng.Med innhaldsmessig likskap mellom parallelle testar tenkjer vi først og fremst på at testane set krav til dei same evner.Dette registrerer vi indirekte ved å korrelera testskårane.Når testskårane korrelerer høgt,vil vi seia at testane er innhaldsmessig tolleg like.Er kor-

relasjonen låg,blir testane å rekna for innhaldsmessig heller ulike.

Ein test kan vera komponert slik at alle items i større eller mindre grad indikerer same trekk eller evne.Dette gjennomgåande trekk,eventuelt desse gjennomgåande trekk,kallar vi ein generell faktor eller generelle faktorar.I eit test kompositum kan vi og finna grupper av items som tappar trekk som andre items ikkje tappar.Vi kan finna fleire slike grupper av relativt høgt-korrelerande items i eit test kompositum,og dei trekk slike itemgrupper tappar,kallar vi gruppefaktorar.Items vil i tillegg til generell(e) faktor(ar) og/eller gruppefaktor(ar) også tappa trekk som er spesifikke for kvart item.

Innhaldsmessig likskap mellom parallelle testar blir såleis eit spørsmål om i kor stor grad slike testar måler dei same faktorane,og denne graden kan vi nærma oss ved korrelasjon.

Vi har før sett at korrelasjonen mellom parallelle testar under klassiske vilkår kan tolkast som forholdet mellom sann varians og testvarians,altså kor stor del av testvariansen som kan tilskrivast sann varians.Saman med sann varians går alltid ein komplementær storleik som vi kallar feilvariansen. Vi har tidlegare funne at testvariansen kan delast opp i to ortogonale (ikkje-korrelerte eller uavhengige) komponentar:

$$\sigma_X^2 = \sigma_T^2 + \sigma_e^2 \qquad (F3)$$

Dersom vi dividerer med σ_X^2 på begge sider i (F3),blir testvariansen lik 1 på venstre side,og på høgre side får vi to proporsjonar:

$$1 = \frac{\sigma_T^2}{\sigma_X^2} + \frac{\sigma_e^2}{\sigma_X^2} \qquad (F34)$$

Som vi ser,er første lekken på høgre side i (F34) ein av våre reliabilitetsdefinisjonar.Ved å venda litt på (F34) kan vi definera reliabilitet på ein alternativ måte.Vi kan skriva:

$$\rho_{XX'} = \frac{\sigma_T^2}{\sigma_X^2} = 1 - \frac{\sigma_e^2}{\sigma_X^2} \qquad (F35)$$

Den siste definisjonen i (F35),som er den nye,vil vi i særleg grad finna nyttig når vi no skal freista leggja meining i reliabilitetsomgrepet.

Ved bruk av parallelle testar for å estimera reliabilitet er det ein ulik diskrepans i intraperson skåreverdi over personar som gjer at korrelasjonen blir mindre enn 1.Denne ulike diskrepansen er grunnlaget for "unreliability".(Legg merke til at diskrepansen i skåreverdiane må variera frå person til person. Gjorde han ikkje det,ville reliabiliteten bli 1.) Alt etter det eksperimentelle design som blir nytta til innsamling av test-data,kan vi resonnera oss fram til dei variasjonskjelder som desse diskrepansane eventuelt kan tilskrivast.

Vi skal nærma oss slike diskrepanskjelder eller feilkjelder ad variansanalytisk veg ved å studera dei to dimensjonane vi kan ta omsyn til ved eksperimentelt design,ekvivalensdimensjonen (like testar/ulike testar) og tidsdimensjonen (samtidig/ ikkje samtidig applikasjon).

Guttman(1945) kom med eit vesentleg bidrag til reliabilitetsdiskusjonen ved sin djuptpløyande analyse av test-retest reliabilitet.Guttman(1945) saman med Cronbach(1947) er basis for vår teoretiske analyse av den semantiske sida ved reliabilitet.

Guttman tek utgangspunkt i feilvariansen.Skal vi analysera reliabilitet,må vi starta med ein presis definisjon av feil, seier han.Hans definisjon er:"The errors with which we are concerned are defined by an indefinitely large universe of trials.(For trials bruker vi applikasjonar.) They are defined separately for each individual in a population for each item (variable) being observed."(Guttman(1945),261)

Guttmans utgangspunkt i eit uendeleg univers av eksperimentelt uavhengige applikasjonar av same test kompositum og ein uendeleg populajon av personar kan illustrerast med følgjande matrise:

$$\left.\begin{array}{ll} (x_{111}+..+x_{1j1}+..+x_{1n1})+..+(x_{11k}+..+x_{1jk}+..+x_{1nk})+.. & T_1 \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & \\ (x_{i11}+..+x_{ij1}+..+x_{in1})+..+(x_{i1k}+..+x_{ijk}+..+x_{ink})+.. & T_i \end{array}\right\} \quad (F36)$$

$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$

$(i = 1,2,....),(j = 1,2,....,n),(k = 1,2,....)$

Vi tenkjer oss i (F36) eit test kompositum med n items applisert uendeleg mange gonger (eksperimentelt uavhengige applikasjonar) på ein uendeleg stor populasjon.Ein kva som helst item-

skåre x_{ijk} er ein observasjon av person i på item j ved applikasjon k.Ein itemskåre for ein person har ein forventa verdi over applikasjonar.Dette kan vi skriva slik:

$$\underset{k}{E}x_{ijk} = X_{ij} \qquad \text{(F37)}$$

Variansen for person i på item j over alle applikasjonar kan no skrivast:

$$\phi^2_{x_{ij}} = \underset{k}{E}(x_{ijk} - X_{ij})^2 \qquad \text{(F38)}$$

Denne variansen kallar Guttman feilvariansen til person i på item j.

Testskåren for person i ved applikasjon k (over items 1,2,..n) kallar vi t_{ik},og vi definerer observert testskåre,

$$t_{ik} = \sum_{j=1}^{n} x_{ijk} \qquad \text{(F39)}$$

Den forventa testskåre for person i over alle applikasjonar kallar vi T_i,og vi skriv,

$$\underset{k}{E}t_{ik} = T_i \qquad \text{(F40)}$$

Guttman definerer no feilvariansen på <u>testen</u> for person i slik:

$$\phi^2_{ti} = \underset{k}{E}(t_{ik} - T_i)^2 \qquad \text{(F41)}$$

(F41) gjev oss den intraindividuelle variasjon over eit univers av eksperimentelt uavhengige test-applikasjonar.

Gjennomsnittet av forventa testskårar,T_i,skriv vi u_T.Den interindividuelle varians av forventa testskårar kan no skrivast:

$$\phi^2_T = \underset{i}{E}(T_i - u_T)^2 \qquad \text{(F42)}$$

(F42) svarar til det vi tidlegare har kalla variansen til sanne skårar.

Endeleg må vi definera totalvariansen til testen over alle applikasjonar og alle personar.Etter som gjennomsnittet til testen over alle applikasjonar og alle personar må bli u_T (grand mean),kan totalvariansen skrivast:

$$\phi^2_t = \underset{ik}{EE}(t_{ik} - u_T)^2 \qquad \text{(F43)}$$

Men

$$\underset{ik}{EE}(t_{ik}-u_T)^2 = \underset{ik}{EE}((t_{ik}-T_i)+(T_i-u_T))^2$$

$$= \underset{ik}{EE}(t_{ik}-T_i)^2+\underset{i}{E}(T_i-u_T)^2 \qquad \text{(F44)}$$

(F44) seier at totalvariansen til testen over alle applikasjonar og alle personar er ein sumvarians av feilskåre og sann skåre. Dette svarar til det vi har funne i (F3).Det viktige for oss no er å leggja merke til kva Guttman i denne analysen reknar for feilvarians.(F41) definerer feilvariansen på testen.Dei minste komponentane i denne feilvariansen er diskrepansen mellom itemskåre og forventa itemskåre slik (F38) viser,$(x_{ijk}-X_{ij})$.Om vi no summerer over items,får vi,når vi bruker (F39) og (F40):

$$\underset{i}{E\phi^2_{ti}} = \underset{ik}{EE}(t_{ik}-T_i)^2 = \underset{ik}{EE}\Big(\sum_{j=1}^{n} x_{ijk}-\sum_{j=1}^{m} X_{ij}\Big)^2$$

$$= \sum_{j=1}^{n} \underset{ik}{EE}(x_{ijk}-X_{ij})^2 \qquad \text{(F45)}$$

Etter (F45) er feilvariansen definert som ein målingsfeil av items (item-feilvarians).

Vi går eit steg vidare.Vi tenkjer oss at vi kompliserer matrisen (F36) ved å dekomponera kvar itemskåre i ein generell komponent (g),ein gruppekomponent (f) og ein spesifikk komponent (s),slik at

$$x_{ijk} = x_{g(ijk)}+x_{f(ijk)}+x_{s(ijk)} \qquad \text{(F46)}$$

Når vi bruker (F39) og (F46),kan testskåren skrivast,

$$t_{ik} = \sum_{j=1}^{n} x_{g(ijk)}+\sum_{j=1}^{n} x_{f(ijk)}+\sum_{j=1}^{n} x_{s(ijk)}$$

$$= t_{g(ik)}+t_{f(ik)}+t_{s(ik)} \qquad \text{(F47)}$$

Den forventa komplekse testskåre over alle applikasjonar for person i skriv vi

$$T_i = T_{g(i)}+T_{f(i)}+T_{s(i)} \qquad \text{(F48)}$$

(F48) gjev oss sann g-skåre,sann f-skåre og sann s-skåre. Over alle applikasjonar får vi ved å nytta (F41),(F47) og (F48) ein intraindividuell varians i generell faktor,gruppefaktor og spesifikk faktor:

$$\phi^2_{ti}=\underset{k}{E}(t_{ik}-T_i)^2=\underset{k}{E}((t_{g(ik)}+t_{g(ik)}+t_{s(ik)})-(T_{g(i)}+T_{f(i)}+T_{s(i)}))^2,$$

som etter ordning gjev

$$\phi^2_{ti} = \underset{k}{E}(t_{ik} - T_i)^2$$

$$= \underset{k}{E}((t_{g(ik)}-T_{g(i)})+(t_{f(ik)}-T_{f(i)})+(t_{s(ik)}-T_{s(i)}))^2. \quad (F49)$$

Av (F49) ser vi at itemfeilvariansen kan skrivast som ein sum-varians av tre komponentar,ein item-feilvarians i måling av g, ein item-feilvarians i måling av f og ein item-feilvarians i måling av s.(For lettare å sjå dette,konferer (F45))

Den forventa interindividuelle varians kan også dekomponerast ved å bruka (F42) og (F48) og ved å bryta opp u_T i faktor-komponentar:

$$\underset{i}{E}(T_i-u_T)^2=\underset{i}{E}((T_{g(i)}+T_{f(i)}+T_{s(i)})-(u_{T(g)}+u_{T(f)}+u_{T(s)}))^2$$

$$=\underset{i}{E}((T_{g(i)}-u_{T(g)})+(T_{f(i)}-u_{T(f)})+(T_{s(i)}-u_{T(s)}))^2 \quad (F50)$$

Den totale testvarians over alle faktorar,over alle applikasjonar og over alle personar kan etter (F44),(F49) og (F50) skrivast:

$$\underset{ik}{EE}(t_{ik}-u_T)^2=\underset{ik}{EE}((t_{ik}-T_i)+(T_i-u_T))^2$$

$$=\underset{ik}{EE}(((t_{g(ik)}+t_{f(ik)}+t_{s(ik)})-(T_{g(i)}+T_{f(i)}+T_{s(i)}))+$$

$$((T_{g(i)}+T_{f(i)}+T_{s(i)})-(u_{T(g)}+u_{T(f)}+u_{T(s)})))^2$$

$$=\underset{ik}{EE}(((t_{g(ik)}-T_{g(i)})+(T_{g(i)}-u_{T(g)}))+$$

$$((t_{f(ik)}-T_{f(i)})+(T_{f(i)}-u_{T(f)}))+$$

$$((t_{s(ik)}-T_{s(i)})+(T_{s(i)}-u_{T(s)})))^2 \quad (F51)$$

I (F51) har vi dekomponert den totale testvariansen til ein sum av intraindividuelle og interindividuelle variansar i generell, gruppe- og spesifikk faktor.

Vi tenkjer oss no at vi har applisert ein og same test ei rekkje gonger til dei same personar.Vi tenkjer oss vidare at faktorskårane er tilgjengelege.Endeleg nå vi tenkja oss at item-feivariansen i dette konkrete tilfelle (sjå (F49)) kan dekomponerst endå meir:Vi kan rekna med mangel på konstans i generell,gruppe- og spesifikk faktor,og vi kan rekna med ein genuin målingsfeil,den tilfellelege feil.Etter dette skulle vi reint hypotetisk kunna sjå på den totale testvariansen som ein varianskomposisjon med følgjande komponentar:

$$\sigma_t^2 = \sigma_{T(g)}^2+\sigma_{T(f)}^2+\sigma_{T(s)}^2+\Sigma\sigma_{t(g)}^2+\Sigma\sigma_{t(f)}^2+\Sigma\sigma_{t(s)}^2+\sigma_e^2 \qquad \text{(F52)}$$

σ_t^2 er variansen til alle observerte testskårar (den totale testvarians) omkring u_T(grand mean).

$\sigma_{T(g)}^2$ er den interindividuelle varians i forventa generell faktorskåre(sann g-varians).

$\sigma_{T(f)}^2$ er den interindividuelle varians i forventa gruppefaktorskåre(sann f-varians).

$\sigma_{T(s)}^2$ er den interindividuelle varians i forventa spesifikk faktorskåre(sann s-varians).

$\sigma_{t(g)}^2$ er den intraindividuelle testvarians i g-faktoren.

$\sigma_{t(f)}^2$ er den intraindividuelle testvarians i f-faktoren.

$\sigma_{t(s)}^2$ er den intraindividuelle testvarians i s-faktorane.

σ_e^2 er item-feilvariansen (residual-variansen).

Dei tre første varianskomponentane representerer differensar mellom personar.Dei tre neste representerer instabilitet i dei tre faktorane.Den siste varianskomponenten i (F52) representerer den eigenlege feilmåling,"errors of measurement of items",som Cronbach kallar denne komponenten.(Cronbach(1947), 13) Denne varianskomponenten er i (F51) samanblanda(confounded) med dei intraindividuelle varianskomponentane.I (F52) skil vi ut item-feilvariansen på reint logisk grunnlag.

Med utgangspunkt i fig.1 og (F52) skulle vi no vera i stand til å leggja ei substansiell meining i reliabilitet alt etter kva slag design som er nytta ved innsamling av test-data.

(a) Vi appliserer same test med tidsintervall mellom.Feilvariansen vil i dette tilfelle måtta inkludera σ_e^2,$\sigma_{t(s)}^2$,$\sigma_{t(f)}^2$ og $\sigma_{t(g)}^2$,dvs.item-feilvariansen og instabilitetsvariansane for dei tre faktorane.

(b) Vi appliserer ulike testar med tidsintervall mellom.Her må vi også inkludera instabilitetsvariansane i feilvariansen i og med at vi ikkje har simultane applikasjonar.Item-feilvariansen er sjølvsagt også inkludert.I tillegg får vi under (b) $\sigma_{T(s)}^2$ med som feilvarians.Det skriv seg frå at vi bruker ulike testar,dvs.testar som har ulike items,slik at dei spesifikke faktorane vil bli ukorrelerte frå test til test.Testane kan også vera meir eller mindre ulike i gruppefaktorane.Er det

tilfelle,blir $\sigma^2_{T(f)}$ delvis å rekna for feilvarians,fordi gruppefaktorane vil bli mindre korrelerte frå test til test enn dei elles ville bli når dei same gruppefaktorane går igjen frå den eine testen til den andre.

(c) Vi appliserer like testar samtidig.I dette tilfelle vil vi ikkje ha noko eksperimentelt grunnlag for indikering av instabilitet.Feilvariansen blir no eit minimum;berre σ^2_e, item-feilvariansen,er med.

(d) Vi appliserer ulike testar samtidig.Heller ikkje under (d) har vi eksperimentelt grunnlag for ei indikering av instabilitet.Vi ser difor bort frå instabilitetsvariansane. Feilvariansen vil då i dette tilfelle inkludera σ^2_e og $\sigma^2_{T(s)}$. Som under (b) kan vi også her ha testar som er meir eller mindre ulike i gruppefaktorane,slik at $\sigma^2_{T(f)}$ delvis kan gå inn i feilvariansen.

Det er med basis i den teoretiske utvikling vi no har gjort greie for,at Cronbach(1947) kan gje tolleg eksakte semantiske definisjonar av reliabilitet.Til kvart av våre fire design svarar ein bestemt definisjon.Dette skulle gjera det heilt klårt at vi har meir enn ein reliabilitet når vi ser semantisk på reliabilitetsproblematikken.

Definisjon (a):
Reliability is the degree to which the test score indicates unchanging individual differences in any traits.

Definisjon (b):
Reliability is the degree to which the test score indicates unchanging individual differences in the general and the group factors defined by the test.

Definisjon (c):
Reliability is the degree to which the test score indicates individual differences in any traits at the present moment.

Definisjon (d):
Reliability is the degree to which the test score indicates the status of the individual at the present instant in the general and group factors defined by the test.

(Cronbach(1947),5-6)

Av dei fire definisjonane som her er sitert,identifiserer vi tre av dei ,(a),(b) og (d),med tre velkjente reliabilitets-

typar.Definisjon (a) svarar til stabilitet,(b) svarar til stabilitet og ekvivalens og (d) svarar til ekvivalens.Det er desse tre standardtypar av reliabilitet som no i lengre tid har vore mest brukt.

Definisjon (c) er ein teoretisk definisjon.Med det meiner vi at vi ikkje har eksperimentelle design til å estimera denne hypotetiske sjølvkorrelasjonen,som Cronbach kallar han.Vi kan berre nærma oss denne reliabilitetskoeffisienten via design som gjev oss eit underestimat,a lower bound,av koeffisienten. Det er Guttman(1945) som introduserer desse lower bounds til den definisjon av reliabilitet som han gjev.Vi trur det kan vera nyttig å sjå nærmare på definisjon (c) endå så unyttig han i og for seg er.Vi gjer det for di vi gjennom definisjon (c) har sjanse til å koma dette vi kallar feilvarians tettare innpå livet.

Definisjon (c) byggjer på eit hypotetisk design som krev samtidig applikasjon av same test,dvs.test-retest til same tid. Skal samtidige applikasjonar av same test vera praktisk gjennomførlege,vil vi ikkje kunna tolka "samtidig" etter bokstaven. Endå med små tidsintervall nå vi prinsipielt tolka korrelasjonen mellom slike parallelle testar som ein stabilitetskoeffisient. Tenkjer vi oss at vi stadig kortar inn tidsintervallet mellom applikasjonar av same test slik at vi nærmar oss genuint samtidige applikasjonar,får vi mindre og mindre grunnlag for ei indikering av instabilitet i testskårane.Diskrepansane vil etter kvart meir og meir kunna tilskrivast ein rein målingsfeil knytt til kvart iten.Vi ser at vi no nærmar oss Guttmans definisjon av feil.Når grensa er nått,genuint samtidige applikasjonar,vil kvart iten alltid få same skåre.Når same test blir gjeven til same tid,kan vi ikkje få nokon feilvarians. "In regard to the nonrepeating event which can be observed only once,reliability has only a theoretical interest. In fact,if one accepts a deterministic position,there is no "error" in a measurement of a unique event.The student's responses and his score are determined by many forces,and we do not know what they are;but the resultant of these forces is a particular act,and the act itself,at this instant and with these particular forces is perfectly reliable. "Chance" and "error" are merely names we give to our ignorance of what determines an event."

(Cronbach(1947),6)

Definisjon (c) svarar til Guttmans definisjon av reliabilitet, som både er ein syntaktisk og ein semantisk definisjon.

Syntaktisk ser Guttmans definisjon av reliabilitet slik ut: $1 - \Sigma\sigma^2_{ti}/\sigma^2_t$, der feilvariansen med basis i (F42) er semantisk definert som gjennomsnittet av dei individuelle feilvariansar over applikasjonar.(Guttman(1945),263)

"In deriving lower-bounds formulas,Guttman deals with hypothetical independent retests in which the mean covariance of two items within trials equals the mean covariance of the same items between trials.Beyond this he makes no assumption. His definition of independence requires that there be no shift in the variables measured between trials;i.e.that the hypothetical trials be simultaneous.Since he is using identical tests simultaneously,he has defined reliability as the hypothetical self-correlation."

