Introduktion til psykometriske begreber – vejledning vedrørende krav til konstruktion af erhvervstest

Forord

Dette er en vejledning i de psykometriske begreber og principper, der er vigtige i forbindelse med udvikling af psykologiske målemetoder, herunder psykologiske test. Vejledningen er udarbejdet af en arbejdsgruppe under Dansk Psykolog Forening.

Arbejdsgruppen har taget udgangspunkt i den amerikanske *Standards for educational and psychological testing*, som er udgivet i 1999 i et samarbejde mellem *American Educational Research Association*, *American Psychological Association* og *National Council on Measurement in Education*. Denne publikation er meget omfattende og detaljeret, og arbejdsgruppen har ikke haft ambition om at gentage det arbejde, men henviser til det for en mere detaljeret gennemgang af de principper, der gennemgås her.

København, april 2004
Carsten Rosenberg Hansen, Jan Ivanouw, Benny Karpatschof og Erik Lykke Mortensen.

Indledning

Formålet med anvendelse af de fleste psykologiske test er at måle eller vurdere psykiske egenskaber. Når man anvender test, er der en række problemer. De fleste angår følgende tre grundlæggende spørgsmål:

I. *Testens konstruktion*
Det drejer sig om problemer vedrørende testens opbygning, fx udvælgelse af og sammensætning af de spørgsmål eller opgaver (kaldet: items), som testen består af, hvilket svarformat der anvendes, og hvorledes testresultaterne skal gøres op. Disse emner har tæt sammenhæng både med testens reliabilitet og validitet, se nedenfor.

II. *Testens målepræcision*.
Det drejer sig her om problemer vedrørende nøjagtigheden af testresultaterne. Disse problemer vedrører *reliabilitet*.

III. *Testens gyldighed*.
Herunder hører dels spørgsmål vedrørende naturen af de psykiske egenskaber, som testen antages at måle, dels spørgsmål vedrørende i hvor høj grad man kan regne med de tolkninger vedrørende personens egenskaber, der kan opstilles ud fra testresultaterne. Eller sagt på en anden måde: hvor velunderbyggede er disse tolkninger? Disse problemstillinger vedrører *validitet*.

A. Testens konstruktion

Ved test, der består af mere end ét item, stilles en række krav til konstruktionen. Fx krav til sammenhæng mellem besvarelser af de enkelte items og krav til sammenhæng mellem den samlede score og besvarelsen af det enkelte item.

1. Beskrivelse af items
Det teoretiske og empiriske grundlag for valg af itemindhold og svarformat (hvordan testpersonen skal give sit svar) skal beskrives i en testmanual. Hvis psykometriske analyser ikke er foretaget ud fra testens oprindelige svarformat, skal dette begrundes. I så tilfælde skal det desuden klart fremgå, at de gennemførte analyser kun indirekte henviser til de oprindelige testbesvarelser.

Et eksempel kan være, at et femtrins format (eksempelvis: 'Angiv, hvor meget du har været generet af dette på en skala fra 0 til 4') ændres til et totrins format. Under bearbejdningen af data slås 1-4 sammen, så testen nu rescores på den måde, at testpersonen enten har angivet slet ikke at have været påvirket, eller også har angivet at have været påvirket i en eller anden ikke nærmere bestemt grad.

2. Endimensionalitet
Hvis testen sigter mod kun at måle én enkelt psykisk egenskab, skal besvarelse af alle testens items alene afhænge af den pågældende egenskab, fx symptomer på et bestemt område eller et bestemt personlighedstræk. Man bør ved konstruktionen af testen udelukke items, som i høj grad afhænger af andre faktorer end den, man ønsker at måle. Eksempelvis vil et item vedrørende kirkeligt engagement nok være mere afhængigt af, i hvilken grad personen er aktivt religiøs end af det generelle sociale engagement, som testen skulle måle.

I praksis kan det være vanskeligt at godtgøre, at en skala kan betragtes som en enkelt dimension. Det er almindeligt, at man bruger beregninger af *intern konsistens* (se også nedenfor under reliabilitet), og det mest almindelige af disse er Cronbachs alfa, som er et generelt mål for, i hvor høj grad testens items korrelerer indbyrdes. Det er imidlertid i almindelighed ikke nok at bruge dette indeks som dokumentation for endimensionalitet.

3. Ekstern itemhomogenitet
For en testskala, der sigter mod kun at måle én enkelt psykisk egenskab, skal besvarelse af alle testens items alene afhænge af et mål for den pågældende egenskab, fx af totalscoren. Når der tages højde for totalscoren på testen, må besvarelse af de enkelte items derfor ikke afhænge af øvrige egenskaber ved de testede personer, fx køn, alder eller uddannelse (som man kan kalde *eksogene baggrundsvariable).* Hvis besvarelserne er afhængige af sådanne eksterne baggrundsvariable, taler man om *differentiel item funktion* (*DIF*) eller *item-bias.*

4. Tekniske krav til scoring af testen
Der er forskellige måder at score forskellige aspekter ved besvarelse af et enkelt item. Man kan fx score besvarelsens korrekthed, hurtighed ved løsningen, eller man kan score forskellige fejltyper. Tilsvarende opgøres ofte forskellige aspekter ved besvarelser af test bestående af flere items, så der for eksempel beregnes det samlede antal rigtige, den gennemsnitlige løsningshastighed og procentvis fordeling på forskellige fejltyper.

De forskellige opgørelsesmetoder udtrykker forskellige aspekter ved testpræstationen, og de sigter ofte mod at måle forskellige psykiske egenskaber. Den eller de anvendte scoringsmetoder skal være teoretisk velbegrundede og i overensstemmelse med resultaterne af relevante psykometriske analyser. Så vidt muligt skal det dokumenteres,

at de afledte testscores ikke systematisk påvirkes af ikke-relevante psykiske egenskaber, og at scoringen sammenfatter de væsentligste informationer ved testbesvarelsen.

5. Multiskala test
Hvis en test består af flere skalaer, som sigter mod at måle forskellige psykiske egenskaber, bør det dokumenteres, at den empiriske sammenhæng mellem de forskellige skala-scores svarer til den teoretisk forventede sammenhæng.

For skalaer, som indgår i en multiskala test, skal der således kunne demonstreres et relevant mønster af konvergent og divergent validitet (jf. afsnittet om validitet). Man bør endvidere være opmærksom på, i hvilken grad der optræder *metodevarians*, dvs. svarmønstre, der går på tværs af skalaerne og mere afhænger af testens indretning. Metodevarians vil kunne vise sig ved høje korrelationer mellem skalaer, der burde vise noget forskelligt, men hvor for eksempel forsøgspersonernes tendens til at vælge enten de lave eller de høje værdier på en vurderingsskala ved besvarelse af de enkelte items slår igennem i den samlede test.

6. Detaljerede krav og retningslinjer
Mere detaljerede krav og principper for testkonstruktion er beskrevet i *Standards for educational and psychological testing*, afsnit 3.

B. Reliabilitet

Der findes en række metoder til at vurdere *reliabilitet* eller målenøjagtighed. For en given test skal de anvendte metoder være relevante i forhold til testens anvendelse og i forhold til de psykiske egenskaber, som testen sigter mod at måle.

