S.T.P.
STIFTELSEN
FÖR TILLÄMPAD
PSYKOLOGI

Läsanvisning för STP:s granskningsrapporter

Mars 2002

innehåll

Läsanvisning för STP:s granskningsrapporter

BAKGRUND

STP granskar svenska arbetspsykologiska test efter kvalitetskriterier som sammanställts av British Psychological Society. Kvalitetskriterierna har anpassats och bearbetats för svenska förhållanden av STP i en arbetsgrupp med bred representation av svenska intressenter som företag, myndigheter, testutgivare, utbildningsinstitutioner, fackförbund och JämO.

Granskningarna genomförs enligt rutiner som garanterar oberoende och opartisk värdering. Granskningsrapporterna publiceras och säljs genom STP, vilket gör det möjligt för testanvändare, testköpare och de som testas att lätt få information om testens kvalitet avseende bl.a. tillförlitlighet och prognostisk förmåga.

Målsättningen med granskningsrapporterna är att de skall kunna läsas av en bred grupp av testanvändare utan djupare psykometriska kunskaper. Granskningsrapporterna innehåller dock med nödvändighet en del fackuttryck vars innebörd inte kan förklaras i varje rapport. Vad kvalitetskriterierna står för, och hur de skall värderas, kan heller inte redogöras för i varje rapport. Med denna läsanvisning är förhoppningen att testanvändaren lättare skall kunna förstå kvalitetskriterierna och de använda fackuttrycken.

Det bör noteras en viktig princip för granskningen, nämligen att det är det presenterade materialet för testet som bedöms. Granskarna bedriver inte egen research kring testet eller utgår från allmänna omdömen om det. Det är det material som är tillgängligt för användaren som bedöms.

Olika kvalitetskriterier för olika typer av test?

De kvalitetskriterier som testen granskas efter är i stort sett desamma vilken typ av test det än rör sig om. Oavsett om testet avser att mäta attityder, personlighetsegenskaper, värderingar, intelligens eller programmeringsskicklighet, är det t.ex. önskvärt att det finns belägg för att testet har en tillfredställande reliabilitet och validitet. Det finns dock en mycket grundläggande skiljelinje mellan två typer av test, som också avspeglar sig i något olika kvalitetskriterier. Denna skiljelinje går mellan test som avser att mäta ”maximum performance” (prestationstest av olika slag, t.ex. begåvningstest och kunskapstest) och sådana som avser att mäta ”typical performance” (personlighetstest, attitydtest osv.). Gemensamt för de test som skall mäta ”maximum performance” är att det för varje uppgift finns minst ett svar som är riktigt och minst ett som är fel och att den testade personen skall sträva efter att få så många riktiga svar som möjligt. Vad gäller de test som skall mäta ”typical performance” finns däremot inga riktiga eller felaktiga svar, utan den testade skall här

försöka svara i enlighet med sin personlighet/övertygelse respektive sitt känslomässiga tillstånd. I de fall kvalitetskriterierna skiljer sig beträffande de test som skall mäta ”maximum performance” och dem som skall mäta ”typical performance” kommer detta att påpekas.

Hur läsanvisningen är upplagd

Läsanvisningen följer strukturen och huvudrubrikerna i granskningsrapporterna och kommenterar termer och begrepp i den ordning de förekommer. Eftersom de uttryck som lekmannen antagligen kommer att ha störst problem med framförallt förekommer under rubriken ”Vetenskaplig dokumentation” i granskningsrapporten, kommer detta avsnitt att kommenteras särskilt ingående.

TESTINFORMATION

Det första avsnittet i granskningsrapporten ger allmän information om testet, som namn, förläggare och vilken version som granskades. Här beskrivs vilken typ av test det är fråga om och vilka eventuella villkor som gäller för användning samt vad testet består av i form av olika formulär eller dataversioner.

HISTORIK OCH BESKRIVNING

Här beskrivs testets bakgrund och vilka typer av data som genereras. Bakgrunden består i en historisk beskrivning där utveckling och teoretiska utgångspunkter redovisas. Det är en faktabeskrivning utifrån materialet vars mål är att ge en bild av hur, på vilka grunder samt utifrån vilka behov testet utvecklats. Huvuddelen av detta avsnitt behöver inga närmare förtydliganden. Punkten ”typ av skala” behöver dock kommenteras:

Typ av skala

När det gäller typical performance-test skiljer man mellan *normativa* och *ipsativa skalor* och för maximum performance-test skiljer man mellan *normativa* och *kriterierelaterade* skalor.

