BEGÅVNINGSTEST VID URVAL AV SÖKANDE TILL HANDELSHÖGSKOLAN I STOCKHOLM

Lennart Sjöberg

Rapport 2012: 2

Reviderad maj 2012

Psykologisk Metod L Sjöberg AB arbetar med utveckling och användning av psykologiska test samt undersökningar av attityder och riskuppfattningar, andra psykologiska utredningar och tillämpad forskning.

Vår affärsidé är att bedriva arbetet i nära anslutning till den aktuella forskningen inom psykologin.

Skrifter utgivna av Psykologisk Metod AB

Sjöberg, L. Bortom Big Five: Konstruktion och validering av ett personlighetstest. Rapport 2008:1.
Sjöberg, L., & Möller, K. (2009). Sociala arbetsfunktioner och personlighet. Rapport 2009:1.
Sjöberg, L. (2009). *UPP*-testet: Kriterierelaterad validitet. Rapport 2009:2.
Sjöberg, L. (2009). *UPP*-testet: Korrektion för skönmålning. Rapport 2009:3.
Sjöberg, L. (2010). *UPP*-testet: Tredje generationens personlighetstest. Rapport 2010: 1.
Sjöberg, L. (2010). *UPP*-testet. Manual, reviderad version, februari 2010 .
Sjöberg, L (2010). *UPP*-testet: Mångfald och jämställdhet gynnas av korrektion för skönmålning. Rapport 2010: 2.
Sjöberg, L. (2010). A third generation personality test. Rapport 2010:3.
Sjöberg, L. (2010). Emotionell intelligens och social förmåga hos ungdomar. Rapport 2010:4.
Sjöberg, L. (2010). Faktorstrukturen hos *UPP*-testet. Rapport 2010:5.
Sjöberg, L. (2010). *UPP*-testet: Användarhandbok, april 2010.
Sjöberg, L. (2010). Teknisk manual, april 2010.
Sjöberg, L. (2010). *UPP*-testet och kundservice: Kriteriestudie. Rapport 2010:6.
Sjöberg, L. (2010). *UPP/Screen*: Ett screeningtest för personlighet och begåvning. Rapport 2010:7.
Sjöberg, L. (2010). Personlighetsdimensioners validitet i arbetslivet: teorier och empiri. Rapport 2010:8
Nilheim, H. (2010). Internetplattformen för *UPP*-testet. Rapport 2010:9.
Sjöberg, L. (2010). Prognos av riksdagsvalet 2010. Rapport 2010: 10.
Sjöberg, L. (2010). Skönmålning på *UPP* hos chefskandidater. Rapport 2010: 11.
Sjöberg, L. (2011). Ökad testvaliditet genom korrektion för skönmålning. Rapport 2011:1.
Sjöberg, L. (2011). *UPP* och *UPP/Screen* i relation till motivation, personlighet och framgång. Rapport 2011:2.
Sjöberg, L. (2012). Skönmålning på ett personlighetstest bland sökande och antagna till bildning. Rapport 2012:1.
Sjöberg, L. (2012). Begåvningstest vid urval av sökande till Handelshögskolan i Stockholm. Rapport 2012:2.

Box 50
182 11 Danderyd

www.psykologisk-metod.se/
info@psykologisk-metod.se

0768 530 885

Innehåll

Förord 1987

Denna rapport är föranledd av att jag blivit ombedd att granska den nuvarande testproceduren vid urval av sökande till Handelshögskolans civilekonomutbildning. Den vänder sig till läsare som kan sakna aktuella kunskaper i testmetodik och jag har därför skrivit relativt utförligt om de grundläggande begreppen, om testningens historik och något om den nuvarande situationen och tänkbara framtida utvecklingen inom området.

Stockholm i februari 1987

Lennart Sjöberg

Förord 2012

Jag skrev denna rapport 1987 på uppdrag av Handelshögskolans dåvarande rektor, Staffan Burenstam-Linder. Den ledde till att jag fick sköta testningarna 1987-91. Rapporten har aldrig publicerats eller spridits på annat sätt än till Högskolans ledning. Att lokalisera den – och kompletterande testdata - krävde tålamod och ibland frustrerande ”elektronisk arkeologi”, men det gick till slut!

Jag har nu redigerat den gamla textfilen så att den kan hanteras i nya datorsystem, lagt till några referenser och snyggat upp graferna. Allt som står i texten håller jag inte med idag, så en del nya fotnoter har varit påkallade. Jag tycker själv att rapporten har ett intresse, som kanske inte enbart är historiskt. Bland annat har ju efterfrågan på begåvningstest ökat kraftigt under det senaste decenniet, efter att länge ha legat i träda på grund av ideologiskt motiverat motstånd. Det finns numera ingen tvekan om att allmänbegåvning, *g*-faktorn eller GMA, är en kraftfull prediktor av arbetsresultat, särskilt inom kvalificerade yrken [31; 32]. Det är heller ingen tvekan om att personlighetsdimensioner ger ett mycket viktigt tillskott [37], liksom emotionell intelligens [7], men det är en annan historia.

Jag konstruerade och tillämpade 1987-91 ett stort antal nya begåvningstest med varierande inriktning. De användes vid antagningarna till HHS. Det var början på min verksamhet som testutvecklare som så småningom har blivit mycket omfattande.

Nerja i maj 2012

Lennart Sjöberg

Sammanfattning

Denna rapport skrevs ursprungligen på beställning av HHS som stod inför att införa ett nya psykologiska begåvningstest för urval av sökande till utbildningen. Den har inte tidigare publicerats. Den innehåller en kortfattad översikt av testområdet samt kommentarer om de test som hade använts fram till 1987. Dessa test studerades närmare empiriskt och visade sig mäta en *g*-faktor, som hade förhållandevis god validitet i relation till studieresultat vid HHS. Förslag till fortsatt utveckling gjordes. Denna version, daterad 2012, skiljer sig från den ursprungliga genom att referenserna uppdaterats och diskussion utvidgats något. Perspektivet på begåvnings- kontra personlighetstest är ett annat: personlighetstest framstår nu som viktigare än de gjorde 1987. Personlighetens betydelse är nu mycket bättre dokumenterad Begåvningstesten, särskilt i form av *g*-test, har fortsatt att visa sig vara mycket användbara som prediktorer av arbets- och utbildningsresultat .Nya utvecklingslinjer vid begåvningstestning har däremot inte riktigt levt upp till de förhoppningar man hade

Inledning

Vid Handelshögskolan används i dag[1] ett system för urval av sökande till civilekonomutbildningen som innebär att en viss del av dem undergår testning. De som lyckas bäst på testet tas sedan in till studierna, tillsammans med dem som kommer in på grundval av höga poäng i andra avseenden (gymnasiebetyg och arbetslivserfarenhet).

Den statliga tillträdesutredningen [41; 42] har föreslagit nya former för urval av sökande till de statliga universiteten och högskolorna. Bl a av det skälet har frågan om Handelshögskolans urvalssystem aktualiserats.

Det nuvarande systemet bygger som sagt till väsentlig del på psykologiska test. Även i framtiden kan det vara aktuellt att utnyttja test i någon form. Frågan uppstår då om man skall fortsätta använda de nuvarande testen eller om man bör söka efter alternativ, eventuellt utveckla egna test. Det är min förhoppning att denna rapport för att kan bidra till beslutsunderlaget i dessa avseenden.

Rapporten ger en historisk exposé av testningens utveckling samt en kortfattad introduktion till de viktigaste begreppen och erfarenheterna inom testteorin. Jag ger därefter en beskrivning av de f n använda testen. Amerikanska metoder för urval till ekonomutbildning beskrivs, liksom forskning angående dessa metoders värde. Rapporten avslutas med vissa rekommendationer angående den framtida användningen av test vid urval bland de sökande till HHS.

Historik

Test har använts för urval i många sammanhang sedan lång tid tillbaka. De användes i stor skala först under första världskriget men liknande procedurer har givetvis tillämpats långt tidigare, t ex i kejsartidens Kina för urval av högre tjänstemän. Det kanske kan ha ett visst intresse att påminna

1 . Alltså 1987.

om att man i Europa under 1800-talet på sina håll framhöll det kinesiska systemet som överlägset det då gängse bördsurvalet i Europa.