Hvis en test anvendes til at måle *varige* psykiske egenskaber (som personlighed eller intelligens), vil retest-reliabilitet således være væsentlig, mens intern konsistens eller parallel-test reliabilitet vil være vigtige for test, som anvendes til at måle *tilstande* (som aktuel symptombelastning).

I testmanualer skal oplysninger om reliabilitet og målingsfejl beskrives tilstrækkeligt detaljeret til, at det er muligt at vurdere usikkerheden ved det enkelte individs testresultater. Det vil i praksis kræve oplysninger om standardmålefejl (standard error of measurement) for testskalaerne, gerne omregnet til eksempler på sikkerhedsintervaller for forskellige måleresultater. Brugeren skal vide, om et måleresultat på fx 20 skal forstås som liggende i et snævert interval på 19.5 til 20.5, eller om måleusikkerheden er større, og resultatet fx skal forstås som liggende i et interval fra 17 til 23.

1. Intern reliabilitet
Intern reliabilitet er beslægtet med intern konsistens, og begge dele vurderes ofte ud fra *Cronbachs koefficient alfa*. Alt andet lige afhænger målepræcisionen af itemhomogenitet (korrelationen mellem items) og antallet af items. Og eftersom koefficient alfa afspejler begge faktorer, anses den traditionelt for et væsentligt mål for skalaens præcision. En tests interne reliabilitet kan imidlertid også vurderes ved *split half-reliabilitet* eller ved korrelation mellem to forskellige, men parallelle udgaver af samme test (*paralleltestreliabilitet*).

2. Retestreliabilitet

Hvis en test antages at måle varige psykiske egenskaber, skal testresultaterne være uafhængige af situationsfaktorer eller kortvarige tilstandsændringer hos de testede personer. Denne form for målepræcision vurderes ved *retestreliabilitet*, som er korrelationen mellem resultaterne, når de samme personer testes på to tidspunkter. Det ideelle interval mellem de to testninger afhænger af flere faktorer, som fx muligheden for at huske tidligere svar, eventuelle andre retesteffekter og ikke mindst stabiliteten af de målte psykiske egenskaber.

Eksempelvis vil en person ikke ændre personlighedstræk i løbet af en måned, og samtidig kan en måneds tid være nok til, at besvarelserne er glemt.

3. Intertester- eller interraterreliabilitet
Et testresultat kan fx afhænge af testerens testadministrationsprocedurer og scoringskriterier. Sådanne former for fejlkilder vurderes ved *interraterreliabilitet*, ved psykologiske test kaldet *intertesterreliabilitet*. Denne form for reliabilitet beskriver, i hvilken grad testresultatet er uafhængig af den person, der tester. Det siger sig selv, at denne problematik er meget væsentlig, hvis testeren skal foretage ratings af fx symptomer eller personlighedstræk. En række danske erfaringer tyder imidlertid på, at *testereffekt* også kan spille en stor rolle ved forskellige former for kognitive test.

4. Intertestreliabilitet
Mange andre fejlkilder end de ovennævnte kan påvirke testresultater. En vigtig faktor er den anvendte *testmetode*, idet fx måling af et bestemt personlighedstræk principielt ikke bør afhænge af den målemetode, der er anvendt (eksempelvis om der er anvendt en projektiv test eller en selvvurderingstest).

En del forskning tyder imidlertid på, at resultaterne ofte bliver forskellige afhængigt af testmetoden. Der kan være flere årsager til dette. En af årsagerne kan være testindstillinger (response style), der påvirker forskellige testmetoder forskelligt. En testperson har fx mere direkte mulighed for at påvirke resultatet af en selvvurderingstest med høj facevaliditet (se nedenfor) end resultatet af en sværere gennemskuelig projektiv test.

En anden årsag kan være, at forskellige opfattelser af testpersonen slår igennem i en selvvurderingstest, i medarbejdernes vurdering, som den viser sig på en observatørrating, eller i en ægtefælles vurdering på en skala til brug for pårørende. Endelig kan manglende overensstemmelse mellem forskellige test, der skal måle den samme egenskab, skyldes, at de i virkeligheden måler noget lidt forskelligt. For eksempel svarer selvoplevet depression, som den viser sig på en selvvurderingstest, måske ikke helt til det depressionsbegreb, som bruges af professionelle, når de vurderer patienten på en observationsbaseret klinisk skala.

6. Detaljerede krav og retningslinjer
Mere detaljerede krav og principper vedrørende reliabilitet er beskrevet i *Standards for educational and psychological testing*, afsnit 2.

<o:p></o:p><o:p></o:p>**C. Validitet**

En relevant testkonstruktion og en rimelig præcision er blandt forudsætningerne for, at

en test kan måle en psykisk egenskab, fx et symptom eller et personlighedstræk. Spørgsmålet om testudsagnenes validitet er imidlertid en overordnet problemstilling, og de øvrige tekniske krav til en test bør vurderes på denne baggrund.

Validitet er den grundlæggende og afgørende betingelse for testens kvalitet. Validitetsbegrebet omfatter alle de nødvendige betingelser for, om testen fungerer efter hensigten, dvs. at den enten måler de psykiske egenskaber, den sigter mod at måle (begrebsvaliditet), eller at testresultaterne empirisk i det mindste kan demonstreres at have sammenhæng med relevante ydre kriterier (kriterievaliditet). Mere specifikt betegner validitet, i hvilken grad empiri og teori støtter den fortolkning af testresultater, som impliceres af de foreslåede anvendelser af testen.

I mange tilfælde bliver resultaterne fra samme test udgangspunkt for flere forskellige fortolkninger. Det kan være, fordi testen indeholder forskellige subskalaer, som gives hver sin fortolkning, eller fordi testresultaterne giver anledning til flere forskellige tolkninger svarende til forskellige anvendelser. Dette indebærer, at validitet i mange tilfælde ikke kan tilskrives en test som helhed, men at man må beskrive validiteten for de forskellige underskalaer og fortolkningsmuligheder for sig.

Der tales ofte om testvaliditet, som om det er en populationsuafhængig egenskab ved testen. Det er derfor vigtigt at gøre sig klart, at resultaterne af de fleste psykometriske analyser principielt må anses for at være udviklet til anvendelse for en bestemt afgrænset gruppe af personer (en population), og at der derfor altid vil være behov for at gentage psykometriske analyser, når testen anvendes over for nye populationer. Det er ikke mindst vigtigt ved testoversættelser.

Som eksempel på, at testresultater er afhængige af den population, som testen er udviklet i forhold til, kan nævnes, at de korrelationskoefficienter, som man beregner ved undersøgelse af reliabilitet og validitet, afhænger af, hvor stor variation der er i det udvalg af personer (sample), som er undersøgt. Et andet eksempel er, at man ikke på forhånd kan være sikker på, at den indre struktur, man har fundet i testen (herunder grupperingen af de forskellige items i underskalaer), også fungerer for en anden population. Et item vedrørende deltagelse i kirkeligt arbejde kan fungere fint i USA som del af en skala til beskrivelse af testpersonens sociale engagement, mens det sikkert ikke vil fungere på samme måde, hvis testen oversættes til brug i Danmark.