Normativa skalor

För det första är det så att ett test som använder en normativ skala mäter *i vilken grad eller utsträckning* den testade besitter en viss egenskap, eller hyser en viss attityd eller vad det nu är som testet försöker mäta. Tanken är alltså att människor kan vara mer eller mindre intelligenta, impulsiva, utåtriktade osv. och att det i princip bör vara möjligt att gradera hur intelligent, impulsiv eller utåtriktad någon är. För det andra är tanken bakom ett test som använder en normativ skala att denna gradering endast kan ske genom att man *ställer den aktuella individen mot en norm* – en grupp av andra individers poäng på testet. Om Kalle får 24 poäng på Stora Musikalitetstestet kan man förstå att Kalle är mycket musikalisk, om man också vet att av 1 000 slumpmässigt utvalda svenska personer är det bara 5% som får 24 poäng eller mer på detta test.

Svarsalternativen som hör till uppgifterna i ett test med normativ skala kan t.ex. vara ”rätt” eller ”fel”, eller ”Ja” eller ”Nej”:

Exempel:
”Hur mycket är 7 X 9 ?”

a) 84
b) 63

Exempel:
”Jag föredrar att sitta hemma med en god bok framför att gå på en stor fest med en massa okända människor”

Ja Nej

Eller så kan svarsalternativen vara utformade enligt en s.k. skattningsskala:

”Jag föredrar att sitta hemma med en god bok framför att gå på en stor fest med en massa okända människor”

Instämmer inte alls 1 2 3 4 5 6 7 Instämmer helt

Testpoängen erhålls genom att antalet rätta svar räknas eller genom att de siffror som ringats in summeras (om skattningsskalor använts). Ju större testpoängen blir desto ”mer” har individen av den aktuella egenskapen, men vad som är en stor och vad som är en liten testpoäng kan endast förstås genom att den relateras till testpoängen från en normgrupp.

Ipsativa skalor

Ipsativa skalor förekommer endast hos test som avser att mäta ”typical performance” och då framför allt personlighetstest. Med ett test som använder en ipsativ skala kan man normalt inte få reda på hur en person med avseende på en enskild egenskap ligger till i förhållande till andra personer. Interindividuella jämförelser är alltså inte möjliga och många psykometriker förhåller sig därför starkt kritiska till denna typ av test. Vanligen mäter ett test med ipsativ skala flera egenskaper, som dessutom har dikotomiserats så att en ”profil” för individen kan tas fram, t.ex.: ”Per är introvert-känslosam-vänlig”, ”Pål är extrovert-rationell-ovänlig.”

Vad beträffar ett test med ipsativ skala svarar man på uppgifterna exempelvis med att välja, från en grupp med påståenden, det som passar bäst respektive det som passar sämst in på en själv. Om testet t.ex. avser att mäta extroversion, envishet och vänlighet, skulle en möjlig fråga kunna se ut så här:

Välj det av följande påståenden som passar bäst och det som passar sämst in på dig själv:

	Passar bäst	Passar sämst
• *Jag föredrar att gå på en stor fest med en massa okända människor framför att sitta hemma med en god bok.*	☐	☐
• *När jag väl bestämt mig för något så ger jag mig inte förrän jag nått mitt mål.*	☐	☐
• *Jag blir sällan irriterad på andra.*	☐	☐

Om den testade väljer att kryssa i ”passar bäst” på det första påståendet ökar hennes extroversionspoäng med +1; väljer hon att kryssa i ”passar sämst” minskar den med –1 och om hon inte kryssar i något av alternativen på första påståendet kvarstår hennes extroversionspoäng oförändrad. Samma princip gäller för envishetspoängen och det andra påståendet samt för vänlighetspoängen och det tredje påståendet. Antag att hela testet består av 20 sådana här uppgifter. Var och en av de tre egenskapspoängerna kan då variera mellan –20 och +20. Men, och detta är viktigt, de tre egenskapspoängerna är *inte oberoende* av varandra, ty om man t.ex. väljer att kryssa i ”passar bäst” på ett av påståendena så har man därigenom valt bort möjligheten att kryssa i ”passar bäst” på de två resterande. Den totala poängen (extroversionspoängen + envishetspoängen + vänlighetspoängen) måste därför alltid bli = 0!