(Cronbach(1947),10)

Coombs(1950) seier at til definisjon (c) svarar ein presisjonskoeffisient.Denne nemninga karakteriserer Guttmans reliabilitetsdefinisjon kanskje betre enn Cronbachs nemning gjer det. Ved eit seinare høve kom Cronbach endå ein gong inn på det vi her kallar definisjon (c).Han har no gått over til å bruka Coombs si nemning,og han definerer presisjonskoeffisienten slik: "A rigorous definition of the coefficient of precision is that it is the limit of the coefficient of stability as the time between testings becomes infinitesimal." (Cronbach (1951),307)

Presisjonskoeffisienten er som sagt ein teoretisk koeffisient som vi berre kan nærma oss med underestimat via inadekvate design.Denne hypotetiske koeffisienten gjev oss den absolutt minimale feil som kan finnast når same test blir applisert to gonger til same person.Det er denne koeffisienten som kjem nærmast ein reliabilitetsdefinisjon i fysisk måling,skulle vi tru,der det er realistisk å rekna med eksperimentelt uavhengige mål og der konstansproblemet ikkje langt frå er under kontroll. Presisjonskoeffisienten er såleis råd å få tak i i fysisk måling,og han gjev oss eit meiningsfylt reliabilitetsmål.

Det er lite truleg at vi i psykologisk måling har bruk for ein presisjonskoeffisient.Det er sjeldan det kan ha mykje for seg å få greie på "the accuracy with which the test measures whatever it measures." Nesten alltid må vi rekna med at det er meir enn item-feilvariansen, σ^2_e i (F52),som bør inkluderast i feilvariansen når vi for praktiske føremål estimerer reliabilitet.Eit psykologisk måleinstrument måler no ein gong ikkje berre ein ting.Vi har nett teke forbss generelle faktorar, gruppefaktorar og spesifikke faktorar i testar.Det er kanskje

ikkje for mykje sagt at det er uråd å konstruera testar som måler berre ein ting.Dette er ikkje så underleg når vi veit kor vanskeleg det er å definera det trekk vi vil måla og kor vanskeleg det er å finna testsampel som representerer det definerte trekk.Det må vera all grunn til å tvila på at vi skulle vera så visshøve ved utveljing av items til ein test at det spesifikke som desse items måler,svarar til vår definisjon av det trekk vi er ute etter.Difor er det svært rimeleg å rekna spesifikk varians for feilvarians,og det endå denne variansen strengt teke er reliabel varians.Cronbach seier: "There is no practical testing problem where the items in the test and only these items constitute the trait under examination."(Cronbach(1951),307) Dette problemet skal vi gå nærmare innpå ved eit seinare høve.

Vi bruker omgrepet feilvarians som ein fellesnemnar for ulike komponentar som går inn i dette vi kallar for feil.(F52) skulle gjera det klårt at dei komponentane som reknast til feilvariansen,representerer ulike typar av feil.Vi har sett at definisjonane (a) og (b) reknar instabilitetsvariansane som feilvarians.Ekvivalenskoeffisientane (b) og (d) reknar spesifikk varians til feilvariansen,stundom også gruppefaktorvarians.Når interindividuelle varianskomponentar går inn som feilvarians,har dette i grunnen ikkje med feilmåling å gjera. Det har samanheng med definisjonsproblemet og med problemet å finna definisjonsrelevante items eller grupper av items. Mangel på kongruens frå test til test (ulike testar) kallar Ekman(1947) for definisjonsfeil.I (F52) reknar vi $ø^2_{T(s)}$ og $ø^2_{T(f)}$ under visse vilkår som feilvarians,og dei representerer då definisjonsfeil.Det er reliabel varians,sann skårevarians i spesifikke faktorar og gruppefaktorar,men ikkje-relevant varians i høve til det trekk vi meiner å måla.

Det kan vera med å gjera feilvarians-omgrepet noko klårare dersom vi greier å skilja mellom målingsfeil og definisjonsfeil.Målingsfeilen representerer for så vidt dei same problem ved måling som andre vitskapar står framfor,t.d.dei fysiske. Definisjonsfeilen er derimot særmerkt for psykologisk måling og knyter seg til det som tidlegare er sagt om måling per definisjon.(Sjå side 5-6)

Vi samlar til slutt i ein tabell det vi har diskutert og sagt om den ulike meining reliabilitetsomgrepet kan ha alt

etter kva slag definisjon vi held oss til eller kva slag framgangsmåte vi bruker når vi estimerer reliabilitet.

Tabell 1

Feilvarians i ulike reliabilitetsdefinisjonar og eksperimentelle design*

	$\sigma^2_{T(g)}$	$\sigma^2_{T(f)}$	$\sigma^2_{T(s)}$	$\sigma^2_{t(g)}$	$\sigma^2_{t(f)}$	$\sigma^2_{t(s)}$	σ^2_e	
Test-retest				x	x	x	x	
Parallelle testar		(x)	x	x	x	x	x	
Parallell split			x	.....			x	
Random split		x	x	.....			x	
KR 20		x	x	.....			x	
Guttman L_2				.....			x	**
Presisjon				.....			x	
Ekvivalens		(x)	x	.....				
Stabilitet				x	x	x	x	
Stab og ekviv		(x)	x	x	x	x	x	

* Ein x indikerer at varianskomponenten er inkludert i feilvariansen

**I formel (31) og (43) set Guttman(1945) opp ulikskapar som overestimerer item-feilvariansen.

..... instabilitet ikkje teken omsyn til

4. Liberaliseringstendensar og retning reformulert reliabilitetsteori.

Kjernen i klassisk reliabilitetsteori er parallellitets- eller ekvivalensomgrepet.Reliabilitet definert som korrelasjonen mellom konkrete testar (Spearman-Yule og Brown-Kelley tradisjonen) krev testar som er like i middelverdi, varians og interkorrelasjonar.Ekvivalenskravet blir også gjort gjeldande for innhald.(Sjå R-22) Reliabilitet definert som korrelasjon mellom konkret test og hypotetisk parallell test krev i tillegg til nyss nemnde krav at halvtestane eller items skal vera parallelle.(Cronbach,Rajaratnam og Gleser(1963),137)

KR20 (og KR21) representerer truleg kulminasjonspunktet i klassisk testteori.Etter Kuder og Richardson kjem litt i senn liberaliseringstendensar til syne.Vi har så vidt nemnt denne tendensen medan han enno var i emning (Flanagans formel),ein tendens til å vilja liberalisera på dei strenge og urealistiske krav til ekvivalens.Men Flanagans liberalisering er for ingen ting å rekna mot det som skulle koma seinare.

Frå no av skal vi konsentrera oss om denne aukande tendens til liberalisering.I først omgang gjeld tendensen berre reliabilitet av type internal consistency (Spearman-Brown tradisjonen).

4.1. Jackson-Ferguson-Gulliksens utvikling av KR20.

Jackson og Fergusons monografi om reliabilitet frå 1941 representerer eit avgjerande steg bort frå dei heilt restriktive krav til parallellitet som ortodokst klassisk teori hevdar for internal consistency reliabilitet.Deira utgreiing gjev oss ingen ny teori.Dei berre viser at KR20 byggjer på vilkår som er tilstrekkelege men ikkje nødvendige.Men vi må ha lov å seia at Jackson og Ferguson går så langt i liberalisering at dei nære på kunne ha enda opp med ein ny teori.Vi reknar Jackson og Fergusons nye utvikling av KR20 historisk sett så interessant og så viktig at vi tek henne med her endå om ho ikkje er noko teoretisk nyskaping.Den utvikling vi viser,svarar til den vi finn i Gulliksen(1950),221-224.Etter det vi kan sjå,er Gulliksens

utvikling godt som lik Jackson og Fergusons.(Lord(1955),325)

Vi tek endå ein gong utgangspunkt i korrelasjonen mellom to summar,her eit konkret test kompositum med k komponentar (items) korrelert med eit hypotetisk parallelt test kompositum med like mange komponentar (items).

$$\rho_{XX'} = \frac{\Sigma(x_1+..+x_i+..+x_j+..+x_k)(x'_1+..+x'_i+..+x'_j+..+x'_k)}{N\sigma_X\sigma_{X'}} \qquad (F53)$$

Dei strenge krav til ekvivalens etter klassisk teori tilseier at items i X' blir matcha med items i X (test X' og test X), dersom vi skulle laga ein test X' som er parallell til test X. Difor er det rimeleg å rekna med at vi i vår hypotetiske korrelasjon skal få større samsvar mellom korresponderande items i test X og test X' enn mellom ikkje-korresponderande i dei to testane.Matcha items eller korresponderande items er items,eitt frå kvar test,med same fotskrift.X_i og X'_i er såleis korresponderande items,medan X_i og X'_j ikkje er det. Det er verdt å merka seg at items frå same test er ikkje-korresponderande items.

Etter dette bør vi i teljaren i (F53) skilja mellom to typar av kovarians.Det er k kovariansar mellom korresponderande eller parallelle items og k(k-1) kovariansar mellom ikkje-korresponderande items.

$$\rho_{XX'} = \frac{\sum_{i=1}^{k}\rho_{ii'}\sigma_i\sigma_{i'} + \sum_{i=1}^{k}\sum_{j=1}^{k}\rho_{ij'}\sigma_i\sigma_{j'}}{\sum_{i=1}^{k}\sigma_i^2 + \sum_{i=1}^{k}\sum_{j=1}^{k}\rho_{ij}\sigma_i\sigma_j} \qquad (F54)$$

Etter som vi korrelerer parallelle testar,er $\sigma_X = \sigma_{X'}$.Difor skriv vi nemnaren i (F54) σ_X^2,som også kan skrivast som ein sum av variansar og kovariansar.

Første lekken i teljaren i (F54) kan skrivast som $\sum_{i=1}^{k}\rho_{ii}\sigma_i^2$, fordi σ_i og σ'_i er like standardavvik i parallelle items og fordi korrelasjonen mellom to parallelle items,$\rho_{ii'}$,er reliabiliteten til eitt item,ρ_{ii}.Det er rimeleg å rekna den gjennomsnittlege kovarians mellom ikkje-korresponderande items lik anten ikkje-korresponderande items er frå same test eller frå ulike testar,slik at $\Sigma\rho_{ij}\sigma_i\sigma_j = \Sigma\rho_{ij'}\sigma_i\sigma_{j'}$. Difor kan siste lekken i teljaren i (F54) skrivast som ein sum av kovariansar mellom ikkje-korresponderande items frå den konkrete testen.

Etter dette kan (F54) skrivast slik:

$$\rho_{XX'} = \frac{\sum_{i=1}^{k}\rho_{ii}\sigma_i^2 + \sum_{i=1}^{k}\sum_{j=1}^{k}\rho_{ij}\sigma_i\sigma_j}{\sum_{i=1}^{k}\sigma_i^2 + \sum_{i=1}^{k}\sum_{j=1}^{k}\rho_{ij}\sigma_i\sigma_j} \qquad \text{(F55)}$$

(F55) estimerer (F53) med statistiske storleikar som alle er henta frå den konkrete testen så nær som ρ_{ii}.

Vi veit at

$\sigma_X^2 = \sum_{i=1}^{k}\sigma_i^2 + \sum_{i=1}^{k}\sum_{j=1}^{k}\rho_{ij}\sigma_i\sigma_j$, dvs. testvariansen er ein sum av kovariansar og variansar. Difor kan vi og skriva

$\sum_{i=1}^{k}\sum_{j=1}^{k}\rho_{ij}\sigma_i\sigma_j = \sigma_X^2 - \sum_{i=1}^{k}\sigma_i^2$. (F55) kan etter dette skrivast

$$\rho_{XX'} = \frac{\sum_{i=1}^{k}\rho_{ii}\sigma_i^2 + \sigma_X^2 - \sum_{i=1}^{k}\sigma_i^2}{\sigma_X^2} \qquad \text{(F56)}$$

Første lekken i (F56) kan ikkje bestemmast etter som vi her har bruk for korrelasjonen mellom parallelle items. Denne korrelasjonen kan vi berre få tak i ved å konstruera ein parallell test, og den prosedyren høyrer ikkje med til denne type reliabilitet. Skal vi koma vidare med (F56), må vi på eikor vis estimera første lekken i teljaren i (F56).

Det er viktig å merka seg kva vilkår Jackson og Ferguson set for å estimera denne første lekken i teljaren i (F56): Noko uventa reknar dei den gjennomsnittlege kovarians mellom korresponderande items lik den gjennomsnittlege kovarians mellom ikkje-korresponderande items:

$$\sum_{i=1}^{k}\rho_{ii}\sigma_i^2/k = \sum_{i=1}^{k}\sum_{j=1}^{k}\rho_{ij}\sigma_i\sigma_j/k(k-1)$$

$$\sum_{i=1}^{k}\rho_{ii}\sigma_i^2 = \sum_{i=1}^{k}\sum_{j=1}^{k}\rho_{ij}\sigma_i\sigma_j/(k-1) \qquad \text{(F57)}$$

(F57) kan også skrivast (kfr. same side lenger oppe)

$$\sum_{i=1}^{k}\rho_{ii}\sigma_i^2 = (\sigma_X^2 - \sum_{i=1}^{k}\sigma_i^2)/(k-1) \qquad \text{(F58)}$$

Ved å nytta (F58) kan (F55) skrivast slik:

$$\rho_{XX'} = \frac{(\sigma_X^2 - \sum_{i=1}^{k}\sigma_i^2)/(k-1) + \sigma_X^2 - \sum_{i=1}^{k}\sigma_i^2}{\sigma_X^2} \qquad (F59)$$

$$= \frac{(\sigma_X^2 - \sum_{i=1}^{k}\sigma_i^2) + (k-1)\sigma_X^2 - (k-1)\sum_{i=1}^{k}\sigma_i^2}{(k-1)\sigma_X^2} \qquad (F60)$$

$$= \frac{\sigma_X^2 - \sum_{i=1}^{k}\sigma_i^2 + k\sigma_X^2 - \sigma_X^2 - k\sum_{i=1}^{k}\sigma_i^2 + \sum_{i=1}^{k}\sigma_i^2}{(k-1)\sigma_X^2} \qquad (F61)$$

$$\rho_{XX'} = \frac{k\sigma_X^2 - k\sum_{i=1}^{k}\sigma_i^2}{(k-1)\sigma_X^2} = \left(\frac{k}{k-1}\right)\left(1 - \frac{\sum_{i=1}^{k}\sigma_i^2}{\sigma_X^2}\right) \qquad (F62)$$

Vi kjenner igjen (F62) som KR20 eller alpha,som vi denne gongen har utvikla på mykje lempelegare vilkår enn tidlegare. Det einaste krav vi no har gjort gjeldande,er at den gjennomsnittlege kovariansen er lik i parallelle testar.

I denne utviklinga er det likevel postulert ein identitet som ikkje er særleg plausibel.Vi hugsar at Jackson og Ferguson hadde vanskar med å estimera kovariansen mellom korresponderande items og at dei måtte ty til ei løysing som knapt nok er godtakande.Med utgangspunkt i "item-parallelle" testar postulerer dei at gjennomsnittleg kovarians mellom korresponderande items (parallelle items) er lik gjennomsnittleg kovarians mellom ikkje-korresponderande items.Det rimelege ville her vera å rekna $\overline{\rho_{ii}\sigma_i^2} > \overline{\rho_{ij}\sigma_i\sigma_j}$, men som vi har sett før,med denne ulikskapen er korrelasjonen mellom item-parallelle testar ikkje bestemmeleg. (F62) må reknast som eit underestimat av (F56).Dersom Jackson og Ferguson hadde greitt å frigjera seg frå item-parallelle testar,kunne vi i denne utviklinga ha fått nye teoretiske synspunkt fram.

Tryons omtale av Gulliksens utvikling av KR20,som vi identifiserer med Jackson og Fergusons utvikling,konkluderer med å seia at "he nearly breaks free,for he derives the KR20 formula without the crippling factorial and other restrictive assumptions of its original authors."(Tryon(1957),247)

4.2. Lords teori om random-parallelle testar.

Medan Jackson og Fergusons utvikling av KR20 er ei reutvikling som med velviljug tolking kanskje kan seiast å implisera nye teoretiske synspunkt,er Lords utvikling av KR21 eit første,eksplisitt gjennombrot av ein ny testteori.(Lord(1955))

Før vi går nærmare inn på Lords synspunkt,skal vi som snarast sjå på KR21 og ein variant av denne formelen.

Vi har tidlegare gjort oss kjende med KR20 i ei litt meir generell form enn den opphavlege.(Sjå R-19) Kuder og Richardson baserte si utvikling på dikotome items og skreiv formel 20 som

$$\rho_{XX'} = \left(\frac{k}{k-1}\right)\left(\frac{\sigma_X^2 - k\overline{pq}}{\sigma_X^2}\right) \qquad \text{(KR20)} \qquad \text{(F63)}$$

der $\overline{pq}$ er den gjennomsnittlege item-varians.

Når item-vanskegrad er stort sett den same i ein test,kan $k\overline{pq}$ tilnærma skrivast som $k\bar{p}\bar{q}$,der $\bar{p}\bar{q}$ er produktet av den gjennomsnittlege p og den gjennomsnittlege q.

$$\rho_{XX'} = \left(\frac{k}{k-1}\right)\left(\frac{\sigma_X^2 - k\bar{p}\bar{q}}{\sigma_X^2}\right) \qquad \text{(KR21)} \qquad \text{(F64)}$$

Andre lekken i teljaren i (F64) kan også skrivast

$$k\bar{p}\bar{q} = k\sum_{i=1}^{n} t_i/kn(kn - \sum_{i=1}^{n} t_i)/kn \qquad \text{(F65)}$$

t_i er testskåre for person i. $\sum_{i=1}^{n} t_i$ er summen av alle testskårar,eller talet på rette items over personar i person-item matrisen.

$$k\bar{p}\bar{q} = \sum_{i=1}^{n} t_i/n(k - \sum_{i=1}^{n} t_i/n)/k \qquad \text{(F66)}$$

$\sum_{i=1}^{n} t_i/n = \bar{t}_i$ (den gjennomsnittlege testskåre)

(F66) kan no skrivast

$$k\bar{p}\bar{q} = \frac{\bar{t}_i(k - \bar{t}_i)}{k} \qquad \text{(F67)}$$

Ved å nytta (F67) kan (F64) skrivast

$$\rho_{XX'} = \left(\frac{k}{k-1}\right)\left(1 - \frac{\frac{\bar{t}_i(k - \bar{t}_i)}{k}}{\sigma_X^2}\right) = \left(\frac{k}{k-1}\right)\left(1 - \frac{\bar{t}_i(k - \bar{t}_i)}{k\sigma_X^2}\right) \qquad \text{(F68)}$$

(F68) er KR21 skriven på ein annan måte enn (F64).Det er ein praktisk variant som er lett å rekna når vi snøgt vil ha eit tilnærma estimat av ein ekvivalenskoeffisient.

Vi skal no sjå korleis Lord kjem fram til (F68) på heilt andre vilkår enn dei Kuder og Richardson byggjer på.
Det er truleg i og for seg ingen original tanke å sjå på items i ein test som eit tilfelleleg utval av items frå eit univers av items.Dette synspunktet kan vi ha møtt lenge før Lord(1955), men han er den første som har nytta dette synspunktet i ein test-teoretisk samanheng.Lord har kome med ein ny definisjon av ekvivalens med utgangspunkt i det han kallar "randomly parallel tests" (heretter kalla random-parallelle testar). Dersom items i to eller fleire testar kan seiast å vera trekte frå eitt og same item-univers (pool of items),blir testane kalla random-parallelle testar.(Lord(1955),328)

I eit univers av dikotome items tenkjer vi oss at vi for ein person i har eit parameter p_i,person i's proporsjon av items som han greier eller får rett svar på.Denne p_i-verdien veit vi ikkje.Vi kan berre estimera verdien ved å finna kor mange items i eit tilfelleleg sampel av items person i greier. Dersom vi har k items i sampelet og person i's skåre (talet på rette items) er t_i,blir den estimerte p_i-verdien lik t_i/k.

Den teoretiske fordeling av skårar basert på eit univers av random-parallelle testar for person i,alle testar med k dikotome items,får etter binomial-teori ein middelverdi lik kp_i og ein varians lik kp_iq_i.Desse parameter kan berre estimerast med statistiske storleikar.

Vi er interesserte i å estimera kp_iq_i,som gjev oss variansen i universet av random-parallelle testar,alle med k items,for person i som i universet har sjanse til å lukkast lik p_i. Denne variansen er ein sampling-varians,eller ein feilvarians for å halda oss til test-teoretisk terminologi.Denne feilvariansen skriv vi her $\sigma^2_{t(i)}$.

$$\sigma^2_{t(i)} = kp_iq_i \qquad \text{(F69)}$$

$$\sigma^2_{t(i)(est)} = kt_i/k(1 - t_i/k) = \frac{t_i(k - t_i)}{k} \qquad \text{(F70)}$$

Eit uhilda estimat (an unbiased estimate) av $\sigma^2_{t(i)}$ får vi ved å multiplisera (F70) med k/(k-1).(Sjå t.d.Guilford(1965))

$$\sigma^2_{t(i)} = \frac{t_i(k - t_i)k}{k(k-1)} = \frac{t_i(k - t_i)}{k-1} \qquad (F71)$$

(F71) er den estimerte feilvarians over random-parallelle testar for person i.