Valideringsprocessen består i en gradvis øgende evidens som baggrund for de foreslåede fortolkninger af testresultaterne. Det første trin i valideringen består i at eksplicitere anvendelsen af testen, og specielt hvilke slutninger man ønsker at drage ud fra testresultaterne.

Nedenfor gennemgås en måde at opdele validitetsbegrebet, som stammer fra Cronbach og Meehl (1955), og som er standard i almindelige tekstfremstillinger på området. Efter en nyere opfattelse, som understreges i *Standards for Educational and Psychological Testing*, er der imidlertid ikke tale om forskellige validitetsformer, men kun om forskellige måder at begrunde validitet på.

1. Facevaliditet
Dette begreb refererer til, om en test umiddelbart virker fornuftig i forhold til sit formål. Traditionelt diskuteres problematikken i relation til de testede personers oplevelse af testen, fordi denne oplevelse kan være vigtig for at sikre, at en testperson er

tilstrækkeligt motiveret for at medvirke til testen, og for at sikre, at testpersonen accepterer de slutninger, der drages ud fra testen.

Også den professionelle testbrugers umiddelbare vurderinger af en test er en vigtigt faktor. Psykologiens historie rummer eksempler på, at facevaliditet spiller en større rolle for testerens valg af test end psykometriske analyser og relevant forskning. Det skal derfor understreges, at det i de fleste tilfælde i praksis er umuligt at vurdere en tests anvendelighed til et bestemt formål, medmindre der foreligger relevante empiriske undersøgelser.

2. Indholdsvaliditet
En psykologisk test skal repræsentere et univers af fx viden, følelser eller adfærd i forskellige situationer. Indholdsvaliditet vedrører spørgsmålet om, hvorvidt de anvendte items udgør et rimeligt repræsentativt udvalg fra dette univers, fx om spørgsmålene til en eksamen er rimeligt repræsentative for det pensum, de studerende skal kunne til den pågældende eksamen. Indholdsvaliditet er også relevant ved personlighedstest. Ved diagnostiske test anvendes for eksempel ofte såkaldte *kritiske items*, som i forhold til bestemte diagnoser anses for afgørende.

3. Kriterievaliditet
Ved kriterievaliditet forstås, i hvilken grad testresultatet stemmer overens med ydre kriterier. Det kan være praktiske ydre kriterier, som når en test for lederegenskaber sammenholdes med arbejdspladsens vurdering af testpersonens succes som leder. Det kan også være ydre kriterier, der ikke har direkte praktisk betydning, men som er med til at indkredse begrebet teoretisk. Som når en testskala, der skal vurdere, hvor selvcentreret en person er, sammenholdes med optællinger af, hvor ofte testpersonen på en videooptagelse af en gruppediskussion anvender ordene ‘jeg’, ‘mig’, ‘min’.

Ved *samtidig kriterievaliditet* sammenholdes testresultater med et på samme tid foreliggende kriterium. I ovennævnte eksempel kunne ledertesten undersøges ved at teste personer, som allerede har vist sig som mere eller mindre velfungerende ledere, og det kriterium, som testresultaterne sammenlignes med, kunne for eksempel være medarbejdernes rating af deres ledere.

Ved prædiktiv eller *fremtidig kriterievaliditet* sammenholdes testresultatet med et relevant kriterium, som foreligger på et senere tidspunkt. Man kunne f.eks. undersøge validiteten af at bruge studentereksamensgennemsnit til at forudsige fremtidige studieforløb ved at sammenligne studentereksamensresultatet med karakterer i det senere studium.

Ved validitetsundersøgelser kan man bruge kontinuerte kriterier, fx resultater fra andre test. Det er også almindeligt at bruge kvalitative kriterier ved at undersøge grupper, som man på forhånd ved er forskellige med hensyn til den pågældende egenskab. Eksempelvis kan en test for grad af selvcentrering undersøges ved at teste en gruppe, der er udvalgt blandt personer, som på forhånd tænkes at være ret selvcentrerede, og en anden gruppe, som kendes som altruistiske og optaget af andre mennesker. Når den sidste type af undersøgelser bruges til støtte for en tests validitet, giver undersøgelser af ekstreme grupper (altså grupper, der er stærkt forskellige på den pågældende egenskab) sjældent tilstrækkelig information til at vurdere testens anvendelighed til undersøgelse af enkeltindivider.

4. Begrebsvaliditet
I praksis kan man ikke forvente, at en test er undersøgt for kriterievaliditet på alle de konkrete områder, hvor man har tænkt sig, at den kan anvendes. En ledertest er måske undersøgt for kriterievaliditet i en bestemt type virksomhed og med en bestemt type ledelsesarbejde. Man kan fortsætte med at undersøge kriterievaliditeten i en anden slags virksomhed og andre slags ledelsesarbejde, men man vil ikke kunne forvente at kunne undersøge, om testen kan bruges i alle typer og undertyper af virksomheder og alle detaljer i ledelsesarbejde.

Ved undersøgelse af begrebsvaliditet går man en omvej, idet man først afklarer et eller flere begreber, som testen skal kunne 'måle', og som er relevante for mange anvendelsesområder. Eksempelvis kan man beskrive ledelsesfunktioner med begreber som 'overblik over sagsgang' og 'evne til at opfatte underordnedes behov'. Disse begreber beskriver nogle (blandt flere) egenskaber, man ønsker at kunne måle med en test. Hvis man kan påvise, at en test har tilstrækkelig validitet til at måle disse egenskaber, vil man derefter kunne anvende den på typer af virksomheder og i lederjob, som ikke er blevet brugt som kriterier i validitetsundersøgelser af den pågældende test, hvis blot man kan definere, hvilket mønster af egenskaber de pågældende lederjob kræver.

Ved kriterievaliditet undersøger man altså den konkrete brug, man har tænkt sig at gøre af testen ved at afprøve den i 'det naturlige miljø'. Ved begrebsvaliditet er opgaven at få testen afprøvet til brug i situationer, man ikke på forhånd kender, men som kan beskrives med de begreber, testen siges at måle. En sådan fremgangsmåde stiller større krav til validiteten, som må undersøges tværgående, hvor mange forskellige kriterier anvendes til at indkredse de begreber, testen skal kunne måle (*konvergent validitet*), og endvidere også de begreber, testen netop ikke skal måle (*divergent validitet*).

Det sidste krav betyder, at en test for depression fx ikke gerne også skal slå ud ved skizofreni, angst og psykopati (hvis den gør det, skal den nok snarere kaldes en 'psykiatrisk patienttest', som kun er i stand til at skelne mellem patienter og ikke-patienter). Arbejdet med begrebsvalidering af test bliver i princippet aldrig færdigt og kan i praksis komme til at ligge tæt på teoretisk arbejde med at definere centrale psykologiske begreber på et bestemt område. Under udviklingen af en test, der skal måle et bestemt begreb, kan man fx komme ud for, at det i løbet af den række undersøgelser, der skal til for at sikre sig begrebsvaliditet, viser sig, at der ikke er tale om et enkelt entydigt begreb, men at det nødvendigvis må opdeles i flere underbegreber. Et eksempel på underopdeling af begreber er opdelingen af angstbegrebet i en state- og en traitversion, som måles med testen State Trait Anxiety Inventory (STAI).