Följden av detta blir att interindividuella jämförelser inte är möjliga vilket inses med hjälp av följande exempel: Antag att Per varken är särskilt extrovert, envis eller vänlig (jämför med andra människor). I själva verket är han extremt oextrovert, oenvis och ovänlig! Han är dock lika envis som han är vänlig och dubbelt så extrovert som han är envis och vänlig. Ett möjligt resultat för Per på det ipsativa testet skulle kunna vara +10 poäng på extroversion, -5 poäng på envishet och –5 poäng på vänlighet. Ponera vidare att Pål är en av de mest extroverta, envisa och vänliga personer som någonsin funnits. Trots att han är extremt envis och vänlig är han ändå dubbelt så extrovert som han är envis och vänlig och han kan därför få precis samma resultat på det ipsativa testet som Per.

Om nu ipsativa skalor omöjliggör interindividuella jämförelser kan man fråga sig varför man överhuvudtaget skall använda sådana skalor. Vad finns det för fördelar med ipsativa skalor egentligen?

Den kanske främsta fördelen är att det är svårare att framställa sig själv i en fördelaktig dager i test som använder sig av ipsativa skalor än i normativa test. Låt oss tänka oss t.ex. en aspirant till den lediga tjänsten som dammsugarförsäljare. Om frågan ovan hade ingått i ett normativt test hade den varit uppdelad i tre frågor och kanske sett ut så här:

	Instämmer inte						Instämmer helt
• *Jag föredrar att gå på en stor fest med en massa okända människor framför att sitta hemma med en god bok.*	*1*	*2*	*3*	*4*	*5*	*6*	*7*
• *När jag väl bestämt mig för något så ger jag mig inte förrän jag nått mitt mål.*	*1*	*2*	*3*	*4*	*5*	*6*	*7*
• *Jag blir sällan irriterad på andra.*	*1*	*2*	*3*	*4*	*5*	*6*	*7*

Det vore nu föga förvånande om aspiranten antar att företaget helst ser en extrovert, envis och vänlig person på den lediga platsen och därför ringar in siffran ”7” för alla tre frågorna, trots att han/hon inte alls är särskilt extrovert, envis eller vänlig. Denna möjlighet – att framställa sig själv i en bättre dager – är kraftigt reducerad på ett test med ipsativ skala. En annan fördel hos test med ipsativa skalor, åtminstone jämfört med normativa test som använder en skattningsskala, är att man inte får snedvridna resultat som en följd av olika personliga svarstendenser (t.ex. att lägga sig långt ut i skalans ändpunkter, att framförallt svara med mittalternativen eller att hålla med om de flesta påståendena oavsett innehåll). Ytterligare en sak som kan vara en fördel med ipsativa test är att man ofta försöker tolka och arbeta med ett *mönster* av hur individen ligger till med avseende på ett flertal egenskaper, snarare än att stirra sig blind på enskilda attribut.

Kriterierelaterade skalor

I ett ”maximum performance-test” är man intresserad av att mäta hur mycket en individ kan prestera i något avseende. Är testet normrelaterat kan man få ett mått på prestationsförmågan genom att relatera individens resultat till hur andra personer lyckas med testet. Exempelvis kan man då säga att ”90% av populationen skriver långsammare på maskin än vad Per kan göra”, eller att ”endast 4% av svenska barn jämnåriga med Pål, 8 år, förstår så många engelska ord som Pål”. Ett alternativt – och i vissa fall mer användbart – sätt att få en uppfattning om prestationsförmågan vore att relatera den till tydliga kriterier för vad personen kan/inte kan göra, så att man exempelvis kan säga: ”Per kan skriva 60 ord i minuten” eller ”Pål förstår 200 engelska ord”. Gör man på detta senare sätt använder man en kriterierelaterad skala. Ett exempel på ett område där man försöker använda en kriterierelaterad skala är skolan där man infört målinriktade betyg. Om eleven klarar ”ditten” får hon betyget ”G”, klarar hon ”ditten” och ”datten” får hon betyget VG och skulle hon klara både ”ditten”, ”datten” och ”dutten” får hon betyget MVG.
Kriterierelaterade skalor förekommer, men är dock sällsynta i samband med arbetspsykologiska test.