Når vi estimerer reliabilitet,er feilvariansen den gjennomsnittlege feilvarians over personar.

$$\sigma^2_e = 1/n \sum_{i=1}^{n} \left(\frac{t_i(k-t_i)}{k-1}\right) = 1/n(k-1) \sum_{i=1}^{n} t_i(k-t_i)$$

$$= 1/(k-1)\left(\frac{\sum_{i=1}^{n} t_i k - \sum_{i=1}^{n} t_i^2}{n}\right) = 1/(k-1)\left(\bar{t}_i k - 1/n \sum_{i=1}^{n} t_i^2\right) \qquad (F72)$$

I (F72) kan $1/n \sum_{i=1}^{n} t_i^2$ skrivast som $\sigma^2_X + \bar{t}_i^2$,når vi i første lekken går frå kvadratsummen til variansen ved å dividera med n i staden for n-1.Difor er σ^2_X i dette tilfellet eit uhilda estimat av populasjonsvariansen berre når n er stor.Vi skriv no (F72)

$$\sigma^2_e = 1/(k-1)\left(\bar{t}_i k - (\sigma^2_X + \bar{t}_i^2)\right) \qquad (F73)$$

Korrelasjonen mellom random-parallelle testar kan no skrivast

$$\rho_{XX'} = 1 - \frac{\sigma^2_e}{\sigma^2_X} = 1 - \left(\frac{1/(k-1)\left(\bar{t}_i k - (\sigma^2_X + \bar{t}_i^2)\right)}{\sigma^2_X}\right)$$

$$= \frac{\sigma^2_X - 1/(k-1)\left(\bar{t}_i k - (\sigma^2_X + \bar{t}_i^2)\right)}{\sigma^2_X}$$

$$= \frac{(k-1)\sigma^2_X - \left(\bar{t}_i k - (\sigma^2_X + \bar{t}_i^2)\right)}{(k-1)\sigma^2_X}$$

$$= \frac{k\sigma^2_X - \sigma^2_X - \bar{t}_i k + \sigma^2_X + \bar{t}_i^2}{(k-1)\sigma^2_X} = \frac{k\sigma^2_X - \bar{t}_i k + \bar{t}_i^2}{(k-1)\sigma^2_X}$$

$$= \frac{k(k\sigma^2_X - \bar{t}_i k + \bar{t}_i^2)}{k(k-1)\sigma^2_X} = \left(\frac{k}{k-1}\right)\left(\frac{k\sigma^2_X - \bar{t}_i k + \bar{t}_i^2}{k\sigma^2_X}\right)$$

$$\rho_{XX'} = \left(\frac{k}{k-1}\right)\left(\frac{k\sigma^2_X - \bar{t}_i(k - \bar{t}_i)}{k\sigma^2_X}\right) = \left(\frac{k}{k-1}\right)\left(1 - \frac{\bar{t}_i(k - \bar{t}_i)}{k\sigma^2_X}\right) \qquad (F74)$$

(F74) er identisk med (F68) som vi utvikla som ein KR21 variant.Dermed har vi vist at det let seg gjera å koma fram til KR21 på heilt andre vilkår enn dei Kuder og Richardson sette.

Medan Kuder og Richardsons vilkår er at alle items har same vanskegrad,byggjer Lords utvikling på følgjande vilkår:

1) Vi er interesserte i korrelasjonen mellom random-parallelle testar,

2) vi er viljuge til å bruka eit hilda estimat av feilvariansen og

3) talet på personar er stort.

(Lord(1955),329)

4.3. Tryons reliabilitetsteori.

Med Tryons reliabilitetsteori tek vi det avgjerande steg som Jackson-Ferguson-Gulliksen ikkje tok:Vi går bort frå og riv oss laus frå omgrepet item-parallelle testar.Det er ei frigjering frå krav som gjeld komponentane i eit test kompositum,med andre ord ei liberalisering av Spearman-Brown tradisjonen.Når komponentane etter denne tradisjonen stetta strenge statistiske krav til ekvivalens,kunne vi med utgangspunkt i desse komponentane estimera korrelasjonen mellom hypotetisk parallelle testar.I Spearman-Yule tradisjonen og i Brown-Kelley tradisjonen blir reliabiliteten estimert ved reint konkret å korrelera parallelle testar.

Etter Tryons teori kan vi estimera korrelasjonen mellom parallelle testar med utgangspunkt i dei observerte statistiske eigenskapar ved komponentane i det konkrete test kompositum. Vi set ikkje krav til komponentane i det heile.Tryons vilkår er berre knytte til hypotetiske komposita.

Tryon gjorde greie for reliabilitetsteorien sin i ein artikkel i Psychological Bulletin i 1957.Vi må vel seia at teorien har vore lite kjent.Først no i seinare år ser det ut til at dei synspunkt som Tryon hevda,har fengt og stimulert til radikal tenking i testteorien.Ser vi Tryons teori i eit historisk perspektiv,er det klårt at denne teorien representerer eit viktig steg i ei utvikling mot eit fullt og heilt liberalisert og reformulert reliabilitetsomgrep,ei utvikling som førebels stoggar ved generalizability.

Vi skal i det følgjande ta for oss Tryons reliabilitetsteori slik han sjølv har gjort greie for teorien (Tryon(1957)) og slik Ghiselli(1964) freistar å forklara Tryons synspunkt meir utførleg enn det var gjort i artikkelen som introduserte dei nye tankane.

Det er naturleg for vårt føremål med ei to-deling av Tryons teori.Hans domene-sampling fell lagleg inn i det vi kan kalla ein KR20 tradisjon som går på dei vilkår denne formelen byggjer på.Tryons domene-validitet er eit nytt omgrep som tilfører tradisjonell reliabilitet ny meining og som i særleg grad peikar framover.

4.3.1. Domene-sampling.

Når ein psykologisk test skal konstruerast,er vi som regel i den situasjon at det vi ønskjer å måla,i verste fall berre er ein intuisjon av noko vi trur kan vera eit trekk eller ein dimensjon,i beste fall ein velfundert og veldefinert dimensjon.I alle fall må dette hypotetiske construct eksplikerast ved å finna åtferdseiningar innanfor ein åtferdsdomene som vi kan rekna med speglar av denne dimensjonen.Ein domene er i Tryons system eit definert område av åtferd der ein bestemt dimensjon er eit gjennomgåande trekk.Når vår dimensjon skal eksplikerast,må vi rekna med at det kan vera så mange måtar å gjera dette på og så mangt og mykje av åtferdseiningar å velja i at dei items som vi av ein eller annan grunn bestemmer oss for å bruka,gjerne kan bli sett på som eit tilfelleleg testsampel frå eit heilt univers av items innanfor vår domene.Dei items vi har valt,er berre eitt av eit utal andre sampel som vi kunne ha valt.Dette eine sampelet av items er det instrument som skal hjelpa oss med å tappa det trekk vi ønskjer å måla.

Det synspunkt vi her hevdar,tykkjest å vera i godt samsvar med Torgersons karakteristikk av den type psykologisk måling som han kallar måling per definisjon.(Torgerson(1958)) (Sjå s.5-6)

Skårane for ulike personar på dette eine sampel av items skal fortelja oss om dei individuelle differensane på det trekket vi måler.Det er naturleg å spørja i kor stor grad vi kan venta at skårane på eit nytt sampel av like mange items vil gje noko nær same rangering av personane som det første sampelet.For å få eit svar på dette spørsmålet ligg den tanke nær at vi korrelerer det første sett av observerte skårar,X_t,med eit nytt sett av skårar,X'_t.

Tryon ser ikkje på dette nye sampel av items,X'_t,som ein konkret test.X'_t er ein teoretisk konstruksjon som er definert med utgangspunkt i visse statis-tiske eigenskapar ved det første sampel,eller det konkrete test koposítum for hand.Definisjonen lyder:

"A comparable X'_t composite is one whose k test-samples (items) vary on the average as much in variances and inter-

correlations as do the k test-samples (items) in the observed X_t composite." (Tryon(1957),231)

Dei definerte krav til eit kommensurabelt kompositum er desse:

$$k' = k \qquad (F75)$$

$$\bar{\sigma}^2_{i'} = \bar{\sigma}^2_i \qquad (F76)$$

$$\bar{c}_{i'j'} = \bar{c}_{ij} \qquad (F77)$$

(F75),(F76) og (F77) seier at eit kommensurabelt kompositum har like mange items,same gjennomsnittlege item-varians og same gjennomsnittlege item-kovarians som det konkrete test kompositum.

Når det observerte kompositum,X_t,er samansett av komponentar som ikkje er ordna eller grupperte,kallar vi testen eit ikkje-stratifisert kompositum.Komponentane får ei tilfelleleg ordning fordi vi reknar som om dei blir sampla frå eit item-univers.

Vi held oss her berre til ikkje-stratifiserte komposita.I vårt hypotetisk kommensurable kompositum,X'_t,vil same tilfellelege ordning gjelda som i det konkrete kompositum,og begge sett av items blir rekna for tilfellelege sampel frå same item-univers.Difor er det rimeleg at den gjennomsnittlege inter-test item-kovarians,$\bar{c}_{ij'}$,blir lik den gjennomsnittlege intra-test item-kovarians,$\bar{c}_{ij}$ og $\bar{c}_{i'j'}$.For ikkje-stratifiserte testar kan etter dette (F77) utvidast til også å gjelda gjennomsnittleg inter-test kovarians:

$$\bar{c}_{i'j'} = \bar{c}_{ij'} = \bar{c}_{ij} \qquad (F78)$$

Konsekvensen av definisjonane (F75),(F76) og (F78) er at variansen i alle kommensurable komposita må vera den same som i det observerte test kompositum:

$$\sigma^2_{X'} = k'\bar{\sigma}^2_{i'} + k'(k'-1)\bar{c}_{i'j'} = k\bar{\sigma}^2_i + k(k-1)\bar{c}_{ij} = \sigma^2_X \qquad (F79)$$

Ein annan konsekvens av definisjonane er at alle kommensurable komposita har like interkorrelasjonar.

Det er ei rimeleg innvending å reisa at det er lite realistisk å rekna med at ei random sampling av items skulle gje komposita med like variansar og interkorrelasjonar.Dette er eit empirisk spørsmål.Tryons kommensurable komposita er

hypotetiske,og ei slik innvending kan ikkje gjelda teorien per se.Teorien definerer eit univers av kompositа som alle har lik varians og like interkorrelasjonar.Dette er ein restriksjon som gjer det nødvendig å definera vår åtferdsdomene ved det sampel av items vi har valt.Vår åtferdsdomene blir såleis å rekna for ein hypotetisk domene som er karakterisert ved eit univers av items som gjennomsnittleg har dei same statistiske eigenskapar som det gjennomsnittlege item i det konkrete kompositum.

"Clearly implicit both in the theory of true and error scores (Spearman-Yule tradisjonen) and in the eclectic concept of true scores and parallel tests(Brown-Kelley tradisjonen) is the notion that it is possible to develop a set of actual tests which precisely meet the mathematical criteria of parallelism.Indeed,a set of actual parallel tests is necessary to estimate reliability....

In the concept of domain sampling,parallel tests are viewed as being intellectual constructs,and a set of actual tests which meet certain criteria of parallelism is not necessary to estimate the degree of reliability of measurement.It is not denied that tests which meet the criteria of parallelism in terms of having precisely equal means,standard deviations, and intercorrelations,as well as the same patterns of correlations with other tests,can exist.This is a matter left open for empirical investigation and has no bearing upon either the theory of reliability or the estimation of the degree of reliability."

(Ghiselli(1964),247)

Ser vi Tryons teori i høve til Spearman-Brown tradisjonen, merkar vi oss at vi har nått fram til full liberalisering: Det knyter seg ingen restriksjonar til komponentane i det observerte kompositum.Ser vi Tryons teori i høve til Spearman-Yule tradisjonen og til Brown-Kelley tradisjonen,merkar vi oss at vilkåra til parallelle testar er just dei same;men klassisk teori krev fleire konkrete testar som fyller desse vilkår,medan Tryons teori berre reint hypotetisk postulerer slike testar.

Korrelasjonen mellom Tryons kommensurable kompositа skriv vi som ein korrelasjon av summar slik vi gjorde det i (F53). Den vesentlege skilnad mellom kovarians-matrisane til Jackson-Ferguson-Gulliksen og Tryon er følgjande:

Jackson-Ferguson-Gulliksen tenkjer seg testar som er parallelle item for item (item-parallelle testar) slik at vi får k par av parallelle items,i mot i',j mot j',og k(k-1) par av ikkje-parallelle items,i mot j',j mot i',medan Tryons

items i X_t alle er tilfellelege i høve til alle items i X_t', i mot i' er eit like tilfelleleg par som i mot j'.Dette vil seia at vi i Tryons tilfelle aldri får par av parallelle items, berre k^2 par av tilfellelege items.Sjølvsagt kan det tenkjast at slumpen vil det slik at vi får par av parallelle items,men slike sjeldne par er like fullt å rekna for tilfellelege.

Etter dette resonnementet blir teljaren i (F53) å skriva som ein sum av ein type kovarians og ikkje to som vi gjorde tidlegare:

$$\rho_{XX'} = \frac{\Sigma \rho_{ij'} \sigma_i \sigma_{j'}}{\sigma_X \sigma_{X'}} \qquad (F80)$$

Ved å nytta (F76) og (F78) kan (F80) skrivast berre med observerte verdiar frå X_t,den konkrete testen:

$$\rho_{XX'} = \frac{\Sigma \rho_{ij} \sigma_i \sigma_j}{\sigma_X^2} \qquad (F81)$$

Men $\Sigma \rho_{ij} \sigma_i \sigma_j = k^2 \overline{\rho_{ij} \sigma_i \sigma_j} = k^2 \bar{c}_{ij}$. Difor kan (F81) skrivast

$$\rho_{XX'} = \frac{k^2 \bar{c}_{ij}}{\sigma_X^2} \qquad (F82)$$

(F82) er Tryons generelle reliabilitetsformel,som gjeld for alle par av kommensurable komposita.Denne formelen er og gjeven av Cronbach(1951) som ein variant til alpha utan at han viser kva for vilkår (F82) byggjer på.

Det er lett å sjå at (F82) er ein lite praktisk formel etter som det krevst mykje reknearbeid for å finna den gjennomsnittlege kovariansen.(F82) er først og fremst ein definisjonsformel,men vi kan med utgangspunkt i denne formelen utvikla andre formlar som er meir praktiske,slik Tryon har vist.Vi skal sjå på eit par av desse.Det er formlar som vi tidlegare har fått fram på andre vilkår enn Tryons.

$$\bar{c}_{ij} = \frac{\Sigma c_{ij}}{k(k-1)} \qquad (F83)$$

Når vi nyttar (F83),kan (F82) skrivast

$$\rho_{XX'} = \frac{k^2 \, c_{ij}}{k(k-1)\sigma_X^2} \qquad (F84)$$

Summen av kovariansane kan og skrivast

$$\Sigma c_{ij} = \sigma_X^2 - \Sigma \sigma_i^2 \qquad \text{(F85)}$$

Difor kan no (F84) skrivast

$$\rho_{XX'} = \frac{k^2(\sigma_X^2 - \Sigma \sigma_i^2)}{k(k-1)\sigma_X^2} = (\frac{k}{k-1})(\frac{\sigma_X^2 - \Sigma \sigma_i^2}{\sigma_X^2})$$

$$= (\frac{k}{k-1})(1 - \frac{\Sigma \sigma_i}{\sigma_X^2}) \qquad \text{(F86)}$$

Vi kjenner (F86) frå før.Det er KR20 i generell form,som vi no kanskje oftast kallar alpha.Tryon har såleis utvikla KR20 på nye vilkår,slik desse er spesifiserte i (F75),(F76) og (F78).Det vil seia at ingen restriktive vilkår er knytte til komponentane i den observerte testen for ei utvikling av KR20 eller alpha.

Ved å skriva testvariansen som ein sum av variansar og kovariansar kan (F82) få denne forma:

$$\rho_{XX'} = \frac{k^2 \bar{c}_{ij}}{k\bar{\sigma}_i^2 + k(k-1)\bar{c}_{ij}} = \frac{k\bar{c}_{ij}}{\bar{\sigma}_i^2 + (k-1)\bar{c}_{ij}} \qquad \text{(F87)}$$

*Tryon kallar (F87) ei kovariansform av den generelle reliabilitetsformelen (F82).*Vi tek med (F87) for å bruka denne formelen som utgangspunkt for ei utvikling av den generelle Spearman-Brown formelen på Tryons vilkår.Denne utviklinga vil gje oss ei tilnærming til (F82).

Vi veit at $\bar{c}_{ij} = \overline{\rho_{ij}\sigma_i\sigma_j}$. Dersom vi postulerer at $\sigma_i\sigma_j = \sigma_i^2$, kan den gjennomsnittlege kovarians,$\bar{c}_{ij}$,skrivast på ein annan måte:$\sigma_i^2\bar{\rho}_{ij}$.No er dette eit urealistisk postulat.I staden for denne måten tek vi den gjennomsnittlege itemvarians og multipliserer med den gjennomsnittlege itemkorrelasjon og får ei tilnærming til den gjennomsnittlege kovarians:

$$\bar{c}_{ij} = \overline{\rho_{ij}\sigma_i\sigma_j} \doteq \bar{\rho}_{ij}\bar{\sigma}_i^2 \qquad \text{(F88)}$$

Vi gjer no bruk av tilnærminga i (F88) ved å skriva (F87)

$$\rho_{XX'} \doteq \frac{k\bar{\rho}_{ij}\bar{\sigma}_i^2}{\bar{\sigma}_i^2 + (k-1)\bar{\rho}_{ij}\bar{\sigma}_i^2} = \frac{k\bar{\rho}_{ij}}{1 + (k-1)\bar{\rho}_{ij}} \qquad \text{(F89)}$$

Som vi ser,har vi i (F89) bruk for den gjennomsnittlege item-korrelasjon.Her er det ikkje gjort krav gjeldande om like interkorrelasjonar,slik klassisk teori gjorde det.

4.3.2. Domene-validitet.

Ein testskåre,$X_{t(i)}$,er berre ein tilfelleleg av eit univers av skårar for person i frå den definerte åtferdsdomene.$X_{t(i)}$ er eit estimat av det vi kan kalla domene-skåren eller meir vanleg:sann skåre.Ein domene-skåre er i Tryons teori definert som ein sum eller gjennomsnitt av skårar på eit uendeleg stort tal av komposita som alle har dei same gjennomsnittlege statistiske eigenskapar som er definert i (F75),(F76) og (F78). Tryons domene-skåre er såleis eit omgrep som er identisk med omgrepet sann skåre slik Brown-Kelley tradisjonen definerer denne skåren.

Eit testresultat er logisk sett reliabelt i den grad ei rangering av personar på den observerte skåre vil gje noko nær same rangering om vi hadde kunna rangert på domene-skåren. Ein korrelasjon mellom observert skåre og domene-skåre ville gje oss eit mål på ein slik reliabilitet.

Det er god meining i den tanke å setja opp ein korrelasjonsmatrise der domeneskåren er prediktorvariabel og den observerte skåre kriteriumvariabel.Logikken i dette er at observert skåre må kunna seiast å vera avhengig av domene-skåren.Regresjonen av observerte skårar på domeneskårar gjev oss ei regresjonsline som representerer predikerte skårar.Avviket frå observerte skårar til predikerte skårar,som her representerer domeneskårar,ser vi på i denne samanheng som feilmåling.Kvadrerer vi desse avviksskårane,summerer og så reknar ut den gjennomsnittlege kvadrerte avviksskåre,får vi feilvariansen.Dei predikerte skårane varierer i sin tur omkring den gjennomsnittlege predikerte skåre,og denne variasjonen representerer den predikerte varians,eller domeneskåre-varians.Dermed har vi delt den observerte skårevarians i sann varians og feilvarians. Reliabiliteten er som kjent forholdet mellom sann varians og observert varians.Vi kan og seia det slik:Reliabiliteten er proporsjonen av observert varians som kan tilskrivast sann varians(predikert varians).

I ein regresjonsteoretisk samanheng er det den kvadrerte koeffisient frå korrelasjonen mellom observert og sann skåre som gjev oss forholdet mellom sann varians og observert varians eller proporsjonen av sann varians. I klassisk testteori får vi eit estimat av denne proporsjonen direkte frå korrelasjonen mellom parallelle testar. Her opererer vi med definisjonar som resulterer i det uvanlege at vi kan tolka ein korrelasjonskoeffisient som ein proporsjon.

Vi skal no korrelera observert skåre med sann skåre. Etter som sann skåre ikkje er tilgjengeleg, må dette bli ein hypotetisk korrelasjon.