6. Detaljerede krav og retningslinjer
Mere detaljerede krav og principper vedrørende validitet er beskrevet i *Standards for educational and psychological testing*, afsnit 1.

D. Niveauer i testbeskrivelser og analyser

Mange test har en kompleks opbygning, og det er derfor nødvendigt at være opmærksom på, at beskrivelser af en tests psykometriske egenskaber kan referere til flere niveauer: Item-, skala- eller profilniveauet. Dertil kommer, at validitet ikke refererer til egenskaber ved testen, men til en bestemt anvendelse af testen, herunder til

beslutningsprocedurer, som helt eller delvist er baseret på testresultaterne. Det betyder, at man ved beskrivelse af en test eller et testbatteri må operere med flere udsagnsniveauer, der hver for sig kan underkastes en validitets- og reliabilitetsvurdering:

1. Item- eller responsniveau
På det laveste niveau er informationsenheden det enkelte item og en given persons enkelte respons på dette. Allerede på dette niveau kan man tale om reliabilitet, idet der fx er mulighed for at undersøge, om de enkelte spørgsmål besvares på samme måde ved retestning. Man har erfaring for, at besvarelse af enkeltspørgsmål i fx personlighedstest er påvirket af mange former for støj, og det er netop en af grundene til, at man afleder skalascores baseret på mange spørgsmål. I almindelighed regner man med, at flere items betyder større reliabilitet.

Man kan også på itemniveau beskæftige sig med validitet, idet man ved vurdering af både indholds- og begrebsvaliditet fokuserer på egenskaber ved de items, som indgår i en test eller delskala. Ved testkonstruktion ser man af og til, at items udvælges på basis af deres relation til eksterne kriterier, fx egenskaber ved erhvervs- eller patientgrupper. Eftersom sådanne kriterier ofte er flerdimensionale, indebærer denne konstruktionsmåde en risiko for, at de resulterende skalaer ikke bliver endimensionale.

2. Skalaniveau
Når man skal udtrykke resultatet af en test, vil man som regel kombinere informationen fra testpersonens besvarelse af de enkelte items i en samlet testscore (eller hvis testen omfatter flere skalaer, en for hver delskala). Den mest almindelige metode er at bruge summen af de items, der indgår i den pågældende skala. I situationer, hvor de enkelte items er blevet besvaret på en vurderingsskala (fx fra 0 til 4), bruges i nogle tilfælde gennemsnittet af scoringerne i stedet for summen, fx for at kunne sammenligne resultater fra skalaer, der består af forskellige antal items. I princippet er der dog stadig tale om, at man anvender summen af scores fra de enkelte items (som så blot divideres med antal items).

Reliabilitetsangivelser henviser oftest til skalaniveau, idet reliabilitet vedrører præcisionen af den enkelte testscore. Tilsvarende beskrives validitet oftest for tolkninger fra de enkelte skalaer.

Det er væsentligt at være opmærksom på, at både reliabilitet og validitet kan variere meget fra delskala til delskala, og at generelle udsagn om reliabilitet og validitet af multidimensionelle test ofte kan dække over store forskelle og derfor vil være problematiske.

3. Profilniveau
Resultatet af en multidimensionel test opgøres ofte ved en profil, der beskriver individets testscores på de dimensioner, som testen tilstræber at måle. Man ser ofte testresultaterne anskueliggjort ved, at de forskellige skalaer er anbragt ved siden af hinanden på den vandrette akse, mens skalascoren er afbildet langs den lodrette akse ved et punkt i feltet. Når man forbinder punkterne fra skalaerne, får man en figur, der minder om en profil, deraf navnet.

Der er her tale om en kompleks information, der kan være vanskelig at overskue og vurdere. Ofte sammenligner man testscores i en sådan profil, uden at man er

tilstrækkeligt opmærksom på, at det, man undersøger, ikke er testscores, men *differencer* mellem testscores. Sådanne differencer har ofte ringere reliabilitet end hver af de indgående skalaer for sig. Som omtalt kan reliabiliteten også udtrykkes i form af den målefejl (standard error), den pågældende skala er behæftet med. Hvis man kender målefejlene for hver af de to testskalaer, bliver målefejlen på differencen mellem skalaerne en form for *sum*. For ukorrelerede skalaer fremkommer denne sum ved først at kvadrere de to målefejl hver for sig, derefter lægge dem sammen og så tage kvadratroden. Hvis man fx har en målefejl på 5 skalapoints på hver skala, bliver målefejlen på forskellen mellem dem kvadratroden af 25 + 25 = 7.1. Der er altså større usikkerhed på differencen mellem skalaerne end på hver skala for sig.

Anbefaler en testkonstruktør profilfortolkninger, bør der derfor i manualen foreligge oplysninger om reliabilitet og validitet af sådanne tolkninger. Som minimum bør manualen indeholde oplysninger, som gør det muligt at vurdere, hvor store differencer mellem scores på to skalaer skal være, for at det er meningsfyldt at fortolke dem.

I forhold til eksterne kriterier (fx jobperformance) bør profilfortolkninger og beslutninger baseret på profiltolkninger evalueres ved empiriske undersøgelser. Validiteten af den konkrete anvendelse af multidimensionelle test skal således dokumenteres.

4. Testbatteriniveau
I mange sammenhænge træffes beslutninger på basis af resultaterne af et helt testbatteri som ofte suppleres med anden relevant information (fx fra sagsakter eller interview). Da der her generelt vil være tale om en uformel og uformaliserbar beslutningsprocedure hos den, der bedømmer, vil det normalt være meget vanskeligt at analysere komponenterne i beslutningsproceduren og at vurdere reliabilitet og validitet. Dog vil en form for global reliabilitet kunne undersøges ved at sammenligne de anbefalinger, som to bedømmere uafhængigt af hinanden ville komme frem til på baggrund af det samme materiale. Ved denne form for personvurdering er det en fordel, at de enkelte dele af forløbet er velbeskrevne og så ensartede som muligt.

Hvis der findes et relevant ydre kriterium, vil det ofte være muligt at foretage en undersøgelse af kriterievaliditet. Hvis undersøgelsesproceduren er usystematisk og forskellig fra undersøger til undersøger, vil værdien af en sådan validitetsundersøgelse dog være begrænset.

Undersøgelser af reliabilitet eller validitet på dette niveau refererer altså til den overordnede beslutningsprocedure og vil derfor ikke give noget godt billede af de psykometriske egenskaber ved de enkelte test eller skalaer, som indgår i beslutningsgrundlaget.