De första moderna testen konstruerades vid sekelskiftet av fransmannen Binet. Hans syfte var att få fram test för bedömning av den intellektuella utvecklingsnivån hos skolbarn. Binets test kom att bli grunden för bestämning av intelligensålder, som är en uppskattning av den nivå till vilken ett barn nått i sin utveckling. Ett naturligt steg var givetvis att dividera intelligensålder med kronologisk ålder varvid man erhöll intelligenskvoten, IK.

Binets test tillåter enbart en grov uppskattning av intellektuell nivå. Detta test har emellertid kommit att få en enorm betydelse och i mångt och mycket präglat den allmänna uppfattningen om vad ett test är. På sina håll har man visat en påtaglig övertro på vad IK säger om en människa, på andra håll en lika överdriven skepsis[2]. Självfallet är det ingeting mystiskt med intelligenstest. Det är helt enkelt fråga om hur duktig man är i att lösa vissa sorters problem som inte kräver specialkunskaper. Det är inte märkvärdigare än så. Icke desto mindre kritiseras psykologer ofta för att de skulle tro att man kan göra någon sorts genomlysning av själens innersta skrymslen med hjälp av test. Om några psykologer eventuellt tror det får skylla sig själva. De har fel, lika fel som kritikerna.

Binet-testet bygger på individualtestning. Under första världskriget kunde man givetvis inte individualtesta alla de tusentals personer som det var fråga om att underkasta någon form av prövning. Därför utvecklade man grupptest för bestämning av den allmänna intellektuella nivån. Efter det att man erhållit gynnsamma erfarenheter av grupptestning för urval till officersutbildning började man på 20-talet att utveckla och tillämpa test för industriella tillämpningar.

Den teoretiska synen på test ändrades under 20- och 30-talen, mycket tack vare L. L. Thurstones epokgörande insatser. Man gick alltmera ifrån Binets begrepp allmän intelligens och började i stället arbeta med nyanserade beskrivningar i termer av intelligensfaktorer. Olika forskare har givetvis utvecklat olika system för en sådan multifaktoriell beskrivning av intelligensen. Thurstones elev Guilford publicerade en teori som på 60-talet åtnjöt en viss populäritet [14] och enligt vilken intelligensen består av 120 faktorer[3]. Denna teori har knappast längre en central roll.

Numera går utvecklingen åt tillämpningar av den kognitiva psykologin; se t ex Glaser [10]. Det kanske främsta namnet i denna riktning som återupplivat intelligensbegreppet inom dem kognitiva psykologin är Sternberg [43] 4. Datortekniken erbjuder många spännande möjligheter för utveckling av, i främsta rummet, individualtest, och detta är en utveckling som nu skjuter fart i USA.

Vad är test och egenskaper?

[2]. Se Sjöberg [36] för en diskussion av testdebatten och i den gängse standardargument som i allmänhet är mer eller mindre felaktiga och oinitierade.

[3]. I en senare utveckling har Guilford ökat budet till 150 faktorer (Guilford, 1982).

4 . Sternbergs stora idé att mäta ”praktisk intelligens” har trots omfattande arbete inte kunnat valideras [11; 12].

Med test menas en standardiserad procedur vars syfte är att observera individers beteende och från dessa observationer dra slutsatser om deras egenskaper. Metoden måste vara standardiserad, d v s tillämpas på samma sätt för alla personer, emedan man önskar mäta olikheterna mellan individerna. Om alla individer konfronteras med samma situation kan man anta att skillnader mellan dem beror på just individfaktorer, inte på skillnader i situationspåverkan. I praktiken är ett test av den typ som vi här har anledning att diskutera helt enkelt en uppsättning av problem som skall lösas av den testade. Varje problem är en testuppgift och i normalfallet består ett test av 20-100 sådana uppgifter som skall lösas på 30-60 minuter.

Test används alltså för att mäta egenskaper. Vad är då en egenskap? Vårt språk är fullt av egenskapstermer: man kan påstås vara mer eller mindre begåvad, försiktig, sparsam, kolerisk o s v. Trots den språkliga nyansrikedomen bör man inte låta sig förledas till att tro att alla egenskaper som det finns ord för också existerar. Alla dessa egenskapstermer är snarast att betrakta som antaganden. De egenskaper de denoterar *kan* existera, men det är också möjligt att de inte gör det.

En egenskap existerar om man kan konstatera att det faktiskt finns en tendens för individer att skilja sig åt på ett konsistent sätt i olka situationer. Är man sparsam med sina hushållsutgifter skall man vara sparsam också med rikets medel i sin gärning som ämbetsman, om sparsamhet skall kunna etableras som en egenskap med någon generalitet. Vanligtvis brukar man beräkna korrelationerna mellan olika situationer för att ta reda på om en egenskap föreligger.

Man missbedömer lätt trovärdigheten hos påståenden om egenskaper. Det har visat sig att de flesta egenskapsbegrepp har en mycket begränsad generalitet. Den som fuskar på en skrivning kan mycket väl vara en ärlig deklarant och tvärtom. Bäst har man lyckats visa att intellektuella egenskaper har en viss generalitet. Kanske man kan lägga till några ytterligare egenskaper, som aggressivitet. För en numera klassisk dokumentation av detta påstående, se Mischel [25]. Många har försökt visa att Mischel hade fel och att det skulle vara möjligt att påvisa existensen av många egenskaper förutom de intellektuella men hittills har ingen lyckats[5].

Däremot kan man visa, vilket bl a just Mischel har gjort [26] att principen "more of the same" har en tendens att gälla, d v s om en person lyckas bra i en viss typ av situation tenderar han att lyckas även i framtiden i snarlika situationer. I denna mening är beteendet predicerbart, och det är en princip som kan tillämpas t ex vid chefsurval när man väljer bland personer som redan fått pröva på chefsbefattningar. Principen är emellertid endast tillämpbar när man har data på beteendet i en snarlik situation. Beteendet i breda klasser av situationer, som dem som tillgodoräknas vid meritpoäng för arbetslivserfarenhet i det svenska systemet för urval till högskolestudier, har troligen inget som helst positivt prognosvärde.[6]

Typer av test

5 . Numera är man inte så negativ, vissa personlighetsegenskaper har t ex påvisats vara ganska invarianta och ha ett betydande genomslag i beteendet.

6. Arbetslivserfarenhet har som bekant en svagt negativ relation till studieresultat, även om gymnasiebetyget konstanthålles statistiskt [16]. Problemet med denna meritgrund är bl a att *allting* räknas, och att man enbart utnyttjar ett enkelt kvantitativt mått på erfarenhetens längd utan att beakta hur väl den sökande klarat av situationen ifråga.

Test kan indelas i tre typer: begåvningstest, kunskapstest och personlighetstest. *Begåvningstest* består av uppgifter som utgörs av någon form av problem. Vissa svar är riktiga och andra är felaktiga. Antalet korrekta svar som en person presterar är ett mått på hans eller hennes begåvning i det aktuella avseendet. Det finns många typer av begåvningstest. De varierar m avs på vilken typ av problem som skall lösas av den testade.

Kunskapstest är snarlika begåvningstest men de skiljer sig i ett viktigt avseende: De kräver att den testade personen har tillägnat sig en viss specifik kunskap utöver den allmänna kompetens som kännetecknar alla "normala" medlemmar av en befolkning eller kultur. Givetvis finns gränsfall som kan vara svåra att klassificera entydigt som kunskaps- eller begåvningstest.

Kunskapstest och begåvningstest brukar samvariera ganska starkt. Mera begåvade personer lär sig mera (men i laboratorieförsök har man haft svårt att visa på ett samband mellan inlärningshastighet och begåvning). Konsekvensen av detta är att utbildning troligen *ökar* skillnaden mellan individerna.

Urvalet till högre utbildning bygger i USA till stor del på begåvningstest. I Sovjetunionen avvisade man begåvningstesten på ideologiska grunder i slutet av 30-talet. För urval till högre utbildning använder man sig där av kunskapstest. Det är emellertid troligt att man i stort sett får samma typ av urval som det som åstadkoms av de amerikanska testen. Skolbetyg är även de högt korrelerade med test, av båda de typer vi hittills diskuterat. För HHS räkning bör man dock påminna om att betygsvariationen bland de antagna är ytterst liten och att den knappast kan förväntas ha något större samband med andra variabler. Mer om detta senare.