Først tek vi for oss korrelasjonen mellom observert skåre og sann skåre slik Spearman-Yule tradisjonen definerer denne skåren.

$$\rho_{XT} = \frac{\Sigma x_X x_T}{N\sigma_X\sigma_T} = \frac{\Sigma(x_T+x_e)x_T}{N\sigma_X\sigma_T} = \frac{\Sigma x_T^2 + x_T x_e}{N\sigma_T(\sigma_T^2+\sigma_e^2)^{1/2}}$$

$$= \frac{\sigma_T^2}{\sigma_T^2(\sigma_T^2+\sigma_e^2)^{1/2}} = \frac{\sigma_T}{(\sigma_T^2+\sigma_e^2)^{1/2}} = \frac{\sigma_T}{\sigma_X} = \left(\frac{\sigma_T^2}{\sigma_X^2}\right)^{1/2} = (\rho_{XX'})^{1/2}$$

$$\rho_{XT} = (\rho_{XX'})^{1/2} \qquad \text{(F90)}$$

(F90) seier at korrelasjonen mellom observert og sann skåre er lik rota av reliabilitetskoeffisienten. Denne korrelasjonen er i tradisjonell testteori kalla reliabilitetsindeksen.

Ved å kvadrera på begge sider i (F90) får vi reliabiliteten uttrykt ved den kvadrerte reliabilitetsindeksen.

$$\rho_{XT}^2 = \frac{\sigma_T^2}{\sigma_X^2} = \rho_{XX'} = 1 - \frac{\sigma_e^2}{\sigma_X^2} \qquad \text{(F91)}$$

I (F91) har vi samla fire formlar som i klassisk testteori alle definerer reliabilitet. Vi skal seinare sjå at desse likskapane ikkje vil gjelda når vi kjem over på generalizability.

I Tryons system kan korrelasjonen mellom observert skåre og domeneskåre estimerast på følgjande måte og etter følgjande resonnement:

Vi tenkjer oss domeneskåren som ein sum av skårar på eit uendeleg tal av testar som stettar Tryons krav (F75),(F76) og (F78).(Vi kan og tenkja oss domeneskåren som den gjennomsnittlege skåre over eit uendeleg tal av testar.Både Gulliksen(1950) og Ghiselli(1964) bruker denne definisjonen og estimerer korrelasjonen mellom observert skåre og sann skåre (domeneskåre) med dette utgangspunktet.)

Vi let k* symbolisera talet på testar,og k* er så stor at

$$1/k^* \longrightarrow 0, \text{og } k^*/(k^*-1) \longrightarrow 1 \qquad \text{(F92)}$$

$$\rho_{XT} = \frac{\Sigma x_1(x_1+x_2+..+x_{(k^*-1)}+x_{k^*})}{N\sigma_X\sigma_T} \qquad \text{(F93)}$$

Når vi utviklar (F93),får vi i teljaren eitt variansuttrykk og k*-1 kovariansuttrykk.I nemnaren kan σ_T skrivast som rota av ein sum av variansar og kovariansar.

$$\rho_{XT} = \frac{\sigma_X^2 + (k^*-1)\rho_{XX'}\sigma_X^2}{\sigma_X(k^*\sigma_X^2 + k^*(k^*-1)\rho_{XX'}\sigma_X^2)^{1/2}} \qquad \text{(F94)}$$

Det er klårt etter definisjonane i (F75),(F76) og (F78) at summen av kovariansane kan skrivast som eit produkt,slik vi har gjort det i teljar og nemnar i (F94),fordi interkorrelasjonane mellom parallelle sampel er like og alle sampel har same varians.

I (F94) kan σ_X^2 i teljaren setjast utanfor parantes.I nemnaren kan σ_X^2 under rotteiknet setjast utanfor parantes,deretter flytter vi σ_X utanfor rotteiknet.Med så gjort kan σ_X^2 og $\sigma_X\sigma_X$ strykast i respektive teljar og nemnar.

$$\rho_{XT} = \frac{1 + (k^*-1)\rho_{XX'}}{(k^* + k^*(k^*-1)\rho_{XX'})^{1/2}} \qquad \text{(F95)}$$

Vi dividerer no i teljar og nemnar i (F95) med k*.For å få k* under rotteiknet i nemnaren må k* kvadrerast.

$$\rho_{XT} = \frac{\frac{1}{k^*} + (\frac{k^*-1}{k^*})\rho_{XX'}}{(\frac{k^*(1 + (k^*-1)\rho_{XX'})}{k^{*2}})^{1/2}} = \frac{\frac{1}{k^*} + (\frac{k^*-1}{k^*})\rho_{XX'}}{(\frac{1}{k^*} + (\frac{k^*-1}{k^*})\rho_{XX'})^{1/2}} \qquad \text{(F96)}$$

Ved å nytta (F92) kan teljar og nemnar i (F96) reduserast

slik at vi kan skriva

$$\rho_{XT} = \frac{\rho_{XX'}}{(\rho_{XX'})^{1/2}} = \left(\frac{\rho^2_{XX'}}{\rho_{XX'}}\right)^{1/2} = (\rho_{XX'})^{1/2} \qquad (F97)$$

Korrelasjonen mellom observert skåre og sann skåre etter klassisk teori og korrelasjonen mellom observert skåre og domeneskåre etter Tryons teori gjev same resultat.Korrelasjonen kan skrivast som rota av reliabilitetskoeffisienten.

Det er Tryons forteneste at reliabilitetsindeksen har fått ein framskoten plass i nyare testteori og meir meining knytt til seg.Det var Truman Kelley som først lanserte reliabilitetsindeksen(Kelley(1916)),som sidan har følgt dei fleste bøker i testteori utan å vera særleg meiningsberande.

Domeneskåren er den skåre ein person ville få dersom vi kunne prøva han på universet av testar.Det er i teorien ein perfekt skåre som det kan vera rimeleg grunn til å rekna for ein kriterieskåre som vi kan validera den observerte skåre mot. Domeneskåren blir slik sett det perfekte kriterium for det definerte trekk eller det hypotetiske construct,og korrelasjonen mellom observert skåre og domeneskåre blir ein validitetskoeffisient.Denne validiteten kallar Tryon behavior domain validity,vi domene-validitet.Tryons domene er eit definert åtferdsunivers der vi samplar items som er meint å tappa det construct vi er interesserte i.Domeneskåren kan difor med rimeleg grunn oppfattast som ein construct-skåre. Såleis blir Tryons domene-validitet ei form for construct validity.Denne samanhengen har Tryon sjølv ikkje peika på.

Når domeneskåren blir oppfatta som det perfekte kriterium, er det lettare å forstå at validiteten har ei øvre grense sett av reliabiliteten.Denne grensa er reliabilitetsindeksen, som er den observerte skåres korrelasjon med domeneskåren. I denne ikkje-empiriske samanheng er domeneskåren det best tenkjelege kriterium.

5. Bruk av variansanalyse i reliabilitetsestimering.

Klassisk reliabilitet har teknisk vore sterkt knytt til produkt-moment korrelasjon. Denne teknikken har tilsynelatande vore godt tenleg til den teori reliabilitetsestimering byggjer på. Mellom anna går teorien ut på at korrelasjonen mellom parallelle testar er lik. Det vil seia at vi greier oss med ein slik korrelasjon basert på to parallelle testar. Etter teorien er denne eine korrelasjonen god for alle.

Likevel må vi seia at produkt-moment korrelasjonen har vore ein alt for enkel teknikk til fullt ut å kunna ta vare på den kompliserte psykometriske feilteori som vi i mange år har operert med på eit omgrepsplan. Ein produkt-moment korrelasjon kan berre gje oss ein estimeringsfeil, eller ein målingsfeil i vår samanheng. Men i teorien identifiserer vi multiple feilkjelder, som alt etter eksperimentelt design kan ha spesifikk verknad på eit testresultat. (Sjå t.d. Magnusson(1966), kap. 8 og 9, Thorndike(1951).) I ein produkt-moment korrelasjon vil simultane feil gå inn som ein, samanblanda (confounded) feil.

Reliabilitetstypen stabilitet og ekvivalens illustrerer dette: Ein korrelasjon mellom to parallelle testar, her parallelle former, administrert med tidsintervall gjev oss reliabiliteten, t.d. $r_{tt} = 0,80$. Her er 20% av testvariansen estimert som feilvarians. Men feilen er i dette test design prinsipielt av to slag: Ein ulik diskrepans mellom skåreverdiar som dels skriv seg frå innhaldsmessig ulike testar og dels frå instabilitet frå gong til gong. I dette tilfelle er vi ikkje komne like langt i den teknikk vi bruker som i den tenking vi har etablert. Vi kunne her ønskja eit design og ein analyseteknikk som gjorde det mogleg å få fram estimat av begge feil, t.d. 0,80 + 0,12 + 0,08, og ikkje som no 0,80 + 0,20.

Variansanalysen brukt som matematisk modell vil kunna oppfylla våre ønskje om differensierte feilestimat. Ein produkt-moment korrelasjon kan ikkje det. Variansanalysen gjer det mogleg å dela opp den totale kvadratsum på delkvadratsummar frå identifiserte variasjonskjelder i ein datamatrise. Dermed kan vi finna kor mykje kvar variasjonskjelde yter til total variasjon.

Det var Hoyt(1941) som for alvor lanserte variansanalysen i reliabilitetsestimering. Hans teknikk kan appliserast på same test design som t.d. KR20,altså ikkje-stratifiserte komposita.Hans estimat gjev også identiske resultat med KR20 eller alpha. Hoyts analyse er teknisk ei nyskaping men teoretisk sett tradisjonell.Hoyts enkle analyse viser heller ikkje dei føremoner variansanalysen kan ha samanlikna med tradisjonelle teknikkar når vi har med meir kompliserte test design å gjera.

Det er først i seinare år at variansanalysen står fram som den desidert mest tenlege reiskap både teknisk og til hjelp for tanken i forskning om kring reliabilitet.Dette kan ha samanheng med to ting: Testteorien er liberalisert,vi krev t.d. ikkje at interkorrelasjonen mellom parallelle testar er like,og vi er meir interesserte i variansstrukturen i kompliserte test design.

Vi skal i det følgjande utvikla to variansanalysemodellar for reliabilitetsestimering som begge byggjer på ikkje-stratifiserte komposita.Seinare skal vi konstruera meir kompliserte modellar på stratifiserte komposita.

5.1. Variansanalysemodell for to-vegs klassifisering.

5.1.1. Oppdeling av total kvadratsum.

Tabell 2. Generell to-vegs datamatrise.

	1 ..... j ..... k	
1	$X_{11} \cdots\cdots X_{1j} \cdots\cdots X_{1k}$	$X_{1.}$
	.....	
i	$X_{i1} \cdots\cdots X_{ij} \cdots\cdots X_{ik}$	$X_{i.}$
	.....	
n	$X_{n1} \cdots\cdots X_{nj} \cdots\cdots X_{nk}$	$X_{n.}$
	$X_{.1} \cdots\cdots X_{.j} \cdots\cdots X_{.k}$	$X_{..}$

Vi tek utgangspunkt i ein datamatrise med n rekkjer og k kolonnar som vist i tabell 2. I matrisen står ein X_{ij} for ein observasjon i i-te rekkje og j-te kolonne. Marginalverdiane $X_{i.}$ og $X_{.j}$ symboliserer M-verdien over alle j i i-te rekkje og M-verdien over alle i i j-te kolonne. $X_{..}$ er M-verdien over rekkjer og kolonnar,total M-verdi.

Ein X_{ij} kan karakteriserast som ein sum av den totale M-verdi, differensen mellom M-verdien for i-te rekkje og M_{tot}, differensen mellom M-verdien for j-te kolonne og M_{tot} og endeleg ein rest. Dette kan vi visa algebraisk:

$$X_{ij} = X_{..} + (X_{i.}-X_{..}) + (X_{.j}-X_{..}) + (X_{ij}-X_{i.}-X_{.j}+X_{..}) \quad (F98)$$

(F98) er ein algebraisk identitet og må vera rett.

Dersom vi flytter $X_{..}$ i (F98) frå høgre til venstre side av likskapsteiknet, får vi:

$$(X_{ij}-X_{..}) = (X_{i.}-X_{..}) + (X_{.j}-X_{..}) + (X_{ij}-X_{i.}-X_{.j}+X_{..}) \quad (F99)$$

(F99) viser at den totale avviksskåren er ein sum av tre komponentar: Ein rekkje-komponent, ein kolonne-komponent og ein residual-komponent.

Residual-komponenten, siste uttrykket til høgre for likskapsteiknet i (F99), er ei samantrengt form og kan ha krav på ei forklaring. Residualen er det som blir igjen når vi frå X_{ij} trekkjer den generelle komponent ($X_{..}$), rekkje-komponenten og kolonne-komponenten. Vi kan kalla residualen e_{ij} og skriv:

$$\begin{aligned} e_{ij} &= X_{ij} - (X_{..} + (X_{i.}-X_{..}) + (X_{.j}-X_{..})) \\ &= X_{ij} - X_{..} - X_{i.} - X_{.j} + X_{..} \\ &= X_{ij} - X_{i.} - X_{.j} + X_{..} \end{aligned} \quad (F100)$$

Vi ser at (F100) er lik siste lekken i (F99).

Vi veit at avviksskåren er utgangspunkt for våre mest vanlege variasjonsmål, som kvadratsum, varians, standardavvik og standardfeil av ulike slag. Dersom vi kvadrerer og summerer over rekkjer og kolonnar i (F99), vil vi få total kvadratsum. Ved desse operasjonane får vi ut i alt seks uttrykk, men tre av dei fell frå fordi dei har avviksskåresummar som faktor. Avviksskåresummar er som kjent lik null. Etter dette vil vi stå att med tre uttrykk, og vi får:

$$(X_{ij} - X_{..})^2 = k(X_{i.}-X_{..})^2 + n(X_{.j}-X_{..})^2 + (X_{ij}-X_{i.}-X_{.j}+X_{..})^2 \quad (F101)$$

(F101) viser at den totale kvadratsum i ein to-vegs datamatrise er samansett av tre komponentar: Ein kvadratsum som

skriv seg frå variasjonen mellom M-verdiar for rekkjer, ein andre frå variasjonen mellom M-verdiar for kolonnar og ein tredje frå variasjonen mellom residualar.

5.1.2. Eksempel.

Vi skal visa med eit eksempel korleis vi kan dekomponera ein to-vegs datamatrise og få fram total kvadratsum som ein sum av tre komponentar.

I tabell 3 har vi ein matrise med 5 rekkjer og 4 kolonnar. Vi har og med M-verdiar for rekkjer og kolonnar og total M-verdi.

Tabell 3.

	1	2	3	4	$X_{i.}$
1	4	5	4	5	4,5
2	3	4	5	4	4,0
3	4	4	3	3	3,5
4	2	3	3	2	2,5
5	1	2	1	2	1,5
$X_{.j}$	2,8	3,6	3,2	3,2	3,2

Etter (F98) kan verdien X_{11}, 4, skrivast:

$$4 = 3,2 + (4,5-3,2) + (2,8-3,2) + (4-4,5-2,8+3,2)$$
$$= 3,2 + 1,3 - 0,4 - 0,1$$

Vi gjer det same for alle X_{ij} i matrisen og kan skriva ein dekomponert datamatrise, slik vi har gjort det i tabell 4 (sjå side 57).

Reknar vi total kvadratsum av matrisen i tabell 3, får vi 29,2. Same kvadratsum får vi om vi reknar kvadratsummen for dei enkelte komponentane i tabell 4 og summerer.

Av tabell 4 går det fram at skåreverdiane er delt opp i 4 komponentar: generell (g), rekkje- (r), kolonne- (k) og residualkomponentar (e_{ij}). Men det er berre tre av desse komponentane som varierer. Som vi ser er den generelle komponent invariant over rekkjer og kolonnar og yter såleis ingen ting

Tabell 4. Dekomponert datamatrise.

		g	r	k	e_{ij}	$X_{i.}$ g	r	k	e_{ij}
	K1	3,2	+1,3	-0,4	-0,1				
	K2	3,2	+1,3	+0,4	+0,1				
R1	K3	3,2	+1,3	+0,0	-0,5	3,2	+1,3	+0,0	+0,0
	K4	3,2	+1,3	+0,0	+0,5				
	K1	3,2	+0,8	-0,4	-0,6				
	K2	3,2	+0,8	+0,4	-0,4				
R2	K3	3,2	+0,8	+0,0	+1,0	3,2	+0,8	+0,0	+0,0
	K4	3,2	+0,8	+0,0	+0,0				
	K1	3,2	+0,3	-0,4	+0,9				
	K2	3,2	+0,3	+0,4	+0,1				
R3	K3	3,2	+0,3	+0,0	-0,5	3,2	+0,3	+0,0	+0,0
	K4	3,2	+0,3	+0,0	-0,5				
	K1	3,2	-0,7	-0,4	-0,1				
	K2	3,2	-0,7	+0,4	+0,1				
R4	K3	3,2	-0,7	+0,0	+0,5	3,2	-0,7	+0,0	+0,0
	K4	3,2	-0,7	+0,0	-0,5				
	K1	3,2	-1,7	-0,4	-0,1				
	K2	3,2	-1,7	+0,4	+0,1				
R5	K3	3,2	-1,7	+0,0	-0,5	3,2	-1,7	+0,0	+0,0
	K4	3,2	-1,7	+0,0	+0,5				
	K1	3,2	+0,0	-0,4	+0,0				
	K2	3,2	+0,0	+0,4	+0,0	$X_{..}$			
$X_{.j}$	K3	3,2	+0,0	+0,0	+0,0	3,2	+0,0	+0,0	+0,0
	K4	3,2	+0,0	+0,0	+0,0				

til variasjonen i matrisen. Rekkjekomponenten varierer over rekkjer men er invariant over kolonnar. Kolonnekomponenten varierer over kolonnar men er invariant over rekkjer. Residualkomponenten varierer over rekkjer og kolonnar.

Alle variable komponentar har M_{tot} ($X_{..}$) lik 0. For å finna kvadratsummen for desse komponentane kan vi difor bruka verdiane i tabell 4 som avviksskårar slik dei står, kvadrera dei og summera over rekkjer og kolonnar for kvar av dei tre komponentane. Desse operasjonane kan vi skriva slik:

$$\mathcal{E}x^2_{tot} = \mathcal{E}x^2_r + \mathcal{E}x^2_k + \mathcal{E}x^2_{res} = 23{,}2 + 1{,}6 + 4{,}4 = 29{,}2$$

Vi får den same totale kvadratsum frå den dekomponerte matrisen som frå den originale.

(F101) viser ein annan måte å gjera dette på. I staden for å summera dei kvadrerte avviksskårane for rekkjekomponenten over rekkjer og kolonnar kan vi multiplisera med k (fordi rekkjekomponenten er den same over kolonnar for ei og same rekkje) og summera over rekkjer. Tilfellet er analogt for avviksskårane til kolonnekomponenten; men no multipliserer vi med n (fordi kolonnekomponenten er den same over rekkjer for ein og same kolonne) og summerer over kolonnar. Denne framgangsmåten er ikkje bunden til "innmaten" i matrisen. Det er nok når vi har marginalverdiane for rekkjer og kolonnar, slik vi har dei i tabell 3. Legg merke til at vi greier dette utan den dekomponerte matrisen. Marginalverdiane $X_{i.}$ i tabell 3 er summen av komponentverdiane under $X_{i.}$ i tabell 4. Som vi ser av tabell 4, kan variasjonen i denne marginalsummen berre tilskrivast rekkjekomponenten. Etter dette kan vi sjå på kvadratsummen til rekkjesummen som k gonger kvadratsummen til M-verdiane for rekkjer slik vi har desse M-verdiane i tabell 3. Marginalverdiane $X_{.j}$ i tabell 3 er summen av komponentverdiane etter $X_{.j}$ i tabell 4. Det er berre kolonnekomponenten som svarar for kolonnevariasjonen. Om vi bruker marginalverdiane for kolonnar i tabell 3 eller kolonnekomponentverdiane i tabell 4, får vi den same kvadratsum for kolonnevariasjon. Difor kan vi sjå på kvadratsummen til kolonnekomponenten som n gonger kvadratsummen til M-verdiane for kolonnar.

5.2. Hoyt-modellen.

Oppdeling av total kvadratsum i additive komponentar for eit to-vegs design gjeld generelt for talverdiar som er ordna i rekkjer og kolonnar. Vi har hittil sett på variansanalysen berre med tanke på å studera den matematiske struktur.

I psykologisk måling blir data ofte ordna i rekkjer og kolonnar. Test-data er til vanleg ordna med skåreverdiar for personar i rekkjer og item-skårar i kolonnar.

Det er formelt ingen ting i vegen for å dela opp den totale kvadratsum i ein person-item matrise i ein person-komponent (rekkje-komponent), ein item-komponent (kolonne-komponent) og ein residual-komponent. (Ei anna sak er kor meiningsfylt det kan vera å tvinga denne additive struktur på våre data. Det problemet tek vi ikkje opp i denne samanhengen. Vi reknar her med at variansanalysen er ein god matematisk modell.) Kvadratsummen for personar tek utgangspunkt i variasjonen i M-verdiane for personar, kvadratsummen for items i variasjonen i M-verdiane for items. Residual-kvadratsummen er resten av totalvariasjonen som ikkje er forklart ved person- og item-variasjonen.