ADMINISTRERING

I detta avsnitt anges om speciell utrustning behövs utöver de normala förutsättningarna, dvs. ett tyst, väl upplyst och ventilerat rum med tillräcklig bordsyta och sittplats för både den testade och testledare. Vidare bedöms hur lång tid det tar att testa en individ med testet. Tiden delas in i fem delar: förberedelsetid, bjudningstid, poängberäkning, analys, presentation.

ANVÄNDARDOKUMENTATION

Här sätter sig utvärderaren in i användarens situation och bedömer det material som medföljer testet. Med användardokumentation menas allt material som medföljer testet även andra källor, samt exempelvis böcker som är lätt tillgängliga för den som använder testet.

Hur lättillgängligt materialet är för den normale användaren betygsätts efter 1) det allmänna intrycket av materialet, och 2) hur lätt det är att sätta sig in i materialet.

VETENSKAPLIG DOKUMENTATION

Större delen av granskningsrapporten ägnas åt att genomlysa testets vetenskapliga dokumentation. Den vetenskapliga dokumentationen handlar i sin tur till största delen om testets *reliabilitet* och *validitet*. Detta är inte så märkligt eftersom kvaliteten hos ett test, dvs. om testet är bra eller dåligt, i grund och botten handlar om dess reliabilitet och validitet. För att man skall förstå granskningsrapporterna är det därför av största betydelse att man har en så exakt uppfattning som möjligt om vad reliabilitet och validitet innebär.

Introduktion till reliabilitet och validitet.

Vi börjar med reliabiliteten och tänker oss att vi har tillgång till en väldigt elastisk linjal, en slags gummisnoddslinjal. Denna linjal vill vi använda för att mäta hur lång Per är. Vi ställer Per mot en vägg, placerar gummisnoddslinjalen intill väggen och avläser: ”182.8 cm”. –Jaha, tänker vi, han var inte längre. Vi anser dock att han ser ovanligt lång ut, så för säkerhets skull gör vi ett nytt försök. Denna gång avläser vi: ”204.7 cm”. Upprepade försök visar att mätningarna med denna gummisnoddslinjal ger högst varierande värden, även då en och samma sak mäts. Slutsatsen vi drar är att gummisnoddslinjalen är ett mätinstrument med mycket dålig reliabilitetet. Så i stället plockar vi fram en mycket exklusiv linjal gjord av titan. Vi ställer Per mot väggen och avläser nu: ”194.6 cm”. –Aha, det verkar rimligt, tycker vi. Men för säkerhets skull gör vi ett försök till. Till vår besvikelse visar det sig vid det andra försöket att det avlästa värdet är ”194.5 cm”. Upprepade försök visar att vi fortfarande får lite varierande värden, men för det mesta håller de sig väldigt nära varandra och vi drar slutsatsen att titanlinjalen är ett mätinstrument med mycket god reliabilitet. Detta exempel illustrerar några viktiga saker:

1. Reliabilitet handlar om slumpmässiga fel. Ju mindre slumpmässiga fel, desto bättre reliabilitet; ju större slumpmässiga fel, desto sämre reliabilitet.

2. Trots att de avlästa värdena varierar så antar vi att Per i ”verkligheten” har en viss bestämd längd.

3. Om mätfelen är slumpmässiga, så att de ibland ger ett lite för högt värde och ibland ett lite för lågt värde, borde de ta ut varandra i det långa loppet. (Alltså är ett test som består av många delar bättre än ett test som bara består av några få delar!)