Personlighetstest är en term som brukar användas för en brokig skara av test som avser att mäta andra egenskaper än de intellektuella. Det finns t ex test som syftar till diagnos av personligheten i psykoanalytiska termer (Rorschach, DMT) medan andra är anpassade till alternativa teoretiska system för mänsklig motivation (TAT, EPPS). Jag skall fatta mig kort om alla dessa test av det skälet att deras prognosvärde i pedagogiska och industriella tillämpningar är försumbart. För aktuell dokumentation av påståendet hänvisas till Schmitt, Gooding, Noe och Kirsch [34], som summerar resultaten från 840 valideringsstudier. Personlighetstest har framför allt ett berättigande inom kliniskt arbete där de kan ge fruktbara uppslag för t ex uppläggningen av psykoterapi7.

Som läsaren säkert inser är detta en ståndpunkt som inte delas av alla psykologer. Vissa personlighetstest, t ex DMT och utvärdering av handstil (grafologi) används i dag på sina håll för chefsurval. Jag kännner dock inte till någon omfattande forskning som kan åberopas som stöd för denna verksamhet. Den som är intresserad av mera information om svagheterna i den industripsykologiska tillämpningen av DMT hänvisas till Sjöberg [35]. För den grafologiskt entusiastiske kan en översikt av empiriska resultat publicerad av Klimoski och Rafaeli [20] vara av värde för att skapa sund skepsis. Att det tycks som om man ibland kan tro att "anything goes" inom området kan läsaren av Svenska Dagbladets I Dag-sida lätt försäkra sig om. Astrologin är just nu mycket aktuell, vill det synas.

7 Senare forskning har gett en helt annan bild: personlighetsvariabler är i det närmaste lika starkt relaterade till kriterier som begåvning är [30].

I detta sammanhang måste jag också säga något om intervjuer och på dem grundade personbedömningar. Det finns en omfattande forskning se t ex Sjöberg & Tollgerdt-Andersson, [39], för vidare diskussion och aktuella referenser) som visat att bedömare starkt överskattar sin förmåga att bedöma människors personlighet. Intervjusituationen har många inbyggda felkällor som sådan, härtill kommer en kognitiv oförmåga att handskas med stora mängder information på ett optimalt sätt. Enkla statistiska modeller är praktiskt taget alltid bättre än mänskliga bedömare för att sammanväga information från, låt oss säga, ett antal test och skolbetyg [13] 8. Problemet för den mänsklige bedömaren är att konsistent väga ihop en stor informationsmängd. Intervjuer har självfallet andra värden men det kan alltså inte påvisas att de har något att ge för att förbättra en prognos, snarare tvärtom.

Hur skall man etablera testens värde?

Det är lätt att bilda sig en ytlig uppfattning om värdet hos en testprocedur. Erfarenheten visar att sådana uppfattningar ofta är vilseledande. Psykologer brukar använda termen "face validity" för att beteckna sådana egenskaper hos testen som förför bedömaren till att tro på att de kan användas för att mäta någon betydelsefull egenskap.

Det är emellertid inte bara testuppgifternas karaktär som kan vara vilseledande. Även informellt samlad kunskap om sambandet mellan testet och andra variabler, t ex studieframgång, kan vara starkt vilseledande. Det finns ett antal faktorer som gör att mänskliga bedömare lätt misstar sig på styrkan av samband mellan variabler. Jag tänker på sådant som att vi minns lyckade prognoser bättre än vår misslyckande, att vi inte har tillgång till information om hur det skulle ha gått för de personer som ej antogs till t ex en utbildning och att positiva förväntningar tenderar att generera självuppfyllande profetior [17]. Det har också visats att människor ofta har en felaktig intuitiv uppfattning om statistiska samband av typ korrelation och att även om man har fullständig information tenderar man att missbedöma korrelationens styrka [40].

Dessa faktorer förklarar att många tror fullt och fast på värdet av test utifrån minnet av sina informella erfarenheter av testet i fråga (dessa brukar med en annan terminologi kallas klinisk erfarenhet). Härtill kommer att personlighetstest oftast bedöms tämligen subjektivt och att olika bedömare ofta kommer till helt olika resultat om en person. (Detta är i själva verket ett skäl till att dessa test har så dålig prognosförmåga). Några viktiga resultat som förklarar tilltro till test trots att de kan sakna validitet är:

1. *Forereffekten.* Utlåtanden på grundval av test som är smickrande och innehåller allmängiltiga påståenden kan upplevas som mycket riktiga, ”på pricken” [8].
2. *Associationer.* Vissa testsvar är starkt associerade med psykologiska begrepp. Exempel: om man i Goodenoughs Draw-a-mantest ritar ett ansikte med stora ögon kan det tyda på paranoia, tror testtolkaren, men ingen sådan validitet finns hos testet [4]. Liknande resultat har erhållits med Rorschachtestet [5].

Alla dessa skäl talar starkt för att man måste etablera testens värde, deras validitet, i skilda praktiska sammanhang genom empiriska studier. I sådana undersökningar insamlas data om "det man vill mäta", alltså i allmänhet skilda mått på framgång (studieframgång, framgång i jobbet) och dessa data, som bildar s k kriterievariabler, relateras till testpoängen. I allmänhet mäter man

8 Undantag från regeln är sällsynta men existerar. Remus och Wong [29] fann t ex att subjektiva bedömningar inte var sämre än statistiska modeller vid urval till en handelshögskoleutbildning.

testens förmåga till prognos av kriterierna med hjälp av korrelationsstatistik. Sambanden brukar sällan överstiga 0.3 i utbildningssammanhang. Detta värde kan emellertid vara vilseledande litet eftersom det med nödvändighet är beräknat på en utvald grupp som har reducerad variation i både test och kriterium. Om man med hjälp av vissa antaganden försöker korrigera för den faktorn kan validiteten ofta uppskattas till ca 0.5. Samma gäller om man bildar relevanta index för att predicera kriteriet [38].

En populär typ av testkritik går ut på att testen är dåliga eftersom de inte korrelerar bättre med kriteriet än vad de gör. Kritikerna förefaller i allmänhet inte känna till att andra metoder nästan alltid är ännu sämre. Sällan ger man uttryck för kunskaper om att även urvalskvoten påverkar resultatet av testanvändningen. (Med urvalskvot avses kvoten mellan antalet antagna och antalet testade). Det praktiska värdet av test beror nämligen dels av sambandet mellan test och kriterium, dels på urvalskvoten och slutligen också på kostnaden för testningen [45].

Testets prognosförmåga m avs på framgång i yrkeslivet, ofta kallad dess validitet, brukar sällan överstiga 0.4, normalt ligger värdet närmare 0.3 (Ghiselli, 1966)[9]. Men även vid en sådan blygsam korrelation mellan test och kriterium kan man visa att testningen är ekonomiskt lönsam om urvalskvoten är tillräckligt liten. I en nyligen publicerad analys av den ekonomiska lönsamheten hos testningar utförda i den amerikanska federala förvaltningen visade Hunter och Hunter [19] att det ekonomiska värdet av de testningar som den federala förvaltningen i USA använder vid beslut om nyanställning uppgår till mycket stora summor. Självfallet är en analys av denna typ diskutabel – den bygger på vissa antaganden som kan framstå som väl starka - men man kan nog sluta sig till, på grundval av Hunters och Hunters studie och andra liknande, att användning av test för urval ofta är en lönsam affär, även om testen har en begränsad validitet.

Vid liten urvalskvot kan man vara ganska säker på att få ett bra urval men samtidigt kommer man att avvisa ett antal kvalificerade sökande. Detta beror givetvis till en god del på att alla kvalificerade sökande inte kan få plats när man bara tar in ett fåtal. Det är också en nödvändig konsekvens av att man inte i förväg kan identifiera de sökande som har den bästa prognosen med absolut säkerhet. Man kan bara göra så gott man kan med hjälp av den information man kan och vill utnyttja. (Det kan givetvis finnas variabler som har prognosvärde men som man av etiska eller praktiska skäl inte vill utnyttja). Det är vanligt att man i diskussioner av urval säger sig vilja basera dessa på subjektiva intryck t ex från intervjuer. All tillgänglig forskning tyder på att detta är en strategi som försämrar urvalet. Detsamma kan sägas om den rika floran av personlighetstest och om grafologi, astrologi, etc.