No er det slik at vi ikkje fullt og heilt fester lit til test-data bygde på eit sampel som grunnlag for ei estimering av dei verdiane vi kan rekna med i populasjonen. Den variasjon vi observerer mellom personar er ikkje berre bestemt av genuine persondifferensar. Vi må rekna med at den observerte person-variasjon er infladert av tilfellelege variasjonskjelder, og på ein eller annan måte må vi freista koma åt denne feilvariasjonen.

Kvadratsummane som kjem fram ved bruk av (F101), er observerte verdiar. Om vi skal ta utgangspunkt i desse kvadratsummane ved ei reliabilitetsestimering, er det spørsmål om vi har variasjonskjelder tilgjengelege som kan vera tenlege som estimat av den tilfellelege variasjon som vi teoretisk har definert inn i dei observerte verdiane.

Dette er ikkje nye tankar; vi kjenner dei frå tradisjonell test-teori. Nytt er det å få desse tankane inn i eit formspråk som ikkje har vore vanleg i test-teorien.

Ein kva som helst skåre i ein person-item matrise, X_{ij}, tenkjer vi oss samansett av ein sann komponent, p_i, og ein feilkomponent, e_{ij}, der sann komponent er definert som gjennomsnittleg item-skåre for person i i universet av items. Vi skriv dette slik:

$$X_{ij} = p_i + e_{ij} \qquad \text{(F102)}$$

I eit sampel av items kan person i's gjennomsnittlege skåre skrivast

$$\frac{\Sigma X_{ij}}{k} = X_{i.} = p_i + e_{i.} \qquad \text{(F103)}$$

$X_{i.}$ står for gjennomsnittet for person i over k items. Verdien p_i er den same over items, og vi treng difor ikkje skriva den som eit gjennomsnitt. Verdien $e_{i.}$ er den gjennomsnittlege feil over items.

Med utgangspunkt i (F101) definerer vi no person-variansen:

$$MS_i = \frac{k\Sigma(X_{i.} - X_{..})^2}{n-1} \qquad \text{(F104)}$$

Vi bør ha det klårt for oss at (F104) ikkje er total test-varians slik denne er definert tidlegare. Variansen til $X_{i.}$, gjennomsnittleg personskåre, er

$$\sigma^2_{X_{i.}} = \frac{\Sigma(X_{i.} - X_{..})^2}{n-1} \qquad \text{(F105)}$$

Variansen til sumskåren kan skrivast

$$\sigma^2_{X.} = \frac{\Sigma(kX_{i.} - kX_{..})^2}{n-1} = \frac{k^2\Sigma(X_{i.} - X_{..})^2}{n-1} \qquad \text{(F106)}$$

Relasjonane mellom testvarians, MS_i og variansen til gjennomsnittleg person-skåre blir då

$$\sigma^2_{X.} = kMS_i = k^2\sigma_{X_{i.}} \qquad \text{(F107)}$$

$$MS_i = k\sigma^2_{X_{i.}} \qquad \text{(F108)}$$

Når vi går ut frå at p_i og $e_{i.}$ er ukorrelerte, kan den forventa varians til gjennomsnittlege personskårar skrivast

$$E(\sigma^2_{X_{i.}}) = \sigma^2_i + \sigma^2_{o_{i.}} \qquad \text{(F109)}$$

Etter (F108) må forventa MS_i bli

$$E(MS_i) = k\sigma^2_i + k\sigma^2_{e_{i.}} \qquad \text{(F110)}$$

$$\text{Men } \sigma_{e_{i.}} = \sigma_e/k^{\frac{1}{2}} \qquad \text{(F111)}$$

Vi kvadrerer i (F111) og får

$$\sigma^2_{e_{i.}} = \sigma^2_e/k \qquad \text{(F112)}$$

Altså,

$$k\sigma^2_{e_{i.}} = \sigma^2_e \qquad \text{(F113)}$$

Vi nyttar no (F113) og skriv (F110) slik:

$$E(MS_i) = k\sigma^2_i + \sigma^2_e \qquad \text{(F114)}$$

Etter (F107) og (F108) blir forventa testvarians

$$E(\sigma^2_t) = k(k\sigma^2_i + \sigma^2_e) = k^2\sigma^2_i + k\sigma^2_e \qquad \text{(F115)}$$

Reliabiliteten til sumskåren blir etter (F115)

$$\rho_k = \frac{k^2\sigma^2_i}{k^2\sigma^2_i + k\sigma^2_e} = \frac{k\sigma^2_i}{k\sigma^2_i + \sigma^2_e} \qquad \text{(F116)}$$

Vi kan gå eit steg vidare med (F116):

$$\rho_k = \frac{\sigma^2_i}{\sigma^2_i + \sigma^2_e/k} = \frac{\sigma^2_i}{\sigma^2_i + \sigma^2_{\bar{e}}} \qquad \text{(F117)}$$

(F116) og (F117) viser at reliabiliteten blir den same anten vi tek utgangspunkt i testvariansen, i den definerte personvariansen (F104) eller i den gjennomsnittlege item-varians.

(F116) og (F117) svarar til tidlegare definisjon av reliabilitet som forholdet mellom sann varians og observert varians.

Forventa varians til eitt item kan skrivast

$$E(\sigma^2_{X_{ij}}) = \sigma^2_i + \sigma^2_e \qquad \text{(F118)}$$

Reliabiliteten til eitt item må etter (F118) bli

$$\rho_1 = \frac{\sigma^2_i}{\sigma^2_i + \sigma^2_e} \qquad \text{(F119)}$$

Etter at vi har definert reliabilitet i variansanalytiske termar, (F116), (F117) og (F119), er problemet no å kunna estimera reliabiliteten frå sampel-data. Dette problemet knyter seg til feilvariansen.

I ein to-vegs analyse er det vanleg å rekna residualen som eit tenleg feilvarians-estimat. Grunngjevinga for dette skal koma etter kvart.

I tabell 5 har vi sett opp ein to-vegs variansanalyse for ein person-item matrise med observert og forventa MS for dei tre variasjonskjeldene.

Tabell 5

Variasjonskjelde	frg	Obs(MS)	E(MS)
Personar	n-1	MS_i	$\sigma^2_e + k\sigma^2_i$
Items	k-1	MS_j	$\sigma^2_e + n\sigma^2_j$
Residual	(n-1)(k-1)	MS_r	σ^e_r

I vår samanheng er det berre obs(MS) og E(MS) for personar og residual som interesserer. Av tabell 5 vil det gå fram at den definerte reliabilitet (F116) kan estimerast slik:

$$\rho_k = \frac{MS_i - MS_r}{MS_i} \qquad \text{(F120)}$$

Eit estimat av (F119), reliabiliteten til eitt item, kjem fram på denne måten:

$$\rho_1 = \frac{1/k(MS_i - MS_r)}{1/k(MS_i - MS_r) + MS_r} = \frac{MS_i - MS_r}{MS_i - MS_r + kMS_r}$$

$$= \frac{MS_i - MS_r}{MS_i + (k-1)MS_r} \qquad \text{(F121)}$$

Eksempel

Vi tenkjer oss at tabell 3 (s. 56) representerer ein person-item matrise. Vi let rekkjene representera fem stilar (skrivne av fem forskjellige personar) og kolonnane fire vurderarar. I tabell 6 presenterer vi variansanalysen av våre hypotetiske data.

Tabell 6

Variasjonskjelde	frg	SS	MS
Stilar	4	23,2	5,800
Vurderarar	3	1,6	0,533
Residual	12	4,4	0,367
Total	19	29,2	

Etter (F120) blir reliabiliteten til sumskåren over fire vurderarar:

$$\rho_4 = \frac{5,800 - 0,367}{5,800} = 0,938$$

Etter (F121) blir reliabiliteten til <u>ein</u> vurderar:

$$\rho_1 = \frac{5,800 - 0,367}{5,800 + 3 \cdot 0,367} = 0,787$$

Det er som sagt residualvariansen vi bruker som estimat av feilvariansen. Variasjonen mellom vurderarar er her halden

utanfor. Denne variasjonen er eit uttrykk for det vi kan kalla vurderar-bias. Vurderarane bruker ikkje skalaen likt, deira referenseramme er ikkje den same. Denne ulike referenseramme er i den modellen vi no har etablert, ikkje definert som feil. Vi tillet at vurderarane bruker skalaen kvar på sin måte. Slik modellen verkar, korrigerer vi for vurderar-bias. Dette har vi vist i tabell 7. Her har vi korrigert for vurderar-bias ved å

Tabell 7. Matrise for korrigert vurderar-bias.

	A	B	C	D	Sum	Gjsn
1	4,4	4,6	4,0	5,0	18	4,5
2	3,4	3,6	5,0	4,0	16	4,0
3	4,4	3,6	3,0	3,0	14	3,5
4	2,4	2,6	3,0	2,0	10	2,5
5	1,4	1,6	1,0	2,0	6	1,5
Sum	16,0	16,0	16,0	16,0	64	
Gjsn	3,2	3,2	3,2	3,2		3,2

addera eller subtrahera ein konstant for kvar vurderar til eller frå den opphavlege skåren. Denne konstanten er differensen mellom vurderargjennomsnitt over rekkjer (kolonnegjennomsnitt) og totalgjennomsnitt. I høve til totalgjennomsnittet vil sjølvsagt somme vurderarar liggja for høgt, andre for lågt i skalaen, og vi korrigerer slik at alle vurderarar får same gjennomsnitt. Som vi ser, blir sumskåren ikkje forandra; og som vi vil forstå, kan ikkje innom-vurderar-variasjonen bli skipla ved denne korreksjonen.

Matrisen i tabell 7 er redusert med mellom-vurderar-variasjon og står igjen med person-variasjon og residual-variasjon, som er dei same som før. Legg no merke til at residualen kan tolkast som ein innom-person-variasjon. I den grad vurderarane er usamde i si karaktergjeving av ein og same stil når dei bruker skalaen likt, i den grad får vi usikker vurdering som vi tolkar som feil. Det er denne feilen som i vår modell går inn som eit estimat av feilvariansen $ø_e^2$.

No er det ikkje sikkert at denne feilen er tilfelleleg. Når vi tidlegare har brukt omgrepet vurderar-bias, har vi med det

meint ein generell tendens hos kvar vurderar til å bruka skalaen høgt eller lågt eller likt i høve til totalgjennomsnittet. Men vi kan og tenkja oss ein spesifikk vurderar-bias, ein systematisk tendens til høg eller låg karakter i møte mellom ein bestemt vurderar og ein bestemt stil. Det er dette som i teknisk terminologi blir kalla person-item interaksjon.

Denne interaksjonen får vi ikkje noko mål på i vårt design, som er eit ikkje-replikert to-vegs design. Det vil seia at vi berre har ein observasjon for kvar av dei kn celler i matrisen. Vår residual er logisk å rekna for ei samanblanding av to typar variasjon: interaksjon og innom-celle variasjon. Denne siste variasjonen kunne vi berre få med minst to observasjonar frå kvar vurderar for kvar stil. Den eine karakteren som kvar vurderar har gjeve ein stil, kan vi sjå på som ein av mange karakterar som denne vurderaren kunne ha gjeve denne stilen under andre vilkår som berre har ein tilfelleleg verknad på karaktergjevinga. Desse hypotetisk mange karakterar innanfor ei og same celle og over alle celler ville vera det beste grunnlag for ein tilfelleleg variasjon, feilvariansen. I denne situasjonen ville vi også kunna rekna med eit celle-gjennomsnitt. Variasjonen mellom cellegjennomsnitt over kn celler ville vera utgangspunkt for ei estimering av den genuine interaksjon.

Etter dette skulle det vera lettare å forstå at residualen i vårt tilfelle er ein samanblanda varians; den "eigenlege" feilvariansen representert ved vurderarfluktuasjon på ein og same stil (vi har fått tak i berre ein av dei mange karakterar som kunne ha vorte gjevne av denne vurderaren på denne stilen) + person-item interaksjon.

I vårt tilfelle er det for så vidt ikkje mykje om å gjera å kunna skilja desse to komponentane som vi no har definert inn i residualen. Ein interaksjon vil nok som regel logisk sett måtta reknast som feilvarians.

Vi kan korrigera matrisen i tabell 7 endå ein gong. Vi kan korrigera for person-variasjon. Det høyrest ikkje rimeleg ut at vi så skulle gjera; det er som kjent person-variasjonen vi er ute etter. Når vi likevel gjer det, er det for, om råd er, å få ei endå betre forståing av feilvariansen, residualen.

Tabell 8 viser oss ein dobbelt-korrigert person-item matrise. Med utgangspunkt i tabell 7 har vi gått vidare og korrigert

Tebell 8. Matrise korrigert for item- og personvariasjon.

	A	B	C	D	Sum	Gjsn
1	3,1	3,3	2,7	3,7	12,8	3,2
2	2,6	2,8	4,2	3,2	12,8	3,2
3	4,1	3,3	2,7	2,7	12,8	3,2
4	3,1	3,3	3,7	2,7	12,8	3,2
5	3,1	3,3	2,7	3,7	12,8	3,2
Sum	16,0	16,0	16,0	16,0	64,0	
Gjsn	3,2	3,2	3,2	3,2		3,2

for "person-bias". Det er gjort ved,for kvar person,å addera eller subtrahera ein konstant til eller frå den korrigerte skåren i tabell 7. Konstanten er differensen mellom person-gjennomsnitt og total gjennomsnitt. Den variasjon som no er att i den dobbelt-korrigerte matrisen, er berre residual-variasjon.

Tabell 8 skulle gjera det konkret kva som i Hoyt-modellen blir rekna som feilvariasjon. Tabellen kan gje oss eit inn-trykk av korleis fire vurderarar med same referenseramme i skalaen ville setja sine karakterar på fem like gode stilar. Alle stilane skulle med sikker vurdering får karakteren 3,2. Variasjonen i tabell 8 er ein variasjon om kring denne"sanne" karakteren og er logisk å rekna som feil. Det er denne va-riasjonen vi har brukt til feilvariansen, $ø_e^2$, i vår relia-bilitetsestimering.

5.3. Reliabilitet basert på ein-vegs variansanalyse. (Webster-modellen)

Hoyt-modellen byggjer på ei tovegs klassifisering der item-variasjonen blir skilt ut og halden utanfor feilvariansen. Dette er ei rimeleg løysing på feilvarians-problemet når vi først og fremst ønskjer ei gjennomsnittleg personrangering

utan omsyn til generell vurderar-bias, om vurderarane bruker skalaen ulikt når det gjeld nivå, eller til den skala dei ulike vurderar måtte bruka, for å halda oss til det eksemplet vi har brukt. Det enkelte items evne til differensiering er sjølvsagt viktig, og Hoyt-modellen er sensitiv for skilnader i innom-item variasjon, i vårt eksempel: innom-vurderar variasjon.

Men det finst også tilfelle der vi ikkje berre ønskjer å sjå i kor stor grad ei person-rangering blir den same over items (vurderarar), men også i kor stor grad items (vurderarar) plasserer personane likt i ein og same skala. Når ei slik plassering er viktig, t.d. i ei evaluering der ein bestemt bruk av ein karakterskala er føreskriven og der den absolutte karakter kan få konsekvensar for personane, blir itemvariasjon (generell vurderarbias) å rekna som feil. Dermed er vi komne dit at vi ønskjer denne variasjonen inn i feilvariansen i tillegg til residualen.

Desse to variasjonskjeldene saman, itemvariasjon og residual, kallar vi innom-person variasjon (wp, within person). Denne variasjonen gjev uttrykk for kor mykje dei ulike items varierer i si "vurdering" av ein og same person når vi ikkje har korrigert for ulike M-verdiar over items.

Tabell 9. Webster-modellen. (Ein-vegs variansanalyse)

Variasjonskjelde	frg	Obs(MS)	E(MS)
Mellom personar	n-1	MS_i	$ø_e^2 + kø_i^2$
Innom personar	n(k-1)	MS_{wp}	$ø_e^2$

Av tabell 9 ser vi at innom-person variasjon er komen i staden for item- og residualvariasjon (sjå tabell 5). Dei forventa variansuttrykk, E(MS), som er relevante for reliabilitetsestimering, er dei same i tabell 5 og tabell 9. Således blir definisjonsformlane for reliabilitet bygde på tabell 9 dei same som reliabilitetsformlane (F116), (F117) og (F119). Men estimeringa blir ikkje den same.

(F120) i Hoyt-modellen svarar til i Webster-modellen

$$\rho_k = \frac{MS_i - MS_{wp}}{MS_i} \qquad (F122)$$

(F121) i Hoyt-modellen svarar til i Webster-modellen

$$\rho_1 = \frac{MS_i - MS_{wp}}{MS_i + (k-1)MS_{wp}} \qquad (F123)$$

I tabell 10 har vi brukt Webster-modellen på våre hypotetiske data frå tabell 3. Samanliknar vi dei to reliabilitetskoeffisientane frå Hoyt-analysen med dei vi no får, ser vi at

Tabell 10.

Variasjonskjelde		frg		SS	MS
Mellom personar		4		23,2	5,8
Innom personar =		15		6,0	0,4
Items	3		1,6		
+Residual	12		4,4		
Total		19		29,2	

Reliabiliteten til sumskåren:

$$\rho_4 = \frac{5,8 - 0,4}{5,8} = 0,931$$

Reliabiliteten til ein vurderar:

$$\rho_1 = \frac{5,8}{5,8 + 3 \cdot 0,4} = 0,771$$

skilnaden er liten. Etter det vi no veit må dette skriva seg frå ein relativt liten itemvariasjon, som i dette tilfelle berre i uvesentleg grad har gjort feilvariansen større.

Konstruksjonen av X_{ij} er i Webster-modellen redusert til tre komponentar: ein generell komponent, ein rekkjekomponent og residualen, som no er ein innom-person variasjon.

Feilkomponenten blir i denne modellen den del av skåren som ikkje er forklart ved generell komponent og rekkjekomponent:

$$e_{ij} = X_{ij} - (X_{..} + (X_{i.} - X_{..}))$$

$$= X_{ij} - X_{..} - X_{i.} + X_{..}$$

$$= X_{ij} - X_{i.} \qquad \text{(F124)}$$

Etter dette skriv vi

$$X_{ij} = X_{..} + (X_{i.} - X_{..}) + (X_{ij} - X_{i.}) \qquad \text{(F125)}$$

I tabell 11 har vi sett opp ein forenkla tabell 4 der generell komponent og rekkjekomponent går inn som ein verdi og der kolonnekomponent og residual går inn som den andre. Etter som

Tabell 11.

	A	B	C	D	Gjsn
1	4,5-0,5	4,5+0,5	4,5-0,5	4,5+0,5	4,5+0,0
2	4,0-1,0	4,0+0,0	4,0+1,0	4,0+0,0	4,0+0,0
3	3,5+0,5	3,5+0,5	3,5-0,5	3,5-0,5	3,5+0,0
4	2,5-0,5	2,5+0,5	2,5+0,5	2,5-0,5	2,5+0,0
5	1,5-0,5	1,5+0,5	1,5-0,5	1,5+0,5	1,5+0,0
	3,2-0,4	3,2+0,4	3,2+0,0	3,2+0,0	3,2+0,0

generell komponent ikkje yter noko som helst til variasjon, vil variasjon i første komponent i tabell 11 berre kunna tilskrivast observert personvariasjon, medan andre-komponent variasjonen kan tilskrivast innom-person variasjon. Dersom vi reknar på desse variasjonane, skulle vi få same resultat som vi fekk i tabell 10.

Det knyter seg elles andre aspekt til denne ein-vegs modellen som vi no har sett på. Det er aspekt som vi skal ta opp i ein annan samanheng og som er sterkt relaterte til ein random-parallell testteori. (Webster(1960))

5.4. Samanhengen mellom KR20 og formlar basert på Hoyt-analyse.

Ein Hoyt-analyse gjev identisk resultat med KR20. Vi skal no visa korleis vi kan utvikla (F120) med utgangspunkt i KR20. Edwards(1959) har synt oss denne samanhengen.

KR20 skriv vi

$$\rho_{XX'} = \left(\frac{k}{k-1}\right)\left(1 - \frac{\Sigma\sigma_i^2}{\sigma_X^2}\right) \qquad \text{(F126)}$$

Vi veit at

$$\Sigma\sigma_i^2 = k\bar{\sigma}_i^2$$

Den gjennomsnittlege item-varians er total innom-kolonne kvadratsum dividert med fridomsgrader k(n-1). Altså,

$$\bar{\sigma}_i^2 = \frac{SS_{wc}}{k(n-1)} = MS_{wc} \qquad \text{(wc = within column)} \qquad \text{(F127)}$$

$$\Sigma\sigma_i^2 = k\bar{\sigma}_i^2 = kMS_{wc} \qquad \text{(F128)}$$

Vi har før sett (F107) at σ_t^2 eller $\sigma_X^2 = kMS_i$. Difor kan vi no skriva (F126) slik:

$$\rho_{XX'} = \left(\frac{k}{k-1}\right)\left(1 - \frac{kMS_{wc}}{kMS_i}\right) = \left(\frac{k}{k-1}\right)\left(1 - \frac{MS_{wc}}{MS_i}\right) \qquad \text{(F129)}$$

(F129) er ein sjeldan brukt variansanalyseformel til reliabilitetsestimering, men det <u>er</u> ein alternativ formel.