E. Normer, standarder og referencedata

Når man skal vurdere et foreliggende prøveresultat, har man brug for et sammenligningsgrundlag. Som sammenligningsgrundlag anvendes ofte såkaldte normer eller standarder. Denne betegnelse er uheldig, dels fordi den kan signalere nogle krav, som den testede person bør leve op til, dels fordi betegnelserne ”normer” og ”standarder” antyder, at der er indsamlet data for et repræsentativt udvalg af befolkningen. Denne type dataindsamling er besværlig og kostbar, og den er stort set

kun forsøgt ved standardisering af intelligenstest og pædagogiske prøver.

I stedet er normer og standarder ofte baseret på mindre og lettere tilgængelige udvalg af personer. Når det er tilfældet, bør sammenligningsgrundlaget mere korrekt betegnes som referencedata eller referencematerialer. Betegnelser som normer, standarder og normative vurderinger bliver imidlertid ofte i praksis anvendt om referencematerialer. Grundlaget for normer og standarder bør derfor beskrives udførligt (herunder kriterier for udvælgelse af referencegruppe). Testmanualer bør oplyse om begrænsninger og fejlmuligheder ved anvendelse af de indsamlede referencematerialer. Til de fleste formål vil det være en fordel at kunne vælge mellem forskellige referencematerialer, så den testede kan sammenlignes med en referencegruppe med samme sociale og demografiske baggrund (fx med hensyn til alder, køn, etnisk baggrund og uddannelse). Testmanualer bør advare mod muligheden for systematisk fejltolkning, hvis der anvendes uhensigtsmæssige referencegrupper.

1. Oversatte test
Generelt gælder, at man ikke uden videre kan gå ud fra, at en oversat test vil fungere efter hensigten her i landet. Den sproglige oversættelse af testen er vigtig og bør vurderes ved tilbageoversættelse til originalsproget. Da formålet med testoversættelser som regel er at få en test med de samme psykometriske egenskaber, som testen har på originalsproget, er den rene sproglige ækvivalens imidlertid ikke tilstrækkelig. Det er nødvendigt også at gennemføre empiriske undersøgelser af de psykometriske egenskaber ved den danske version af testen, ganske som ved udvikling af en ny test. Endvidere vil man ikke uden videre kunne anvende udenlandske referencematerialer. Hvis ikke man empirisk kan godtgøre, at disse vil fungere tilfredsstillende i Danmark, må man indsamle danske referencematerialer.

2. Normative og ipsative vurderinger
Der er to måder at foretage vurderinger af en persons testresultater, den normative og den ipsative.

Ved den *normative vurdering* af testresultater sammenlignes personens resultater med en norm eller standard baseret på en referencegruppe. Dette er tilfældet ved de fleste kognitive test og personlighedstest. Testresultatet omdannes ofte til en score, der er lettere at bruge end den oprindelige råscore. Dette sker ved at beskrive, hvorledes testresultatet er placeret i forhold til referencegruppen. Man kan fx angive, hvor stor del af referencegruppen der har et tilsvarende eller lavere testresultat. Dette kaldes percentilen. Man kan også angive, hvor langt fra referencegruppens gennemsnit testresultatet ligger, og udtrykke denne afstand som andele af referencegruppens standardafvigelse. Dette kaldes z-scoren, som har gennemsnit 0 og standardafvigelse 1. Da denne måde at gøre resultatet op på medfører tal med decimaler og negative tal, foretrækker nogle at anvende T-scores (som har gennemsnit 50 og standardafvigelse 10), eller andre omdannelser af z-scoren (STEN-scores, STANINE-scores). En hyppigt anvendt omregning anvendes ved IQ-sores, som har gennemsnit 100 og standardafvigelse 15. På grundlag af sådanne normative scores er det muligt foretage interpersonelle sammenligninger.

Ved den *ipsative vurdering* er der derimod principielt tale om, at man så at sige sammenligner personen med sig selv (jf. profiltolkninger). Man sammenligner ikke testresultatet i forhold til et referencemateriale af andre personer, men sammenligner forskellige træk ved den samme person. Herved forsvinder viden om personens resultat

i forhold til andre mennesker. Et eksempel på denne opgørelsesform er testen Inventory of Interpersonal Problems (IIP). Testpersonen angiver på en skala fra 0 til 4, i hvor høj grad vedkommende har bestemte interpersonelle vanskeligheder. De 64 items er opdelt i 8 forskellige typer af interpersonelle problemer (8 items i hver). Testen kan opgøres normativt ved at udregne gennemsnittet for hver af de 8 skalaer for sig. Testen opgøres traditionelt imidlertid også ipsativt, idet der fra hver af personens 64 items trækkes gennemsnittet for personens svar på alle 64 items, som divideres med standardafvigelsen for de 64 items. Derefter udregnes gennemsnittet igen for hver af de 8 skalaer. Derved forsvinder information om, hvor stærkt belastet personen i det hele taget mener at være af interpersonelle problemer. Tilbage bliver en 'renset' information, der fortæller, hvilke typer interpersonelle problemer der er sværest for den pågældende person.

Anvendt ved en intelligenstest ville den ipsativiserede profil altså ikke vise noget om det generelle intelligensniveau, men derimod noget om stærke og svage sider af den enkelte persons kognitive funktion. Den ipsativiserede profil ville fx ikke vise, at en person havde en høj verbal og en lav matematisk intelligens, men derimod at den pågældende er stærkere i den verbale end i den matematiske del af intelligenstesten.

3. Ipsativt konstruerede test
I forrige afsnit blev beskrevet måder at behandle resultater fra test, hvor personen besvarer hvert item for sig. Særligt inden for erhvervstestområdet anvendes imidlertid ofte ipsativt *konstruerede* test, hvor man stiller testpersonen over for en valgsituation. Personen skal prioritere mellem alternative svarmuligheder. Fx skal personen i en personlighedstest vælge, hvilket af tre udsagn der passer bedst på vedkommende, og hvilket der passer dårligst.

Der vil være tilfælde, hvor testpersonen mener, at alle tre sætninger passer godt, men alligevel skal angive, hvilken der passer bedst, og hvilken dårligst. Ligesom der vil være tilfælde, hvor personen ikke synes, at nogen af sætningerne passer, men alligevel er nødt til at angive, hvilken der trods alt passer bedre, og hvilken der passer dårligere end de andre. Testen er konstrueret således, at de valgmuligheder, der sammenlignes, repræsenterer forskellige underskalaer. For hver delskala udregnes en score ved at optælle, hvor mange gange de items, der tilhører den pågældende delskala, er blevet prioriteret højt, henholdsvis lavt. Denne metode anvendes fx ved undersøgelse af en persons værdier.

Det ligger i testens konstruktion, at den samlede sum af scores for alle delskalaer ligger fast. Dette udelukker således, at testen kan vise, at en person ligger over eller under gennemsnittet på samtlige skalaer. Om dette er en begrænsning, afhænger af begrebsindholdet af testen. At man i princippet kan beregne en persons score i den sidste skala, hvis man kender resultatet fra de andre, er blevet kritiseret af statistikere, idet det medfører, at resultaterne fra delskalaerne ikke er uafhængige af hinanden, som de bør være i mange former for statistiske beregninger.