4. Reliabiliteten blir antagligen sämre då man mäter ett psykologiskt attribut än då man mäter ett fysikaliskt.

Titanlinjalen gav mindre slumpmässiga fel än gummisnoddslinjalen och hade alltså en bättre reliabilitet. Ponera nu att vi tar den här utmärkta titanlinjalen och försöker använda den för att mäta Pers vikt. Vi ställer Per mot väggen och avläser ”194.5”, och så drar vi slutsatsen att Per väger 194.5 kilo. Detta vore dock mindre lyckat, för om vi använder titanlinjalen för att mäta vikten har vi ett mätinstrument med mycket dålig *validitet*. En personvåg skulle däremot vara ett mätinstrument med mycket god validitet i det här sammanhanget. Alltså:

1. Validitet handlar om rimligheten i slutsatsen att vi verkligen mäter det vi avser att mäta. Om det är rimligt att tro att vi mäter det vi är ute efter med vårt mätinstrument, kan vi säga att våra mätningar är av god validitet. Om det finns anledning att betvivla att vi mäter det vi är ute efter kan vi säga att våra mätningar är av en sämre validitet.

2. Ett mätinstrument kan egentligen inte *i sig självt* sägas ha god eller dålig validitet. Titanlinjalen ger mätningar av god validitet om vi använder den för att mäta längd, men om vi använder den för att mäta vikt ger den mätningar av usel validitet. På samma sätt förhåller det sig med psykologiska test. Ett test kan ha utmärkt validitet i ett sammanhang men vara helt värdelöst i ett annat. Ett intelligenstest kan exempelvis ha god validitet om det används för att välja ut personer som söker till en akademisk utbildning, men ha låg validitet om det används för att välja ut svenska representanter till OS-laget i femkamp.

3. Validiteten blir antagligen mycket svårare att bestämma då man mäter ett psykologiskt attribut än då man mäter ett fysikaliskt. Huruvida en viss linjal verkligen mäter längd verkar inte särskilt svårt att avgöra. Men att ett visst psykologiskt test verkligen mäter ”ledarskapsförmåga” eller ”intelligens” är förstås mer osäkert och mycket svårare att dra några slutsatser om.

Kan man säga att den ena av de här sakerna, reliabilitet eller validitet, på något sätt är viktigare än den andra? Ja, i en bemärkelse kan man faktiskt påstå att reliabiliteten är viktigare än validiteten. För har jag ingen reliabilitet så kan jag aldrig hävda

att jag har någon god validitet. Jag kan inte hävda: ”Ja, visserligen ger mina mätningar helt slumpmässiga värden, men de mäter i alla fall rätt sak!”; det vore ett ganska meningslöst påstående. Å andra sidan kan man invända att ingen kommer att använda ett instrument med usel reliabilitet och att det därför inte kan göra någon skada. Men om instrumentet har god reliabilitet, men i själva verket mäter någonting helt annat än man tror så kan konsekvenserna bli ödesdigra.

I granskningsrapporterna tas beläggen för testets validitet upp före beläggen för dess reliabilitet. Av pedagogiska skäl vänder vi på ordningen här och börjar med att titta närmare på reliabiliteten.

RELIABILITET

Medan det ofta är mycket svårt att bestämma validiteten i ett test så är reliabiliteten jämförelsevis ganska lätt att mäta. En definition på reliabilitet är: den grad av information (varians) i testet som är systematisk (sann) – och därmed är tolkbar. Vid bestämningen av ett tests reliabilitet bygger de flesta metoderna på korrelationer. En korrelation är ett statistiskt mått på den grad av samband som mäts med en korrelationskoefficient (som ofta betecknas r). Denna korrelationskoefficient kan anta värden mellan –1 och +1. När man beräknar testreliabiliteten är det dock sällsynt att korrelationen blir negativ, och den kommer därför vanligen att få ett värde mellan 0 och +1; ju högre värdet blir desto bättre är reliabiliteten.

Reliabilitet handlar som sagt om mängden slumpmässiga mätfel som påverkar resultatet. Slumpmässiga mätfel kan dock komma från olika källor och beroende på vilken källa som är i fokus skiljer man mellan två typer av reliabilitet, *stabilitet* och *homogenitet*.

Stabilitet

Ett mått på stabiliteten i ett test får man genom att beräkna korrelationen mellan testpoängen vid tillfälle 1 och testpoängen vid tillfälle 2. Det kan gå till så att vi har ett intelligenstest t. ex., som vi använder på, låt oss säga, 200 vuxna personer. För var och en av personerna beräknar vi poängen på testet. Testet är konstruerat så att en hög poäng antas indikera att individen är mycket intelligent medan en låg poäng indikerar att individen är mindre intelligent. Vi antar också att intelligens är en någotsånär stabil egenskap, som inte varierar särskilt mycket från tillfälle till tillfälle. Efter att ha testat de 200 personerna, väntar vi i t.ex. ett halvår, eller åtminstone så lång tid att majoriteten kan förväntas ha glömt de flesta uppgifterna, och sedan testar vi samma personer igen, med samma test.