Intelligenstest och skolbetyg brukar korrelera ganska högt, säg 0.7. De brukar också korrelera med senare studieframgång på ungefär samma nivå. Om de används tillsammans för prognos blir i allmänhet prognosen bättre än om de används var för sig.

Varför test?

Test har onekligen ett värde vid urval, antingen de nu används tillsammans med annan information eller ej. Test av den typ som här avses är objektiva, d v s olika testledare kommer till samma slutsats om en person. De är också ekonomiskt fördelaktiga. Som påvisades ovan är

[9]. Som påpekats ovan ligger värdet ofta något högre när det gäller prognos av studieframgång.

testningar ofta mycket kostnadseffektiva. I HHS-fallet är det givetvis omöjligt att göra en exakt analys av vad test skulle medföra i termer av ekonomiska vinster (men kanske kan man ändå försöka sig på det!) men det finns utifrån de samlade erfarenheterna på området alltså goda skäl att tro på testningens ekonomiska värde.

Test är även rättvisa. Alla som gått i skola vet att betygens grad av rättvisa är, milt talat, diskutabel. I dagarna har den ständigt pyrande betygsdebatten blossat upp på nytt och man har påvisat att olika gymnasier tenderar att ha olika normer för betygssättning. I själva verket är det förhållanden av denna typ som motiverat den omfattande testningsverksamheten i USA när det gäller urval till college-utbildning.

Testen råkade i vårt land i blåsväder i slutet av 60-talet och testverksamheten avtog kraftigt. Man trodde att testen missgynnade vissa grupper, att "fel" egenskaper testades o dyl. Jag har sammanfattat denna debatt på annat ställe [36]. Forskningen har entydigt visat att kritiken var obefogad. Det är visserligen sant att testpoäng ofta samvarierar med socialgruppstillhörighet men det betyder inte att testen är missvisande. För det första är överlappningen mellan socialgrupperna stor, för det andra har ingen bevisat att alla skillnader mellan personer från olika socialgrupper beror på uppväxtmiljöerna, inte heller att att alla sådana skillnader "sitter på ytan" t ex i den meningen att det enbaqrt rör sig om att olika socialgrupper har skilda språkvanor.

För svensk del hade debatten den tråkiga konsekvensen att forsknings- och utvecklingsarbetet inom testområdet avstannade. Många av de test som nu används (verksamheten har tagit fart igen) är 25- 30 år gamla och manualerna hänvisar till undersökningar från 60-talet. Som vi skall se är läget inte bättre för HHS testens del. Den internationella forskningen, som hela tiden varit livlig och som lett till många viktiga insikter, tycks inte fånga många praktikers uppmärksamhet.

Träningseffekter

Det finns i USA och andra länder (Japan brukar nämnas som ett extremfall) en omfattande verksamhet med träning inför urvalstest. Testkonstruktörer brukar vara bekymrade över sådan träning därför att den kan antas vara specifik och enbart påverka testpoängen som sådan, inte den förmåga som testet skall mäta. Den missgynnar också sökande som saknar tillgång till övningsprogram.

Har träning någon effekt? Kulik, Bangert-Drowns och Kulik (1984) gav en sammanfattning av forskning på området. De fann att den genomsnittliga effekten varierade med typ av test. På ett vanligt testbatteri (SAT) var effekten beskedlig, 0.15 standardavvikelser[10], medan den på andra test kunde vara tre gånger så stor och således knappast längre beskedlig. Nyligen har Powers (1986) försökt finna vilka egenskaper hos testuppgifter som påverkar träningseffekterna. Självfallet beror träningseffekterna även av omfattning och typ av träning.

I Sverige kan vi nu vänta oss en testträningsverksamhet växa upp i och med att högskoleprovet skall få tas hur många gånger som helst. Hur är det med HHS nuvarande testuppsättning?

---

[10]. Det är i allmänhet lämpligt att tala om effektstorlekar i termer av standardavvikelser eftersom testpoäng som sådan är en oftast godtycklig variabel, till stor del en funktion av antalet uppgifter i testet.

Professor Ruist (professor i statistik vid HHS) har studerat förekomsten av omtestningar vid HHS och funnit de resultat som framgår av Tabell 1.

Resultat vid omtestning med testbatteriet vid HHS					
År	Totalt antal testade	Omtestade	Antal antagna totalt	Antagna omtestade	Poängökning vid omtestning
1983	247	6	56	1	6.2
1984	244	26	99	13	4.1
1985	237	31	121	18	5.0
1986	233	14	98	9	8.0

Som synes förekommer omtestning ganska ofta och en träningseffekt kan spåras. De omtestade antas i större utsträckning än dem som testas för första gången: 53% antagna över de fyra åren mot 38% av dem som testats första gången. Det är tydligen ingen föraktlig fördel att ha blivit testad tidigare, särskilt inte med tanke på att de som testas en andra eller tredje gång givetvis vid första testningen hade ett mindre gott resultat - annars skulle de ju ha blivit antagna redan då. Vi vet inte om testens validitet sänks på grund av träningseffekt men det är rimligt att anta att så är fallet.

Test vid urval till handelshögskolor

Det finns en omfattande forskning om värdet av test för urval till högre utbildning och olika slags yrkesuppgifter. Rent generellt kan man säga att:

- abilitetstest ger ett visst tillskott för prognos av framgång i högre utbildning, utöver värdet hos den prognos som grundas på skolbetyg.

- det finns samband, om än svagare, mellan test och betyg å ena sidan, yrkesframgång i kvalificerade yrken å den andra [2].

- det tycks som om validiteten hos testen är "trubbig", d v s testens prognosvärde tenderar att vara oberoende av vilken typ av yrkesuppgifter som de används för att prognosticera [33]. Detta något förbluffande resultat kan bero på att den viktiga komponenten i testframgång är psykisk "energi" snarare än abilitet[11].

När det så gäller urval till ekonomutbildning tycks ingen aktuell svensk forskning finnas som varit specifikt inriktad på detta problem. Tillträdesutredningen initierade forskning om samband mellan studieresultat och betyg, resp. arbetslivserfarenhet och högskoleprovet. Utförliga data beträffande sambandet mellan studieresultat och högskoleprovet föreligger såvitt bekant

[11]. Tyvärr har testforskningen i stort sett försummat frågan om motivation i testningen. Denna kan givetvis variera högst avsevärt, jfr t ex den militära inskrivningsprövningen med urvalstestningen för HHS.

ännu ej, men det har påvisats att arbetslivserfarenhet har en svagt negativ korrelation med studieresultat [16] [12].

Från Educational Testing Service i Princeton föreligger ett omfattande material avseende de urvalstest som de administrerar för urval till amerikanska universitets MBA program (GMAT).

Utvecklingen av GMAT beskrivs av Hecht och Schrader (1986). Testet gavs första gången 1954 och innehöll redan då både ett verbalt och ett kvantitativt avsnitt. Hecht och Schrader beskriver hur testet gradvis förändrats under tiden efter 1954. De påpekar att testet fortfarande har samma grundläggande karaktär av blandat verbalt och kvantitativt testbatteri. I dagsläget ingår 6 deltest, som vart och ett tar 30 minuter. Swinton och Powers (1981) har med faktoranalys bl a visat på skillnader mellan de kvantitativa testen, som tycks fungera olika om de var mera abstrakta eller praktiska.

Mycket omfattande valideringsstudier redovisades av Wightman och Leary [46]. De fann, i sin översikt av 123 valideringsstudier utförda under 1980-talet, att GMAT bidrog till en klar förbättring av prognosstyrkan jämfört med enbart collegebetyg (kriterium: första årets studieresultat i MBA- program). Många andra typer av prediktorer prövades, bl a bedömningar baserade på intervjuer samt skilda former av arbetslivserfarenhet, men i intet fall kunde man belägga en tydlig och replikerbar förbättring av prognosstyrkan utöver vad som kunde åstadkommas med hjälp av college-betyg och test. Av speciellt intresse är att den verbala delen av testet tycks fungera minst lika bra som den kvantitativa. De multipla korrelationerna för prognos av studieresultat i MBA-utbildning utifrån test och college-betyg ligger i snitt (median) på 0.41, för testet enbart på 0.33. Resultatet är ungefär vad man kan vänta sig utifrån generella kunskaper om tests prognosvärde.