Vi går eit steg vidare med (F129). Den totale innom-kolonne kvadratsum er samansett av kvadratsummen mellom rekkjene og residualkvadratsummen,

$$SS_{wc} = SS_i + SS_r \qquad \text{(F130)}$$

Vi kan difor skriva (F129)

$$\rho_{XX'} = \left(\frac{k}{k-1}\right)\left(1 - \frac{SS_i + SS_r}{k(n-1)MS_i}\right) \qquad \text{(F131)}$$

Vi utviklar og reduserer (F131):

$$\rho_{XX'} = \left(\frac{k}{k-1}\right)\left(\frac{k(n-1)MS_i - SS_i - SS_r}{k(n-1)MS_i}\right)$$

$$\rho_{XX'} = \frac{k(n-1)MS_i - SS_i - SS_r}{(k-1)(n-1)MS_i}$$

$$= \frac{k(n-1)MS_i}{(k-1)(n-1)MS_i} - \frac{SS_i}{(k-1)(n-1)MS_i} \quad \frac{SS_r}{(k-1)(n-1)MS_i}$$

$$= \left(\frac{k}{k-1}\right) - \left(\frac{1}{k-1}\right) - \frac{MS_r}{MS_i}$$

$$= 1 - \frac{MS_r}{MS_i} = \frac{MS_i - MS_r}{MS_i} \qquad \text{(F132)}$$

(F132) er den same som (F120),og dermed har vi vist at KR20 og Hoyt-koeffisienten er identiske.

5.5. Utvikling av Spearman-Browns generelle formel på variansanalyse--vilkår.

Vi har tidlegare funne at reliabiliteten til det gjennomsnittlege item, eller reliabiliteten til summen av k items, kan estimerast med variansforholdet

$$\rho_k = \frac{MS_i - MS_r}{MS_i} \text{ eller } \rho_k = \frac{MS_i - MS_{wp}}{MS_{wp}} \qquad \text{(F133)}$$

alt etter om vi bruker ein to-vegs eller ein ein-vegs modell. Vi har vidare funne at reliabiliteten til eitt item også kan skrivast som eit variansforhold

$$\rho_1 = \frac{MS_i - MS_r}{MS_i + (k-1)MS_r} \text{ eller } \rho_1 = \frac{MS_i - MS_{wp}}{MS_i + (k-1)MS_{wp}} \qquad \text{(F134)}$$

alt etter modellen som vi nyttar.

Vi vil no visa at vi frå (F133) kan utvikla Spearman-Browns generelle profeti-formel ved å bruka (F134) i staden for ein produkt-moment korrelasjon, som har vore det vanlege, og på vilkår som er svært mykje annleis enn dei Spearman og Brown opphavleg bygde på. Vi viser denne utviklinga for Hoyt-modellen, men utviklinga er og gyldig for Webster-modellen.

Vi multipliserer først teljar og nemnar i Hoyt-formelen i (F133) med k og skriv

$$\rho_k = \frac{k(MS_i - MS_r)}{kMS_i} \qquad (F135)$$

Men kMS_i kan også skrivast

$$kMS_i = MS_i + (k-1)MS_r + (k-1)(MS_i - MS_r) \qquad (F136)$$

Vi nyttar (F136) og skriv (F135) slik:

$$\rho_k = \frac{k(MS_i - MS_r)}{MS_i + (k-1)MS_r + (k-1)(MS_i - MS_r)} \qquad (F137)$$

Vi dividerer teljar og nemnar i (F137) med $MS_i + (k-1)MS_r$, og vi får

$$\rho_k = \frac{\dfrac{k(MS_i - MS_r)}{MS_i + (k-1)MS_r}}{\dfrac{MS_i + (k-1)MS_r + (k-1)(MS_i - MS_r)}{MS_i + (k-1)MS_r}} \qquad (F138)$$

Vi kan og skriva (F138) slik:

$$\rho_k = \frac{\dfrac{k(MS_i - MS_r)}{MS_i + (k-1)MS_r}}{1 + \dfrac{(k-1)(MS_i - MS_r)}{MS_i + (k-1)MS_r}} \qquad (F139)$$

Om vi no gjer oss nytte av (F134), vil vi kunna sjå at (F139) kan skrivast

$$\rho_k = \frac{k\,\rho_1}{1 + (k-1)\,\rho_1} \qquad (F140)$$

(F140) er Spearman-Browns formel. Denne formelen har vi no fått over på variansanalyse-vilkår, som er mykje meir liberale enn dei vi til vanleg gjer gjeldande for denne formelen. Det er viktig å merka seg at vi i (F140) nyttar ein intraklassekorrelasjon. Vi er elles vane med at vi nyttar ein

interklassekorrelasjon i Spearman-Browns formel. Desse to korrelasjonsomgrep skal vi sjå nærare på.

Dei vilkår (F140) byggjer på er etter Winer(1962) "that the error of measurement is uncorrelated with the true score, that the sample of n people on whom the observations are made is a random sample from the population to which inferences are to be made, that the sample of k measuring instruments used is a random sample from a population of comparable instruments, and that the within-person variance may be pooled to provide an estimate of σ_e^2." (Winer(1962),127)

5.6. Intraklassekorrelasjon og interklassekorrelasjon.

Når vi estimerer reliabilitet med variansanalyse, kan vi forsvara å seia at vi samanliknar mellom-person variasjon med innom-person variasjon. Høg reliabilitet er i denne samanheng karakterisert ved relativt stor mellom-person variasjon og relativt liten innom-person variasjon. Med basis i testteori tolkar vi dette dit at ein stor del av den observerte person-variasjon kan tilskrivast ein genuin eller ein sann person-differense.

Vi uttrykkjer reliabilitet med eit variansforhold. Dette forholdet har noko av korrelasjonsomgrepet i seg når vi ser det slik at reliabilitet er eit uttrykk for at innom-person observasjonar går saman, dei er i stor grad like; medan mellom-person observasjonar er ulike, dei varierer mykje.

Eit variansforhold som uttrykkjer korrelasjon, blir gjerne kalla ein intraklassekorrelasjon. Det er her spørsmål om samgang innanfor ein og same klasse av observasjonar. Ein annan type korrelasjon har vi i den meir tradisjonelle produkt-moment korrelasjon, som vi gjerne kan kalla ein interklasse-korrelasjon for di det her er spørsmål om observasjonar går saman over klassar. Høgd og vekt er observasjonar av ulikt slag (heilt ulike måle-einingar). Dei representerer to klassar av observasjonar, og korrelasjonen seier oss i kor stor grad mykje eller lite i den eine klassen går saman med mykje eller lite i den andre klassen.

Ein interklassekorrelasjon knyter seg til ein observasjon for kvart element, for kvar person t.d., for kvar av to klassar, og berre to klassar. Difor kan vi seia at interklassekorrelasjon representerer ein bivariat teknikk. Ein intraklassekorrelasjon er univariat i den forstand at vi ser på innom-klasse observasjonar som tilhøyrande ein og same variabel. Men ein intraklassekorrelasjon er ikkje bunden til berre to innom-klasse observasjonar for kvart element slik interklassekorrelasjon er bunden til ein observasjon for kvar av to variable for kvart element. I så måte er intraklassekorrelasjon eit meir generelt omgrep enn interklassekorrelasjon.

For å få eit mål på korrelasjonen mellom vurderarar i det store og heile i våre hypotetiske data i tabell 3 må vi rekna seks interkorrelasjonar, og gjennomsnittet av desse seks korrelasjonane kan gje oss ein tokke av kva som ligg i omgrepet intraklassekorrelasjon.

Når vi jamfører intra- og interklassekorrelasjon,kan vi kanskje våga oss på ein analogi i forholdet mellom variansanalyse og t-test.Variansanalysen er ein heilt generell test til prøving av differensar mellom M-verdiar. Det er ein all over test.Derimot maktar ikkje t-testen meir enn to M-verdiar om gongen. Ein intraklassekorrelasjon er ein all over korrelasjon for å sjå korleis innomklasse observasjonar går saman.

Vi veit at ein produkt-moment korrelasjon prøver kor godt standardskårane på dei to variable går saman. Det er god forklaring på at vi kan korrelera observasjonar frå to klassar som bruker heilt ulike måle-einingar. Ved ein produktmoment korrelasjon transformerer vi råskårefordelingar automatisk til fordelingar med $M = 0$ og $s = 1$. Dette vil seia at ein produkt-moment korrelasjon ikkje tek omsyn til, i vårt eksempel t.d., at vurderarar sprcier karakterane ulikt og at dei bruker skalaen ulikt med omsyn til generelt nivå.

I mange høve er vi interesserte i å bruka teknikkar som er sensitive for denne slags informasjon i data. I reliabilitetsestimering kan dette vera om å gjera. Difor er ein produktmoment korrelasjon ikkje generelt tenleg til dette føremålet, etter som vi ikkje kan rekna med at innom-kolonne variansar

er like og at mellom-kolonne M-verdiar er dei same. Klassisk testteori postulerer slike statistiske eigenskapar, og då er sjølvsagt ein produkt-moment korrelasjon på sin plass. I ein liberalisert testteori vil ein intraklassekorrelasjon vera generelt tenleg, men ikkje ein interklassekorrelasjon.

Vi plar definera ein intraklassekorrelasjon slik vi tidlegare har definert reliabiliteten til eitt item,

$$\rho_1 = \frac{\sigma_i^2}{\sigma_i^2 + \sigma_e^2} \qquad \text{(F141)}$$

(F141) seier oss kor stor del av ein gjennomsnittleg observasjons varians som kan tilskrivast sytematisk varians.

Det er interessant å sjå på dei matematiske relasjonane mellom intraklassekorrelasjon og interklassekorrelasjon. Vi symboliserer her intraklassekorrelasjon med R, den gjennomsnittlege interklassekorrelasjon med $\bar{r}$, og vi skal visa at $R = \bar{r}$ på visse vilkår.

For å gjera dette provet enklast råd er transformerer vi først alle skårar i ein matrise til avviksskårar frå respektive $X_{.j}$, dvs. M-verdiar for respektive kolonnar. Dette resulterer i ein matrise der alle $X_{.j} = 0$. Det totale gjennomsnittet, $X_{..}$, blir også lik null. Vi har med dette ikkje gjort noko som kan forandra korrelasjonen mellom kolonnane, etter som ein produkt-moment korrelasjon automatisk transformerer til like M-verdiar. Også Hoyt-koeffisienten byggjer på kolonnar med like M-verdiar. Det veit vi fordi vi held kolonnevariasjonen utanfor. Vår transformerte matrise er såleis framleis generelt gyldig som utgangspunkt både for interklassekorrelasjon og intraklassekorrelasjon (Hoyt-analyse).

Vi illustrerer den skisserte framgangsmåten så langt ved å transformera matrisen i tabell 3. Det har vi gjort i tabell 12 på side 76.

Vi går vidare og reknar ut kvadratsummar og produktsummar med basis i tabell 12 og set desse verdiane inn i ein kovarians matrise. (Ein varians-kovarians matrise kan kallast ein kovarians matrise når vi ser variansen som eit spesialtilfelle av kovariansen.) Dette er gjort i tabell 13.

Tabell 12.

	A	B	C	D	Sum
1	+1,2	+1,4	+0,8	+1,8	+5,2
2	+0,2	+0,4	+1,8	+0,8	+3,2
3	+1,2	+0,4	-0,2	-0,2	+1,2
4	-0,8	-0,6	-0,2	-1,2	-2,8
5	-1,8	-1,6	-2,2	-1,2	-6,8
$X_{.j}$	0,0	0,0	0,0	0,0	0,0

Tabell 13. Kovarians matrise (kvadratsummar og produktsummar).

	A	B	C	D
A	6,8	5,6	5,2	5,2
B	5,6	5,2	5,4	5,4
C	5,2	5,4	8,8	5,8
D	5,2	5,4	5,8	6,8

Totalsummen av kvadratsummar og produktsummar i tabell 13 blir 92,8, som er kvadratsummen i fordelinga av sumskårar i tabell 7 på side 64.

Vi finn at i ein variansanalyse blir, når vi nyttar (F107),

$$SS_i = SS_{X_i}/k$$
$$SS_{X_{i.}} = SS_{X_i}/k^2 \qquad \text{(F142)}$$

I variansanalysen er kvadratsummen for personar lik 1/k av kvadratsummen for sumskårane, og kvadratsummen for gjennomsnittsskårane er $1/k^2$ av sumskårekvadratsummen. Altså,

$SS_{X_i} = 92,8$

$SS_i = 23,2$

$SS_{X_{i.}} = 5,8$

Tabell 13 let seg med enkle operasjonar omgjerast til ein interkorrelasjonsmatrise, som er oppstilt i tabell 14.

Tabell 14. Interkorrelasjonsmatrise.

	B	C	D
A	0,942	0,671	0,765
B		0,798	0,908
C			0,748

$\bar{r} = 4,832/6 = 0,805$

For å ta dette generelt: Det er i alt k(k-1) interkorrelasjonar i ein korrelasjonsmatrise når vi reknar jl og lj for ulike par. Den gjennomsnittlege produkt-moment korrelasjon i ein slik matrise kan skrivast

$$\bar{\rho}_{jl} = \frac{1}{k(k-1)} \left(\frac{\Sigma x_j x_l}{N\sigma_j \sigma_l}\right) \qquad (F143)$$

Vi postulerer no like variansar i kolonnane slik at $\sigma_j = \sigma_l$. (F143) kan då skrivast

$$\bar{\rho}_{jl} = \frac{k}{k(k-1)} \left(\frac{\Sigma x_j x_l}{N\sigma_j^2}\right) = \frac{1}{k(k-1)} \left(\frac{\Sigma x_j x_l}{\Sigma x_j^2}\right) \qquad (F144)$$

I siste uttrykket i (F144) er teljaren lik produktsummen i ein kovarians matrise. I vårt språk kan denne produktsummen skrivast

$$\Sigma x_j x_l = kSS_i - SS_{tot} \qquad (F145)$$

Det er verdt å merka seg at i matrisen i tabell 12 er summen av kolonne-kvadratsummane lik total kvadratsum, fordi alle kolonnegjennomsnitt og total gjennomsnitt er lik null.

I siste uttrykk i (F144) kan nemnaren skrivast som gjennomsnittet av total kvadratsum, fordi alle kolonne-kvadratsummar er postulert like.

Altså,

$$\Sigma x_j^2 = 1/k(SS_{tot}) \qquad (F146)$$

Etter dette kan vi skriva (F144)

$$\bar{\rho}_{jl} = \frac{1}{k(k-1)}\left(\frac{kSS_i - SS_{tot}}{1/k(SS_{tot})}\right) \qquad (F147)$$

Vi veit at $SS_{tot} = SS_i + SS_j + SS_r$. I vårt tilfelle, sjå tabell 12, er $SS_j = 0$. Difor kan vi skriva $SS_{tot} = SS_i + SS_r$. (F147) kan etter dette skrivast

$$\bar{\rho}_{jl} = \frac{1}{k(k-1)}\left(\frac{kSS_i - (SS_i + SS_r)}{1/k(SS_i + SS_r)}\right)$$

$$= \frac{(k-1)SS_i - SS_r}{(k-1)(SS_i + SS_r)} \qquad (F148)$$

Vi dividerer no både teljar og nemnar i (F148) med (k-1)(n-1) og får

$$\bar{\rho}_{jl} = \frac{\frac{(k-1)SS_i}{(k-1)(n-1)} - \frac{SS_r}{(k-1)(n-1)}}{\frac{(k-1)SS_i}{(k-1)(n-1)} + \frac{(k-1)SS_r}{(k-1)(n-1)}} = \frac{MS_i - MS_r}{MS_i + (k-1)MS_r} \qquad (F149)$$

(F149) er identisk med (F121). (F149) er estimasjonsformelen for intraklassekorrelasjon når vi held kolonnevariasjon utanfor. Når vi postulerer like kolonnevariansar, har vi no vist at intraklassekorrelasjonen er lik den gjennomsnittlege interklassekorrelasjon (produkt-moment korrelasjon).

Vi har tidlegare funne at intraklassekorrelasjonen for våre data estimert med (F149) blir 0,787. Den gjennomsnittlege produkt-moment korrelasjon har vi fått til 0,805. At dei ikkje er like må skriva seg frå at vurderar-variansane (altså innomkolonne-variansane) ikkje er dei same.

Det er lite å finna om intraklassekorrelasjon i læreboker. Hays(1963) har litt med. Vi viser til Haggard(1958) og Ljung (1960)

6. G-teori for ikkje-stratifiserte komposita.

Vi har sett kor elegant Tryon kunne kvitta seg med dei strenge statistiske krav som klassisk teori sette til komponentane i ein test for å estimera reliabilitet på internal consistency basis. Hans domenesampling-teori byggjer på dei gjennomsnittlege statistiske eigenskapane til testen eller rettare komponentane, og han finn at vi framleis kan estimera reliabilitet med alpha-koeffisienten. Tryons krav er knytte til komposita og ikkje til komponentar. Med det har han greitt å koma seg unna å setja krav til observerte data. Det bør vi kunna rekna som ein stor føremon.

Tryons restriktive krav til komposita har gjort at han har måtta avgrensa sitt univers av testar til berre å gjelda testar med like variansar og interkorrelasjonar. På ein måte sett er vi då ikkje komne lenger enn vi var: Vi tenkjer på Spearman-Yule tradisjonen og på Brown-Kelley tradisjonen med omgrepet parallelle testar. Det som er nytt hos Tryon i forhold til denne klassiske tradisjonen er at han bruker internal consistency utan vilkår for å estimera den hypotetiske korrelasjon mellom parallelle testar, medan klassikarane reint konkret måtte korrelera parallelle testar. Spearman-Brown tradisjonen byggjer og på internal consistency, men på strenge vilkår. Vi må difor med god grunn ha lov til å rekna Tryon som den mest liberale innanfor klassisk tradisjon, om vi i det heile skal rekna han til denne tradisjonen.

Likevel kan vi seia at konsekvensane av eit random sampling synspunkt ikkje har fått fritt spelerom med Tryons teori. Den fulle konsekvens av random sampling av items vil gjelda både komponentar og komposita, ikkje berre komponentane som hos Tryon. Testar som er sammensette av like mange tilfeldelege items, vil ha ulik varians og ulike interkorrelasjonar. Det vil med andre ord seia at den fulle konsekvens av ein random sampling teori ikkje gjev oss parallelle testar i klassisk forstand.

Med det må vi rekna med at vi er komne i ein vanskeleg situasjon for reliabilitetsestimering, som tilsynelatande har vore heilt avhengig av konstruksjonen parallelle testar. Ei løysing ligg nær, men det er ei lite tilfredsstillande løysing reint praktisk: Vi tek eit sampel av random-parallelle testar, interkorrelerer

dei og reknar ut den gjennomsnittlege korrelasjon mellom desse testane. Denne løysinga er upraktisk for di vi no aller helst skulle korrelera meir enn to konkrete testar. Vi vil sjølvsagt fram til ein tolleg påliteleg gjennomsnittleg korrelasjon, og då krevst det mange testar. Her ser vi føremonen med klassisk definisjon av parallelle testar: Korrelasjonen mellom to testar er nok for di alle interkorrelasjonar er like.

Det store ved Tryons reliabilitet er at han bruker internal consistency utan krav til komponentane for å estimera korrelasjonen mellom parallelle testar. Det er dette som er den elegante løysing både teoretisk og praktisk. Det er denne løysinga vi gjerne ville bruka i ein ny situasjon med random-parallelle testar.

6.1. Reliabilitet redefinert.

Det skulle vera klårt at vi innanfor ein random sampling teori vanskeleg kan forsvara å definera reliabilitet som korrelasjonen mellom parallelle testar, slik klassisk teori gjer det. Heilt grunnleggjande har denne definisjonen ikkje vore, kan vi vel seia. Det vesentlege ved reliabilitet er å få estimert kor stor del av den observerte skårevarians som vi kan tilskriva sann skårevarians. Det er dette som er den konstitutive reliabilitetsdefinisjonen, medan korrelasjonen mellom parallelle testar like gjerne kan kallast ein operasjonell definisjon når klassiske krav er gjort gjeldande.

Vi har tidlegare sett at under klassiske vilkår kan vi bruka fleire alternative definisjonar av reliabilitet:

$$\rho_{XX'} = \frac{\sigma_T^2}{\sigma_X^2} = \rho_{XT}^2 \qquad \text{(F150)}$$

Under random sampling vilkår vil den kvadrerte korrelasjon mellom observerte og sanne skårar vera ein fullt tenleg definisjon av reliabilitet. Denne definisjonen har vore lansert for mange år sidan, Kelley (1924), Cureton (1931); men har ikkje vorte allmont akseptert som definisjon, endå om denne definisjonen er den mest generelle. Grunnen må vel vera at så lenge vi vedkjenner oss klassisk teori, er det unødvendig å skifta til ny definisjon.