Nu förväntar vi oss att de personer som hade en hög poäng på testet vid första tillfället, även vid det andra tillfället skall ha en hög poäng. Och vi förväntar oss också att de personer som hade en låg poäng på testet vid det första tillfället, även vid det andra tillfället skall ha en låg poäng. I vilken grad personer som har höga poäng vid tillfälle 1 också har höga poäng vid tillfälle 2, och personer som har låga poäng vid tillfälle 1 också har låga poäng vid tillfälle 2, räknar vi ut med korrelationskoefficienten. Mycket grovt kan man säga att en korrelation på 0.8 är bra och en korrelation på 0.9 är utmärkt.

Den reliabilitet som man får fram när man beräknar testets stabilitet fokuserar på mätfel som har att göra med själva testtillfället eller testsituationen. En individ kanske får lite sämre resultat än hon egentligen skulle ha haft beroende på att hon har huvudvärk just vid testtillfället, en annan kanske får lite bättre resultat än hon egentligen skulle ha haft eftersom hon var ovanligt frisk vid testtillfället! Ytterligare en individ får lite sämre resultat än hon egentligen skulle ha haft därför att det var en fluga som störde henne, och en annan får lite bättre än hon borde ha därför att flugan fick henne att tänka på en sak som gjorde att hon klarade en av uppgifterna! Ju mer sådana här tillfälligheter påverkar resultatet, desto lägre kommer korrelationen mellan mättillfällena att bli, och desto sämre blir alltså reliabiliteten.

Homogenitet

En annan källa till slumpmässiga mätfel är själva uppgifterna (”items”) i testet. Även om alla uppgifter skall mäta samma sak, t.ex. intelligens, kan slumpen göra att en individ klarar vissa frågor men inte andra. När det gäller mätfel som har med själva uppgifterna att göra, är det framförallt homogenitetsmetoden och ett statistiskt mått som kallas Cronbachs alfa som används. Cronbachs alfa är antagligen det mest använda reliabilitetsmåttet (och det inte bara i testsammanhang). Tyvärr är detta också ett mått som blivit mycket missförstått och missbrukat. Det ligger utanför ramarna för denna läsanvisning att ge någon ingående förklaring till hur Cronbachs alfa skall förstås. För detta krävs djupare kunskaper i statistisk analys än vad som kan förutsättas. Ett försök att ge en grov bild av alfa skall dock göras genom att vi går igenom vilka faktorer som påverkar storleken på detta mått. *Generellt ökar alfa a) med antalet items b) med ökade inter-itemkorrelationer och c) med färre dimensioner i testet:*

- Att alfa bör öka med antalet items är inte så konstigt. En av slutsatserna vi drog av vårt exempel med gummisnoddslinjalen var att slumpmässiga mätfel borde ta ut varandra i det långa loppet och att ett test som består av många delar därför borde ha högre reliabilitet än ett test som bara består av några få delar. Så ju fler items vi har desto mer borde de enskilda frågornas slumpfel ta ut varandra och desto högre bör reliabiliteten och därmed alfa bli. Om ett test består av väldigt många frågor behöver man därför inte låta sig imponeras av dess höga alfa.

- Att alfa blir större med ökade inter-itemkorrelationer är heller inte att undra på. Fokus är, som sagt, på mätfel som har med själva uppgifterna att göra. Om items är högt korrelerade med varandra har inte slumpen getts särskilt stort utrymme att påverka hur individerna reagerat på de olika frågorna. Höga inter-itemkorrelationer är i allmänhet bra och något som testkonstruktören strävat efter, men om inter-itemkorrelationerna är *mycket* höga finns det anledning att ifrågasätta testet. Risken är nämligen stor att testet då täcker in ett alldeles för snävt område av vad det egentligen är ute efter. Antag att jag på ett test som skall mäta extroversion bara har frågor av följande typ:

  - *Jag föredrar klart att gå på en stor fest med en massa okända människor framför att sitta hemma med en god bok.*

- *Jag föredrar definitivt att gå på en stor fest med en massa okända människor framför att sitta hemma med en god bok.*

- *Jag föredrar verkligen att gå på en stor fest med en massa okända människor framför att sitta hemma med en god bok.*

Då lär inter-itemkorrelationerna bli mycket höga, men frågan är om testet verkligen mäter extroversion. Det kanske vore riktigare att säga att det mäter ”Festprisseaktighet” eller ”Bokläsarbenägenhet”.