Prognosvärdet hos GMAT har dokumenterats även av andra än testets administratörer [6; 9; 28]. En ganska överdrivet kritisk diskussion av GMAT ger Benson [3], som i en egen valideringsstudie fann samband mellan test och kriterium på samma nivå, kanske något lägre, som den normala i dessa fall (r = ca 0.3). Självfallet är detta en i absolut mening låg nivå på sambandet. Med risk för att upprepa mig måste jag dock påpeka att resultatet måste relateras till vilka alternativen är och likaså till urvalskvotens storlek. Inget av detta gör Benson.

[12]. Som ett kuriosum kan nämnas att Henrysson och medarbetare rapporterar data på arbetslivserfarenhet och studieframgång i ekonomutbildning vid universiteten i Stockholm och Lund. Resultat: inget samband i Stockholm och märkligt nog ett svagt positivt samband i Lund, som tycks ha berott på speciell kvotgruppssammansättning.

Kommentarer till test använda vid HHS

Inledning

F n13 används sju test vid urvalet till HHS, nämligen

1. Uppskattning
2. Matriser 64
3. Planlösning
4. Slutsatser AROS
5. Talserier PA
6. Ovanliga användningssätt
7. R22

Normalt skall det vid testanvändning föreligga en manual för varje test. Manualen skall ange på vilken teoretisk grund testet bygger och var man kan finna vetenskapliga undersökningar till stöd för denna grund. Vidare skall anges vissa psykometriska basdata såsom upplysningar om testets grad av mätprecision för olika grupper (reliabilitet), dess giltighet för prognos av skilda slag av kriterier och för olika grupper (validitet) samt upplysningar om dess känslighet för övning och om de förhållanden och för vilka grupper det utprövats (standardiserats).

Jag har inte kunnat lokalisera några manualer för de vid HHS använda testen, med undantag för testet Matriser 64, däremot har jag för de 7 testen fått en del upplysningar om vissa av de moment som skall ingå i en manual. På dessa upplysningar bygger nedanstående kommentarer, som börjar med en allmän karaktäristik.

Allmänt

Urvalet av dessa test synes ej ha grundats på en ingående teoretisk eller empirisk analys av vilka test som lämpligen bör ingå vid urval av sökande till en handelshögskola, inte heller på uppföljning av resultat och erfarenheter utifrån HHS:s egna testningar under tidigare perioder[14]. HHS-batteriet har också kommit att avvika på ett delvis anmärkningsvärt sätt från den bästa utländska förebilden, amerikanska GMAT. (Jag är den siste att föreslå att allt i Sverige skall göras efter amerikanska mönster men faktum är att GMAT bygger på en oerhört omfattande forskning medan det vid HHS använda testbatteriet inte bygger på någon forskning alls). Bristen på manualer är besvärande. Det material som jag bygger på här är till synes tämligen föråldrat. Jag övergår nu till att kommentera enskilda test.

13 1987

[14] Under tiden före 1980 utfördes dessa testningar av dåvarande psykotekniska institutet vid Stockholms universitet. När institutet upplöstes och till delar inlemmades i AMS tycks materialet rörande HHS-testningarna ha förkommit.

Kommentarer till enskilda test

Uppskattning. Testet består av 75 uppgifter. Varje uppgift definieras av en aritmetisk operation vars resultat alltså skall "uppskattas", d v s den testade uppmanas att gissa approximativt vad svaret kan vara och får inte använda papper och penna samt avråds från att försöka beräkna det exakt riktiga svaret. Varje uppgift har tre svarsalternativ och den testade skall pricka för ett av dessa.

Testet är konstruerat vid dåvarande PA-rådet och uppgiftsanalys av har utförts på en grupp "UMS-tekniker". Reliabiliteten har beräknats på ett för snabbhetsprov felvisande sätt, jag kan inte bedöma den på tillgänglig information[15].

Vad beträffar testets giltighet anges endast att det har ett visst prognosvärde vid utbildning av postkassörsaspiranter. Testet tycks, föga förvånande, samvariera rätt kraftigt med andra test på aritmetisk förmåga.

Omdöme[16].- Data om reliabilitet är otillräckliga för slutsats om mätprecision. Validitetsdata föreligger ej för kvalificerat arbete. Teorianknytning saknas helt.

Talserier PA. Detta test, också från PA-rådet, är en svensk version av ett traditionellt test på s k induktiv begåvning, av Thurstone på 30-talet förslaget som en grundläggande begåvningsfaktor. Testet består som namnet säger av talserier som är uppbyggda av upp till 8 en- eller två-siffriga tal. Den testade skall försöka finna en regel med vars hjälp han kan fullfölja serien, vilket sker genom att man väljer ett av 7 svarsalternativ för varje uppgift. 29 uppgifter ingår. Den svenska versionen standardiserades ursprungligen på elever vid handelshögskolor, men några uppgifter från dessa standardiseringar eller om när de utfördes har ej varit tillgängliga. Reliabiliteten tycks vara acceptabel, dock saknas uppgift om grad av snabbhetsladdning.

266 "abiturienter" i Borås 1967 testades med detta och andra test. Dessa andra test beskrivs inte närmare i den stencil jag fått varför resultatet av en faktoranalys (som inte heller beskrivs närmare, det sägs t ex inte om man använt oberoende eller beorende faktorer eller hur antalet faktorer fastställts) inte är speciellt upplysande. Vissa skillnader förelåg mellan reallinje, latinlinje och allmän linje. Programmerare följdes under tiden 1960-64. De som klarade testet bättre bedömdes som mera framgångsrika i jobbert, i genomsnitt. Korrelationskoefficient rapporteras emellertid ej.

Omdöme. Ett välkänt test som mäter den induktiva begåvningsfaktorn. Troligen acceptabel reliabilitet och troligen av visst prognosvärde för HHS del. Testet är tydligen konstruerat på 50-talet.

---

[15] Man brukar skilja mellan snabbhetstest (speed tests) och "power tests" (svensk term saknas). I ett utpräglat snabbhetstest klarar alla de testade alla de uppgifter de hinner fram till. Om man uppskattar reliabiliteten t ex med hjälp av korrelationen mellan antal rätt lösta uppgifter med jämna ordningstal och dem med udda ordningstal får man ett missvisande högt värde. Det krävs i stället att man ger testet två gånger till samma personer med två jämförbara versioner, s.k. parallella test, för att man skall få en rättvisande uppfattning om testets mätprecision. Man kan uppskatta hur väl ett test approximerar ett power-test men sådana uppskattningar redovisas ej i materialet. Denna kritik träffar de flesta av de granskade testen.

[16] Dessa omdömen är preliminära och grundar sig på den magra information jag haft tillgänglig, en nypa salt är påkallad.

Matriser 64. Detta är en svensk version av ett bekant test från 1938 som brukar kallas Ravens matriser och som betraktas som ett användbart mått på allmänbegåvningen, särskilt i den övre delen av begåvningsfördelningen. Testet standardiserades omkring 1963 av PA-rådet på skilda studerandegrupper.

Testuppgifterna kan sägas kräva en kombination av rumslig och induktiv förmåga (se talserietestet för ett renare mått på induktiv förmåga). Givet ett visst mönster samt en borttagen del skall den testade lokalisera den borttagna delen bland flera förslag. Rätt svar erhåller man genom att iaktta vissa invarianser i det givna mönstret men också vissa variationsregler. I testet Matriser 64 ingår 48 uppgifter och det torde vara snabbhetsladdat.

Reliabiliteten förefaller av manualen att döma att vara god. Validiteten har inte undersökts annat än för "skolbetyg". De standardiseringsgrupper som man arbetade med på PA-rådet förefaller mindre kvalificerade än HHS-gruppen.

Omdöme. Ett välkänt test som i den svenska versionen nog har sitt värde men som möjligen kunde ha anpassats bättre till den aktuella populationen.

Planlösning. Detta test har konstruerats vid dåvarande psykotekniska institutet i Stockholm. Det gäller i testet att placera vissa givna ytor i en ram så att det villkoret uppfylls att vissa av ytorna har minst en kvadratsida gemensam. Uppgifterna kräver att man kan handskas väl med spatiala relationer, det är ett utpräglat spatialt test som kan bli hur svårt som helst.