En vanskelighed ved denne type test er endvidere, at en komplet sammenligning af alle skalaer (gennem de items, der indgår i hver af valgsituationerne) kan blive meget omfattende, hvis der også skal være et rimeligt antal items til hver skala. I praksis anvendes derfor det, der svarer til 'reducerede systemer' i tipning. Nogle skalaer sammenlignes flere gange indbyrdes end andre, og dette kan skævvride resultaterne. En person kan således vanskeligere samtidig score højt på skalaer, som sammenlignes

hyppigt end på skalaer, der ikke sammenlignes så hyppigt i testen (selv om personen gerne ville prioritere sætninger fra to bestemte delskalaer højt, er vedkommende jo nødt til at vælge mellem dem, hvis de optræder i samme valgsituation).

Ligesom ved ipsativt opgjorte test er de ipsativt konstruerede test udviklet til at sammenligne egenskaber ved den samme person, men i praksis anvendes ipsativt konstruerede test også til interpersonelle sammenligninger. Dette har været kritiseret, men nyere data tyder på, at resultater fra korrekt ipsativt konstruerede test ikke adskiller sig væsentligt fra resultater fra konventionelt konstruerede test.

4. Detaljerede krav og retningslinjer
Mere detaljerede krav og principper vedrørende normer og standarder er beskrevet i *Standards for educational and psychological testing*, afsnit 4.

F. Konstruktion af test-skalascores

1. Simpel binær scoring
Den simpleste form for skalascores består i at sammenfatte en række binære items (der alle kun har to svarkategorier) på den måde, at den ene svarkategori tildeles værdien 0 og den anden 1. Scoren bliver altså identisk med antal svar, der har værdien 1. Sådanne skalascores har selvfølgelig kun en klar psykologisk mening, hvis alle items tilhører en og samme psykometriske dimension. Der skal altså være tale om en endimensional skala.

2. Ækvidistant scoring
Når der er mere end to svarmuligheder ved hvert item, bliver det et særligt spørgsmål, om man kan regne med, at afstandene mellem de forskellige svarmuligheder skal forstås som lige store (ækvidistante). Ved en Likert-skala kan svarmulighederne fx være: helt uenig, delvis uenig, neutral, delvis enig, helt enig, men der forekommer også tilfælde, hvor kun skalaens yderpunkter er beskrevet (fx 1 = helt uenig og 7 = helt enig). Svarkategorierne tildeles normalt fortløbende heltalsværdier, og skalascores beregnes som en simpel sum af scores på et antal items. Er der fx 20 items, hver med svarscores 0-4, vil totalscoren kunne variere fra 0 til 80.

Ud over at skalaen skal være endimensional, forudsætter fremgangsmåden, at svarkategorierne er ækvidistante i forhold til den relevante psykologiske egenskab, idet scoringerne fra de enkelte items behandles som en *intervalskala*. Der er ofte blevet stillet spørgsmål ved antagelsen om den lige store afstand mellem svarmulighederne. I dag er der imidlertid statistiske metoder, der kan anvendes til at undersøge antagelsen om ækvidistance i en bestemt test. Antagelsen vil næppe nogen sinde være helt korrekt, men for skalaer bestående af mange items udjævnes støjen, og totalscoren har ofte vist sig at være reliabel.

3. Vægtet scoring
Undertiden har testkonstruktøren den opfattelse, at de forskellige items skal tillægges forskellig vægt, fordi der er nogle items, der er mere informative end andre. Inden sammentællingen af itemscores multipliceres disse derfor med en vægt, der kan være 1 for items med begrænset information og større end 1 for items med en større information.

Medmindre vægtning af items er solidt empirisk funderet, forringer differentieret vægtning ofte skalaers reliabilitet og validitet. Tit har differentieret vægtning ikke vist sig robust ved den krydsvalidering, som er nødvendig, når testen anvendes på nye populationer.

Differentieret vægtning alene på basis af teoretiske overvejelser vil yderst sjældent kunne forsvares.

4. Koblede skalaer
I de tre allerede nævnte tilfælde er det forudsat, at hvert item kun indgår i scoringen af en delskala. Der findes imidlertid psykologiske begreber, som har en række fællestræk samtidig med at være distinkte enheder. Angst og depression er adskilte fænomener, som imidlertid ofte optræder sammen, og som også deler nogle egenskaber, fx optræder koncentrationsvanskeligheder ved begge lidelser.

Derfor har testkonstruktører i nogle tilfælde opbygget test, så nogle items tæller med i flere delskalaer. Det konstruktionsprincip medfører imidlertid, at skalaerne bliver indbyrdes koblede og dermed korrelerede. Scores fra skalaer, der er korrelerede, er altså ikke uafhængige af hinanden, hvilket kan give statistiske problemer.

Eksempelvis kan det være vanskeligt at undersøge, om en behandlingsform mest hjælper på angsten eller på de depressive træk hos deltagerne, hvis man vurderer effekten ved hjælp af en test, der på forhånd har høj korrelation mellem skalaerne for angst og for depression.

G. Hyppigt anvendte psykometriske metoder

Inden for psykometrien anvendes de almindelige statistiske metoder, som kendes fra mange andre sammenhænge. Dertil kommer imidlertid specielle metoder, som især anvendes ved testkonstruktion eller ved undersøgelse af sammenhænge mellem testscores:

1. Traditionelle itemanalyseteknikker:
Helt fra testpsykologiens begyndelse har man interesseret sig for svarfordelinger på de enkelte items. For items med to svarkategorier har man beregnet andelen af korrekte svar eller andelen af ja-svar, medens man for items i Likert-format har beregnet den procentvise andel i de forskellige svarkategorier og desuden fordelingens gennemsnit og varians. Ved traditionel testkonstruktion undersøges, om hvert item diskriminerer tilstrækkeligt mellem testpersonerne, om det i tilstrækkelig grad varierer i sværhedsgrad, og man beregner som tidligere beskrevet, i hvilken grad hvert item korrelerer med testens totalscore (til det formål anvendes ofte Cronbachs koefficient alfa).

De traditionelle itemanalyse-teknikker kan gennemføres på grundlag af relativt få testbesvarelser. De vil derfor ofte være at foretrække ved indledende analyser og ved test, hvor det i praksis ikke kan lade sig gøre at indhente et stort antal besvarelser. Det er tilfældet for mange kliniske test, og det er også tilfældet for fx erhvervstest af typen observatørbaseret rating. Endelig vil mange foretrække traditionel itemanalyse ved test, som tydeligvis ikke er konstrueret med henblik på at opbygge endimensionale skalaer.

2. Formaliserede itemresponseanalyser
Siden halvtredserne har man brugt nye statistiske modeller, der har det til fælles, at de beskriver egenskaber ved items og testpersoner på samme skala.

I traditionel testning får man et tal, der beskriver. i hvor høj grad testpersonen har en bestemt egenskab. Ved disse itemresponseanalyser får man ud over et måleresultat, der karakteriserer den enkelte person, også tal, der beskriver, i hvor høj grad bestemte items 'indeholder' den bestemte egenskab. Ved en færdighedstest beskriver hver testpersons talværdi (personparameter), hvor dygtig vedkommende er med hensyn til den pågældende færdighed, mens et tal kan beskrive hvert items sværhed (itemparameter). Modellen kan indrettes, så et items sværhedsgrad er den samme som dygtighedsgraden hos den person, der har 50 % chance for at besvare det pågældende item rigtigt.