Seinare års nyorientering innanfor rein testteori viser likevel at eit skifte er no i ferd med å skje i definisjonsspørsmålet. Både Cronbach,Rajaratnam,Gleser(1963) og Lord og Novick(1966) definerer reliabilitet som den kvadrerte korrelasjon mellom observerte og sanne skårar, altså ρ^2_{XT}. Denne determinasjonskoeffisienten seier oss kor stor del av den observerte skårevarians som lineært kan predikerast frå sann skårevarians. Dermed tilfredsstiller denne definisjonen også tradisjonelle krav til ein reliabilitetskoeffisient.

6.2. Alpha er i meste fall lik den definerte reliabilitet.

Under klassiske vilkår, med Tryon som den mest liberale, har vi sett at vi kan estimera den kvadrerte korrelasjon mellom observerte og sanne skårar med alpha, etter som alpha og determinasjonskoeffisienten, ρ^2_{XT}, på slike vilkår er like. Under random sampling vilkår vil ikkje denne likskapen lenger gjelda. Alpha blir i den nye situasjonen eit underestimat av reliabiliteten. Så langt vi veit er dette den einaste måten å estimera reliabiliteten på, når vi ønskjer å byggja estimatet på internal consistency.

Vi skal no visa at alpha i meste fall er lik reliabiliteten. Provet vårt gjer vi så enkelt som råd er; det vil seia vi tenkjer oss at vi berre har to komponentar i vårt kompositum, slik at $X = Y_1 + Y_2$.

Vi har tidlegare sett at alpha-koeffisienten bygt på eit kompositum av to komponentar er Guttmans formel.(Sjå s.17 og s.20.) Vi skal altså visa at reliabiliteten er lik eller større enn Guttmans koeffisient:

$$\rho^2_{XT} \geq 2\left(1 - \frac{\sigma(Y_1)^2 + \sigma(Y_2)^2}{\sigma^2_X}\right) \qquad \text{(F151)}$$

Vi presiserer med ein gong at vi ikkje rekmar Y_1 og Y_2 for parallelle mål i klassisk forstand. Y_1 og Y_2 er to tilfeldelege mål med kvar sin sann skåre, T_1 og T_2, som ikkje er postulert like.

Vi skal først visa at kovariansen mellom Y_1 og Y_2 er lik kovariansen mellom T_1 og T_2.

$$\begin{aligned} \varrho(Y_1,Y_2) &= \varrho(T_1,T_2) \\ &= \varrho((T_1 + E_1),(T_2 + E_2)) \\ &= \varrho(T_1,T_2) + \varrho(T_1,E_2) + \varrho(T_2,E_1) + \varrho(E_1,E_2) \\ &= \varrho(T_1,T_2) \end{aligned} \qquad (F152)$$

Vi har i utviklinga av (F152) postulert uavhengighet mellom sann komponent og feilkomponent og mellom feilkomponentar.Med det fell tre av uttrykka frå, og vi står att med kovariansen mellom dei sanne komponentane. Dette "vesle" provet får vi bruk for litt lenger ute i det "store" provet.

Vi slår fast ein sjølvsagt ting:

$$(\varrho(T_1) - \varrho(T_2))^2 \geq 0 \qquad (F153)$$

(F153) seier ikkje anna enn at anten er standardavviket til T_1 og T_2 like eller så er dei ulike. Vi kan utvikla (F153),og vi får

$$\varrho(T_1)^2 + \varrho(T_2)^2 - 2\varrho(T_1)\varrho(T_2) \geq 0$$

$$\varrho(T_1)^2 + \varrho(T_2)^2 \geq 2\varrho(T_1)\varrho(T_2) \qquad (F154)$$

I den matematiske statistikk finst der ein ulikskap som blir kalla Cauchy-Schwartz ulikskapen, som i vår samanheng kan skrivast

$$\varrho(T_1)\varrho(T_2) \geq |\varrho(T_1,T_2)| \geq 0 \qquad (F155)$$

(F155) seier at produktet av standardavviket til T_1 og T_2 i det minste er lik den absolutte verdi av kovariansen mellom T_1 og T_2, som igjen i det minste er lik null. Det er ikkje utan vidare lett å sjå at dette må så vera. Lettare blir det når vi dividerer ulikskapane med $\varrho(T_1)\varrho(T_2)$. Då får vi

$$\frac{\varrho(T_1)\varrho(T_2)}{\varrho(T_1)\varrho(T_2)} \geq \frac{|\varrho(T_1,T_2)|}{\varrho(T_1)\varrho(T_2)} \geq 0 \qquad (F156)$$

Andre lekken i ulikskapen i (F156) er kovariansen til T_1 og T_2 dividert med produktet av standardavviket til T_1 og T_2, og det er som kjent ein produkt-moment korrelasjon. (F156) kan difor skrivast

$$1 \geq |\rho(T_1,T_2)| \geq 0 \qquad \text{(F157)}$$

(F157) seier ikkje meir enn vi har visst frå før. Men kanskje det no skulle vera lettare å akseptera (F155), når vi ser at det går ei utvikling frå (F155) til (F157).

Ved å gjera oss nytte av (F154) og (F155) kan vi skriva

$$\sigma(T_1)^2 + \sigma(T_2)^2 \geq 2|\sigma(T_1,T_2)| \geq 2\sigma(T_1,T_2) \qquad \text{(F158)}$$

Når summen av variansane til T_1 og T_2 er minst like stor som 2 gonger produktet av standardavviket til T_1 og T_2 (F154), så må etter (F155) summen av dei to variansane også bli minst like stor som 2 gonger den absolutte verdi av kovariansen mellom T_1 og T_2, som igjen må vera minst like stor som kovariansen mellom T_1 og T_2, som kan vera negativ.

Til summen av Y_1 og Y_2 svarar observert kompositum skåre X, og denne X har ein sann skåre T som er ein sum av T_1 og T_2. Variansen til sann skåre på testen er ein sum av komponentvariansane og 2 gonger kovariansen mellom komponentane.

$$\sigma(T)^2 = \sigma(T_1)^2 + \sigma(T_2)^2 + 2\sigma(T_1,T_2) \qquad \text{(F159)}$$

Dersom vi tek bort andre lekken i (F158) og legg til $2\sigma(T_1,T_2)$ på begge sider av den ulikskapen som er igjen, får vi

$$\sigma(T_1)^2 + \sigma(T_2)^2 + 2\sigma(T_1,T_2) \geq 2\sigma(T_1,T_2) + 2\sigma(T_1,T_2) \qquad \text{(F160)}$$

Venstre side av (F160) er som vi ser, sann skårevarians på testen slik den er skriven i (F159). Høgre side av ulikskapen (F160) er sjølvsagt lik 4 gonger kovariansen mellom T_1 og T_2. Altså

$$\sigma(T)^2 \geq 4\sigma(T_1,T_2) \qquad \text{(F161)}$$

Men (F161) kan etter (F152) også skrivast

$$\sigma(T)^2 \geq 4\sigma(Y_1,Y_2) \qquad (F162)$$

Dersom vi dividerer begge sider av ulikskapen (F162) med $\sigma(X)^2$, vil vi sjå at venstre side blir reliabiliteten. Vi får

$$\frac{\sigma(T)^2}{\sigma(X)^2} = \rho(X,T)^2 \geq \frac{4\sigma(Y_1,Y_2)}{\sigma(X)^2} \qquad (F163)$$

Høgre side av ulikskapen (F163) kan skrivast om. Kovariansen mellom Y_1 og Y_2 kan vi og skriva som differensen mellom kompositum-varians og summen av komponentvariansane. Dermed skriv vi

$$\rho(X,T)^2 \geq 2\left(\frac{2\sigma(Y_1,Y_2)}{\sigma(X)^2}\right) = 2\left(\frac{\sigma(X)^2 - (\sigma(Y_1)^2 + \sigma(Y_2)^2}{\sigma(X)^2}\right) \qquad (F164)$$

(F164) kan så skrivast

$$\rho(X,T)^2 \geq 2\left(1 - \frac{\sigma(Y_1)^2 + \sigma(Y_2)^2}{\sigma(X)^2}\right) \qquad (F165)$$

Vi er komne fram til (F165) som er identisk med (F151). Dermed har vi vist at (F151) er rett.

Det ligg nær å tenkja seg at det må vera råd å føra eit generelt prov for at reliabiliteten er minst like stor som alpha. Dette generelle provet finn vi i Novick og Lewis(1967). Det er noko meir komplisert enn det vi her har ført. Vi viser til dette generelle provet og slår fast at

$$\rho(X,T)^2 \geq \left(\frac{k}{k-1}\right)\left(1 - \frac{\sigma(Y_i)^2}{\sigma(X)^2}\right) \qquad (F166)$$

6.3. G-koeffisienten.

I eit uendeleg item-univers vil vi kunna trekkja uendeleg mange testar(komposita), kvar med k items. Desse testane har ikkje like interkorrelasjonar, og dei vil gje oss ulike estimat av reliabiliteten. Det er verdt å merka seg at (F166)

gjev oss ein spesifikk reliabilitet: Kor nøyaktig kan vi slutta oss til sann skåre når vi observerer med just dette sampel av items,med testen X_1 t.d.? Det er truleg så at vi berre sjeldan er interesserte i ein slik spesifikk reliabilitet,som berre gjeld ein bestemt test av uendeleg mange liknande, men ikkje nødvendigvis parallelle testar. Som regel vil vi vera interesserte i å vita reliabiliteten til ein eller annan test meir eller mindre tilfellelg trekt frå universet av testar. Det er denne (generiske) forventa reliabilitet vi vil vera best tent med, altså den gjennomsnittlege reliabilitet i universet av testar. Denne forventa reliabilitet kan vi skriva $E\rho(X,T)^2$.

Det er denne gjennomsnittlege kvadrerte korrelasjonskoeffisient mellom tilfellelege testar trekte frå universet av testar og gjennomsnittet av alle testar i universet som Cronbach, Rajaratnam,Gleser(1963) har definert som G-koeffisienten (the generalizability coefficient). Denne G-koeffisienten indikerer kor nøyaktig vi jamtover kan slutta oss til individuelle differensar i universskåren (sann skåre) frå individuelle differensar som er observerte med ein eller annan test trekt frå universet.(Cronbach,Ikeda,Avner(1964),728)

G-koeffisienten kan sjølvsagt ikkje bestemmast direkte. Han må estimerast, og vi kan bruka intraklassekorrelasjon for dette føremålet. Alpha gjev oss, slik tilfellet var med den spesifikke reliabiliteten, eit underestimat av G-koeffisienten. Provet for dette finn vi i Rajaratnam,Cronbach,Gleser(1965),41. Det kan minna noko om det provet vi nett førte for å visa at den spesifikke reliabiliteten er minst like stor som alpha, når eit kompositum berre har to komponentar.

6.4. Reliabilitet reformulert.

$E\rho(X,T)^2$, G-koeffisienten, gjev oss den gjennomsnittlege proporsjon av variansen i observerte skårar som lineært kan predikerast frå universskårane. Det er ikkje postulert ekvivalente mål i denne G-teorien som fører fram til denne type koeffisient. Eit mål blir sett på som om det var sampla frå eit univers av liknande men ikkje nødvendigvis ekvivalente mål. Det representerer eit sant mål, ein universskåre. Vi ønskjer å vita kor godt eit slikt observert mål represen-

terer universskåren, eller i tråd med det vi har sagt tidlegare, kor godt slike mål i det store og heile representerer universskåren. Vi ønskjer å generalisera frå ein observasjon til universet av observasjonar. Det er dette som er kjernen i G-teorien (generalizability theory), lansert av Cronbach,Rajaratnam,Gleser i ein rapport i British Journal of Statistical Psychology i 1963. G-teorien er ei reformulering av klassisk reliabilitet som både har syntaktiske og semantiske implikasjonar.

Utgangspunktet for denne reformulering av reliabilitetsteorien er dei restriktive krav som i klassisk teori blir sette til observasjonar, at dei skulle ha like M-verdiar, like variansar og like interkorrelasjonar. Slike urealistiske krav kan berre stettast i den mest omhyggelege test-konstruksjon. Her kan vi greia å laga ekvivalente målingsinstrument. Men vi bruker og har bruk for psykologiske og pedagogiske mål som i praksis ikkje stettar klassiske krav. Det kan t.d. gjelda karaktergjeving og klasseromsobservasjon. Også slike mål må vurderast med omsyn til kor godt dei tener sitt føremål. Men for slike mål har vi vanta grunnlaget for ei vurdering. Vi har ikkje hatt nokon teori for denne type observasjonar. Det klassiske grunnlaget gjeld ikkje for mål som så opplagt ikkje er ekvivalente.

"Raters have different means and standard deviations and unequal intercorrelations. However, in some way reliability of ratings must be established. So, what has to be done is to get away of parallelism of classical theory. In this context the concept of universe or domain will be helpful. In the rater-situation, for instance, to ask about rater agreement is to ask how well we can generalize from one set of ratings to ratings by other raters. The observations obtained are regarded as members of a postulated class of observations to which we can generalize. A particular observation fits within many different universes. It is necessary to indicate the universe of generalization, rather than to imply that there is one underlying "true" score. In order to make the concept of universe stick, the universe has to be described, i.e.the universe of conditions of observations over which we want to generalize has to be specified unambiguously." (Cronbach (1966),116)

Dette er eit nytt test-teoretisk feste for reliabilitets-estimering med liberale og realistiske krav til data. I denne test-teoretiske nyorientering ligg dei semantiske implikasjonane for reliabilitetsomgrepet.

CRG(1963) finn ein syntaktisk rasjonale for sitt test-teoretiske grunnlag i Cornfield og Tukey(1956). Cornfield og Tukey har gjort greie for ein noko liberalisert(?) variansanalyse-modell, som dei kallar ein "duebur"-modell (a pigeon-hole model). Dei tenkjer seg ein to-vegs variansanalyse med eit univers av rekkjer og eit univers av kolonnar med "duebur" i kryssingane mellom rekkjer og kolonnar. I desse cellene (duebura) tenkjer dei seg vidare eit univers av potensielle observasjonar. Vi samplar n rekkjer og k kolonnar, og innanfor kvar av dei kn celler vi dermed får,samplar vi m observasjonar, X_{ij}. Cornfield og Tukey byggjer på vilkår som i vår samanheng vil seia random sampling av personar (rekkjer) og items (kolonnar) og korrelerte interaksjonar mellom personar og items, i.e. "the cell effects are tied to the corresponding row and column effects in the sense that when a particular person and condition (item) have been selected all three values are determined."(CRG(1963),150) Cornfield og Tukey kjem fram til nett dei same forventa variansuttrykk som tidlegare hadde vore brukt på meir restriktive(?) modellar.

Med basis i denne variansanalyse-modellen viser CRG korleis vi kan bruka intraklassekorrelasjonen i reliabilitetsanalysar utan andre vilkår enn at personar og items er sampla tilfelleleg (random) og uavhengig frå ein populasjon av personar og frå eit univers av items.

Endå ei "liberalisering" skjer i utviklinga av G-teorien. CRG viser at vi kan forkasta kravet om unit-rank, kravet om at eit kompositum berre skal måla ein generell faktor. Dei seier:

"To make clear the extent to which we break away from restrictive assumptions, we may set up a model comparable to $X_{pi} = M + (M_p - M) + (M_i - M) + e_{pi}$ but far more general. Scores are considered to be determined by the following factors:

The first centroid factor of the covariances between conditions in the universe; the person's score on this factor is the universe score M_p.

$F_1F_2...$; other centroid factors required to account for covariances between conditions. For convenience, scores X_{pF} on any factor are standardized with variance one.

d_i; residual variance after removal of the centroid factors. No restriction is placed on $V_{d_{pi/i}}$, the variance of d_{pi} with i fixed.

Now, introducing factor loadings $b_i ø_{M_p}$ and a_{Fi}, we may write

$$X_{pi} = M + b_i(M_p-M) + (M_i-M) + {}_F(a_{Fi}X_{pF}) + d_{pi}$$

The mean of a_{Fi} for any F is zero, and the mean of the b_i is one. It is now evident that we have discarded the unit-rank assumption and the Spearman assumption that the regression of X_{pi} on M_p is the same for all i." CRG(1963),147) Som vi ser, vil vi enda opp med $X_{pi} = M + (M_p-M) + (M_i-M) + d_{pi}$, som er lik den modellen vi har gjeve før.(Sjå s.55)

Denne modellen, som er den same som Hoyts og som vel Jackson (1939) først lanserte, er tilsynelatande mykje ulik Spearmans observerte skåre som har to komponentar, ein sann skåre og ein feilskåre. Det er likevel ein nær samanheng mellom desse to modellane. Vi kan merka oss at (M_i-M) under klassiske vilkår alltid blir null og kan såleis reknast som definert bort frå modellen for di alle testar har same M-verdi. Vidare bør vi merka oss at M, total M-verdi, er invariant over personar og items: Den yter ingen ting til skårevariasjon.(Sjå s.56) Variansanalyse-modellen som vi etablerer her, degenererer til Spearmans modell når klassiske vilkår er gjort gjeldande.

Det er interessant å konstatera at CRG på sine liberale vilkår kjem fram til akkurat dei same formlane for G-estimering som vi kan utvikla for reliabilitetsestimering utan å basera oss på nokon ny teori. Vi viser til kapittel 5 og reliabilitetsformlane der, som no også gjeld som G-formlar. Berre formel 29 i CRG(1963) er ny. Denne formelen krev ein kommentar.

6.5. G-studie og D-studie.

CRG skil mellom det dei kallar ein G-studie og ein D-studie. Ein G-studie har til føremål å prøva ut eit instrument eller kanskje rettare ein type instrument, ved å sjå kor godt det

let seg gjera å generalisera frå observert skåre til universskåre, eller også å sjå i kor stor grad det postulerte eller definerte univers av observasjonar "heng saman".

Ein D-studie har til føremål å koma fram til resultat som kan vera grunnlag for ei avgjerd (decision) som gjeld personar eller grupper av personar t.d. Her er det spørsmål om kor godt eit instrument bestemt til praktisk bruk vil predikera universskåren.

6.6. Test design.

Ein G-studie og ein D-studie kan ha ulikt design. I denne samanhengen, vi held oss her til ikkje-stratifiserte komposita, vil ulikt design seia to ting: 1) at k varierer frå G til D og 2) at vi bruker crossed eller nested design.

6.6.1. k varierer frå G til D.

I ein G-studie kan vi bruka fleire items enn i ein D-studie t.d. Vi prøver ut eit observasjonsskjema med 4 observatørar (G-studie), men vi bruker skjemaet i praksis med 2 observatørar (D-studie). Sett at vi skal finna ein G-koeffisient for D-data med utgangspunkt i G-data. Vi har k items i G-data, og vi ønskjer å bruka k' items i D-data. Kor stor blir G-koeffisienten for D-data? Vi kan utvikla følgjande estimasjonsformel:

Den uvekta universskårekomponenten frå G-data er 1/k av $MS_i - MS_r$. Vi vektar så denne komponenten med k'. Den observerte skårevarians får feilvariansen, MS_r, i tillegg. Vi skal fram til eit estimat av alpha(k'), som vi skriv

$$\text{alpha}(k') = \frac{k'\sigma_i^2}{k'\sigma_i^2 + \sigma_e^2} \qquad \text{(F167)}$$

(F167) blir estimert slik:

$$\text{alpha}(k') = \frac{k'/k(MS_i - MS_r)}{k'/k(MS_i - MS_r) + MS_r}$$

$$= \frac{k'(MS_i - MS_r)}{k'(MS_i - MS_r) + kMS_r}$$

$$\text{alpha}(k') = \frac{k'(MS_i - MS_r)}{k'MS_i + kMS_r - k'MS_r}$$

$$= \frac{k'(MS_i - MS_r)}{k'MS_i + (k-k')MS_r} \qquad \text{(F168)}$$

(F168) er CRG's formel (29) som gjev oss G-koeffisienten til D-data med utgangspunkt i G-data. Så vidt vi kjenner til, er denne formelen ei nyskaping.

Vi kan og koma fram til (F168) ved å ta utgangspunkt i den gjennomsnittlege sumskåre over k' items. Når vi går frå eitt item til k' items, vil feilvariansen i den gjennomsnittlege skåre bli 1/k' av feilvariansen til eitt item. Her følgjer vi same tankegang som når vi reknar standardfeilen til M. (Sjå s.61)

Vi ønskjer å estimera $\sigma_i^2/(\sigma_i^2 + (1/k')\sigma_e^2)$, og det gjer vi på følgjande måte:

$$\text{alpha}(k') = \frac{1/k(MS_i - MS_r)}{1/k(MS_i - MS_r) + (1/k')MS_r} \qquad \text{(F169)}$$

Når vi utviklar (F169), kjem vi fram til (F168).