Ytterst sporadiska informationer tyder på acceptabel reliabilitet samt någon grad av validitet för programmerare (troligen samma grupp från 60-talets början som tidigare refererats). Märkligt nog är ingen korrelationskoefficient för mätning av validiteten angiven.

OmdömeFel! Bokmärket är inte definierat.. Detta är troligen ett användbart test för mätning av spatial begåvning som i sin tur är en av Thurstones begåvningsfaktorer. Inga data anförs som skulle tyda på att det är direkt av betydelse för HHS-studerande med denna typ av begåvning.

Slusatser AROS. Detta är en svensk version av ett amerikanskt test, översatt och försvenskat vid ASEA 1959. I testet ingår 67 uppgifter. Testet innehåller uppgifter som består av att man skall ta ställning till om slutsatser i syllogismer är korrekta.

Reliabiliteten är troligen acceptabel. Vad gäller validiteten anges endast en viss korrelation för konstruktörer vid ASEA.

Omdöme. Testets värde för syftet urval av sökande till handelshögskolan är ej etablerat.

R 22. Detta är ett allmänbegåvningstest som innehåller en blandning av uppgifter som mäter ordkunskap, förmåga att upptäcka logiska principer, induktiv förmåga, mm. Provet avser att differentiera på högre begåvningsnivåer. Det innehåller 52 uppgifter av flervalstyp.

Testets reliabilitet tycks vara hög, återigen dock med reservation för att man ej tagit fram data på parallelltest. (Rätt stark snabbhetsladdning tycks föreligga). När det gäller validitet anges vissa data från SJ-anställda på 60-talet, bl a resebyråföreståndare. Inga närmare detaljer om hur

kriteriet konstruerats. "Normer" för handelshögskolestuderande (N=53). "Acceptabel" validitet för urval av programmerare (det anges ej vad som menas med detta). Höga korrelationer med bl a matristestet.

Omdöme. Detta test är förmodligen användbart även om man i stort sett saknar data till stöd för det antagandet. Det innehåller vissa verbala uppgifter vilket gör att det framstår som ett tillskott utöver övriga test, icke desto mindre är det i tämligen hög grad överlappande med dem.

Ovanliga användningssätt. Detta test instruerar den testade att finna på upp till 6 alternativa användningssätt av vardagliga föremål. Sex uppgifter ges. Testet bygger på Guilfords teorier om "divergent tänkande". Guilford föreslog på 60-talet att den här typen av test skulle mäta kreativitet men detta antagande har knappast verifierats av senare forskning [21] 17. Det aktuella testet är en översättning av ett av Guilfords test, standardiserat på "systemaspiranter" och "administrativa assistenter". Korrelationer med intelligenstest och betyg förefaller vara låga, varför testet ur den synpunkten har en chans att bidra med ett unikt tillskott till övriga test. Reliabiliteten är troligen acceptabel. Inga data ges dock på bedömarreliabiliteten, vilket hade varit önskvärt med tanke på att testets utvärdering har ett subjektivt moment.

Omdöme. Ett test av ytterst tveksamt värde i sammanhanget, eftersom relationen till kreativitet bygger på teoretisk spekulation och aldrig övetygande dokumenterats – men se not 17.

Interkorrelationer mellan testen

Interkorrelationerna mellan testen för testade 1982 har varit tillgängliga. De tyder på att testens inbördes samband styrs av tre faktorer: spatial, induktiv och aritmetisk förmåga. Härtill kommer testet Ovanliga användningssätt som, vilket var väntat, samvarierar med övriga test endast i ringa utsträckning, se Tabell 2. En *g*-faktor framgår tydligt.

Tabell 2. Korrelationer mellan de använda testen.

	Upp-skattning	Matriser	Planslös-ning	Slutsatser	Talserier	Ovanliga använd-ningssätt	R22
Upp-skattning	1.00	0.22	0.24	0.26	0.48	0.27	0.41
Matriser	0.22	1.00	0.40	0.24	0.46	0.14	0.48
Plan-lösning	0.24	0.40	1.00	0.25	0.44	0.06	0.50
Slutsatser	0.26	0.24	0.25	1.00	0.23	0.18	0.42
Talserier	0.48	0.46	0.44	0.23	1.00	0.13	0.52
Ovanliga använd-ningssätt	0.27	0.14	0.06	0.18	0.13	1.00	0.10
R22	0.41	0.48	0.50	0.42	0.52	0.10	1.00

17 Guilfords idéer ligger till grund för de välbekanta och om debatterade Torrance-testen, som i nutida diskussioner får ett mycket positivt omdöme [1].

Även det första testet, uppskattning, har förhållandevis låga samband med övriga test. Nu behöver givetvis inte låga samband med övriga test vara en nackdel, det kan tvärtom tyda på att testet bidrar med något unikt och därför med ett större tillskott till prognosen än övriga test.

Valideringsstudie

Data har förelegat som gör det möjligt att studera sambandet mellan de vid HHS använda testen och studieframgång för inskrivna 1982. Data avser antalet tenterade studieenheter under 3 läsår, samt uppgift om avbruten utbildning.

Vid tolkning av validitetssiffror bör givetvis beaktas att flera faktorer utom testets värde påverkar storleken på korrelationen mellan test och kriterium. Baird [2] pekar särskilt på spridningen (ju mera reducerad variation i test och kriterium, desto mindre korrelation) samt heterogeniteten i utbildningsprogrammet (ju större heterogenitet, desto mindre korrelation).

Bland de frågor som är viktiga att studera är frågan om könsskillnader. Maccoby och Jacklin [24] ger en utförlig genomgång av de könsskillnader som kunnat påvisas i bl a psykologiska test. Dessa skillnader är mindre än vad man i allmänhet tror, men när det gäller verbala test är kvinnor i genomsnitt bättre än män, som i sin tur excellerar i spatiala och matematiska test. Eftersom HHS-testen är så utpräglat kvantitativa och spatiala i sin uppläggning finns det anledning att tro att kvinnliga sökande missgynnats av dem.

Data föreligger från totalt 92 studerande som togs in HT 1982 efter goda resultat på testningen. Av dessa var 29 kvinnor. Gymnasiebetyg har rapporterats för 73 av dem.

Det är möjligt att bearbeta två, som det kommer att visa sig tämligen olika kriterier: studieavbrott och antal avklarade studieenheter. De som registrerat att de upphört med studierna eller som trots att sådan registrering ej förelåg ej redovisade en enda avklarad studieenhet på 4 år räknades som avbrottsfall. Övriga räknades som aktiva studerande.[18]

[18] Man kunde givetvis ha satt gränsen för att räkna någon som avbrott till något värde >0 studieenheter. Det är inte troligt att resultaten skulle ha ändrats radikalt av ett sådant beslut. Jag har kontrollräknat en del av analyserna med gränsen satt till 20 studieenheter snarare än 0 men inga väsentliga skillnader kunde noteras.

I det följande har jag i huvudsak använt mig av vanlig produktmomentkorrelation, men även kontrollerat resultaten med hjälp av rangkorrelation som torde vara mera robust mot uteliggare. Inga viktiga skillnader uppkom mellan dessa två metoder varför jag i det följande håller mig till produktmomentkorrelationen. En speciell aspekt är den könsskillnad som kan noteras, mera om den nedan.

Korrelationerna mellan kriterierna antal avklarade studieenheter efter 4 år och avbrott samt de olika testen och deras genomsnitt ges i Tabell 3. Dessa värden är främst intressanta om vi ser till den totala predicerbarheten av kriterierna och till skillnaderna mellan män och kvinnor. Antal avklarade studieenheter prediceras på s a s normal nivå i detta material (om kring 0.4) medan avbrott knappast kan prediceras alls. För kvinnor är emellertid testens värde mycket tveksamt. Det finns en intresseväckande tendens som tyder på att de mera "testbegåvade" kvinnorna avbryter medan avbrott hos männen i någon mån är relaterade till gymnasiebetyg. De svaga sambanden med gymnasiebetyg hänger troligen ihop med att variationen i denna variabel var starkt reducerad.

Tabell 3. Korrelationer mellan test, betyg och studieresultat samt avbrott/ej avbrott. N=92.