- För att förstå påståendet att ”alfa ökar med färre dimensioner” måste man förstå vad som menas med en ”dimension”, vilket i sin tur egentligen kräver insikt i en statistisk metod som kallas *faktoranalys*. Denna metod är tämligen komplicerad, men eftersom faktoranalys används mycket vid konstruktion och utvärdering av psykologiska test skall vi göra en ”intuitiv faktoranalys”:

  Med faktoranalys reduceras en (stor) uppsättning variabler till en mindre mängd bakomliggande ”faktorer/dimensioner”. Ponera att vi har delat ut ett test bestående av sex items till en stor grupp personer. Item 1 mäter förmågan att förstå synonymer (S), item 2 mäter förmågan att förstå antonymer (A), item 3 mäter förmågan att förstå homonymer (H), item 4 mäter förmågan att lösa aritmetiska problem (Ar), item 5 mäter förmågan att lösa algebraiska problem (Al) och item 6 mäter förmågan att lösa geometriska problem (G).

  Korrelationerna mellan poängen på dessa test blev som följer:

	S	A	H	Ar	Al	G
S		**.89**	**.77**	.24	.19	.11
A			**.81**	.20	.22	.13
H				.14	.09	.18
Ar					**.92**	**.79**
Al						**.80**
G						

Av korrelationsmatrisen att döma hänger de tre första testen ihop på något sätt, liksom de tre sista. En ”intuitiv faktoranalys” säger oss att här finns två bakomliggande faktorer, eller dimensioner: Verbal förmåga och matematisk förmåga. Testet är alltså ”tvådimensionellt”. Hade bara de språkliga frågorna ingått i testet, eller bara de matematiska, så hade det varit endimensionellt. Om testet är endimensionellt (vilket man normalt strävar efter) får man i allmänhet ett högre alfa än om det är flerdimensionellt (t.ex. tvådimensionellt). Ett högt alfa är dock ingen som helst garanti för att testet är endimensionellt. Med många items och/eller höga inter-itemkorrelationer kan man erhålla ett högt Cronbachs alfa även för ett flerdimensionellt test. Den generella regeln är dock att ett lågt

alfa för ett test bör leda till att dimensionaliteten i testet undersöks med faktoranalys.

VALIDITET

Man brukar skilja mellan ”empirisk validitet” och ”begreppsvaliditet”.

Empirisk validitet

Empirisk validitet handlar om att man beräknar korrelationen mellan testpoäng och poäng på något yttre kriterium (som helst är väldefinierat och konkret). Ju högre korrelation, desto bättre empirisk validitet. Empirisk validitet är särskilt viktig då syftet med testet är att predicera (alltså förutsäga) något framtida kriterium. Eftersom syftet med högskoleprovet är att det skall mäta kunskaper och färdigheter som behövs om man skall klara av sina högskolestudier bör den som inte klarar högskoleprovet särskilt bra inte heller klara sina högskolestudier särskilt bra, medan den som klarar högskoleprovet med bravur även bör klara sina högskolestudier med bravur! Med andra ord bör det finnas en positiv korrelation mellan höga poäng på högskoleprovet och senare framgångar på högskolan. Vad beträffar arbetspsykologiska test, som ofta används i samband med urval och rekrytering, är förstås den empiriska validiteten av helt avgörande betydelse för om testet skall kunna sägas vara till någon som helst nytta. Om syftet med testet är att välja ut de mest lämpliga för en viss sysselsättning måste det ju kunna diskriminera mellan mer och mindre lämpliga personer. Här är det dock viktigt att vara medveten om urvalskvotens betydelse. Även ett test som korrelerar svagt (t.ex. r = 0.2) med ett mått på senare yrkesframgång kan vara mycket värdefullt om antalet sökanden är många och antalet platser är få, medan ett test som korrelerar starkt (t.ex. r = 0.5) med senare yrkesframgång inte är till någon större nytta om antalet sökanden är få och antalet lediga plaser många (alla kan ju komma att tas in).