Det kan udtrykkes på en anden måde: For at finde sandsynligheden for rigtig besvarelse, skal vi bruge forholdet mellem dygtighedsgrad og sværhedsgrad (som vi kan kalde for k). Vi har altså brøken k = dygtighedsgrad/sværhedsgrad. Sandsynligheden for en rigtig besvarelse er k/(1+k). Hvis denne kvotient er lig med 1, betyder det, at opgavens sværhedsgrad modsvarer personens dygtighedsgrad, og at sandsynligheden for et positivt udfald netop bliver 50 %. Hvis opgavens sværhedsgrad er lavere end personens dygtighedsgrad, bliver sandsynligheden større end 50 %, og hvis den ligger højere, bliver den mindre end 50 %. Er sværhedsgraden fx det dobbelte af dygtighedsgraden, bliver sandsynligheden for rigtig besvarelse (1/2)/(1+1/2) = 1/3.

På denne måde kan personparametre (dygtighedsgrader) og itemparametre (sværhedsgrader) udtrykkes på samme skala. Denne model har en række fordele. Når man ganske præcist kan beskrive sværhedsgraden af et item, kan man vælge den bedst mulige sammensætning med items, der spreder sig med sværhedsgrader i det område af dygtighedsgrader, som man vil måle. Dette kan forfines til brug for computeradministreret testning (CAT), således at programmet udsætter testpersonen for items, der nogenlunde svarer til vedkommendes færdighedsniveau, således at vedkommende hverken skal bruge tid på for lette eller for svære opgaver.

En anden fordel ved denne model er, at det bliver muligt at beregne måleusikkerhed ved målemetoden ikke kun for skalaen som helhed, men således at usikkerheden varierer i skalaens forløb (som regel er der mindre usikkerhed midt på skalaerne end i yderenderne).

Måleusikkerheden har blandt andet sammenhæng med, hvor meget information det har været muligt at uddrage fra den gruppe personer, der er anvendt ved udviklingen af testen. Der skal dels være et tilstrækkeligt antal personer, dels skal der være en tilstrækkelig repræsentation af dygtighedsgrader for hele det færdighedsområde, der skal dækkes af testen.

Der skal altså både være tilstrækkeligt med personer med lav og med høj dygtighedsgrad. Derimod er det ikke vigtigt, at persongruppen i streng forstand er repræsentativ for den population, som testen skal kunne bruges over for.

Situationen er ikke anderledes ved andre typer af test, eksempelvis personlighedstest. Her bliver personparameteren blot i stedet udtryk for, i hvor høj grad personen har den pågældende egenskab, mens itemparameteren viser, hvor 'meget' af den pågældende egenskab en person skal have for at svare positivt på det pågældende item.

De fleste itemresponsemodeller er baseret på en antagelse om endimensionalitet, dvs. at alle items i en test eller skala måler den samme psykiske egenskab, og derfor estimeres kun en personparameter. Derimod er itemresponsemodeller forskellige med hensyn til antallet af itemparametre, der estimeres. Det er den mest enkle itemresponsemodel, der er introduceret ovenfor. Den kaldes også *enparametermodellen*, eller *Raschmodellen*, opkaldt efter den danske statistiker Rasch, og omfatter altså kun en enkelt karakteristik af det enkelte item, nemlig dets sværhedsgrad.

En mere kompleks model antager, at det også er nødvendigt at beskrive, hvor godt det enkelte item er til at diskriminere mellem dygtige og mindre dygtige personer. Et item, der diskriminerer godt, vil typisk ikke blive besvaret rigtigt af personer med dygtighed under et bestemt niveau, men bliver stort set besvaret rigtigt af alle personer med dygtighed over dette niveau. Ved item, der ikke diskriminerer så godt, vil også en del personer med ringe dygtighed kunne besvare det rigtigt, mens nogle af de dygtigere personer alligevel ikke kan besvare det rigtigt. Der er altså mere glidende overgang, og det pågældende item er ikke placeret så præcist på sværhedsgradskalaen (eller dygtighedsgradskalaen som er det samme).

Ved *toparametermodellen* er der altså to tal, som karakteriserer hvert item, et tal for sværhed og et for diskriminationsevne. I *treparametermodellen* anvendes også et tredje tal, som beskriver, hvor let det er at svare rigtigt på det pågældende item ud fra rent gæt (som eksempelvis ved multiple choice items med 5 svarmuligheder, hvor man vil have 20 % chance for et rigtigt svar ved rent gæt).

Der er forskellige meninger, men nogle teoretikere fastholder, at i stedet for at finde en model (en-, to- eller treparameter), der kan passe til en given test, bør man konstruere test således, at de faktisk passer til Raschmodellen, fordi denne har en række fordele. Den vigtigste er nok, at den medfører, at en person kan karakteriseres entydigt alene ud fra summen af scores fra testens items, mens dette ikke er tilfældet ved de andre modeller. Sumscoren omsættes i Raschmodellen entydigt til en personparameter, men ikke nødvendigvis som en ækvidistant omregning. Det kan fx godt være, at tre yderligere besvarede items giver en større tilvækst i personparameteren for en person, der allerede har 20 rigtige end for en person, der kun har 10 rigtige.

De nævnte tre itemresponsemodeller er oprindeligt udviklet til at beskrive test bestående af items med to svarkategorier. Der er imidlertid senere udviklet en lang række modeller, som fx kan anvendes til at beskrive items i Likert-format eller items med en kontinuert responseskala, og det er i dag muligt at finde pc-programmer, der kan analysere næsten ethvert tænkeligt itemformat.

Det er muligt ud fra sikkerhedsgrænser for personparameteren at få et fingerpeg om testens anvendelighed. Hvis man som testbruger ved, hvor store forskelle man gerne vil kunne måle mellem forskellige personer (eller mellem den samme persons scores på forskellige tidspunkter), og hvor på måleskalaen disse personer nogenlunde vil befinde sig, kan man ud fra opgørelsen af målefejl på de forskellige dele af skalaen (hvilket skal være angivet i manualen) let se, om testen er tilstrækkelig præcis til formålet.

En vanskelighed ved itemresponsemodeller er, at de ofte kræver mere af datamaterialet end den traditionelle analyse. Det vil tydeligt give sig udtryk, hvis enkelte items ikke passer tilstrækkeligt godt til de skalaer, de indgår i. Modellerne kræver endvidere

generelt en højere grad af endimensionalitet af måleskalaerne, hvilket ofte ikke findes ved personlighedstest, som måler bredt definerede personlighedstræk. Disse metoder forudsætter endvidere ret store grupper af personer under testudviklingen.