	Alla		Män		Kvinnor	
	Antal avklarade enheter	Av-brott/ej avbrott	Antal avklarade enheter	Av-brott/ej avbrott	Antal avklarade enheter	Av-brott/ej avbrott
Test/pre-diktor						
1. Uppskatt-ning	0.29	0.08	0.27	0.18	0.35	-0.25
2. Matriser	0.20	-0.08	0.25	-0.06	0.13	-0.06
3. Planlösning	0.20	0.06	0.24	0.05	0.11	0.04
4. Slutsatser	0.16	0.02	0.27	0.12	-0.08	-0.24
5. Talserier	0.17	-0.08	0.21	-0.06	0.06	-0.18
6. Ovanliga användnings-sätt	0.06	-0.15	0.13	-0.15	-0.12	-0.11
7. R22	0.30	-0.06	0.28	0.02	0.34	-0.25
Genomsnitt-liga testpoäng, lika vikter	0.39	-0.04	0.48	0.06	0.23	-0.26
Gymnasie-betyg, genom-snitt	0.10	-0.15	0.07	-0.20	0.17	0.08

Materialet är egentligen för litet för en fullständig multipel regressionsanalys men här är i alla fall resultatet på hela gruppen med test 1-7 som prediktorer av antala avklarade studieenheter. Den multipla korrelationen blev=0.42, en ytterst obetydlig ökning jämfört med prognosvärdet hos den genomsnittliga testpoängen som ju var 0.39, se Tabell 3. Största beta-vikterna fick test 2, 3 och 7. Test 5 fick t o m en negativ vikt.

Ett visst intresse har onekligen kriteriet avbrott. Jag har gjort en diskriminantanalys med avbrott som beroende variabel och de 7 testen som prediktorer. En viss uppfattning av testens värde (en överskattning) kan man få av Tabell 4 som ger resultatet av en korsklassificering av faktisk och predicerad grupptillhörighet m avs på avbrott/icke avbrott.

Tabell 4. Preduktion av avbrott, diskriminantanalys

	Predicerade ej avbrott	Predicerade avbrott
Faktiska ej avbrott	51	25
Faktiska avbrott	5	11

Det kan alltså noteras att 36 avbrott prediceras, av dessa är 25 missar. Samtidigt avbryter 5 som prediceras som icke avbrott. Om man så vill kan man säga att man utifrån testen kunde ha predicerat 11/16 av avbrotten, men till priset av att avvisa 25 studerande som skulle ha fullföljt.

Det är fullt möjligt att avbrottskriteriet är alltför komplext för att det skall särskilt väl kunna prediceras med hjälp av linjära modeller. Det är ju tänkbart att vissa avbrytare kan räknas till gruppen som har svaga studieförutsättningar medan andra kanske avbryter för att de i stället antas på en för dem mera attraktiv studiebana. Materialet är tyvärr för litet för en analys av dessa intressanta frågor, liksom det är för litet för separata multivariata analyser av män och kvinnor.

Det kan ha ett intresse att gå litet djupare in på hur sambanden mellan test, detta fall det genomsnittliga testvärdet, och kriteriet är konstituerat. Jag har till att börja med delat in materialet i fyra under grupper, nämligen:

A. Över genomsnittet i både test och betyg.
B. Över genomsnittet i betyg, under i test.
C. Under genomsnittet i betyg, över i test.
D. Under genomsnittet både i test och betyg.

För var och en av dessa grupper har jag studerat andelen den utgör av totalgruppen samt sannolikheten för ett studieresultat bättre än genomsnittet. Intressant här är att andelen i grupp B är så mycket högre än C. Detta är inte möjligt att förklara enbart som ett utslag av låg korrelation mellan test och betyg, normalt borde dessa andelar vara ungefär lika. Intressant är också tendensen att framgången är sämre i grupp B än i övriga grupper, även om tendensen inte är särskilt stark. En indikation får vi dock om att vissa personer med höga betyg och låga testpoäng inte klarar sig så bra på Handelshögskolan.

Sammanfattning och slutsatser

Från psykometrisk synpunkt är de nuvarande HHS testens egenskaper ofullständigt dokumenterade. Man kan dock anta att mätprecisionen i de flesta fall är tillfredsställande medan validiteten, speciellt för HHS syften, varit så gott som okänd före de studier som presenteras här.

Testen är från 50-talet eller förra hälften av 60-talet, möjligen med undantag för Guilford-testet Ovanliga användningssätt, som härstammar från 60-talets mitt. Jämfört med GMAT är avsaknaden av verbalt laddade test påtaglig. Det är en ganska ensidig dominans för induktiv och spatial begåvning med vissa inslag av mera rutinbetonade funktioner.

Inga försök tycks ha gjorts att anpassa testen för den aktuella gruppen. Man har dock valt test för att försöka differentiera i den högre delen av begåvningsfördelningen. Det kan också noteras att validiteten mot antal avklarade studieenheter är acceptabel, särskilt som den inte korrigerats för begränsad spridning i testen. Det är intressant att se att R22, som är ett *g*-faktortest har bästa validiteten, och att Uppskattningar har nästan lika bra validitet. Kanske var utbildningen på HHS mycket kvantitativt orienterad. Bäst prognos hade dock den samlade testpoängen, som förmodligen var det mest effektiva sättet att mäta *g*-faktorn. Det finns nu mycket omfattande forskning som stödjer användningen av *g*-faktortest [22].

Säkerheten i testproceduren vid HHS synes ej ha varit hög då detta är test som inte torde vara alltför svåra att identifiera och ta del av. Testen ges i exakt samma form och med samma uppgifter varje år. Som vi sett förekommer troligen ganska betydande träningseffekter.

Om man vill utveckla en ny uppsättning av test för urval till HHS anser jag att det är lämpligt att arbeta med ett bredare spektrum av test än vad som nu används. Vissa av de testvariabler som nu används synes av tveksamt värde, jag tänker på Guilfordtestet och vissa mera rutinmässiga uppgifter. Med tanke på att framtidens testning troligen kommer att bli alltmer datoriserad, liksom ekonomutbildningen är på väg att bli det, bör det allvarligt övervägas om inte testningen i sin helhet bör datoriseras så snart som möjligt.

För att undvika träningseffekter bör särskilda åtgärder vidtas, t ex nykonstruktion av åtminstone en del av uppgifterna varje år. Levine och Drasgowe [23] ger en god översikt av metoder för att identifiera sådana testade vars svarsprofiler avviker från profiler som kan förväntas enligt testmodellen. Man kan anta att en del av dessa personer t ex har tränat på att memorera svaret på vissa uppgifter eller av andra skäl besvarat testuppgifterna på andra grunder än de väntade.

Det är angeläget att testkonstruktionen baseras på psykologisk teori och forskning av aktuellt slag och att forskningen följs kontinuerligt så att HHS testning kan tillgodogöra sig den internationella erfarenheten på området, varvid främst utvecklingen inom den kognitiva psykologin och dess tillämpningar inom testningen synes vara av intresse. En utmärkt sammanställning av den kognitiva psykologins möjligheter på detta område ges i en av Sternberg redigerad bok [44]. Sternbergs egen intelligensmodell [23] representerar det hittills kanske mest lovande och intresseväckande försöket att komma vidare på den stig som upptrampats av de första generationernas testpsykologer.

Den traditionella psykometriska testteorin [15] ersätts nu gradvis av modernare stokastiska modeller, se t ex Hulin, Drasgow och Parsons [18]. Här finns mycket att hämta för utveckling av nya och förhoppningsvis bättre test. Datorisering av testning är en utveckling som nu är tillräckligt väl etablerad (i USA) (Green, Bock, Humphreys, Linn & Reckase, 1984) för att ha avkastat en del intressanta forskningsresultat som tyder på att datoriserade test kan göras avsevärt kortare än traditionella test utan förlust av mätprecision [27].

Till slut: vilka förskjutningar har inträffat i perspektivet sedan 1980-talet?

- Personlighetens betydelse är nu mycket bättre dokumenterad
- Begåvningstesten, särskilt i form av *g*-test, har fortsatt att visa sig vara mycket användbara som prediktorer av arbets- och utbildningsresultat
- Nya utvecklingslinjer vid begåvningstestning har däremot inte riktigt levt upp till de förhoppningar man hade

Referenser

[1]. Baer, J. (2011). How divergent thinking tests mislead us: Are the Torrance Tests still relevant in the 21st century? The Division 10 debate. [doi:10.1037/a0025210]. *Psychology of Aesthetics, Creativity, and the Arts, 5*, 309-313.