Det er ikkje lett å sjå at (F168) er lik Spearman-Browns generelle profetiformel, men det kan vi visa. Likskapen er ein likskap i form. Spearman-Brown byggjer som kjent på interklassekorrelasjon, medan vi i G-teori held oss til intraklassekorrelasjon.

Alpha blir symbolisert med a, og vi startar med Spearman-Browns formel, som vi skriv

$$a(k') = \frac{na(k)}{1 + (n-1)a(k)} \qquad \text{(F170)}$$

I (F170) er $n = k'/k$, forholdet mellom talet på items i D-data og talet på items i G-data. Altså

$$a(k') = \frac{(k'/k)a(k)}{1 + (k'/k - 1)a(k)} \qquad \text{(F171)}$$

a(k) er Hoyt-koeffisienten, og (F171) kan difor skrivast:

$$a(k') = \frac{(k'/k)(MS_i - MS_r)/MS_i}{1 + ((k'/k)-1)(MS_i - MS_r)/MS_i}$$

$$= \frac{(k'/k)(MS_i - MS_r)}{MS_i + ((k'-k)/k)(MS_i - MS_r)}$$

$$= \frac{k'(MS_i - MS_r)}{kMS_i + (k'-k)(MS_i - MS_r)} \qquad \text{(F172)}$$

Når vi multipliserer ut andre lekken i nemnaren i (F172), ordnar og stryk, får vi

$$a(k') = \frac{k'(MS_i - MS_r)}{k'MS_i + (k'-k)MS_r}$$, som er (F168).

Når vi skal estimera G-koeffisienten for D-data med utgangspunkt i G-data, er (F168) den generelle formel. Det er likevel to spesialtilfelle av denne formelen som det kan vera nyttig å kjenna til. Desse to spesialtilfella har vi utvikla tidlegare i ein annan samanheng.

Det er forholdet mellom k og k' som er avgjerande for kva formel vi skal bruka.

G-data	D-data	Formel
$k \neq k'$		$a(k') = \frac{k'(MS_i - MS_r)}{k'MS_i + (k'-k)MS_r}$, som er (F168).
$k = k'$		$a(k') = \frac{MS_i - MS_r}{MS_i}$, som er Hoyt-formelen, (F120).
$k > 1$,	$k'=1$	$a(k') = \frac{(MS_i - MS_r)}{MS_i + (k-1)MS_r}$, som er (F121).

CRG(1963) seier at desse formlane er rettferdig-gjorde "without assuming equivalence, unit-rank or uniformity of the within cell variances σ_e^2. Without equivalence, however, we can only say that the intraclass correlation alpha is approximately equal to $E\rho_{ii}$, and a more-or-less close lower bound to $E\rho_{Mi}{}^2$." (CRG(1963), 151)

6.6.2. Crossed og nested design.

Når alle personar blir prøvde med dei same items, får vi det CRG(1963) kallar matcha data. Vi samplar eit sett av items, og berre eitt, og gjev dette sett av items til alle personar. Dette er tradisjonell test-type. Vi har bruk for k items.

Men vi kan og sampla eit sett av items frå itemuniverset like mange gonger som vi har personar. Når kvar person får sitt sampel av items, har vi med umatcha data å gjera. Vi har no bruk for kn items.

I variansanalyse-terminologi vil desse to typar av data svara til crossed og nested design. Når vi kryssar personar og items, vil det seia det same som at alle personar får dei same items. Når personar får ulike sett av items, vil det seia at items er spesifikke for kvar person, eit sett av items går ikkje igjen frå person til person. Vi seier at items er "nested within persons". I dette tilfellet kan vi berre identifisera innom-person variasjon i tillegg til mellom-person variasjon, medan vi for crossed design identifiserer både variasjon mellom items og interaksjon mellom personar og items i tillegg til mellom-person variasjon.

Desse to typar av design, crossed og nested, svarar til dei to modellane for reliabilitetsestimering, Hoyt-modellen og Webster-modellen, som vi har sett på tidlegare. (Sjå s.59 og s.66.)

I testing har nested design ikkje vore brukt, og det ville vanskeleg kunna forsvarast å bruka dette designet innanfor klassisk test-teori. Derimot vil det vera i fullt samsvar med G-teori å nytta nested design.

Frå brukar-synspunkt kan det by på store føremoner med nested design i testing. Det er eit avgjort aber med standardiserte testar og standpunktprøver som blir brukte om igjen og om igjen, at dei kan bli kjente for dei som skal få dei. Ved å nytta oss av nested design ville kvar elev kunna få sitt sett av items til ei standpunktprøving, eit random sampel frå eit univers av slike items, og aldri det same sett av items om igjen ved ei retesting t.d. Med elektroniske hjelpemidlar er ikkje dette lenger ein heilt urealistisk tanke, og det er grunn til å merka seg med stor interesse at Educational Testing Service

visstnok arbeider med konkrete planar for denne type testing. (Cronbach(1966),134) For andre typar av instrument vil nested design vera meir vanleg, observasjon av born til ulike tider, karaktergjeving med ulike sensorar t.d. Men vi har knapt nok hatt teori til denne type data. Det reknar vi med at vi no har.

6.6.3. Eksempel.

Det hypotetiske materialet som vi presenterte på s.56 og som vi rekna reliabilitet på på s.63, kan vera G-data. Vi ønskjer G-koeffisienten for D-data når vi bruker a) 2 vurderarar, dei same 2 for alle 5 stilar og b) 8 vurderarar, dei same 8 for alle 5 stilar. Vi bruker (F168).

a)

$$\text{alpha}(2) = \frac{2(5{,}800 - 0{,}367)}{2 \cdot 5{,}800 + (4-2)0{,}367} = \frac{10{,}866}{12{,}334} = 0{,}881$$

b)

$$\text{alpha}(8) = \frac{8(5{,}800 - 0{,}367)}{8 \cdot 5{,}800 + (4-8)0{,}367} = \frac{43{,}464}{44{,}932} = 0{,}967$$

Alpha(2) og alpha(8) er estimat av $\rho(X,T)^2$ for 2,respektive 8, vurderarar. Med 2 vurderarar kan vi rekna med at i det minste 88% av sumskårevariansen er forklart ved universskårevariansen. For 8 vurderarar er tilsvarande tal 97%. Dette er vårt mål på kor godt vi kan generalisera frå observert skåre til universskåre.

Vårt illustrasjonsmateriale er matcha data. Våre data skriv seg frå crossed design.

På basis av crossed design kan vi estimera G-koeffisienten for nested design. Det kan vi forklara på denne måten: Ved å addera kvadratsummen for mellom items og person-item interaksjon får vi innom-person kvadratsum. Dette er ein samanblanda (confounded) kvadratsum som vi også kunne rekna med å få om våre data var baserte på nested design. Nested design i denne samanheng vil seia at dei fem stilane er vurderte av forskjellige vurderarar alle fire gongene, om vi vil, fem forskjellige sensorlag. I dette tilfellet har vi brukt for 4,5-20 tilfeldeleg valde vurderarar mot fire

vurderarar, eller eitt sensorlag når designet er crossed. Vår innom-person kvadratsum er eit mål på det gjennomsnittlege sensorlags semje i si karaktergjeving, her uttrykt ved karakterspreiing, for kvar stil.

For å estimera ein G-koeffisient for denne type data når G-studien byggjer på matcha data, tek vi utgangspunkt i den analysen vi gjorde med Webster-modellen på s.68, som er ein nested-design-modell. Vi bruker no innom-person variansen som feilvarians. Vi ønskjer G-koeffisienten når vi bruker a) 10 forskjellige vurderarar og b) 40 forskjellige vurderarar. Vi får no

a)

$$\text{alpha}(2) = \frac{2(5,8 - 0,4)}{2 \cdot 5,8 + (4-2)0,4} = \frac{10,8}{12,4} = 0,871$$

b)

$$\text{alpha}(8) = \frac{8(5,8 - 0,4)}{8 \cdot 5,8 + (4-8)0,4} = \frac{43,2}{44,8} = 0,964$$

Når 10 forskjellige vurderarar gjev sin karakter til 5 forskjellige stilar, kan vi rekna med at 87% av observert skårevarians kan tilskrivast universskårevarians. Når vi bruker 40 forskjellige vurderarar, vil 96% av observert skårevarians vera forklart ved universskårevariansen.

Når vi bruker denne modellen, seier vi ikkje at minst 87%, respektive minst 96%, av observert skårevarians er forklart ved universskårevariansen. Med nested design er intraklassekorrelasjonen ikkje eit underestimat av den kvadrerte korrelasjon mellom observert skåre og universskåre. Intraklassekorrelasjonen er i dette tilfellet lik den kvadrerte korrelasjonen. Vi går ikkje nærmare inn på dette her, men vi viser til Webster(1960) og til CRG(1963).

Om estimering av G-koeffisienten for nested design på basis av crossed design, sjå Rajaratnam(1960).

6.7. Generaliseringsuniverset.

I definisjonen av G-koeffisienten står universskåren sentralt. Universskåren er den gjennomsnittlege skåre ein person ville oppnå om han vart prøvt på alle items i universet, altså ein skåre basert på universet av potensielle observasjonar.

Denne universskåren kan punktestimerast, om vi er interesserte i det.(Lord og Novick(1966),Cronbach(1966))

I G-teori er vi meir interesserte i å vita kor godt dei observerte skårar korrelerer med universskårane for dermed å få eit mål på kor godt vi jamtover kan estimera universskåre på basis av observert skåre. Vi vil vita i kor stor grad vi har grunnlag for å generalisera til universskåren, den perfekte skåre, eller som vi tradisjonelt har sagt, den sanne skåre.

"When an investigator makes an observation, he never regards that measurement as meaningful in its own right. Rather, he regards it as a sample from a universe of observations that might have been made with other instruments, other observers, or on other occasions. Any conclusions he draws from his own data will (likewise) be generalized over some universe of comparable observations. In order to judge the dependability of such a generalization, he must determine how closely a sample of behavior such as his agrees with the result to be expected from making all possible measures in the universe."
(Rajaratnam, Cronbach, Gleser(1965),41)

Den type univers av observasjonar vi held oss til i denne omgang, er eit ikkje-stratifisert univers av potensielle observasjonar. Det vil seia at vi bruker berre eitt identifikasjonsattributt ved items, slik at alle observasjonar i vårt univers kan gå inn under ei og same klassifisering. Sjølvsagt vil observasjonar nærsagt alltid kunna klassifiserast etter fleire identifikasjonsattributt. Ein observasjon kan representera både eit univers av former og eit univers av applikasjonar (trials). Difor er det så viktig at det univers av observasjonar vi ønskjer å generalisera til, blir definert på førehand. Det er eit absolutt krav i G-teori at ein slik definisjon blir eksplisitt gjeven. I klassisk teori har generaliseringsuniverset fått liten og ingen plass, og det trass i at sann

skåre har hatt ein mykje framskoten plass.

"Since a given measure may reasonably be generalized to many different universes, the investigator must specify the universe which is of interest to him before he can begin to study generalizability. This consideration is omitted from the classical theory, where it is implied that every test has a true score, belongs to only one family of parallel tests, and has only one "reliability coefficient".(CRG(1963),144)

Ein test vil i G-teori kunna tilhøyra mange familiar av random-parallelle testar. Difor må vi også rekna med at ein test potensielt har mange G-koeffisientar. Denne rasjonale krev difor at vi startar ein G-studie med ein definisjon av det som interesserer oss, og at vi så får tak i to eller fleire uavhengig selekterte observasjonar innanfor dette universet.

Guttman(1953) peika på kor lite dette synspunktet har vore påakta i klassisk teori, og Cronbach seier i denne samanheng:

"The crucial notion in the classical theory is that a test has a true score that enters into all tests "parallel" to it. This concept was severely criticized by Guttman. He referred to the example of the Thurstone fourletter fluency test, for which at least 3 dissimilar parallel tests could be constructed. Each would lead to a different reliability coefficient. In terms of generalizability theory, however, we would rather say that these represent different universes we might generalize to. So it is more appropriate and more reasonable to investigate how well to specify a universe of particular test forms (generally speaking, a universe of conditions)over which we want to generalize." (Cronbach(1966),124)

Vi har tidlegare sett på dei konvensjonelle typar av reliabilitet: ekvivalens, stabilitet og ekvivalens-og-stabilitet. (sjå s.31) På ein måte kan vi sjå denne klassifisering som ein grov skisse av tre generaliseringsunivers. Ekvivalens og stabilitet representerer ikkje-stratifiserte univers,medan ekvivalens-og-stabilitet i grunnen representerer eit stratifisert univers etter som vi har brukt to identifikasjons-attributt. (For di vi ikkje bruker eit "fullbore" design på ekvivalens-og-stabilitet, får vi ein reliabilitetskoeffisient

som blandar saman dei to univers: Vi administrerer form 1 ein dag, form 2 ei veke seinare t.d.. Hadde vi administrert både form 1 og form 2 same dag og form 1 og 2 same dag ei veke seinare t.d., ville designet vore så vidt godt at vi kunne tillata oss å generalisera over dei to univers samstundes.)

Men desse meir klassiske "definisjonar" av universet er så vage at dei ikkje tilfredsstiller dei presise og eksplisitte definisjonar som G-teori krev. Dette er sikkert motiveringa for at Standards for educational and psychological tests and manuals(1966) rår til at vi ikkje lenger skal bruka termane ekvivalens,stabilitet og ekvivalens-og-stabilitet, men i kvart tilfelle så presist og eintydig som råd er definera det spesifikke univers vi har i tankar.

"Designating a coefficient as "an estimate of the expected coefficient over (e.g.) test forms and trials" permits us to dispense with the unwieldy nomenclature of the Technical Recommendations for Psychological Tests(1954) - viz.,"coefficient of stability","coefficient of equivalence",and "coefficient of internal consistency". Any idea that could be expressed by naming coefficients distinctively can be expressed more precisely by designating the universe to which each coefficient refers." (CRG(1963),159)

Vi har tidlegare peika på at Tryons omgrep domenevaliditet er nært i slekt med omgrepet "construct validity". Også G-teori står i nært slektskap til desse to omgrep.

"Since the universe is a construct that he (the investigator) introduces because he thinks it has explanatory or predictive power, an investigation of generalizability is seen to be an investigation of the "construct validity" of the measure. Thus the theory of "reliability" and the theory of "validity" coalesce; the analysis of generalizability indicates how validly one can interpret a measure as representative of a certain set of possible measures." (CRG(1963),157)

Det er viktig å ha det klårt for seg at i G-teori er univers og koeffisient "avhengige" av kvarandre. Frå "klassisk" bruk av reliabilitet kan vi finna døme på, truleg ikkje få, at ein reliabilitetskoeffisient blir gjeven som ikkje gjev oss relevant informasjon om testen i den samanheng han blir brukt.

Kor ofte kan vi ikkje sjå at det blir brukt split-half reliabilitet (internal consistency) når det er så opplagt at vi hadde bruk for test-retest reliabilitet, for di den første som regel er lettare å få tak i enn den andre, og for di vi av ein eller annan grunn har fått det slik for oss at same kva for koeffisient vi bruker, så seier han oss noko om denne tests reliabilitet, reliabiliteten med stor R. G-teori presiserer at ein slik reliabilitet ikkje finst,

"The reinterpretation of "reliability" theory as a theory of generalizability removes many confusions from the application of measurement theory. The semantic problems of interpreting "reliability", "true score", and "error" reduce to mere matters of syntax when we introduce the word "generalizability". To speak of the generalizability of a measure is obviously an incomplete statement until the speaker indicates what construct is being generalized to; he is forced to be explicit about what has often been implicit and therefore lost from sight. The so-called error of measurement becomes a discrepancy between the measurement and the universe score, and the question, "What universe?" follows naturally." (CRG(1963),156)

LITTERATURLISTE

Brown, W. (1910). Some experimental results in the correlation of mental abilities. British Journal of Psychology, 3, 296-322.

Coombs, C. H. (1950). The concepts of reliability and homogeneity. Educational and Psychological Measurement, 10, 43-56.

Coombs, C. H. (1966). Scaling and data theory. Ann Arbor.Stensil.

Cornfield, J. and Tukey, J. W. (1956). Average values in mean squares in factorials. Annals of Mathematical Statistics, 27, 907-949.

Cronbach, L. J. (1947). Test reliability: Its meaning and determination. Psychometrika, 12, 1-16.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika, 16, 297- 334.

Cronbach, L. J. (1966). Mental test theory and decision theory. I Psychological Measurement Theory, proceedings of the NUFFIC international summer session in science at "Het Oude Hof", The Hague, July 14-28, 1966.

Cronbach, L. J. and Meehl, P. E. (1955). Construct validity in psychological tests. Psychological Bulletin, 52, 282-302.

Cronbach, L. J.,Gleser, G. C. and Rajaratnam, N. (1963). Theory of generalizability: A liberalization of reliability theory. British Journal of Statistical Psychology, 16, 137-163.

Cronbach, L. J., Ikeda, H. and Avner, R.A. (1964). Intraclass correlation as an approximation to the coefficient of generalizability. Psychological Reports, 15, 727-736.

Cureton, E.E. (1931). Errors of measurement and correlation. Archives of Psychology, 14, 715-738.

Edwards, A. L. (1959). A note on Tryon's measure of reliability. Psychometrika, 24, 257-260.

Ekman, G. (1947). Reliabilitet och konstans. Uppsala: Almqvist & Wiksell.

Ghiselli, E. E. (1964). Theory of psychological measurement. New York: McGraw-Hill.

Guilford, J. P. (1965). Fundamental statistics in psychology and education. Fourth edition. New York: McGraw-Hill.

Gulliksen, H. (1950). Theory of mental tests. New York: Wiley

Guttman, L. (1945). A basis for analyzing test-retest reliability. Psychometrika, 10, 255-282.

Guttman, L. (1953). A special review of Harold Gulliksen: Theory of mental tests. Psychometrika, 18, 123-130.

Haggard, E. A. (1958). Intraclass correlation and the analysis of variance. New York: Dryden.

Hays, W. L. (1963). Statistics. New York: Holt, Rinehart and Winston.

Hoyt, C. (1941). Test reliability estimated by analysis of variance. Psychometrika, 6, 153-160.

Jackson, R. W. B. (1939). Reliability of mental tests. British Journal of Psychology, 29, 267-287.

Jackson, R. W. B. and Ferguson, G. A. (1941). Studies on the reliability of tests. Bulletin 12, Department of educational research. Toronto: University of Toronto.

Kelley, T. L. (1924). Note on the reliability of a test. Journal of Educational Psychology, 15, 193-204.

Kelley, T. L. (1924). Statistical method. New York: Macmillan.

Kuder, G. F. and Richardson, M. W. (1937). The theory of the estimation of test reliability. Psychometrika, 2, 151-160.

Ljung, B. O. (1960). Intraklasskorrelation - en översikt av metoder och ett program för IBM 650. Rapport från Pedagogisk-psykologiska Institutionen, Lärarhögskolan i Stockholm.

Lord, F. M. (1955). Estimating test reliability. Educational and Psychological Measurement, 15, 325-336.

Lord, F. M. and Novick, M. R. (1966). Statistical theories of mental test scores. Princeton: Educational Testing Service. Stencilert.

Magnusson, D. (1966). Testteori. Andra utgåve. Stockholm: Almqvist & Wiksell.

Novick, M. R. and Lewis, C. (1967). Coefficient alpha and the reliability of composite measurements. Psychometrika, 32,1-13.

Rajaratnam, N. (1960). Reliability formulas for independent decision data when reliability data are matched. Psychometrika, 25, 261-271.

Rajaratnam, N., Cronbach, L. J. and Gleser, G. C. (1965). Generalizability of stratified-parallel tests. Psychometrika, 30, 39-56.

Rulon, P. J. (1939). A simplified procedure for determining the reliability of a test by split halves. Harvard Educational Review, 9, 99-103.

Spearman, C. (1910). Correlation calculated with faulty data. British Journal of Psychology, 3, 271-295.

Standards for educational and psychological tests and manuals. (1966). Prepared by a joint committee of the APA, the AERA, and the NCME. Washington D.C.: American Psychological Association.

Sutcliffe, J. P. (1965). A probability model for errors of classification. 1. General considerations. Psychometrika, 30, 73-96.

Technical recommendations for psychological and diagnostic techniques. (1954). Washington D. C.: American Psychological Association.

Thorndike, R. L. (1951). Reliability. I Lindquist, E. F.(Editor): Educational measurement. Washington D. C.: American Council on Education.

Torgerson, W. S. (1958). Theory and methods of scaling. New York: Wiley.

Tryon, R. C. (1957). Reliability and behavior domain validity: Reformulation and historical critique. Psychological Bulletin, 54, 229-249.

Webster, H. (1960). A generalization of Kuder-Richardson reliability formula 21. Educational and Psychological Measurement, 20, 131-138.

Winer, B. J. (1962). Statistical principles in experimental design. New York: McGraw-Hill.

Yule, G. U. (1922). An introduction to the theory of statistics. London: Griffin.