Begreppsvaliditet

Man kan i och för sig tänka sig fall där kriterievaliditeten är den enda typ av validitet som räknas. Från en hög poäng på högskoleprovet drar vi slutsatsen att individen kommer att lyckas bra med sina studier, från en låg poäng däremot drar vi slutsatsen att individen kommer att lyckas mindre bra. Visar sig korrelationen mellan högskoleprovet och studieframgången vara positiv och hög är sådana slutsatser uppenbarligen rimliga, vilket är allt vi behöver veta. I många fall finns dock inget relevant kriterium för testet, t.ex. om det avser mäta intelligens, intolerans mot tvetydighet, självmedvetande eller neuroticism. Då räcker det inte med en simpel korrrelation mellan testpoäng och ett mer eller mindre lämpligt ”kriterium”. Det behövs något mer, nämligen begreppsvaliditet. För att förstå vad begreppsvaliditet är måste man först inse att det inte beträffande någon av de nämnda personlighetsegenskaperna räcker det att bara stipulera en operationell definition av egenskapen i fråga. Intelligens ÄR ju inte betyg, poäng på ett intelligenstest eller den hastighet med vilken man löser ekvationer med två obekanta. Snarare bör man betrakta intelligens som en latent variabel på vilken dessa mått kan sägas vara mer eller mindre lyckade indikatorer. Man måste också inse att ett begrepps innebörd (t.ex. innebörden i begreppet ”intelligens”) beror på dess förhållanden till andra begrepp, eller m.a.o.

en teori. Begreppsvaliditet handlar således om samspelet teori-empiri-test. För att begrepp och test skall få ökad begreppsvaliditet krävs att ”alla bitarna faller på plats”: Teorin ger upphov till hypoteser och hypoteserna testas mot verkligheten (empirin) med ett eller flera test. Om hypoteserna slår in får både teorin och testen ökad trovärdighet; slår hypoteserna inte in kan det vara fel på antingen teorin, testen eller bådadera.

Begreppsvalidering av test som används i arbetspsykologiska sammanhang kan göras på många sätt. Ofta handlar det om att man korrelerar testet med andra test, som antas mäta samma eller liknande egenskaper. Höga korrelationer antas tala för god begreppsvaliditet. Ett annat sätt är att faktoranalysera testet. Om testet t.ex. avser att mäta extroversion och neuroticism och faktoranalysen visar på två faktorer – en där alla extroversionsfrågorna laddar högt och en där alla neuroticismfrågorna laddar högt – då tas detta som indikation på god begreppsvaliditet.

Normer

Som tidigare nämnts får den poäng som en individ erhållit på ett test som använder en normativ skala sin innebörd genom att poängen jämförs med den fördelning av poäng som en normgrupp fått på testet. För den aktuella individen kan då exempelvis sägas att ”80% av normgruppen hade en lägre poäng”. Värdet av dylika påståenden är förstås helt avhängigt normgruppens sammansättning och storlek. Om de normer som fanns att tillgå för högskoleprovet endast byggde på en normgrupp bestående av 10 barn i stället för 1 000 studenter skulle det vara svårt att tolka en viss erhållen poäng.

I många fall kan en mer nyanserad tolkning av testpoängen göras om normer finns uppdelade på kön, nationalitet, yrke etc. och inte bara från en blandad grupp. Om man t.ex. får veta att 90% av alla dammsugarförsäljare är mer påstridiga än vad Pål är, får man onekligen en annan uppfattning om Påls lämplighet som dammsugarförsäljare än om man får veta att endast 5% av den vuxna befolkningen är mer påstridiga än Pål.

SAMMANFATTANDE UTLÅTANDE

Här ges en mer allmän beskrivning av utvärderingen och de betyg som getts kommenteras. Utvärderingen söker besvara frågan om hur väl ett instrumentet lever upp till sina egna anspråk och i vilken grad dokumentationen visar detta.