3. Faktoranalyse
Faktoranalyse bruges til at reducere information, så det bliver lettere at få overblik over den. Hvis man tænker sig scores fra en længere række færdighedsopgaver, vil der i mange tilfælde være korrelationer mellem disse scores. Dette udtrykker blot, at hvis man kan løse den ene opgave, vil man også med en vis sandsynlighed kunne løse den anden. Det interessante er, at de pågældende opgaver måske falder i nogle grupper, således at hvis man kan løse opgave A, har man også let ved opgave B og D, men måske ikke opgave C.

Hvis man derimod har let ved at løse opgave C, har man måske også let ved at løse opgave E og F, men ikke nødvendigvis opgave A. Dette er let forståeligt, hvis man fx forestiller sig, at opgave A, B og D især trækker på verbale færdigheder, mens opgave C, E og F mere er matematiske. Vi kan i denne situation muligvis få et bedre overblik over personens færdigheder ved i stedet for at huske seks testresultater kun at skulle huske to (verbal og matematisk).

Ved faktoranalyse undersøger man korrelationerne mellem en række målinger, fx scores på en række opgaver. Gennem faktoranalysen får man belyst det mønster af sammenhænge, der er i korrelationsmatricen (korrelationerne opstillet i en firkant). Metoden munder ud i et forslag til at samle måleresultaterne i et (færre) antal dimensioner, kaldet faktorer. Den simpleste form for faktoranalyse er *principal component* analyse, der er en ren matematisk reduktion af datamaterialet til faktorer (principal components).

Ved andre former for faktoranalyse underforstås en antagelse om, at målingerne er udtryk for nogle latente (psykologiske) variable, og at det nærmere valg af statistisk metode til at foretage reduktionen kan få betydning for resultatet. Der findes forskellige kriterier for, hvor mange faktorer man opstiller (uddrager), men det er i sidste ende ofte også et teoretisk spørgsmål, der vedrører det psykologiske emneområde.

Jo færre faktorer man nøjes med, desto mere information kasserer man for at fremme overblikket, og omvendt jo flere faktorer man tager med, desto flere detaljer og desto bedre beskrivelse får man af det konkrete datamateriale. Da man imidlertid ikke er interesseret i at beskrive egenskaber ved det aktuelle datamateriale, som er særlige for dette, og som man ikke kan regne med at genfinde ved andre tilsvarende materialer, er der tale om en afvejning, når man vælger antal faktorer.

Når man har valgt den 'faktorløsning', man er tilfreds med, giver faktoranalysen også metoder til at omregne resultater fra de oprindelige målinger (fx færdighedsopgaver) til faktorscores (fx matematikscore). Selv om man har foretaget en rimelig datareduktion, er det imidlertid ikke sikkert, at faktorerne giver klare mål for de psykologiske egenskaber, som man er ude efter. I mange tilfælde kan man forbedre dette ved at ændre på omregningerne fra itemscores til faktorscores, så de enkelte items samler sig bedre om meningsfulde faktorer.

Sådanne omregninger kan man tillade sig, hvis de foretages matematisk konsistent. Denne omregning kaldes 'faktorrotation', fordi den kan ses på en figur, hvor man lader

scores blive liggende i feltet, mens man 'drejer' de akser, der udgøres af faktorerne. Der er forskellige måder at rotere faktorerne. Ved nogle metoder bliver der ingen korrelation mellem faktorerne (ortogonal, eller retvinklet, rotation), mens man ved andre metoder accepterer, at der bliver indbyrdes korrelation mellem faktorerne (oblique, eller skrå rotation). I sidste tilfælde kan man opfatte faktorerne som en form for 'mellembegreber', som faktisk igen kan faktoranalyseres for at opnå nogle overordnede faktorer.

Man kan, som gennemgået, bruge faktoranalyse til at finde dimensioner, der kan beskrive resultaterne fra enkeltitems (itemfaktoranalyse), men man kan også bruge faktoranalyse til at reducere antal dimensioner, fx kan man i en personlighedstest med mange delskalaer forsøge at samle disse til nogle færre overordnede skalaer, jf. oblique skalaer ovenfor.

På trods af at det kan være problematisk at basere analyser på korrelationskoefficenter ved items med få svarkategorier, har faktoranalyse ofte været anvendt til at analysere store puljer af spørgsmål inden for personlighedsområdet. Hensigten har været at inddele den totale pulje af items i grupper af korrelerede items og på denne baggrund at konstruere personlighedsskalaer. Personlighedstest er på denne måde ofte blevet konstrueret på basis af faktoranalyse, men mange vil i dag foretrække at opfatte itemfaktoranalyse som en foreløbig analyse, der kan anvendes til at inddele items i grupper, som potentielt kan danne basis for konstruktion af skalaer. Den endelige skalakonstruktion kan så ske på basis af mere detaljerede psykometriske analyser, herunder anvendelse af item response modeller.

Den traditionelle faktoranalyse er *eksplorativ* og i høj grad baseret på skøn og vurderinger, jf. valg af metode til at uddrage faktorer og (især) rotationsmetoder. I de senere år har man udviklet metoder til såkaldt *konfirmatorisk* faktoranalyse, hvor man statistisk tester faktormodeller, som er opstillet i forvejen, enten på baggrund af tidligere eksplorative faktoranalyser eller direkte ud fra psykologisk teori. Denne form for faktoranalyse spiller i dag en stor rolle ved undersøgelser af teoretisk forventede sammenhænge mellem fx kognitive test eller mellem forskellige personlighedsskalaer. Hidtil har konfirmatorisk faktoranalyse imidlertid kun spillet en mindre rolle ved analyser på itemniveau og ved konstruktion af test eller skalaer.

Eksplorativ faktoranalyse er en hjælp til at få et første overblik over dimensionerne i et stort antal items, som ikke umiddelbart kan grupperes ved indholdsanalyse eller på basis af teoretiske overvejelser. Eksplorativ faktoranalyse er også relevant ved testoversættelser, hvor man for den danske version kan være interesseret i at foretage indledende faktoranalyser på item- og skalaniveau til sammenligning med de oprindelige, som blev foretaget på den originale version. Antages skalaerne i en oversat multidimensionel personlighedstest imidlertid, som det almindeligvis er tilfældet at have en bestemt faktorstruktur, vil konfirmatorisk faktoranalyse kunne anvendes til at teste strukturen i den danske version af testen.

Afslutning

For den læser, der ønsker en yderligere indføring i emnet, henvises som undervejs nævnt til den amerikanske *Standards for educational and psychological testing*, udgivet i 1999. Nedenstående supplerende litteratur kan også anbefales. For definition af

fagterminologi henvises til gængs psykologisk litteratur.

Supplerende litteratur

Embretson, S.S. & Reise, S.P. (2000) *Item Response Theory for Psychologists.* NJ: Erlbaum.
Gregory, R.J. (2004) *Psychological Testing. 4th ed.* Boston: Allyn and Bacon.
Jackson, C. (2002) *Psykologisk testning.* København: Psykologisk Forlag.
Kaplan, R.M. & Saccuzzo, D.P. (2001) *Psychological Testing. 5th ed.* CA: Wadsworth.
Mabon, H. (2004) *Arbetspsykologisk testning – om urvalsmetoder i arbetslivet. 2:a reviderade upplagan.* Psykologiförlaget AB.