[2]. Baird, L. L. (1985). Do grades and tests predict adult accomplishment? *Research in Higher Education, 23*, 3-85.

[3]. Benson, G. (1983). *GMAT -- fact or fiction. A look at the validity of the exam* Paper presented at the Annual Meeting of the Rocky Mountains Educational Research Association, Tuscon, AZ, November 2-5.

[4]. Chapman, L. J., & Chapman, J. P. (1967). Genesis of popular but erroneous psychodiagnostic observations. *Journal of Abormal Psychology, 73*, 193-204.

[5]. Chapman, L. J., & Chapman, J. P. (1969). Illusory correlation as an obstacle to the use of valid psychodiagnostic signs. *Journal of Abnormal Psychology, 74*, 271-280.

[6]. Deckro, R. F., & Woundenberg, H. W. (1977). M. B. A. admission criteria and academic success. *Decision Sciences, 8*, 765-769.

[7]. Engelberg, E., & Sjöberg, L. (2005). Emotional intelligence and interpersonal skills. In R. D. Roberts & R. Schulze (Eds.), *International handbook of emotional intelligence* (pp. 289-308). Cambridge MA: Hogrefe.

[8]. Forer, B. R. (1949). The fallacy of personal validation: a classroom demonstration of gullibility. *Journal of Abnormal & Social Psychology, 44*, 118-123.

[9]. Gayle, J. B., & Jones, T. H. (1973). Admission Standards for Graduate Study in Management. *Decision Sciences*421-425.

[10]. Glaser, R. (1981). The future of testing: A research agenda for cognitive psychology and psychometrics. [doi:10.1037/0003-066X.36.9.923]. *American Psychologist, 36*, 923-936.

[11]. Gottfredson, L. S. (2003). Dissecting practical intelligence theory: Its claims and evidence. *Intelligence, 31*, 343-397.

[12]. Gottfredson, L. S. (2003). g, jobs and life. In H. Nyborg (Ed.), *The scientific study of general intelligence: A tribute to Arthur R. Jensen* (pp. 293-342). Oxford, UK: Pergamon.

[13]. Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective,

impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical-statistical controversy. *Psychology, Public Policy, and Law, 2,* 293-323.

[14]. Guilford, J. P. (1967). *The nature of human intelligence*. New York, NY, US: McGraw-Hill.

[15]. Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley.

[16]. Henrysson, S., Lexelius, A., & Kriström, M. (1984). *Meritvärdering och studieprognos: några undersökningar av antagningssystemets effekter*. Stockholm: UHÄ.

[17]. Hogarth, R. M. (1981). Beyond discrete biases: functional and dysfunctional aspects of judgmental heuristics. *Psychological Bulletin, 90,* 191-217.

[18]. Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory: Application to psychological measurement*. Homewood, IL: Dow Jones-Irwin.

[19]. Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96,* 72-98.

[20]. Klimoski, R. J., & Rafaeli, A. (1983). Inferring personal qualities through handwriting analysis. *Journal of Occupational Psychology, 56,* 191-202.

[21]. Kogan, N., & Pankove, E. (1974). Long-term Predictive Validity of Divergent-thinking Tests: Some Negative Evidence. [doi:10.1037/h0021521]. *Journal of Educational Psychology, 66,* 802-810.

[22]. Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2004). Academic Performance, Career Potential, Creativity, and Job Performance: Can One Construct Predict Them All? [doi:10.1037/0022-3514.86.1.148]. *Journal of Personality and Social Psychology, 86,* 148-161.

[23]. Levine, M. V., & Drasgow, F. (1982). Appropriateness measurement: Review, critique and validating studies. [doi:10.1111/j.2044-8317.1982.tb00640.x]. *British Journal of Mathematical and Statistical Psychology, 35,* 42-56.

[24]. Maccoby, E. E., & Jacklin, C. N. (1974). *The psychology of sex differences* (Vol. 1). Stanford, CA: Stanford University Press.

[25]. Mischel, W. (1968). *Personality and assessment*. New York: Wiley.

[26]. Mischel, W., & Peake, P. K. (1982). Beyond deja vu in the search for cross-situational consistency. *Psychological Review, 89,* 730-755.

[27]. Moreno, K. E., Wetzel, C. D., McBride, J. R., & Weiss, D. J. (1984). Relationship between corresponding armed services vocational aptitude battery (ASVAB) and computerized adaptive testing (CAT) subtests. *Applied Psychological Measurement, 8,* 155-163.

[28]. Paolillo, J. G. (1982). The predictive validity of selected admissions variables relative to grade point average earned in a Master of Business Administration Program. [doi:10.1177/001316448204200423]. *Educational and Psychological Measurement, 42*, 1163-1167.

[29]. Remus, W., & Wong, C. (1982). An evaluation of five models for the admission decision. *College Student Journal, 16*, 53-59.

[30]. Roberts, B. W., Kuncel, N. R., Shiner, R., Caspi, A., & Goldberg, L. R. (2007). The power of personality: The comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. *Perspectives on Psychological Science, 2*, 313-345.

[31]. Salgado, J. F., Anderson, N., Moscoso, S., Bertua, C., de Fruyt, F., & Rolland, J. P. (2003). A Meta-Analytic Study of General Mental Ability Validity for Different Occupations in the European Community. [doi:10.1037/0021-9010.88.6.1068]. *Journal of Applied Psychology, 88*, 1068-1081.

[32]. Schmidt, F. L., & Hunter, J. (2004). General Mental Ability in the World of Work: Occupational Attainment and Job Performance. [doi:10.1037/0022-3514.86.1.162]. *Journal of Personality and Social Psychology, 86*, 162-173.

[33]. Schmidt, F. L., Hunter, J. E., & Pearlman, K. (1981). Task differences as moderators of aptitude test validity in selection: A red herring. [doi:10.1037/0021-9010.66.2.166]. *Journal of Applied Psychology, 66*, 166-185.

[34]. Schmitt, N., Gooding, R. Z., Noe, R. A., & Kirsch, M. (1984). Metaanalyses of validity studies published between 1964 and 1982 and the investigation of study characteristics. [doi:10.1111/j.1744-6570.1984.tb00519.x]. *Personnel Psychology, 37*, 407-422.

[35]. Sjöberg, L. (1981). Värdet av DMT vid urval av flygförare. *Nordisk Psykologi, 33*, 241-248.

[36]. Sjöberg, L. (1982). Test, betyg och urval. In L. Sjöberg (Ed.), *Upplevelse och prestation. Några utvecklingslinjer inom teoretisk psykologi* (pp. 83-125). Lund: Doxa.

[37]. Sjöberg, L. (2010). *Personlighetsdimensioners validitet i arbetslivet: teorier och empiri* (SSE/EFI Working Paper Series in Business Administration No. 2010:6). Stockholm: Stockholm School of Economics.

[38]. Sjöberg, L., Bergman, D., Lornudd, C., & Sandahl, C. (2011). *Sambandet mellan ett personlighetstest och 360-graders bedömningar av chefer i hälso- och sjukvården*. Stockholm: Karolinska Institutet, Institutionen för lärande, informatik, management och etik (LIME).

[39]. Sjöberg, L., & Tollgerdt-Andersson, I. (1985). *Vad är personkemi? Socialpsykologisk forskning*

om attraktivitet. (What is person chemistry? Social psychological research on attractiveness). Stockholm: Scandinavian Executive Search.

[40]. Smedslund, J. (1963). The concept of correlation in adults. *Scandinavian Journal of Psychology, 4*, 165-173.

[41]. Statens offentliga utredningar. (1985). *Prov för urval till högskolan, Nr. 59.*

[42]. Statens offentliga utredningar. (1985). *Tillträde till högskolan, Nr. 57.*

[43]. Sternberg, R. J. (Ed.). (1982). *Handbook of human intelligence.* Cambridge: Cambridge University Press.

[44]. Sternberg, R. J. (Ed.). (1985). *Human abilities. An information processing approach.* New York: Freeman.

[45]. Taylor, H. C., & Russell, J. T. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection: Discussion and tables. *Journal of Applied Psychology, 23*, 565-578.

[46]. Wightman, L. E., & F., L. L. (1985). *GMAC validity study service: A three-year summary*. Princeton, NJ: Graduate Management Admission Council and Educational Testing Service.